Doctoral Dissertation

STUDY OF SUSTAINABLE ROUTE PLANNING FOR MULTIPLE TOURISTS WITH MULTI-AGENT REINFORCEMENT LEARNING

KONG Yuntao

Supervisor: NGUYEN, Minh Le

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
[Information Science]

December 2024

Abstract

The rapid expansion of the tourism industry has led to the growing research focused on tourist route planning. However, most existing studies concentrate on individual tourist routing, leaving a significant gap in addressing scenarios that involve multiple tourists. Traditional approaches for multi-tourist route planning, often adapted from single-tourist models, tend to emphasize tourist preferences and advantages. This has resulted in challenges such as popularity-biased route planning, which intensifies issues like overtourism in highly popular areas and hinders sustainable tourism practices. To overcome these challenges, we propose a multi-agent reinforcement learning (MARL) framework for planning routes for multiple tourists, integrating tourist distribution into the process. Our method comprises two essential components: first, a novel reinforcement learning environment tailored for tourism, allowing interactions with multiple tourists; second, a dual-congestion model that accounts for both localized congestion at attractions and the broader city-wide distribution of tourists. This dual-congestion concept formulates the reward structure within our MARL framework. We validate our approach through extensive experiments using real-world human mobility data from Kyoto, a renowned global tourist destination. The results demonstrate that our model outperforms current approaches in optimizing route rewards while managing tourist distribution effectively. Furthermore, we conducted a user study to assess the impact of our congestion-aware mechanism on tourist experiences. The findings suggest that while our dual-congestion model may slightly impact tourists who favor popular destinations, it underscores the generally conflicting relationship between sustainable tourism and individual tourist preferences. Importantly, our model shows potential in transforming this conflict into a more cooperative interaction.

Additionally, we explore multi-agent communication protocols. To alleviate the non-stationary problem in MARL, we employ techniques to denoise irrelevant information and perform information fusion effectively. Our implementation of two types of selectors and three attention-based methods shows the framework's capability to handle large-scale agents' interaction. Moreover, experiments indicate that traditional methods provide limited improvements for the non-stationary challenges in our scenario, pointing to future research directions focusing on sequential actions of agents and the adaptation of joint optimization in collaborative-adversarial scenarios. Next, we reveal the similarity between multi-agent communication and Multi-

hop Question Answering (QA), and apply our proposed communication framework on Multi-hop QA. We develop the advancements in Multi-hop QA by developing the "Answer Multi-hop questions by Single-hop QA" (AMS) system. This innovative approach employs a denoise component and a single-hop QA model adopting the co-attention and self-attention architecture. Our AMS system outperforms existing GNN-based models on the HotpotQA dataset, showcasing improvements in Joint EM and Joint F1 scores while using fewer resources. It illustrates our framework's effectiveness in other complicated task.

In summary, this research advocates a comprehensive approach for multiple tourists route planning with MARL. This work establishes a robust and collaborative framework for addressing the complex issue of popularity-biased tourists route planning, significantly advancing the capabilities for achieving sustainable tourism and efficient information sharing in complex environments.

Keywords: Multiple Tourists Route Planning, Multi-agent Reinforcement Learning, Multi-agent System Communication, Unbiased Route Planning, Sustainable Tourism Sightseeing.

Acknowledgment

Throughout the journey of my doctoral research, I have been fortunate to receive support and assistance from many individuals. I would like to extend my sincerest gratitude to everyone who has contributed to this process.

I would like to express my heartfelt thanks to my advisor, Professor NGUYEN, Minh Le. Your academic guidance has been instrumental in shaping my research. Your broad academic vision and dedication to excellence have been a source of lifelong inspiration for me.

I am also deeply grateful to Professor Ma Qiang and every member of his research team. It has been a privilege to collaborate and exchange ideas with all of you throughout my research journey.

I must also express my gratitude to my family and friends. Your understanding, support, and patience have allowed me to fully dedicate myself to my academic pursuits.

To all who have been a part in my journey, I am profoundly grateful.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

1.1 Motivation

Tourism not only fosters job creation by stimulating local economies through associated industries but also serves as a crucial bridge for cultural exchange and understanding by exposing travelers to diverse cultural heritages. Recently, tourism industry has been acting an indispensable part of the global economy, promoting economic development in numerous countries and regions. Annually, the increasing number of global tourists and the growth in contribution to the GDP highlight the industry's significant influence. After COVID-19 pandemic, the growth in tourism expenditure is seven times higher than nominal GDP growth. Figure 1.1 shows the statistical comparison between global nominal GDP growth and nominal tourism spending growth from Tourism Economics [1].

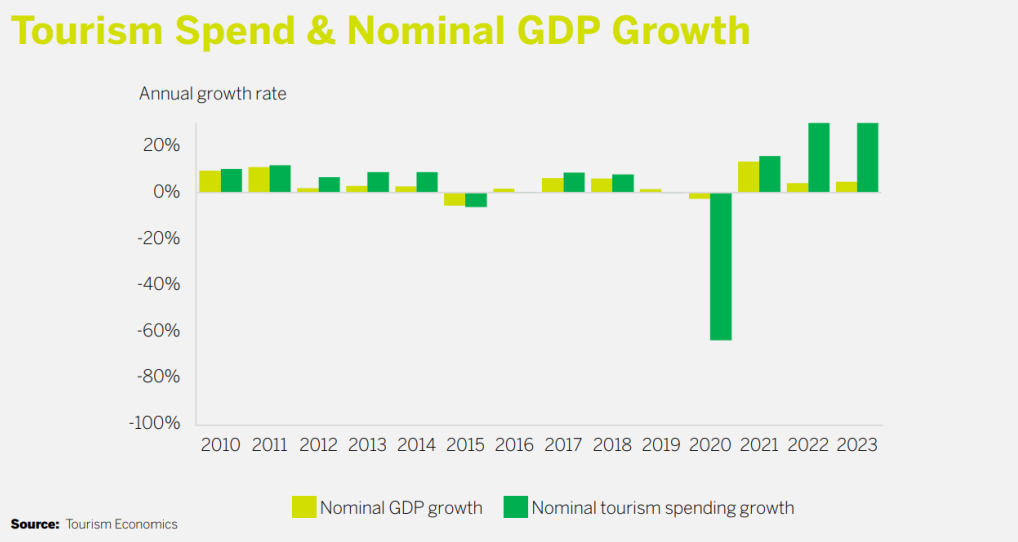


Figure 1.1: Comparison between global nominal GDP growth and nominal tourism spending growth from Tourism Economics [1].

Figure 1.2: Overtourism problem in Kyoto (Japan) and Rome (Italy).

However, the rapid expansion of tourism presents some challenges in tourism destinations, including environmental pollution, cultural conflicts, and resource over-exploitation, which are generally caused by overtourism (tourism pollution). The phenomenon of overtourism, where visitor numbers greatly exceed a destination's capacity, leads to negative impact on the local residents, environment and culture, which ultimately harms both the visitor experience and tourism sustainable development. Conversely, under-tourism, a strongly related issue with overtourism, represents a significant issue in lesser-known tourist spots that fail to attract adequate visitor numbers due to insufficient promotion or resource allocation, thus limiting economic growth and leading to under-investment and infrastructural deficiencies. The root of these issues stems from the uneven distribution of tourists, largely a result of biased-sightseeing influenced by media portrayals and promotional strategies that favor certain regions. This bias not only limits tourist choices but also exacerbates pressures on popular destinations while hindering the development of other lesser-known ones. Figure 1.2 shows overtourism in world famous tourism spots.

Research of tourist route planning has been boosted by the rapid growth of tourism in the last decades. However, there are three limitations in existing research regarding the above issues:

- Existing research only focuses on popular point-of-interest (POI) congestion, addressing crowding at famous destinations but often neglecting less popular ones, thereby overlooking under-tourism. It results in a missed opportunity to balance regional development and enhance local economies. At the core of the problem is the inequitable distribution of tourism resources, which can lead to environmental destruction and degradation of cultural heritage, thereby contributing to unsustainable tourism practices.

- Existing research on tourist route planning predominantly focuses on single tourist planning, resulting in homogenized itineraries and POI congestion. While some recent studies have begun to explore multiple tourist planning, they typically prioritize tourist benefits without addressing the interests of the destinations. Our research aims to systematically resolve bias issues by considering both tourist and destination interests, thereby promoting sustainable tourism.

- Multi-Agent Reinforcement Learning (MARL) has not been extensively explored in the route planning for multiple tourists, where existing studies relying on single-agent frameworks yield suboptimal results. Our research is dedicated to the application of MARL in planning routes for multiple tourists, employing Kyoto as a practical example to tackle real-world challenges.

Our research proposes a systematic approach to alleviate overtourism by addressing the biased-sightseeing issue, aiming to achieve equitable distribution of tourists. This not only facilitates tourists in discovering new destinations but also contributes to the sustainability of the tourism and provides new strategies for local governments and tourism organizations. Consequently, the objective of our research is to establish a scalable and robust model utilizing MARL to effectively manage route planning for multiple tourists. This model takes into account both sustainable tourism and tourists' preferences, striving to achieve unbiased sightseeing and promote sustainable tourism practices.

1.2 Research problems and Objectives

We focus on three research problems as follows:

RQ1. How to solve the unbiased-sightseeing problem with MARL?

Objective: we model the multiple tourists route planning with MARL, where each agent represents a tourist. We propose a mechanism that considers both local congestion at tourist spots and overall city-wide tourist' distribution to formulates the reward system in our MARL framework. This mechanism should guide the MARL algorithm to generate routes with an even and equatable distribution of tourists.

RQ2. How to conduct communication in multi-agent system considering agents' scalability?

Objective: we propose an method that can improve the efficiency in large-scalar agents' communication. Our model consists of two main parts that first denoise irrelevant agents and then conduct information sharing.

RQ3. Could we extend our multi-agent communication framework to other complicated task?

Objective: we abstract the similarity between multi-agent communication and other complicated task (Multi-hop QA). We study the application of our communication framework on the Multi-hop Question task.

1.3 Contributions

In this work, we make following contributions:

- We introduce the critical issue of popularity-biased sightseeing, a prevalent challenge arising from neglecting lesser-known POIs. We address it by a novel tourism MARL framework, including considering concentration of tourists at individual POIs and their distribution across the entire network of POIs. (Chapter 3)

- We conduct extensive experiments using a Kyoto real human-mobility dataset. Our model consistently generates tourist routes that achieve a more equitable distribution of visits across POIs, a more balanced distribution of tourists, and higher rewards compared to existing models. This performance is particularly notable in large-scale tourists scenarios. (Chapter 3)

- We conduct a survey to empirically investigate the impact of our method on tourists' experience, which gives reference for practical implementation for sustainable tourism. (Chapter 4)

- We analyze tourists' preference and sustainable tourism development based on game theory, including the discussion of our method in transforming the relationship between them from a non-cooperative relationship to a cooperative one. (Chapter 4)

- We propose an intention-aware communication protocol for MARL. Our proposed intention-aware communication protocol can effectively enhance RPMTD performance in Kyoto real human-mobility dataset for better achievement in the sustainable development goal of tourism. (Chapter 5)

- We reveal the similarity between multi-agent communication and Multi-hop QA, and extend the communication framework to the QA task. We investigate an effective method AMS to answer multi-hop questions, which incorporates single-hop QA models with a document filter. And it is further enhanced with two-step tuning. AMS outperforms other sophisticated GNN-based models in HotpotQA dataset, while it requires less computational resource.(Chapter 6)

1.4 Organization of this Thesis

The rest of this thesis is structured as follows:

- Chapter 3 presents our proposed RPMTD model, which incorporates dual congestion-aware mechanism for tourists route planning by MARL.

- Chapter 4 presents an empirical user study of several congestion-aware route planning methods.

- Chapter 5 presents a collaborative and intention-aware Multi-agent framework for scalable agents.

- Chapter 6 presents a new model, called AMS, for multi-hop QA, which shares the similar framework with our communication method.

- Chapter 7 concludes this thesis by summarizing our key contributions and highlighting various directions for future work.

Chapter 2

Background

2.1 Orienteering Problem

The Orienteering Problem (OP) is a combinatorial optimization challenge that originates from the sport of orienteering. In this sport, participants aim to navigate between checkpoints in unfamiliar terrain, under time constraints, to maximize their score. The orienteering problem abstracts this by presenting a scenario where one must select a subset of available locations to visit, each with an associated score, and determine the most profitable route that respects a given time or distance limit. The main goal is to maximize the total score obtained from visiting various checkpoints. There is a strict limit on the total traveling time or distance, which cannot be exceeded. Traveling Salesman Problem (TSP) is similar with this problem with the additional complexity of score maximization under constraints.

TSP is a highly recognized and widely studied problem in both operations research and theoretical computer science. It involves determining the shortest possible route that visits a given set of cities exactly once before returning to the starting point. The primary objective is to minimize the total distance or travel time. As a fundamental problem in combinatorial optimization, TSP aims to find the optimal loop that ensures each city is visited only once while reducing the overall travel cost, whether measured by distance or time.

2.2 Tourist Route Planning

Tourist Route Planning is a specialized area within route optimization that focuses on developing itineraries for tourists based on specific criteria such as minimizing travel time, maximizing the attractiveness or satisfaction of the visit, and adhering to constraints like opening hours of attractions and the available time of the tourist. The goal is to create an optimal visiting plan that enhances the tourist's experience by effectively scheduling visits to various Points of Interest (POIs) within a given timeframe. The objective

is to optimize the tourist's route so that it maximizes their satisfaction from visiting various attractions while considering personal preferences and practical constraints. These constraints can include the operating hours of POIs, geographic distances between locations, the preferred duration of stay at each site, budget limits, and the total time available for touring. Like TSP, Tourist Route Planning involves finding an efficient path among a set of locations. The basic structure of Tourist Route Planning can be viewed as a variant of TSP where each city in TSP is similar to a tourist attraction. Different from TSP, Tourist Route Planning incorporates more complex constraints and objectives, such as matching the tourist's interests and preferences, which are not considerations in the classic TSP. Like TSP and OP, Tourist Route Planning is generally NP-hard, particularly due to its additional constraints and the need for personalization. Effective solutions often employ sophisticated algorithms, including heuristic methods, genetic algorithms, and machine learning approaches, to approximate the best routes. Recently, there are works utilizing the reinforcement learning to train Pointer Networks for the TSP and other related combinatorial optimization problems, and the promising results are obtained [2–4].

2.3 Reinforcement Learning

Reinforcement Learning (RL) is a type of machine learning that trains agents by feedback from environment. Unlike supervised learning, where training data come with labels indicating the correct action, in reinforcement learning, an agent learns the policy by interacting with an environment and rewards based on its actions. This learning paradigm is heavily inspired by behavioral psychology and how entities learn from the consequences of their actions in real-world scenarios. Following shows some key concepts in RL:

- Agent: decision-maker normally a neural network model in deep learning.

- Environment: the parameters the agent interacts with.

- Actions: what the agent can do. Each action affects the environment.

- State: the current situation returned by the environment.

- Observation: piece of information that an agent perceives about the environment at any given time.

- Reward: a feedback from the environment. Rewards can be positive (reinforcing a behavior) or negative (discouraging a behavior).

- Policy: a strategy that the agent determines its actions based.

- Value Function: it predicts the long-term return expected from a state or state-action pair, used to evaluate the quality of states and guide the policy.

In RL, the agent learns to a policy to make actions based on the given state in a way that maximizes cumulative reward. It begins often without prior knowledge, and through repeated interactions (trial and error). This experience helps the agent to refine its policy, increasingly favoring actions that lead to higher rewards. Genially, there are three type of RL methods: value-based, policy-based and actor-critic methods.

Value-based method

Value-based methods focus on learning the value function with states or state-action pairs. The main concept is to estimate the value of being in a specific state by considering the expected cumulative long-term rewards, which is termed as return in RL. Following shows key concepts of value-based methods: The value function assigns a value to each state (or state-action pair) that represents the return over the future, starting from that state or after taking a particular action in that state. The main types of value functions include:

- **State Value Function** ($V(s)$): Represents the expected return (rewards) starting from state s, and following a particular policy π.
- **Action Value Function** ($Q(s, a)$): Represents the expected return starting from state s, taking an action a.

Policy-based method

Policy-based methods focus on learning a policy function that directly maps states to a probability distribution over actions. Unlike value-based approaches, which first learn a value function and then derive the policy, policy-based methods optimize the policy directly. These methods often exhibit more stable convergence compared to value-based techniques. Due to their reliance on probability distributions, policy-based updates are generally smoother. The policy, expressed as $\pi(a|s)$, is usually represented as a

parameterized function with parameters θ, which defines the likelihood of selecting action a in state s.

The learning process involves adjusting the parameters θ of the policy function $\pi_\theta(a|s)$ to maximize some objective function, typically the expected return from the start distribution, $J(\theta)$. The updates to θ are usually done using gradient ascent on $J(\theta)$:

$$\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta) \tag{2.1}$$

where α is the learning rate.

Followings shows some popular policy-based algorithms

- **REINFORCE**: This algorithm uses the return from a complete episode to update the policy gradient. It estimates the gradients using complete episodes. This indicates it calculates the total reward from each state-action pair at the end of an episode, which it uses to update the policy gradient. The update for REINFORCE is given by:

$$\theta \leftarrow \theta + \alpha \sum_{t=0}^{T} G_t \nabla_\theta \log \pi_\theta(a_t|s_t) \tag{2.2}$$

  where θ represents the policy parameters, α is the learning rate, G_t is the return from time step t, and $\nabla_\theta \log \pi_\theta(a_t|s_t)$ is the gradient of the logarithm of the policy's probability of taking action a_t in state s_t.
- **Proximal Policy Optimization (PPO)**: PPO is a more advanced policy gradient method that addresses some of the practical issues of earlier algorithms like REINFORCE and TRPO (Trust Region Policy Optimization). It's designed to take multiple steps of optimization using the same batch of data. One of the key innovations of PPO is its objective function, which includes a clipping mechanism to prevent excessively large policy updates. This is mathematically represented as following:

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t \left[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t) \right] \tag{2.3}$$

  where $r_t(\theta)$ is the ratio of the new policy probability to the old policy probability of taking action a_t at state s_t, $\hat{A}_t$ is the advantage estimate at time t, and ϵ is a small number.

Actor-critic Method

The Actor-Critic method is a fundamental technique in reinforcement learning that combines both policy-based and value-based methods. It consists of

two main components: the actor determines actions given the current state; it is is usually parameterized by θ. The policy function $\pi_\theta(a|s)$ determines the likelihood of selecting action a when in state s; the critic evaluates the proposed actions by the value function. Its evaluation is typically conducted using a value function, either a state-value function $V(s)$ or an action-value function $Q(s, a)$, which is generally parameterized by w. Together, these two components work to optimize the policy and the value function simultaneously. The actor and the critic are updated as following:

- **Critic Update**:

$$\delta = r + \gamma V(s') - V(s) \tag{2.4}$$

$$w \leftarrow w + \alpha^{\text{critic}} \delta \nabla_w V(s) \tag{2.5}$$

  where r is the reward, γ is the discount factor, s' is the next state, and α^{critic} is the learning rate for the critic.

- **Actor Update**:

$$\theta \leftarrow \theta + \alpha^{\text{actor}} \delta \nabla_\theta \log \pi_\theta(a|s) \tag{2.6}$$

  where α^{actor} is the learning rate.

2.4 Multi-agent Reinforcement Learning

Multi-agent Reinforcement Learning (MARL) expands upon the conventional single-agent framework of reinforcement learning by introducing environments where several agents operate and interact at the same time. In MARL, each agent learns to make decisions based on its observations and the shared environment, often with the goal of maximizing their individual or collective rewards. This field is crucial in understanding and designing intelligent systems where multiple decision-makers interact. Following shows the key concepts in MARL:

- **Multiple Agents**: Unlike single-agent environments, MARL involves multiple agents, each with their capacity to learn and act independently. These agents may have either cooperative, competitive, or mixed objectives.
- **Environment Dynamics**: The inclusion of multiple agents introduces certain complexity to the interaction and environment. The result of one agent's action may be influenced not only by the environmental state but also by the actions of other agents. This inter-agent dependence can create a non-stationary environment from the viewpoint of an individual agent.

- **Communication and Coordination**: Agents may need to communicate or coordinate their actions, especially in cooperative tasks. Effective strategies for communication and coordination are key research areas in MARL.

There are three learning types in MARL:

- **Independent Learning**: Each agent perceives the other agents as part of its environment and learns its own policy independently. While this method is straightforward, it may encounter difficulties in adapting to the non-stationary conditions caused by the dynamic behaviors and learning processes of the other agents.
- **Joint Action/Centralized Learning**: Agents consider the joint actions of all agents in their decision-making process. This approach can become inpractical and infeasible with increasing numbers of agents.
- **Centralized Training with Decentralized Execution**: Agents are trained in a centralized method where global information may be used to optimize performance, but they execute their learned policies independently.

2.5 MARL Cooperation Mode

MARL is a complex field because it includes varying dynamics among agents within the same environment. These dynamics can be broadly categorized into cooperative, competitive, and hybrid (both cooperative and competitive) modes, each presenting unique challenges and requiring different strategies.

Cooperative MARL
In cooperative scenarios, agents work together towards a common goal, often sharing a collective reward based on the overall performance of the group. This requires agents to learn behaviors that not only benefit themselves but also complement and enhance the actions of their peers. Coordination is crucial, and agents often benefit from some level of communication, allowing them to align their strategies and optimize joint actions. Training in these environments might involve centralized learning processes where a global view helps to optimize collective strategies, aiding in more synchronized and efficient teamwork. Such cooperation is evident in tasks like multi-robot systems in logistics, where harmonious collaboration directly translates to increased productivity.

Competitive MARL

Competitive MARL places agents in opposition, with each agent striving to maximize its own individual rewards, often at the expense of others. This creates an adversarial environment where agents must continuously adapt to the strategies of their opponents, leading to a complex dynamic of actions and counteractions. The challenge here is for each agent to outmaneuver or outperform the others, which can resemble playing a high-stakes game where each player seeks to win over the others. Competitive settings are typical in games like poker or strategic games where the success of one agent means a loss for another, pushing each agent to refine their strategies in a continual arms race.

Mixed MARL
Mixed interaction modes present a scenario where agents must both cooperate with some agents and compete against others. This mode is highly reflective of many real-world situations where interactions are not strictly adversarial or collaborative. Agents in such environments must navigate complex social landscapes, forming temporary alliances or competing as dictated by their objectives, which may change based on the context or phase of the task at hand. This requires agents to develop nuanced strategies that can adapt to shifting alliances and oppositions, often necessitating a delicate balance between self-interest and collective goals. Mixed modes are common in scenarios like team sports or corporate strategies where entities might collaborate on some projects while competing in others.

2.6 Challenges in MARL

Challenges in Multi-agent Reinforcement Learning (MARL) stem from the complexities caused by interactions from multiple agents within a shared environment. These challenges differentiate MARL from single-agent reinforcement learning, requiring distinct approaches and solutions. Here are several key challenges encountered in MARL:

Non-stationarity
When each agent learns and updates its policy, the environment change becomes unstable for an individual agent. This is because the environment is updated with joint actions from multi-agent. And the actions and strategies of other agents, which are part of the environment's dynamics, are continually evolving. Thus, what might appear as an optimal action at one point can become suboptimal as other agents adjust their behaviors.

Scalability
The complexity of MARL problems increases exponentially with the number of agents. This issue arises because the joint action space grows exponentially with the number of agents involved. Managing and computing optimal policies in such large-scale systems become computationally infeasible with standard algorithms, necessitating scalable solutions that can effectively approximate solutions without exhaustive computation.

Partial Observability
In many real-world MARL scenarios, agents do not have access to complete information about the environment or the states of other agents. This partial observability adds another layer of complexity, as agents must make decisions based on limited, and often noisy, information. The challenge lies in designing agents that can effectively infer the necessary information from their observations and predict the actions of other agents to make optimal decisions.

Multi-Agent Communication
Effective communication strategies can significantly enhance coordination and learning in MARL. However, designing these communication protocols (what information to share, when, and with whom) introduces additional complexity. Communication can lead to better-informed decisions and more coherent group behavior, but it also incurs costs in terms of bandwidth, privacy, and computational overhead.

Exploiting and Exploring
In MARL, agents are required to balance between exploitation, which involves using existing knowledge, and exploration, where new actions are taken to gather more information about the environment. This balance becomes more complex in a dynamic context where the actions of other agents also influence the environment. The exploratory actions of one agent can have a significant effect on the rewards that other agents receive.

2.7 Multi-hop QA in Nature Language Processing

Multi-hop QA in Natural Language Processing (NLP) is a complex task that involves synthesizing and connecting information from multiple data sources to answer queries that require more than one inferential step. This type of

QA challenges systems to not only locate relevant information but also to link disparate pieces of information logically to construct an accurate answer.

In multi-hop QA, the questions are inherently intricate, often necessitating an understanding of various aspects of the content spread across different sections of a text or across multiple documents. The system must identify these relevant segments, determine their interrelationships, and derive conclusions that are not explicitly stated but implied by the interconnected data.

The process begins with a deep analysis of the question to understand the underlying requirements. The system then scans through large datasets to pull together bits of information that appear relevant to the query. Each piece of information is like a clue in a larger puzzle; the system must not only collect these clues but also figure out how they fit together to form a complete work. With the development of deep learning, advance models like Transformers, Graph Neural Networks, and Memory Networks have promoted the capabilities of multi-hop QA systems. These technologies enable more dynamic handling of data connections and memory, facilitating deeper understanding and more coherent responses to multifaceted questions.

Chapter 3

Dual Congestion-Aware Route Planning for Tourists by Multi-agent Reinforcement Learning

3.1 Introduction

As a worldwide leisure activity, tourism has developed into a significant source of revenue and one of the world's largest industries in recent decades, boosting tourist route planning research in past years. Sophisticated models have been developed to enhance tourist route planning from various aspects, such as popular route optimization and route personalization [5–8]. Existing studies have primarily focused on single tourist route planning, yet little attention has been paid to multiple tourists route planning. This results in homogenization of route planning and points-of-interest (POIs) congestion problem when route planning is performed for a large number of tourists [9].

Recently, although some studies [9–13] have attempted to diversify multiple tourists' routes, they were still developed from the perspective of single

Figure 3.1: An example of real popularity-biased sightseeing in Kyoto, Japan; there is no tourists in the lesser-known attraction, but the popular attraction is crowded with tourists.

tourist planning methods. That is, as only the benefits of tourists are considered in existing methods, visits are mainly planned on popular POIs, and some minor POIs remain unvisited at the end of the trip [9]. We term this problem as popularity-biased route planning, which implies that particular popular POIs attract excessive visits, whereas other POIs do not attract visits. In reality, famous POIs attract a large number of tourists, causing overtourism. However, some less well-known POIs suffer from under-tourism because of the shortage of tourists [14–16]. Figure 3.1 illustrates the problem of Kyoto popularity-biased sightseeing in reality. This problem not only poses a burden on the popular POIs but also discourages the development of sustainable tourism [17]. It has become a global issue, especially in famous tourism cities, such as Kyoto (Japan), Paris (France), and Rome (Italy) [18]. Recently, some countries and organizations (e.g., European Union, France, Japan, New Zealand) have implemented tourists' distribution policies to alleviate this problem [19–22].

