POI Recommendation via Multi-Objective Adversarial Imitation Learning

Zhenglin Wan[*1, 3, 5], **Anjun Gao**[*2], **Xingrui Yu**[3], **Pingfu Chao** [† 2], **Jun Song**[4], **Maohao Ran**[4, 5]

[1]School of Data Science, The Chinese University of Hong Kong-Shenzhen, Shenzhen, China
[2]School of Computer Science and Technology, Soochow University, Suzhou, China
[3]Centre for Frontier AI Research, Agency for Science, Technology and Research (A*STAR), Singapore
[4]Department of Geography, Hong Kong Baptist University, Hong Kong
[5]Metasequoia Intelligence, Shenzhen, China
carlos@metaseq.ai, kalis1e7@gmail.com, yu_xingrui@cfar.a-star.edu.sg, pfchao@suda.edu.cn, junsong@hkbu.edu.hk, maohao@life.hkbu.edu.hk

Abstract

Point-of-Interest (POI) recommendation aims to predict users' future locations based on their historical check-ins. Despite the success of recent deep learning approaches in capturing POI semantics and user behavior, they continue to face the persistent problem of data sparsity and incompleteness. In this paper, we introduce Multi-Objective Adversarial Imitation Recommender (MOAIR), a novel framework that integrates Generative Adversarial Imitation Learning with multi-objectives to address this issue. MOAIR effectively captures user behavior patterns and spatial-temporal contextual information via graph-enhanced self-supervised state encoder and overcomes data sparsity by robustly learning from limited data and generating diverse samples. By accommodating diverse user patterns in the training data, the framework also mitigates the typical mode-collapse issue in generative adversarial learning and thus enhances the overall performance. MOAIR employs a multi-objective imitation learning architecture where the imitation learning agent (IL agent) explores the POI space and receives multifaceted reward signals. Utilizing the Paralleled Proximal Policy Optimization (3PO) framework to optimize multi-objectives, the IL agent ensures efficient and stable policy updates. Additionally, to address the issue of high noise in POI recommendation scenarios, we use a novel generative way to define our policy net and incorporate a variational bottleneck for regularization to enhance the stability of adversarial learning. Comprehensive experiments reveal the superior performance for MOAIR compared with baselines, especially with sparse training data.

Introduction

Nowadays, with the flourishing of Location-Based Services (LBS), the next POI (point-of-interest) recommendation system has become an essential component in many spatial-temporal applications. These systems enhance the users' experience by suggesting locations, such as restaurants, museums, parks, and historical sites, that they prefer to visit next. Therefore, the objective of the POI recommendation system is to predict the user's next POI visit based on individual preferences and the most recently visited POI trajectories (Sánchez and Bellogín 2022).

Deep-learning-based methods have proven to be effective in POI recommendations. Compared to original RNN models, ST-RNN (Liu et al. 2016) and HST-LSTM (Kong and Wu 2018) focus on capturing different spatial-temporal features using recurrent-neural-network architecture. Recently, the attention mechanism (Vaswani et al. 2017a) demonstrates powerful performances on sequential modeling, STAN (Luo, Liu, and Liu 2021) addresses the challenges of temporal relations in location-based applications by leveraging a bi-layer attention mechanism. Furthermore, graph-based deep learning (Liu et al. 2017; Xiong et al. 2020) helps to enhance the representation of POI data, and some studies construct graphs to exploit the global transition patterns for check-in interactions across different trajectories, such as GETNext (Yang, Liu, and Zhao 2022).

However, despite the previous success in modeling transition patterns and contextual dependency, their deep-learning-based methods largely rely on training data quality and completeness. On the other hand, we found that the data quality of the POI recommendation task is problematic. The statistics show that the average time interval between two consecutive check-ins in the Gowalla Dataset is approximately 51.28 hours (Zhuang et al. 2024), suggesting severe conditions of missing check-ins and incomplete trajectories. Furthermore, the POI recommendation problem is further complicated by the vast number of POIs and users, along with extremely sparse data and pervasive noise (Liu and Wu 2024). Most existing methods struggle in this context because they assume the trajectory is noise-free and complete, leading to a high risk of overfitting to insufficient training data and being significantly impacted by noisy data. As a result, the performance of existing work is likely to suffer significant degradation due to poor data quality. Additionally, these works fail to integrate spatial-temporal factors involved in the attention mechanism.

To address the challenges of data quality in POI recommendation, we propose **MOAIR**, a novel framework that combines Generative Adversarial Imitation Learning (GAIL) (Ho and Ermon 2016) with **3PO (Paralleled Proximal Policy Optimization)** and a graph-enhanced spatial-temporal self-supervised learning module as its state en-

[*]These authors contributed equally.
[†]Corresponding author

coder. GAIL provides robustness against limited and noisy data due to their adversarial learning approach, while the diverse user patterns in POI datasets help alleviate the mode collapse issue typically seen in adversarial learning framework (Mangalam and Garg 2021). The active exploration by the IL agent also mitigates data sparsity by generating varied training samples, thus bringing much robustness to this task. In MOAIR, to further stabilize the policy improvement, the IL agent's objective is composed of multiple synergistic MDP (Markov Decision Process) terms with multi-reward functions. The primary reward is derived from a discriminator, with auxiliary terms based on the statistical properties of the training set. These diverse reward signals guide the IL agent's policy improvement, optimized through 3PO with a gradient-aggregation mechanism, where each MDP is managed by a parallel PPO unit. To address the imbalance between the discriminator and generator in GAIL and handle high noise in complex POI environments, we define the policy network generatively, introducing a latent variable z that generates the policy. A variational information bottleneck (VIB) (Alemi et al. 2016) inspired by VAE (Kingma and Welling 2013) is incorporated to ensure z preserves the most crucial information, regularizing the discriminator. By collectively optimizing these MDPs, MOAIR effectively captures user preferences and contextual information, resulting in accurate and robust POI recommendations.

