STM-GAIL: Spatial-Temporal Meta-GAIL for Learning Diverse Human Driving Strategies

Yingxue Zhang
Binghamton University
yzhang42@binghamton.edu

Yanhua Li
Worcester Polytechnics Institute
yli15@wpi.edu

Xun Zhou
The University of Iowa
xun-zhou@uiowa.edu

Ziming Zhang
Worcester Polytechnics Institute
zzhang15@wpi.edu

Jun Luo
Lenovo Group Limited
jluo1@lenovo.com

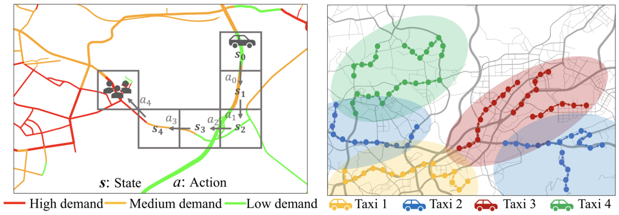
Abstract

With large amounts of human-generated spatial-temporal urban data (*e.g.*, GPS trajectories of vehicles, passengers' trip data on buses and trains, *etc.*), human urban strategy analysis has become an important problem in many urban scenarios. This problem is hard to solve due to two major challenges: (1) data scarcity (*i.e.*, each human agent can only provide limited observations) and (2) data heterogeneity (*i.e.*, having mixed observations from many different human agents). Most of the existing works on this problem usually require a large amount of historical observations aiming to correctly infer a human agent's urban strategy and thus fail to properly address both challenges at the same time. To solve the human urban strategy analysis problem in case of data scarcity and data heterogeneity, we design a novel learning paradigm — Spatial-Temporal Meta-GAIL (STM-GAIL), which can successfully learn diverse human urban strategies from heterogeneous human-generated spatial-temporal urban data. STM-GAIL models the human decision processes as variable length Markov decision processes (VLMDPs) and incorporates the surrounding spatial feature patterns (*e.g.*, traffic volume patterns, *etc.*) into states to better capture the spatial-temporal dependencies of human decisions. Besides, STM-GAIL learns diverse human urban strategies from the meta-learning perspective, and can distinguish various human urban strategies by adding an inference network on top of the standard GAIL. STM-GAIL can be quickly adapted to a new human expert's urban strategy with a single trajectory. Extensive experiments on real-world human-generated spatial-temporal dataset are performed.

Keywords—Generative adversarial imitation learning; meta-learning; human behavior analysis.

1 Introduction

In human urban decision-making processes (*e.g.*, taxi drivers' passenger-seeking processes as shown in Figure 1(a), *etc.*), human agents devise their own "strategies" to optimize their objectives (*e.g.*, maximizing revenue, minimizing travel time, *etc.*). These strategies are usually implicit to observers and even the agents themselves, which govern the daily mobility patterns of the human agents. *Human urban strategy analysis* aims at extracting and understanding how the human decisions are made using the observed human-generated spatial-temporal urban data (*e.g.*, GPS trajectories of taxis and personal vehicles, passengers' trip data on buses and trains, *etc.*). This is an important problem in many urban intelligence scenarios such as ride-sharing vehicle dispatching, public transportation management, and autonomous driving, *etc.*



(a) Taxi driver's passenger-seeking trajectory (b) Data scarcity and heterogeneity

Figure 1: Examples of taxi driver's decision-making process (left), data scarcity and heterogeneity challenges (right).

Challenges. The human urban strategy analysis problem is challenging due to the following two reasons. Firstly, learning from observations usually requires large amounts of training data from the agent being studied. However, in many urban cases, it is hard to collect abundant mobility data from a single human agent (*i.e.*, data scarcity challenge). Second, data collected from urban scenarios is often heterogeneous, meaning that it records behaviors of many different expert agents following various urban strategies (*i.e.*, data heterogeneity challenge, See Figure 1(b) for example). This makes it even harder to extract explicit and reliable strategies of a target agent.

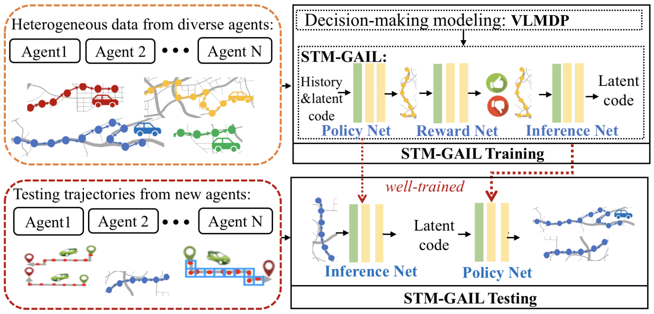


Figure 2: Solution framework.

Prior works. Over the last few years, many imitation learning algorithms have been proposed to conduct the human urban strategy analysis. For example, Ho et al. proposed generative adversarial imitation learning (GAIL) [10] which can successfully learn human decision-making strategies and accurately mimic human behaviors in various scenarios using deep neural networks (DNNs). Pan et al. proposed an Explainable xGAIL [19] to demonstrate the learning processes of GAIL in many real world cases. Zhang et al. extended the standard GAIL to conditional GAIL (cGAIL) [27] to unveil taxi drivers' driving policies by transferring knowledge across taxi drivers. Moreover, TrajGAIL [28] proposed by Zhang et al. incorporates the self-attention mechanism into GAIL to capture the long-term decision dependencies and learn the human decision strategies.

However, all these prior methods fail to properly address the above aforementioned challenges at the same time. These methods learn human urban strategies from scratch and require a large amount of historical behaviors of a single human agent, aiming to correctly infer her urban strategies. In case of data scarcity (*i.e.*, each agent has limited observations) and heterogeneity (*i.e.*, having mixed observations from many agents), such as inferring the driving strategies of a new taxi driver based on all drivers' trajectories, all these methods would fail.

