STCD-HAWKES: SPATIO-TEMPORAL CAUSAL DEBIASING WITH HAWKES PROCESS FOR NEXT POI RECOMMENDATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Next-point-of-interest (POI) recommendation is essential for location-based services yet remains challenging due to the complex, spatio-temporal nature of user mobility. Prior methods often treat time and space as static features or perform simple fusion, ignoring two critical challenges: the distributional heterogeneity between temporal triggers and spatial choices, and the causal entanglement where contextual constraints (e.g., distance) mask intrinsic user preferences. To address these gaps, we propose **STCD-Hawkes**, a unified probabilistic framework for **S**patio-**T**emporal **C**ausal **D**ebiasing with **Hawkes** Processes. Specifically, we decompose the recommendation problem into two coupled generative processes. First, to model *"when"* a user moves, we employ a Time-Aware Hawkes Process with a rotation-based adaptation mechanism, capturing the continuous evolution of temporal intensity. Second, to predict *"where"* the user goes, we introduce a Score-Based Latent Diffusion Model that generates optimal spatial representations from noise, strictly conditioned on the learned temporal intensity. Third, to resolve confounding effects, we design a Causal Disentanglement Module using a Disentangled Causal Mixture-of-Experts (MoE) architecture with orthogonal regularization, which explicitly separates invariant user interests from dynamic contextual sensitivities. Extensive experiments on three real-world datasets demonstrate that STCD-Hawkes significantly outperforms state-of-the-art baselines, offering both superior accuracy and interpretable insights into the causal factors of mobility.

1 INTRODUCTION

Point-of-Interest (POI) recommendation (Cheng et al., 2013; Sánchez & Bellogín, 2022) is a pivotal task in location-based services (LBSNs), aiming to predict a user's next destination based on their historical check-in trajectories. Accurate recommendations not only enhance personalized user experiences but also support critical applications in urban planning, tourism, and mobility analysis. Despite its importance, modeling user mobility remains a formidable challenge, primarily due to the entangled nature of intrinsic user preferences and dynamic spatiotemporal constraints, which often obscures the true causal drivers of check-in behaviors.

Capturing complex spatio-temporal dependencies remains a central difficulty in POI recommendation. Early methods, such as FPMC (Koren et al., 2009) and STRNN (Liu et al., 2016), emphasize transition order but reduce spatial influence to discrete location identities or simple distance constraints, thereby losing richer geographical context. Therefore, later studies attempt to alleviate particular aspects of this limitation through attention or graph structures (Huang et al., 2022; Yin et al., 2023), interval-aware sequence modeling (Wang et al., 2022a), and neural point processes (Shchur et al., 2020; Chen et al., 2021). These advances, however, do not resolve the underlying mismatch between temporal triggers and spatial choices: the two factors are still commonly learned in separate modules and combined through feature fusion or static graphs (e.g., DePOI (Zhou et al., 2025)), leaving their continuous conditional interaction insufficiently modeled.

The first challenge is the **distributional heterogeneity between temporal triggers and spatial choices**. In real-world scenarios, user movement is governed by two fundamental, intertwined prob-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

abilistic processes: *'when to go"* (temporal intensity) and *'where to go"* (spatial distribution). Empirically, the cross-dataset analysis in Appendix C shows that temporal intervals vary markedly across datasets, whereas spatial transitions consistently concentrate within short distances, indicating that the two factors follow distinct generative distributions. However, most POI models treat time and space merely as discrete contextual features or static embeddings, directly fusing them for prediction. This approach overlooks the intrinsic distributional difference between temporal triggers and spatial choices, failing to model how spatial distributions are dynamically conditioned on the user's instantaneous temporal state. Consequently, they lack the flexibility to generate precise predictions in continuous time and space.

The second challenge is the **causal entanglement where contextual constraints mask intrinsic user preferences**. A user's visit to a POI is usually driven by a mixture of *intrinsic preferences* (e.g., a love for coffee) and *contextual constraints* (e.g., visiting a nearby cafe simply because it is 8:00 AM). The strong concentration of consecutive check-ins within short distances further exposes a pronounced proximity bias. Since the data is heavily biased towards spatially proximate locations, standard representation learning models tend to overfit to this "distance bias." They mistake a context-driven visit (convenience) for a genuine user preference (interest). For example, a model might recommend a location solely because it is nearby, ignoring whether it matches the user's actual taste. To achieve robust and explainable recommendations, it is essential to disentangle these confounding factors, ensuring that predictions are driven by causal user interests rather than purely coincidental situational constraints.

To address these challenges, we propose a novel framework named **S**patio-**T**emporal **C**ausal **D**ebiasing with **Hawkes** Process (**STCD-Hawkes**). Unlike prior works that perform simple feature fusion, we conceptually decompose the POI recommendation problem into two coupled sub-tasks: modeling the temporal intensity of user activity and generating the conditional spatial probability of the next destination. Specifically, we first employ a **Time-Aware Hawkes Process** equipped with a rotation-based history adaptation mechanism. This module captures the continuous evolution of user activity intensity, effectively answering 'when" a user is likely to move, regardless of the varying time intervals observed in different datasets. To answer 'where", we introduce a **Score-Based Latent Diffusion Model**. Instead of selecting from a fixed set of locations, this module learns to generate the optimal spatial representation from Gaussian noise, strictly conditioned on the user's temporal state. Furthermore, to resolve the entanglement issue, we design a **Causal Disentanglement Module** based on a Mixture-of-Experts (MoE) architecture with orthogonal regularization. This component explicitly separates user history into intrinsic interests and contextual sensitivities, allowing the model to leverage unbiased causal signals for final ranking.

Our main contributions are summarized as follows:

- We propose a probabilistic decomposition framework that jointly models user mobility through a Time-Aware Hawkes Process and a Latent Spatial Diffusion Model. This approach captures the continuous interplay between temporal intensity and spatial distribution, overcoming the limitations of static feature fusion.
- We introduce a Causal Disentanglement mechanism using a Disentangled Causal Mixture-of-Experts architecture. By enforcing orthogonality between intrinsic preferences and contextual constraints, we effectively remove spurious correlations and enhance the robustness of user representations.
- We conduct extensive experiments on three real-world datasets. The results demonstrate that STCD-Hawkes significantly outperforms state-of-the-art baselines in terms of recommendation accuracy and provides interpretable insights into the causal factors of user mobility.

2 METHODOLOGY

This section details the proposed STCD-Hawkes framework. We address the challenge of modeling continuous spatio-temporal dynamics for discrete POI recommendation by decomposing the problem into temporal intensity modeling, conditional spatial generation, and causal debiasing. The overall framework is illustrated in Figure 1.

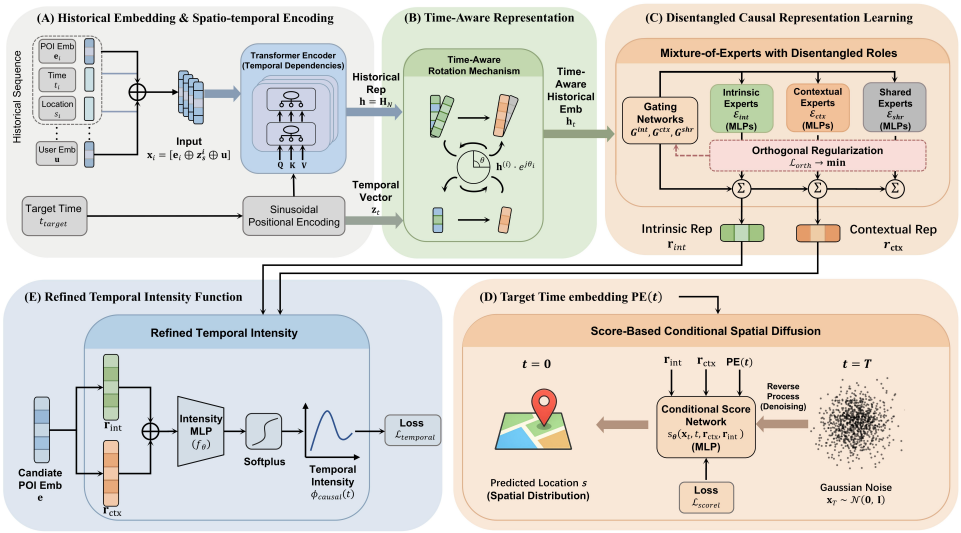


Figure 1: The overall architecture of our proposed method.