We summarize the main contributions of our work as:

- We model the POI recommendation problem in a sequential-decision-making way, for the first time, and employ a generative adversarial imitation learning architecture, leveraging its strength in handling limited and sparse training data in this context.
- We propose Nifty GCNs that optimize the aggregation of graph structure and spatial-temporal-aware attention to make spatial and temporal factors participate in calculating attention scores, serving as part of the state encoder of the IL agent.
- To mitigate the high noise issue in POI recommendation and the potential convergence difficulty, we introduce a novel generative approach for modeling the policy function. This involves incorporating a latent variable z and applying a variational information bottleneck on it to regularize and filter out noise.
- Additionally, we utilize a multi-reward mechanism to stabilize policy improvement of the IL agent. The combination of multi-objective learning and VIB provides more stable policy learning in a highly noisy and complicated POI environment. Moreover, we implement a novel 3PO architecture with a gradient-aggregation mechanism to effectively perform robust multi-objective learning.

Related Work

Related Imitation Learning Work

GAIL and VAIL Generative Adversarial Imitation Learning (GAIL) combines generative adversarial networks (GANs) with imitation learning to train an agent that mimics expert behavior. GAIL uses a discriminator to distinguish between expert and agent-generated trajectories, providing feedback as a reward signal for policy improvement using any reinforcement learning (RL) algorithm. Variational Adversarial Imitation Learning (VAIL) extends GAIL by introducing a latent variable z and optimizes the following objective (Peng et al. 2018):

$$\begin{aligned}\min_{D,E}\max_{\beta\geq 0}\ &\mathbb{E}_{s\sim\pi^*(s)}\left[\mathbb{E}_{z\sim E(z|s)}\left[-\log(D(z))\right]\right]\\ &+\mathbb{E}_{s\sim\pi(s)}\left[\mathbb{E}_{z\sim E(z|s)}\left[-\log(1-D(z))\right]\right]\\ &+\beta\left(\mathbb{E}_{s\sim\tilde{\pi}(s)}\left[\mathrm{KL}\left[E(z|s)\parallel r(z)\right]\right]-I_c\right),\end{aligned}\tag{1}$$

where D is the discriminator, and E maps state s to the latent variable distribution $P(z|s)$.

Recent POI Recommendation Work

In POI recommendation, a core assumption is that users' future movements are influenced by their recent check-ins. Over the past decade, research has evolved from LSTM-based models, which capture sequential patterns often enhanced with attention mechanisms (Kong and Wu 2018; Li, Shen, and Zhu 2018; Wu et al. 2019, 2020; Liu et al. 2020), to graph-based approaches that model spatial-temporal relationships among POIs (Xie et al. 2016; Liu et al. 2017; Xiong et al. 2020; Christoforidis et al. 2021). More recently, attention and transformer-based models have gained traction. For instance, STAN (Luo, Liu, and Liu 2021) employs self-attention to capture spatial-temporal interactions, AGRAN (Wang et al. 2023) uses an adaptive POI graph with attention for dynamic modeling, and GETNext (Yang, Liu, and Zhao 2022) integrates global transition patterns with spatial-temporal context and category embeddings. However, the aforementioned models rely heavily on the quality of training data and are significantly affected by missing data points, data sparsity, and noise. In this paper, we propose a novel imitation learning framework that is robust to sparse and noisy data in the POI recommendation scenario, preventing our model from suffering significant performance degradation due to poor data quality.

Preliminaries

Problem Definitions

The next POI recommendation starts with POI trajectories, so we first give the definition of it:

Definition 1 (**Check-in**). *A **Check-in** is defined as an event where a user visits a specific POI and records the time of the visit. Formally, a check-in can be represented as a tuple (u_i, v_j, t) where $u_i \in U$ represents the set of all users, $v_j \in V$ denotes the set of all possible POIs, and t is the timestamp representing the exact time of the visit.*

Definition 2 (**Trajectory**). *A trajectory L_j^i is represented as a sequence of chronologically ordered check-ins from user $u_i \in U$ starting at time t_j, denoted as $L_j^i = \{(v_1, t_j) \rightarrow (v_2, t_{j+1}) \rightarrow \cdots \rightarrow (v_n, t_{j+n})\}$. Therefore, the full trajectory set of a user u_i is denoted as LS^i, i.e. $\forall L_j^i, L_j^i \in LS^i$.*

Hence, our problem can be defined as:

Definition 3 (**Next POI Recommendation**). *Given the current trajectory* $L_j^i = \{(v_1, t_j) \rightarrow (v_2, t_{j+1}) \rightarrow \cdots \rightarrow (v_n, t_{j+n})\}$ *from a user* u_i*, the next POI recommendation aims to recommend top-k POIs that the user is the most likely to visit next.*

Usually, the next POI recommendation relies on the user's full trajectory set LS^i as a reference, and the top-k POIs are selected by calculating probabilities for all possible candidates and ranking. Hence, in MOAIR, we define the state and action in the IL environment state as: 1) **state**: we define the state as (u_i, L_j^i) since we are using user pattern and historical check-in trajectory to predict the next POI. 2) **action**: all possible next-POI candidates v_{t+1}. In MOAIR, the expert data in the IL context is equivalent to the training dataset.

