Two-Level Routed Sparse Mixture-of-Experts with Spatiotemporal Self-Attention for Next POI Recommendation

Qiang Li*, Lei Zhang*, Bailong Liu*, and Fuyang Zhang†
*School of Computer Science and Technology / School of Artificial Intelligence
China University of Mining and Technology, Xuzhou, 221116, Jiangsu, China
Engineering Research Center of Mine Digitalization, Ministry of Education, Xuzhou, 221116, Jiangsu, China
†Beijing University of Posts and Telecommunications, Beijing, China
Emails: {niubili, zhanglei, liubailong}@cumt.edu.cn, zhangfuyang@bupt.edu.cn

***Abstract*—Next Point-Of-Interest (POI) recommendation predicts a user's next visited location from historical check-in sequences with spatio-temporal context, and serves as a core task in location-based services and urban computing. However, large-scale irregular check-ins still exhibit multi-scale spatio-temporal dependencies and heterogeneous long-tail mobility patterns, making dense shared Feed-Forward Networks (FFNs) prone to pattern entanglement under limited compute. To address these issues, we present BiRtSTMoE, which augments the two-stage spatio-temporal attention framework as follows: (i) We design multi-channel, multi-scale relation biases and inject stage-appropriate biases into the two attention stages: temporal, spatial, and semantic biases for history aggregation, and candidate-history spatial and semantic biases for candidate matching. (ii) We replace the dense FFN in each Transformer block with a two-level routed sparse Mixture-of-Experts (MoE) feed-forward network. A mechanism gate performs coarse routing based on transition dynamics and periodic rhythm features to select an expert group, and a scene gate performs fine-grained routing within the selected mechanism using POI scene semantics to sparsely activate sub-experts. We add hierarchical load-balancing regularization to stabilize training and avoid expert collapse. Experiments on common LBSN benchmarks (NYC and TKY) show consistent gains over strong baselines, and improved robustness on long-tail time-gap/distance slices; routing analyses also support interpretability.**

***Index Terms*—Next point-of-interest recommendation, spatio-temporal attention, mixture of experts (MoE), sparse routing, mobility modeling**

I. Introduction

Modeling and forecasting human mobility remains a central problem in urban computing. With the wide deployment of smart mobile devices and Location-Based Social Networks (LBSNs), large-scale spatio-temporal trajectory collection has become feasible, which in turn enables high-accuracy Point-Of-Interest (POI) recommendation. However, real-world check-in sequences still exhibit pronounced spatio-temporal heterogeneity and long-range dependence: transition regularities vary across time scales, while POI visitation is highly long-tailed, where popular venues dominate training and tail venues lack supervision. These factors jointly increase the difficulty of next POI recommendation under large candidate spaces.

Recent next-POI models have evolved from neural sequential recommenders such as DeepMove [5] and GeoSAN [6] to attention-based and graph-enhanced architectures, including STAN [4], STiSAN [1], ReHDM [2], and MCGT [3]. These studies show the benefit of explicitly modeling spatio-temporal relations, while TiSASRec [7] further suggests that interval-aware attention is effective for irregular sequential recommendation. In parallel, sparse conditional computation with Mixture-of-Experts (MoE), from sparsely gated MoE [8] and Switch Transformer [9] to recommendation-oriented designs such as M3oE [10] and FAME [11], provides a practical way to increase model capacity and reduce interference among heterogeneous behaviors.

Despite this progress, three limitations remain. First, single-scale time-gap and distance biases are often insufficient for jointly modeling short-range transitions and long-range dependencies in irregular mobility sequences. Second, user mobility exhibits both mechanism-level variation (e.g., commuting vs. exploration) and scene-level variation, yet a shared dense feed-forward network can entangle these patterns and cause negative transfer. Third, increasing capacity usually raises computation, while long-tail POIs remain hard to model robustly.

We therefore propose **BiRtSTMoE**, a STAN-based next POI model that injects stage-appropriate multi-channel relation biases into the two attention stages and replaces dense FFNs with a mechanism-to-scene two-level routed sparse MoE.

Our main contributions are as follows:

- We present BiRtSTMoE for next POI recommendation, combining multi-scale relation biases with a two-level routed sparse MoE on top of the STAN two-stage spatio-temporal attention backbone.
- We design a multi-channel, multi-scale relation injection strategy with stage-appropriate biases for history aggregation and candidate matching.
- We introduce a mechanism-to-scene two-level routed MoE with hierarchical load-balancing regularization, en-

abling sparse conditional specialization for heterogeneous mobility patterns with modest computational overhead.