To tackle the challenges of popularity-biased route planning and POI congestion, we introduce a new route planning model, the **R**oute **P**lanning **M**odel with consideration of **T**ourist **D**istribution (RPMTD). Our approach is two-fold: Initially, we develop a novel tourism multi-agent reinforcement learning (RL) environment capable of interacting with multiple tourists. Building upon this environment, we introduce the RPMTD model, distinguished by its dual-congestion awareness. This model strategically assesses tourists' distribution to optimize route planning. In detail, the RPMTD model incorporates two types of congestion rewards: local and global. The local congestion reward measures the level of crowding at individual POIs, helping to avoid over-concentration of tourists at specific locations. In contrast, the global congestion reward assesses the overall distribution of tourists across all POIs, aiming for a more balanced and equitable allocation of tourists. This dual-congestion approach ensures that tourism's individual and collective impacts are considered, promoting sustainable and enjoyable tourism experience. Figure 3.2 illustrates the comparison of our method with existing single and multiple tourists route planning methods.

In our reinforcement learning (RL) environment, we leverage authentic human mobility data from Kyoto, Japan, renowned for its popularity and representativeness in global tourism. To capture the intricate dynamics of human movement, we consider three distinct types of mobility data: (i) the mobility patterns of residents, (ii) the movements of tourists, and (iii) pseudo-mobility data, which simulates a broader spectrum of real-world scenarios. We apply our model, RPMTD, to 72 POIs in Kyoto. The model's performance is meticulously evaluated based on several critical metrics: the equitable distribution of visits across POIs, the overall evenness in the

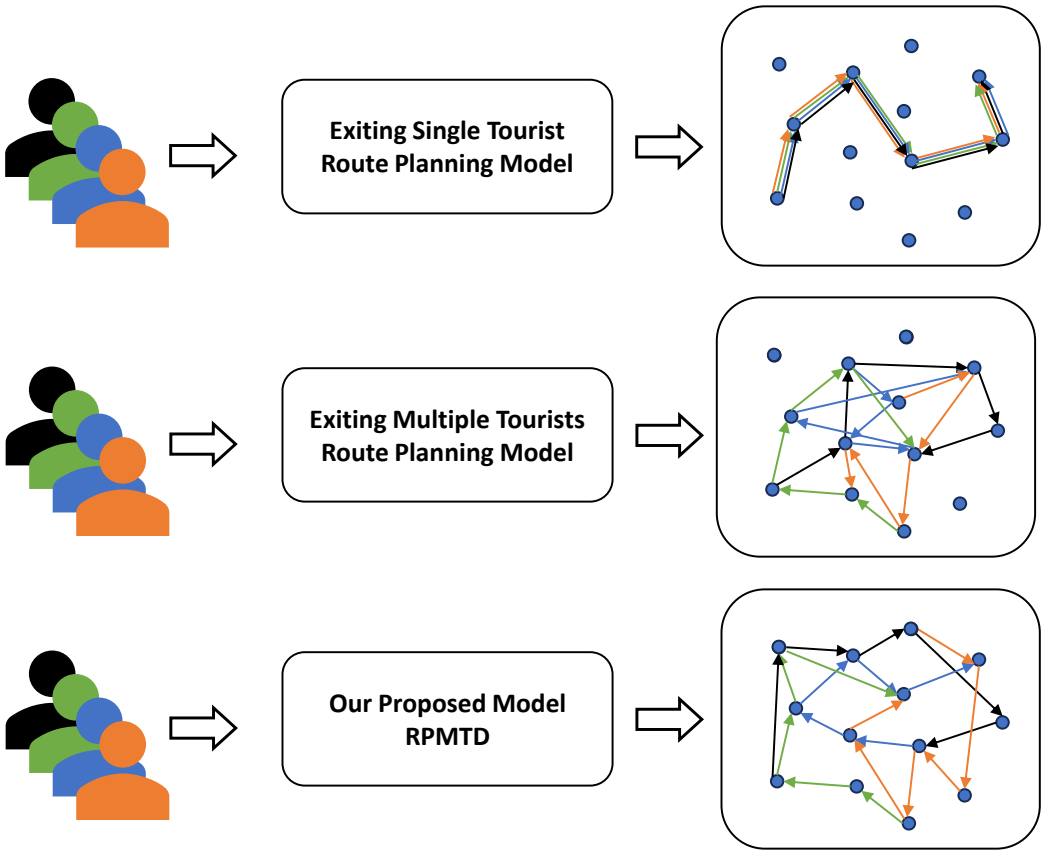


Figure 3.2: Comparison of our method with existing single and multiple tourists route planning methods in terms of multiple tourists planning. Single tourist planning methods generate homogenized route for all tourists. Existing multiple tourists planning methods generate routes biased on popular POIs. Our method generates routes with a balanced distribution of visits.

distribution of tourists, and the total reward achieved. The results from these evaluations demonstrate that our model not only surpasses competing models in performance but also exhibits considerable robustness across various tourists scenarios.

The main contributions of our study are summarized as follows:

1. We introduce the popularity-biased sightseeing problem and propose a novel tourism multi-agent RL framework for it, including an environment for interaction with tourists and a model with consideration of both local and global congestion.
2. Unlike existing methods, we implement three types of mobility data to

simulate the RL environment: 1) mobility of residents, 2) mobility of tourists, and 3) pseudo-mobility generated automatically. This fashion makes our RL environment more close to the real world.

3. We conduct extensive experiments to validate our method with real datasets including 72 POIs in Kyoto. The experimental results demonstrate that our model generates routes with fairer POIs' visits, more even tourists' distribution and higher reward than other existing models, specifically in large-scale tourists planning.

3.2 Related Work

Single tourist planning: Numerous works have developed single tourist route planning schemes considering different aspects. Some studies aim to maximize trip rewards. Hsieh et al. [23] consider the tourist route planning problem as an orienteering problem and generate the route with specified starting and ending POI. Gunawan et al. [24] utilize integer linear programming to model the problem. Gama et al. [4] construct a sophisticated model based on the pointer network and apply RL for the model training. Some works focus on personalized route planning. Duan et al. [5] personalize the route recommendation by analyzing the user's history of interest from check-in behavior. Gionis et al. [25] customize the recommendation by consideration of the order of POI's category. Taylor et al. [26] consider mandatory POIs visiting in route recommendation. Some works enhance performance based on location-based social network services and social media. Xu et al. [27] and Padia et al. [7] discriminate the POIs by sentiment analysis based on users' reviews. Seo et al. [28] take into account the anchoring effect in route planning by weighting the tourist's initial check-in data. Qian et al. [29] and Yu et al. [30] consider contextual factors in POI recommendation. Zhang. et al [31] propose incorporating a visual model to improve POI recommendation. Gao et al. [32] use the social information from an expanded set of friends to deal with the data sparsity problem. Cheng et al. [33] measure the geographical importance and fuse it with social information for route recommendation. Liu et al. [34] and Chen et al. [6] recommend topics and POIs to tourists by utilizing POIs' textual information. Besides, there are also some works [35–37] studying other aspects, such as privacy protection, user transition pattern analysis, and mixed styles of sightseeing.

Multiple tourists planning: Conversely, few studies focus on multiple tourists. Sylejmani et al. [10] plan a group of tourists considering individual

preferences and mutual social relationships. Lim et al. [12] and Elmi et al. [13] cluster tourists into groups and apply a single tourist method to recommend routes for each group. Sarkar et al. [11] apply Subgame-Perfect Nash equilibrium of game theory to recommend various routes for a group of tourists. Kong et al. [9] apply the single-agent RL framework to diversify multiple tourist routes.

Other congestion-related problems: Lee et al. [38] and Normoy et al. [39] study the navigation of agent's behavior in crowds to avoid obstacles. Li et al. [40] and Sridhar et al. [41] investigate traffic-flow management to prevent road congestion. Unlike them, our work focuses on the spots' congestion problem and tourist' distribution.

Difference with previous works: Our work differs from previous works in the following aspects: (i) previous works generate the route only from the benefits of tourists, e.g., optimized route recommendation and route personalization, whereas we consider the POIs' benefits; (ii) previous works apply either heuristic methods or single-agent RL for multiple tourists route planning, whereas, we propose multi-agent RL framework for multiple tourists planning; (iii) previous works are based on simple and small-scale tourist scenarios, while our work is validated on real and large-scale mobility, which is more realistically applicable;

3.3 Preliminaries and RL Environment

In this section, we first summarize the notations used in our work in Table 3.1. Subsequently, we define our problem and the proposed RL Environment.

3.3.1 Basic concept

We represent the set of POIs in the target city as:

$$\mathcal{P} = \{p_1, ..., p_n\} \tag{3.1}$$

where n is the number of POIs. Each POI p_i is associated with capacity c_i, location loc_i, time cost $cost_{p_i}$, score scr_i, visit number vn_i^t, and number of tourists num_i^t. We represent the set of tourists as:

$$\mathcal{T} = \{t_1, ..., t_m\} \tag{3.2}$$

where m is the number of tourists. Each tourist t_j is associated with activate time I_j^{act}, remaining time b_j^{re}, time budget b_j, starting POI p_j^s, and ending

Table 3.1: Description of notations.

	Notation	Description	Notation	Description
	$\mathcal{P}$	The set of all POIs	$\mathcal{T}$	The set of all tourists
	p_i	A specific POI, where $p_i \in \mathcal{P}$	t_j	A specific tourists, where $t_j \in \mathcal{T}$
	c_i	Capacity (max number of tourists) of p_i	sp_j	Moving Speed of t_j
POI	loc_i	Coordinate of p_i, i.e., (latitude, longitude)	b_j	Time budget for t_j
&	$cost_{p_i}$	The amout of time needed to visit p_i	p_j^s	Starting POI of t_j
Tourist	scr_i	Score of p_i, denoting the attractiveness of p_i	p_j^e	Ending POI of t_j
	vn_i^t	Number of p_i has been visited at time t	Q_j	Query of t_j, including b_j, p_j^s and p_j^e
	num_i^t	Number of people (population) in p_i at time t	S_j	Route of t_j, a sequence of POIs
	r_i	Reward of visiting p_i		
	$agent_j$	A specific agent in RL, representing a tourists t_j		
	M	Mobility Matrix, representing POIs' population change over one day		
	I^{cur}	Indicator of current time slot in M at each interaction		
	I_j^{act}	Indicator of activate time for t_j/$Agent_j$. (t_j is unstarted or in sightseeing before activate time.)		
RL	b_j^{re}	Remaining time budget for t_j		
	tm_{ij}	The amount of time needed to move from t_j's current POI to p_i		
	O_j^i	Observation of p_i for t_j (including c_i, $cost_{p_i}$, scr_i, vn_i^t, num_i^t, b_j^{re}, b_j, tm_{ij})		
	O_j	Global observation for t_j, stacking O_j^i of all POIs together		
	A_j	Action of $Agent_j$ given observation O_j, next visiting POI		

POI p_j^e. For t_j, the query Q_j is given as a tuple:

$$Q_j = (b_j, p_j^s, p_j^e) \tag{3.3}$$

where b_j, p_j^s and p_j^e denotes the time budget, starting and ending POI, respectively. Based on Q_j, the model returns a route S_j for t_j, which includes k POIs. In S_j, the tourist must begin at p_j^s and terminate at p_j^e, while keeping the traveling time within b_j.

The reward of a tourist visiting a POI dynamically depends on the visited POI's crowdness and the global tourists' distribution. The objective is to maximize the sum of all tourists' rewards accumulated from tourists sequentially visiting POIs in their routes.

3.3.2 Problem Formulation

3.3.2.1 General Tourists Route Planning

We first define the General Route Planning *without* consideration of dual-congestion awareness.

Definition Definition 3.1 *General Route Planning for Tourists*

Given $\mathcal{T}$, the query $Q_j = (b_j, p_j^s, p_j^e)$ is specified for t_j. A route S_j should be generated for t_j from $\mathcal{P}$. The objective function of maximizing the sum of all tourists' rewards is represented as follows:

$$Max \sum_{j=1}^{m} \sum_{i=1}^{n} r_i \cdot \mathbf{X}_j(p_i),\ where\ \mathbf{X}_j(p_i) = \begin{cases} 1,\ if\ p_i \in S_j \\ 0,\ otherwise \end{cases} \tag{3.4}$$

r_i is the reward of visiting p_i; most existing works use scr_i and users' preference to estimate it. Function $\mathbf{X}_j(p_i)$ calculates if p_i in S_j. Eq. (1) is solved by following constraints:

Constraint 1:

$$S_j = (p_1^j, ..., p_k^j)$$
$$where$$
$$\begin{cases} p_1^j, ..., p_k^j \in \mathcal{P} \\ p_1^j = p_j^s\ and\ p_k^j = p_j^e \\ |\left\{p_1^j, ..., p_k^j\right\}| = k \end{cases} \tag{3.5}$$

Constraint 2:

$$\sum_{i=1}^{k-1} D_i^j \leq b_j,\ where\ D_i^j = cost_{p_i^j} + travel\_time(p_i^j, p_{i+1}^j) \tag{3.6}$$

where $travel\_time(p_i^j, p_{i+1}^j)$ calculates the time needed to travel from POI p_i^j to the next POI p_{i+1}^j, and $p_i^j, p_{i+1}^j \in S_j$. This is the approximated time based on the POIs' locations and sp_j.

Constraint 3:

$$\begin{aligned} 2 \leq \mathbf{H}(p_x^j), \mathbf{H}(p_y^j) \leq k, \forall x, y = 2, ..., k \\ \mathbf{H}(p_x^j) - \mathbf{H}(p_y^j) + 1 \leq (k-1)(1 - \mathbf{F}(p_x^j, p_y^j)) \end{aligned} \tag{3.7}$$

where $\mathbf{H}(p_x^j)$ calculates the position of p_x^j in S_j. $\mathbf{F}(p_x^j, p_y^j)$ calculates the consecutivity of p_x^j and p_y^j in S_j. $\mathbf{F}(p_x^j, p_y^j) = 1$ if p_y^j is visited after p_x^j, otherwise $\mathbf{F}(p_x^j, p_y^j) = 0$.

Constraint 1 ensures that (i) all visited POIs are in the set $\mathcal{P}$; (ii) the route must begin at the starting POI and terminate at the ending POI; (iii) no POI can be visited more than once. Constraint 2 ensures that for t_j, the travel duration should not exceed b_j. Constraint 3 ensures that sub-tours are eliminated.

3.3.2.2 Dual-congestion aware Route Planning

As one class of General Routes Planning, we define the dual-congestion reward of visiting a POI as follows:

Definition Definition 3.2 *Local Congestion Reward*

The local congestion reward of visiting p_i evaluates the congestion of p_i, which depends on c_i, num_i^t, vn_i^t and scr_i.

$$r_i^{local} = f_{local}(c_i, num_i^t, scr_i, vn_i^t) \tag{3.8}$$

Definition Definition 3.3 *Global Congestion Reward*

The global congestion reward of visiting p_i evaluates the global tourists' distribution, which depends on all POIs' capacity and number of tourists.

$$r_i^{global} = f_{global}(c_1, ..., c_n, num_1^t, ..., num_n^t) \tag{3.9}$$

Definition Definition 3.4 *Dual-congestion Reward*

The dual-congestion reward of visiting the POI p_i is the sum of the global and local congestion rewards.

$$r_i = sum(r_i^{global}, r_i^{local}) \tag{3.10}$$

Dual congestion evaluates global and local congestion simultaneously. The objective function also maximizes the sum of all tourists' rewards while subject to the same constraints.

3.3.3 Multi-agent Reinforcement Learning Environment

Our proposed environment consists of the mobility matrix, tourists, and POIs. The $U \times V$ mobility matrix M presents the POIs' population change over one day, where U is the number of POIs; V is the number of time slots representing one day. Cell_{uv} of M indicates the number of people in POI u at the specific time slot v. There is a current time indicator I^{cur} for M, indicating the current interaction time slot and moving to the next time slot after one interaction. For t_j, there is an indicator of activate time I_j^{act}. Before the activation time, t_j does not interact with the environment. Specifically, t_j is either in sightseeing or unstarted before the activation time. For p_i, num_i^t is set according to the current time slot of the mobility matrix in each interaction.

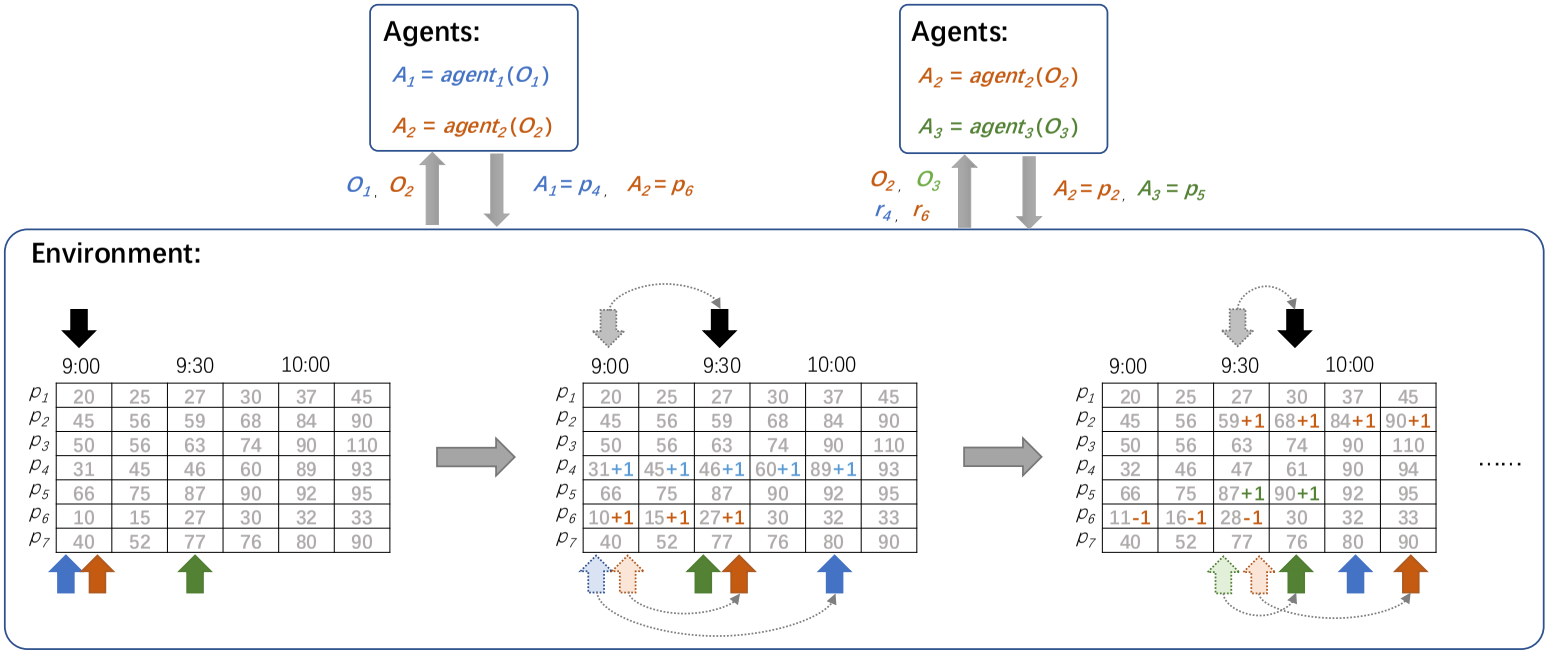


Figure 3.3: Example of interaction between environment and three agents (tourists). The black arrow on the mobility matrix is the current time indicator I^{cur}; colorful arrows under the mobility matrix are agents' activate time indicator I^{act}.

I^{cur} is initialized by the tourists' starting time. The POIs' popularity is set according to the current time slot from the mobility matrix M. Environment generates the observation O_j for t_j, which is stacking of each POI p_i's observation O_j^i. In O_j^i, there are 8 features: c_i, $cost_{p_i}$, scr_i, vn_i^t, num_i^t, b_j^{re}, b_j, and tm_{ij}. Given O_j, agent$_j$ outputs the action A_j, the next visiting POI. Based on A_j, M and I_j^{act} is updated, and rewards are given.

Figure 3.3 shows an example of interaction. The black arrow on the mobility matrix M is the current time indicator I^{cur}; blue, red and green arrows under M are the activate time indicators of $agent_1$, $agen_2$, $agent_3$, respectively. $agent_1$, $agen_2$ start at 9:00, and $agent_3$ starts at 9:30. At the beginning, I^{cur} is set at 9:00, and each POI's population is set according to the 9:00 slot of M. The environment generates the observations O_1, O_2 at 9:00 for $agent_1$ and $agent_2$, respectively, based on which the two agents output actions (next POIs). $agent_1$ will stay in p_4 from 9:00 to 10:00 and $agent_2$ will stay in p_6 from 9:00 to 9:30. M and two agents' activation time is updated accordingly. The rewards of visiting p_4 and p_6 are given. Then I^{cur} skips 9:15 and moves to 9:30, as no agents are activated at 9:15. Observations O_2, O_3 at 9:30 are generated for the $agent_2$ and $agent_3$, respectively. $agent_2$ and $agent_3$ output actions, based on which M and two agents' activate time is updated. Then I^{cur} moves to 9:45 where $agent_3$ is activated. Interaction repeats until all tourists finish their trips.

3.4 Dual-congestion aware Routes Planning Model

In this section, we first illustrate the RL implementation. Subsequently, our Dual-congestion mechanism is described.

3.4.1 Multi-agent Reinforcement Learning Implementation

Our work is based on the fully decentralized approach to adapt our model to scalable tourists scenarios. The actor-critic method is adopted, and independent PPO (IPPO) [42] is utilized for training. Since all the agents are homogeneous, meaning they have the same state space, action space, and optimization objective, parameter sharing is conducted.

In our scenario, the model is required to handle interactions among a variable number of tourists. IPPO is employed due to its intrinsic capability to address such dynamic challenge. This dynamic challenge is attributed

from two factors: (i) the model should generate routes for various number of tourists, instead of only for a constant number of tourists; (ii) in each interaction, not all agents participate in the interaction. Because some agents

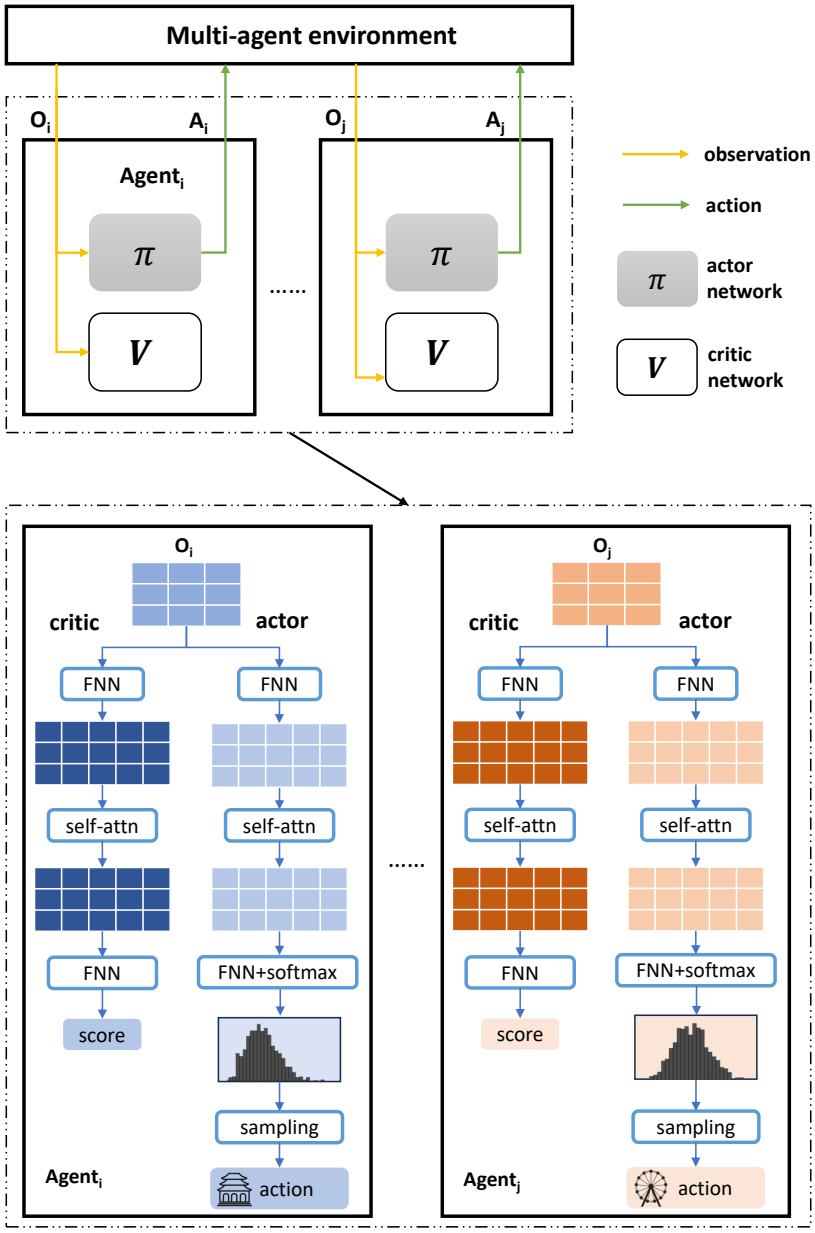


Figure 3.4: Schematic diagram of multi-agent IPPO algorithm and model structure of actor and critic networks.

are inactive and in sightseeing.

In the actor-critic method, the actor outputs the action and learns a policy with the critic's guide. The observation $\mathbf{O}$ (we omit the subscript for a general denotation) is given for an agent from the environment in each interaction. Specifically, $\mathbf{O} = \{O^1, ..., O^n\} \in \mathbb{R}^{n \times d}$, where n and d denotes the number of POIs and the POI's feature dimension. The actor first maps $\mathbf{O}$ into the hidden space with a feed-forward network (FNN). The hidden representation $\mathbf{O}'$ is obtained as following:

$$\mathbf{O}' = \mathrm{FFN}_1(\mathbf{O}) \in \mathbb{R}^{n \times h} \tag{3.11}$$

where h denotes the hidden dimension.

Then, the self-attention is applied to update hidden representation $\mathbf{O}'$, which is motivated by its effectiveness in combinatorial optimization studies [43]. Specifically, it learns the relations between each pair of POIs and updates each POI's representation from all POIs. We use the transformer encoder [44] as a self-attention module.

$$\mathbf{H} = \mathrm{TransformerEncoder}(\mathbf{O}') \in \mathbb{R}^{n \times h} \tag{3.12}$$

The updated representation $\mathbf{H}$ is further fed into another FNN to calculate the logits of each POI, and then is normalized by softmax. The final action is obtained by random sampling.

$$\mathbf{H}' = \mathrm{softmax}(\mathrm{FFN}_2(\mathbf{H})) \in \mathbb{R}^n \tag{3.13}$$

The critic's structure is similar to the actor except that it finally calculates a scalar value. Figure 3.4 shows the structures of the actor and critic networks.