Contributions. To solve the human urban strategy analysis problem in case of data scarcity and data heterogeneity, in this work, a novel learning paradigm —Spatial-Temporal Meta-GAIL (STM-GAIL) is proposed, which can successfully learn diverse human urban strategies from heterogeneous human-generated spatial-temporal urban data. Our solution framework is shown in Figure 2. STM-GAIL contains three model components including a policy network, a reward network and an inference network. It models the human urban decision-making processes as variable length Markov decision processes (VLMDPs). STM-GAIL learns diverse human urban strategies from a meta-learning perspective. Our main contributions can be summarized as follows:

- We make the first attempt to learn diverse human urban decision-making strategies in case of data scarcity and data heterogeneity, and propose the Spatial-Temporal Meta-GAIL (STM-GAIL), which incorporates the spatial-temporal dependencies of human decisions into GAIL framework by modeling the human urban decision-making processes as variable length Markov decision processes (VLMDPs) and taking the surrounding spatial feature patterns (*e.g.*, traffic volume patterns, travel demand patterns, *etc.*) as part of the states.
- STM-GAIL learns diverse human strategies from the meta-learning perspective, novel objective, architecture and algorithms are designed. In STM-GAIL, an inference network is designed on top of the standard GAIL, which infers the latent variables of diverse human strategies in an unsupervised way by maximizing the mutual information between the latent space and trajectories. STM-GAIL can be generalized to a new human expert's urban strategy with a single trajectory.
- Extensive experiments on real-world human-generated spatial-temporal dataset are performed to validate the effectiveness of our STM-GAIL. The experimental results show that our STM-GAIL has significant improvement compared to state-of-the-art baselines when learning human urban strategies.

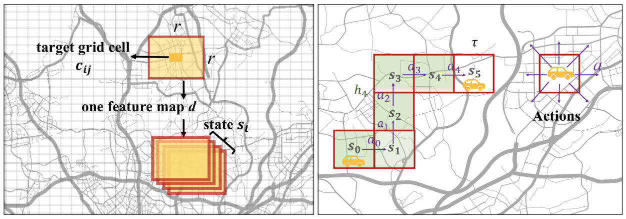
2 Preliminaries

Human-generated spatial-temporal urban data is collected from expert human agents to learn human urban decision-making strategies. In general, human-generated spatial-temporal urban data is a set of human mobility trajectories (*e.g.*, taxi GPS trajectories) which contains sequences of states and actions. In this section, we formally define our problem.

Definition 1 (Grid cells). A city is partitioned into $m_1 \times m_2$ grid cells, each grid cell has equal side-length in latitude and longitude. The set of grid cells of a city is defined as $\mathcal{C} = \{c_{ij}\}$, where $1 \le i \le m_1$ and $1 \le j \le m_2$. Each grid cell is associated with a set of features (*e.g.*, traffic speed, traffic volume, *etc.*) indicating the current status of the grid cell [29–31].

Definition 2 (States). As illustrated in Figure 3(a), a state at time t is defined as a multi-dimensional tensor $\boldsymbol{s}_t \in \mathbb{R}^{m \times r \times r}$, which is composed of m different feature maps $\boldsymbol{d} \in \mathbb{R}^{r \times r}$, each element $d_c \in \mathbb{R}$ inside a feature map $\boldsymbol{d}$ indicates a feature (*e.g.*, traffic speed, travel demand, *etc.*) of a grid cell at time t. The set of states is defined as $\mathcal{S} = \{\boldsymbol{s}_t\}$.

Definition 3 (Actions). An action a_t is a decision made by a human agent at state $\boldsymbol{s}_t$, which is governed by a specific urban strategy. By following an action a_t,



(a) Map gridding and state (b) Action, trajectory and history

Figure 3: Definition illustrations.

the human agent transits from the current state $\boldsymbol{s}_t$ to the next state $\boldsymbol{s}_{t+1}$ as shown in Figure 3(b). The set of actions is defined as $\mathcal{A} = \{a_t\}$.

Definition 4 (Trajectories and History). A trajectory τ is a sequence of states and actions that a human agent traverses and takes when completing a task, *i.e.*, $\tau = (\boldsymbol{s}_0, a_0, \cdots, \boldsymbol{s}_T, a_T)$. The history of a trajectory τ at time step t includes all states and actions prior to t, *i.e.*, $h_{t-1} = (\boldsymbol{s}_0, a_0, \cdots, \boldsymbol{s}_{t-1}, a_{t-1})$. The set of trajectories is defined as $\mathcal{T} = \{\tau\}$. The set of histories is defined as $\mathcal{H} = \{h\}$.

Definition 5 (Policy). The policy function $\pi : \mathcal{S} \times \mathcal{H} \mapsto [0, 1]$ controls what action to perform in each state, which is a probability distribution defined as $\pi(a_t \mid \boldsymbol{s}_t, h_{t-1})$ indicating the probabilities of choosing different actions given the current state $\boldsymbol{s}_t$ and the history h_{t-1}.

Definition 6 (Reward). The reward function is defined as $r : \mathcal{S} \times \mathcal{A} \times \mathcal{H} \mapsto \mathbb{R}$, *i.e.*, $r(\boldsymbol{s}_t, a_t \mid h_{t-1})$, which provides a numerical score based on a state $\boldsymbol{s}_t$ and an action a_t given the history h_{t-1}, and incentivizes a human agent to achieve a goal in a task.

Problem Statement. Given a set of heterogeneous trajectories $\mathcal{T}$ generated by a wide range of expert human agents, we aim to learn the diverse urban decision-making strategies of human experts, *i.e.*, the policy function $\pi(a \mid \boldsymbol{s}, h)$.

3 Methodologies

Built upon the standard generative adversarial imitation learning (GAIL [10]), we propose a novel STM-GAIL to learn diverse human urban strategies. STM-GAIL takes the spatial and temporal dependencies into consideration by incorporating the feature maps of surrounding areas into states and modeling the human decision-making processes as VLMDPs (See Section 3.1 and 3.3). Moreover, to tackle data scarcity and heterogeneity challenges, STM-GAIL learns diverse human urban strategies from the meta-learning perspective, an inference network is designed on top of the standard GAIL, which infers the latent variables of diverse human strategies in an unsupervised way. STM-GAIL can be generalized to a new human urban strategy with a single trajectory (See Section 3.2, 3.3 and 3.4).