II. Problem Definition

Definition 1 (POI Check-in Sequence)**.** Let $\mathcal{U}$ denote the user set and $\mathcal{V}$ denote the POI vocabulary. For user $u \in \mathcal{U}$, the time-ordered historical prefix is $\mathbf{x}_{1:L}^u = \{(v_i, t_i, \ell_i, c_i, r_i)\}_{i=1}^L$, where $v_i \in \mathcal{V}$ is the visited POI, t_i is the timestamp, $\ell_i \in \mathbb{R}^2$ is the coordinate, and c_i, r_i denote the optional category and region tokens; r_i is an offline-mapped region token obtained by discretizing the POI coordinate into a coarse geographic region.

Definition 2 (Relative Spatio-Temporal Relations)**.** For any two historical check-ins i, j, the relative time gap and geographic distance are $\Delta t_{i,j} = |t_i - t_j|$ and $\Delta d_{i,j} = \mathrm{dist}(\ell_i, \ell_j)$. For candidate POI $v \in \mathcal{V}$ and history position j, the candidate–history spatial relation is

$$N_{v,j}^d = \mathrm{dist}(\ell_v, \ell_j). \tag{1}$$

where ℓ_v is the coordinate of v, and v is associated with optional category and region tokens c_v, r_v. The distance $\mathrm{dist}(\ell_i, \ell_j)$ is computed by the Haversine great-circle distance.

Definition 3 (Next POI Recommendation). Given the time-ordered historical prefix $\mathbf{x}_{1:L}^u$ of user $u \in \mathcal{U}$, the next POI recommendation task aims to predict the ground-truth next visited POI $v_{L+1} \in \mathcal{V}$. Specifically, the model learns a scoring function over candidate POIs conditioned on the historical prefix, and ranks the full POI set $\mathcal{V}$ according to these scores. During inference, POIs with higher scores are considered more likely to be the user's next destination.

III. Method

A. Overall Framework

The architecture of BiRtSTMoE is shown in Fig. 1. Built on STAN's two-stage backbone, it introduces multi-channel relation biases in attention logits and a two-level routed sparse MoE in place of dense FFNs. The model consists of representation construction, a two-level gated MoE Transformer, and two-stage aggregation/decision that produces $\mathbf{g}$, $\mathbf{c}_v$, and $s(v)$.

B. Representation Construction and Embedding

This module constructs token embeddings, relation features for two-stage attention, and routing features for sparse expert selection.

a) Construction of base token representations.: For the i-th check-in token, we use additive POI, semantic, temporal/positional, and user embeddings:

$$\mathbf{e}_i = \mathbf{e}_{v_i}^{\text{poi}} + \mathbf{e}_{c_i}^{\text{cat}} + \mathbf{e}_{r_i}^{\text{reg}} + \mathbf{e}^{\text{time}}(t_i) + \mathbf{e}_i^{\text{pos}} + \mathbf{e}_u^{\text{user}}, \tag{2}$$

where $\mathbf{e}^{\text{poi}}, \mathbf{e}^{\text{cat}}, \mathbf{e}^{\text{reg}}$ are POI/category/region embeddings, respectively; $\mathbf{e}^{\text{time}}(\cdot)$ encodes periodicity, $\mathbf{e}^{\text{pos}}$ is the sequential positional encoding, and $\mathbf{e}_u^{\text{user}}$ is the user embedding.

b) Relative spatio-temporal relation construction.: To inject explicit spatio-temporal inductive bias into two-stage attention, we construct two types of relative relation features. Continuous time/distance values are bucketized later when relation biases are looked up in Sec. III-D: (i) within-trajectory token–token relations: for any $i, j \in [1, L]$, compute the time interval $\Delta t_{i,j} = |t_i - t_j|$ and the spatial distance $\Delta d_{i,j}$ (Haversine), used as bias terms for stage-1 intra-sequence self-attention (history $\rightarrow$ history); (ii) candidate–trajectory relations: for a candidate POI v and a history token j, construct the spatial relation $N_{v,j}^d$, together with region/category relations, for stage-2 candidate-aware attention (candidate $\rightarrow$ history) (see Sec. III-D).