2.1 Probabilistic Spatio-temporal Decomposition

To capture the complex temporal dynamics, we formulate the next POI recommendation problem as a Marked Temporal Point Process (MTPP). Let a user's historical trajectory be denoted as $\mathcal{H}_t = \{(l_i, t_i, s_i) \mid t_i < t\}$, where l_i represents the POI identity, t_i is the check-in timestamp, and $s_i \in \mathbb{R}^2$ denotes the geographical coordinates (latitude and longitude).

Unlike prior methods that treat time and space as independent features or generate raw coordinates directly, we propose to model the joint intensity function in a latent feature space. Specifically, we employ GeoEncoder (Lian et al., 2020) to map the raw coordinates s into a dense spatial embedding $\mathbf{z}^s \in \mathbb{R}^d$. The joint intensity $\lambda(t, l)$ of visiting POI l at time t is decomposed into a temporal intensity component and a conditional spatial density:

$$\lambda(t, l) = \lambda(t \mid \mathcal{H}_t) \cdot p(\mathbf{z}_l^s \mid t, \mathcal{H}_t), \tag{1}$$

where $\lambda(t \mid \mathcal{H}_t)$ represents the instantaneous intensity of the user's intent to engage in a check-in behavior at time t (driven by the user's temporal interest), and $p(\mathbf{z}_l^s \mid t, \mathcal{H}_t)$ is the probability density of the target POI's spatial embedding condition on the current time and history. This decomposition allows us to leverage a Hawkes Process for continuous time modeling and a Latent Diffusion Model for generating the optimal spatial representation.

The log-likelihood of observing a sequence of N check-ins is formulated as:

$$\begin{aligned} \log \mathcal{L}(\mathcal{H}) = &\sum_{n=1}^{N} \log \lambda(t_n \mid \mathcal{H}_{t_n}) - \int_0^{T_{\max}} \lambda(\tau)\, d\tau \\ &+ \sum_{n=1}^{N} \log p(\mathbf{z}_{l_n}^s \mid t_n). \end{aligned} \tag{2}$$

2.1.1 Spatio-temporal History Embedding

To effectively condition both the temporal and spatial modules, we construct a unified historical representation. For each check-in event, we fuse the POI's ID embedding $\mathbf{e}_l \in \mathbb{R}^d$, its spatial embedding $\mathbf{z}^s = \text{GeoEnc}(s)$, and the user embedding $\mathbf{u} \in \mathbb{R}^d$:

$$\mathbf{x}_i = \mathbf{W}_{in}[\mathbf{e}_{l_i} \oplus \mathbf{z}_i^s \oplus \mathbf{u}] + \mathbf{b}_{in}, \tag{3}$$

where $\oplus$ denotes concatenation, and $\mathbf{W}_{in}, \mathbf{b}_{in}$ are learnable projection parameters.

The sequence of event embeddings $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ is then processed by a Transformer encoder to capture long-term sequential dependencies. To incorporate temporal information explicitly, we employ a sinusoidal time encoding $\text{PE}(t_i)$ added to the input:

$$\mathbf{H} = \text{TransformerEncoder}(\mathbf{X} + \text{PE}(\mathbf{t})), \tag{4}$$

where $\mathbf{H} \in \mathbb{R}^{N \times d}$ represents the sequence of hidden states. The final hidden state $\mathbf{h} = \mathbf{H}_N$ serves as the summarized historical context $\mathcal{H}_t$ for downstream modules.

2.2 Time-Aware Hawkes Process

This module models the temporal intensity $\lambda(t \mid \mathcal{H}_t)$, quantifying the user's instantaneous desire to initiate a check-in at time t, given their history. Traditional RNN-based point processes often use a static vector to represent history, potentially failing to capture how the influence of past events decays or oscillates continuously over the interval (t_N, t). To address this, we introduce a ***Time-Aware Rotation Mechanism*** that dynamically adapts the historical representation to the query time t.

2.2.1 Rotation-based Temporal Adaptation

Let $\mathbf{h} = \mathbf{H}_N \in \mathbb{R}^D$ be the static historical embedding obtained from the Transformer encoder, summarizing the user's sequence up to the last event t_N. To predict the intensity at a future time $t > t_N$, we must encode the elapsed time $\Delta t = t - t_N$. Inspired by Rotary Positional Embeddings (RoPE), we propose to modulate the historical vector via a complex-valued rotation dependent on Δt.

We treat the embedding dimension D as $D/2$ complex subspaces. The j-th subspace corresponds to the feature pair (h_{2j-1}, h_{2j}):

$$\mathbf{h}^{[j]} = h_{2j-1} + \mathrm{i} \cdot h_{2j}, \quad j = 1, \ldots, D/2, \tag{5}$$

where i is the imaginary unit. We define a time-dependent rotation angle $\theta_j(\Delta t) = \Delta t \cdot \omega_j$, where ω_j is a learnable frequency for the j-th subspace. The time-adapted history $\mathbf{h}_t$ is computed by rotating $\mathbf{h}$ in the complex plane:

$$\mathbf{h}_t^{[j]} = \mathbf{h}^{[j]} \cdot e^{\mathrm{i}\theta_j(\Delta t)}. \tag{6}$$

In real-valued form, this operation is equivalent to:

$$\begin{pmatrix} h_{t,2j-1} \\ h_{t,2j} \end{pmatrix} = \begin{pmatrix} \cos(\omega_j \Delta t) & -\sin(\omega_j \Delta t) \\ \sin(\omega_j \Delta t) & \cos(\omega_j \Delta t) \end{pmatrix} \begin{pmatrix} h_{2j-1} \\ h_{2j} \end{pmatrix}. \tag{7}$$

This rotation mechanism allows the model to capture periodic temporal patterns and time-decaying dependencies naturally within the latent space, without destroying the norm of the historical vector.

2.2.2 Intensity Function Estimation

Based on the time-adapted history $\mathbf{h}_t$, we model the conditional intensity function $\lambda(t \mid \mathcal{H}_t)$. Since this term represents the user's generic activity level (independent of specific POI choices), it is formulated as:

$$\lambda(t \mid \mathcal{H}_t) = \text{softplus}\left(\mathbf{w}_\lambda^\top \text{MLP}_\lambda(\mathbf{h}_t \oplus \mathbf{u})\right), \tag{8}$$

where $\mathbf{u}$ is the user embedding, and $\text{softplus}(x) = \log(1 + e^x)$ ensures the intensity is positive. The inclusion of $\mathbf{u}$ allows the base activity rate to be personalized (e.g., some users are naturally more active than others).

This decoupled $\lambda(t)$ enables the model to distinguish between temporal intensity and spatial diffusion, adhering to the decomposition in Eq. (1).

2.3 Score-Based Conditional Spatial Diffusion

Modeling the spatial distribution of the next POI is challenging due to the complex nature of human mobility. Instead of assuming a simple parametric distribution (e.g., Gaussian) over coordinates, we propose a conditional diffusion model operating in the latent spatial feature space. Our goal is to learn the conditional density $p(\mathbf{z}^s \mid t, \mathcal{H}_t)$ of the target POI's spatial embedding, guided by the user's historical context.

2.3.1 FORWARD DIFFUSION PROCESS IN LATENT SPACE

Let $\mathbf{z}_0^s \in \mathbb{R}^d$ be the clean spatial embedding of the target POI (obtained via the GeoEncoder defined in Sec. 3.1). We define a forward diffusion process that progressively injects Gaussian noise into $\mathbf{z}_0^s$ over a continuous diffusion time horizon $\tau \in [0, 1]$. This process is modeled as a Stochastic Differential Equation (SDE):

$$d\mathbf{z}_\tau^s = \mathbf{f}(\mathbf{z}_\tau^s, \tau)d\tau + g(\tau)d\mathbf{w}_\tau, \tag{9}$$

where $\mathbf{f}(\cdot)$ is the vector-valued drift coefficient, $g(\tau)$ is the scalar diffusion coefficient, and $\mathbf{w}_\tau$ is a standard Wiener process. We adopt the Variance Preserving (VP) SDE formulation, which ensures the distribution of $\mathbf{z}_1^s$ approaches a standard Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

Thanks to the properties of linear SDEs, the transition kernel $p(\mathbf{z}_\tau^s \mid \mathbf{z}_0^s)$ is Gaussian, allowing us to sample the noisy latent state $\mathbf{z}_\tau^s$ directly at any diffusion step τ:

$$\mathbf{z}_\tau^s = \alpha_\tau \mathbf{z}_0^s + \sigma_\tau \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{10}$$

where α_τ and σ_τ are noise schedule parameters derived from the SDE coefficients.

