Enhancing Long-and Short-Term Representations for Next POI Recommendations via Frequency and Hierarchical Contrastive Learning

Jiajie Chen[1*], Yu Sang[2*], Peng-Fei Zhang[3], Jiaan Wang[1], Jianfeng Qu[1, 6†], Zhixu Li[4, 5]

[1]School of Computer Science and Technology, Soochow University, Suzhou, China
[2]School of Artificial Intelligence and Computer Science, Jiangnan University
[3]School of Electrical Engineering and Computer Science, University of Queensland
[4]School of Information, Renmin University of China, Beijing, China
[5]International College (Suzhou Research Institute), Renmin University of China, Suzhou, China
[6]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University
{jjchen.ada, mima.zpf, jawang.nlp}@gmail.com, 7213107001@stu.jiangnan.edu.cn, jfqu@suda.edu.cn, zhixuli@ruc.edu.cn

Abstract

Next POI recommendation aids users in predicting their destinations of interest and plays an increasingly vital role in location-based social services. Recent works focus on analyzing both long-term and short-term interests in POI recommendation to gain a deeper understanding of user profiles. However, these methods for modeling long-term user's sequences primarily rely on the Transformer model, which functions as a low-pass filter, often leading to the loss of high-frequency information. Additionally, long-term and short-term sequences are typically modeled independently, with short-term sequences often defined solely by the most recent check-ins, overlooking their interactions and dependencies. Therefore, we propose Enhancing Long-and Short-Term Representations for Next POI Recommendations via **F**requency and **H**ierarchical **C**ontrastive Learning (FHCRec). FHCRec captures both high-frequency and low-frequency information in long-term sequences to model richer long-term user's preference representations. Moreover, it harnesses the characteristics of the short-term subsequences embedded within long-term sequences to enhance short-term preference characterization via local and global hierarchical contrastive learning, resulting in more personalized short-term preferences. The enhanced long-term and short-term preferences are integrated to improve model recommendation performance. Extensive experiments on three real-world datasets demonstrate the effectiveness of our method.

Introduction

With the rapid advancement of location-based social media services, vast amounts of data on people's interactions with nearby places have been recorded (Li et al. 2022). This valuable information has significantly advanced the research on POI recommendations, providing convenience to both users and businesses. Typically, POI recommendation methods aim to model a user's historical check-in records and preferences to make accurate predictions about they current check-in behavior. Early methods focused on single-sequence modeling, utilizing techniques such as Markov chain-based stochastic models (Ye, Zhu, and Cheng 2013), matrix factorization models (Lian et al. 2014), and RNN models (Liu et al. 2016; Yang et al. 2020). These approaches have been extensively explored and yielded notable results.

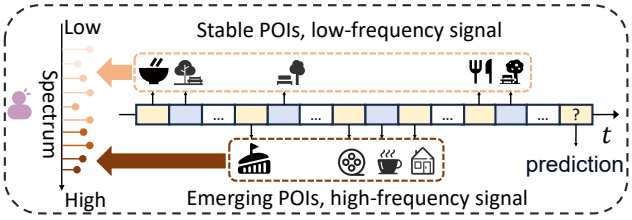


Figure 1: Frequency signal characterization of long-term POI sequence

Recent studies have transitioned from single-sequence modeling to a paradigm that incorporates both long-term and short-term sequence modeling. Long-term sequence modeling captures a user's overall historical travel preferences, while short-term sequence modeling reflects they recent behavioral habits. By considering both aspects, the impact of recent noise and outdated preferences is mitigated, thereby enhancing recommendation accuracy (Zhao et al. 2018).

For **long-term sequence modeling**, SLi-Rec (Yu et al. 2019) enhances traditional RNNs with time-aware and content-aware controllers to incorporate contextual information from long-term sequences. STGN (Zhao et al. 2020) introduces spatiotemporal gates in RNNs to update long-term interests. Compared to RNNs, Transformer models possess stronger long-range modeling capabilities and better comprehension of the global context, making them popular for modeling long-term user sequence. PLSPL and CLSR (Wu et al. 2020; Zheng et al. 2022) both employ self-attention mechanisms to capture user preferences in long-term modules. Additionally, CFPRec and CLSPRec (Zhang et al. 2022a; Duan et al. 2023a) use bidirectional Transformer to incorporate sequence contextual relationships, offering a more accurate model of past long-term sequences. However, Transformer-based methods suffer from inherent over-

*These authors contributed equally.
†Corresponding author.

smoothing (Shin et al. 2024), where increasing the number of model layers causes feature representations to become overly similar. While these less distinguishable features effectively capture low-frequency information, they may hinder the model's ability to capture high-frequency information. As illustrated in Figure 1, a user's check-in sequence indicates a preference for going to the park after eating during a specific time period. This relatively stable overall preference is reflected as a low-frequency signal in the frequency-domain spectrum. However, the user's behavior might vary, such as going to the park after exercising or choosing to watch a movie, drink coffee and go home during a similar period of time. These rapidly changes or emerging preferences represent high-frequency information. Thus, users' long-term preferences are complex and diverse, rather than static. Optimal representation modeling requires accounting for both low-frequency and high-frequency information in the user's long-term preferences.