Proposed Model: MOAIR

In this section, we introduce our proposed model: MOAIR, which consists of a sophisticated graph-based spatial-temporal state encoder and multi-objective adversarial imitation learning framework. We acknowledge that previous work has used a sequential-decision-making framework and graph-based encoder to model human mobility (Wang et al. 2020, 2021a, 2022). However, their design principle largely differs from MOAIR and this section will give a detailed explanation of our method.

Graph Enhanced Spatial-Temporal State Encoder

This component is designed to model representations of check-in sequences. Two key modules form the State Encoder: GCNs for embedding POIs and attention-based models for representing sequences of check-ins. As relationships among POIs are comprehensive and complex, the GNN is a suitable data structure to capture them. Thanks to the powerful self-attention mechanism (Vaswani et al. 2017b), its long sequence modeling and extensibility are greatly improved, which makes self-supervised learning possible.

Nifty GCNs We still employ a dual-graph structure to model the POIs embedding, i.e., geographical graph and transition graph. Inspired by the ideas of simplifying the GCN propagation (He et al. 2020), we decided to omit non-linear activation functions. However, we retain the learned matrix W_h while disabling the bias. This decision is based on the complex semantic nature of POIs and the need to project high-dimensional vectors into another latent space to effectively model relationships such as geographical distance and transition frequency. Additionally, we employ row normalization instead of symmetric normalization to streamline the propagation process. Consequently, the aggregation process of Nifty GCNs (N-GCNs) can be succinctly represented as follows:

$$H = D^{-1}\hat{A}XW_h, \tag{2}$$

where H is the output feature matrix after propagation, capturing transformed node features. D is the diagonal degree matrix. $\hat{A}$ is the adjacency matrix that adds self-loops. X is the input feature matrix, with rows as node feature vectors.

Spatial-Temporal-Aware Attention In the scenario of POI Recommendations, each check-in contains complex semantic meanings, with spatial-temporal factors playing a crucial role in analyzing user patterns. Many previous models treat each check-in within a trajectory as tokens for calculating attention scores. While some works acknowledge the importance of spatial and temporal factors, they often map distance and time intervals into hidden vectors and concatenate them with sequence representations through an external module. Inspired by the model from (Kong and Wu 2018), we propose integrating spatial and temporal factors directly into the attention mechanism, leading to the development of spatial-temporal-aware attention (st-attention). Firstly, for a hidden representation of trajectory $h \in \mathcal{R}^{L\times D}$, there are $E_d \in \mathcal{R}^{L\times L}$ and $E_t \in \mathcal{R}^{L\times L}$ represent the distance intervals and time intervals between every pair of POIs, respectively. Given the significance of relative intervals in calculating attention scores, we apply max-min scaling to each row of E_d and E_t to ensure that both distance and time intervals fall within the range $[0, 1]$. The common attention mechanism can be represented as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{3}$$

To effectively ***disentangle*** spatial and temporal features, we apply transformation matrices W_s and W_t to the hidden representations. The resulting vectors are then concatenated to form the final representation. This approach ensures that spatial and temporal factors are directly involved in the attention mechanism, enhancing the model's ability to capture intricate patterns in user behavior. Overall, we have:

$$S = \text{softmax}\left(\frac{QK^T \cdot E_d}{\sqrt{d_k}}\right)VW_s, \tag{4}$$

$$T = \text{softmax}\left(\frac{QK^T \cdot E_t}{\sqrt{d_k}}\right)VW_t, \tag{5}$$

$$\text{ST-Attention} = S \oplus T, \tag{6}$$

where Q, K, V represent a query, key, and values respectively. In the phase of supervised learning, we discard the self-supervised modules and simply average a trajectory derived from st-attention $h_{u_i}^n$, which represents a hidden vector of user i, and n represents the number of layers in the encoder. The key propagation in State Encoder can be formulated as:

$$h_{u_i}^0 = W_g(\text{N-GCN}(\mathcal{G}_t, h_e) \oplus \text{N-GCN}(\mathcal{G}_d, h_e)), \tag{7}$$

$$h_{u_i}^j = \text{ST-Attention}^{L_j}(h_{u_i}^{j-1}), \tag{8}$$

$$h_{u_i} = \frac{1}{K}\sum_{r=1}^{K} h_{u_i}^n, \tag{9}$$

where $\mathcal{G}_t$ and $\mathcal{G}_d$ stand for the transition graph and distance graph of POIs, W_g represents weights transforming concatenated graph enhanced embedding, h_e represents initial POIs embedding and other semantic embedding like categories, L_j represents the j-th layer of st-attention encoder, and K represents the length of the sequence.

Self-supervised Objective In the phase of self-supervised learning in the State Encoder, selecting an appropriate masking objective is crucial. While span masks are widely used in the NLP field and have been shown to be effective (Raffel et al. 2020), POIs contain complex spatial-temporal and semantic information, making span masks less suitable. Therefore, we slightly modify the Masked Language Model (MLM) (Devlin et al. 2018) by allowing the masking probability to evolve over the training period. This adjustment helps the model to more effectively capture the underlying patterns within check-ins. The specific settings are presented in the Experiment Settings. Finally, the objective function can be represented as:

$$\mathcal{L} = -\sum_{u_i \in \mathcal{U}} \sum_{j=1}^{n} y_{p_j}^{u_i} \log(\hat{y}_{p_j}^{u_i}), \tag{10}$$

where $y_{p_j}^{u_i}$ denotes the true label of the j-th POI for user u_i.

Multi-Objective Adversarial Imitation Learning

Once the state encoder is well-trained, it provides a vector representation for the state setting (u_i, L_j^i). In the subsequent imitation learning stage, this representation is used to compute the policy distribution and discriminator output, fully representing the state (u_i, L_j^i) and conveying the user patterns and contextual information of L_j^i.