The actor network and the critic network are represented as $\pi_\theta(a_i|o_i)$ and $V_\omega(o_i)$, where θ and ω are learnable parameters in the networks. Given the trajectory from interactions $\{(o_1, a_1, r_1, o_1'), \ldots, (o_M, a_M, r_M, o_M')\}$, where o_i, a_i, r_i, o_i' are the current observation, action, reward, and next observation at i-th step, the critic network is learned by minimizing following loss:

$$L_{critic} = \frac{1}{M} \sum_{i=1}^{M} (r_i + \gamma V_\omega(o_i') - V_\omega(o_i))^2 \tag{3.14}$$

where γ is the discount factor. The actor network is learned by maximizing the following objective:

$$L_{actor}^{clip} = \mathbb{E}[\min(l_i A_i, \mathrm{clip}(l_i, 1-\epsilon, 1+\epsilon) A_i)] \tag{3.15}$$

where $l_i = \frac{\pi_\theta(a_i|o_i)}{\pi_{\theta_{old}}(a_i|o_i)}$ is the likelihood ratio; $\mathrm{clip}(x, k, h)$ clips x in $[k, h]$; A_i is the advantage estimate; ϵ is the hyperparameter that controls the clipping ratio.

3.4.2 Dual-congestion mechanism

We propose a dual-congestion mechanism for unbiased route planning to construct a reward function. Specifically, local congestion considers visited POI crowdedness, and global congestion considers the evenness of tourists' overall distribution. The total reward is the weighted sum of both rewards:

$$Reward_{total} = \omega_1 Reward_{global} + \omega_2 Reward_{local} \tag{3.16}$$

where ω_1 and ω_2 are weights for local and global congestion rewards.

Reward based on Local Congestion: $Reward_{local}$ evaluates the visited POI's crowdness. The intuition is that fewer tourists pose less burden on the POI and make tourists more satisfied, which results in higher rewards. We use a linear function to define the local congestion reward without considering negative reward as follows:

$$Reward_{local} = \max(score_{sc} \cdot (1 - \frac{num^t}{c}), 0) \tag{3.17}$$

where $score_{sc}$ is the scaled POI score, which depends on each POI's visit number and overall trip process, aiming to improve the minor POIs' visiting. The intuition is that POIs with fewer visits will have higher scores.

$$\begin{aligned} score_{sc} &= \max(\mathrm{I}(x) \cdot \mathrm{Sc}(num^t), 1) \cdot score \\ \mathrm{I}(x) &= \frac{1}{1 + e^{-20(x-\lambda)}} \\ \mathrm{Sc}(num^t) &= 1 + \frac{20}{(num^t + 1)^4} \end{aligned} \tag{3.18}$$

where $\mathrm{I}(x)$ is the trigger function. x is the percentage of total tour process. $\mathrm{Sc}(num^t)$ is the score scaling function. $\mathrm{I}(x)$ triggers score scaling function after a certain time. λ is a hyperparameter set to 0.6 as default, indicating that the score scaling function is triggered from 60% of the total touring process. The score scaling function scales the POI's score according to the cumulative visit number. For instance, the scores of POIs with zero, one, two, and three visits are multiplied by 21, 2.25, 1.25, and 1.08, respectively.

Reward based on Global Congestion: To more fairly distribute the tourists over all the POIs, we propose the global congestion reward $Reward_{global}$, which is inversely proportional to the variance of attendance percentages across all POIs:

$$Reward_{global} = \frac{1}{Var(ap_1, ..., ap_n)} \tag{3.19}$$

where ap_i is the attendance percentage of p_i, $ap_i = \frac{num_i^t}{c_i}$.

3.5 Experiment and Result

3.5.1 Target City Background

Our experiments are conducted on a real-human mobility dataset collected from Kyoto, Japan, one of the world's most famous and typical tourism destinations.

Kyoto accounts for 3.5% of the world's tourists each year, which is similar to other popular tourism cities, such as New York (4.4%), London (1.9%), and Paris (2.9%). More than 10% local people are employed in tourism, and

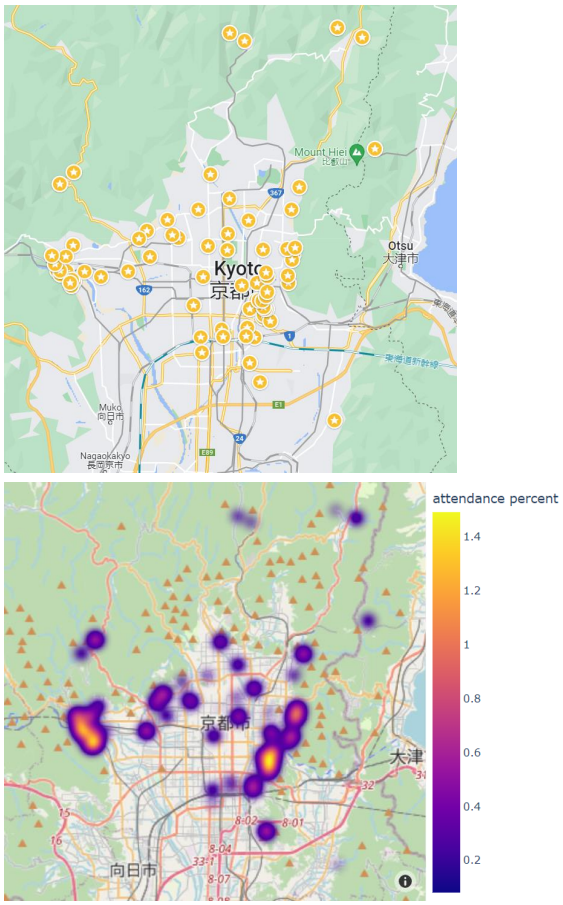


Figure 3.5: Locations distribution of 72 POIs of Kyoto in our experiment (upper) and popularity-biased tourists' distribution based on real mobility data at rush hour (lower). Each circular yellow mark represents one POI's location.

the tourism value was about 12 billion U.S. dollars in 2018, making up about 17% gross domestic product of this city. [45]. Thus, Kyoto is a typical tourism city, and proper distribution of tourists is crucial for both tourism and the development of the city. There are numerous temples, shrines, and historical houses in Kyoto. Our experiments are conducted on 72 POIs, including 14 famous World Cultural Heritage Sites. Figure 3.5 shows the locations of 72 POIs and popularity-biased tourists' distribution at rush hour.

3.5.2 Experiment Setting

Mobility Data Preparation To simulate the real world and validate our model robustness, we conduct experiments on two mobility datasets: F-Data and FYP-Data. The data is collected from two resources: (i) Flickr, a widely used image hosting and sharing platform worldwide; (ii) Yahoo! JAPAN, a popular portal service in Japan, provides various daily services, including search engine, news, entertainment, weather forecast, etc. Details are given as follows:

- **F-Data**: We count the number of Flickr users who post photos on Flickr in each POI and different time slots. Laplace smoothing is conducted for the POIs whose visit number is zero, adding a minimum non-zero number for all POIs' visit numbers. It represents the Kyoto tourists' mobility; the maximum is about 4000.
- **Y-Data**: This dataset is provided by LY Corporation (previous Yahoo Japan Corporation). It includes 6667031 trajectories of 757878 mobile users using Yahoo! JAPAN services in 16 days. The data is pre-anonymized and random noise is added. We count the number of mobile users in each POI every half hour to obtain the mobility data and average them for further processing. It represents the Kyoto residents' mobility, and the maximum mobility is about 25000.
- **P-Data**: Based on Y-data, we use a density model [46] to generate trajectory data of 500 pseudo-users to simulate the mobility of random tourists in the same day.
- **FYP-Data**: This data is the combination of **Y-Data, F-Data** and **P-Data**, which is closer to reality.

We obtain the mobility data for each time slot and generate the mobility matrix M for experiments.

Tourists Setting We generate 100, 200 tourists as small-scale tourists and 500, 1000 as large-scale tourists. In F-Data experiment, we only consider small-scale tourists. In FYP-Data, we consider both small-scale and

large-scale tourists. Starting POIs are selected from Kyoto's three popular sightseeing POIs and destinations are selected as the city center and Kyoto station. Each tourist's time budget is randomly set between 6 and 8 hours.

Implementation details The FFNs in both actor and critic networks are set as 2-layer with 0.2 dropout. For the transformer self-attention module, we stack 2 layers and 8 heads. In the main experiments, λ, ω_1, and ω_2 are set as 0.6, 0.1, and 1, respectively. In RL training, discount factor γ, clip control factor ϵ, and learning rate for actor and critic are set as 0.99, 0.2, 3e-4, and 1e-3, respectively. We implement experiments on one A100 GPU server.

3.5.3 Baseline Setting

We select two models based on RL as our baseline, and implement them on our proposed environment for comparison.

- MARLRR [9] diversifies tourists' routes by dynamic reward function from a tourism economic model and applies the single-agent RL method DQN for multiple tourists route planning.
- Pointer-NN [4] constructs model with pointer networks and dynamic graph self-attention. The model is trained by the REINFORCE algorithm. It is one of the state-of-the-art models for single tourist route planning.

We also select other multi-agent reinforcement learning (MARL) algorithms as our baseline to compare with IPPO. Specifically, we consider independent learning and centralized policy gradient methods. Value decomposition methods are not considered as they specialize in decomposing a single joint reward for multiple agents, whereas in my scenario each agent gets an individual reward from the environment.

- IQL [47]is a commonly-used independent learning algorithm and developed from Q-Learning. In IQL, each agent learns independently and perceives the other agents as part of the environment. Therefore, it can be directly applied for our scenario.
- MADDPG [48] is a centralized policy gradient algorithm based on centralized training with decentralized execution (CTDE) framework. In MADDPG, each agent is trained with a deep deterministic policy gradient approach while considering the presence of other agents.
- MAPPO [49] is also a CTDE algorithm. Similar with IPPO, it is an extension of the Proximal Policy Optimization algorithm to multi-agent environments. The difference is that the learning of agents in MAPPO is based on global information.

However, MADDPG and MAPPO cannot be directly applied in our dynamic scenario due to their intrinsic limitations of critic network in processing inputs of variable dimensions. To solve the variable-length input problem, we utilize self-attentive embedding [50] to convert the variable-length input into a fixed-length embedding.

Specifically, we firstly simplify the observations of agents. Each agent's observation contains POI-specific information and tourist-specific information. As the POI-specific information is consistent across all agents, we only

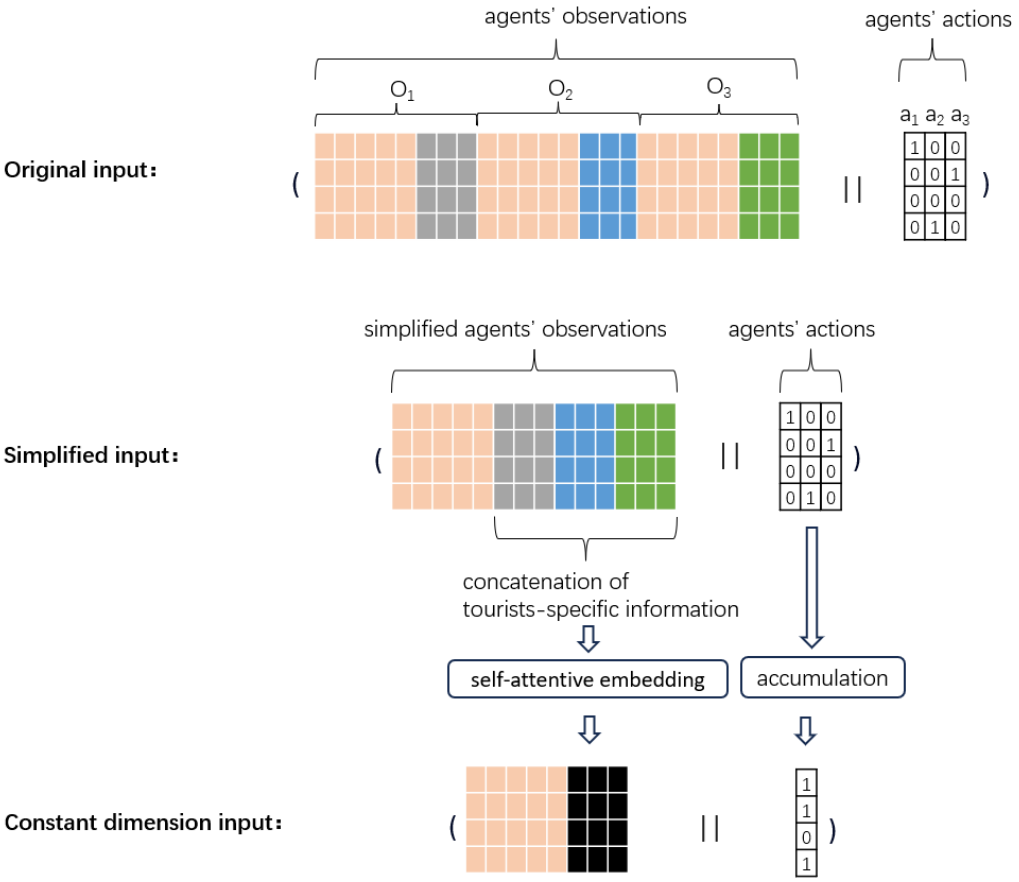


Figure 3.6: An example of converting three agents' input into a constant dimension input. (||) denotes the concatenation function. The POI-specific information is in orange; the tourist-specific information is in gray, blue and green; the multiple tourists-specific information processed by self-attentive embedding is in black.

concatenate tourist-specific information and convert it into a fixed-length embedding with self-attentive embedding. Furthermore, one-hot vector is employed to represent agent's action. We accumulate each agent's action one-hot vector along dimension of each POI to represent the joint actions. Through this method, a constant dimension input is obtained to apply MADDPG and MAPPO for our scenario. Figure 3.6 shows an example of converting three agents' input into a constant dimension input.

3.5.4 Evaluation Metrics

To evaluate the performance, the following evaluation metrics are utilized:

(1) Maximum and average variance of all POIs' attendance percentage (Var_{max} and $\text{Var}_{average}$) throughout the whole trip are utilized to measure the bias/imbalance of tourists' distribution.
(2) Gini coefficient [51] of POIs (Gini_{POI}) and tourists ($\text{Gini}_{tourist}$) are utilized to measure the fairness among POIs' visiting and tourists' reward, respectively. A higher Gini coefficient indicates greater unfairness.
(3) Total static reward (Reward_{static}) and total congestion-aware/dynamic reward ($\text{Reward}_{dynamic}$) are utilized to measure the total reward of all tourists.
(4) Average of edit distance (ED_r) is utilized to measure the diversity of planned routes. Higher edit distance indicates greater diversity.

3.5.5 Experimental Results

We first present the comparison between IPPO and other MARL algorithms on F-Data and FYP-Data.

Table 3.2 shows the comparison result of MARL algorithms. As IQL, MADDPG and MAPPO all fail to converge in the scenarios of FYP-data 500 and 1000 tourists, the comparison is not shown in the table. IPPO outperforms other MARL algorithms in both F-data and FYP-data. IQL performs the worst and fails in convergence in scenario of more than 100 tourists. MADDPG and MAPPO fail in convergence in scenarios of 500 and 1000 tourists in FYP-data. The reason could be that in large-scale agents' interaction, current method cannot effectively represent the global observations of all agents; thus, the central critic network cannot accurately provide guidance to each agent's actor network.

We present the comparison between our model RPMTD and other two route planning models on F-Data and FYP-Data.

Table 3.2: Result of MARL algorithms based on F-data and FYP-data.

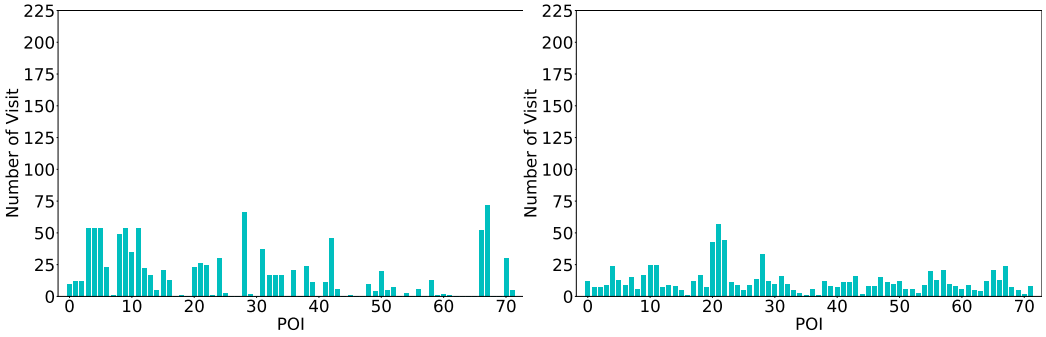
		Gini_{POI}	$\text{Gini}_{tourist}$	Var_{ave}	Var_{max}	ED_r
F-data 100 tourists	IPPO	**0.309**	**0.119**	**0.075**	**0.094**	1265.281
	MADDPG	0.473	0.136	0.079	0.113	**1274.67**
	MAPPO	0.488	0.124	0.087	0.096	1264.725
	IQL	0.503	0.137	0.082	0.128	1270.169
F-data 200 tourists	IPPO	**0.321**	**0.12**	**0.063**	**0.093**	**2214.93**
	MADDPG	0.628	0.131	0.093	0.152	2039.486
	MAPPO	0.735	0.146	0.119	0.173	2184.971
	IQL	-	-	-	-	-
FYP-data 100 tourists	IPPO	**0.612**	**0.168**	**0.036**	**0.067**	**1315.29**
	MADDPG	0.685	0.193	0.046	0.105	1154.658
	MAPPO	0.719	0.204	0.054	0.129	1243.854
	IQL	0.758	0.181	0.049	0.094	1268.541
FYP-data 200 tourists	IPPO	**0.625**	**0.189**	**0.032**	**0.071**	**2265.35**
	MADDPG	0.752	0.191	0.062	0.138	2146.347
	MAPPO	0.827	0.215	0.071	0.113	2048.807
	IQL	-	-	-	-	-

Table 3.3: Result of route planning methods based on F-Data.

		Gini_{POI}	$\text{Gini}_{tourist}$	Var_{ave}	Var_{max}	ED_r
F-Data 100 tourists	RPMTD	**0.309**	0.119	**0.075**	**0.094**	1265.281
	MARLRR	0.555	0.121	0.079	0.099	**1320.825**
	Pointer-NN	0.675	**0.113**	0.157	0.485	91.512
F-Data 200 tourists	RPMTD	**0.321**	0.120	**0.063**	**0.093**	**2214.927**
	MARLRR	0.617	0.126	0.081	0.110	2042.472
	Pointer-NN	0.717	**0.118**	0.184	0.427	173.947

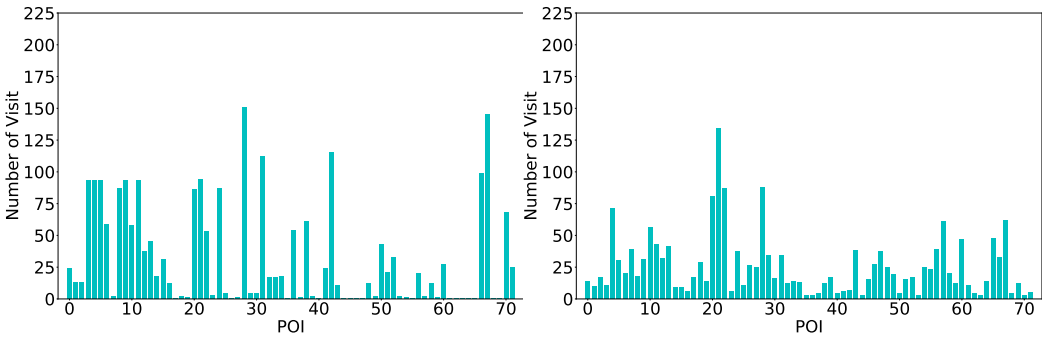
Table 3.3 shows the result on F-Data. Pointer-NN shows the worst result except $\text{Gini}_{tourist}$, as it generates homogenized routes, which causes POI congestion problems in multiple tourists planning. Generally, RPMTD shows better result in Gini_{POI}, Var_{ave}, and Var_{max}, while MARLRR shows comparable performance with RPMTD in ED_r. Additionally, Gini_{POI} of all three models with 200 tourists is higher than that for 100 tourists, indicating that the unfairness of POI grows with increasing tourists.

Figure 3.7 shows each POI's visit number at the end of the trip. MARLRR shows a skewed distribution, where the popular POIs attract most



((a)) Visit distribution of MARLRR with 100 tourists on F-Data

((b)) Visit distribution of RPMTD with 100 tourists on F-Data



((c)) Visit distribution of MARLRR with 200 tourists on F-Data

((d)) Visit distribution of RPMTD with 200 tourists on F-Data

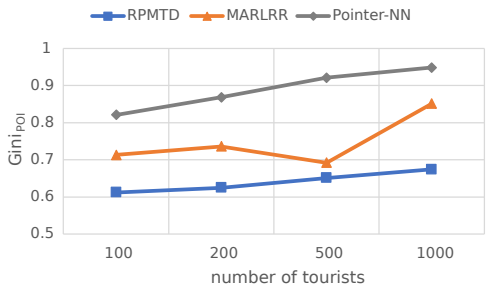
Figure 3.7: Comparison of visits distribution at the end of the trip between RPMTD and MARLRR based on F-Data.

tourists, and some minor POIs have zero visits. In contrast, our method distributes tourists more evenly, and all the POIs have been visited. Our model retains the characteristics of tourism that popular POIs still have more tourists than others. Comparison between 100 and 200 tourists shows that popular POIs are more likely to attract visits when the number of tourists increases, which is consistent with the Table 3.3 result that the unfairness of POI grows with increasing tourists.

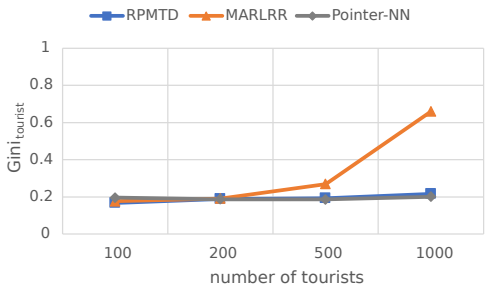
Similarly, we conduct experiments on FYP-Data, and Table 3.4 shows the results. Compared with F-Data, the FYP-Data result shows a generally larger Gini coefficient, which is due to the data difference. RPMTD shows the smallest Gini_{POI}, while MARLRR and Pointer-NN show much larger Gini_{POI}, indicating stronger unfairness of POI. Pointer-NN shows the smallest $\text{Gini}_{tourist}$ and ED_r, which means that tourists' rewards are similar and routes are homogeneous. Meanwhile, this results in the low $\text{Reward}_{dynamic}$ in Table 3.6. RPMTD and MARLRR show comparable $\text{Gini}_{tourist}$ in small-scale

Table 3.4: Result of route planning methods based on FYP-Data.

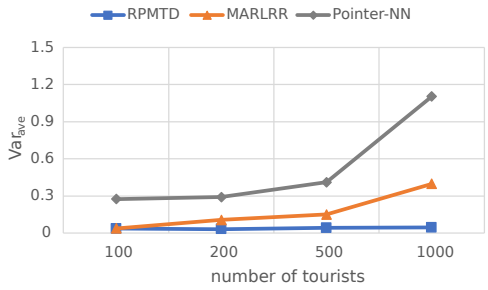
		Gini_{POI}	$\text{Gini}_{tourist}$	Var_{ave}	Var_{max}	ED_r
FYP-Data 100 tourists	RPMTD	**0.612**	**0.168**	**0.036**	**0.067**	**1315.286**
	MARLRR	0.713	0.176	0.038	0.071	1287.936
	Pointer-NN	0.821	0.196	0.275	0.356	69.328
FYP-Data 200 tourists	RPMTD	**0.625**	0.189	**0.032**	**0.071**	2265.351
	MARLRR	0.736	0.191	0.107	0.181	**2313.025**
	Pointer-NN	0.868	**0.187**	0.292	0.478	93.836
FYP-Data 500 tourists	RPMTD	**0.651**	0.193	**0.043**	**0.081**	**3619.283**
	MARLRR	0.692	0.269	0.151	1.498	3592.875
	Pointer-NN	0.921	**0.186**	0.412	1.754	151.802
FYP-Data 1000 tourists	RPMTD	**0.674**	0.216	**0.045**	**0.096**	**5812.231**
	MARLRR	0.851	0.659	0.398	0.967	5762.922
	Pointer-NN	0.948	**0.201**	1.104	3.257	184.064



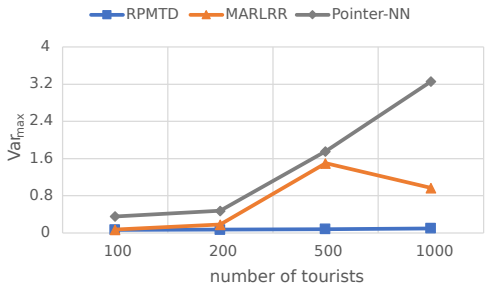
((a)) Gini_{POI} comparison



((b)) $\text{Gini}_{tourist}$ comparison



((c)) Var_{ave} comparison



((d)) Var_{max} comparison

Figure 3.8: Comparison of Gini coefficient and variance based on FYP-Data
.

Table 3.5: Result of ablation study on FYP-Data with 1000 tourists.

ω_1	ω_2	Gini_{POI}	$\text{Gini}_{tourist}$	Var_{ave}	Var_{max}	ED_r
0.01	1	0.784	-	0.081	0.192	6059.608
0.1	1	0.674	-	0.045	0.096	5812.231
1	1	0.667	-	0.041	0.091	5864.692
10	1	0.670	-	0.042	0.089	5901.156

tourists scenarios. However, $\text{Gini}_{tourist}$ of MARLRR increases significantly in large-scale tourists scenarios. In terms of Var_{ave} and Var_{max}, RPMTD shows the best performance. Pointer-NN performs much worse than the other two models because tourists are planned in similar POIs.

Figure 3.8 illustrates the comparison of Gini coefficient and variance on FYP-Data. Three models show the same tendency in Gini_{POI} that unfairness always grows with increasing tourists. Except $\text{Gini}_{tourist}$, Pointer-NN performs much worse than RPMTD and MARLRR, especially in the case of 500 and 1000 tourists, which indicates that the existing single tourist planning methods cannot be applied for large-scale tourists planning. The performance of MARLRR is generally worse than that of RPMTD and unstable. In contrast, RPMTD shows better and more robust performance with scalable tourists.

3.5.6 Ablation Study

Different ratios of dual congestion will affect the policy learning, ultimately affecting the distribution of tourists. We study the impact of different ratios of dual congestion on model performance by fixing the weight of local congestion ω_2 as 1 and setting different weights of global congestion ω_1. The ratio of dual congestion will affect the calculation of the tourist's reward. This implies that under different ratios, the tourists' rewards cannot be standardized, making $\text{Gini}_{tourist}$ comparison inapplicable. Thus, $\text{Gini}_{tourist}$ is not compared here. The experiment is conducted on FYP-Data with 1000 tourists, and the weights are normalized in implementation. Table 3.5 shows the result.