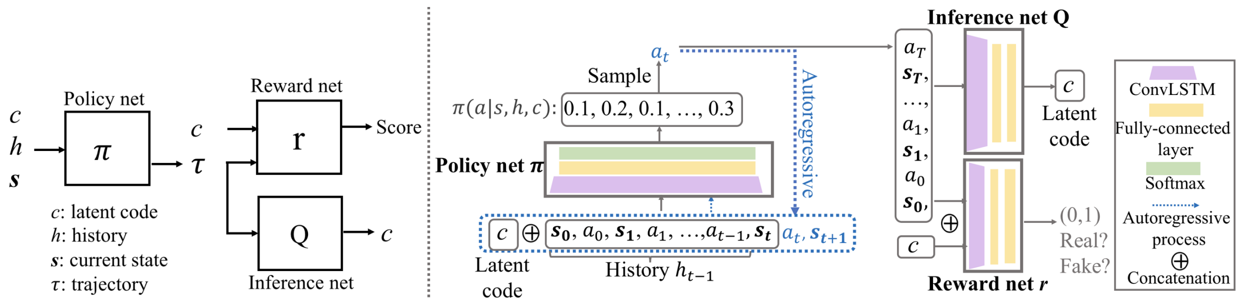
3.1 Modeling Human Sequential Decision-Making Process as VLMDP Over recent years, a large amount of works have focused on learning human decision-making strategies by modeling decision-making processes as Markov decision processes (MDPs) [22], which have a strong Markovian assumption [4], namely, an agent makes an action a_t only based on the current state $\boldsymbol{s}_t$ instead of any prior states and actions (*i.e.*, history h_{t-1}). However, in many urban scenarios, the Markov property does not hold. For example, as illustrated in Figure 1(a), when looking for a new passenger, a taxi driver's decisions of which direction to go not only depend on his current and previous locations, but also depend on the surrounding travel demand. Such spatial-temporal dependencies of human mobility are complicated and hard to capture when learning human urban strategies.

To capture the long-term dependency of human decisions, we model the decision-making process as a variable length Markov decision process (VLMDP), which includes an agent as the decision maker and an environment that interacts with the agent. A VLMDP is defined as a 5-tuple $\langle \mathcal{S}, \mathcal{A}, P, r, \gamma \rangle$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space; P denotes the transition function, *e.g.*, $P(\boldsymbol{s}_t | h_{t-1})$ is the transition probability of transiting to state $\boldsymbol{s}_t$ by following the history h_{t-1}; $r : \mathcal{S} \times \mathcal{A} \times \mathcal{H} \mapsto \mathbb{R}$ is the bounded reward function that outputs a reward value for a given state-action-history triple; $\gamma \in (0, 1]$ is a discount factor. The initial states are determined by the distribution $p(\boldsymbol{s}_0) : \mathcal{S} \mapsto [0, 1]$. The actions are chosen through a stationary and stochastic policy $\pi : \mathcal{S} \times \mathcal{H} \mapsto [0, 1]$. A decision making process forms a trajectory $\tau = (\boldsymbol{s}_0, a_0, \cdots, \boldsymbol{s}_T, a_T)$, where T is the terminal time step.

In this work, we use expectation with respect to a policy π to denote the expectation with respect to the trajectories it generates. For instance, $\mathbb{E}_\pi[r(\boldsymbol{s}, a \mid h)] = \mathbb{E}_{\boldsymbol{s}_t, h_{t-1}, a_t \sim \pi} \left[\sum_{t=0}^{T} \gamma^t r\left(\boldsymbol{s}_t, a_t \mid h_{t-1}\right) \right]$ denotes the following sampling processes including $\boldsymbol{s}_0 \sim p(\boldsymbol{s}_0), a_t \sim \pi\left(\cdot \mid \boldsymbol{s}_t, h_{t-1}\right)$, and $\boldsymbol{s}_t \sim P\left(\boldsymbol{s}_t \mid h_{t-1}\right)$. Each agent aims to maximize its expected cumulative reward $\mathbb{E}_\pi[r(\boldsymbol{s}, a \mid h)]$ by optimizing the policy π.

3.2 Objective In many previous MDP works [10, 19, 27, 32], the human strategy learning problem can be modeled as a constrained optimization problem as below:

$$
\begin{aligned}
& \max_{r} \min_{\pi} : -H(\pi), \\
(3.1) \qquad & \quad \text{s.t.} : \mathbb{E}_\pi[r(\boldsymbol{s}, a)] = \mathbb{E}_{\pi_E}[r(\boldsymbol{s}, a)], \\
& \qquad \sum_{a \in \mathcal{A}} \pi(a \mid \boldsymbol{s}) = 1, \quad \forall \boldsymbol{s} \in \mathcal{S}.
\end{aligned}
$$



(a) STM-GAIL Overview

(b) STM-GAIL Architecture

Figure 4: STM-GAIL structure.

However, Eq 3.1 does not consider any temporal dependencies of decisions. In this work, to incorporate the long-term temporal dependencies of human decisions and adapt the human strategy learning problem to VLMDP, the problem is re-formulated as Eq 3.2 [28]:

$$\begin{aligned}
&\max_r \min_\pi : -H(\pi),\\
&\qquad \text{s.t.} : \mathbb{E}_\pi[r(\boldsymbol{s},a \mid h)] = \mathbb{E}_{\pi_E}[r(\boldsymbol{s},a \mid h)],\\
&\qquad\qquad \sum_{a\in\mathcal{A}} \pi(a \mid \boldsymbol{s},h) = 1, \quad \forall \boldsymbol{s} \in \mathcal{S}.
\end{aligned} \tag{3.2}$$

In Eq 3.2, $H(\pi)$ is a γ-discounted causal entropy, which measures the uncertainty of a policy distribution $\pi(a \mid \boldsymbol{s},h)$, *i.e.*, $H(\pi) = \sum_{t=0}^{T}\sum_{h_t}\gamma^t\pi\left(a_t \mid \boldsymbol{s}_t,h_{t-1}\right)\log\pi\left(a_t \mid \boldsymbol{s}_t,h_{t-1}\right)$. π_E is the empirical policy observed from the collected human expert's mobility data. Eq 3.2 aims to find the policy $\pi(a \mid \boldsymbol{s},h)$ with maximum causal entropy $H(\pi)$, and find the reward function $r(\boldsymbol{s},a \mid h)$ such that the expected reward of a trajectory under π matches that under the empirical policy π_E.