Policy Network In traditional reinforcement learning (RL), the policy is defined as $\pi(a|s)$, representing the probability of taking action a when in state s. However, applying this directly to POI recommendation systems exposes the model to high levels of noise, which can degrade performance. To address this challenge, we introduce a novel generative approach within our MOAIR framework to model the policy distribution more robustly. Our strategy involves incorporating a latent variable $\mathbf{z}$, sampled from a prior distribution $\mathcal{N}(0, I)$, to generate the policy distribution $\pi_\theta(a|s, \mathbf{z})$ given the state s. By imposing constraints on the mutual information between the prior and posterior distributions, we ensure that the posterior distribution $P_\phi(\mathbf{z}|s, a)$ captures the most essential information within the (s, a) pairs, effectively filtering out noise.

On the other hand, recognizing the complexity and vastness of the POI space, which poses significant cold-start (Wang et al. 2021b) and slow-adaptation challenges for the IL agent, we propose a pre-training phase. Before engaging in multi-objective imitation learning, we employ Behavior Cloning (BC) (Torabi, Warnell, and Stone 2018) with Maximum Likelihood Estimation to train $\pi_\theta(a|s, \mathbf{z})$, as illustrated in Figure 1, thereby addressing the cold-start issue:

$$\max_\theta \mathbb{E}_{(s,a)\sim\mathcal{D}, \mathbf{z}\sim\mathcal{N}(0,I)}[\pi_\theta(a|s, \mathbf{z})], \tag{11}$$

where $\mathcal{D}$ means expert dataset. To facilitate this, we utilize variational inference by introducing a posterior network $P_\phi(\mathbf{z}|s, a)$. This transforms our objective into co-train the parameters ϕ and θ by maximizing the Evidence Lower Bound (ELBO):

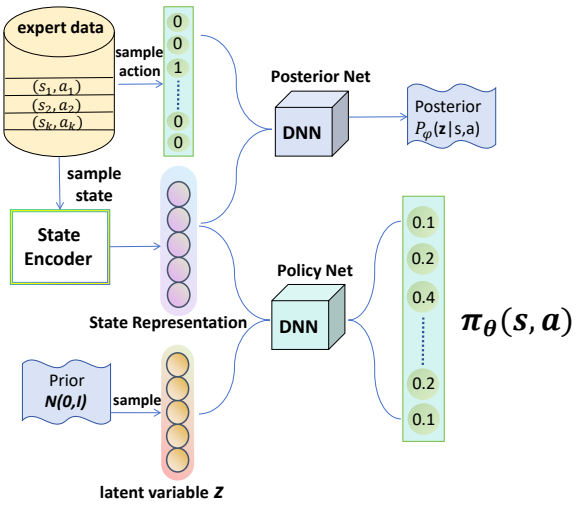

Figure 1: Behavior cloning addresses the cold-start problem by co-training the policy and posterior networks to minimize the ELBO. DNN means deep-neural-network.

$$\mathbb{E}_{(s,a)\sim\mathcal{D}}\Big[\mathbb{E}_{\mathbf{z}\sim P_\phi(\mathbf{z}|s,a)}\big[\pi_\theta(a|s, \mathbf{z})\big] - D_{\mathrm{KL}}\big(P_\phi(\mathbf{z}|s, a)\|P(\mathbf{z})\big)\Big]. \tag{12}$$

Optimizing equation (12) serves two purposes jointly:

1. **Policy Network Pretraining:** This step leverages expert data to pretrain the policy network, enabling the IL agent to effectively navigate the complex POI environment and mitigate the cold-start issue.
2. **Posterior Network Formation:** At the same time, it yields a posterior network $P_\phi(\mathbf{z}|s, a)$ that encapsulates the most crucial and noise-free information within the (s, a) pairs in expert dataset. This posterior network plays a pivotal role in subsequent adversarial training, acting as a bottleneck within the GAIL framework. We will discuss this in *adversarial training* section.

Multi-Reward Setting In MOAIR, we observe that a single reward signal is insufficient and unstable to effectively guide the IL agent's improvement, particularly in complex POI environments and state space. To address this, we implement a multi-reward strategy, which collectively enhances the agent's performance and significantly increases the robustness of training. This is because different MDPs can mutually correct each other, stabilizing the agent when one MDP encounters instability (Hayes et al. 2022). Specifically, we define the reward vector as $\mathbf{r}(s, a) = \big[r_0(s, a), r_1(s, a), \ldots, r_m(s, a)\big]^T$, where $r_0(s, a)$ is the primary reward from the discriminator feedback, and $r_k(s, a)$ $(k = 1, 2, \ldots, m)$ are auxiliary rewards defined by the statistical properties of expert data. These auxiliary rewards are

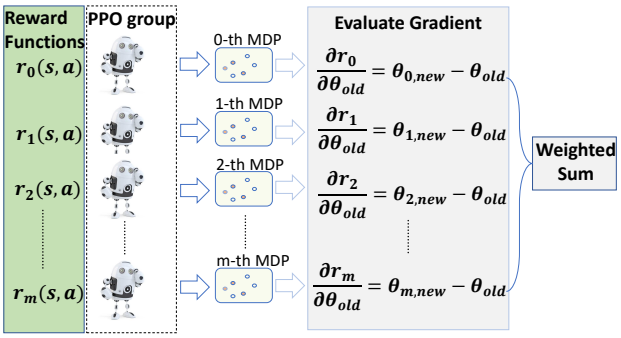

Figure 2: Overview of 3PO architecture.

rule-based and heuristic. For example, in MOAIR, $r_1(s,a)$ is defined as the fitness of the POI selected v with respect to the user u, calculated as $\frac{N_{u,i,\mathcal{D}}}{N_{u,\mathcal{D}}}$, where $N_{u,i,\mathcal{D}}$ means the number of visit times of v by the user u, and $N_{u,\mathcal{D}}$ means the total number of checks ins of the user u. Various definitions of this *fitness* can yield multi-faceted reward functions. The rule-based rewards should meet two criteria: (1) they should encourage the agent to select POIs with high fitness for specific users, and (2) they should be Markovian (i.e., related only to the current state s and action a), as non-Markovian processes cannot leverage RL for gradient estimation.