C. Two-Level Gated MoE Transformer Enhancement

We keep multi-head attention with relation biases in Transformer blocks and replace the feed-forward sublayer with a two-level routed sparse MoE-FFN to enable conditional computation.

a) Transformer block.: A block contains multi-head attention and MoE-FFN sublayers (both with residual connections and layer normalization) [12]:

$$\tilde{\mathbf{h}}_i = \mathrm{LN}\big(\mathbf{h}_i + \mathrm{MHA}(\mathbf{h}_i)\big), \tag{3}$$

$$\mathbf{h}_i' = \mathrm{LN}\big(\tilde{\mathbf{h}}_i + \mathrm{MoE}(\tilde{\mathbf{h}}_i)\big). \tag{4}$$

b) Routing inputs: mobility routing features.: To make routing explicitly align with the hierarchical decision process of dynamic intent $\rightarrow$ semantic scene, we construct lightweight feature vectors required by two-level gating for each token position $i \in [1, L]$. The first-level gate focuses on transition dynamics and rhythmic signals, while the second-level gate further introduces weak semantics such as region/category:

$$\begin{aligned} \mathbf{z}_i^{(1)} &= \big[\tilde{\mathbf{h}}_i;\ \psi_t(\Delta t_{i,i-1});\ \psi_d(\Delta d_{i,i-1});\ \mathbf{e}^{\text{time}}(t_i)\big], \\ \mathbf{z}_i^{(2)} &= \big[\mathbf{z}_i^{(1)};\ \mathbf{e}_{c_i}^{\text{cat}};\ \mathbf{e}_{r_i}^{\text{reg}}\big]. \end{aligned} \tag{5}$$

Here $\Delta t_{i,i-1}$ and $\Delta d_{i,i-1}$ are the time interval and spatial distance between two consecutive visits ($i = 1$ is padded with 0 or uses a special token embedding of the start token); ψ_t and ψ_d can be bucket embeddings or small MLPs over scalar inputs. The above features contain no label information to avoid leakage.

c) Two-level gated sparse MoE-FFN.: Each expert is a two-layer MLP: $f_{m,e}(\mathbf{x}) = \mathbf{W}_{m,e}^{(2)} \mathrm{GELU}(\mathbf{W}_{m,e}^{(1)} \mathbf{x})$. We adopt an organization of single expert pool + two-level nested gating: suppose there are M mechanism groups, and group m contains E_m scene experts; routing first selects mechanism groups and then selects experts within each selected group, achieving a coarse-to-fine division of labor.

Level-1 (mechanism) routing:

$$\begin{aligned} \mathbf{p}_i^{(1)} &= \mathrm{softmax}(\mathbf{W}^{g1} \mathbf{z}_i^{(1)}) \in \mathbb{R}^M, \\ \mathcal{M}_i &= \mathrm{TopK}(\mathbf{p}_i^{(1)}, k_1). \end{aligned} \tag{6}$$

Level-2 (scene) routing: For each selected $m \in \mathcal{M}_i$:

$$\begin{aligned} \mathbf{p}_{i,m}^{(2)} &= \mathrm{softmax}(\mathbf{W}_m^{g2} \mathbf{z}_i^{(2)}) \in \mathbb{R}^{E_m}, \\ \mathcal{E}_{i,m} &= \mathrm{TopK}(\mathbf{p}_{i,m}^{(2)}, k_2). \end{aligned} \tag{7}$$