To solve the human strategy learning problem defined in Eq 3.2, Zhang et.al [28] prove it is equivalent to solving a min-max problem as Eq 3.3:

$$\begin{aligned}
\min_{\pi\in\Pi}\max_r\, &-\lambda_1 H(\pi) + \mathbb{E}_{\pi_E}[\log(r(\boldsymbol{s},a \mid h))]\\
&+\mathbb{E}_\pi[\log(1-r(\boldsymbol{s},a \mid h))].
\end{aligned} \tag{3.3}$$

Apparently, Eq 3.3 is similar to the objective of generative adversarial networks (GANs) [8,16,17], and it is natural to employ the GAN framework, where the policy function π and the reward function r can be viewed as a generator and a discriminator, respectively. Eq 3.3 can capture the spatial-temporal dependencies of decisions, however, it cannot deal with the data scarcity and heterogeneity problems. Eq 3.3 can learn a single human expert's urban strategy with access to abundant trajectories, once we cannot collect enough historical trajectories from the human expert, this method would fail. Moreover, when facing the data heterogeneity problem, namely, the trajectories are collected from different human experts, Eq 3.3 would simply assume all trajectories are produced by one expert and fail to learn different urban strategies.

Thus, to deal with the data scarcity and heterogeneity problems, we introduce a latent variable c to our policy and reward functions, *i.e.*, $\pi(a|\boldsymbol{s},h,c)$ and $r(\boldsymbol{s},a \mid h,c)$, respectively. In general, $c \sim p(c)$ would be a latent vector representing a specific strategy of a human expert in the latent space. To enable the latent variable c to identify different strategies, we propose to add a mutual information regularizer to Eq 3.3 to encourage strong connections between c and the generated human trajectories. The mutual information between the latent variable and trajectories is denoted as $I(c;\tau)$, the objective with the mutual information regularizer is as follows:

$$\begin{aligned}
\min_{\pi\in\Pi}\max_r &\mathbb{E}_{\pi_E}[\log(r(\boldsymbol{s},a \mid h,c))] + \mathbb{E}_\pi[\log(1-r(\boldsymbol{s},a \mid h,c))]\\
&-\lambda_1 H(\pi) - \lambda_2 I(c;\tau).
\end{aligned} \tag{3.4}$$

In Eq 3.4, the latent variable c helps to identify different strategies in a heterogeneous dataset and also enable fast generalization to new strategies with few samples. However, it is hard to directly maximize the mutual information $I(c;\tau)$ without the access to the posterior distribution $P(c|\tau)$. Instead, we calculate the variational lower bound [2,21] of $I(c;\tau)$ and use an auxiliary distribution $Q(c|\tau)$ to approximate the true posterior $P(c|\tau)$:

$$\begin{aligned}
I(c;\tau) &= H(c) - H(c \mid \tau)\\
&= \mathbb{E}_{\tau\sim\pi(\cdot|\boldsymbol{s},h,c),c'\sim P(c|\tau)}\left[\log P\left(c' \mid \tau\right)\right] + H(c)\\
&= \mathbb{E}_{\tau\sim\pi(\cdot|\boldsymbol{s},h,c)}[\underbrace{D_{\mathrm{KL}}(P(\cdot \mid \tau)\|Q(\cdot \mid \tau))}_{\geq 0}\\
&\quad+\mathbb{E}_{c'\sim P(c|\tau)}[\log Q(c' \mid \tau)]] + H(c)\\
&\geq \mathbb{E}_{c\sim p(c),\tau\sim\pi(\cdot|\boldsymbol{s},h,c)}[\log Q(c \mid \tau)] + H(c)\\
&= L_I(\pi,Q),
\end{aligned} \tag{3.5}$$

where $p(c)$ is a prior distribution, Q is the auxiliary distribution, and we can treat Q as an inference neural network, which uses τ to infer c. As a result, the final objective for our Spatial-Temporal Meta-GAIL (STM-

GAIL) is as Eq 3.6:

$$\min_{\pi,Q}\max_{r}\mathbb{E}_{\pi_E}[\log(r(\boldsymbol{s},a\mid h,c))]+\mathbb{E}_{\pi}[\log(1-r(\boldsymbol{s},a\mid h,c))] - \lambda_1 H(\pi)-\lambda_2 L_I(\pi,Q). \tag{3.6}$$

3.3 Model Architecture In our final objective Eq 3.6, a policy network π, a reward network r and an inference network Q are required. Figure 4 shows the detailed architecture of STM-GAIL, which applies ConvLSTM [25] inside each model component to better capture the spatial-temporal dependencies of human decisions in a trajectory.

The policy network π outputs an action distribution $\pi(a_t|\boldsymbol{s}_t,h_{t-1},c)$ based on the current state $\boldsymbol{s}_t$, the history h_{t-1} and the latent vector c. A specific action a_t will be sampled from the distribution, *i.e.*, $a_t \sim \pi(a_t|\boldsymbol{s}_t,h_{t-1},c)$. Given the sampled action a_t, the next state $\boldsymbol{s}_{t+1}$ is directly provided by the environment (through the transition function $P(\boldsymbol{s}_{t+1}|h_t)$[1]), which is combined with the extended history h_t and latent vector c as the new input of the policy network, *i.e.*, $a_{t+1} \sim \pi(a_{t+1}|\boldsymbol{s}_{t+1},h_t,c)$. Thus, the policy network works in an auto-regressive way. Inside the policy network π, the current state $\boldsymbol{s}_t$ and the latent vector c are concatenated together and pass a ConvLSTM, the history h_{t-1} is stored within the hidden states [25] of ConvLSTM. The output of the ConvLSTM passes a fully-connected layer and a softmax function [15] to get the probabilities of choosing different actions.