As for **short-term sequence modeling**, SLi-Rec (Yu et al. 2019) identifies recent check-in behaviors as short-term sequences, while CLSR (Zheng et al. 2022) emphasizes the impact of the most recent interaction. Both methods ultimately represent the user's overall interest by linearly combining long and short-term preferences. STGN (Zhao et al. 2020) uses a time gate for modeling long-and short-term interests and a distance gate to differentiate them. CLSPRec (Duan et al. 2023a) utilizes a shared encoder to process both long-and short-term sequences, attempting to concatenate these preferences for the final prediction. However, these approaches often focus solely on the most recent check-in sequences as the short-term sequence and model long- and short-term sequences separately, neglecting their interaction. Since long-term sequences inherently contain short-term subsequences, which offer insights into the intrinsic and potential relationships between short-term interests. Ignoring the diverse short-term interest patterns within long-term sequence and focusing only on the most recent short-term sequences results in a narrow feature representation.

To address above challenges, we propose the FHCRec model: Enhancing Long-and Short-Term Representations for Next POI Recommendations via Frequency and Hierarchical Contrastive Learning. For long-term sequence modeling, we employ a bidirectional Transformer enhanced with frequency-domain exploration. Both low- and high-frequency information in long-term sequence are captured and adaptively integrated to relatively balance stable and dynamic interests. This approach enriches the user's long-term preference representation. For short-term sequence modeling, rather than focusing solely on the current short-term sequence as previous methods do, we segment the long-term sequence within the same sample into meaningful short-term subsequences based on dates. Historical short-term patterns embedded within long-term sequence are encoded to derive local representations, which are then aggregated to form a global sequence representation. Both local and global representations are further optimized using hierarchical contrastive learning, enhancing the accuracy of short-term preference characterization and resulting in a more personalized representations of short-term interests. Finally, the enhanced long-term and short-term interest representations are integrated to predict the user's next POI. Our experiments validate the model's recommendation capabilities and the effectiveness of its components. The main contributions are summarized as follows:

- Our work explores the issue of over-smoothing in long-term sequence modeling and the simplistic definition of short-term sequence that results in insufficient interaction between long-and short-term sequences.
- To the best of our knowledge, this is the first study in the POI recommendation to leverage frequency-domain information to enhance long-term sequence representation. We also account for the short-term subsequences embedded within long-term sequences and design hierarchical contrastive learning from both local and global perspectives, thus improving the representation of short-term preferences.
- Our proposed model significantly outperforms state-of-the-art sequential POI recommendation methods across three real-world datasets, demonstrating its effectiveness.

Preliminaries

Problem Formulation

Let $\mathcal{U} = \{u_1, u_2, \cdots, u_M\}$ represent a set of users, $\mathcal{P} = \{p_1, p_2, \cdots, p_N\}$ represent a set of POIs, where M, N are the total number of users and POIs, respectively. A check-in record $r = (u, p, t)$ means that user u visited POI p at time t. We follow the method (Zhang et al. 2022a; Duan et al. 2023a) to divide each user's sequences and redefine it as:

Definition 1: Short-term sequence. According to time t, the user's check-in records are divided into days. We can naturally define the short-term sequence as the check-in records of a certain day $S_{Short} = \{r_1, r_2, r_3 \dots r_{|S_{Short}|}\}$.

Definition 2: Valid check-in sequence sample. A valid check-in sequence is a combination of consecutive short-term sequences $S = \{S_{short\_1}, S_{short\_2}, \dots, S_{short\_N}\}$.

Definition 3: Current short-term sequence. A current short-term sequence is closely related to the POI we want to predict and distinguished from other short-term sequence $S_{Current} = S_{short\_N}$.

Definition 4: Long-term sequence. A long-term sequence is a coherent check-in record in the valid check-in sequence that does not include the current short-term sequence $S = \{S_{short\_1} \oplus S_{short\_2} \oplus \dots \oplus S_{short\_(N-1)}\}$.

1-D Discrete Fourier Transform

The Fourier Transform is a mathematical tool that converts a signal from the time domain to the frequency-domain. For discrete signals, the Discrete Fourier transform (DFT) is commonly applied. In practical applications, to simplify calculations and reduce redundancy, the Fast Fourier Transform (FFT) is typically employed to implement the DFT. For complex signals, let $x[n]$ represent the original discrete signal with a length of N. We use orthogonal normalization to

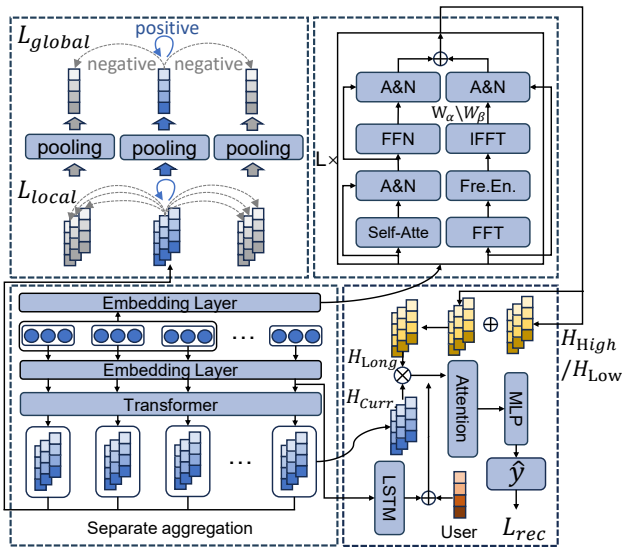


Figure 2: Framework of our proposed model. A&N stands for Add&Norm, and Fre.En stands for Frequency Enhanced

convert the signal to the frequency-domain:

$$X[k] = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x[n] e^{-i\frac{2\pi}{N}kn}, \quad k = 0, 1, \ldots, N-1. \tag{1}$$

The signal can also be converted back to the time domain using an Inverse Fast Fourier Transform (IFFT):

$$x[n] = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} X[k] e^{i\frac{2\pi}{N}kn}, \quad n = 0, 1, \ldots, N-1. \tag{2}$$

Orthogonal normalization ensures the preservation of signal energy, making it independent of signal length:

$$x[n] = \text{IFFT}(\text{FFT}(x[n])). \tag{3}$$

Method

In this section, we thoroughly present our model diagram. The detailed model diagram is shown in Figure 2.