2.3.2 REVERSE DENOISING AND SCORE MATCHING

The generative process reverses the forward diffusion to recover the target spatial embedding $\mathbf{z}_0^s$ from pure noise, conditioned on the user's current temporal state. This is governed by the reverse-time SDE:

$$\begin{aligned} d\mathbf{z}_\tau^s = &\left[\mathbf{f}(\mathbf{z}_\tau^s, \tau) - g(\tau)^2 \nabla_{\mathbf{z}^s} \log p_\tau(\mathbf{z}_\tau^s \mid \mathbf{h}_t)\right] d\tau \\ &+ g(\tau) d\bar{\mathbf{w}}_\tau. \end{aligned} \tag{11}$$

Here, the core challenge is to estimate the score function $\nabla_{\mathbf{z}^s} \log p_\tau(\mathbf{z}_\tau^s \mid \mathbf{h}_t)$. We employ a time-dependent neural network $S_\theta(\mathbf{z}_\tau^s, \tau, \mathbf{c})$ to approximate this score. Crucially, the condition vector $\mathbf{c}$ is derived from the time-aware historical embedding $\mathbf{h}_t$ (from Eq. 7) and the user embedding $\mathbf{u}$:

$$\mathbf{c} = \text{MLP}_{cond}(\mathbf{h}_t \oplus \mathbf{u}). \tag{12}$$

The objective is to minimize the Fisher divergence between the estimated score and the true score of the perturbation kernel:

$$\mathcal{L}_{diff} = \mathbb{E}_{\tau, \mathbf{z}_0^s, \boldsymbol{\epsilon}} \left[\| \sigma_\tau S_\theta(\mathbf{z}_\tau^s, \tau, \mathbf{c}) + \boldsymbol{\epsilon} \|_2^2 \right]. \tag{13}$$

By minimizing $\mathcal{L}_{diff}$, the network learns to "push" noisy embeddings toward the manifold of plausible POI spatial representations consistent with the user's current context. This learned score field effectively models the spatial probability component $p(\mathbf{z}^s \mid t, \mathcal{H}_t)$ in our decomposition.

2.4 DISENTANGLED CAUSAL REPRESENTATION LEARNING

A core challenge in POI recommendation is distinguishing a user's *intrinsic preference* (e.g., a general fondness for coffee shops) from *contextual influence* (e.g., visiting a nearby cafe simply because it is 8:00 AM). Standard embeddings typically entangle these factors, causing the model to overfit spurious correlations (e.g., mistaking a rush-hour visit for a strong preference). To address this, we introduce a causal disentanglement module that factorizes the time-aware historical state $\mathbf{h}_t$ into invariant intrinsic interests and dynamic contextual sensitivities.

2.4.1 MIXTURE-OF-EXPERTS WITH DISENTANGLED ROLES

We employ a Disentangled Causal Mixture-of-Experts architecture to project the entangled history $\mathbf{h}_t$ (from Eq. 7) into distinct causal subspaces. We define three groups of expert networks:

- **Intrinsic Experts** ($\mathcal{E}_{int}$): Capture stable, long-term user preferences that remain invariant across different temporal contexts.
- **Contextual Experts** ($\mathcal{E}_{ctx}$): Capture the user's sensitivity to dynamic constraints, such as time of day and current location.
- **Shared Experts** ($\mathcal{E}_{shr}$): Capture confounding factors that influence both intrinsic preferences and contextual decisions.

Let the input be $\mathbf{x} = \mathbf{h}_t$. The output of the k-th expert in group $g \in \{int, ctx, shr\}$ is $E_k^g(\mathbf{x}) = \text{MLP}_k^g(\mathbf{x})$. We use gating networks $G^g(\mathbf{x})$ to dynamically aggregate these experts:

$$\mathbf{r}_{int} = \sum_k G_k^{int}(\mathbf{x}) E_k^{int}(\mathbf{x}) + \sum_k G_k^{shr}(\mathbf{x}) E_k^{shr}(\mathbf{x}), \tag{14}$$

$$\mathbf{r}_{ctx} = \sum_k G_k^{ctx}(\mathbf{x}) E_k^{ctx}(\mathbf{x}) + \sum_k G_k^{shr}(\mathbf{x}) E_k^{shr}(\mathbf{x}). \tag{15}$$

Here, $\mathbf{r}_{int}$ serves as the debiased representation of user taste, while $\mathbf{r}_{ctx}$ encapsulates the "treatment" effect of the current spatio-temporal context.

2.4.2 Orthogonal Regularization

To ensure the learned representations are statistically independent, we impose an orthogonal regularization constraint. This prevents information leakage between the intrinsic and contextual subspaces:

$$\begin{aligned}\mathcal{L}_{orth} = \sum_{l=1}^{L} (&\|(\mathbf{W}_{int}^l)^\top \mathbf{W}_{shr}^l\|_F^2 + \|(\mathbf{W}_{ctx}^l)^\top \mathbf{W}_{shr}^l\|_F^2 \\ &+ \|(\mathbf{W}_{int}^l)^\top \mathbf{W}_{ctx}^l\|_F^2),\end{aligned} \tag{16}$$

where $\mathbf{W}_g^l$ denotes the weight matrices of the expert group g at layer l, and $\|\cdot\|_F$ is the Frobenius norm.

2.4.3 Causal Injection into Spatio-Temporal Processes

With the disentangled representations, we refine the inputs to the temporal and spatial modules defined in Sections 3.2 and 3.3 to ensure causal correctness.

1. Refined Temporal Intensity: The user's decision to act (intensity) is driven by both their inherent habits and current context. We update the intensity function in Eq. (9) to explicitly use the disentangled features:

$$\lambda(t \mid \mathcal{H}_t) \leftarrow \text{softplus}\Big(\mathbf{w}_\lambda^\top [\mathbf{r}_{int} \oplus \mathbf{r}_{ctx} \oplus \mathbf{u}]\Big). \tag{17}$$

2. Context-Guided Spatial Diffusion: The spatial diffusion process, which generates the likely location, should be heavily conditioned on contextual constraints (e.g., distance, time). We update the condition vector $\mathbf{c}$ in the score network S_θ (Eq. 13) to emphasize the contextual representation:

$$\mathbf{c} \leftarrow \text{MLP}_{cond}(\mathbf{r}_{ctx} \oplus \mathbf{r}_{int}). \tag{18}$$

By explicitly feeding $\mathbf{r}_{ctx}$ to the diffusion model, the generated spatial embeddings are more responsive to the immediate spatio-temporal constraints, while $\mathbf{r}_{int}$ ensures the semantic consistency of the recommended POI.

2.5 Model Optimization

To train the proposed STCD-Hawkes framework, we jointly optimize the prediction accuracy, temporal consistency, and spatial generation quality.

The total loss function is a weighted sum of four components, corresponding to the recommendation task, temporal modeling, spatial generation, and causal disentanglement:

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \gamma_1 \mathcal{L}_{temporal} + \gamma_2 \mathcal{L}_{diff} + \gamma_3 \mathcal{L}_{orth}, \tag{19}$$

where $\gamma_1, \gamma_2, \gamma_3$ are hyperparameters balancing the multi-task objectives.

Recommendation Loss: We treat the next POI prediction as a multi-class classification problem. The probability of visiting POI l is computed via a softmax over the semantic and contextual matching scores:

$$\hat{y}_l = \frac{\exp(\mathbf{e}_l^\top (\mathbf{r}_{int} + \mathbf{r}_{ctx}))}{\sum_{j \in \mathcal{V}} \exp(\mathbf{e}_j^\top (\mathbf{r}_{int} + \mathbf{r}_{ctx}))}, \tag{20}$$

Table 1: Experimental results on three datasets. Best results are in boldface, and second best results are underlined.