The reward network r can be viewed as a discriminator, which aims to distinguish the positive data from the negative data by giving high scores if the input τ is collected from expert human agents, and giving low scores if the input τ is generated by the policy network. The input of the reward network includes i) the current state $\boldsymbol{s}_t$ and action a_t, ii) the history h_{t-1} and iii) the latent vector c. The output of the network is a score from 0 to 1. Inside the reward network r, all the states, actions and the latent vector are concatenated together and pass a ConvLSTM and two fully-connected layers, the output is activated by Sigmoid function [9].

The inference network Q aims to infer the distribution of latent vector c using the generated trjectory τ. Q takes a trajectory generated by the policy network as the input, and outputs a latent vector c. Inside the inference network Q, the input trajectory passes a ConvLSTM and two fully-connected layers activated by hyperbolic tangent function [11].

3.4 Training and Testing Algorithms To optimize Eq 3.6, novel training and testing algorithms are proposed.

STM-GAIL Training algorithm. In Eq 3.6, a prior distribution $p(c)$ is required. However, for most urban scenarios, we do not have access to $p(c)$ but instead have human agent trajectories sampled from $\mathcal{T}$, we use the following generative process:

$$\tau \sim \mathcal{T}, c \sim Q(c \mid \tau) \tag{3.7}$$

to synthesize latent variables, which approximates the prior distribution when π and Q are trained to optimality, the effectiveness has been validated by Yu et al. [26]. The detailed training process is in Algorithm 1.

Algorithm 1 STM-GAIL Training Process

```
Input: Trajectories collected from diverse human experts $\mathcal{D} = \{\tau_i\}$, initial parameters of policy network, reward network and inference network $\boldsymbol{\theta}_0$, $\boldsymbol{\omega}_0$, $\boldsymbol{\psi}_0$.
Output: Learned policy network $\pi_{\boldsymbol{\theta}}$, reward network $r_{\boldsymbol{\omega}}$ and inference network $Q_{\boldsymbol{\psi}}$.
repeat
    Sample two batches of trajectories $\tau_E$ and $\tau'_E$: $\tau_E, \tau'_E \sim \mathcal{D}$
    Infer a batch of latent codes: $c \sim Q_{\boldsymbol{\psi}}(c \mid \tau_E)$.
    Sample trajectories $\tau$ using the policy network $\pi_{\boldsymbol{\theta}}$ with the latent code fixed during each rollout, i.e. $\tau \sim \pi_{\boldsymbol{\theta}}(\tau \mid c)$.
    Update $\boldsymbol{\omega}$ to maximize Eq. 3.8 with $\tau'_E$ and $\tau$.
    Update $\boldsymbol{\psi}$ to minimize Eq. 3.9 with $\tau$ .
    Update $\boldsymbol{\theta}$ with TRPO [24] to minimize Eq. 3.10.
until Convergence
```

Based on Eq. 3.6, we can get the objective functions for π, r and Q separately. Denote $\boldsymbol{\omega}$ as the parameters of reward network r, η as the learning rate, we update the reward network with Eq. 3.8:

$$\begin{aligned}&\mathcal{L}_r(\boldsymbol{\omega}) = \mathbb{E}_{\pi_E}[\log(r_{\boldsymbol{\omega}})] + \mathbb{E}_{\pi}[\log(1-r_{\boldsymbol{\omega}})],\\ &\boldsymbol{\omega} = \boldsymbol{\omega} + \eta\nabla_{\boldsymbol{\omega}}\mathcal{L}_r(\boldsymbol{\omega}).\end{aligned} \tag{3.8}$$

Denote $\boldsymbol{\psi}$ as the parameters of the inference network Q, we update Q with Eq. 3.9:

$$\begin{aligned}&\mathcal{L}_Q(\boldsymbol{\psi}) = -\lambda_2\mathbb{E}_{c\sim p(c),\tau\sim\pi(\cdot|\boldsymbol{s},h,c)}[\log Q_{\boldsymbol{\psi}}(c \mid \tau)],\\ &\boldsymbol{\psi} = \boldsymbol{\psi} - \eta\nabla_{\boldsymbol{\psi}}\mathcal{L}_Q(\boldsymbol{\psi}).\end{aligned} \tag{3.9}$$

Denote $\boldsymbol{\theta}$ as the parameters of the policy network π, our goal is to minimize the objective for $\pi_{\boldsymbol{\theta}}$ using Trusted Region Policy Optimization (TRPO) [24], the objective for $\pi_{\boldsymbol{\theta}}$ is as below:

$$\mathcal{L}_\pi(\boldsymbol{\theta}) = \mathbb{E}_{\pi_{\boldsymbol{\theta}}}[\log(1-r)] - \lambda_1 H(\pi_{\boldsymbol{\theta}}) - \lambda_2 L_I(\pi_{\boldsymbol{\theta}}, Q). \tag{3.10}$$

[1] In this work, we are in a model-free setup, thus, we do not need access to the transition function [7].

Algorithm 2 STM-GAIL Testing Process

```
Input: Trajectories $\mathcal{D}_{\text{Test}} = \{\tilde{\tau}_i\}$ collected from diverse new human experts (each expert only provides one single trajectory $\tilde{\tau}_i$), learned policy network $\pi_{\boldsymbol{\theta}}$ and inference network $Q_{\boldsymbol{\psi}}$.
Output: Generated trajectories for each expert.
repeat
    Infer the latent code $\tilde{c}_i$ from $\tilde{\tau}_i$: $\tilde{c}_i \sim Q_{\boldsymbol{\psi}}(c \mid \tilde{\tau}_i)$.
    Generate trajectories $\hat{\tau}$ for the human expert using $\pi_{\boldsymbol{\theta}}$ with $\tilde{c}_i$ fixed during each rollout, i.e. $\hat{\tau} \sim \pi_{\boldsymbol{\theta}}(\tau \mid c_i)$.
until Testing finished for $\mathcal{D}_{\text{Test}}$
```

STM-GAIL Testing algorithm. During the testing process, we have the trajectories collected from diverse new human experts. We first use the well-trained $Q_{\boldsymbol{\psi}}$ to infer the corresponding latent vector from a trajectory, and then use the learned $\pi_{\boldsymbol{\theta}}$ and the latent vector to produce more trajectories which are similar to the real ones governed by the real policy. The detailed testing algorithm is in Algorithm 2.