Embedding Layer

We use a simple embedding layer to obtain embeddings, given the input user check-in sequence $S_i^u = \left\{r_1, r_2 \ldots r_{|S_i^u|}\right\}$, it is encoded as follows:

$$\mathbf{e}_r = \mathbf{u} \oplus \mathbf{l}, \tag{4}$$

where $\mathbf{e}_r \in \mathbf{R}^D$; $\oplus$ represents the concatenation operation; $\mathbf{u}, \mathbf{l} \in \mathbf{R}^{\frac{D}{2}}$ represents the embedded representation of the user and POI. Therefore, the check-in sequence is embedded as $\mathbf{E}_{S_i^u} = \left[\mathbf{e}_{r_1}, \mathbf{e}_{r_2}, \mathbf{e}_{r_3} \cdots, \mathbf{e}_{r_{|S_i^u|}}\right]$. We will abbreviate all the embedding matrices as $\mathbf{E}$.

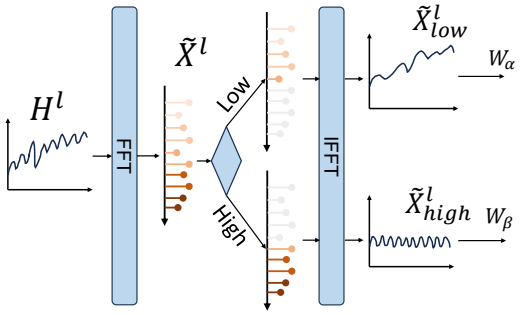


Figure 3: Frequency enhancement. Gray indicates removed information

Long-Term Modeling with Frequency Enhancement

We apply our frequency-domain enhanced Transformer encoder to the representation of long-term sequence based on the embedding layer. Each layer of the encoder contains a frequency enhancement step, which works in parallel with the standard self-attention sub-layer. This design aims to use frequency-domain information to supplement the information of the self-attention mechanism in the time domain to alleviate over-smoothing. We will enhance the representation of long-term sequence from low-frequency and high-frequency layers respectively.

Given an input sequence representation matrix $\mathbf{H}^l \in \mathbb{R}^{N \times D}$ for the l-th layer and the input of layer 0 is $\mathbf{H}^0 = \mathbf{E}^0$, the matrix is then fed into the encoder.

Self-Attention Sub-Layer. This sub-layer is a standard Transformer self-attention mechanism that captures global dependencies by weighted summing over different positions in the input sequence. After projection of the embedding $\mathbf{H}^l$, we obtain the query $\mathbf{Q}^l \in \mathbb{R}^{N \times D}$, key $\mathbf{K}^l \in \mathbb{R}^{N \times D}$, value $\mathbf{V}^l \in \mathbb{R}^{N \times D}$, and perform the attention calculation:

$$\text{Attention } (\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{D}}\right)\mathbf{V}. \tag{5}$$

The final output after the multi-head attention mechanism is:

$$\begin{aligned} &\text{MultiHead } (\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \\ &\text{Concat } (\text{ head }_1, \ldots, \text{ head }_h)\, \mathcal{W}_O \\ &\text{where head }_i = \text{ Attention } (\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i)\,. \end{aligned} \tag{6}$$

We also retain the dropout, residual and normalization process, the final representation as:

$$\begin{aligned} \tilde{\mathbf{X}}^l_{Att} &= \text{MultiHead}\left(\mathbf{Q}^l, \mathbf{K}^l, \mathbf{V}^l\right), \\ \mathbf{H}^l_{Att} &= \text{LayerNorm}\left(\tilde{\mathbf{X}}^l_{Att} + \text{Dropout}\left(\mathbf{H}^l\right)\right). \end{aligned} \tag{7}$$

Frequency-Domain Enhancement. This sub-layer captures the frequency-domain information of the input sequence, enhancing the output of the self-attention sub-layer.

First, for a given input matrix $\mathbf{H}^l$ in Figure 3, we perform a Fourier Transform operation on the item dimension level to convert the input matrix into the frequency-domain:

$$\tilde{\mathbf{X}}^l = \text{FFT}(\mathbf{H}^l) \in \mathbb{C}^{M \times D}. \tag{8}$$

Due to the conjugate symmetry in the frequency-domain, the length of the spectrum M is:

$$M = \lceil N/2 \rceil + 1. \tag{9}$$

Since the length of the long-term sequence of different samples is uncertain, we determine the high and low frequency thresholds δ according to the ratio:

$$\delta = \frac{N}{d}, \tag{10}$$

where d is a hyperparameter that can be adjusted. When we want to perform low-frequency enhancement, we retain the low-frequency information in the spectrum:

$$\tilde{\mathbf{X}}^l_{low} = \tilde{\mathbf{X}}^l[:\delta, :]. \tag{11}$$

Afterwards, we perform IFFT on it and use learnable weight parameters to adaptively adjust the amount of frequency-domain information:

$$\tilde{\mathbf{H}}^l_{\text{low}} = \mathbf{W}_\alpha \cdot \text{IFFT}\left(\tilde{\mathbf{X}}^l_{\text{low}}\right). \tag{12}$$