Generally, global congestion contributes to Gini_{POI}, Var_{ave}, and Var_{max}. When ω_1 increases from 0.01 to 0.1, Gini_{POI}, Var_{ave}, and Var_{max} increase by 14%, 44%, and 49% respectively. The global congestion effect is very weak when ω_1 is 0.01; the total reward is dominated by local congestion, and the model degenerates into a single-congestion model. Less well-known POIs do

Table 3.6: Comparison based on total static and dynamic reward on FYP-Data

		Reward_{static}	$\text{Reward}_{dynamic}$
200 tourists	RPMTD	10366	**5582**
	MARLRR	10537	4621
	Pointer-NN	**13205**	342
1000 tourists	RPMTD	49792	**241847**
	MARLRR	42377	150854
	Pointer-NN	**65830**	910

not have visits. When ω_1 increases to 0.1, the total reward is contributed by global and local congestion. Less well-known POIs have visits. When ω_1 increases from 1 to 10, performance is relatively stable, indicating global congestion becomes dominant.

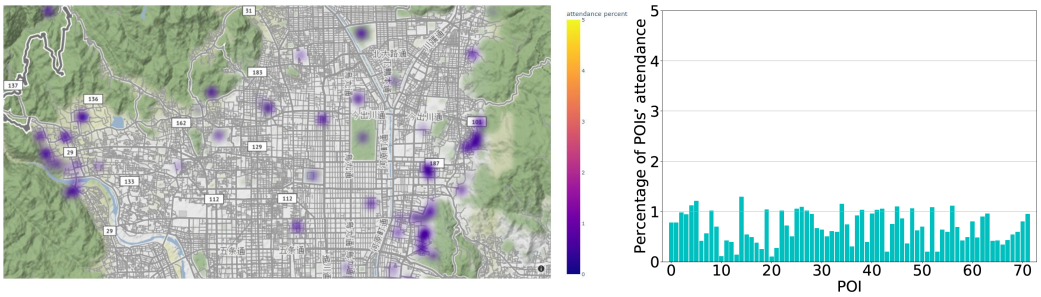
3.5.7 Scalability

Each model uses an individual reward policy, which is unified. Thus, we make a comparison based on both static and dynamic/congestion rewards. Table 3.6 shows the results, which are calculated with small-scale and large-scale tourists on FYP-Data.

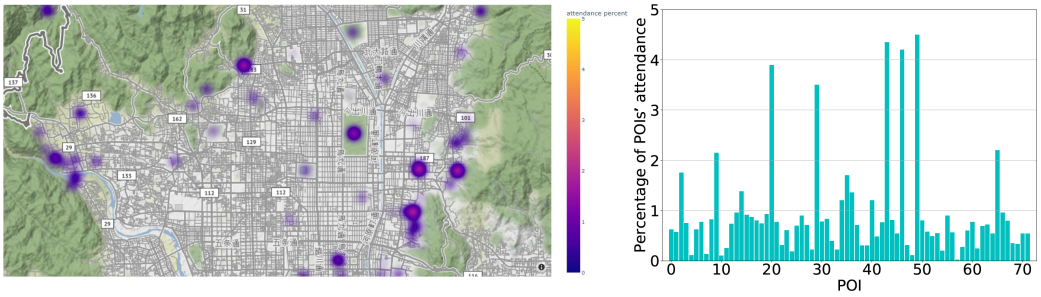
Pointer-NN shows the best performance in Reward_{static} but the worst performance in $\text{Reward}_{dynamic}$. The reason is that homogeneous route planning leads to POIs' congestion and low dynamic reward. This problem is exacerbated by large-scale tourists. MARLRR and RPMTD show comparable results in small-scale tourists scenarios. In large-scale tourists scenarios, RPMTD significantly outperforms MARLRR in $\text{Reward}_{dynamic}$. This result is aligned with the observation in Figure 3.8 that MARLRR performs unstably in large-scale tourists scenarios.

3.5.8 Case Study

Both RPMTD and MARLRR achieve Var_{max} around 15:00. Figure 3.9 shows the comparison and visualization based on FYP-data with 1000 tourists. RPMTD maintains each POI's attendance percentage generally under 100%. MARLRR result shows several POIs exceeding the capacity significantly. In MARLRR planning, tourists in the same region would be planned to the POI with the highest score, which might be the reason for its unstable performance



((a)) POIs' attendance percentage of RPMTD and the visualization



((b)) POIs' attendance percentage of MARLRR and the visualization

Figure 3.9: Comparison and visualization of all POIs' attendance percentage at 15:00 in FYP-Data.

with large-scale tourists. RPMTD balances between the tourists' distribution and the high visiting score.

3.6 Summary

In this section. we propose the popularity-biased routes planning problem. We solve this problem by introducing multi-agent RL framework, which includes an environment for interaction with multiple tourists and a dual-congestion aware model. In our experiment, we consider three novel mobility data to make it more close to the reality. Extensive experiments and comparison with baseline models are conducted, the result shows our model superior performance and robustness. Our work is the first to conduct multiple tourists' routes planning from the interest of POIs. The proposed RL environment provides a prerequisite for other study of multi-agent RL in this domain. Meanwhile, we find the new problem that bias grows with increasing tourists, which could be our further study. Additionally, our work could be incorporated with existing models for the sustainable development goal of tourism. We leave these as our further work.

Chapter 4

User Study of Congestion-aware Route Planning

Our work RPMTD considers the interests of POIs in route planning, which mainly benefits the sustainable development of tourism and local residents. We evaluate the effectiveness of our work on the distribution of tourists. However, tourism is a service-oriented industry, which highlights the critical significance of the tourists' experience. Considering the selfish nature of tourists, those who are scheduled away from popular POIs may feel less satisfied and unfairly treated. In other words, the relationship between sustainable tourism and tourists' preferences could be adversarial. A sustainable congestion-aware route planning model should, on the one hand, cater to tourists' satisfaction, and on the other hand, take into account the benefits of POIs when planning routes. Therefore, a survey is conducted to empirically study the impact of congestion-awareness on user experience, which could serve as reference for its future practical implementation.

4.1 Methodology

Our user study is part of Yi et al. [52]. Unlike theirs, our analysis is based on game theory, and we focus on discussion of cooperative and non-cooperative relationships.

A questionnaire is assigned to 41 participants in Kyoto for user experience evaluation. Four route planning models with different degrees of congestion awareness are examined, which is aligned with our previous experiments.

- **MARLRR** considers single dynamic congestion for individual visited POI.
- **RPMTD** considers both local congestion of visited POI and global congestion of tourists' distribution.
- **Non-Dual RPMTD** only considers local congestion of visited POI, and global congestion is not considered.
- **Point-NN** does not consider any congestion at all.

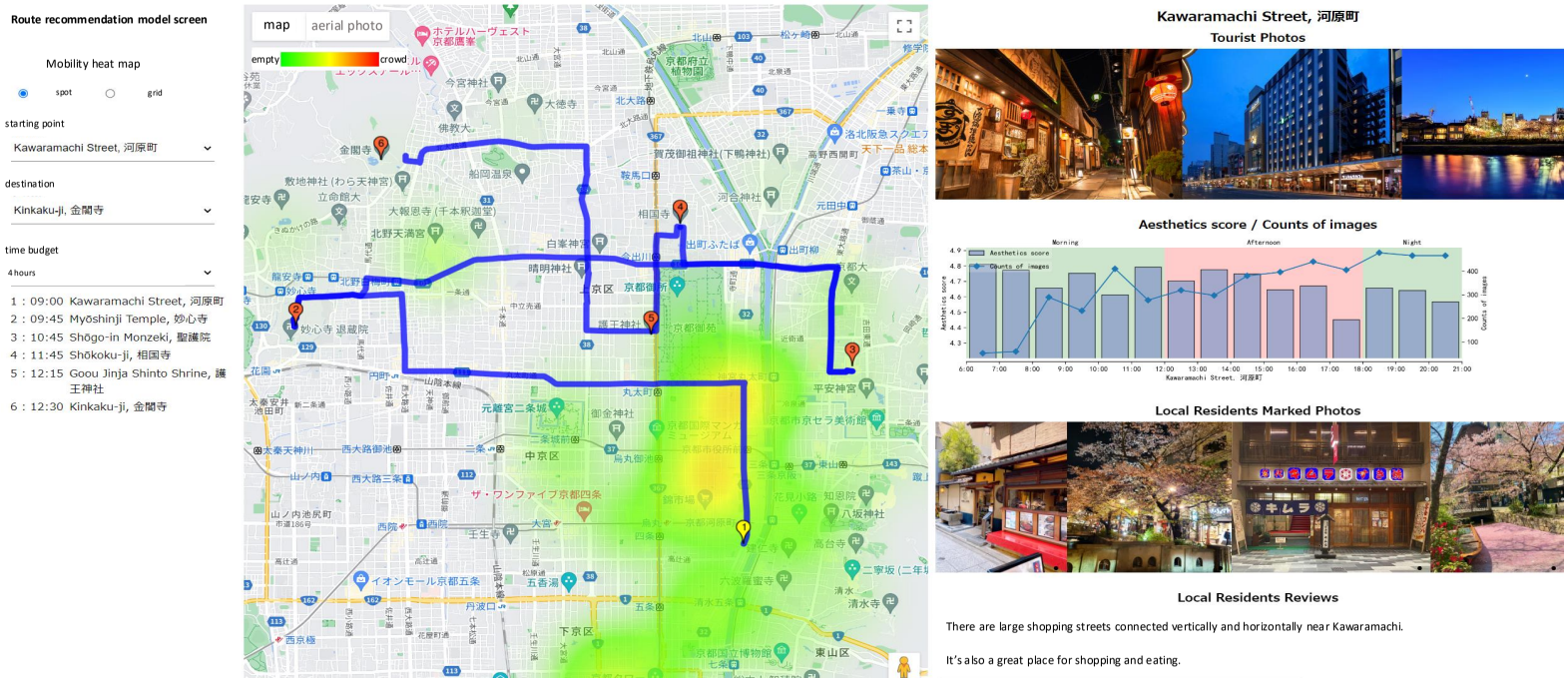


Figure 4.1: An example of questionnaire screen for user study of congestion-aware route planning in Kyoto, which includes three sections: (i) query and planning details on the left; (ii) visualization of planned routes and congestion level in the middle; (iii) POI-specific information on the right

Five popular queries are conducted in the survey:

- (4 hours, Kyoto Tower, Kawaramachi)
- (6 hours, Kyoto Tower, Kawaramachi)
- (8 hours, Kyoto Tower, Kawaramachi)
- (4 hours, Kawaramachi , Arashiyama Station)
- (6 hours, Kawaramachi , Kinkakuji Temple)

Each model generates one route for each query. In total, 20 routes are generated.

In the questionnaire, participants are asked to evaluate each route based on the information shown on a screen. Figure 4.1 shows an example of the screen, which is divided into three sections. The left section illustrates the details of the query and planned route. The middle section visualizes the planned route and congestion level of the city on the map. The right section shows the information on visited POI, including photos, comments, aesthetics score, and POI's congestion level.

Participants evaluate each route from five aspects:

- **Time Scheduling** evaluates whether the time planned for sightseeing and moving is reasonable.
- **Visiting Order** evaluates whether the order of sightseeing POIs is reasonable.
- **Traveling Distance** evaluates whether the distance between planned POIs is reasonable.
- **Traveling Comfort** evaluates comfort level regarding congestion in POI sightseeing.
- **Overall Satisfaction** evaluates comprehensive satisfaction.

Each aspect is rated from 1 to 5, where 1 indicates very unsatisfied; 2 indicates unsatisfied; 3 indicates neutral; 4 indicates satisfied; 5 indicates very satisfied. Participants could give comments/reasons for their evaluation.

4.2 Survey Result

Figure 4.2 illustrates the details of satisfaction levels for all models.

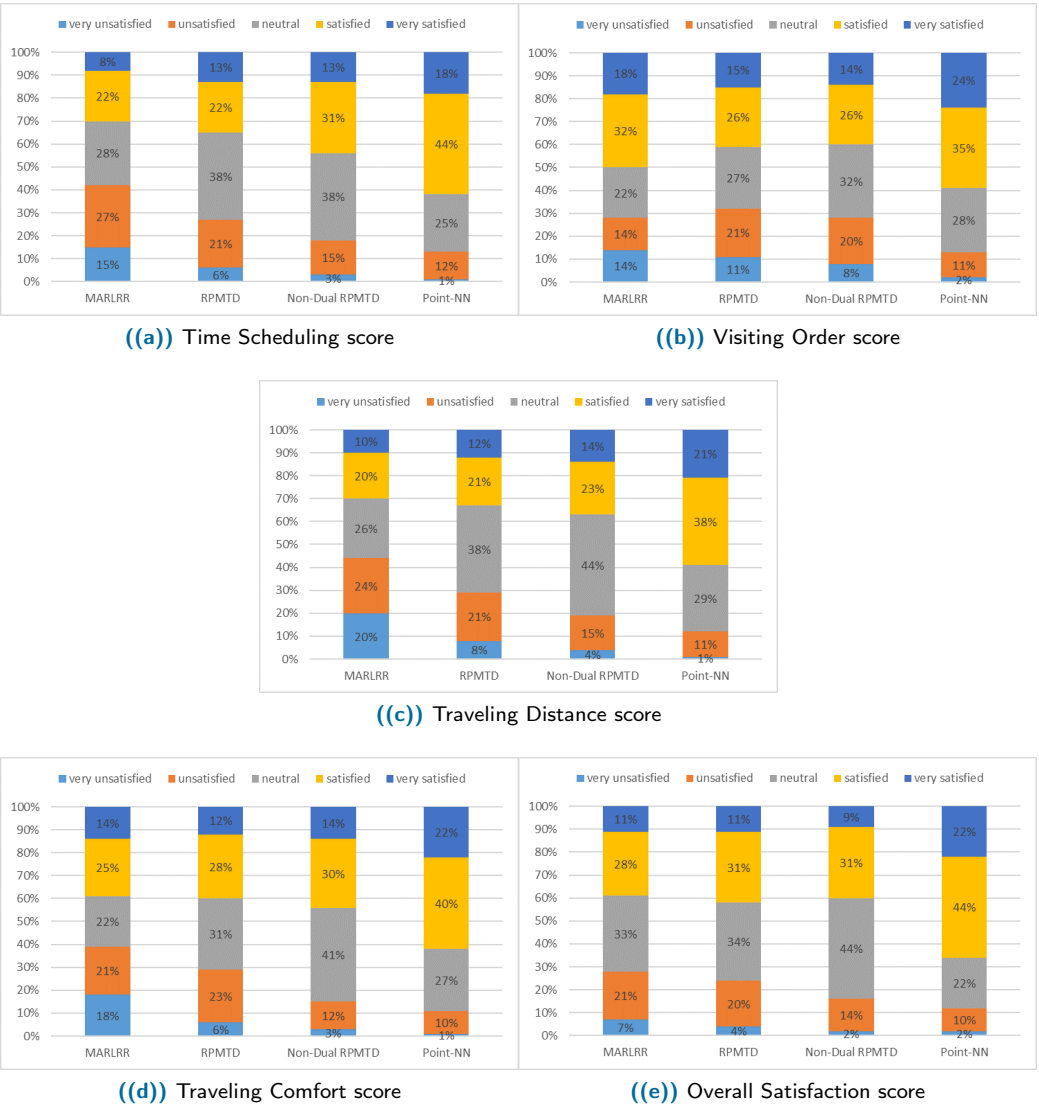
Figure 4.2(a) shows time scheduling evaluation. Point-NN achieves most "satisfied" or "very satisfied" (62%). MARLRR achieves the most "very unsatisfied" (15%). RPMTD has higher "very unsatisfied" or "unsatisfied" than Non-Dual RPMTD.

Figure 4.2(b) shows the visiting order evaluation. Point-NN achieves the most "satisfied" or "very satisfied" (59%), followed by MARLRR (50%).

RPMTD and Non-Dual RPMTD show comparable results that achieve around 40%.

Figure 4.2(c) shows the traveling distance evaluation. Point-NN achieves the most "satisfied" or "very satisfied" (59%), followed by Non-Dual RPMTD (37%). RPMTD and MARLRR show comparable results achieving around 30%. MARLRR achieves the most "unsatisfied" or "very unsatisfied" (44%).

Figure 4.2(d) shows the traveling comfort evaluation. Point-NN achieves



((a)) Time Scheduling score

((b)) Visiting Order score

((c)) Traveling Distance score

((d)) Traveling Comfort score

((e)) Overall Satisfaction score

Figure 4.2: Survey result for all methods on aspects of time scheduling, visiting order, traveling distance, traveling comfort, and overall satisfaction.

Table 4.1: Weighted score of aspects for each method.

	MARLRR	RPMTD	Non-Dual RPMTD	Point-NN
Time Scheduling	2.81	3.15	3.42	**3.66**
Visiting Order	3.25	3.13	3.25	**3.68**
Traveling Distance	2.76	3.08	3.34	**3.67**
Traveling Comfort	2.96	3.17	3.46	**3.72**
Overall Satisfaction	3.15	3.25	3.37	**3.74**

the most "satisfied" or "very satisfied" (52%), followed by Non-Dual RPMTD (44%). RPMTD and MARLRR show comparable results that achieve around 30%. MARLRR achieves the most "unsatisfied" or "very unsatisfied" (39%)

Figure 4.2(e) shows the overall satisfaction evaluation. Point-NN achieves the most "satisfied" or "very satisfied" (66%). Non-Dual RPMTD, RPMTD and MARLRR show comparable results that achieve around 40%.

The result is summarized as the weighted score for each aspect and model in Table 4.1.

4.3 Discussion

4.3.1 Model Comparison

Point-NN vs. Congestion-aware Models: Point-NN outperforms congestion-aware models in terms of weighted scores of all aspects. Specifically, regarding the weighted score of overall satisfaction, POINT-NN outperforms MARLRR, RPMTD, Non-Dual RPMTD by 15%, 13%, and 10%, respectively. There could be two reasons: (i) congestion-aware models avoid visiting the most popular POIs, while Point-NN mainly travels to the most popular POIs. Although tourists have different preferences for POIs, satisfaction will be greatly reduced if the must-see POIs are not included in planned routes; (ii) congestion-aware models affect other aspects, which finally results in reduced overall satisfaction.

RPMTD vs. NON-Dual RPMTD: Global congestion awareness considers the overall distribution of tourists, representing the fairness of POIs. RPMTD tends to plan tourists to less well-known POIs, which results in Non-Dual RPMTD being 8% better than RPMTD on the weighted score of time scheduling, traveling distance, and overall satisfaction. This is also aligned with our previous thought that tourists would be less satisfied if they are scheduled to visit less well-known POIs. Therefore, determining congestion

level and balancing it with tourists' satisfaction is crucial for sustainable route planning models.

MARLRR vs. Non-Dual RPMTD: Both MARLRR and Non-Dual RPMTD only consider single congestion. However, Non-Dual RPMTD performs better than MARLRR in all aspects. Such differences should come from the reward design and the RL framework. Our multi-agent RL framework has an advantage in multiple tourists planning, especially in large-scale tourists scenarios.

4.3.2 Inconsistency of Evaluations

The objective evaluation of experiments shows the effectiveness of our model for proposed problem; while subjective evaluation shows decline in tourists' satisfaction, which indicates the inconsistency between two evaluations. In our opinion, the major reason is the different perspectives of these evaluations. Objective evaluation is based on global view, while the subjective evaluation of users is only based on the individual view of local POI. Therefore, tourists' evaluation is not comprehensive and cannot fully represents the performance of the model.

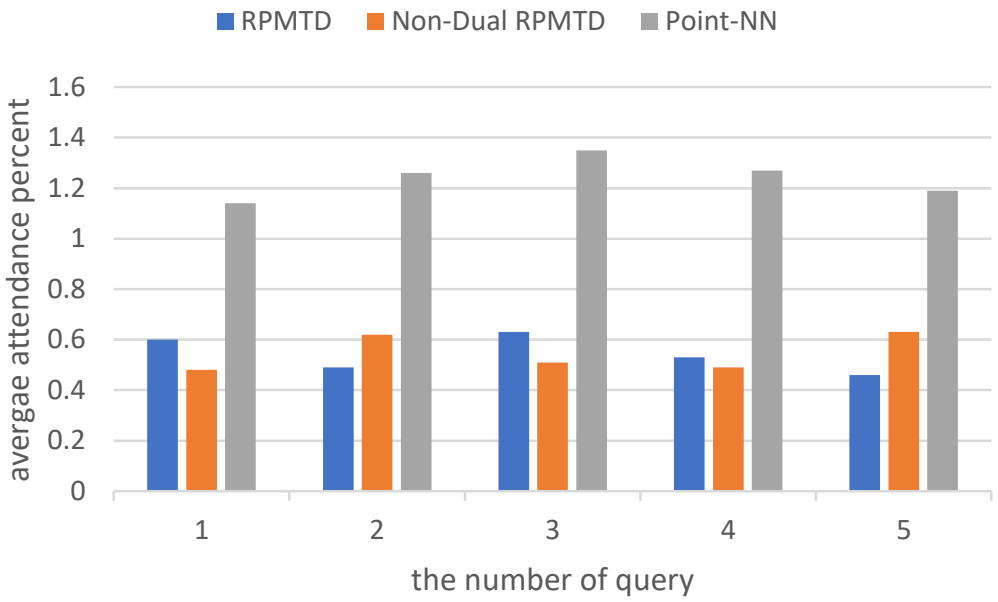


Figure 4.3: Comparison of average attendance percent of the POIs in the routes generated by our model and Point-NN based on 5 queries.

Additionally, Point-NN unexpectedly outperforms our model in terms of traveling comfort, which contradicts the result of our experiment. Thus, we compare the average attendance percent of the POIs in the routes generated by our model and Point-NN based on 5 queries. Figure 4.3 shows the comparison result, which illustrates that the congestion level of our model's route is better than that of Point-NN. It illustrates the problem that user's perception contradicts actual congestion level, and our analysis of this gap stems from the following two factors:

- The survey was conducted in 2023. As COVID-19 was still spreading in Japan, we utilized the virtual questionnaire on iPad instead of on-site sightseeing questionnaire. Participants rated the congestion score according to the heatmap on screen, which cannot make participants perceive the real-world congestion level. The virtual sightseeing questionnaire may have certain limitations, and on-site sightseeing will be conducted in our future surveys.
- In cognitive science, user's experience would be affected by prior knowledge [53, 54]. Thus, we suspect that tourists' perception of congestion could be related to their prior knowledge of POIs. For instance, the prevailing perception of tourists regarding popular POIs is their massive visitor volumes, which could cause a psychological adjustment and a higher congestion tolerance for these POIs. Conversely, lesser-known POIs are typically perceived as less crowded, which would cause heightened perceptions of crowding when there are slightly more tourists. This a novel problem identified in our research. Future investigations could delve into the effects of tourists' prior knowledge on their experience, extending our understanding of the cognitive mechanisms in the congestion-aware route planning.

4.3.3 Lesser-known POIs' attractiveness

Survey results show the decline of tourists' satisfaction as a result of being planned to lesser-known POIs. The observed decline in tourists' satisfaction should not be attributed to lesser-known POIs' low attractiveness. Conversely, we think that lesser-known POIs possess their unique attractiveness. To clarify this, a statistical analysis is conducted to compare the ratings of POIs in Kyoto, as sourced from Google Maps, against their attendance percentage during peak periods. The result is shown in Figure 4.4. POIs with an attendance percentage below 0.3 are categorized as lesser-known; POIs with an attendance percentage exceeding 1 are classified as popular; and the rest are classified as usual. The analysis reveals ratings of lesser-known POIs

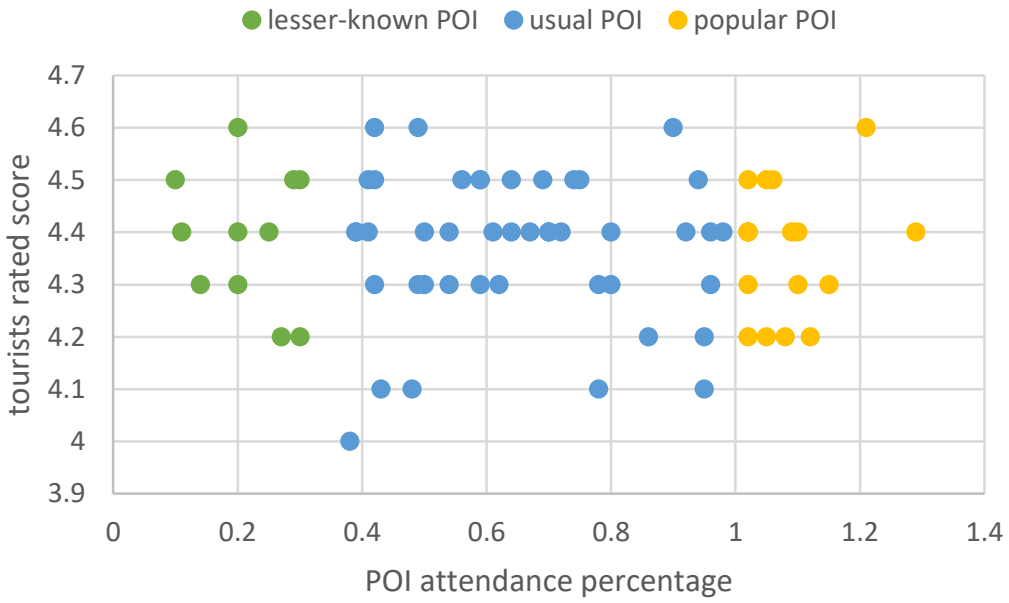


Figure 4.4: Statistics analysis between the scores of Kyoto POIs from Google Maps and the attendance percentage of each POI during rush hours.

are comparable to those of popular ones. The mean score of lesser-known POIs is 4.38, and the mean score of popular POIs is 4.37, indicating that lesser-known POIs could be attractive as popular POIs.

4.3.4 Non-cooperative vs. Cooperative Relation

In this study, we find that tourists tend to feel less satisfaction when they are scheduled to visit less popular POIs. Conversely, the less popular POIs experience more equitable and sustainable development due to the reasonable distribution of tourists. It illustrates an adversarial relationship in tourism [55, 56]. There is also research showing the same adversarial relationship in other recommendation systems, indicating recommending fair content harms user satisfaction [57]. In general, such a relationship is termed as *non-cooperative* relation in game theory.

It is important to emphasize that tourists feel lower satisfaction with less popular POIs, not because less popular POIs are not worth visiting or have low aesthetic value. In fact, many less popular POIs are just as beautiful and interesting as popular POIs. However, due to the Matthew effect [17, 58], popular POIs become famous and attract more tourists, and the influx of

more tourists brings even greater fame. On the contrary, less popular POIs have fewer visitors and gradually become less well-known. This can create a vicious cycle where popular POIs become even more popular, while less popular POIs become even less known. Our congestion-aware mechanism can help to break this vicious cycle by planning tourists to visit less popular POIs.

By helping to break the vicious cycle, our congestion-aware mechanism can benefit both tourists and the sustainable development of tourism. Tourists' satisfaction with less popular POIs is initially low due to their lack of understanding of these places. However, as less popular POIs receive more visits, their popularity increases, and thus their attraction to tourists and tourists' satisfaction will increase. In this process, tourists and the sustainable development of tourism will find mutual benefits, and their relationship transforms from the initially *non-cooperative* relationship to a *cooperative* one, which will benefit both parties [59].