Methods	NYC				TKY				CA			
	Acc@1	Acc@5	Acc@10	MRR	Acc@1	Acc@5	Acc@10	MRR	Acc@1	Acc@5	Acc@10	MRR
FPMC	0.1003	0.2126	0.2970	0.1701	0.0814	0.2045	0.2746	0.1344	0.0383	0.0702	0.1159	0.0911
PRME	0.1159	0.2236	0.3105	0.1712	0.1052	0.2278	0.2944	0.1786	0.0521	0.1034	0.1425	0.1002
LSTM	0.1305	0.2719	0.3283	0.1857	0.1335	0.2728	0.3277	0.1834	0.0665	0.1306	0.1784	0.1201
STRNN	0.1483	0.2923	0.3622	0.2198	0.1409	0.3022	0.3577	0.2212	0.0799	0.1423	0.1910	0.1429
STGN	0.1716	0.3381	0.4124	0.2312	0.1610	0.3391	0.3846	0.2311	0.0810	0.1462	0.1940	0.1429
GETNext	0.2342	0.4767	0.5834	0.3510	0.2189	0.4006	0.4884	0.2796	0.1342	0.2819	0.3864	0.2493
AGRAN	0.2121	0.4519	0.5529	0.3179	0.1748	0.4003	0.4640	0.2471	0.1180	0.2818	0.4017	0.2140
PLSPL	0.1917	0.3678	0.4523	0.2806	0.1889	0.3523	0.4150	0.2542	0.1072	0.2278	0.2995	0.1847
STAN	0.2231	0.4582	0.5734	0.3253	0.1963	0.3798	0.4464	0.2852	0.1104	0.2348	0.3108	0.1869
CLSRec	0.1784	0.3830	0.4591	0.2691	0.1453	0.3394	0.4106	0.2340	0.0891	0.1815	0.2013	0.1302
DePOI	0.2569	0.4963	0.5916	0.3773	0.2093	0.4131	0.5027	0.3196	0.1647	0.3484	0.4058	0.2520
Ours	**0.3110**	**0.5349**	**0.6244**	**0.4180**	**0.2511**	**0.4663**	**0.5379**	**0.3520**	**0.2038**	**0.3882**	**0.4317**	**0.2874**
Imprv.	21.1%	7.8%	5.5%	10.8%	14.7%	12.9%	7.0%	10.1%	23.7%	11.4%	6.4%	14.0%

where $\mathbf{e}_l$ is the embedding of candidate POI l. We minimize the Cross-Entropy loss: $\mathcal{L}_{rec} = -\sum_i \log \hat{y}_{l_{true}}$.

Temporal Intensity Loss: To ensure the temporal point process accurately models the user's activity rhythm, we maximize the log-likelihood of the observed event times t_i using the intensity $\lambda(t)$ defined in Eq. (11):

$$\mathcal{L}_{temporal} = -\sum_i \log \lambda(t_i \mid \mathcal{H}_{t_i}) + \int_0^{T_{max}} \lambda(\tau \mid \mathcal{H}_\tau) d\tau. \tag{21}$$

The integral term is estimated via Monte Carlo sampling.

3 EXPERIMENTS

We evaluate STCD-Hawkes on three real-world datasets through overall performance comparisons and sequence-length analyses. Our experiments address the following research questions:

- **RQ1:** How does STCD-Hawkes perform compared to current state-of-the-art POI recommendation methods?
- **RQ2:** How does STCD-Hawkes perform across different trajectory lengths, and what do the results suggest about the roles of temporal and spatial information?
- **RQ3:** Are the individual modules of STCD-Hawkes effective?

The main text reports the core empirical evidence for all three research questions. Appendix E complements RQ3 with complete dataset-wise ablation visualizations and an in-depth analysis of component contributions. Appendix F assesses whether the learned representations respond to the intended historical and contextual factors, whereas Appendix H examines sensitivity to the embedding dimension and number of diffusion steps. Appendix D documents the datasets, evaluation protocols, and implementation settings to support reproducibility.

3.1 COMPARISON OF PERFORMANCE (RQ1)

Table 1 summarizes the performance comparison among all baselines and our proposed model on the three benchmark datasets: NYC, TKY, and CA. For all three datasets, we adopt a chronological 80%/10%/10% split for training, validation, and testing, respectively. The compared methods include traditional sequential models (FPMC(Koren et al., 2009), PRME(Rendle et al., 2010)), recurrent neural models (LSTM(Graves, 2012), STRNN(Liu et al., 2016), STGN(Zhao et al., 2020)),

Table 2: Performance comparison on different POI sequence lengths across three datasets (NYC, TKY, CA). Metrics include Acc@1 and MRR.

Model	NYC		TKY		CA	
	Acc@1	MRR	Acc@1	MRR	Acc@1	MRR
Short sequences (<10 POIs)						
GETNext	0.2272	0.3338	0.2101	0.2726	0.1229	0.2479
AGRAN	0.2109	0.3239	0.1739	0.2322	0.1140	0.2086
CLSRec	0.1400	0.2202	0.1320	0.2149	0.0639	0.1032
DePOI	0.1906	0.2900	0.1792	0.3033	0.1270	0.2135
Ours	**0.2852**	**0.3916**	**0.2281**	**0.3459**	**0.1919**	**0.2955**
Medium sequences (10–20 POIs)						
GETNext	0.2369	0.3543	0.2318	0.2826	0.1303	0.2583
AGRAN	0.2211	0.3062	0.1832	0.2618	0.1243	0.2287
CLSRec	0.1910	0.2526	0.1426	0.2250	0.0738	0.1032
DePOI	0.2815	0.4106	0.2303	0.3430	0.1767	0.2727
Ours	**0.3047**	**0.4126**	**0.2481**	**0.3547**	**0.1999**	**0.2754**
Long sequences (>20 POIs)						
GETNext	0.2376	0.3638	0.2105	0.2826	0.1507	0.2387
AGRAN	0.2013	0.3275	0.1645	0.2424	0.1136	0.1998
CLSRec	0.2000	0.3400	0.1622	0.2651	0.1347	0.1932
DePOI	0.2904	0.4202	0.2114	0.3047	0.1864	0.2629
Ours	**0.3452**	**0.4516**	**0.2781**	**0.3545**	**0.2209**	**0.2953**

graph-enhanced models (GETNext(Yang et al., 2022), AGRAN(Wang et al., 2023d)), attention-based models (PLSPL(Wu et al., 2020), STAN(Luo et al., 2021), CLSRec(Duan et al., 2023), DePOI(Zhou et al., 2025)), and our proposed method. Across all three datasets, our model achieves the best performance on all evaluation metrics, including Acc@1, Acc@5, Acc@10, and MRR.

For the traditional sequential models (FPMC, PRME) model either long-term or short-term transitions but fail to integrate contextual and spatial dependencies, resulting in the lowest performance. RNN Models (LSTM, STRNN, STGN) capture sequential dependencies better, but their limited ability to model long-range dependencies or spatial causality restricts their expressiveness. Attention and Graph-Based Models (GETNext, AGRAN, PLSPL, STAN, CLSRec) introduce attention and graph structures to enhance spatio-temporal reasoning. Nevertheless, they still rely on static representations or manually constructed graphs, leading to suboptimal performance on complex, dynamic mobility data. As the most recent and strongest baseline, DePOI disentangles causal relationships in spatiotemporal graphs. However, its performance is constrained by the explicit graph construction and the lack of continuous temporal modeling. Our model surpasses DePOI by introducing a neural temporal point process and score-based spatial diffusion modeling, providing a principled probabilistic formulation for spatiotemporal dynamics.

3.2 In-depth Analysis on Sequence Length (RQ2)

To better understand the effect of POI sequence length on recommendation performance, we conduct an in-depth analysis by partitioning user trajectories into three groups based on sequence length: short sequences (< 10 POIs), medium sequences ($10 \sim 20$ POIs), and long sequences (> 20 POIs). This allows us to examine how temporal and spatial factors contribute differently across sequences of varying lengths.

Table 2 presents the comparison results of our STCD-Hawkes model against representative baseline models on three datasets.

- **Short sequences:** In this setting, spatial factors dominate the prediction, as users provide limited temporal context. Our model outperforms both Transformer-based and graph-based methods, benefiting from the conditional spatial distribution module that effectively leverages location proximity and contextual embeddings. Graph-based models also perform

reasonably well due to their strong spatial reasoning, whereas Transformer-based models struggle with sparse temporal signals.

- **Medium sequences:** Here, both temporal and spatial information contribute to performance. STCD-Hawkes maintains superior accuracy by jointly modeling temporal intensity and spatial distribution while adaptively balancing their contributions through the causal debiasing mechanism. Transformers begin to show improved performance due to richer temporal patterns, while graph-based models remain limited by insufficient sequence length to fully capture temporal dependencies.
- **Long sequences:** In long sequences, temporal dynamics play a critical role. Our neural Hawkes process captures recurring patterns and periodicity effectively, which allows STCD-Hawkes to significantly outperform other models. Transformer-based models also benefit from longer sequences and achieve reasonable performance, while graph-based models underperform because their spatial focus cannot fully exploit long-range temporal dependencies.