The extracted frequency information is added to the output of the attention sub-layer to finally obtain the low-frequency enhanced output of the l-th layer:

$$\mathbf{H}^{l+1} = \mathbf{H}^l_{Att} + \tilde{\mathbf{H}}^l_{\text{low}}. \tag{13}$$

After l layers, the final low-frequency enhanced transformer long-term hidden state is expressed as:

$$\mathbf{H}_{Low} = \mathbf{H}^l. \tag{14}$$

We input the sample into the encoder twice, once for low-frequency enhancement and once for high-frequency enhancement. The high-frequency enhancement is the same, and the eq (11) and eq (12) is changed to:

$$\begin{aligned} \tilde{\mathbf{X}}^l_{high} &= \tilde{\mathbf{X}}^l[\delta:, :], \\ \tilde{\mathbf{H}}^l_{\text{high}} &= \mathbf{W}_\beta \cdot \text{IFFT}\left(\tilde{\mathbf{X}}^l_{\text{high}}\right). \end{aligned} \tag{15}$$

After normalization, the final high-frequency enhanced representation is recorded as $\mathbf{H}_{High}$. As we have adapted the low and high frequencies separately in the encoder, the low- and high-frequency enhanced representations are directly incorporated as the representation of the long-term sequence:

$$\mathbf{H}_{Long} = \mathbf{H}_{High} + \mathbf{H}_{Low}. \tag{16}$$

Contrastive Strategies for Short-Term Modeling

Because short sequences lack sufficient frequency-domain information, we do not apply frequency-domain enhancement to them. Instead, we enhance the representational capacity of short-term sequences by applying hierarchical contrastive learning to local and global objectives.

Positive and negative sample selection. For the encoder, we use the standard transformer, which is abbreviated as $\text{Trans}_\theta(\cdot)$. Given a user check-in sequence $S^u = \{S^u_1, S^u_2, \ldots, S^u_{N-1}, S^u_N\}$, we omit the subscript $short$. For the selection of positive samples, we enhance the check-in sequence from the semantic aspect and use minimal data enhancement dropout. In most cases, different users have unique check-in activities on various dates. Therefore, we select different users as negative samples and segment the sequences by date to enrich the short-term sequence samples. We approach the contrastive learning tasks from both local and global perspectives in a hierarchical manner, enhancing the encoder's capacity to model short-term sequence representations. This improvement leads to better short-term sequence representations, resulting in superior recommendation results. And the negative sample sequence is represented as $S^{u_1}, S^{u_2}, \ldots, S^{u_k}$.

Local Contrastive Learning. We utilize the standard Transformer model to represent short-term sequences: $h^u_i = \text{Trans}_\theta(S^u_i)$, the positive sample is the representation enhanced by dropout $h^{u^+}_i = \text{Trans}_{\theta^+}(S^u_i)$. The goal of local contrastive learning is to refine the personalized representation of user's short-term sequences within the same day and to distinguish the representations of different short-term sequences across various users:

$$\mathcal{L}_{local} = -\log \frac{e^{\text{sim}\left(\mathbf{h}^u_i, \mathbf{h}^{u^+}_i\right)/\tau}}{e^{\text{sim}\left(\mathbf{h}^u_i, \mathbf{h}^{u^+}_i\right)/\tau} + \sum_{u_k \neq u} e^{\text{sim}\left(\mathbf{h}^u_i, \mathbf{h}^{u_k}_*\right)/\tau}}. \tag{17}$$

Global Contrastive Learning. We aggregate the short-term sequences of users to represent h^u, h^{u^+}, h^{u^k}, respectively. The objective of global contrastive learning is to improve the comprehensive representation of users' short-term interest preferences over an extended period. During the aggregation process, we perform weighted aggregation based on the length of different short-term sequences.

$$\mathcal{L}_{global} = -\log \frac{e^{\text{sim}\left(\mathbf{h}^u, \mathbf{h}^{u^+}\right)/\tau}}{e^{\text{sim}\left(\mathbf{h}^u, \mathbf{h}^{u^+}\right)/\tau} + \sum_{u_k \neq u} e^{\text{sim}\left(\mathbf{h}^u, \mathbf{h}^{u^k}\right)/\tau}}. \tag{18}$$

Joint Loss. We use the hyperparameter α to balance the strengths of the global and local contrastive learning losses:

$$\mathcal{L}_{CL} = \alpha \cdot \mathcal{L}_{local} + (1-\alpha) \cdot \mathcal{L}_{global}. \tag{19}$$

Model Training

We retain the user enhancement mechanism from CLSPRec to capture the users' current state using an LSTM network. This state is then combined with the user representation to obtain an enhanced user representation:

$$\mathbf{u} = \text{Embeddinglayer}(u) + M \cdot \text{Mean}\left(\text{LSTM}\left(S^u_{\text{Current}}\right)\right), \tag{20}$$

where M is a learnable parameter.