4 Evaluation

In this section, we introduce the real-world dataset, baseline models and the metrics that we use to evaluate our STM-GAIL, and present extensive experimental results.

4.1 Data and Experiment Description In our experiment, we aim to learn the taxi drivers' passenger-seeking strategies from the collected passenger-seeking trajectories.

Dataset Description. The passenger-seeking trajectories are collected from 17,877 taxis in Shenzhen, China from July 1 to Sep 31, 2016. Each passenger-seeking trajectory is formed by multiple GPS records of a taxi. A GPS record includes five attributes including the taxi plate ID, longitude, latitude, time stamp and passenger indicator which is a binary value indicating whether a passenger is on board.

State Space. We first partition the Shenzhen City into 40 × 50 equal-sized grid cells with a side-length $l_1 = 0.0084°$ in latitude and $l_2 = 0.0126°$ in longitude. And we divide the time of a day into five-minute time slots. A state of a grid cell is defined as a multi-dimensional tensor which is composed of different feature maps of its neighboring 5 × 5 grid cells in a specific time slot.

Action Space. When a taxi is in a specific state, the taxi driver has 10 actions to choose from, including going to 8 neighboring grid cells, staying at the current grid cell, and terminating the trip.

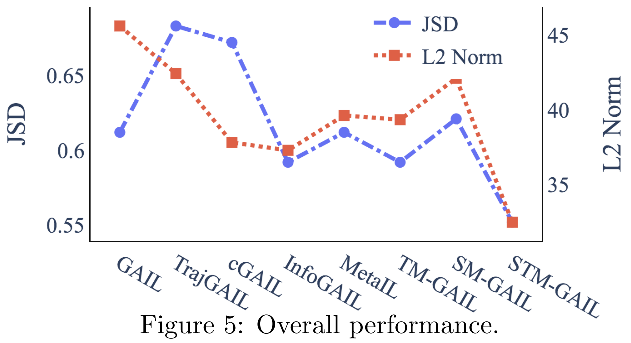


Figure 5: Overall performance.

Experiment Description. In this experiment, we study how taxi drivers make decisions when seeking passengers. Given the historical trajectories of different expert drivers in Shenzhen, China, the state space and the action space, we aim to learn the passenger-seeking strategies for diverse taxi drivers. All the expert drivers and their historical trajectories are randomly split into training set (85%) and testing set (15%).

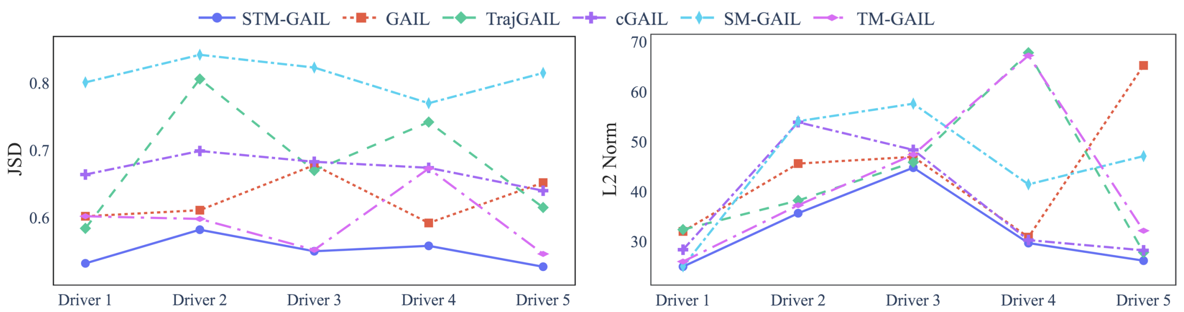
4.2 Baselines To evaluate our model, we compare STM-GAIL with state-of-the-art imitation learning methods. Firstly, to validate that the standard imitation learning methods cannot learn diverse human decision-making strategies, we compare our proposed STM-GAIL with standard **GAIL** [10] and **TrajGAIL** [28]. Next, we compare our STM-GAIL with state-of-the-art meta imitation learning methods including **cGAIL** [27], **InfoGAIL** [14] and **MetaIL** [6], which do not consider the spatial-temporal dependencies in the human decision-making processes. Moreover, to validate both spatial and temporal dependencies are important when learning diverse human strategies, we have two baseline models including **Temporal Meta-GAIL (TM-GAIL)** [14,28] and **Spatial Meta-GAIL (SM-GAIL)** [14,28]. TM-GAIL and SM-GAIL has the same objective as our STM-GAIL. However, TM-GAIL ignores the spatial patterns in the decision-making processes, SM-GAIL ignores the long-term dependencies of decisions,

4.3 Evaluation Metrics In our experiment, we use two metrics to evaluate our STM-GAIL including i) Jensen–Shannon (JS) divergence ii) L_2-Norm:

Jensen–Shannon divergence. JS divergence is a method of measuring the similarity between two probability distributions P and Q:

$$(4.11)\quad JSD(P||Q) = H(\frac{P+Q}{2}) - \frac{1}{2}(H(P) + H(Q)),$$

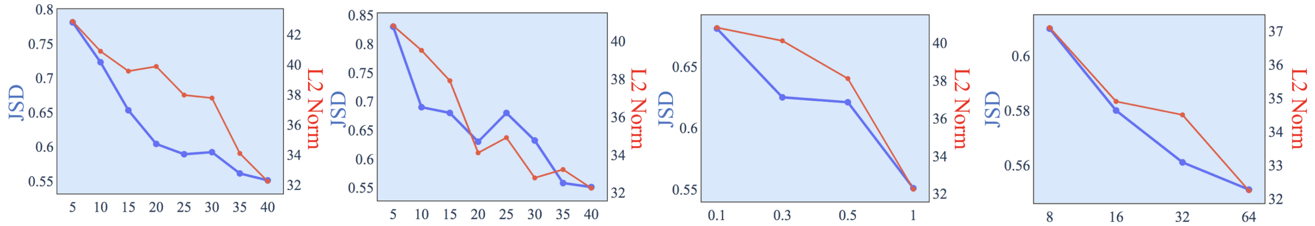
where $H(P)$ is the Shannon entropy for distribution P. In our experiments, JS divergence is used to measure the similarity between the learned policy (*i.e.*, π) and the empirical ground-truth policy (*i.e.*, π_E).