3PO Architecture In MOAIR, our objective is to optimize the policy π_θ to maximize multiple objectives simultaneously. Recognizing that these reward functions are synergistic and that each reward function guides the policy along a distinct improvement path, we designed a gradient-aggregation mechanism, and maintained a center-policy that is shared by each PPO unit to optimize these objectives. This mechanism aggregates the policy gradients from each objective into a center policy-gradient, which is then used to perform unified policy improvement to center policy.

Our approach leverages the robustness of PPO in dynamic environments and noisy objectives (Batra et al. 2023), making it particularly suitable for scenarios where the reward $\mathbf{r}_0$ from the discriminator evolves over time. To implement this, we developed a Parallel PPO system (3PO) within MOAIR, comprising multiple PPO units. In this system, each reward function corresponds to an independent MDP and is managed by a separate PPO unit in each round. Each MDP is equipped with its own actor-critic pair, where the critic conducts policy evaluation and the actor handles policy improvement. The combined objective of the 3PO framework in MOAIR is formulated as:

$$\max_{\pi_\theta} \mathbb{E}_{\pi_\theta}\left[\sum_{k=0}^{T-1}\gamma^k\left(r_0(s_k,a_k)+\sum_{i=1}^{m}c_i r_i(s_k,a_k)\right)\right], \tag{13}$$

where c_i is a hyperparameter between 0 and 1, representing the weight of the i-th MDP, γ is the decay rate of RL, and T is the length of one episode.

As illustrated in Figure 2. At the beginning of each round, each PPO unit shares the same actor (i.e., policy network). Each PPO unit uses this policy network as the *old* policy and performs importance sampling using this *old* policy as the behavior policy to interact with the environment. Different PPO units receive different reward signals and train different critic networks independently based on the trajectories collected by the *old* policy. Each PPO unit performs multiple rounds of policy updates, resulting in new policy parameters. We get the gradient w.r.t each MDP component in the objective by subtracting the new policy parameter in corresponding PPO units by old policy parameter: $\frac{\partial r_i}{\partial \theta_{\text{old}}} = \theta_{i,new} - \theta_{old}$ (i.e., the gradient of each component in the objective (13)), where $\theta_{i,new}$ is the new policy obtained by the $i-th$ MDP.

Then, the overall gradient of θ is given by:

$$\langle 1, c_1, c_2, \ldots, c_m\rangle \cdot \left\langle \frac{\partial r_0}{\partial \theta_{\text{old}}}, \frac{\partial r_1}{\partial \theta_{\text{old}}}, \frac{\partial r_2}{\partial \theta_{\text{old}}}, \ldots, \frac{\partial r_m}{\partial \theta_{\text{old}}} \right\rangle,$$

where the first vector contains the coefficients $\langle 1, c_1, c_2, c_3, \ldots, c_m\rangle$, and the second vector consists of the partial derivatives of the objective of each MDP with respect to the old parameters $\left\langle \frac{\partial r_0}{\partial \theta_{\text{old}}}, \frac{\partial r_1}{\partial \theta_{\text{old}}}, \frac{\partial r_2}{\partial \theta_{\text{old}}}, \frac{\partial r_3}{\partial \theta_{\text{old}}}, \ldots, \frac{\partial r_m}{\partial \theta_{\text{old}}} \right\rangle$.

During training, we use a synchronous way: after **all** PPO units have finished computing their gradients, MOAIR updates its center-policy parameter π_θ using a linear combination of gradients from multiple PPO units. In next round, we copy the center policy parameter to each PPO unit as their *old* policy and make them share the same updated policy parameters. Different PPO units in the same round can be paralleled, improving efficiency. Derived from the discriminator, the reward function r_0 is continuously updated through adversarial training while other reward functions are static.

Adversarial Training The primary term r_0 in Objective (13) is provided by a discriminator. We employ a modified adversarial training method to optimize this discriminator. In adversarial learning frameworks, it is widely recognized that the discriminator often tends to overfit, producing extreme outcomes (close to 0 or 1). This overfitting results in non-informative gradients and leads to instability in adversarial training (Peng et al. 2018). This issue is further exacerbated in the POI recommendation, where the training data is often noisy and sometimes incomplete.