In the previous section, we derived a frequency-enhanced long-term interest representation. Simultaneously, hierarchical contrastive learning was employed to enhance the encoder's ability to capture short-term interests, ensuring accurate modeling of the user's current short-term sequence representation $\mathbf{H}_{Current} = \text{Trans}_\theta(S^u_{Current})$. We then derive the user's preference representation $\mathbf{H}_{S^u} = [\mathbf{H}_{Long}, \mathbf{H}_{Current}]$ and use an attention mechanism to de-

rive the user's final preference:

$$\begin{aligned} \mathbf{h}^u &= \sum_{i=1}^{|H|} \omega_i \mathbf{h}_i, \mathbf{h}_i \in \mathbf{H}_{S^u}, \\ \omega_i &= \frac{\exp\left(\mathbf{u}^\top \mathbf{h}_i\right)}{\sum_{i'=1}^{|H|} \exp\left(\mathbf{u}^\top \mathbf{h}_{i'}\right)}, \end{aligned} \tag{21}$$

where ω_i is computed by the attention score between the user embedding $\mathbf{u}$ and the user's various hidden states $\mathbf{h}_i$. The final user preference is derived by applying softmax to $\mathbf{h}^u$ to obtain the probability distribution over POIs:

$$\hat{y} = \operatorname{softmax}\left(\mathbf{h}^u \mathbf{W}\right), \tag{22}$$

where $\hat{y}$ represents the probability distribution of the POI to be predicted, and $\mathbf{W}$ is a learnable matrix. Cross-entropy is used as the loss function for POI prediction:

$$\mathcal{L}_{\text{poi}} = -\sum_{i \in \mathcal{N}} \log\left(\hat{y}_i\right), \tag{23}$$

where $\mathcal{N}$ denotes the training sample set, and $\hat{y}_i$ represents the predicted probability of the ground truth POI for the i-th training sample in $\mathcal{N}$.

Final loss: We employ a multi-task learning strategy, recommendation loss is combined with the contrastive learning loss to compute the final loss:

$$\mathcal{L} = \mathcal{L}_{\text{poi}} + \mathcal{L}_{\text{CL}}. \tag{24}$$

Experiments

Datasets

We utilize three datasets from Foursquare, i.e., Singapore (SIN), New York City (NYC) and Phoenix (PHO), as detailed in Table 2. Interaction records are sorted based on the user's check-in times. Following the methodology in (Zhang et al. 2022a; Duan et al. 2023a), we filter out POIs with fewer than 10 occurrences, remove inactive users with fewer than 5 check-ins, and discard trajectories shorter than 3 check-ins. The sorted interaction records are divided into training, validation, and test sets in a ratio of 8: 1: 1.

Baselines

Nine baseline models are selected for comparison. (1) **ST-RNN** (Liu et al. 2016) models specific spatio-temporal context distances in check-in sequences using an RNN. (2) **ATST-LSTM** (Huang et al. 2019) assigns different weights to historical check-ins in a spatiotemporal LSTM using an attention mechanism. (3) **MCARNN** (Liao et al. 2018) employs a multi-task learning framework with RNN to predict the next user activity and location. (4) **PLSPL** (Wu et al. 2020) combines an attention mechanism and LSTM to learn user's past and current preferences. (5) **iMTL** (Zhang et al. 2021) captures relationships between user's activities and locations using a dual-channel encoder and specific decoder. (6) **CTLE** (Lin et al. 2021) models contextual position embeddings using a bidirectional Transformer model. (7) **CFPRec** (Zhang et al. 2022a) uses an attention mechanism and LSTM to model user's past, current, and future preferences separately. (8) **ContraRe**c (Zhang et al. 2022b) is a SOTA sequence-based method that performs contextual contrastive learning tasks on sequences of the same target item. (9) **CLSPRec** (Duan et al. 2023a) is a SOTA model that uses a shared trajectory encoder to contrast long and short-term representations for the next POI recommendation task.

Model Settings: Our method is implemented in PyTorch on a cluster server equipped with V100/A100 GPUs. The final performance is determined by averaging the results from three rounds of experiments. For the PHO, NYC, and SIN datasets, the embedding dimensions D are set to 120, 80, and 120, respectively. The number of negative samples for contrastive learning is set to 5 for all datasets. A dropout rate of 0.5 is applied in the Transformer encoder to generate positive samples. The contrast learning temperature τ is set to 0.2. Finally, the Adam optimizer is used with a learning rate of 1e-4. The hyperparameter settings will be analyzed in the following sections.

Evaluation Metrics: To validate the performance of our method, we employed two widely recognized metrics in our experiments: (1)HR@K and (2)NDCG@K. We report our performance with K set to 5 and 10 for a fair comparison.

Overall Performance

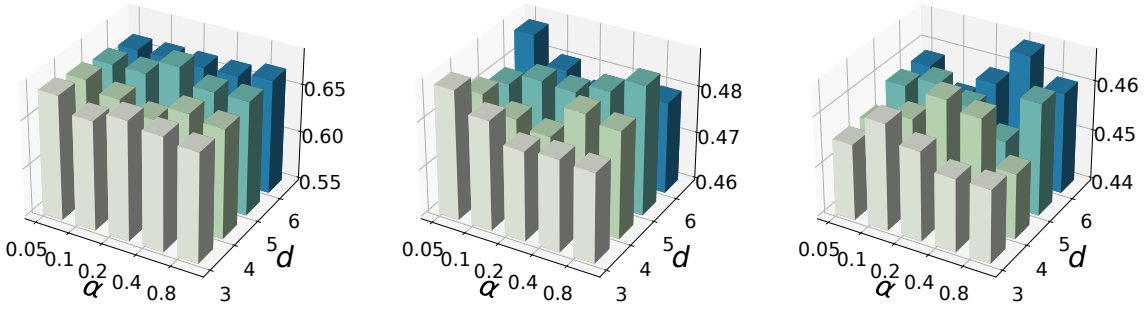
We compare the proposed model with baseline models across three datasets. Table 1 presents the best performance of all models on the three datasets. Our model demonstrates significant performance improvements across all datasets. These experiments validate the effectiveness of our proposed model for the next POI ecommendation task, with improvements ranging from 7% to 21.6% across the three datasets. Among traditional sequence-based methods, ATST-LSTM and PLSPL outperform ST-RNN, indicating the effectiveness of attention mechanisms in sequence modeling. The success of MCARNN and iMTL highlights the importance of considering user activities and locations in recommendation system. CTLE, CFPRec, and CLSPRec significantly outperform traditional sequence modeling methods, demonstrating the bidirectional Transformer model's capability in long sequence modeling. However, Transformer-based models tend to produce oversmoothed item representations, leading to performance that is inferior to our model. ContraRec and CLSPRec both utilize contrastive learning, resulting in significant performance improvements. but they primarily focus on modeling global contrast targets and tend to overlook local information. Our FHCRec model enhances the bidirectional Transformer model by capturing frequency-domain information in long-term sequence modeling to enrich embedding representations. It also considers both global and local short-term sequence representations through hierarchical contrastive learning, thus achieving the best results.