(a) JS Divergence for different testing drivers

(b) L2 Norm for different testing drivers

Figure 6: Performance on learning diverse strategies.



(a) Number of drivers in training set

(b) Number of training trajectories per driver

(c) λ_2

(d) Batch size

Figure 7: Impact of hyper-parameters on urban strategies learning with STM-GAIL.

L_2-Norm. L_2-Norm is used to measure the distance between the trajectories generated by the learned policy (*e.g.*, $P = (p_1, \cdots, p_n)$) and the trajectories sampled from the empirical ground-truth policy (*e.g.*, $Q = (q_1, \cdots, q_n)$). L_2-Norm is defined as below:

$$L_2(P, Q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}. \tag{4.12}$$

4.4 Experimental Settings

In the experiment, we parametrize the auxiliary distribution $Q(c \mid \tau)$ as a neural network, and its form depends on the true posterior $P(c \mid \tau)$. We found that simply treating $Q(c \mid \tau)$ as a factored Gaussian distribution is sufficient. For all experiments, we use Adam [12] for online optimization. During training, the batch size is set to 64, and the learning rate is 1×10^{-5}.

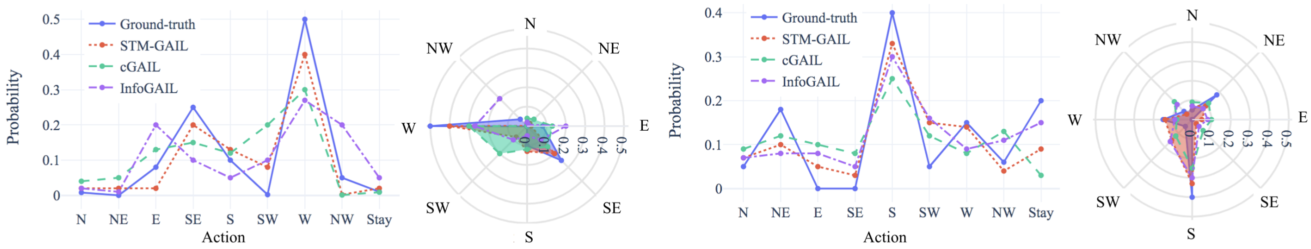
4.5 Results

4.5.1 Overall performance

We first present the overall performance of our STM-GAIL compared with all baseline models when learning taxi drivers' diverse passenger-seeking strategies. As shown in Figure 5, compared with our STM-GAIL, we find the imitation learning methods including GAIL and TrajGAIL have higher errors for both metrics, which indicates the two models cannot distinguish different strategies and fail to learn diverse human strategies, since they assume all training trajectories are from one single expert driver. Besides, the meta imitation learning methods including cGAIL, InfoGAIL and MetaIL do not present decent performance, since they simply model the human decision-making processes as MDPs and ignore the complex spatial-temporal dependencies of human decisions in the urban scenario, which usually leads to poor performance when learning diverse urban strategies. The higher errors of TM-GAIL and SM-GAIL indicate both spatial and temporal dependencies are important when learning diverse human urban strategies. In our STM-GAIL, it can successfully capture spatial-temporal dependencies of taxi drivers' passenger-seeking decisions and also distinguish different human expert strategies with very limited data.

4.5.2 Performance on learning diverse strategies.

Next, we validate whether STM-GAIL can accurately learn different human urban strategies for each individual. As shown in Figure 6, we compare our STM-GAIL with some competitive baseline models. For each testing driver in our testing set, STM-GAIL presents the lowest errors for both metrics (see Figure 6(a) and Figure 6(b)). GAIL and TrajGAIL have relatively higher errors in JS Divergence and L_2 Norm compared with our STM-GAIL, since they do not address the data heterogeneity problem; cGAIL still presents high errors in both metrics, since it cannot learn the unstructured patterns and connections among strategies. Besides, SM-GAIL and TM-GAIL ignore either the temporal dependencies or spatial patterns in the decision-making process and thus produces poor performance. By contrast, our STM-GAIL learns the unstructured patterns of di-



(a) Case 1: action selection for driver 1

(b) Case 2: action selection for driver 2

Figure 8: Case studies: learned policies vs. ground-truth policies for two taxi drivers in two cases.

verse strategies using an inference network, and models the decision-making processes as VLMDPs, which guarantee the good performance.

4.5.3 Ablation Study We also study how hyperparameters influence the strategy learning performance in our STM-GAIL. As shown in Figure 7(a), we find if the training trajectories are collected from more taxi drivers, the learned policy would be better adapted to different testing drivers' strategies. In Figure 7(b), we find if each driver provides more trajectories in the training process, STM-GAIL can learn diverse driving strategies better and thus produce lower errors. In Figure 7(c), we can find the performance is sensitive to the value of λ_2, λ_2 should be chosen based on the loss scale to ensure the whole loss scale keeps the same, in our experiments, the best choice of λ_2 should be 1. In Figure 7(d), we find large batch size results in good performance in our experiments.

4.5.4 Case Study To further investigate how STM-GAIL performs when learning different urban strategies in different scenarios and urban states, we have a few representative case studies. We first select two different taxi drivers and get their empirical policies from their mobility data. For a specific state, we present the probabilities of choosing different actions using the learned policies of STM-GAIL, cGAIL and InfoGAIL. We find for both drivers in two different states (see Figure 8(a) and Figure 8(b)), baseline models do not present stable performance in action selection, and action-chosen probabilities produced by their policies are greatly different from the ground-truth policy. By contrast, the policies learned by STM-GAIL match the ground-truth policies very well, which indicate our STM-GAIL can successfully learn diverse urban strategies.