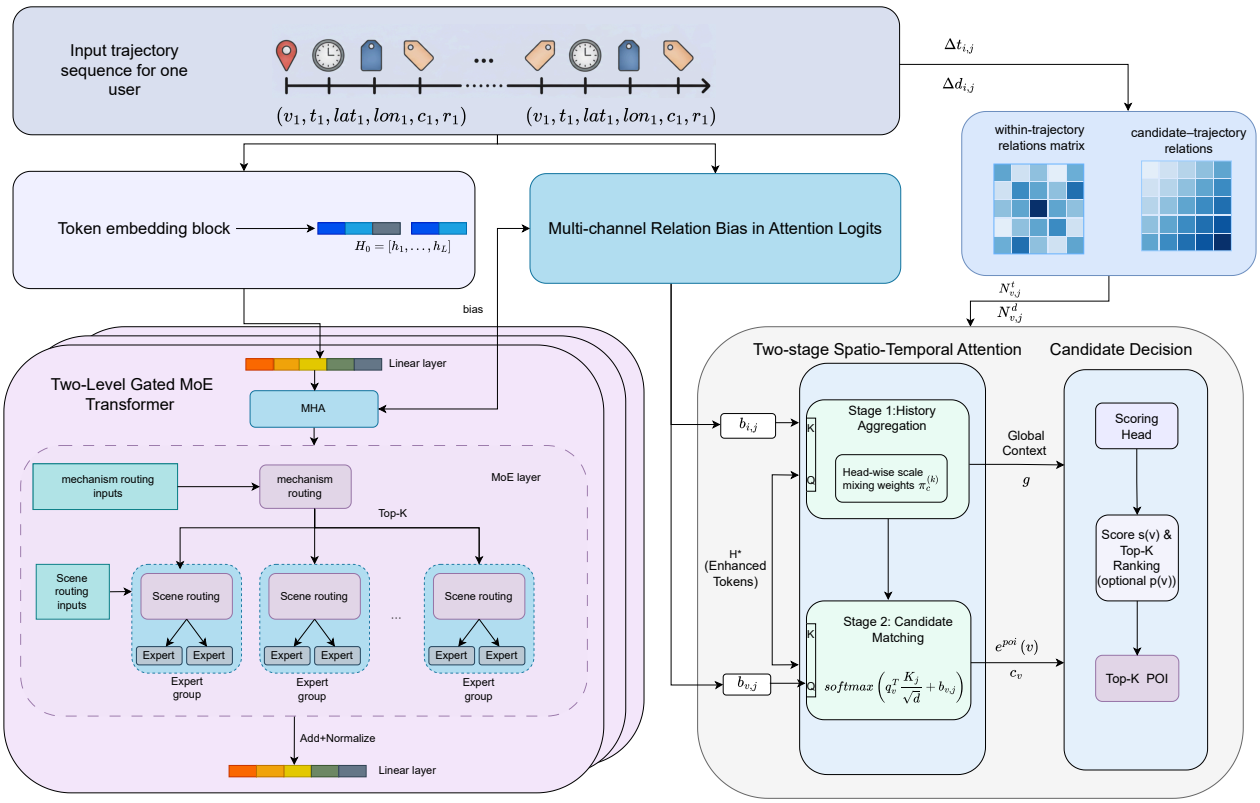

Fig. 1. Overall framework of BiRtSTMoE: two-stage spatio-temporal attention with a two-level routed sparse MoE-FFN.

The two-level gating probabilities are combined as conditional probabilities and normalized over the activated expert set:

$$w_{i,m,e} = p^{(1)}_{i,m}\, p^{(2)}_{i,m,e},$$
$$\tilde{w}_{i,m,e} = \frac{w_{i,m,e}}{\sum_{m' \in \mathcal{M}_i} \sum_{e' \in \mathcal{E}_{i,m'}} w_{i,m',e'}}. \tag{8}$$

Note that $\mathbf{p}^{(1)}$ and $\mathbf{p}^{(2)}$ are routing distributions for expert dispatch, and are different from the recommendation probability $p(v_{L+1} \mid \cdot)$ over POIs.

For experts not selected, we set $\tilde{w}_{i,m,e} = 0$. The final output is the weighted mixture of activated experts (where $\mathbf{x}_i$ is the MoE input, and we set $\mathbf{x}_i = \tilde{\mathbf{h}}_i$ in this paper):

$$\mathrm{MoE}(\mathbf{x}_i) = \sum_{m \in \mathcal{M}_i} \sum_{e \in \mathcal{E}_{i,m}} \tilde{w}_{i,m,e}\, f_{m,e}(\mathbf{x}_i), \tag{9}$$

D. Multi-Scale Spatio-Temporal Aggregation and Candidate Decision

This module follows STAN's *two-stage attention* idea: it first aggregates the history globally to preserve general evidence, then performs context matching between each candidate and the history to avoid compressing the entire history into a single vector and losing information. To better capture correlations between *non-adjacent locations* and *non-consecutive visits*, we inject multi-channel, multi-scale spatio-temporal relation biases into attention logits.

a) Multi-channel, multi-scale relation bias.: The additive bias $b_{i,j}$ for attention from query i to key j is decomposed as a sum of channels:

$$b_{i,j} = b^{t_h}(\Delta t_{i,j}) + b^{t_d}(\Delta t_{i,j}) + b^{t_w}(t_i, t_j) + b^{d}(\Delta d_{i,j}) + b^{r}(r_i, r_j) + b^{c}(c_i, c_j). \tag{10}$$

where b^{t_h} and b^{t_d} characterize hour/day-scale gaps, b^{t_w} characterizes within-week phase (cyclic) differences, and b^d characterizes spatial transition intensity; b^r and b^c are region/category semantic relation biases, respectively.

b) Bucketization of time and distance.: We apply truncated logarithmic bucketization to time gaps:

$$\mathrm{bucket}_t(\Delta t) = \min\left(B_t - 1, \lfloor \log(\Delta t + 1)/\tau_t \rfloor\right), \tag{11}$$

Distance bucketization $\mathrm{bucket}_d(\Delta d)$ is defined similarly. We use separate bucket tables for different time scales, set $b^d(\Delta d) = u^d_{\mathrm{bucket}_d(\Delta d)}$, where u^d is a learnable scalar bias table, and obtain $b^{t_w}(t_i, t_j)$ from a cyclic hour-of-week lookup.