To address this, we pretrain the policy network and obtain a posterior network in the initial step. The posterior distribution $P_\phi(\mathbf{z}|s,a)$ preserves the most crucial information in the state-action pair (s,a), as ensured by the KL-divergence term in the objective function (12), which limits the mutual information between the prior and posterior distributions of $\mathbf{z}$. Building on this, we use the posterior $\mathbf{z}$ as the input to the discriminator, and the objective of the discriminator is:

$$\begin{aligned}\max_{D,\phi} \quad & \mathbb{E}_{(s,a)\sim\pi_E}\left[\mathbb{E}_{\mathbf{z}\sim P_\phi(\mathbf{z}|s,a)}[\log D(\mathbf{z})]\right] \\ & + \mathbb{E}_{(s,a)\sim\pi_\theta}\left[\mathbb{E}_{\mathbf{z}\sim P_\phi(\mathbf{z}|s,a)}[\log(1-D(\mathbf{z}))]\right] \\ & - \eta\left(\mathbb{E}_{(s,a)\sim\pi_{\text{mix}}}\left[D_{\text{KL}}\left[P_\phi(\mathbf{z}|s,a)\parallel P(\mathbf{z})\right]\right]-c\right).\end{aligned} \tag{14}$$

This approach regularizes the discriminator by preventing it from overfitting to irrelevant or noisy details within the state-action pair (s,a). Moreover, the inclusion of the La-

grangian term continues to limit the mutual information between $\mathbf{z}$ and (s, a) during discriminator training, effectively serving as a form of regularization.

Experiment

Datasets

We utilize check-in data from both *Foursquare* and *Gowalla*, as these datasets are commonly employed in prior studies. The *Foursquare* data covers Tokyo (TKY) and New York (NYC), while we incorporate *Gowalla* data due to its substantial volume. Each dataset is organized by user, sorted chronologically, and split with the first 80% used for training and the remaining 20% for testing. POIs with fewer than 10 check-ins are filtered out, and trajectories are segmented to ensure no more than a one-day gap between consecutive check-ins. Table 2 shows the statistics of the datasets.

Baselines

We selected a mix of traditional and state-of-the-art models for comparison: 1) **LSTM** (Hochreiter and Schmidhuber 1997): a modified RNN with forget gates and memory cells; 2) **ST-RNN** (Liu et al. 2016): extends RNN with time and distance transition matrices; 3) **HST-LSTM** (Kong and Wu 2018): integrates spatial and temporal influences in LSTM with a hierarchical approach; 4) **STAN** (Luo, Liu, and Liu 2021): employs self-attention to capture spatial-temporal interactions in trajectories; 5) **AGRAN** (Wang et al. 2023): uses an adaptive POI graph and attention mechanism for dynamic geographical dependency modeling; 6) **GETNext** (Yang, Liu, and Zhao 2022): a transformer-based model incorporating global transition patterns, spatial-temporal context, and category embeddings; 7) **Graph-FlashBack** (Rao et al. 2022): an RNN-based model leveraging a weighted POI transition graph to capture sequential patterns.

Evaluation Metrics

We assess MOAIR using three widely-used metrics in previous work: Accuracy@k (Acc@k), Precision@k (Prec@k), and Mean Reciprocal Rank (MRR). Acc@k measures the proportion of correctly predicted next POIs, Prec@k calculates the ratio of relevant locations in the top-K results, and MRR evaluates the ranking of the first relevant location in the list. We select K=5,10 to compare the performance.

Settings

In our experiments, we uses four *A40 48G* GPUs, an *AMD EPYC 7543P 32-core* CPU, and a Linux operating system. The embedding dimension for the state is set at *128*. The dimension of latent variable z is set to 32, the learning rate and PPO-clip hyper-parameter of each PPO unit are set to be 3×10^{-4} and 0.1, and the decay rate γ of IL agent is 0.99. The masking probability is set to 0.15 at first and finally achieves 0.25 along with the epoch increasing.

Overall Performance

We compare our model, *MOAIR*, with other baselines on two metrics with *K @ 5, and 10* respectively. As Table 1 illustrates, our model achieves the SOTA performances. Among the three datasets, the New York dataset demonstrated the best performance, followed by Tokyo and Gowalla. This ranking aligns with the complexity of the POI space, as the IL agent's efficiency is affected by environmental complexity. Key observations include: 1). **LSTM** under-performed due to its inability to effectively utilize spatial-temporal features and model user preferences. In contrast, the **ST-RNN** and **HST-LSTM** models outperformed LSTM, with HST-LSTM particularly excelling due to its global context encoding, which captures the periodicity of visit sequences. However, these RNN-based models struggle with sequence tasks compared to self-attention mechanisms. 2). **Graph-Flashback** excelled among the baselines, likely due to its use of a weighted transition graph. **GETNext**, while slightly less effective than Graph-Flashback, still outperformed other baselines, highlighting the importance of category information alongside transition patterns. However, these models fail to effectively analyze users' patterns from historical check-ins and face overfitting risks due to their heavy dependence on training data, especially when it is sparse. 3). **MOAIR** incorporates the strengths of the above models in its state encoder while mitigating their weaknesses with a well-designed IL agent and a reward mechanism suited for POI recommendation tasks. This approach offers robustness to sparse data and enhances user-aware prediction.

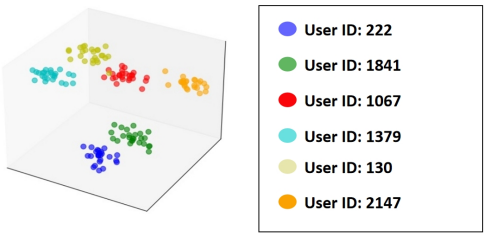

Figure 3: Visualization of High Dimensional user-trajectory representation by t-SNE.

Visualization of User-Preference Modeling

To further demonstrate MOAIR's effectiveness in modeling user patterns, Figure 3 uses t-sne (Van der Maaten and Hinton 2008) to visualize the representations of multiple users and their check-in trajectories, as obtained by our state encoder. Since visualizing all trajectories is impractical, we randomly selected a few users and included all their trajectories from the Tokyo dataset. The visualization reveals a clear clustering of different check-in trajectories for the same user, indicating that the user-trajectory representations in high-dimensional space are significantly influenced by user patterns. Each cluster center represents a distinct user pattern.