Overall, these results show that the framework adapts to sequence length: spatial information is more influential for short sequences, whereas temporal dependencies become crucial for longer ones. This supports the effectiveness of the proposed spatiotemporal decomposition and causal debiasing across varying sequence lengths.

3.3 Ablation Study (RQ3)

Table 3: Ablation results on three datasets. The best results are in boldface.

Variant	NYC		TKY		CA	
	Acc@1	MRR	Acc@1	MRR	Acc@1	MRR
Ours	**0.3110**	**0.4180**	**0.2511**	**0.3520**	**0.2038**	**0.2874**
w/o t	0.2751	0.3730	0.1913	0.3016	0.1548	0.2448
w/o s	0.3079	0.3953	0.2352	0.3334	0.1816	0.2696
w/o causal	0.2870	0.3860	0.2144	0.3273	0.1663	0.2497

To isolate the contribution of each component, we construct three ablated variants:

- **w/o t** removes the Neural Hawkes Process and represents time intervals using fixed discrete steps, thereby eliminating adaptive temporal-intensity modeling.
- **w/o s** replaces the score-based conditional spatial distribution with a Gaussian model, reducing the spatial component to a unimodal density.
- **w/o causal** removes the causal debiasing module and, consequently, the disentanglement of intrinsic preferences from contextual correlations.

As shown in Table 3, the complete model consistently achieves the best performance across all datasets and metrics, with the most pronounced deterioration observed after removing temporal modeling. These results provide preliminary evidence that the three components make complementary contributions. Appendix E presents the dataset-wise visualizations and detailed analysis.

4 Conclusion

In this paper, we proposed **STCD-Hawkes**, a unified framework that addresses distributional heterogeneity and causal entanglement by decomposing mobility into coupled *"when"* and *"where"* processes. Our model integrates a Time-Aware Hawkes Process for temporal intensity and a Score-Based Latent Diffusion Model for spatial generation further enhanced by a Causal Disentanglement Module to separate intrinsic preferences from contextual constraints. Extensive experiments on three real-world datasets demonstrate that STCD-Hawkes significantly outperforms state-of-the-art baselines in terms of recommendation accuracy.

AI USE STATEMENT

Generative AI tools were not used in conducting the research or preparing this manuscript. This includes the formulation of research questions and hypotheses, methodology and experiment design, data processing and analysis, implementation, interpretation of results, literature review, figure creation, and manuscript writing or editing. The authors take full responsibility for the content of this work.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide detailed dataset statistics, evaluation metrics, implementation settings, ablation studies, and hyperparameter analyses in the appendix. The main paper presents the complete model formulation and optimization objectives. All reported results are evaluated using consistent data-processing protocols and experimental settings across the three datasets.

REFERENCES

Gang Cao, Shengmin Cui, and Inwhee Joe. Improving the spatial–temporal aware attention network with dynamic trajectory graph learning for next point-of-interest recommendation. *Information Processing & Management*, 60(3):103335, 2023.

Ricky T. Q. Chen, Brandon Amos, and Maximilian Nickel. Neural spatio-temporal point processes. In *International Conference on Learning Representations*, 2021.

Chen Cheng, Haiqin Yang, Michael R Lyu, and Irwin King. Where you like to go next: Successive point-of-interest recommendation. In *IJCAI*, volume 13, pp. 2605–2611, 2013.

Eunjoon Cho, Seth A Myers, and Jure Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1082–1090, 2011.

Shuiguang Deng, Dongjing Wang, Xitong Li, and Guandong Xu. Exploring user emotion in microblogs for music recommendation. *Expert Systems with Applications*, 42(23):9284–9293, 2015.

Chenghua Duan, Wei Fan, Wei Zhou, Hu Liu, and Junhao Wen. Clsprec: Contrastive learning of long and short-term preferences for next poi recommendation. In *Proceedings of the 32nd acm international conference on information and knowledge management*, pp. 473–482, 2023.

Alex Graves. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pp. 37–45, 2012.

Zheng Huang, Jing Ma, Yushun Dong, Natasha Zhang Foutz, and Jundong Li. Empowering next poi recommendation with multi-relational modeling. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pp. 2034–2038, 2022.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

Yang Li, Tong Chen, Yadan Luo, Hongzhi Yin, and Zi Huang. Discovering collaborative signals for next poi recommendation with iterative seq2graph augmentation. *arXiv preprint arXiv:2106.15814*, 2021.

Defu Lian, Yongji Wu, Yong Ge, Xing Xie, and Enhong Chen. Geography-aware sequential location recommendation. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2009–2019, 2020.

Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. Predicting the next location: A recurrent model with spatial and temporal contexts. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Yingtao Luo, Qiang Liu, and Zhaocheng Liu. Stan: Spatio-temporal attention network for next location recommendation. In *Proceedings of the web conference 2021*, pp. 2177–2185, 2021.

Michael J Pazzani and Daniel Billsus. Content-based recommendation systems. In *The adaptive web: methods and strategies of web personalization*, pp. 325–341. Springer, 2007.

Yifang Qin, Yifan Wang, Fang Sun, Wei Ju, Xuyang Hou, Zhe Wang, Jia Cheng, Jun Lei, and Ming Zhang. Disenpoi: Disentangling sequential and geographical influence for point-of-interest recommendation. In *Proceedings of the sixteenth ACM international conference on web search and data mining*, pp. 508–516, 2023a.

Yifang Qin, Hongjun Wu, Wei Ju, Xiao Luo, and Ming Zhang. A diffusion model for poi recommendation. *ACM Transactions on Information Systems*, 42(2):1–27, 2023b.

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

Pablo Sánchez and Alejandro Bellogín. Point-of-interest recommender systems based on location-based social networks: a survey from an experimental perspective. *ACM Computing Surveys (CSUR)*, 54(11s):1–37, 2022.

Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pp. 285–295, 2001.

Oleksandr Shchur, Marin Biloš, and Stephan Günnemann. Intensity-free learning of temporal point processes. In *International Conference on Learning Representations*, 2020.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Hongjin Tao, Jun Zeng, Ziwei Wang, Min Gao, and Junhao Wen. Next poi recommendation based on spatial and temporal disentanglement representation. In *2023 IEEE International Conference on Web Services (ICWS)*, pp. 84–90. IEEE, 2023.

En Wang, Yiheng Jiang, Yuanbo Xu, Liang Wang, and Yongjian Yang. Spatial-temporal interval aware sequential poi recommendation. In *2022 IEEE 38th international conference on data engineering (ICDE)*, pp. 2086–2098. IEEE, 2022a.

Xinfeng Wang, Fumiyo Fukumoto, Jin Cui, Yoshimi Suzuki, Jiyi Li, and Dongjin Yu. Eedn: Enhanced encoder-decoder network with local and global context learning for poi recommendation. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pp. 383–392, 2023a.

Xinfeng Wang, Fumiyo Fukumoto, Jiyi Li, Dongjin Yu, and Xiaoxiao Sun. Statrl: Spatial-temporal and text representation learning for poi recommendation. *Applied Intelligence*, 53(7):8286–8301, 2023b.

Xingliang Wang, Dongjing Wang, Dongjin Yu, Runze Wu, Qimeng Yang, Shuiguang Deng, and Guandong Xu. Intent-aware graph neural network for point-of-interest embedding and recommendation. *Neurocomputing*, 557:126734, 2023c.

Yifan Wang, Yifang Qin, Fang Sun, Bo Zhang, Xuyang Hou, Ke Hu, Jia Cheng, Jun Lei, and Ming Zhang. Disenctr: Dynamic graph-based disentangled representation for click-through rate prediction. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pp. 2314–2318, 2022b.

Zhaobo Wang, Yanmin Zhu, Qiaomei Zhang, Haobing Liu, Chunyang Wang, and Tong Liu. Graph-enhanced spatial-temporal network for next poi recommendation. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(6):1–21, 2022c.

Zhaobo Wang, Yanmin Zhu, Chunyang Wang, Wenze Ma, Bo Li, and Jiadi Yu. Adaptive graph representation learning for next poi recommendation. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pp. 393–402, 2023d.