c) Relation-biased attention.: We use multi-head attention with relation biases [13]. For query i and key j:

$$\alpha_{i,j} = \mathrm{softmax}_j\left(\frac{(\mathbf{W}_Q \mathbf{h}_i)^\top (\mathbf{W}_K \mathbf{h}_j)}{\sqrt{d}} + b_{i,j}\right), \tag{12}$$

and output $\mathbf{o}_i = \sum_j \alpha_{i,j} \mathbf{W}_V \mathbf{h}_j$.

d) Stage 1: history aggregation (head-wise multi-scale mixing).: To explicitly encourage different heads to specialize in different scales, we perform head-wise mixing over the multi-scale main channel set $\mathcal{C}_{\mathrm{ms}} = \{t_h, t_d, t_w, d\}$. For head h, we introduce learnable parameters $\{\omega_k^{(h)}\}_{k\in\mathcal{C}_{\mathrm{ms}}}$ and apply softmax to obtain mixing weights:

$$\pi_k^{(h)} = \frac{\exp(\omega_k^{(h)})}{\sum_{k'\in\mathcal{C}_{\mathrm{ms}}} \exp(\omega_{k'}^{(h)})}, \qquad k \in \mathcal{C}_{\mathrm{ms}}. \tag{13}$$

Thus, the relation bias of head h is

$$b_{i,j}^{(h)} = \sum_{k\in\mathcal{C}_{\mathrm{ms}}} \pi_k^{(h)}\, b_{i,j}^{k} \;+\; b^r(r_i, r_j) \;+\; b^c(c_i, c_j), \tag{14}$$

where $b_{i,j}^k$ denotes the bias contribution of channel $k \in \mathcal{C}_{\mathrm{ms}}$ (i.e., $b_{i,j}^{t_h} = b^{t_h}(\Delta t_{i,j})$, $b_{i,j}^{t_d} = b^{t_d}(\Delta t_{i,j})$, $b_{i,j}^{t_w} = b^{t_w}(t_i, t_j)$, and $b_{i,j}^{d} = b^{d}(\Delta d_{i,j})$). Let $i = L$ denote the current last observed check-in; the stage-1 global context is

$$\mathbf{g} = \sum_{j=1}^{L} \alpha_{L,j}^{(1)} \mathbf{h}_j, \tag{15}$$

where $\alpha^{(1)}$ denotes the attention weights after stage-1 aggregation.

e) Stage 2: candidate matching and decision (logit-wise multi-channel summation).: For candidate POI v (with embedding $\mathbf{e}_v^{\mathrm{poi}}$), we compute candidate-aware attention:

$$\beta_{v,j} = \mathrm{softmax}_j \left(\frac{(\mathbf{W}'_Q \mathbf{e}_v^{\mathrm{poi}})^\top (\mathbf{W}'_K \mathbf{h}_j)}{\sqrt{d}} + b_{v,j} \right), \tag{16}$$

where the candidate-side bias keeps the spatial and semantic channels:

$$b_{v,j} = b^d(N_{v,j}^d) + b^r(r_v, r_j) + b^c(c_v, c_j). \tag{17}$$

Then $\mathbf{c}_v = \sum_j \beta_{v,j} \mathbf{h}_j$, and we output the score using a global–candidate joint representation:

$$s(v) = \mathbf{w}^\top \mathrm{GELU}\left(\mathbf{W}_s[\mathbf{g}; \mathbf{c}_v; \mathbf{e}_v^{\mathrm{poi}}]\right). \tag{18}$$

E. Loss and Load Balancing

BiRtSTMoE is trained with a sampled-softmax recommendation loss and a two-level routing regularizer. Negatives are drawn from a smoothed popularity distribution $q(v) \propto f(v)^\gamma$, where $f(v)$ is the POI frequency in the training set and $\gamma \in (0, 1)$.

Let v^+ be the ground-truth next POI and $\mathcal{N}$ the sampled negative set. The recommendation loss is

$$\mathcal{L}_{\mathrm{rec}} = -\log \frac{\exp(s(v^+))}{\exp(s(v^+)) + \sum_{v^-\in\mathcal{N}} \exp(s(v^-))}. \tag{19}$$