Ablation Study

As illustrated in Figure 4, we compared the performance of various ablation modules on the TKY and NYC datasets using Acc@5, Prec@5, and MRR as metrics. Our designed modules were shown to be essential. The ablation studies include: 1) w/o Multi-reward: using a single reward function from discriminator and reducing PPO units to 1, causing

Model	TKY			NYC			Gowalla		
	Acc@5/Prec@5	Acc@10/Prec@10	MRR	Acc@5/Prec@5	Acc@10/Prec@10	MRR	Acc@5/Prec@5	Acc@10/Prec@10	MRR
LSTM	0.220/0.658	0.301/0.563	0.148	0.224/0.596	0.315/0.495	0.145	0.090/0.474	0.125/0.410	0.071
ST-RNN	0.245/0.684	0.326/0.587	0.161	0.247/0.621	0.332/0.523	0.162	0.095/0.491	0.134/0.425	0.078
HST-LSTM	0.273/0.710	0.351/0.618	0.175	0.279/0.654	0.365/0.557	0.175	0.115/0.522	0.151/0.458	0.094
STAN	0.331/0.782	0.416/0.684	0.215	0.333/0.729	0.424/0.612	0.223	0.156/0.578	0.197/0.496	0.125
AGRAN	0.334/0.789	0.415/0.688	0.219	0.334/0.733	0.426/0.620	0.226	0.158/0.582	0.201/0.503	0.129
GETNext	0.374/0.846	0.468/0.752	0.258	0.389/0.778	0.479/0.665	0.269	0.182/0.610	0.237/0.545	0.139
Graph-Flashback	0.381/0.854	0.475/0.761	0.267	0.401/0.795	0.491/0.671	0.274	0.191/0.625	0.246/0.557	0.144
MOAIR	**0.415/0.896**	**0.511/0.811**	**0.282**	**0.436/0.830**	**0.538/0.715**	**0.290**	**0.225/0.663**	**0.273/0.582**	**0.161**

Table 1: Performance Comparison Across Different Models and Datasets.

Datasets	Users	POIs	Check-ins	Sparsity
Tokyo	2292	7873	447488	97.52%
New York	1082	5135	147735	97.34%
Gowalla	10923	9054	306687	99.69%

Table 2: Datasets Statistics

a 17% performance drop (Acc@5 in TKY); 2) w/o BC: removing behavior cloning stage, leading to a 9% performance drop; 3) w/o latent variable: using direct mapping instead of generative policy, resulting in a 5% performance drop; 4) w/o st-attention: using distance intervals and time intervals embedding instead, leading to a 10% performance drop.

Influence of Data Sparsity

To validate MOAIR's robustness against data sparsity, we systematically reduced the training dataset size. Figure 5 shows the performance decline as the number of check-in sequences per user decreased. We compared MOAIR with four competitive models using training set sizes of 90%, 75%, 60%, 45%, and 30% of the original data. On the NYC dataset (measured by Acc@5), MOAIR was the least affected by reduced training samples, particularly when the reduction wasn't severe. In contrast, deep learning models showed significant performance drops, highlighting their sensitivity to data sparsity. We attribute MOAIR's robustness to several key factors: the robustness of adversarial

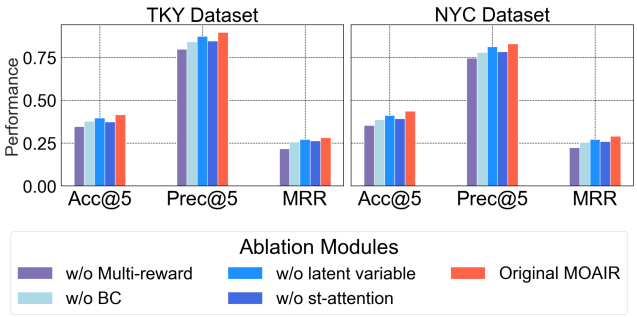

Figure 4: Ablation Study on Four Key Modules in MOAIR.

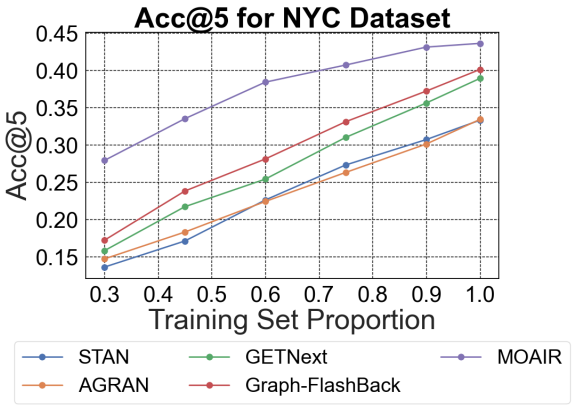

Figure 5: Influence of Data Sparsity.

learning, the IL agent's active exploration of the POI and user space, the variational bottleneck's role in regularization and preventing overfitting to limited data, and the strength of 3PO in steadily adapting robustly to dynamic environments.

Conclusion

We propose a novel framework leveraging robust imitation learning for the next POI recommendation. The MOAIR model functions through a multi-objective IL agent, guided by an adversarial discriminator and multi-faceted, synergistic rewards. This IL agent is equipped with a self-supervised state encoder that provides informative representations of user-trajectory data, which utilizes N-GCNs and st-attention to model the user preferences. Additionally, we integrate a latent variable into the policy network and apply a variational information bottleneck to filter out noise and optimize each objective simultaneously through a novel 3PO architecture. Extensive experiments validate the effectiveness of MOAIR, which offers promising insights for future recommender systems, especially in situations where data availability and quality are not fully guaranteed.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 62102277 and Natural Science Foundation of Jiangsu Province under Grant No. BK20210703.