Xuemei Wei, Chunli Liu, Yezheng Liu, Yang Li, and Kai Zhang. Next location recommendation: a multi-context features integration perspective. *World Wide Web*, 26(4):2051–2074, 2023.

Yuxia Wu, Ke Li, Guoshuai Zhao, and Xueming Qian. Personalized long-and short-term preference learning for next poi recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 34(4):1944–1957, 2020.

Xiaodong Yan, Tengwei Song, Yifeng Jiao, Jianshan He, Jiaotuan Wang, Ruopeng Li, and Wei Chu. Spatio-temporal hypergraph learning for next poi recommendation. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pp. 403–412, 2023.

Dingqi Yang, Daqing Zhang, Vincent W Zheng, and Zhiyong Yu. Modeling user activity preference by leveraging user spatial temporal characteristics in lbsns. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(1):129–142, 2014.

Dingqi Yang, Benjamin Fankhauser, Paolo Rosso, and Philippe Cudre-Mauroux. Location prediction over sparse user mobility traces using rnns. In *Proceedings of the twenty-ninth international joint conference on artificial intelligence*, pp. 2184–2190, 2020.

Song Yang, Jiamou Liu, and Kaiqi 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*, pp. 1144–1153, 2022.

Feiyu Yin, Yong Liu, Zhiqi Shen, Lisi Chen, Shuo Shang, and Peng Han. Next poi recommendation with dynamic graph and explicit dependency. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 4827–4834, 2023.

Pengpeng Zhao, Anjing Luo, Yanchi Liu, Jiajie Xu, Zhixu Li, Fuzhen Zhuang, Victor S Sheng, and Xiaofang Zhou. Where to go next: A spatio-temporal gated network for next poi recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 34(5):2512–2524, 2020.

Hailun Zhou, Jiajie Xu, Qiaoming Zhu, and Chengfei Liu. Disentangled graph debiasing for next poi recommendation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1779–1788, 2025.

Roadmap for the appendix.

- Appendix A places STCD-Hawkes in the broader POI recommendation literature. It reviews classical preference- and context-based paradigms as well as recent graph, disentanglement, diffusion, and intention-modeling approaches, and clarifies the remaining need to integrate continuous temporal intensity, conditional spatial generation, and causal debiasing.
- Appendix B formalizes the next-POI recommendation task and introduces the notation for users, POIs, timestamps, spatial contexts, and check-in sequences used throughout the paper.
- Appendix C analyzes temporal intervals and spatial transition distances on NYC, TKY, and CA. These cross-dataset patterns establish the distributional mismatch between temporal triggers and spatial choices and show why dominant proximity signals can obscure intrinsic preferences.
- Appendix D documents the datasets, chronological data splits, evaluation metrics, optimization settings, and principal hyperparameter ranges, enabling the reported comparisons to be reproduced under a consistent protocol.
- Appendix E provides component-wise ablations of the Time-Aware Hawkes Process, conditional spatial diffusion model, and causal debiasing module. The results quantify the contribution of each component and test whether their integration is responsible for the full model's gains.
- Appendix F presents a counterfactual sensitivity analysis of the learned contextual representation. It examines whether predictions depend on plausible recent and historical causes rather than future information or incidental correlations, complementing the aggregate accuracy results with an interpretability check.
- Appendix G summarizes the conventional Hawkes process and its history-dependent intensity, providing the point-process background for the temporal module developed in the main paper.
- Appendix H studies the effects of embedding dimension and the number of diffusion steps across all three datasets. These experiments assess robustness and explain the default choices in terms of predictive accuracy, model capacity, and inference efficiency.

A Further Related Work

Early research on Point-of-Interest (POI) recommendation mainly relied on traditional recommendation paradigms, including collaborative filtering, content-based, context-aware, and hybrid approaches. Collaborative filtering (CF) (Sarwar et al., 2001) includes user- and item-based methods. They infer preferences from similarities among users or items, respectively. In contrast, content-based (CB) techniques (Pazzani & Billsus, 2007) emphasize learning from item attributes such as metadata, keywords, or textual descriptions to generate personalized recommendations. Context-aware models (Wang et al., 2023b) extend these ideas by considering external factors that may influence user decisions, such as time, mood, or location. For instance, Deng et al. (2015) integrated users' emotional states from social media with contextual information to enhance music recommendation. To address the shortcomings of single-strategy systems, hybrid models have been developed to merge different recommendation mechanisms for improved robustness and adaptability.

With the development of deep learning techniques (Yang et al., 2020), researchers have increasingly turned to neural architectures to capture complex behavioral patterns. Graph-based frameworks (Wang et al., 2022c;b; Li et al., 2021) have further advanced this direction by leveraging structural relations among users and POIs. For example, Cao et al. (2023) introduced a spatio-temporal attention network that integrates trajectory graphs to enhance next POI prediction. Tao et al. (2023) proposed a disentangled modeling strategy to separately capture spatial and temporal dependencies. Wei et al. (2023) highlighted semantic relations among POIs and incorporated multiple contextual signals for more accurate recommendations. Similarly, Qin et al. (2023a) constructed dual graphs to jointly represent sequential and geographic relations, applying contrastive learning for feature disentanglement. In a related study, Qin et al. (2023b) adopted diffusion-based generation to unify model training and inference for better generalization, building on score-based generative modeling that learns reverse-time dynamics through score estimation (Song et al., 2021). Other recent efforts include the intention-aware model by Wang et al. (2023c), which identifies latent user

purposes within check-in behaviors, and the hypergraph-based approach by Wang et al. (2023a), which simultaneously captures global and local dependencies across user trajectories.

B Problem Formulation

In this section, we present the basic concepts and notations used in our formulation of the POI recommendation problem. Let the user collection be represented as $U = \{u_1, u_2, \ldots, u_{|U|}\}$ and the set of POIs as $L = \{l_1, l_2, \ldots, l_{|L|}\}$. Each check-in record contains two essential contextual aspects: the spatial context, referring to the location of the POI and denoted by S, and the temporal context, referring to the corresponding timestamp and denoted by T. By integrating both spatial and temporal contexts, the sequence of check-in behaviors for a specific user u can be formally expressed as:

$$H_u = \left[(l_1, t_1, s_1), \ldots, (l_i, t_i, s_i), \ldots, (l_n, t_n, s_n)\right]. \tag{22}$$

C Data Analyses

To thoroughly investigate the temporal and spatial challenges in POI recommendation scenarios, we employ three widely used real-world datasets: NYC, TKY, and CA, each containing rich user check-in records with corresponding temporal and spatial information. Following common practices in POI recommendation, user trajectories are divided into continuous visit sequences. To ensure temporal coherence, any two consecutive check-ins with a time gap exceeding 24 hours are assigned to separate sequences. Based on the results, we can draw the following observations:

C.1 Temporal Interval Distribution

Figure 2(a) presents the distribution of time intervals between consecutive check-ins across the three datasets. Overall, a large number of check-ins occur within short time spans, particularly within 2 hours, which constitutes the highest proportion in all datasets. For NYC and TKY, the frequency decreases steadily as the time interval increases, showing a typical decay pattern. However, in contrast, the CA dataset exhibits a distinct trend: its highest proportion appears within the 12-24 hour range, rather than the shorter intervals.

This observation indicates that user check-in behaviors in different datasets vary significantly in their temporal patterns. In CA, users tend to have longer gaps between visits, suggesting that the next POI is not always directly influenced by the previous one. Therefore, assuming strong temporal continuity across all datasets can be misleading. The time gap between POIs plays a crucial role and should be explicitly modeled in predicting the next location.

C.2 Spatial Distance Distribution

Figure 2(b) illustrates the spatial distance distribution between consecutive POIs. Across all datasets, most check-ins occur within 0-1 km, while the proportion sharply drops for larger distances. This pattern indicates that users frequently move between geographically close POIs, and that short-distance movements dominate user behavior.

However, the dominance of nearby check-ins also implies that spatial proximity alone is insufficient to capture true spatial dependencies. Simply learning from distance-based statistics may fail to represent the complex spatial interactions that influence user mobility.

C.3 Spatio-Temporal Correlation

By comparing temporal and spatial patterns, we find that temporal intervals exhibit a relatively balanced distribution beyond the 0-2 hour range, while spatial distances are heavily concentrated within short ranges. This reveals a fundamental difference between temporal and spatial dynamics: temporal behavior varies widely across datasets, whereas spatial behavior is highly localized. Consequently, time and space should not be treated as directly interchangeable or simply concatenated in modeling. Instead, they need to be jointly modeled while respecting their distinct characteristics.