Ablation Study

To verify the effectiveness of our proposed model, we conduct ablation experiments by removing different components to create the following variants: (V1) The frequency-enhanced Transformer encoder is replaced with a standard

	PHO				NYC				SIN			
	H@5	H@10	N@5	N@10	H@5	H@10	N@5	N@10	H@5	H@10	N@5	N@10
ST-RNN	0.1240	0.2028	0.0802	0.1229	0.1347	0.1826	0.0593	0.1303	0.0959	0.1370	0.0655	0.0794
ATST-LSTM	0.1579	0.2377	0.1033	0.1385	0.1667	0.2031	0.0912	0.1638	0.1296	0.1933	0.1027	0.1476
MCARNN	0.1905	0.2726	0.1264	0.1617	0.1835	0.2397	0.1036	0.1870	0.1608	0.1862	0.1169	0.1591
PLSPL	0.1775	0.2569	0.1285	0.1538	0.1741	0.2413	0.0961	0.1825	0.1447	0.1719	0.1126	0.1384
iMTL	0.1830	0.2747	0.1301	0.1632	0.1789	0.2422	0.0989	0.1861	0.1505	0.1801	0.1051	0.1423
CTLE	0.2632	0.3605	0.1995	0.2068	0.2421	0.3205	0.1513	0.1841	0.2041	0.2784	0.1315	0.1556
CFPRec	0.3421	0.4253	0.2432	0.2730	0.2771	0.3606	0.1971	0.2190	0.2310	0.3085	0.1588	0.1836
ContraRec	0.3381	0.3680	0.2843	0.2939	0.1951	0.2368	0.1425	0.1560	0.2047	0.2710	0.1454	0.1660
CLSPRec	0.5368	0.6368	0.3811	0.4175	0.3545	0.4352	0.2653	0.2871	0.3544	0.4093	0.2794	0.2942
FHCRec	**0.6000**	**0.6842**	**0.4257**	**0.4481**	**0.3917**	**0.4658**	**0.2877**	**0.3099**	**0.4219**	**0.4873**	**0.3369**	**0.3562**
Improve	11.8%	7.4%	11.7%	7.3%	10.5%	7.0%	8.4%	7.9%	19.0%	19.1%	20.6%	21.1%

Table 1: Performance comparison in HR@K and NDCG@K on three datasets, "H" stands for HR, and "N" stands for NDCG.



(a) PHO (b) SIN (c) NYC

Figure 4: Hyperparameter sensitivity analysis on HR@10

	#User	#POI	#Check-in	#Density
PHO	2946	7247	47980	0.22%
NYC	16387	56252	511431	0.06%
SIN	8648	33712	355337	0.12%

Table 2: Datasets statistics

	PHO		NYC		SIN	
	H@5	N@10	H@5	N@10	H@5	N@10
V	**0.6000**	**0.4481**	**0.3917**	**0.3099**	**0.4219**	**0.3562**
V1	0.5579	0.4454	0.3856	0.2937	0.4051	0.3561
V2	0.5737	0.4428	0.3784	0.3061	0.4145	0.3552
V3	0.5789	0.4421	0.3773	0.3068	0.4119	0.3539
V4	0.5526	0.4411	0.3706	0.3054	0.4093	0.3549

Table 3: Ablation experiments on three datasets

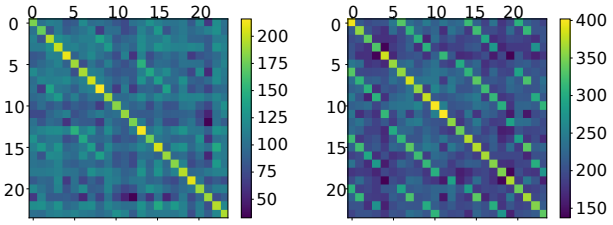
encoder to model long-term sequences. (V2) Global contrastive learning is excluded. (V3) Local contrastive learning is excluded. (V4) The contrastive learning module is entirely removed. The HR@5 and NDCG@10 results are reported on three datasets, as shown in the table 3. From the results, we can first observe that our proposed model significantly outperforms the variants. Compared to V1, the results indicate the frequency enhanced encoder effectively supplements frequency-domain information, resulting in more robust embedding representations. Additionally, our model surpasses V4, indicating that contrastive learning effectively learns better short-term sequence representations. Finally, by comparing V2, V3, and V4, we observe that relying solely on either local or global contrastive learning cannot achieve optimal performance. A proper combination of both approaches offers greater advantages.

Hyperparameter Sensitivity Analysis

We discuss two critical hyperparameters in the model. We search for d, which controls the frequency threshold, within the range $\{3, 4, 5, 6\}$ and α, which controls the strength of local and global contrast learning, within the range $\{0.05, 0.1, 0.2, 0.4, 0.8\}$. We exhaustively explore all combinations across the three datasets to obtain the best performance, and the final results are presented in Figure 4. In practical experiments, we set d and α to 5 and 0.8 for the SIN, 6 and 0.4 for the NYC, and 5 and 0.2 for the PHO.