4.3.5 Decline of Tourists' Satisfaction

In our opinion, the decline in tourists' satisfaction is predicted. This anticipated decline can be attributed to two factors. First, due to the adversarial relationship as discussed above, satisfaction of those tourists who are planned to lesser-known POIs will decline. Second, current evaluation is based on the perspective of single tourist. Our model specializes in multiple tourists planning and prioritizes collective benefits over individual preference. As pursuit of collective interests would result in compromising individual interests [60], our model is less satisfying compared with single-tourist planning model.

The survey results indicate about 15% decline in satisfaction. This is an empirical value in our research rather than the value in actual implementation. The primary objective of this survey is to gain an empirical understanding of the impact of our method on tourists' satisfaction, providing a foundational reference for its practical application. In various real-world scenarios, the effect of dual-congestion should be balanced by adjusting the weight of global congestion. Moreover, we think that the decline in tourists' satisfaction is a transient effect because our method could transform their relationship from non-cooperative to cooperative. Anticipating the long-term implications of this paradigm shift, we hypothesize a sustained improvement in tourists' satisfaction. Our future surveys will substantiate this hypothesis.

4.3.6 Unilateral Bias

This survey concentrates on tourists to understand their experience of congestion-aware route planning. However, only surveying tourists represents a limited perspective within the tourism industry. Other groups, such as local residents and the local governments, are also involved in this domain.

The impact of the congestion awareness mechanism on the entire tourism industry is a complex issue involving multiple stakeholders. Tourists are the obvious stakeholders, as they are the ones who first experience the congestion. However, local residents and businesses are also affected, as they may experience noise, pollution, and other negative impacts from congestion. The government also has a vested interest in congestion awareness, as it can contribute to reducing congestion and improving the overall quality of life.

Future surveys should include all of these stakeholders in order to get a more comprehensive and objective understanding of the impact of congestion awareness on tourism. This will help to ensure that any policies implemented to address congestion are effective and consider the needs of all stakeholders.

4.4 Summary

This user study is conducted to analyze the our method's impact on users' experience, and the result indicates that our model could transform the non-cooperative relationship between users and tourism into a cooperative one. This relationship transformation represents a significant advancement towards achieving sustainable tourism that caters to both tourists' preferences and broader environmental and social concerns. A more comprehensive and long-term survey regarding the relationship between tourists and tourism should also be further investigated.

Chapter 5

Collaborative and Intention-aware Communication for Scalable Multi-agent Framework

5.1 Introduction

Multiple tourists route planning requires a dynamic interaction of variable tourists as we discussed in previous chapters. Therefore, in our scenario, application of independent learning is effective because of its simplicity and adaptability. However, in independent learning, each agent is executed with a single-agent algorithm, indicating there is no communication between agents and each agent makes action only based on the their individual observation. In other words, each agent independently updates its policy based on its own experience, without considering the existence or strategies of other agents. This would cause the non-stationary problem, meaning that when an agent is learning to adapt to the environment, the parallel learning and strategy adjustments of other agents would cause the dynamics of the entire environment, making the state transition of the environment unstable and unpredictable for any single agent.

Table 5.1 shows an example of non-stationary problem in MARL. Black and blue agents obtain their observations at k step. The actions of black and blue agents are moving object to right and down, respectively. Their joint actions move the object to right-down position. In the next step, environment generates observations for both agents, which would confuse them, as the object is not on the their expecting position. In our scenario, non-stationary problem indicates that the change of mobility matrix is unpredictable for every single agent in current interaction because all agents jointly update the mobility matrix. Such a problem would discourage model's convergence and performance.

In MARL, the non-stationary problem usually could be solved by following aspects:

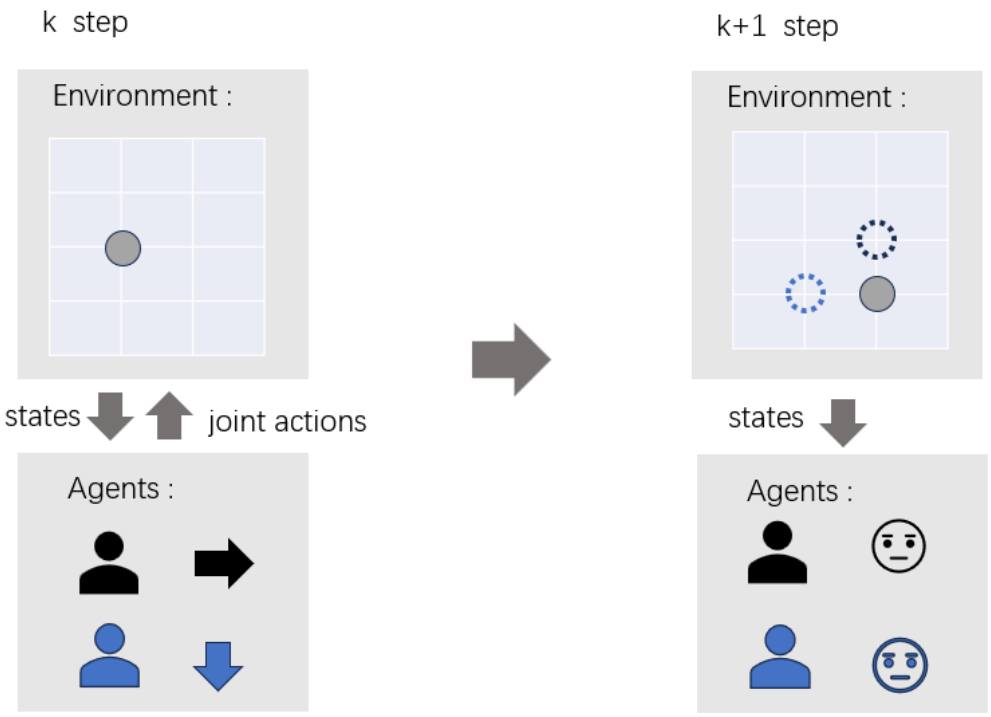


Figure 5.1: An example of non-stationary problem in MARL.

- **centralized training:** using the centralized training strategy, the information of all agents is used in the training phase to learn how to deal with uncertainty and dynamic changes in the environment. In MARL, apart from independent learning, there are other two paradigms: Fully Centralized (FC) and Centralized Training with Decentralized Execution(CTDE) methods. Both of these two methods apply a centralized controller for information sharing and coordination, which enables the individual agent access the global information for better decision making, accordingly alleviating the non-stationary problem. However, these two paradigms cannot be applied in our scenario because they cannot fulfill our dynamic requirement. FC methods face flexibility challenges and CTDE methods face the scalability problem.
- **communication mechanism:** establishing an effective communication protocol between agents so that they can share useful information and coordinate actions to respond to changes in the environment. Current MARL communication works are mostly developed from the CTDE methods. Because agents only communicate with limited-

bandwidth channel decentralized execution, and centralized training provide rich information for communication. Methods with communication protocol generally outperform the communication-free methods.

In this section, we focus on the development of communication protocol for our independent learning MARL framework, further improving our model performance in multiple tourists route planning. We consider two factors in our scenario: (i) large-scale agents communication; (ii) efficient information fusion.

We propose an intention-aware communication protocol for MARL based on existing attention methods. Our work is developed from two aspects:

- **intention communication:** each agent takes action at the same round, which means that everyone will not know the actions taken by other agents. Therefore, getting everyone's intentions and communicating before each agent takes action is significantly helpful to each agent's action making and coordination.

- **communicatee type:** in multi-agent communication, we need to answer two key questions: *(i) whom should communicate with; (ii) how to conduct communication*, especially in large-scale agents scenarios. Broadcasting with every agent is unnecessary and inefficient.

Our method consists of 3 components: (i) retrieval/denoise: individual agent selects other agents related to its own; (ii) information fusion: the single agent communicates with the selected collective agents; (iii) action: a single agent learns from other agents' intentions and make a decision. Specifically, in the first step, we propose intention-related and distance-related methods to select the most relevant agents. For the distance-related method, the agent only interacts with agents within a certain range around it. This is practical in other MARL environments, such as StarCraft. In our scenario, the intuition is that, tourists who are in the same region would have similar visiting plans. In the second step, we propose attention-based methods to obtain collective intentions and fuse the collective intentions with individual's. Finally, agents make action based on the fused intentions. We conduct experiments on our Kyoto tourism mobility dataset FYP-data and compare with our proposed baseline model RPMTD. Results shows that the proposed communication protocol can effectively promote model performance.

5.2 Related Works

5.2.1 Multi-agent Communication

Communication plays a crucial role in MARL. By understanding and communicating with other agents, an agent can better perceive changes in the environment. With the rapid development of CTDE framework, most existing works focus on the communication enhancement conditioned on this framework. VDN [61] and QMIX [62] are value decomposition methods, which specialize in decomposing a single joint reward for multiple agents in cooperative scenarios. These methods compose implicit communication learning; namely, agents do not explicitly communicate with others, however they could implicitly learn other agents' policies during training. Majority of researches focus on explicit communication. RIAL and DIAL [63] learn to pass message from current step to next step. CommNet [64] conducts communication by aggregating hidden layer representations of agents' respective networks. IC3Net [65] introduces the gating mechanism that determines whether an agent should communicate with other agents at given time step. Therefore, it contributes to scenarios where agents need to decide when to communicate and when to act. BiC-Net [66] utilizes Bi-directional Recurrent Neural Networks to process information between agents. This structure allows information to flow in both directions between agents, enabling agents to more fully understand the entire environment and the status of other agents. ATOC [67] combines attention mechanisms and communication protocols to dynamically decide when and how to pass information between agents. This approach allows an agent to adapt its communication behavior based on the needs of the current environment and the status of other agents. These methods communicate with all agents and totally depend on the central critic networks, which is highly cost and inefficient. Some works attempt agents selection and peer communication. TMC [68] proposes the mutual information to measure the information sharing in agents and minimize the entropy of message. TraMAC [69] use an additional key from the sender to the receiver to calculate the information importance. However, these methods are based on CTDE framework and unpractical for our scenario. Additionally, our model should process large-scale agents communication; however, these models work on small-scale scenario.

5.2.2 Intention of Agents

In cognitive science, especially in Theory of Mind, *intention* refers to the mental state of an individual planning to achieve a certain goal. This is a

core concept in understanding the behavior of others, as it involves predicting and interpreting the actions of others based not only on their current behavior but also on their purposes. In MARL, understanding and utilizing intentions can improve the efficiency of collaboration and coordination between agents. Qi et al. [70] attempts a linear function approximation of the utility function with consideration of the belief in the planning process. Fang et al. [71] consider a multi-agent reinforcement learning method for multi-order execution in finance and proposes a learnable communication protocol, involving the intention sharing with other agents. Xu et al. [72] let agents infer the intentions of nearby agents by their local observations and integrate it into their decision making in traffic planning. Kim et al. [73] propose an intention-sharing communication protocol based on the iteratively generated pseudo trajectory.

5.3 Proposed Method

We develop a multi-agent communication protocol which consists of three steps: (i) selection of most relevant agents; (ii) sharing and integration information with the selected agents; (iii) making an action based on the fused information. Our communication module is integrated into the decision-making pipeline of each agent's actor network. Specifically, it works prior to the final action determination stage. The module processes the intentions of each agent, facilitating inter-agent communication, before these

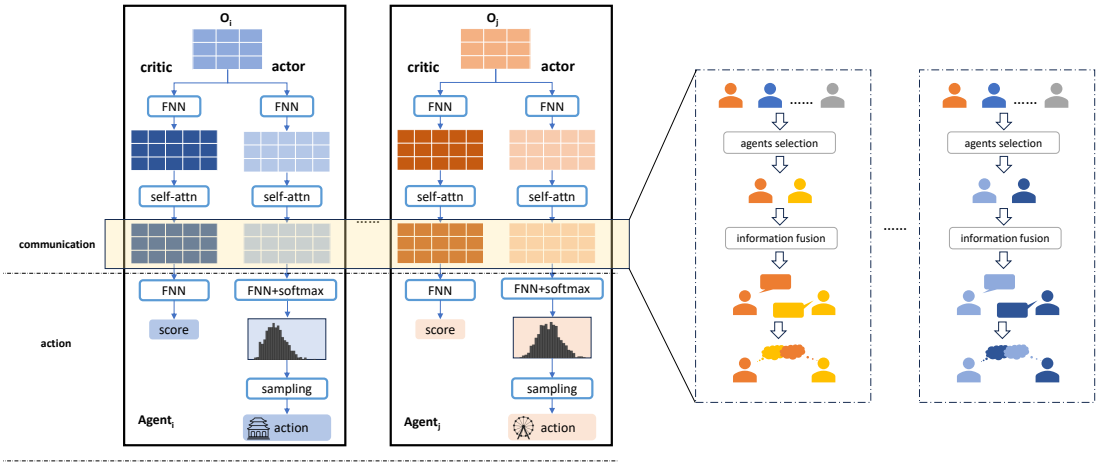


Figure 5.2: Communication module integrated in RPMTD.

intentions are passed to the softmax and sampling layers for action making. In our method, *intention* refers to the pre-determined cognitive state of each agent, representing their preliminary assessment prior to engaging with the probability distribution of all POIs. After several rounds of information sharing and fusion, each agent makes an action. Figure 5.2 shows the communication module integrated in RPMTD.

5.3.1 Communication Mechanism Structure

Given the specific agent's intention and the collective agents' intentions, we select agents from collective ones. Then the selected agents' intentions are fed

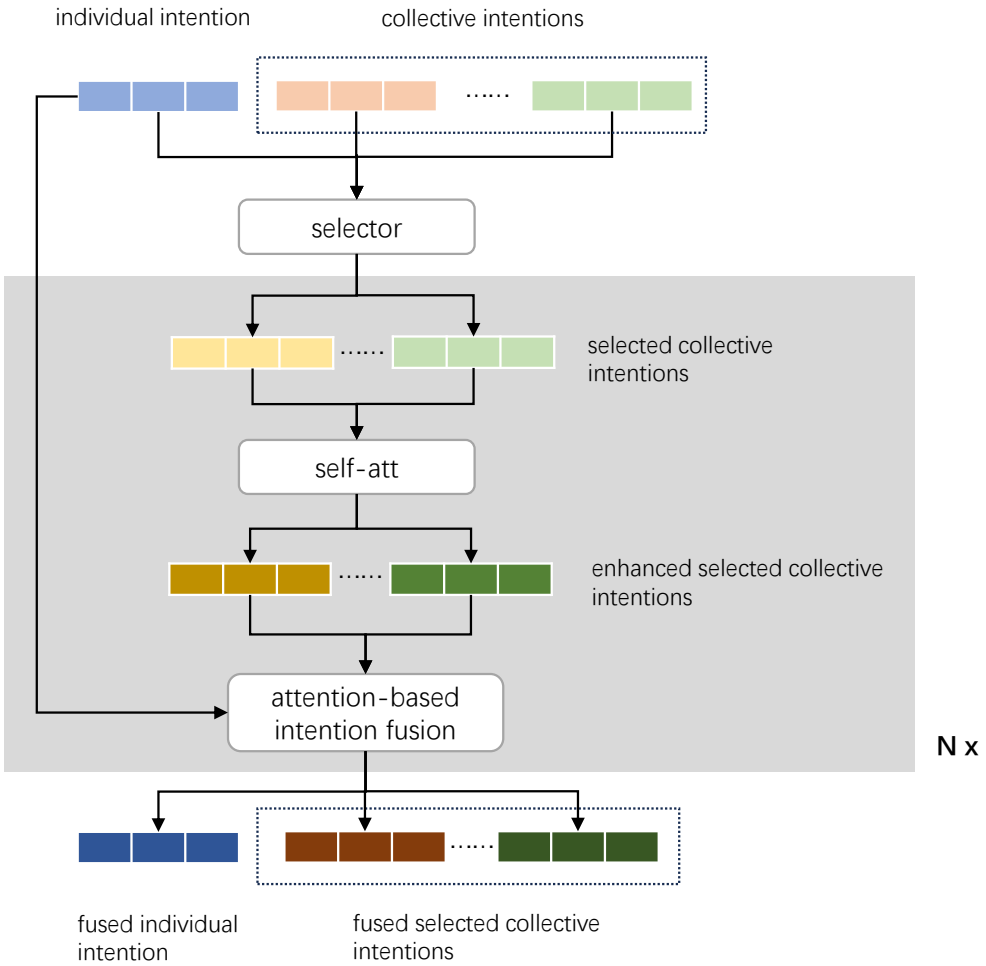


Figure 5.3: The structure of our communication mechanism.

in a self-attention layer to enhance their collective features. Subsequently, the feature-enhanced collective intentions are fed into the attention-based intention fusion layer together with the specific agent's intention for information sharing and integration. Afterwards, the fused individual intention and the selected collective intentions are obtained. The process of self-attention feature enhancement and attention-based intention fusion repeats for N rounds for better information aggregation. Finally, the action is made based on the fused individual intention. Figure 5.3 shows the structure of our communication mechanism.

5.3.2 Agents Selector

For a specific agent, given the collective intentions of all agents, we first select the relevant agents for that agent. The selector is the crucial part of communication mechanism and we implement two types of agents selector: intention-based and distance-based selector. In our methods, we implement either of them.

- The intention-based selector aims to neglect the agents who have irrelevant route planning intentions and focus on the ones who have the similar planning. The intuition is that the agents having the similar planning are more likely to visit the same POI and cause congestion. We firstly use a FNN to map the intention tensor into the hidden space and then apply cosine similarity to calculate the similarity between the specific agent's and other ones'. The cosine similarity equals to 1 meaning that the intentions of the two agents are exactly the same; the cosine similarity equals to 0 meaning orthogonality and the intentions of the two agents are not related; the cosine similarity equals to -1 meaning that the intentions of the two agents are opposite. Agents selection is based on the threshold of cosine similarity which is set as a hyperparameter θ. We select agents whose cosine similarity with the specific agent is greater than θ.
- The distance-based selector indicates that agents only communicate with ones who are in a certain range. This is also aligned with other MARL environment such as Starcraft that each agent can only observe information within a certain range around it, forcing the agent to make decisions based on partial information and learn how to coordinate actions with other agents nearby. Agents selection is according to the distance hyperparameter k. We select agents who are within k distance from the specific agent.

5.3.3 Attention-based Intention Fusion

The selected collective intentions $\mathbf{E}_c \in \mathbb{R}^{l_c \times d}$ and individual intention $\mathbf{E}_i \in \mathbb{R}^{1 \times d}$ are given, where l_c is the number of selected agents. d denotes the size of intention. Then the individual intention needs to be intensified by the selected collective intentions. For this purpose, we apply the attention mechanism to learn the relationship between them. To validate the generality of attention model's effectiveness, we conduct experiments with three kinds of attention mechanisms: co-attention (pre-self-attention), self-attention and cross-attention (post-self-attention). In our method, either of them could be implemented:

$$\mathbf{I}', \mathbf{C}' = \text{attention\_fusion}(\mathbf{E}_i, \mathbf{E}_c) \tag{5.1}$$

where $\mathbf{I}'$ is the fused individual intention and $\mathbf{C}'$ is fused (selected[1]) collective intentions.

5.3.3.1 Co-attention

Co-attention [74] is a vital model for QA. It enables the question (individual) and context (collective) to attend mutually, and also learns the question-aware context representation iteratively. We implement it as follows: collective intentions $\mathbf{E}_c$ and individual intention $\mathbf{E}_i$ is mapped into the hidden dimension by FFNs. Affinity vector $\boldsymbol{A}$ is the product of collective intentions $\mathbf{C}$ and individual intention $\mathbf{I}$. In vector $\boldsymbol{A}$, each value is the related score of individual intention and other intention from the collective ones:

$$\mathbf{C} = \text{FFN}_1(\mathbf{E}_c) \in \mathbb{R}^{l_c \times h} \tag{5.2}$$

$$\mathbf{I} = \text{FFN}_2(\mathbf{E}_i) \in \mathbb{R}^{1 \times h} \tag{5.3}$$

$$\boldsymbol{A} = \text{softmax}(\mathbf{C}\mathbf{I}^\top) \in \mathbb{R}^{l_c \times 1} \tag{5.4}$$

By multiplying $\boldsymbol{A}$ with collective intentions $\mathbf{C}$, we can obtain the individual intention $\mathbf{I}'$ attended by the collective intentions. Similarly, we derive the collective intentions $\mathbf{S}_c$ attended by the individual intention as follows:

$$\mathbf{I}' = \boldsymbol{A}^\top \times \mathbf{C} \in \mathbb{R}^{1 \times h} \tag{5.5}$$

$$\mathbf{S}_c = \boldsymbol{A} \times \mathbf{I} \in \mathbb{R}^{l_c \times h} \tag{5.6}$$

where $\top$ denotes the transpose.

Let the updated individual intention $\mathbf{I}'$ attend collective intentions again with $\boldsymbol{A}$. In addition, the attended collective intentions is further fed into a BiGRU as follows:

$$\mathbf{D}_c = \text{BiGRU}(\boldsymbol{A} \times \mathbf{I}') \in \mathbb{R}^{l_c \times h} \tag{5.7}$$

[1]From here on, "selected" is omitted for simple expression.

$\mathbf{D}_c$ and $\mathbf{S}_c$ are collective intentions intensified by the individual intention. Finally, they are concatenated and further applied with the FFN_d to transform into the original length:

$$\mathbf{C}' = \text{FFN}_d([\mathbf{D}_c||\mathbf{S}_c]) \in \mathbb{R}^{l_c \times h} \tag{5.8}$$

where $[\cdot||\cdot]$ denotes the concatenation function.

5.3.3.2 Self-attention

Self-attention (Transformer) [44] is a revolutionary model in NLP development. It enables a model to weigh the importance of different words in a sentence relative to each other. The result is a dynamic attention model that adjusts based on the input, allowing for a context-aware representation that enhances tasks like translation, summarization and QA. Self-attention block consists of several components: Multi-head attention, Add & Norm layer, FFN and Residual Connection.

The core component of self-attention is Multi-head attention. The input x is mapped in to three spaces: key ($\mathbf{K}$), query ($\mathbf{Q}$) and value ($\mathbf{V}$) by three linear transformations:

$$\mathbf{Q} = xW^Q, \quad \mathbf{K} = xW^K, \quad \mathbf{V} = xW^V \tag{5.9}$$

For each head, attention scores are computed using the dot product of queries and keys, scaled by the dimensionality of the keys ($\sqrt{d_k}$), where d_k is the dimension of the key. The softmax function is applied to obtain the weights on the values:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{5.10}$$

The model projects the queries, keys, and values several times with different linear transformations. This results in multiple sets of $\mathbf{K}$, $\mathbf{Q}$ and $\mathbf{V}$ for each head. The attention function is applied independently on each set of projections, allowing the model to jointly attend to information from different representation subspaces at different positions. The outputs of each head are concatenated and then projected once more with another learned weight matrix W^O:

$$\textbf{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{5.11}$$

where $\text{head}_i = \text{Attention}(\mathbf{Q}W_i^Q, \mathbf{K}W_i^K, \mathbf{V}W_i^V)$.

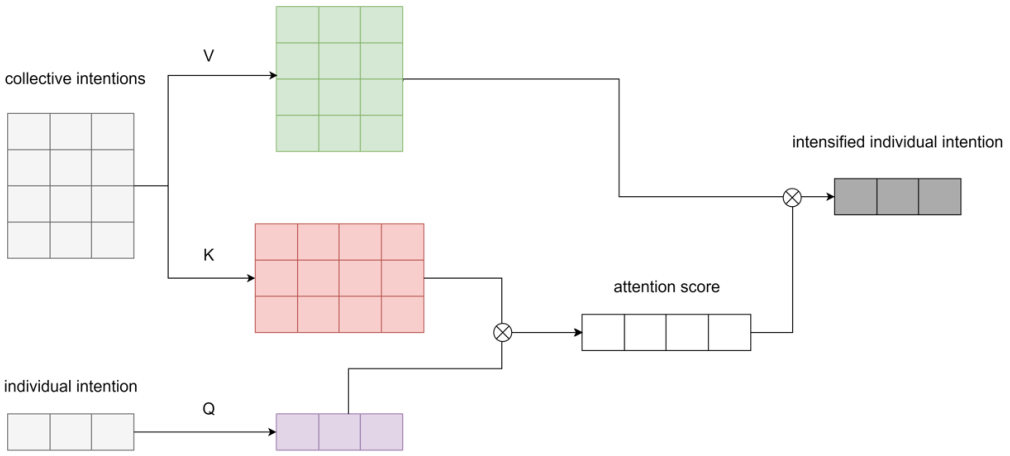


Figure 5.4: The interaction of **K**, **Q** and **V** in cross-attention.

In our method, we firstly concatenate the individual intention $\mathbf{E}_i$ and collective intentions $\mathbf{E}_c$, and then feed them into self-attention block for information fusion:

$$\mathbf{I}', \mathbf{C}' = \text{Self\_attention}(\mathbf{E}_i || \mathbf{E}_c) \tag{5.12}$$

5.3.3.3 Cross-attention

Cross-attention [44] is also proposed in Transformer. The structure of cross-attention is similar with self-attention. The difference is that self-attention focuses feature enhancement within the same sequence to model relationships among its elements. Cross-attention focuses on elements from one sequence in relation to another sequence. Therefore, it specializes in two resources information fusion, which properly meet the requirement in our communication method. Specifically, in our scenario, the key (**K**), value (**V**) come from the collective intentions, and query (**Q**) comes the individual intention. The remaining of operations are same with self-attention. Figure 5.4 shows the interaction of **K**, **Q** and **V** in cross-attention. The input is collective intentions and individual intention, and the output is intensified individual intention or collective intentions. We utilize two cross-attention blocks to intensify individual and collective intentions, respectively:

$$\mathbf{I}' = \text{Cross\_attention}_1(\mathbf{E}_c, \mathbf{E}_i) \tag{5.13}$$

$$\mathbf{C}' = \text{Cross\_attention}_2(\mathbf{E}_i, \mathbf{E}_c) \tag{5.14}$$

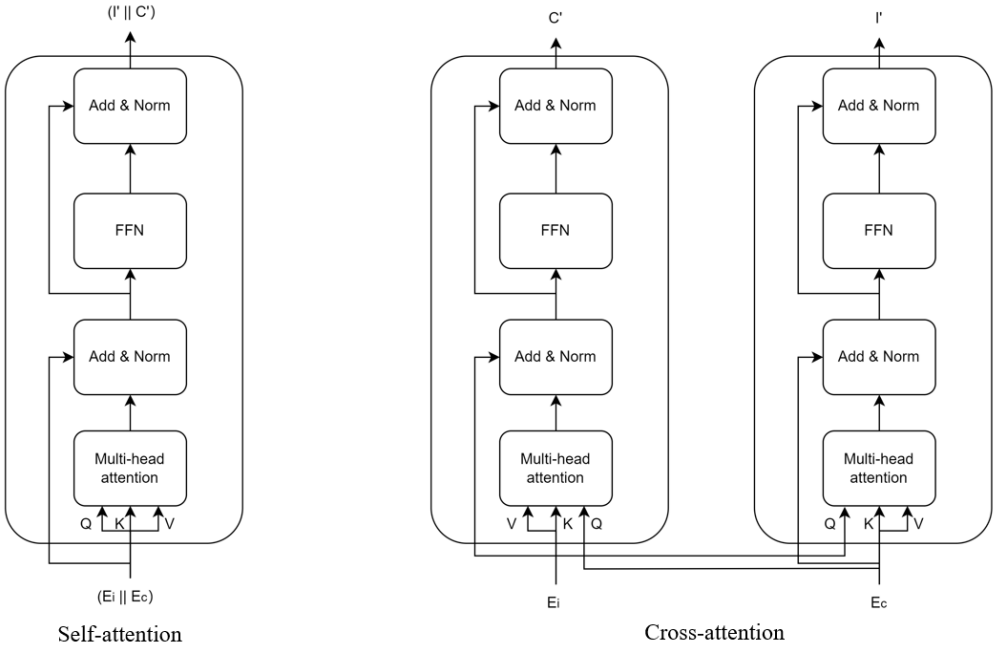


Figure 5.5: Comparison of self-attention and cross-attention.