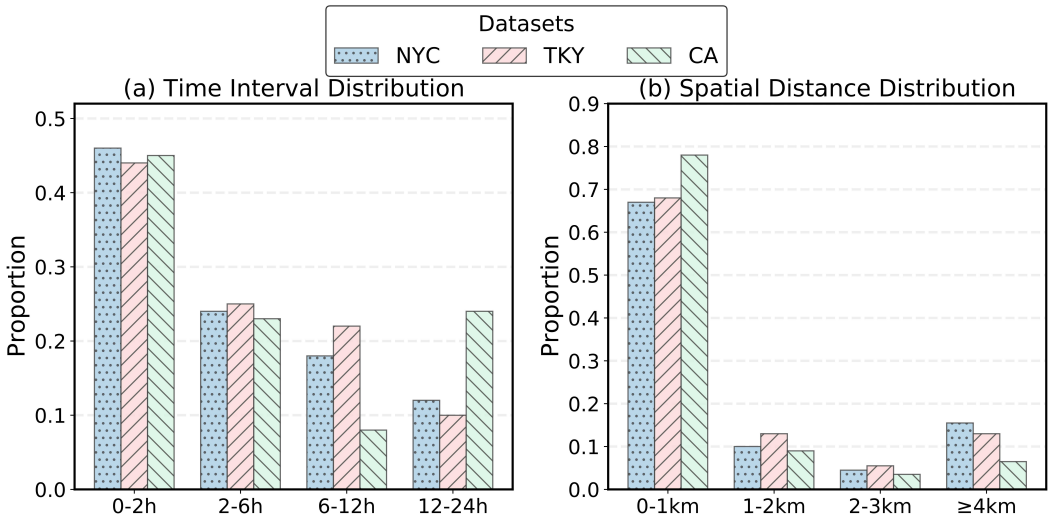


Figure 2: Data analyses.

Together, these contrasting patterns motivate the two challenges considered in this work: the distributional heterogeneity between temporal triggers and spatial choices, and the risk that dominant proximity signals obscure users' intrinsic preferences.

D Additional Experimental Details

D.1 Dataset Details

To evaluate the proposed model, we employ three publicly available real-world datasets: NYC (Yang et al., 2014), TKY (Yang et al., 2014), and CA (Cho et al., 2011). These datasets have been widely adopted in previous POI recommendation studies for their rich temporal and spatial check-in information. The statistics of these datasets are summarized in Table 4. We adopt the same preprocessing procedure as described in (Yan et al., 2023). All check-in records are arranged in chronological order, and each user's sequence is partitioned based on visit continuity. Specifically, the earliest 80% of data is used for model training, the following 10% for validation, and the remaining 10% for testing.

Table 4: Statistics of the datasets

Dataset	#User	#poi	#cat	#checkin	#trajectory
NYC	1,075	5,099	318	104,074	14,160
TKY	2,281	7,844	291	361,430	44,692
CA	4,318	9,923	301	250,780	32,920

D.2 Experimental Settings

Following prior work (Yang et al., 2022), we adopt two standard metrics to assess recommendation performance: Acc@k and Mean Reciprocal Rank (MRR). Acc@k measures whether the target POI appears among the top-k predicted locations, providing a direct measure of ranking accuracy. MRR considers the reciprocal rank of the target POI in the predicted list, thereby reflecting how highly the correct location is ranked on average.

D.3 Implementation Details

The embedding dimension d is searched within $\{32, 64, 128, 256, 512\}$, and we set $d = 256$ as the default for valid comparisons across datasets. We utilize the Adam optimizer for model training, with

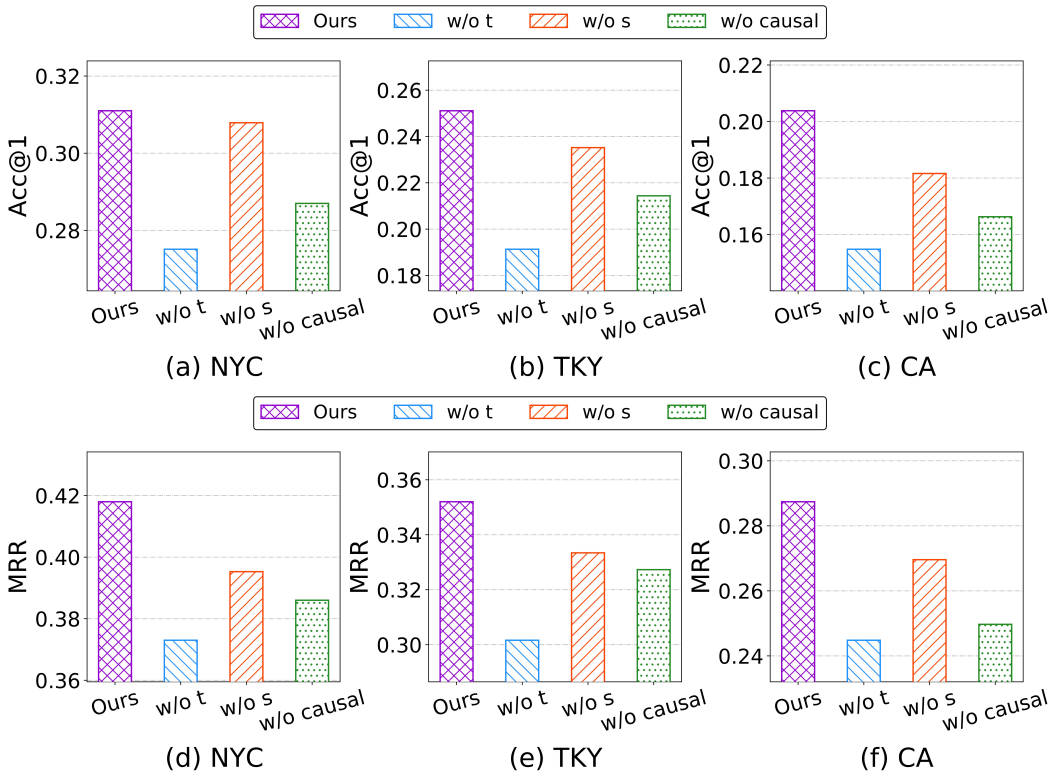


Figure 3: Ablation study. The top row (a–c) shows Acc@1 performance, and the bottom row (d–f) shows MRR performance across three datasets.

the learning rate searched on a logarithmic scale from 1×10^{-4} to 1×10^{-2}. The batch size is set to 128 for three datasets. The number of diffusion steps T is set to 10. The temporal intensity weight (γ_1), the spatial diffusion weight (γ_2), and the orthogonal regularization weight (γ_3)—are carefully searched within the range $[10^{-3}, 1.0]$. For the Disentangled Causal Mixture-of-Experts module, the number of experts for each group (intrinsic, contextual, and shared) is tuned from $\{4, 8, 16\}$.

E Additional Experimental Analyses

E.1 Ablation Study

We conduct ablation experiments to verify the effectiveness of each core design in our proposed framework. Specifically, we construct three model variants by removing or replacing key components, aiming to demonstrate the necessity of temporal modeling, spatial modeling, and causal debiasing within our proposed STCD-Hawkes model. The detailed results are summarized in Figure 3.

The results of the ablation study are reported in Figure 3. Several key observations can be drawn:

Removing the temporal modeling module causes the most significant performance drop across all datasets. This highlights the importance of accurately capturing continuous-time event dynamics, as discrete formulations cannot represent irregular intervals or temporal self-excitation effects inherent in spatiotemporal data.

When the score-based diffusion module is replaced by a Gaussian distribution, performance consistently declines on all datasets. This confirms that the proposed diffusion-based spatial generator provides a more expressive and flexible mechanism to model spatial transitions conditioned on time and context. The Gaussian simplification, by contrast, fails to model geographic movement patterns.