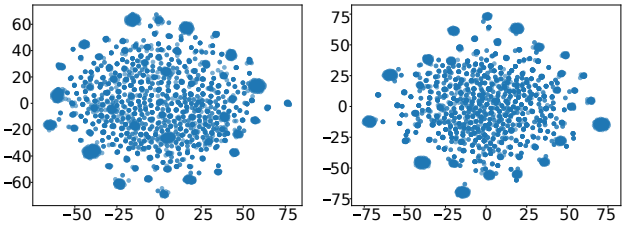
Visualization

Visualization of Long-Term Sequence Representation. We extract a sequence of user check-ins from the SIN dataset and compare the similarity of check-in embed-



(a) No enhancement (b) With enhancement

Figure 5: Similarity matrix heatmap



(a) No CL (b) With CL

Figure 6: Visualization of current short-term sequence representation

dings using the traditional Transformer and the frequency-domain enhanced Transformer. We visualize the results as a heat map Figure 5 and observe that the traditional Transformer exhibits consistent color changes suggesting an over-smoothing problem, which restricts the model's expressive capacity. In contrast, the similarity matrix after frequency-domain enhancement display more pronounced changes between adjacent check-ins, indicating the embeddings are more discriminative. This demonstrates the effectiveness of the frequency-domain enhanced Transformer.

Visualization of Current Sequence Representation. Contrastive learning has been shown to make the distribution of embedded representations more uniform and tightly clustered, thereby enhancing performance in various deep learning tasks (Wang and Isola 2020; Li et al. 2020; Wu, Xiao, and Vydiswaran 2023). To demonstrate the effectiveness of our proposed contrastive learning, as shown in Figure 6, we use t-SNE (Van der Maaten and Hinton 2008) dimensionality reduction to visualize the learned current short-term embeddings from NYC dataset, which are closely related to the POIs to be predicted. The final visualization results indicate that the embedding distribution learned through contrastive learning is denser and more centralized. Additionally, the small-scale clustering effect is more pronounced, which is beneficial for deep learning tasks.

Related Work

Early POI recommendation methods primarily focus on the influence of recent check-in behavior on subsequent visits. Traditional models, such as the Markov random model (Ye, Zhu, and Cheng 2013), predicts the category of the user's next check-in based on category distribution, the predict the final location. The collaborative filtering model (Lian et al. 2014) employs weighted matrix factorization to model implicit behavioral feedback. These approaches are heavily dependent on feature engineering. Compared with traditional models, deep learning models(Xu et al. 2023; Qu et al. 2021) are gaining popularity because of their ability to automatically extract features.

In deep learning models, various RNN architectures and their variants have been used to model user's preferences (Hochreiter and Schmidhuber 1997; Duan et al. 2023b). Later, researchers emphasize the importance of modeling long-term preferences and integrating them with short-term preferences. For example, LSPTM (Sun et al. 2020) models both long-term and short-term preferences using a geographically enhanced LSTM model. STGN (Zhao et al. 2020) integrates time gates and distance gates to model and distinguish both user's long-term and short-term interests. RNN models struggle with modeling long-term sequences, so DeepMove (Feng et al. 2018) employs GRU to model current sequence information while utilizing attention mechanisms to capture relevant information from past trajectories.

Given the significant potential of the Transformer architecture(Li et al. 2021), STAN (Luo, Liu, and Liu 2021) employs a two-layer attention mechanism to account for spatiotemporal correlations within trajectories, aiding in the prediction of the next check-in. GETNext (Yang, Liu, and Zhao 2022) incorporates global trajectory information, user general preferences, temporal perception, and categorical perception into the Transformer, while also mitigating the cold start problem. CFPRec (Zhang et al. 2022a) employs a bidirectional Transformer to model past preferences, integrating them with current and future preferences to create a more expressive user representation. Unlike CFPRec, CLSPRec (Duan et al. 2023a) utilizes a shared encoder to model both long-term and short-term sequences, enabling the encoder to iteratively learn common patterns within these sequences. AGRAN (Wang et al. 2023) learns an adaptive POI graph matrix and then use spatiotemporal self-attention mechanism to capture user interests. CLLP (Zhou et al. 2024) enhances the Transformer with local window attention and incorporates spatiotemporal factors.

Conclusion

In this paper, we propose a novel model FHCRec to enhance and jointly recommend user's long and short-term preferences. We model user's long-term sequences using a frequency-domain enhanced Transformer to capture richer long-term preferences from low and high-frequency perspectives. For short-term sequence modeling, we utilize not only the current short-term sequence but also the implicit short-term subsequences within the long-term sequence, employing hierarchical contrastive learning to refine the modeling from local and global viewpoints. Experiments on three real-world datasets validate the effectiveness of our model. Additionally, we demonstrate the effectiveness of our model components through comprehensive ablation studies, hyperparameter analysis, and visualization experiments.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No.62072323, 62102276), China Postdoctoral Science Foundation (Grant No. 2023M732563), the Fundamental Research Funds for the Central Universities, JLU and the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

Duan, C.; Fan, W.; Zhou, W.; Liu, H.; and Wen, J. 2023a. Clsprec: Contrastive learning of long and short-term preferences for next poi recommendation. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, 473–482.

Duan, J.; Zhang, P.-F.; Qiu, R.; and Huang, Z. 2023b. Long short-term enhanced memory for sequential recommendation. *World Wide Web*, 26(2): 561–583.