Figure 5.5 shows the comparison of self-attention and cross-attention implemented in our method.

5.3.3.4 Other Implementation

Apart from the our proposed communication mechanism, we also consider other tricks which are widely used to counter on non-stationary problem and improvement of multi-agent cooperation:

- **implicit state update**: the uncertainty of non-stationary dynamic in our scenario is the mobility information updated by all activated agents. To alleviate this problem, we propose to update the observation of next agent, when one agent makes an action. This operation is performed sequentially according to the agent sequence. Although our agents are not a sequence but a set, this trick is expected to reduce impact of non-stationary problem.
- **joint optimization**: in our scenario, each agent makes an action according to its observation and obtains a reward from environment individually. In cooperative scenarios, multiple agents receive a shared reward based on the collective actions and outcomes of the entire group.

Joint reward is used to promote cooperation among agents, guiding them to work together towards a common goal. This mechanism is especially valuable in scenarios where agent collaboration is essential for success. Thus, we consider the mean of rewards from all activated agents in one interaction:

$$reward = \frac{1}{N} \sum_{1}^{N} Reward(agent_i(O_i)) \tag{5.15}$$

where N denotes the number of activated agents in the interaction.

5.4 Experiments and Results

5.4.1 Experiment setting

Mobility Data Preparation In order to be consistent with previous experiments, we conduct our experiments on our Kyoto real-human mobility dataset FYP-Data.

Tourists Setting We also keep consistent with previous experiments. We generate 100, 200 tourists as small-scale tourists and 500, 1000 as large-scale tourists.

Implementation details The FFNs implemented in this section are all set as 2-layer with 0.2 dropout. Dropout in BiGRU of co-attention is set as 0.3. Transformer encoder is adopted for self-attention module. We stack 2 layers and 8 heads for both self-attention and cross-attention blocks. In the intention-based selector, θ is selected from { -0.3, **0** (default) and 0.3}. In distance-based selector, k is selected from {1km, **1.5km** (default) and 2km}. Repeat number N of self-attention feature enhancement and attention-based intention fusion is selected from {1, 2 and **3** (default)}. Other hyperparameters are set the same with previous implementation. We implement experiments on one A100 GPU server.

Baseline setting Since our communication mechanism design is based on our previous work, we select RPMTD as our baseline model, which adopts IPPO as MARL algorithm. Additionally, we also consider IQL which fails in model convergence to validate the effectiveness of our communication.

5.4.2 Evaluation Metrics

We follow our previous study evaluation metrics, except that total static rewards is not considered:

(1) Maximum and average variance of all POIs' attendance percentage (Var_{max} and $\text{Var}_{average}$) throughout the whole trip
(2) Gini coefficient [51] of POIs (Gini_{POI}) and tourists ($\text{Gini}_{tourist}$)
(3) Total congestion-aware/dynamic reward ($\text{Reward}_{dynamic}$[2])
(4) Average of edit distance (ED_r)

5.4.3 Experimental Results

5.4.3.1 Main Results

We first validate our communication method on FYP-Data with IPPO as MARL algorithm implemented. For intention-based and distance-based selectors, θ and k is set as 0 and 1.5, respectively. Information fusion repeat N is set as 3. For the attention-based intention fusion model, cross-attention is implemented. Table 5.1 shows the result of RPMTD with two types of selectors communication methods. Both types of selectors can improve the model performance. We find that distance-based method is outperformed by intention-based method in small-scale tourists scenarios. Since distance-based selector might be influenced by the distribution of tourists. It strongly depends on the tourists around the agent. In small-scale tourists scenarios, there could be very few or even no other ones in the agent's communication range. In large-scale tourists scenarios, the distance-based method shows better performance in terms of Var_{ave} and ED_r. Generally, the intention-based method outperforms distance-based one.

We also validate our communication method on FYP-Data with IQL as MARL algorithm implemented. Without communication, IQL-based RPMTD fails in convergence in scenarios of more than 100 tourists. Result is shown in Table 5.2. With the intention-based communication, model converges with 200 tourists, even though the performance is much worse than IPPO-based RPMTD. And in large-scale (500 and 1000) tourists, IQL-based model still shows convergence failure. Such results indicate two points: (i) our two types of selector-based communication methods can both promote model performance, and intention-based selector shows better enhancement; (ii) IQL does not work well for large-scale agent scenarios with our communication method. Compared with IQL, IPPO performs better in our scenario. Thus, we implement IPPO in following experiments.

[2]For simple expression, the subscript *'dynamic'* is omitted in following part.

Table 5.1: Result of IPPO-based RPMTD with communication based on FYP-Data. "/" means RPMTD without communication.

	selector	Gini_{POI}	$\text{Gini}_{tourist}$	Var_{ave}	Var_{max}	ED_r	Reward
FYP-Data 100 tourists	/	0.612	0.168	0.036	0.067	1315.286	3041
	intention-based	**0.582**	**0.159**	**0.034**	**0.064**	**1400.356**	**3554**
	distance-based	0.601	0.161	0.035	0.066	1350.294	3221
FYP-Data 200 tourists	/	0.625	0.189	0.032	0.071	2265.351	5582
	intention-based	**0.594**	**0.180**	**0.030**	**0.068**	**2418.873**	**6273**
	distance-based	0.603	0.182	0.031	0.069	2364.115	5748
FYP-Data 500 tourists	/	0.651	0.193	0.043	0.081	3619.283	131464
	intention-based	**0.586**	**0.174**	**0.039**	**0.073**	3671.201	**143498**
	distance-based	0.593	0.176	0.044	0.075	**3806.247**	138488
FYP-Data 1000 tourists	/	0.674	0.216	0.045	0.096	5812.231	241847
	intention-based	**0.606**	**0.195**	0.043	**0.087**	**6389.954**	**265225**
	distance-based	0.641	0.206	**0.041**	0.091	6202.306	253692

Table 5.2: Result of IQL-based RPMTD with communication methods based on FYP-Data. "/" means RPMTD without communication.

	selector	Gini_{POI}	$\text{Gini}_{tourist}$	Var_{ave}	Var_{max}	ED_r	Reward
FYP-Data 100 tourists	/	0.758	0.181	0.049	0.094	1268.541	2145
	intention-based	**0.712**	**0.169**	**0.046**	**0.088**	**1370.344**	**3182**
	distance-based	0.737	0.173	0.048	0.092	1320.443	3122
FYP-Data 200 tourists	/	-	-	-	-	-	-
	intention-based	**0.797**	**0.178**	**0.060**	**0.137**	**2301.671**	**4863**
	distance-based	-	-	-	-	-	-

5.4.3.2 Hyperparameter Search

We further investigate the hyperparameters impact on model performance. For intention hyperparameter θ, we search it from {-0.3, 0, 0.3}, with repeat hyperparameter N set as 3. We conduct experiments in scenarios of 200 and 1000 tourists and the results are shown in Table 5.3. It shows that $\theta = 0$ outperforms other hyperparameters in scenario of 200 tourists. For $\theta = 0.3$, threshold is set relatively high, agents may rarely find others, leading to insufficient communication and coordination. For $\theta = -0.3$, agents may communicate too frequently, with those whose intentions are not aligned enough. In scenario of 1000 tourists, results of $\theta = 0.3$ and 0 are similar, indicating that agents could have sufficient communication although threshold is high. This may also indicate that communication has reached the upper limit of the model, and interacting with more agents does not lead

Table 5.3: Result of intention hyperparameter θ search based on FYP-Data. rand_0.5 indicates random communication with possibly of 50%.

	θ	Gini_{POI}	$\text{Gini}_{tourist}$	Var_{ave}	Var_{max}	ED_r	Reward
FYP-Data 200 tourists	rand_0.5	0.639	0.207	0.047	0.092	2124.721	5481
	-0.3	0.631	0.191	0.053	0.103	2206.927	5594
	0	**0.594**	**0.180**	**0.030**	**0.068**	**2418.873**	**6273**
	0.3	0.623	0.190	0.035	0.072	2204.190	5645
	1	0.625	0.189	0.032	0.071	2265.351	5582
FYP-Data 1000 tourists	rand_0.5	0.648	0.254	0.060	0.103	5641.285	218745
	-0.3	0.652	0.212	0.056	0.114	5748.357	236162
	0	**0.606**	**0.195**	0.043	**0.087**	6389.954	**265225**
	0.3	0.637	0.205	**0.041**	0.092	**6470.855**	261963
	1	0.674	0.216	0.045	0.096	5812.231	241847

to more effective communication. Results of $\theta = -0.3$ and 1[3] are similar , and $\theta = -0.3$ are even worse than $\theta = 1$ in terms of variance, which means that agents' communication with irrelevant ones could bring noise for decision making. 50% randomness communication is also conducted as comparison, which underperforms other parameters. This data is obtained by averaging 5 experiments. Generally, $\theta = 0$ shows more stable performance than others; thus, considering both small and large-scale tourists, $\theta = 0$ is set as default.

We conduct similar experiments for distance hyperparameter k. We search it from {1, 1.5, 2}, with repeat hyperparameter N set as 3. The results are shown in Table 5.4. It tells the similar story that if agents communicate only within limited range, they may not receive sufficient information. Conversely, if agents communicate over a large range, it can lead to information overload and bring noise. From k=1 to k=1.5, the reward increases by more than 10%. From k=1.5 to k=2, the performance is stable. Therefore, k is set as 1.5 as default.

To verify the number of communication, we search hyperparameter N with two types of selectors from {1, 2, 3}, with k set as 1.5 and θ set as 0. Results are shown in Table 5.5 and Table 5.6. For intention-based model, the increase of communication does not bring significant improvement in model performance. From one to three-hop communication, the reward is improved by within 4%. Differently, the increase of communication enhances distance-based model performance by 12% in 1000 tourists scenario. Especially when the hop of communication increases from one to two-hop, reward is improved by 10%. Generally, the intention-based model shows more stable and better performance than distance-based one. The distance-based model could be

[3] $\theta = 1$ means no communication.

sensitive with the number of communication, number of tourists and tourists' distribution.

Table 5.4: Result of distance hyperparameter k search based on FYP-Data.

	k	Gini_{POI}	$\text{Gini}_{tourist}$	Var_{ave}	Var_{max}	ED_r	Reward
FYP-Data 200 tourists	1	0.673	0.193	0.055	0.105	1946.181	5152
	1.5	**0.603**	**0.182**	**0.031**	**0.069**	**2364.115**	**5748**
	2	0.624	0.184	0.037	0.075	2292.656	5575
FYP-Data 1000 tourists	1	0.691	0.220	0.048	0.103	5595.451	217502
	1.5	**0.641**	**0.206**	**0.041**	0.091	**6202.306**	**253692**
	2	0.622	0.208	0.047	**0.089**	6181.908	229207

Table 5.5: Result of number of communication hyperparameter N search for distance-based model based on FYP-Data.

	N	Gini_{POI}	$\text{Gini}_{tourist}$	Var_{ave}	Var_{max}	ED_r	Reward
FYP-Data 200 tourists	1	0.687	0.236	0.048	0.113	1829.847	5430
	2	0.619	0.192	0.037	0.081	**2455.285**	5597
	3	**0.603**	**0.182**	**0.031**	**0.069**	2364.115	**5748**
FYP-Data 1000 tourists	1	0.732	0.283	0.049	0.129	5977.493	225019
	2	0.710	0.241	0.047	0.134	6104.472	247108
	3	**0.641**	**0.206**	**0.041**	**0.091**	**6202.306**	**253692**

Table 5.6: Result of number of communication hyperparameter N search for intention-based model based on FYP-Data.

	N	Gini_{POI}	$\text{Gini}_{tourist}$	Var_{ave}	Var_{max}	ED_r	Reward
FYP-Data 200 tourists	1	0.601	0.182	0.034	0.070	2361.045	6174
	2	0.598	0.184	0.035	**0.065**	2401.038	**6316**
	3	**0.594**	**0.180**	**0.030**	0.068	**2418.873**	6273
FYP-Data 1000 tourists	1	0.611	0.202	0.044	0.090	6270.031	257268
	2	0.612	**0.193**	0.047	0.092	6319.452	257026
	3	**0.606**	0.195	**0.043**	**0.087**	**6389.954**	**265225**

Table 5.7: Result of three intention fusion models based on intention selector.

	attention type	Gini_{POI}	$\text{Gini}_{\text{tourist}}$	Var_{ave}	Var_{max}	ED_r	Reward
FYP-Data 200 tourists	/	0.625	0.189	0.032	0.071	2265.351	5582
	co-attention	0.620	0.186	0.031	0.065	2379.293	5741
	self-attention	0.607	0.183	**0.029**	**0.063**	2350.328	6097
	cross-attention	**0.594**	**0.180**	0.030	0.068	**2418.873**	**6273**
FYP-Data 1000 tourists	/	0.674	0.216	0.045	0.096	5812.231	241847
	co-attention	0.652	0.210	**0.040**	0.092	5979.462	243145
	self-attention	0.619	**0.192**	0.044	0.089	**6423.507**	245803
	cross-attention	**0.606**	0.195	0.043	**0.087**	6389.954	**265225**

5.4.3.3 Comparison of Intention Fusion Models

In previous experiments, only cross-attention model is considered. We compare it with other two models: co-attention and self-attention based on the intention selector. k, θ and N are set as 1.5, 0, 3, respectively. Results are shown in Table 5.7. In 200 tourists scenario, self-attention and cross-attention shows comparable performance. In 1000 tourists scenario, cross-attention outperforms self-attention in terms of Gini_{POI} and Reward; in other metrics, two models' performance is similar. Generally three intention fusion models can bring gain in RPMTD performance; cross-attention outperforms self-attention in large-scale scenario; co-attention shows less improvement compared with other two.

5.4.3.4 Trick Validation

In our main experiments, tricks of implicit state update (ISU) and joint optimization (JO) are not implemented. We investigate its effect based on our intention selector and cross-attention. Results of ISU analysis are shown in Table 5.8. In scenario of 200 tourists, the implementation of the ISU does not enhance model performance, except for marginal improvements observed in $\text{Gini}_{tourist}$ and Reward. In the case of 1000 tourists, the implementation of ISU appears to detrimentally affect performance, with the exception of Var_{ave}. Overall, ISU cannot enhance the performance of our model. It is suspected that ISU is effective in sequential decision-making models, which requires that agents act in a specific order. Contrary to this, in our scenario, agents are required to act simultaneously, without any inherent sequential properties. Additionally, the non-stationary level in the sequence of decision making is different for each agent. Namely, the former agents have a greater impact on the state update, but at the same time they suffer from a higher

Table 5.8: Result of implicit state update (ISU) effect analysis based on intention selector.

	ISU	Gini_{POI}	$\text{Gini}_{tourist}$	Var_{ave}	Var_{max}	ED_r	Reward
FYP-Data	w/o	0.594	**0.180**	**0.030**	**0.068**	**2418.873**	6273
200 tourists	with	**0.591**	0.187	0.035	0.075	2404.158	**6292**
FYP-Data	w/o	**0.606**	**0.195**	0.043	**0.087**	**6389.954**	**265225**
1000 tourists	with	0.617	0.206	**0.041**	0.093	6021.741	240267

Table 5.9: Result of joint optimization (JO) effect analysis based on intention selector with FYP-Data 1000 tourists.

ω_1	ω_2	JO	Gini_{POI}	$\text{Gini}_{tourist}$	Var_{ave}	Var_{max}	ED_r	Reward
0.01	1	w/o	**0.714**	-	**0.075**	**0.178**	**6451.420**	-
0.01	1	with	0.735	-	0.081	0.203	6012.835	-
0.1	1	w/o	**0.606**	-	**0.043**	**0.087**	**6389.954**	-
0.1	1	with	0.674	-	0.051	0.105	6101.709	-
1	1	w/o	**0.604**	-	0.040	0.082	6437.424	-
1	1	with	0.610	-	**0.034**	**0.061**	**6914.726**	-

level of uncertainty. Due to the dynamic issue in our scenario, the decision order of the same agent is uncertain in different interactions. This leads to inconsistent level of stabilisation states perceived by the same agent, which could be worsen when the decision-making sequence is long. It could be the reason that model performance generally degrades in 1000 tourists scenario. What is more, parameter sharing is conducted in our all experiments. It could be difficult for the same model to learn from different levels of non-stationary state. Therefore, how to implement ISU in scenario like ours is a novel problem and could be future investigated.

We also investigate the joint optimization effect based on our intention-based model with different dual-congestion ratios. We fix the weight of local congestion ω_2 as 1 and set different weights of global congestion ω_1 in Equation 3.16. ω_1 default value is 0.1 in our main experiments and the weights are normalized in implementation. Different dual-congestion ratios represent different cooperation modes. As the ratio of dual-congestion directly influences the reward, the comparison of $\text{Gini}_{tourist}$ and Reward is not conducted. When ω_1=0.01, reward is dominated by local congestion. Agents primarily learn to maximise individual gain; The relation between agents

is almost competitive. Joint optimization does harm model performance. Individual agent could have conflicting goals. The use of joint optimisation may lead to reward inconsistency, i.e. optimisation by one agent may negatively affect other agents. When ω_1=0.1, reward is contributed by both local and global congestion. Agents learn to compromise individual interest with collective interests. The relation between agents is collaborative-adversarial (Hybrid). Similar with competitive one, joint optimization would decrease model performance. When ω_1=1, reward is dominated by global congestion. Agents are more likely to learn maximising collective interests. The relation between agents is almost cooperative. Joint optimization improves the Var_{ave}, Var_{max} and ED_r by 15%, 26% and 7%. This results validate the joint optimization effectiveness in cooperative mode. However, ratio of dual-congestion changes according to real practice. Considering route planning model primarily serves the tourists, improvement on collaborative-adversarial (Hybrid) model performance is necessary. Thus, we need study when to communicate and what information to share. Adaption of joint optimization in our scenario could be the further study.

5.5 Summary

In this section, we introduce a multi-agent communication protocol that considers agents' intention and information fusion to address the non-stationary problem inherent in dynamic environments. We first retrieve the most related agents from collective agents. Subsequently, an attention-based information fusion technique is implemented to utilize the relevant data effectively. This approach enhances the agents' ability to make informed decisions in complex scenarios. Two types of selector and three attention-based methods are implemented in this work and results show the effectiveness of this framework. We further investigate the widely used tricks against the non-stationary problem in our method. Limited improvement is observed. We conclude that study of agents acting in a specific order in our scenario and adaption of joint optimization in collaborative-adversarial mode could be our future work.

Chapter 6

From Multi-agent Communication to Multi-hop comprehension: An Effective Method for Answering Multi-hop Questions with Single-hop QA System

6.1 Introduction

The core of this section is *"apply Multi-agent communication to Multi-hop comprehension"*, which is inspired by the intuition that Multi-hop QA and multiple agents' communication could share the same framework. In multiple agents' communication, the irrelevant agents would be firstly denoised, and the single agent selectively communicate with the retrieved agents by attention-based information module. This could be aligned with the Multi-hop QA task, where the QA system could firstly denoise most of irrelevant information from a large amount of text and retrieve paragraphs most related to the query. Additionally, the communication between a single agent and other agents is similar to the interaction between the query and context in the QA system. Specifically, a single agent share its intention with other agents, and it obtains collective intention from other agents. In the QA system, the query interacts with the retrieved information, so that the two parties' information can be fused with each other. Figure 6.1 shows the similar frameworks of Multi-hop QA and multi-agent communication.

As a popular task in Natural Language Processing (NLP), much effort has been made to the development of question answering (QA) systems, due to the release of many large-scale and high-quality datasets such as [75–77]. Early on, these datasets mainly concentrate on single-hop questions, in which an answer can be retrieved from a single paragraph and only one fact is involved. With the recent explosion of success of deep learning techniques,

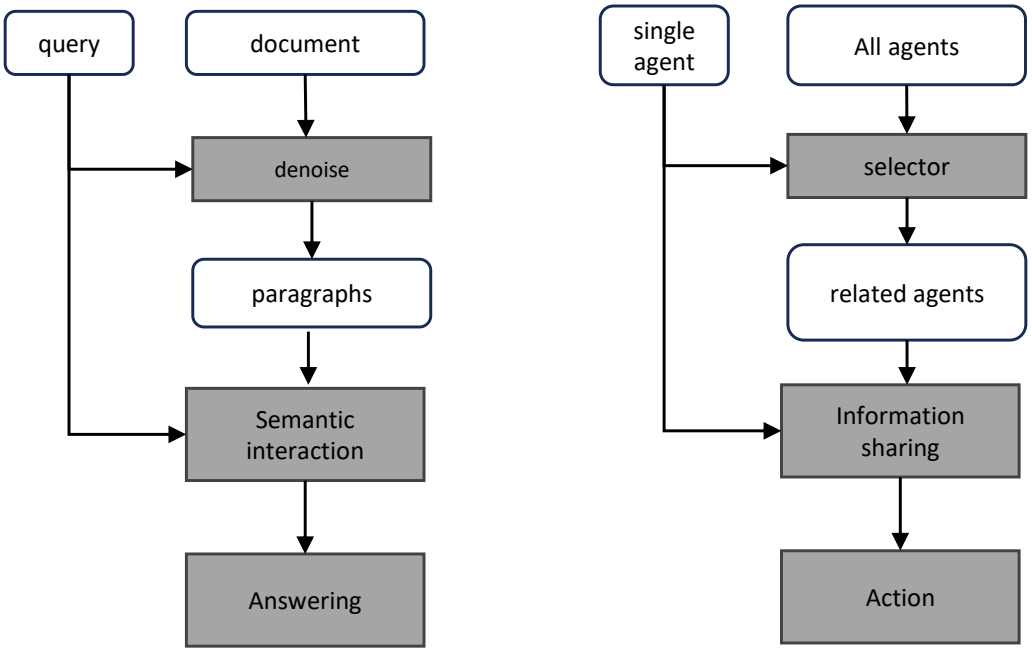


Figure 6.1: Framework comparison of Multi-hop QA (left) and multi-agent communication (right).

QA models such as [78,79] have correspondingly improved and have achieved super-human performance, especially in SQuAD 2.0. More recently, multi-hop QA datasets including [80–82] have gained increasing attention. These datasets require models to answer a more complicated question by integrating information from multiple paragraphs and facts.

Figure 6.2 shows an example from HotpotQA [82], which is a popular multi-hop QA dataset. Here, given a complex question and a document, the question is the composition of two single-hop sub-questions: (i) 'Who is the author of "Armad"?' (the answer is Ernest Cline) and (ii) 'Which novel by Ernest Cline will be adapted as a feature film by Steven Spielberg?'. The document contains 10 paragraphs but only two paragraphs are related to the question. Models are required to aggregate information from scattered facts across multiple paragraphs, and predict both the answer and supporting facts (i.e., sentences showing evidences of the answer).

Regarding the current research line, there has been a trend of exploiting graph neural network (GNN) for multi-hop QA [83–85]. Investigation of the graph construction and applying GNN reasoning has been explored. GNN-based models intuitively consider answering multi-hop questions as reasoning process on a document graph. Specifically, the document is first modeled

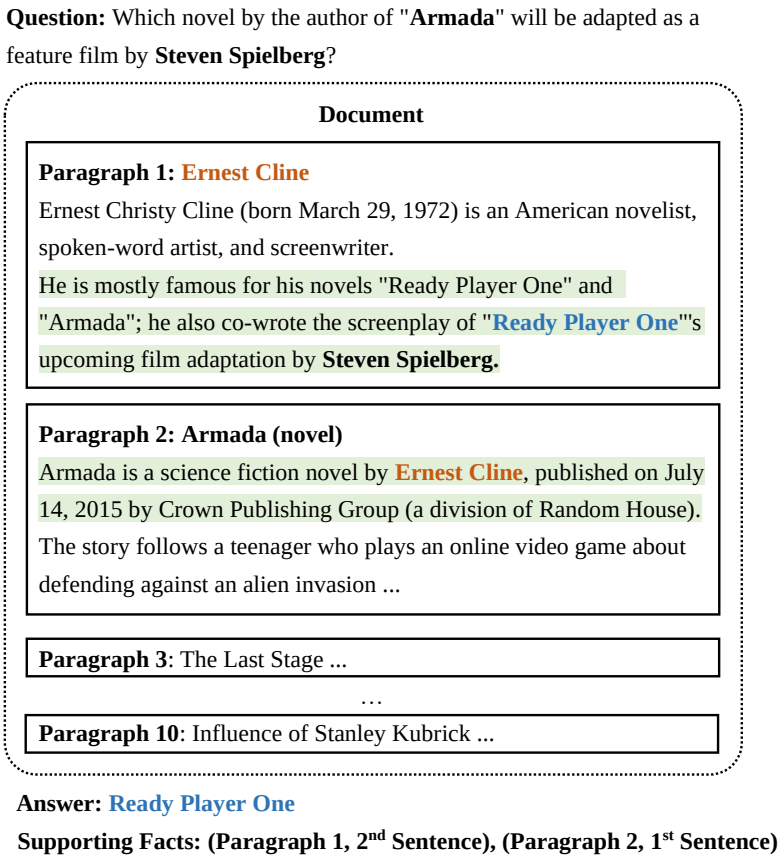


Figure 6.2: An example from HotpotQA. A document and A compositional question are given. Both the answer and supporting facts (in green background) should be predicted.

into a graph, and then GNN is applied for information propagation and aggregation. The updated graph state is expected to have the semantics of each node with its neighbors, which would be used for the final prediction. However, it has been studied that the computation of GNN is usually expensive and the graph construction strongly depends on prior knowledge [86].

Recently, document filters [83, 84, 87] are proposed to denoise any document by selecting the most relevant paragraphs inside it. Table 6.1 shows promising performance of the filter from Hierarchical Graph Network (HGN) [84]. For 2-paragraph selection, both precision and recall can achieve around 95%. For 4-paragraph selection, recall is nearly 99%. We observe that such performance can effectively neglect irrelevant information while keeping necessary evidences, making it acceptable to utilize single-hop QA model for

multi-hop QA.

Table 6.1: Performance of HGN's document filter.