Removing the causal debiasing component leads to moderate but consistent declines across datasets. This demonstrates that eliminating spurious correlations between contextual and spatio-temporal

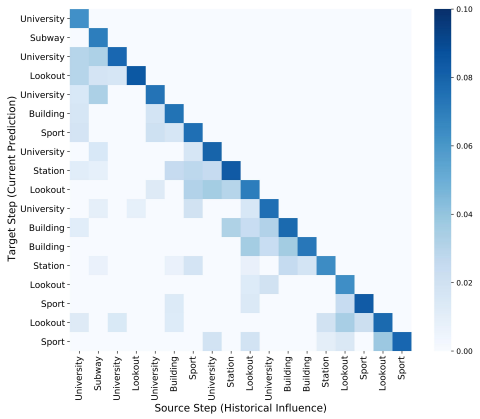


Figure 4: The causal strength of temporal and spatial factors on an individual POI sequence from the NYC dataset.

variables allows the model to focus on causally relevant dependencies, improving both generalization and interpretability.

The complete model achieves the best performance on all datasets, outperforming every ablated variant by a clear margin. The integration of the temporal process, score-based spatial diffusion, and causal debiasing enables the model to jointly reason about dynamic time intervals, complex spatial diffusion, and unbiased causal relations.

F Counterfactual Sensitivity Analysis

To validate whether STCD-Hawkes captures genuine spatio-temporal dependencies rather than spurious correlations, we visualize the dynamic causal relationships for a representative user sequence in the NYC dataset. Since our framework relies on a disentangled representation rather than a static graph, we derive the *causal heatmap* (Figure 4) via Counterfactual Sensitivity Analysis.

Specifically, we utilize the disentanglement capability of the MoE module to isolate the *contextual representation* r_{ctx}, which governs spatio-temporal decisions. The causal strength $C_{i,j}$ from a historical step j to the current step i is calculated as the sensitivity of the contextual output at step i with respect to the input features at step j:

$$C_{i,j} = \left\| \frac{\partial E_{ctx}(h_i)}{\partial x_j} \right\|_F \tag{23}$$

This formulation measures the counterfactual impact: how much the model's assessment of the spatio-temporal context at step i would change if the event at step j were perturbed.

From Figure 4, we observe distinctive patterns that align with our causal modeling:

- **Strong Local Sensitivity:** The heatmap exhibits high intensity along the diagonal. This indicates that the Contextual Experts (E_{ctx}) are most sensitive to the immediate time and location inputs, confirming that the current spatio-temporal state is the primary "treatment" driving the visit.
- **History-Dependent Decay:** We observe causal influence propagating from recent historical steps (near the diagonal) to the current prediction, while distant history shows weaker influence. This reflects the Hawkes process's ability to capture the decaying nature of temporal excitement.
- **Absence of Spurious Future Correlations:** The causality is strictly confined to the lower triangle ($j \leq i$), demonstrating that our model adheres to physical causality where future events do not influence past representations.

Overall, this visualization confirms that STCD-Hawkes effectively disentangles causal factors, relying on specific historical triggers and immediate contexts to guide recommendations.

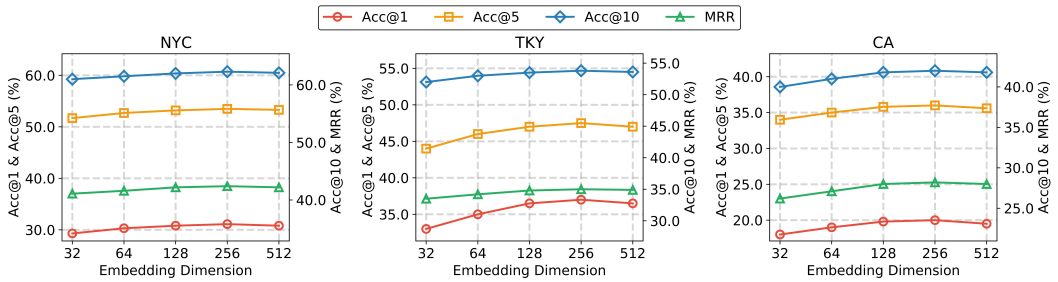


Figure 5: Hyperparameter analysis for different embedding dimensions across three datasets (NYC, TKY, CA).

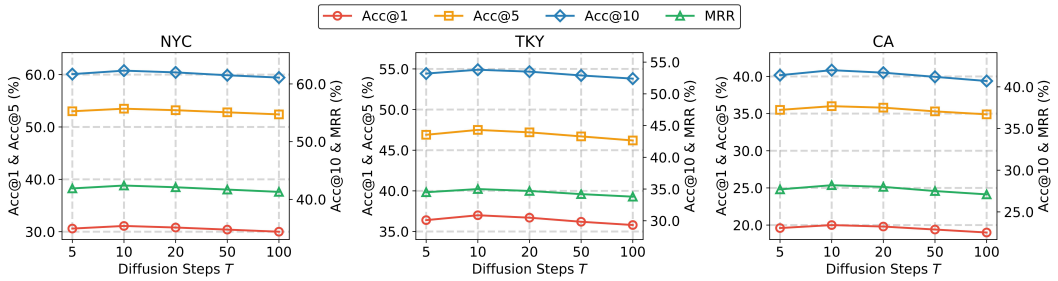


Figure 6: Impact of diffusion steps T on recommendation performance across three datasets (NYC, TKY, CA).

G Hawkes Process

The point process provides a framework for modeling the occurrence of events along a temporal axis. Among its variants, the Hawkes process is a particular case characterized by self-excitation, meaning that past events can stimulate future ones. In a conventional point process, the event intensity is typically treated as a fixed value, independent of previous occurrences. However, in the Hawkes process, each past event contributes to the likelihood of subsequent events, resulting in a dynamic and history-dependent intensity. Formally, the event intensity at time t for type k can be expressed as:

$$\lambda(k,t) = \mu(k) + \sum_{h:t_h<t} \alpha_{k_h,k} \exp\left(-\delta_{k_h,k}\left(t - t_h\right)\right), \tag{24}$$

where $\lambda(k,t)$ denotes the instantaneous intensity of event k at time t, $\mu(k)$ is the base rate independent of history, $\alpha_{k_h,k}$ quantifies the influence of a past event k_h on the current one, and $\delta_{k_h,k}$ controls the rate of exponential decay over time. Here, t_h indicates the timestamp of a preceding event, and t represents the current moment. In the context of this study, each event corresponds to a user's interaction behavior within the POI sequence.

H Hyperparameter Experiment

H.1 Impact of Embedding Dimension

The embedding dimension plays a key role in our STCD-Hawkes model, as it determines the capacity for capturing spatial and temporal patterns from user trajectories. We conduct experiments by varying the embedding size across five values: 32, 64, 128, 256, and 512, and evaluate performance on three datasets (NYC, TKY, CA) using Acc@1, Acc@5, Acc@10, and MRR metrics. The results are visualized in Figure 5.

For the NYC dataset, performance is largely stable across different dimensions, showing that a small embedding is sufficient to capture the necessary information for accurate POI recommendation. In contrast, for TKY and CA, increasing the embedding dimension from 32 to 256 yields noticeable

improvements, indicating that larger embeddings better capture complex user patterns in bigger datasets. However, beyond a dimension of 256, the performance gain becomes marginal for all metrics, suggesting that excessively large embeddings provide limited additional benefits. Overall, these results imply that while larger embeddings can enhance model capacity, the STCD-Hawkes model achieves high performance even with moderate embedding sizes.

H.2 Impact of Diffusion Steps (T)

The number of diffusion steps, denoted as T, is a critical hyperparameter in our score-based generative framework. It governs the granularity of the noise addition and denoising processes, directly influencing the quality of the generated spatial coordinates. To investigate its impact, we conduct experiments by varying T across the set $\{5, 10, 20, 50, 100\}$ and evaluate the performance on the NYC, TKY, and CA datasets. The experimental results are illustrated in Figure 6.

As shown in the figure, we observe a consistent trend across all three datasets. Performance initially improves significantly as T increases from 5 to 10, reaching an optimal peak at $T = 10$. This indicates that a very small number of steps (e.g., $T = 5$) causes large discretization errors during the reverse sampling process, making it difficult for the model to accurately reconstruct the fine-grained spatial distribution from Gaussian noise.

However, as T increases further to 20, 50, and 100, the performance exhibits a slight but steady decline. This suggests that while sufficient steps are necessary for effective denoising, excessive diffusion steps may lead to over-smoothing of the spatial probability density or introduce unnecessary stochastic noise, which hinders the precise ranking of specific POIs. Furthermore, a larger T significantly increases the inference latency. Consequently, we select $T = 10$ as the default setting for our model, as it achieves the best balance between recommendation accuracy and computational efficiency.