Feng, J.; Li, Y.; Zhang, C.; Sun, F.; Meng, F.; Guo, A.; and Jin, D. 2018. Deepmove: Predicting human mobility with attentional recurrent networks. In *Proceedings of the 2018 World Wide Web Conference*, 1459–1468.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.

Huang, L.; Ma, Y.; Wang, S.; and Liu, Y. 2019. An attention-based spatiotemporal lstm network for next poi recommendation. *IEEE Transactions on Services Computing*, 14(6): 1585–1597.

Li, J.; Zhou, P.; Xiong, C.; and Hoi, S. C. 2020. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*.

Li, Y.; Chen, T.; Zhang, P.-F.; Huang, Z.; and Yin, H. 2022. Self-supervised graph-based point-of-interest recommendation. *arXiv preprint arXiv:2210.12506*.

Li, Y.; Chen, T.; Zhang, P.-F.; and Yin, H. 2021. Lightweight self-attentive sequential recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 967–977.

Lian, D.; Zhao, C.; Xie, X.; Sun, G.; Chen, E.; and Rui, Y. 2014. GeoMF: joint geographical modeling and matrix factorization for point-of-interest recommendation. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 831–840.

Liao, D.; Liu, W.; Zhong, Y.; Li, J.; and Wang, G. 2018. Predicting Activity and Location with Multi-task Context Aware Recurrent Neural Network. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, 3435–3441.

Lin, Y.; Wan, H.; Guo, S.; and Lin, Y. 2021. Pre-training context and time aware location embeddings from spatial-temporal trajectories for user next location prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 4241–4248.

Liu, Q.; Wu, S.; Wang, L.; and Tan, T. 2016. Predicting the next location: A recurrent model with spatial and temporal contexts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 194–200.

Luo, Y.; Liu, Q.; and Liu, Z. 2021. Stan: Spatio-temporal attention network for next location recommendation. In *Proceedings of the Web Conference 2021*, 2177–2185.

Qu, J.; Hua, W.; Ouyang, D.; and Zhou, X. 2021. A noise-aware method with type constraint pattern for neural relation extraction. *IEEE transactions on knowledge and data engineering*, 35(2): 1134–1148.

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*, 8984–8992.

Sun, K.; Qian, T.; Chen, T.; Liang, Y.; Nguyen, Q. V. H.; and Yin, H. 2020. Where to go next: Modeling long-and short-term user preferences for point-of-interest recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 01, 214–221.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, T.; and Isola, P. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, 9929–9939. PMLR.

Wang, Z.; Zhu, Y.; Wang, C.; Ma, W.; Li, B.; and Yu, J. 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*, 393–402.

Wu, Y.; Li, K.; Zhao, G.; and Qian, X. 2020. Personalized long-and short-term preference learning for next POI recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 34(4): 1944–1957.

Wu, Z.; Xiao, C.; and Vydiswaran, V. 2023. HiCL: Hierarchical Contrastive Learning of Unsupervised Sentence Embeddings. *arXiv preprint arXiv:2310.09720*.

Xu, T.; Qu, J.; Hua, W.; Li, Z.; Xu, J.; Liu, A.; Zhao, L.; and Zhou, X. 2023. Evidence Reasoning and Curriculum Learning for Document-Level Relation Extraction. *IEEE Transactions on Knowledge and Data Engineering*.

Yang, D.; Fankhauser, B.; Rosso, P.; and Cudre-Mauroux, P. 2020. Location prediction over sparse user mobility traces using rnns. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2184–2190.

Yang, S.; Liu, J.; and Zhao, K. 2022. GETNext: trajectory flow map enhanced transformer for next POI recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1144–1153.

Ye, J.; Zhu, Z.; and Cheng, H. 2013. What's your next move: User activity prediction in location-based social networks. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, 171–179. SIAM.

Yu, Z.; Lian, J.; Mahmoody, A.; Liu, G.; and Xie, X. 2019. Adaptive User Modeling with Long and Short-Term Preferences for Personalized Recommendation. In *Proceedings of the International Joint Conference on Artificial Intelligence*, volume 7, 4213–4219.

Zhang, L.; Sun, Z.; Wu, Z.; Zhang, J.; Ong, Y. S.; and Qu, X. 2022a. Next Point-of-Interest Recommendation with Inferring Multi-step Future Preferences. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, 3751–3757.

Zhang, L.; Sun, Z.; Wu, Z.; Zhang, J.; Ong, Y. S.; and Qu, X. 2022b. Next Point-of-Interest Recommendation with Inferring Multi-step Future Preferences. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, 3751–3757.

Zhang, L.; Sun, Z.; Zhang, J.; Lei, Y.; Li, C.; Wu, Z.; Kloeden, H.; and Klanner, F. 2021. An interactive multi-task learning framework for next POI recommendation with uncertain check-ins. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, 3551–3557.

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

Zhao, W.; Wang, B.; Ye, J.; Gao, Y.; Yang, M.; and Chen, X. 2018. Plastic: Prioritize long and short-term information in top-n recommendation using adversarial training. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 3676–3682.

Zheng, Y.; Gao, C.; Chang, J.; Niu, Y.; Song, Y.; Jin, D.; and Li, Y. 2022. Disentangling long and short-term interests for recommendation. In *Proceedings of the ACM Web Conference 2022*, 2256–2267.

Zhou, H.; Jia, Z.; Zhu, H.; and Zhang, Z. 2024. CLLP: Contrastive Learning Framework Based on Latent Preferences for Next POI Recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1473–1482.