Filter	Precision	Recall
2-paragraph selection	94.53	94.53
4-paragraph selection	49.45	98.74

Inspired by this, our work proposes an effective method to **A**nswer **M**ulti-hop questions by **S**ingle-hop QA system (**AMS**). We consider HGN [84], one of state-of-the-art (SOTA) models, with its document filter as our baseline. Our AMS exploits existing single-hop QA models based on the attention mechanism and integrates with the HGN's document filter. Since the prediction of supporting facts is also required, additional layers are incorporated for related sub-tasks to adapt multi-task learning. Besides, two-step tuning is proposed to enhance model's performance, which is based on transfer learning from other QA datasets. We conduct comprehensive experiments on five datasets to study how two-step tuning impacts on the model's performance. To validate our method, we focus on the HotpotQA dataset distractor setting [82]. The result shows that AMS can outperform the strong baseline model, and decrease both model's size and computational resource by around 80% and 23%, respectively. Moreover, AMS also outperforms other sophisticated GNN-based models.

To conclude, our contributions are threefolds. First, we propose an effective method (AMS) to answer multi-hop questions, which incorporates single-hop QA models with a document filter. Second, the proposed model outperforms the strong baseline and other sophisticated GNN-based models, while it requires less computational resource. Lastly, we propose a new two-step fine-tuning scheme that can improve the overall performance. We experimentally study its effectiveness with diverse datasets to analyze their effect on the model's performance.

6.2 Related Work

GNN-based Multi-hop QA GNN-based models attempt to construct a graph based on entities or other levels of granularity in text, which could bridge scattered information in different paragraphs. For instance, MHQA-GRN [88] integrates evidence by constructing an entity-based graph and investigates two GNNs to update graph state. Entity-GCN [89] refines entity-based graphs with different edges representing different relations. HDE-Graph [90] constructs a heterogeneous graphs by introducing the entity

and document nodes. CogQA [91] imitates human reasoning to construct a cognitive graph and predicts both possible answer spans and next-hop answer spans. DFGN [83] proposes a RoBERTa-based document filter to select the most relevant paragraphs and develops a dynamic entity-based graph interacting with context. SAE [87] improves the document filter by considering information between paragraphs. HGN [84] utilizes Wikipedia's hyperlinks to retrieve more paragraphs and proposes a hierarchical graph consisting of entity, sentence, paragraph and question nodes. BFR-Graph [85] constructs a weighted graph by relational information and poses restrictions on information propagation to improve the efficiency of graph reasoning.

No-GNN-based Multi-hop QA There are also attempts to address multi-hop QA by exploiting the existing NLP methods. For instance, Coref-GRU [92] extracts entities and their coreference from different paragraphs, and aggregates the information by using multi-GRU layers with a gated-attention reader. CFC [93] employs the hierarchical attention to construct the coarse and fine module for two-stage scoring. QFE [94] follows an extractive summarization work and incorporates an additional sentence prediction layer for multi-task learning. C2F Reader [95] considers the graph-attention as a special kind of self-attention, and argues that GNN may be unnecessary for multi-hop reasoning. Compared with the above methods, our work takes a step forward to effectively utilize existing single-hop QA models, and shows better performance than sophisticated GNN-based models.

Fine-tuning for NLP Tasks ULMFiT [96] proposes the discriminative fine-tuning that employs layer-wise learning rates, and slanted triangular learning rates with a sharp increase and a gradual decrease of the learning rates. [97] compare the performance of feature extraction and fine-tuning, and demonstrates that the distance between pre-training and the target task can impact on their relative performance. [98] explores a general scheme to fine-tune BERT for text classification, ranging from in-domain tuning, multi-task learning, to the fine-tuning in the target task. [99] proposes compact adapter modules for the text Transformer. Above works explore general fine-tunning schemes or study on a specific task. However, to the best of our knowledge, there is no work focusing on multi-hop QA.

6.3 Proposed Model

We select HGN [84], which is the SOTA approach for HotpotQA, as our strong baseline. Inspired from HGN, our model is the integration of its document filter and single-hop QA models. In our approach, the document is first denoised by the filter and then is fed into the attention-based single-hop QA model for the sub-tasks prediction and multi-task learning. Figure 6.3 shows an overview of our model.

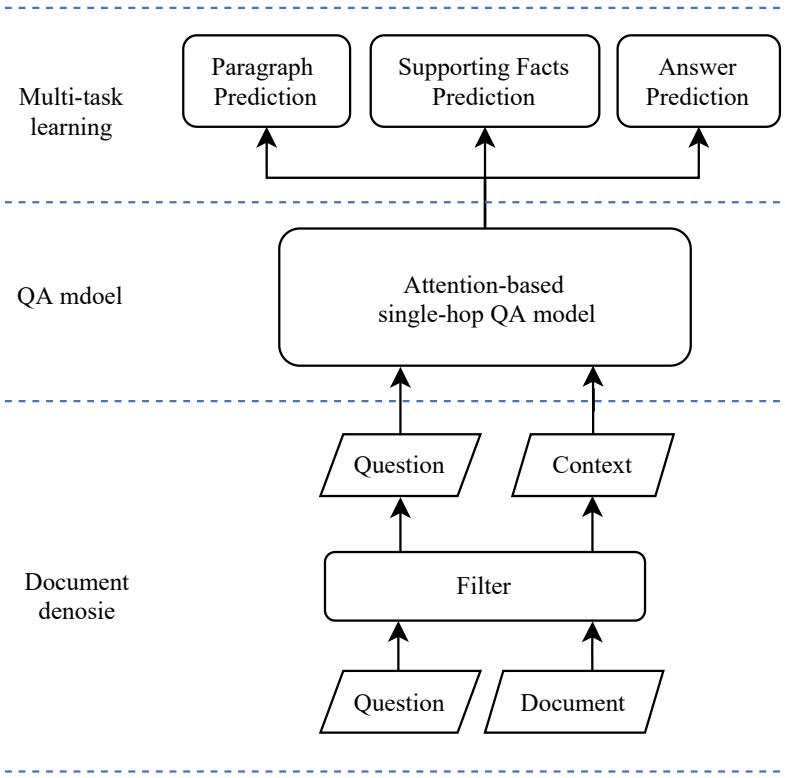


Figure 6.3: Overview of our model. Answer prediction includes answer span prediction and answer type prediction.

6.3.1 Document Denoise

The filter plays a crucial role in our work and we follow HGN's filter consisting of three components:

- Paragraph Ranker: It is trained based on RoBERTa and followed by a binary classification layer to calculate the probability of whether each paragraph contains supporting facts.

- Title Matching: It searches for paragraphs whose title exactly match any phrase with the question.
- Entity Matching: It searches for paragraphs which contain any entity exactly that appears in the question.

HGN's filter selects paragraphs within two steps. In the first step, it retrieves paragraphs by Title Matching. If multiple paragraphs are returned, two paragraphs yielding the highest score from Paragraph Ranker are selected. If it fails to retrieve any paragraphs, it further searches for paragraphs by Entity Matching. If it also fails, the paragraph yielding the highest score from the Paragraph Ranker is thus selected. In the second step, the filter retrieves additional paragraphs by Wikipedia's hyperlinks from the paragraphs identified by first step.

Table 6.1 show the performance of the adopted filter. According to the table, we select four paragraphs from the total ten paragraphs since it achieves high recall (98.74%). The retrieved paragraphs are concatenated and used as context. Figure 6.4 shows the distribution of token length of the context, indicating that around 94% token length is within 500. Such performance can effectively reduce the input length and keep necessary information. At this stage, the output is the question and context denoised from the filter:

$$\textit{Question}, \textit{Context} = \text{Filter}(\textit{Question}, \textit{Document}) \tag{6.1}$$

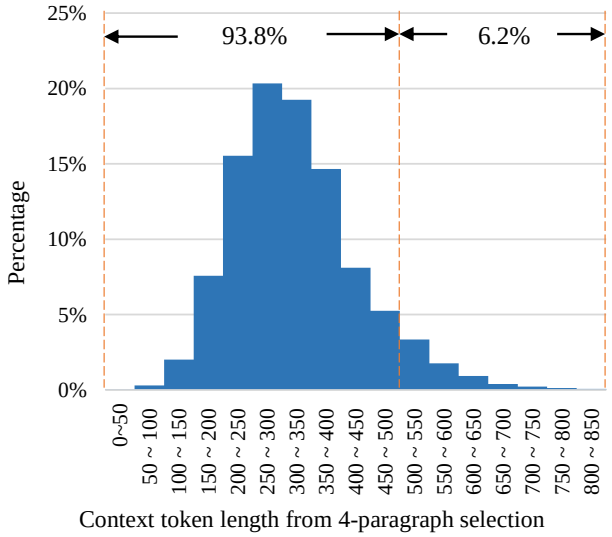


Figure 6.4: Distribution of context token length from 4-paragraph selection.

6.3.2 QA Model

With the promising performance of the document filter, we propose a single-hop QA model to eliminate the burden of GNN in the multi-hop QA task. Figure 6.5 illustrates the proposed single-hop QA model architecture, which performs the following steps.

First, it feeds the question and the context into the RoBERTa-large model to obtain question embeddings $\mathbf{E}_q \in \mathbb{R}^{l_q \times d}$ and context embedding $\mathbf{E}_c \in \mathbb{R}^{l_c \times d}$, where l_c and l_q are the length of context and question. d denotes the size of RoBERTa-large embedding.

After the representation of each context and question is extracted, the context embedding needs to be intensified by the question embedding. For this purpose, we apply the attention mechanism to learn the relationship between them. To show the generality of our single-hop QA model's effectiveness, we conduct experiments with two kinds of attention mechanisms: co-attention and self-attention. As a result, context can be updated by either of them:

$$\mathbf{C}' = \text{attention}(\mathbf{E}_q, \mathbf{E}_c) \in \mathbb{R}^{l_c \times h} \tag{6.2}$$

where h denotes the hidden dimension. The detail is explained in the subsequent sections.

6.3.2.1 Co-attention

Co-attention [74] is a vital model for single-hop QA. It enables the question and context to attend mutually, and also learns the question-aware context representation iteratively. We implement it as follows: Embedding $\mathbf{E}_c$ and $\mathbf{E}_q$ is mapped into a hidden dimension by two-layer feed-forward networks (FFNs[1]). Affinity matrix $\boldsymbol{A}$ is the product of context representation $\mathbf{C}$ and question representation $\mathbf{Q}$. In matrix $\boldsymbol{A}$, each value is the related score of one word from the question and one word from the context:

$$\mathbf{C} = \text{FFN}_c(\mathbf{E}_c) \in \mathbb{R}^{l_c \times h} \tag{6.3}$$

$$\mathbf{Q} = \text{FFN}_q(\mathbf{E}_q) \in \mathbb{R}^{l_q \times h} \tag{6.4}$$

$$\boldsymbol{A} = \mathbf{C}\mathbf{Q}^\top \in \mathbb{R}^{l_c \times l_q} \tag{6.5}$$

We normalize matrix $\boldsymbol{A}$ row-wise by softmax, so that each row indicates how much one word from the context is attended by all words from the question. By multiplying it with context representation $\mathbf{C}$, we can obtain

[1]All FFNs in this work includes two linear transformations with ReLU, Layer Normalization and Dropout in between.

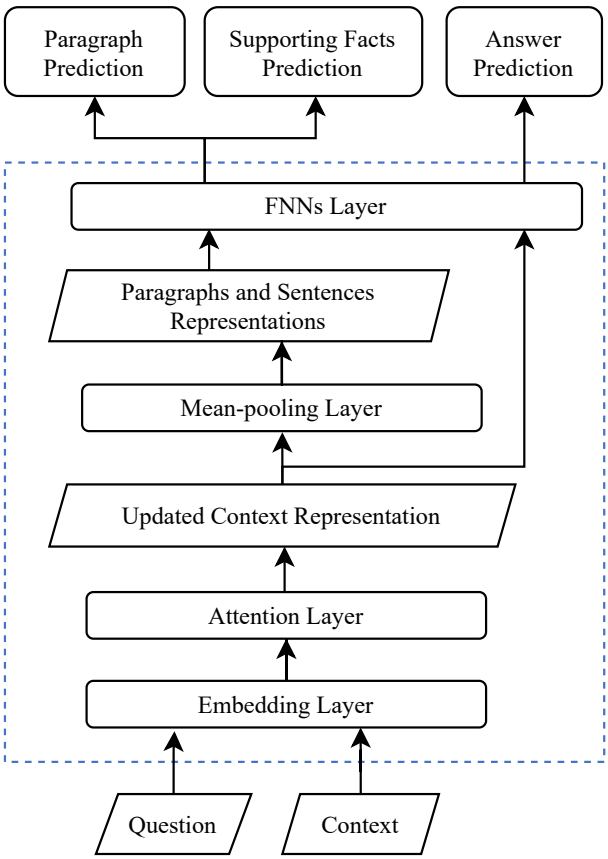


Figure 6.5: Architecture of proposed attention-based single-hop QA model.

the question representation $\mathbf{S}_q$ attended by the context. Similarly, we derive the context representation $\mathbf{S}_c$ attended by the question as follows:

$$\mathbf{S}_q = \text{softmax}(\boldsymbol{A}^\top) \times \mathbf{C} \in \mathbb{R}^{l_q \times h} \tag{6.6}$$

$$\mathbf{S}_c = \text{softmax}(\boldsymbol{A}) \times \mathbf{Q} \in \mathbb{R}^{l_c \times h} \tag{6.7}$$

where softmax$(\cdot)$ denotes the normalization column-wise and $\top$ denotes the matrix transpose.

Let the updated question $\mathbf{S}_q$ attend context again with the matrix $\boldsymbol{A}$. In addition, the attended context is further fed into a BiGRU as follows:

$$\mathbf{D}_c = \text{BiGRU}(\text{softmax}(\boldsymbol{A}) \times \mathbf{S}_q) \in \mathbb{R}^{l_c \times h} \tag{6.8}$$

$\mathbf{D}_c$ and $\mathbf{S}_c$ are context representations intensified by the question. Finally, they are concatenated and further applied with the FFN_d to transform into

the original document's length:

$$\mathbf{C}' = \text{FFN}_d([\mathbf{D}_c||\mathbf{S}_c]) \in \mathbb{R}^{l_c \times h} \tag{6.9}$$

where $[\cdot||\cdot]$ denotes the concatenation function.

6.3.2.2 Self-attention

We use a Transformer encoder [44] for defining self-attention, including a linear layer that maps the representation into the hidden dimension. It can capture relations between each pair of words from the query and the context. We set 8-head attention and stack two encoder layers to keep the model's size smaller than HGN.

$$\mathbf{C}' = \text{Self-attention}([\mathbf{E}_q||\mathbf{E}_c]) \in \mathbb{R}^{l_c \times h} \tag{6.10}$$

6.3.2.3 Prediction

After the attention module, updated context $\mathbf{C}'$ is sent to a mean-pooling layer to extract the representations of paragraphs and sentences:

$$\mathbf{P} = \text{Mean-pooling}(\mathbf{C}', \mathit{start}_p, \mathit{end}_p) \tag{6.11}$$

$$\mathbf{S} = \text{Mean-pooling}(\mathbf{C}', \mathit{start}_s, \mathit{end}_s) \tag{6.12}$$

where start_p and start_s denote the starting positions of each paragraph and each sentence, respectively. Similarly, end_p and end_s denote the ending positions.

Unlike the conventional single-hop QA, additional layers are employed for sub-tasks. the paragraphs' representation $\mathbf{P}$ is sent to a FFN for binary classification to calculate the probability that they contain supporting facts or not. Similarly, the sentences' representation $\mathbf{S}$ is sent to a FNN to calculate the probability that they are supporting facts or not.

$$\boldsymbol{o}_{para} = \text{FNN}_1(\mathbf{P}) \tag{6.13}$$

$$\boldsymbol{o}_{sent} = \text{FNN}_2(\mathbf{S}) \tag{6.14}$$

On the other hand, updated context $\mathbf{C}'$ is directly sent to other FFNs to predict the starting and ending positions of the answer span:

$$\boldsymbol{o}_{start} = \text{FNN}_4(\mathbf{C}') \tag{6.15}$$

$$\boldsymbol{o}_{end} = \text{FNN}_5(\mathbf{C}') \tag{6.16}$$

Since the answer type could be "yes", "no" or an answer span, 3-way classification is conducted. If the prediction is "yes" or "no", the answer

is directly returned. Otherwise, the answer span is returned. Similar with HGN, we also use the first hidden representation for answer type classification.

$$\boldsymbol{o}_{type} = \text{FNN}_6(\mathbf{C}'[0]) \tag{6.17}$$

6.3.3 Multi-task Learning

Finally, an answer type, an answer span with the starting and ending positions, gold paragraphs, and support facts are jointly predicted for multi-task learning. The cross-entropy loss is used for each task. Thus, the total loss ($\mathcal{L}_{total}$) is a weighted sum of each loss and each weight λ_i is our hyper-parameter:

$$\begin{aligned}\mathcal{L}_{total} =& \lambda_1\mathcal{L}_{type} + \lambda_2\mathcal{L}_{start} + \lambda_3\mathcal{L}_{end} \\ &+ \lambda_4\mathcal{L}_{para} + \lambda_5\mathcal{L}_{sent}\end{aligned} \tag{6.18}$$

6.4 Two-step Tuning

BERT-based language models [100, 101] are pre-trained on the large-scale corpora to learn universal semantics. But for a specific task, such as multi-hop QA, the data distribution can be different. More tuning on a related domain is expected to bring improvement as also investigated in [98, 99]. Therefore, we propose two-step tuning with an in-task distribution and a cross-task distribution for enhancing the model's performance. To study its effectiveness based on diverse datasets, we experiment with five datasets: SQuAD [102], NewsQA [103], TweetQA [104], CoLA [105], IMDB [106].

In-task Tuning: In this scenario, language model is first tuned in a QA dataset[2], including SQuAD, NewsQA and TweetQA[3]. Then, we use the tuned language model as an embedding in our proposed AMS and perform the second tuning in HotpotQA.

Cross-task Tuning: In this scenario, the first tuning dataset is not a QA dataset. Specifically, CoLA is a grammatical classification dataset and IMDB is a sentimental classification dataset. The second tuning process is the same as the in-task tunning.

[2]We only tune the the language model, instead of the entire model, in first tuning. It enables us to study its effectiveness from cross-task datasets.

[3]There is no annotated answer span in TweetQA. We retrieve the span with the best BLUE-1 score for training.

6.5 Experiment

6.5.1 Dataset

HotpotQA [82] is a popular multi-hop QA dataset, which is constructed from Wikipedia. There are two sub-datasets: the distractor setting and the fullwiki setting. For each case in the distractor setting, a compositional question and a document containing 10 paragraphs are given. In the document, only 2 paragraphs are related with the question and other 8 paragraphs are distractions. The gold paragraphs, supporting facts and ground-truth answers are annotated. The QA system is required to predict both an answer and supporting facts. In the fullwiki setting, the answer should be retrieved from the whole Wikipedia. In this work, we focus on the distractor setting. Official evaluation metrics are considered, i.e., EM (exact match) and the F1 score for the individual and joint evaluations of both the answer and supporting facts.

6.5.2 Experimental Setting

We conduct experiments based on a Quadro RTX 8000 GPU. We train the model for 8 epochs, and set learning rate as 1e-5 with batch size 8. For the hyper-parameters in our multi-task learning, we search λ_1, λ_2, λ_3 and λ_4 from {**1**,3,5} and λ_5 from {5, 10, 15, **20**}, in which the boldface indicates the best setting.

6.5.3 Experimental Result

6.5.3.1 Comparison with Baseline

We reproduce HGN with its source code and the result is based on RoBERTa-large. The upper part of Table 6.2 shows the comparison between our proposed AMS and HGN on the development set. According the table, the co-attention based model ($\text{AMS}_{\text{co-attention}}$) underperforms HGN within 1.0 point. The self-attention based model ($\text{AMS}_{\text{self-attention}}$) yields the better performance and especially outperforms HGN by 0.89 points for Joint EM.

6.5.3.2 Comparison based on Two-step Tuning

Table 6.3 shows the comparison between the original RoBERTa-large embedding and our two-step tuning embedding. This result is based on $\text{AMS}_{\text{co-attention}}$, demonstrating the following information:

Table 6.2: Comparison between HGN and AMS on dev set. The upper part is based on original RoBERTa-large embedding, which means the RoBERTa-large embedding from HuggingFace without two-step tuning. The lower part is based on SQuAD tuning embedding, which means two-step tuning based on SQuAD. 'Ans' indicates 'Answer' and 'Sup' indicates 'Supporting facts'. Δ = model's performance - HGN (reproduced) performance with original RoBERTa-large.

Embedding	Model	Ans		Sup		Joint	
		EM	F1	EM	F1	EM	F1
Original RoBERTa-large	HGN (reproduced)	68.33	82.04	62.89	88.53	45.78	74.06
	$AMS_{co\text{-}attention}$	67.85	81.55	63.28	87.7	46.35	73.58
	Δ	-0.48	-0.49	0.39	-0.83	0.57	-0.48
	$AMS_{self\text{-}attention}$	68.87	82.14	63.20	88.45	46.67	74.21
	Δ	0.54	0.10	0.31	-0.08	0.89	0.15
SQuAD tuning	HGN (reproduced)	69.14	82.55	63.24	88.82	46.75	74.75
	Δ	0.81	0.51	0.35	0.29	0.97	0.69
	$AMS_{co\text{-}attention}$	69.21	82.48	63.7	88.62	47.33	74.41
	Δ	0.88	0.44	0.81	0.09	1.55	0.35
	$AMS_{self\text{-}attention}$	69.26	82.51	64.4	88.63	47.56	74.62
	Δ	0.93	0.47	1.51	0.1	1.78	0.56

Table 6.3: Comparison between different embeddings.

Embedding	Ans		Sup		Joint	
	EM	F1	EM	F1	EM	F1
Original RoBERTa-large	67.85	81.55	63.28	87.7	46.35	73.58
SQuAD tuning	**69.21**	**82.48**	**63.7**	88.62	**47.33**	**74.41**
TweetQA tuning	67.87	81.79	63.52	88.62	46.84	73.93
NewsQA tuning	68.28	82.09	63.65	**88.77**	47.24	74.21
CoLA tuning	67.86	81.44	63.59	87.29	46.84	73.29
IMDB tuning	67.56	81.43	63.66	87.31	46.65	73.15

- In-task tuning can improve overall performance.
- SQuAD tuning yields the best improvement and TweetQA yields the smallest improvement. Potential reasons could be: (i) SQuAD and HotpotQA are all constructed from Wikipedia; thus, they may share the same resource and most relevant data distribution. (ii) TweetQA is more oral-style than other datasets. And the retrieved answer for training in TweetQA could be incomplete.
- Cross-task tuning can improve Sup EM but cannot benefit the answer prediction. We hypothesize that this is because supporting facts prediction is closely aligned with the classification task.

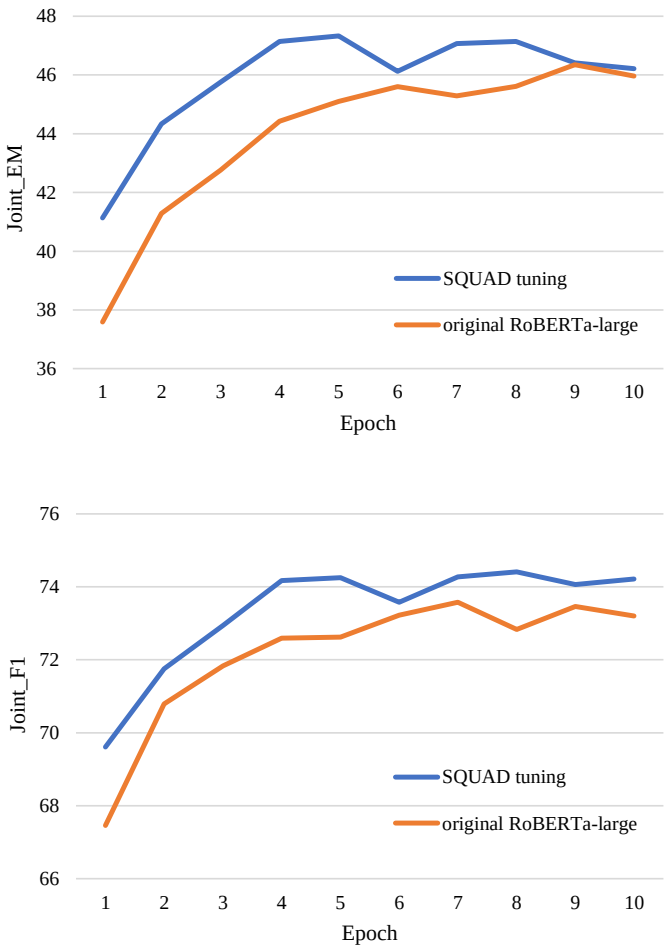


Figure 6.6: Comparison between original RoBERTa-large and SQuAD tuning on Joint EM (upper) and Joint F1 (lower).

The lower part of Table 6.2 illustrates that both HGN and AMS can be overall enhanced by SQuAD tuning (two-step tuning based on SQuAD). Compared with the reproduced HGN, AMS with SQuAD tuning can outperform it obviously in Sup EM and Joint EM. Furthermore, under the condition of both AMS and HGN using SQuAD tuning, their performances are quite competitive.

Figure 6.6 shows curve comparisons between the original RoBERTa-large and the SQuAD tuning based on Joint F1 (bottom) and Joint EM (top). From the figure, the SQuAD tuning curve is initially better than the original RoBERTa-large curve and it converges around 4^{th} epoch. This is faster than the original RoBERTa-large, showing the power of transfer learning in multi-hop reasoning.

6.5.3.3 Comparison with Related Work

We make comparisons with GNN-based models that use the BERT-based language model and the document filter. Table 6.4 shows the comparison result on the development set. According to the table, our proposed method outperforms GNN-based models with both BERT-base and RoBERTa-large, and $\text{AMS}_{\text{self-attention}}$ yields the best performance.

6.5.4 Comparison of Model's size and Computational Resource

Table 6.5 shows the comparison of the model's size, computational resource and performance. The result is based on RoBERTa-large. $\text{AMS}_{\text{co-attention}}$ model's size is only about 20% of HGN and $\text{AMS}_{\text{self-attention}}$ model's size is

Table 6.4: Comparison with related work on dev set. AMS result is based on SQuAD tuning and HGN result is without SQuAD tuning.

Embedding	Model	Ans		Sup		Joint	
		EM	F1	EM	F1	EM	F1
Bert-base	DFGN	55.66	69.34	53.10	82.24	33.68	59.86
	HGN	60.23	74.49	56.62	85.91	38.16	66.20
	$\text{AMS}_{\text{co-attention}}$	61.39	75.39	58.78	**85.93**	40.04	67.03
	$\text{AMS}_{\text{self-attention}}$	**62.11**	**75.76**	**59.20**	85.78	**40.73**	**67.39**
RoBERTa-large	SAE	67.70	80.75	63.30	87.38	46.81	72.75
	HGN	68.33	82.04	62.89	88.53	45.78	74.06
	$\text{AMS}_{\text{co-attention}}$	69.21	82.48	63.70	88.22	47.33	74.41
	$\text{AMS}_{\text{self-attention}}$	**69.26**	**82.51**	**64.40**	**88.63**	**47.56**	**74.62**

Table 6.5: Comparison of model's size, computational resource and performance.