References

Alemi, A. A.; Fischer, I.; Dillon, J. V.; and Murphy, K. 2016. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*.

Batra, S.; Tjanaka, B.; Fontaine, M. C.; Petrenko, A.; Nikolaidis, S.; and Sukhatme, G. 2023. Proximal policy gradient arborescence for quality diversity reinforcement learning. *arXiv preprint arXiv:2305.13795*.

Christoforidis, G.; Kefalas, P.; Papadopoulos, A. N.; and Manolopoulos, Y. 2021. RELINE: point-of-interest recommendations using multiple network embeddings. *Knowledge and Information Systems*, 63: 791–817.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Hayes, C. F.; Rădulescu, R.; Bargiacchi, E.; Källström, J.; Macfarlane, M.; Reymond, M.; Verstraeten, T.; Zintgraf, L. M.; Dazeley, R.; Heintz, F.; et al. 2022. A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems*, 36(1): 26.

He, X.; Deng, K.; Wang, X.; Li, Y.; Zhang, Y.; and Wang, M. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, 639–648.

Ho, J.; and Ermon, S. 2016. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29.

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

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kong, D.; and Wu, F. 2018. HST-LSTM: A Hierarchical Spatial-Temporal Long-Short Term Memory Network for Location Prediction. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, 2341–2347. International Joint Conferences on Artificial Intelligence Organization.

Li, R.; Shen, Y.; and Zhu, Y. 2018. Next Point-of-Interest Recommendation with Temporal and Multi-level Context Attention. In *2018 IEEE International Conference on Data Mining (ICDM)*, 1110–1115.

Liu, B.; Qian, T.; Liu, B.; Hong, L.; You, Z.; and Li, Y. 2017. Learning spatiotemporal-aware representation for POI recommendation. *arXiv preprint arXiv:1704.08853*.

Liu, C. H.; Wang, Y.; Piao, C.; Dai, Z.; Yuan, Y.; Wang, G.; and Wu, D. 2020. Time-aware location prediction by convolutional area-of-interest modeling and memory-augmented attentive lstm. *IEEE Transactions on Knowledge and Data Engineering*, 34(5): 2472–2484.

Liu, Q.; Wu, S.; Wang, L.; and Tan, T. 2016. Predicting the Next Location: A Recurrent Model with Spatial and Temporal Contexts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

Liu, X.; and Wu, L. 2024. FAGRec: Alleviating data sparsity in POI recommendations via the feature-aware graph learning. *Electronic Research Archive*, 32(4): 2728–2744.

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

Mangalam, K.; and Garg, R. 2021. Overcoming mode collapse with adaptive multi adversarial training. *arXiv preprint arXiv:2112.14406*.

Peng, X. B.; Kanazawa, A.; Toyer, S.; Abbeel, P.; and Levine, S. 2018. Variational discriminator bottleneck: Improving imitation learning, inverse rl, and gans by constraining information flow. *arXiv preprint arXiv:1810.00821*.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

Rao, X.; Chen, L.; Liu, Y.; Shang, S.; Yao, B.; and Han, P. 2022. Graph-flashback network for next location recommendation. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 1463–1471.

Sánchez, P.; and Bellogín, A. 2022. 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.

Torabi, F.; Warnell, G.; and Stone, P. 2018. Behavioral cloning from observation. *arXiv preprint arXiv:1805.01954*.

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

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017a. Attention is all you need. *Advances in neural information processing systems*, 30.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017b. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wang, D.; Liu, K.; Xiong, H.; and Fu, Y. 2022. Online POI recommendation: Learning dynamic geo-human interactions in streams. *IEEE Transactions on Big Data*, 9(3): 832–844.

Wang, D.; Wang, P.; Liu, K.; Zhou, Y.; Hughes, C. E.; and Fu, Y. 2021a. Reinforced Imitative Graph Representation Learning for Mobile User Profiling: An Adversarial Training Perspective. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5): 4410–4417.

Wang, L.; Jin, B.; Huang, Z.; Zhao, H.; Lian, D.; Liu, Q.; and Chen, E. 2021b. Preference-adaptive meta-learning for cold-start recommendation. In *IJCAI*, 1607–1614.

Wang, P.; Liu, K.; Jiang, L.; Li, X.; and Fu, Y. 2020. Incremental mobile user profiling: Reinforcement learning with spatial knowledge graph for modeling event streams. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 853–861.

Wang, Z.; Zhu, Y.; Wang, C.; Ma, W.; Li, B.; and Yu, J. 2023. Adaptive Graph Representation Learning for Next POI Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 393–402.

Wu, Y.; Li, K.; Zhao, G.; and Qian, X. 2019. Long-and short-term preference learning for next POI recommendation. In *Proceedings of the 28th ACM international conference on information and knowledge management*, 2301–2304.

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

Xie, M.; Yin, H.; Wang, H.; Xu, F.; Chen, W.; and Wang, S. 2016. Learning graph-based poi embedding for location-based recommendation. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, 15–24.

Xiong, X.; Xiong, F.; Zhao, J.; Qiao, S.; Li, Y.; and Zhao, Y. 2020. Dynamic discovery of favorite locations in spatio-temporal social networks. *Information Processing & Management*, 57(6): 102337.

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

Zhuang, Z.; Wei, T.; Liu, L.; Qi, H.; Shen, Y.; and Yin, B. 2024. TAU: Trajectory Data Augmentation with Uncertainty for Next POI Recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20): 22565–22573.