To stabilize sparse routing, we apply load balancing at both the mechanism level and the intra-group expert level. At each level, *importance* is the average routing probability and *load* is the fraction of tokens activated by top-k routing. We define

$$\mathcal{L}_{\mathrm{bal}}^{(1)} = M \sum_{m=1}^{M} \mathrm{Imp}_m^{(1)}\, \mathrm{Load}_m^{(1)}, \tag{20}$$

$$\mathcal{L}_{\mathrm{bal}}^{(2)}(m) = E_m \sum_{e=1}^{E_m} \mathrm{Imp}_{m,e}^{(2)}\, \mathrm{Load}_{m,e}^{(2)}, \tag{21}$$

and the overall balancing term is

$$\mathcal{L}_{\mathrm{bal}} = \lambda_1 \mathcal{L}_{\mathrm{bal}}^{(1)} + \lambda_2 \sum_{m=1}^{M} \mathcal{L}_{\mathrm{bal}}^{(2)}(m). \tag{22}$$

The final objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{rec}} + \mathcal{L}_{\mathrm{bal}} + \lambda_{\mathrm{wd}} \|\Theta\|_2^2. \tag{23}$$

IV. Experiments

A. Datasets and Evaluation Protocol

We evaluate on two widely used Foursquare check-in benchmarks: **NYC** and **TKY**. Following standard preprocessing, we remove low-frequency users with fewer than 10 check-ins and low-frequency POIs visited fewer than 5 times. For each user, we sort check-ins chronologically and further split sessions: if the time gap between two consecutive check-ins exceeds a threshold (24 hours in our implementation), we split into a new session; to reduce noise and unlearnability from extremely short sequences, we discard sessions with length smaller than 3. We then construct supervised samples within each session using a sliding window of length L, predicting the next POI from the history prefix. We split train/validation/test for each user in chronological order with an 8:1:1 ratio to ensure no future information leakage. After preprocessing, NYC contains 1,083 users, 38,363 POIs, and 227,428 check-ins with an average session length of 5.93, while TKY contains 2,293 users, 61,858 POIs, and 573,703 check-ins with an average session length of 9.27. We report Recall@K and NDCG@K for $K \in \{5, 10\}$ [14], and rank *over the full POI set* during evaluation (no approximate candidate retrieval) to ensure fair comparison. For robustness, we repeat training with multiple random seeds and report the mean.

B. Baseline Methods and Experimental Settings

Baselines include: **ST-RNN** [15] (a classic method with spatio-temporal recurrent modeling), **STGCN** [16] (introducing graph convolution to model spatio-temporal dependencies), **STAN** [4] (a two-stage spatio-temporal attention framework), **STiSAN** [1] (self-attention enhanced with interval/relation modeling), **GETNext** [17] (a stronger baseline for next POI representation/dependency modeling), **NE-DCHL** [18] (a recent strong baseline). We train and tune all baselines under the same data split and evaluation protocol, and apply early stopping on validation metrics.

Unless otherwise specified, we use $d = 50$, $d_{\mathrm{ff}} = 200$, 4 Transformer layers, and $H = 2$ attention heads. Time/distance bucketization uses $(B_t, B_d) = (32, 32)$. For BiRtSTMoE, we set $M = 3$ mechanism groups and $E_m = 4$ experts per group

(12 experts in total) with top-$(k_1, k_2) = (1, 2)$. We train with Adam (learning rate 0.003), batch size 1, early stopping on validation NDCG@10, and dropout in $[0.01, 0.05]$.

C. Overall Results

Table I reports the overall performance. We can see that **BiRtSTMoE (Ours)** achieves the best results on both NYC and TKY, indicating that (i) multi-channel multi-scale relation biases better handle irregular sampling and long-distance/long-interval dependencies; and (ii) two-level routed sparse MoE-FFN improves modeling of heterogeneous mobility patterns via mechanism-first, scene-level conditional specialization. Compared with strong baselines, especially GETNext and NE-DCHL, Ours improves R@10 / N@10 by $+0.0886/+0.0684$ on NYC and by $+0.0550/+0.0416$ on TKY.

D. Ablation Study

We evaluate: (A1) removing multi-channel multi-scale biases (bias=0); (A2) keeping only a single-channel bias; (A3) two-level MoE $\rightarrow$ single-level MoE; (A4) removing load balancing $\mathcal{L}_{\text{bal}}$; (A5) reversing the two-level routing order (scene-first then mechanism); (A6) removing routing input features (using only token representations). For consistency with Table I, we report R@10 and N@10. From the design motivation, these variants respectively weaken (i) multi-scale relation injection for irregular sampling and long-range dependencies, (ii) conditional specialization of sparse experts, or (iii) routing training stability. Therefore, the overall trend should show degradation to different degrees; removing relation biases and simplifying two-level routing to single-level are usually more significant, while removing the balancing regularizer more easily leads to unbalanced expert utilization during training.