5 Related Work

Imitation Learning. Imitation learning aims to learn the policies from expert demonstrations. Most of the imitation learning methods model the decision making processes as Markov Decision Processes (MDPs) [1,20]. For example, GAIL [10] borrows the generative adversarial networks (GANs) framework to learn the policies of experts from demonstrations. Many works extend the GAIL framework to diverse urban applications. For example, Kuefler et al. try to use GAIL to imitate the drivers' behaviors in autonomous driving [13]. Zhang et al. extend the standard GAIL to conditional GAIL (cGAIL) [27]. TrajGAIL [28] models the decision making processes as variable length Markov decision processes (VLMDPs) to capture the long-term decision dependencies. However, all these methods learn strategies from scratch and require a large amount of demonstrations, and cannot learn the diverse urban strategies directly.

Meta Learning. Meta learning [5] tries to learn a generalized model from training tasks which can be fast adapted into new related tasks with a few samples. Meta learning has been applied to many areas including imitation learning. In meta imitation learning, many prior works focus on learning diverse tasks from mixed experts' demonstrations [18, 23]. Moreover, one-shot imitation learning [3,6] demonstrates impressive results on learning new tasks using a single demonstration, however, it requires a large amount of training tasks and needs prior knowledge on the task distribution. All these works did not consider the uniqueness of learning urban strategies, and cannot successfully capture the spatial-temporal dependencies of human decisions.

6 Conclusion

In this paper, we make the first attempt to solve the human urban strategy analysis problem in case of data scarcity and data heterogeneity, and propose a novel imitation leaning paradigm —Spatial-Temporal Meta-GAIL (STM-GAIL), which can successfully learn diverse human urban strategies from heterogeneous human-generated spatial-temporal urban data. In our STM-GAIL, we incorporate the spatial-temporal dependencies of human decisions into GAIL framework, and propose to learn diverse human urban strategies from the meta-learning perspective. STM-GAIL can be generalized to a new human urban strategy with a single trajectory. Extensive experiments on real-world dataset are performed to prove the effectiveness of STM-GAIL.

References

[1] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. ICML '04, 2004.

[2] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *CoRR.*

[3] Y. Duan, M. Andrychowicz, B. C. Stadie, J. Ho, J. Schneider, I. Sutskever, P. Abbeel, and W. Zaremba. One-shot imitation learning, 2017.

[4] R. Durrett. *Probability: Theory and Examples.* 2010.

[5] C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks, 2017.

[6] C. Finn, T. Yu, T. Zhang, P. Abbeel, and S. Levine. One-shot visual imitation learning via meta-learning. In *Proceedings of the 1st Annual Conference on Robot Learning.*

[7] J. Gläscher, N. Daw, P. Dayan, and J. P. O'Doherty. States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron.*

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NeurIPS*. 2014.

[9] J. Han and C. Moraga. The influence of the sigmoid function parameters on the speed of backpropagation learning. In *Proceedings of the International Workshop on Artificial Neural Networks.*

[10] J. Ho and S. Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, 2016.

[11] B. Kalman and S. Kwasny. Why tanh: choosing a sigmoidal function. In *IJCNN International Joint Conference on Neural Networks.*

[12] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[13] A. Kuefler, J. Morton, T. Wheeler, and M. Kochenderfer. Imitating driver behavior with generative adversarial networks. In *2017 IEEE Intelligent Vehicles Symposium (IV).*

[14] Y. Li, J. Song, and S. Ermon. Infogail: Interpretable imitation learning from visual demonstrations. NIPS'17, 2017.

[15] W. Liu, Y. Wen, Z. Yu, and M. Yang. Large-margin softmax loss for convolutional neural networks.

[16] M. Mirza and S. Osindero. Conditional generative adversarial nets. In *CoRR abs/1411.1784*, 2014.

[17] O. Mogren. C-RNN-GAN: continuous recurrent neural networks with adversarial training. *CoRR*, 2016.

[18] T. Munkhdalai and H. Yu. Meta networks. In *ICML.*

[19] M. Pan, W. Huang, Y. Li, X. Zhou, and J. Luo. Xgail: Explainable generative adversarial imitation learning for explainable human decision analysis. In *KDD, 2020.*

[20] D. A. Pomerleau. Efficient training of artificial neural networks for autonomous navigation. *Neural Computation.*

[21] B. Poole, S. Ozair, A. van den Oord, A. A. Alemi, and G. Tucker. On variational bounds of mutual information, 2019.

[22] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* John Wiley amp; Sons, Inc.

[23] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap. Meta-learning with memory-augmented neural networks. In *ICML.*

[24] J. Schulman, S. Levine, P. Moritz, M. Jordan, and P. Abbeel. Trust region policy optimization. ICML'15.

[25] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W. chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 2015.

[26] L. Yu, T. Yu, C. Finn, and S. Ermon. Meta-inverse reinforcement learning with probabilistic context variables. In *NeurIPS.*

[27] X. Zhang, Y. Li, X. Zhou, and J. Luo. cgail: Conditional generative adversarial imitation learning—an application in taxi drivers' strategy learning. *IEEE TBD.*

[28] X. Zhang, Y. Li, X. Zhou, Z. Zhang, and J. Luo. Trajgail: Trajectory generative adversarial imitation learning for long-term decision analysis. In *2020 ICDM.*

[29] Y. Zhang, Y. Li, X. Zhou, X. Kong, and J. Luo. Curbgan: Conditional urban traffic estimation through spatio-temporal generative adversarial networks. In *KDD, 2020.*

[30] Y. Zhang, Y. Li, X. Zhou, X. Kong, and J. Luo. TrafficGAN: Off-deployment traffic estimation with traffic generative adversarial networks. In *ICDM*, 2019.

[31] Y. Zhang, Y. Li, X. Zhou, Z. Liu, and J. Luo. C^3-GAN: Complex-condition-controlled urban traffic estimation through generative adversarial networks. In *2021 ICDM.*

[32] B. D. Ziebart, A. Maas, J. A. Bagnell, and A. K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd National Conference on Artificial Intelligence.*