E. Long-Tail Analysis

Robustness under long-tail intervals is analyzed by bucketing test cases by time gap Δt/distance Δd and reporting Recall@10 and NDCG@10. Performance consistently decreases as Δt and Δd grow, but BiRtSTMoE shows a slower degradation trend on both NYC and TKY.

F. Training Dynamics and Expert Loads

The benefits of sparse MoE depend on a proper trade-off between expert specialization and load balancing: if routing collapses too early to a few experts, training may become unstable and waste capacity; if routing is overly uniform, specialization across heterogeneous mechanisms may be weakened. We therefore inspect the token load shares of 12 experts across epochs. The maximum expert share is higher in early training and gradually stabilizes later, while non-uniform expert loads remain after convergence, suggesting that the model reaches a balanced yet specialized routing pattern.

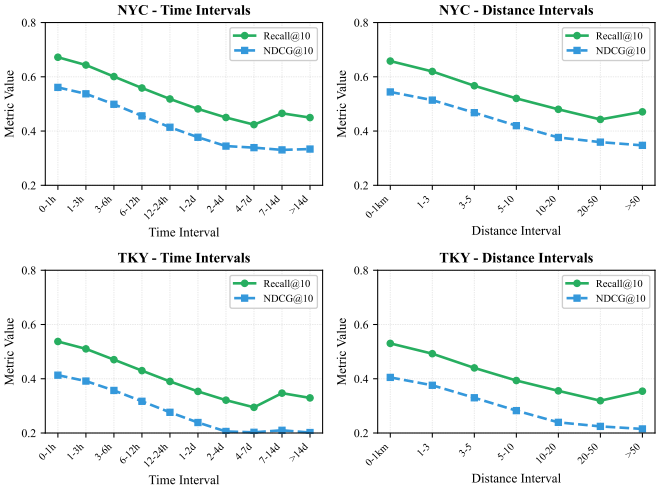

Fig. 2. Results over time-gap/distance buckets (NYC/TKY).

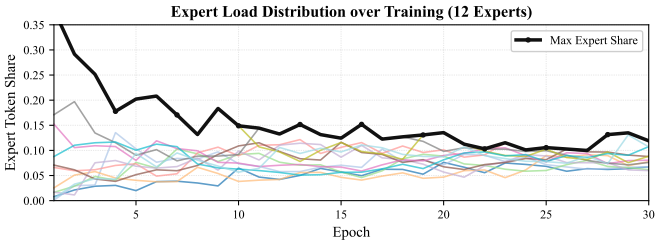

Fig. 3. Expert load dynamics of BiRtSTMoE during training. We track the token share of 12 experts across epochs, along with the maximum expert share per epoch.

G. Efficiency

Capacity is improved while computation is controlled, at the cost of higher parameter overhead: multi-scale relation injection adds only lightweight bias lookups, while two-level routed sparse MoE activates only top-(k_1, k_2) experts per token. In theory, attention remains $O(L^2 d)$ and the sparse MoE cost is $O(Lkdd_{\text{ff}})$ with $k = k_1 k_2$. In practice, the sparse BiRtSTMoE uses 2.24M parameters and keeps FLOPs close to STAN (1.04$\times$), with a moderate runtime overhead of 1.18$\times$ per 1k samples; the dense variant uses 2.01M parameters with 1.10$\times$ FLOPs and 1.07$\times$ runtime.

V. Conclusion

In this paper, we propose BiRtSTMoE for Next POI recommendation. By leveraging multi-channel, multi-scale relation biases and a mechanism-to-scene two-level routed sparse Mixture-of-Experts (MoE) with hierarchical load-balancing regularization, our model addresses multi-scale irregular spatio-temporal dependencies and heterogeneous long-tail mobility patterns. The framework consists of a two-stage spatio-temporal attention backbone (history aggregation and candidate matching) and a two-level routed sparse MoE-FFN for conditional specialization. Results on NYC and TKY indicate superiority over strong baselines, including NE-DCHL,

TABLE I
OVERALL NEXT POI RECOMMENDATION PERFORMANCE (HIGHER IS BETTER).

	NYC				TKY			
Model	R@5	R@10	N@5	N@10	R@5	R@10	N@5	N@10
ST-RNN	0.1718	0.2818	0.1460	0.1973	0.1117	0.1832	0.0838	0.1191
STGCN	0.2371	0.3594	0.1761	0.2709	0.2112	0.3287	0.1482	0.2389
STAN	0.4023	0.4927	0.3225	0.3437	0.2821	0.3317	0.2074	0.2189
STiSAN	0.4453	0.5108	0.3520	0.3680	0.2805	0.3218	0.2250	0.2320
GETNext	0.4172	0.5466	0.3779	0.3941	0.3486	0.3982	0.2812	0.2947
NE-DCHL	0.4469	0.5027	0.3978	0.4165	0.3702	0.4180	0.2975	0.3132
BiRtSTMoE (Ours)	**0.5281**	**0.5913**	**0.4647**	**0.4849**	**0.3925**	**0.4730**	**0.3380**	**0.3548**

TABLE II
ABLATION STUDY.

Variant	NYC		TKY	
	R@10	N@10	R@10	N@10
Full BiRtSTMoE	0.5913	0.4849	0.4730	0.3548
(A1) w/o bias	0.5480	0.4480	0.4390	0.3260
(A2) single-channel bias	0.5635	0.4620	0.4520	0.3390
(A3) single-level MoE	0.5540	0.4725	0.4600	0.3460
(A4) w/o $\mathcal{L}_{\mathrm{bal}}$	0.5830	0.4780	0.4680	0.3515
(A5) reverse routing	0.5790	0.4750	0.4650	0.3490
(A6) w/o routing features	0.5690	0.4665	0.4560	0.3405

with improved robustness on long-tail time-gap/distance slices. Limitations include evaluation on two LBSN benchmarks and the additional routing/dispatch overhead introduced by sparse MoE. Future research will focus on broader datasets and domains and more efficient routing implementations.

REFERENCES

[1] E. Wang, Y. Jiang, Y. Xu, L. Wang, and Y. Yang, "Spatial-temporal interval aware sequential POI recommendation," in *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, 2022, pp. 2086–2098.

[2] X. Li, Z. Gu, R. Yao, Y. Zhou, H. Zhu, J. Zhao, and W. Du, "Beyond individual and point: Next POI recommendation via region-aware dynamic hypergraph with dual-level modeling," in *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 2025, pp. 3081–3089.

[3] X. Qiu, Z. Wang, Z. Zang, C. Yuan, and S. Sun, "MCGT: Multi-class graph model driven transformer for next POI recommendation," *Neurocomputing*, p. 130773, 2025.

[4] Y. Luo, Q. Liu, and Z. Liu, "STAN: Spatio-temporal attention network for next location recommendation," in *Proceedings of The Web Conference 2021 (WWW '21)*, 2021, pp. 2177–2185.

[5] J. Feng, Y. Li, C. Zhang, F. Sun, F. Meng, A. Guo, and D. Jin, "DeepMove: Predicting human mobility with attentional recurrent networks," in *Proceedings of the 2018 World Wide Web Conference (WWW '18)*, 2018, pp. 1459–1468.

[6] D. Lian, Y. Wu, Y. Ge, X. Xie, and E. Chen, "Geography-aware sequential location recommendation," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '20)*, 2020, pp. 2009–2019.

[7] J. Li, Y. Wang, and J. McAuley, "Time interval aware self-attention for sequential recommendation," in *Proceedings of the 13th ACM International Conference on Web Search and Data Mining (WSDM '20)*, 2020, pp. 322–330.

[8] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," in *International Conference on Learning Representations (ICLR)*, 2017.

[9] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *Journal of Machine Learning Research*, vol. 23, no. 120, pp. 1–39, 2022.

[10] Z. Zhang, S. Liu, J. Yu, Q. Cai, X. Zhao, C. Zhang, Z. Liu, Q. Liu, H. Zhao, L. Hu, P. Jiang, and K. Gai, "M3oE: Multi-Domain Multi-Task Mixture-of Experts Recommendation Framework," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*, 2024, pp. 893–902.

[11] M. Liu, S. Zhang, and C. Long, "Facet-aware multi-head mixture-of-experts model for sequential recommendation," in *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining (WSDM '25)*, 2025, pp. 127–135.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 5998–6008.

[13] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2018, pp. 464–468.

[14] K. J"arvelin and J. Kek"al"ainen, "Cumulated gain-based evaluation of IR techniques," *ACM Transactions on Information Systems*, vol. 20, no. 4, pp. 422–446, 2002.

[15] Q. Liu, S. Wu, L. Wang, and T. Tan, "Predicting the next location: A recurrent model with spatial and temporal contexts," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2016, pp. 194–200.

[16] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, 2018, pp. 3634–3640.

[17] S. Yang, J. Liu, and K. Zhao, "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 (SIGIR '22)*, 2022, pp. 1144–1153.

[18] H. Zhang, G. Wang, and X. Yan, "NE-DCHL: Nonlinear enhanced disentangled contrastive hypergraph learning for next point-of-interest recommendation," *Information*, vol. 16, no. 12, p. 1086, 2025.