	Baseline	Proposed model	
	HGN	$\text{AMS}_{\text{co-attention}}$	$\text{AMS}_{\text{self-attention}}$
Model's size	31.61M	**6.30M**	30.83M
RoBERTa-large	355M	355M	355M
Training time	191 min	**148 min**	160 min
Joint EM	45.78	47.33	**47.56**
Joint F1	74.06	74.41	**74.62**

close to HGN. For computational resource, $\text{AMS}_{\text{co-attention}}$ and $\text{AMS}_{\text{self-attention}}$ is 77.5% and 83.8% of HGN, respectively. Since RoBERTa-large (355M) dominates the total model's size, training time is not reduced significantly. The computational resource is expected to further decrease by incorporating a lighter language model. Generally, both proposed models show better performance and use less computational resource.

6.5.5 Error Analysis

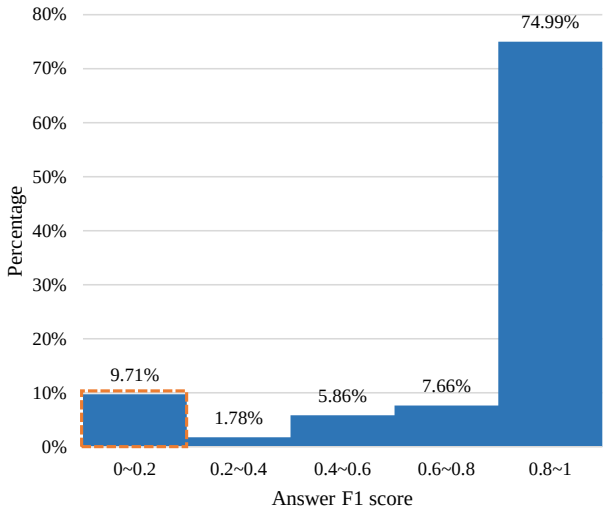


Figure 6.7: Answer F1 score distribution on dev set. There are almost 10% answer F1 score less than 0.2

We analyse the answer F1 score on the development set. Figure 6.7 illustrates its distribution. Almost 10% of the answer F1 score is less than 0.2,

in which 9.7% answer F1 score is 0. Further improvement can be considered from this error. Similar with HGN, we randomly sample 100 examples with answer F1 score as 0 and they are categorized as follows:

- **Multi-answer** (12%): There are multiple gold answers and the predicted answer is different from the annotation. For example, the annotation is 'National Broadcasting Company' and the predicted answer is 'NBC'.
- **Multi-hop** (28%): The supporting facts prediction is incorrect, from which the model fails to predict the right answer. For example, the supporting facts are the 1st and the 2nd sentences, but the model predicts the 3rd and the 4th sentences as supporting facts and retrieves answer from them. Accordingly, the answer prediction is incorrect.
- **MRC** (38%): The supporting facts' prediction is right but the answer prediction is wrong. For example, the supporting facts are the 1st and the 2nd sentences. The model predicts them correctly. But the final answer prediction is wrong.
- **Comparison** (22%): The model fails to do numerical operations that involves information aggregation. For example, the question is ' The CEO of Walmart and the CEO of Apple, who is older?'

Multi-hop and MRC account for more than 50%, which indicates that the performance could be further improved by more advanced QA models.

Another tricky error is that there are 1,322 cases, about 17% of the

Table 6.6: Some examples that supporting facts F1 is 0 but answer F1 is 1.

ID	Answer	Supporting Facts	Predicted Answer	Predicted Supporting Facts
5ae180195542 9901ffe4aec4	Battle Creek, Michigan	[['Adventures of Superman (TV series), 2], [' Kellogg's', 0], ['Kellogg's', 2]]	Battle Creek, Michigan	[['Cocoa Krispies', 0],['Adventures of Superman (TV series)', 0]]
5ae1fa2b5542 997f29b3c1df	Eminem	[['Mack 10 discography', 2], ['Numb (Rihanna song)', 0]]	Eminem	[['The Monster (song)', 0], ['Numb (Rihanna song)', 1]]
5ae18d615542 997283cd2229	mixed martial arts	[['Liz McCarthy (fighter)', 0], ['Atomweight (MMA)', 0]]	mixed martial arts	[['Atomweight', 0], ['Amber Brown (fighter)', 0]]

development set, that supporting fact F1 is 0 but answer F1 is 1. This means that the supporting facts prediction is wrong but the answer prediction is right. Table 6.6 shows some examples of this case. Such interpretable problem may occur when the answer is not directly retrieved from predicted sentences. It could be further studied by considering supporting facts prediction's restrictions for the answer prediction.

6.6 Summary

In this section, we propose a simple yet effective model, called AMS, for multi-hop QA. AMS is the integration of HGN's document filter and single-hop QA models, which shares the similar framework with our MARL communication framework. We also introduce a new fine-tuning scheme for improving its performance. The result shows that AMS can outperform the strong baseline HGN with less amount of computational resource. Furthermore, AMS can achieve the better performance than other sophisticated GNN-based models. In contrast to GNN-based methods, our method can effectively leverage existing single-hop QA models and does not require any auxiliary tool, such as NER, which should gain more attention in the further research.

Chapter 7

Conclusion

This research aims to solve the overtourism and under-tourism caused by biased-sightseeing problem in route planning for multiple tourists. Our work undertakes a exploration of MARL and communication protocol, addressing critical challenges through innovative frameworks that reflect an integrated approach to problem-solving.

We begin with introducing the popularity-biased route planning problem, resolved through our MARL framework incorporating a dual congestion-aware model. Our model RPMTD evaluates both the crowdedness of visited POIs and the distribution of tourists, leveraging novel mobility data to create a realistic interaction environment. The experimental results underscore the model's superior performance and robustness in managing tourist distribution, highlighting its practical potential for urban tourism management. Notably, the user study reveals that our model transforms non-cooperative relationships between users and tourism into cooperative ones, marking a significant advancement towards sustainable tourism that balances tourists' preferences with broader environmental and social concerns.

Next, we propose multi-agent communication protocols which aims to mitigate the non-stationary problem, employing methods to denoise irrelevant information and perform information fusion. By implementing two types of selectors and three attention-based methods, we demonstrate the framework's effectiveness. However, experiments reveal limited improvements from widely used tricks against the non-stationary problem. Future research will focus on the sequential actions of agents and the adaptation of joint optimization in collaborative-adversarial scenarios.

Finally, we extend our communication framework to multi-hop QA and propose the AMS model for this task, which integrates HGN's document filter with single-hop QA models and introduces a novel fine-tuning scheme. Our findings demonstrate that AMS not only outperforms the strong baseline HGN with fewer computational resources but also surpasses other sophisticated GNN-based models. This approach focuses on existing single-hop QA models without auxiliary tools like NER, suggesting significant potential for further research.

We also points out several limitations of our research: (i) this work majorly considers the interests of POIs, and the interests of tourists such as must-see POIs are not considered. The tourism industry should satisfy tourists and balance two parts in reality. Cooperating with other tourist-oriented work could be further work; (ii) we only planned maximum 1000 tourists in our scenario. However, in reality, there are far more than this amount of tourists. Development of a practical method for real number of users is necessary; (iii) in real tourist route planning, the randomness of tourists exists. Modeling these randomness will be a step forward to practical and real tourist route planning.

In summary, this study advocates a comprehensive approach to addressing the real-world biased-sightseeing problem utilizing MARL. By incorporating innovative models with practical, real-world mobility data and emphasizing both cooperation and competition among agents, this research establishes a foundational framework for MARL studies focused on route planning for multiple tourists, fostering a more adaptive, efficient, and scalable system. This work marks a development of sustainable and efficient tourist route planning methodologies, which signifies a considerable advancement toward achieving sustainable tourism that balances the preferences of tourists with broader environmental and social considerations.

References

[1] Tourism Economics, "WTM Global Travel Report," 2024, accessed: May 16, 2024. [Online]. Available: https://www.wtm.com/content/dam/sitebuilder/rxuk/wtmkt/documents/WTM-Global-Travel-Report-v4.pdf.coredownload.990096961.pdf

[2] M. Deudon, P. Cournut, A. Lacoste, Y. Adulyasak, and L.-M. Rousseau, "Learning heuristics for the tsp by policy gradient," in *Integration of Constraint Programming, Artificial Intelligence, and Operations Research: 15th International Conference, CPAIOR 2018, Delft, The Netherlands, June 26–29, 2018, Proceedings 15*. Springer, 2018, pp. 170–181.

[3] I. Bello, H. Pham, Q. V. Le, M. Norouzi, and S. Bengio, "Neural combinatorial optimization with reinforcement learning," *arXiv preprint arXiv:1611.09940*, 2016.

[4] R. Gama and H. Fernandes, "A reinforcement learning approach to the orienteering problem with time windows," *Computers & Operations Research*, vol. 133, p. 105357, 2021.

[5] Z. Duan, Y. Gao, J. Feng, X. Zhang, and J. Wang, "Personalized tourism route recommendation based on user's active interests," in *2020 21st IEEE International Conference on Mobile Data Management (MDM)*. IEEE, 2020, pp. 729–734.

[6] L. Chen, L. Zhang, S. Cao, Z. Wu, and J. Cao, "Personalized itinerary recommendation: Deep and collaborative learning with textual information," *Expert Systems with Applications*, vol. 144, p. 113070, 2020.

[7] P. Padia, K. H. Lim, J. Cha, and A. Harwood, "Sentiment-aware and personalized tour recommendation," in *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 2019, pp. 900–909.

[8] K. H. Lim, J. Chan, S. Karunasekera, and C. Leckie, "Personalized itinerary recommendation with queuing time awareness," in *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, 2017, pp. 325–334.

[9] W. K. Kong, S. Zheng, M. L. Nguyen, and Q. Ma, "Diversity-oriented route planning for tourists," in *International Conference on Database and Expert Systems Applications*. Springer, 2022, pp. 243–255.

[10] K. Sylejmani, J. Dorn, and N. Musliu, "Planning the trip itinerary for tourist groups," *Information Technology & Tourism*, vol. 17, pp. 275–314, 2017.

[11] J. L. Sarkar and A. Majumder, "gtour: Multiple itinerary recommendation engine for group of tourists," *Expert Systems with Applications*, vol. 191, p. 116190, 2022.

[12] K. H. Lim, J. Chan, C. Leckie, and S. Karunasekera, "Towards next generation touring: Personalized group tours," in *Proceedings of the International Conference on Automated Planning and Scheduling*, vol. 26, 2016, pp. 412–420.

[13] S. Elmi and K.-L. Tan, "Next pois prediction for group recommendations: Influence-based deep learning model," in *International Conference on Database and Expert Systems Applications*. Springer, 2023, pp. 295–309.

[14] S. Barač-Miftarević, "Undertourism vs. overtourism: A systematic literature review," *Tourism: An International Interdisciplinary Journal*, vol. 71, no. 1, pp. 178–192, 2023.

[15] C. Milano and K. Koens, "The paradox of tourism extremes. excesses and restraints in times of covid-19," *Current Issues in Tourism*, vol. 25, no. 2, pp. 219–231, 2022.

[16] M. Blázquez-Salom, M. Cladera, and M. Sard, "Identifying the sustainability indicators of overtourism and undertourism in majorca," *Journal of Sustainable Tourism*, vol. 31, no. 7, pp. 1694–1718, 2023.

[17] G.-J. Hospers, "Overtourism in european cities: From challenges to coping strategies," in *CESifo Forum*, vol. 20, no. 03. München: ifo Institut–Leibniz-Institut für Wirtschaftsforschung an der . . . , 2019, pp. 20–24.

[18] A. Dhiraj and S. Kumar, "Overtourism: Causes, impacts and solution," in *Overtourism as Destination Risk: Impacts and Solutions*. Emerald Publishing Limited, 2021, pp. 49–56.

[19] Japan National Tourism Organization, "Sustainable: Travel Japan," https://www.japan.travel/en/sustainable/, accessed: March 28, 2024.

[20] European Commission, "An EU initiative to reward innovative and smart tourism in European cities!" https://smart-tourism-capital.ec.europa.eu/index_en, accessed: March 28, 2024.

[21] Alliance France Tourisme, "Destination France : Quelle régulation face à la surfréquentation touristique," https://www.alliance-france-tourisme.fr/posts/destination-france-quelle-regulation-face-a-la-surfrequentation-touristique, accessed: March 28, 2024.

[22] Ministry of Business, Innovation and Employment, "New Zealand-Aotearoa Government Tourism Strategy," https://www.mbie.govt.nz/dmsdocument/5482-2019-new-zealand-aotearoa-government-tourism-strategy-pdf, accessed: March 27, 2024.

[23] H. Hsieh, C. Li, and S. Lin, "Triprec: recommending trip routes from large scale check-in data," in *WWW*, 2012, pp. 529–530.

[24] A. Gunawan, Z. Yuan, and H. Lau, "A mathematical model and metaheuristics for time dependent orienteering problem," in *PATAT*, 2014.

[25] A. Gionis, T. Lappas, K. Pelechrinis, and E. Terzi, "Customized tour recommendations in urban areas," in *ACM WSDM*, 2014, pp. 313–322.

[26] K. Taylor, K. Lim, and J. Chan, "Travel itinerary recommendations with must-see points-of-interest," in *Companion Proceedings of The The Web Conference 2018*, 2018, pp. 1198–1205.

[27] G. Xu, B. Fu, and Y. Gu, "Point-of-interest recommendations via a supervised random walk algorithm," *IEEE Intelligent Systems*, vol. 31, pp. 15–23, 2016.

[28] Y. Seo and Y. Cho, "Point of interest recommendations based on the anchoring effect in location-based social network services," *Expert Systems with Applications*, vol. 164, p. 114018, 2021.

[29] T. Qian, B. Liu, Q. Nguyen, and H. Yin, "Spatiotemporal representation learning for translation-based poi recommendation," *ACM Transactions On Information Systems (TOIS)*, vol. 37, pp. 1–24, 2019.

[30] D. Yu, W. Wanyan, and D. Wang, "Leveraging contextual influence and user preferences for point-of-interest recommendation," *Multimedia Tools and Applications*, vol. 80, pp. 1487–1501, 2021.

[31] Z. Zhang, C. Zou, R. Ding, and Z. Chen, "Vcg: Exploiting visual contents and geographical influence for point-of-interest recommendation," *Neurocomputing*, vol. 357, pp. 53–65, 2019.

[32] R. Gao, J. Li, X. Li, C. Song, J. Chang, D. Liu, and C. Wang, "Stscr: Exploring spatial-temporal sequential influence and social information for location recommendation," *Neurocomputing*, vol. 319, pp. 118–133, 2018.

[33] C. Cheng, H. Yang, I. King, and M. Lyu, "Fused matrix factorization with geographical and social influence in location-based social networks," in *AAAI*, vol. 26, 2012, pp. 17–23.

[34] B. Liu and H. Xiong, "Point-of-interest recommendation in location based social networks with topic and location awareness," in *SDM 2013*, 2013, pp. 396–404.

[35] S. Moritake, H. Kasahara, and Q. Ma, "Merihari-area tour planning by considering regional characteristics," in *International Conference On Database And Expert Systems Applications*, 2023, pp. 49–64.

[36] K. Wang, X. Wang, and X. Lu, "Poi recommendation method using lstm-attention in lbsn considering privacy protection," *Complex & Intelligent Systems*, vol. 9, pp. 2801–2812, 2023.

[37] J. Sun, C. Zhuang, and Q. Ma, "User transition pattern analysis for travel route recommendation," *IEICE TRANSACTIONS on Information and Systems*, vol. 102, pp. 2472–2484, 2019.

[38] J. Lee, J. Won, and J. Lee, "Crowd simulation by deep reinforcement learning," in *ACM SIGGRAPH*, 2018, pp. 1–7.

[39] A. Normoyle, M. Likhachev, and A. Safonova, "Stochastic activity authoring with direct user control," in *Proceedings of the 18th Meeting of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, 2014, pp. 31–38.

[40] K. Li, L. Chen, and S. Shang, "Towards alleviating traffic congestion: optimal route planning for massive-scale trips," in *IJCAI*, 2021, pp. 3400–3406.

[41] B. Sridhar, S. Grabbe, and A. Mukherjee, "Modeling and optimization in traffic flow management," *Proceedings of the IEEE*, vol. 96, pp. 2060–2080, 2008.

[42] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *ArXiv Preprint ArXiv:1707.06347*, 2017.

[43] N. Mazyavkina, S. Sviridov, S. Ivanov, and E. Burnaev, "Reinforcement learning for combinatorial optimization: A survey," *Computers & Operations Research*, vol. 134, p. 105400, 2021.

[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Kaiser, and I. Polosukhin, "Attention is all you need," vol. 30, 2017.

[45] K. government, "Kyoto city economy 2022," https://www.city.kyoto.lg.jp/sankan/page/0000310040.html, 2023.

[46] L. Pappalardo, F. Simini, G. Barlacchi, and R. Pellungrini, "scikit-mobility: A python library for the analysis, generation, and risk assessment of mobility data," *ArXiv Preprint ArXiv:1907.07062*, 2019.

[47] I. Kostrikov, A. Nair, and S. Levine, "Offline reinforcement learning with implicit q-learning," *ArXiv Preprint ArXiv:2110.06169*, 2021.

[48] R. Lowe, Y. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Advances In Neural Information Processing Systems*, vol. 30, 2017.

[49] C. Yu, A. Velu, E. Vinitsky, J. Gao, Y. Wang, A. Bayen, and Y. Wu, "The surprising effectiveness of ppo in cooperative multi-agent games," in *Advances In Neural Information Processing Systems*, vol. 35, 2022, pp. 24 611–24 624.

[50] Y. Zhu, T. Ko, D. Snyder, B. Mak, and D. Povey, "Self-attentive speaker embeddings for text-independent speaker verification." in *Interspeech*, vol. 2018, 2018, pp. 3573–3577.

[51] L. Ceriani and P. Verme, "The origins of the gini index: extracts from variabilità e mutabilità (1912) by corrado gini," *Journal of Economic Inequality*, vol. 10, pp. 421–443, 2012.

[52] K. Yi, X. Jin, Z. Bai, Y. Kong, and Q. Ma, "An empirical user study on congestion-aware route recommendation," in *Proc. of ENTER e-Tourism Conference 2024(to appear)*, 2024.

[53] M. Zarour and M. Alharbi, "User experience aspects and dimensions: systematic literature review," *International Journal Of Knowledge Engineering*, vol. 3, pp. 52–59, 2017.

[54] W. Wattanacharoensil and D. La-ornual, "A systematic review of cognitive biases in tourist decisions," *Tourism Management*, vol. 75, pp. 353–369, 2019.

[55] A. Capocchi, C. Vallone, M. Pierotti, and A. Amaduzzi, "Overtourism: A literature review to assess implications and future perspectives," *Sustainability*, vol. 11, p. 3303, 2019.

[56] M. Żemła, "Reasons and consequences of overtourism in contemporary cities—knowledge gaps and future research," *Sustainability*, vol. 12, p. 1729, 2020.

[57] R. Mehrotra, J. McInerney, H. Bouchard, M. Lalmas, and F. Diaz, "Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2018, pp. 2243–2251.

[58] R. Merton, "The matthew effect in science: The reward and communication systems of science are considered," *Science*, vol. 159, pp. 56–63, 1968.

[59] F. Muros, "Cooperative game theory tools in coalitional control networks," in *Springer*, 2019, pp. 9–11.

[60] E. Ostrom, "Collective action and the evolution of social norms," *Journal of Economic Perspectives*, vol. 14, pp. 137–158, 2000.

[61] P. Sunehag, G. Lever, A. Gruslys, W. M. Czarnecki, V. Zambaldi, M. Jaderberg, M. Lanctot, N. Sonnerat, J. Z. Leibo, K. Tuyls *et al.*, "Value-decomposition networks for cooperative multi-agent learning," *arXiv preprint arXiv:1706.05296*, 2017.

[62] T. Rashid, M. Samvelyan, C. S. De Witt, G. Farquhar, J. Foerster, and S. Whiteson, "Monotonic value function factorisation for deep multi-agent reinforcement learning," *Journal of Machine Learning Research*, vol. 21, no. 178, pp. 1–51, 2020.

[63] J. Foerster, I. A. Assael, N. De Freitas, and S. Whiteson, "Learning to communicate with deep multi-agent reinforcement learning," *Advances in neural information processing systems*, vol. 29, 2016.

[64] S. Sukhbaatar, R. Fergus *et al.*, "Learning multiagent communication with backpropagation," *Advances in neural information processing systems*, vol. 29, 2016.

[65] A. Singh, T. Jain, and S. Sukhbaatar, "Learning when to communicate at scale in multiagent cooperative and competitive tasks," *arXiv preprint arXiv:1812.09755*, 2018.

[66] P. Peng, Y. Wen, Y. Yang, Q. Yuan, Z. Tang, H. Long, and J. Wang, "Multiagent bidirectionally-coordinated nets: Emergence of human-level coordination in learning to play starcraft combat games," *arXiv preprint arXiv:1703.10069*, 2017.

[67] J. Jiang and Z. Lu, "Learning attentional communication for multi-agent cooperation," *Advances in neural information processing systems*, vol. 31, 2018.

[68] S. Q. Zhang, Q. Zhang, and J. Lin, "Succinct and robust multi-agent communication with temporal message control," *Advances in neural information processing systems*, vol. 33, pp. 17 271–17 282, 2020.

[69] A. Das, T. Gervet, J. Romoff, D. Batra, D. Parikh, M. Rabbat, and J. Pineau, "Tarmac: Targeted multi-agent communication," in *International Conference on machine learning*. PMLR, 2019, pp. 1538–1546.

[70] S. Qi and S.-C. Zhu, "Intent-aware multi-agent reinforcement learning," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 7533–7540.

[71] Y. Fang, Z. Tang, K. Ren, W. Liu, L. Zhao, J. Bian, D. Li, W. Zhang, Y. Yu, and T.-Y. Liu, "Learning multi-agent intention-aware communication for optimal multi-order execution in finance," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023, pp. 4003–4012.

[72] X. Wu, R. Chandra, T. Guan, A. Bedi, and D. Manocha, "Intent-aware planning in heterogeneous traffic via distributed multi-agent reinforcement learning," in *Conference on Robot Learning*. PMLR, 2023, pp. 446–477.

[73] W. Kim, J. Park, and Y. Sung, "Communication in multi-agent reinforcement learning: Intention sharing," in *International Conference on Learning Representations*, 2020.

[74] C. Xiong, V. Zhong, and R. Socher, "Dynamic coattention networks for question answering," *arXiv preprint arXiv:1611.01604*, 2016.

[75] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, "Teaching machines to read and comprehend," *Advances in neural information processing systems*, vol. 28, pp. 1693–1701, 2015.

[76] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for squad," *arXiv preprint arXiv:1806.03822*, 2018.

[77] M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer, "Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension," *arXiv preprint arXiv:1705.03551*, 2017.

[78] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

[79] Z. Zhang, J. Yang, and H. Zhao, "Retrospective reader for machine reading comprehension," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 16, 2021, pp. 14 506–14 514.

[80] D. Khashabi, S. Chaturvedi, M. Roth, S. Upadhyay, and D. Roth, "Looking beyond the surface: A challenge set for reading comprehension over multiple sentences," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 252–262.

[81] J. Welbl, P. Stenetorp, and S. Riedel, "Constructing datasets for multi-hop reading comprehension across documents," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 287–302, 2018.

[82] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, and C. D. Manning, "HotpotQA: A dataset for diverse, explainable multi-hop question answering," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels,

Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 2369–2380. [Online]. Available: https://aclanthology.org/D18-1259

[83] L. Qiu, Y. Xiao, Y. Qu, H. Zhou, L. Li, W. Zhang, and Y. Yu, "Dynamically fused graph network for multi-hop reasoning," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 6140–6150.

[84] Y. Fang, S. Sun, Z. Gan, R. Pillai, S. Wang, and J. Liu, "Hierarchical graph network for multi-hop question answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 8823–8838.

[85] Y. Huang and M. Yang, "Breadth first reasoning graph for multi-hop question answering," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 5810–5821.

[86] L. Wu, Y. Chen, K. Shen, X. Guo, H. Gao, S. Li, J. Pei, and B. Long, "Graph neural networks for natural language processing: A survey," *arXiv preprint arXiv:2106.06090*, 2021.

[87] M. Tu, K. Huang, G. Wang, J. Huang, X. He, and B. Zhou, "Select, answer and explain: Interpretable multi-hop reading comprehension over multiple documents," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 9073–9080.

[88] L. Song, Z. Wang, M. Yu, Y. Zhang, R. Florian, and D. Gildea, "Exploring graph-structured passage representation for multi-hop reading comprehension with graph neural networks," *arXiv preprint arXiv:1809.02040*, 2018.

[89] N. De Cao, W. Aziz, and I. Titov, "Question answering by reasoning across documents with graph convolutional networks," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 2306–2317.

[90] M. Tu, G. Wang, J. Huang, Y. Tang, X. He, and B. Zhou, "Multi-hop reading comprehension across multiple documents by reasoning over heterogeneous graphs," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2704–2713.

[91] M. Ding, C. Zhou, Q. Chen, H. Yang, and J. Tang, "Cognitive graph for multi-hop reading comprehension at scale," *arXiv preprint arXiv:1905.05460*, 2019.

[92] B. Dhingra, Q. Jin, Z. Yang, W. W. Cohen, and R. Salakhutdinov, "Neural models for reasoning over multiple mentions using coreference," *arXiv preprint arXiv:1804.05922*, 2018.

[93] V. Zhong, C. Xiong, N. S. Keskar, and R. Socher, "Coarse-grain fine-grain coattention network for multi-evidence question answering," *arXiv preprint arXiv:1901.00603*, 2019.

[94] K. Nishida, K. Nishida, M. Nagata, A. Otsuka, I. Saito, H. Asano, and J. Tomita, "Answering while summarizing: Multi-task learning for multi-hop qa with evidence extraction," *arXiv preprint arXiv:1905.08511*, 2019.

[95] N. Shao, Y. Cui, T. Liu, S. Wang, and G. Hu, "Is graph structure necessary for multi-hop question answering?" in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 7187–7192.

[96] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 328–339. [Online]. Available: https://aclanthology.org/P18-1031

[97] M. E. Peters, S. Ruder, and N. A. Smith, "To tune or not to tune? adapting pretrained representations to diverse tasks," *arXiv preprint arXiv:1903.05987*, 2019.

[98] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune bert for text classification?" in *China National Conference on Chinese Computational Linguistics*. Springer, 2019, pp. 194–206.

[99] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2790–2799.

[100] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL*, 2019.

[101] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[102] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for SQuAD," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 784–789. [Online]. Available: https://aclanthology.org/P18-2124

[103] A. Trischler, T. Wang, X. Yuan, J. Harris, A. Sordoni, P. Bachman, and K. Suleman, "Newsqa: A machine comprehension dataset," in *Proceedings of the 2nd Workshop on Representation Learning for NLP*, 2017, pp. 191–200.

[104] W. Xiong, J. Wu, H. Wang, V. Kulkarni, M. Yu, S. Chang, X. Guo, and W. Y. Wang, "Tweetqa: A social media focused question answering dataset," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 5020–5031.

[105] A. Warstadt, A. Singh, and S. R. Bowman, "Neural network acceptability judgments," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 625–641, 2019.

[106] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, 2011, pp. 142–150.

Publications

[1] Y. Kong, K. Yi, L. Wang, C. Peng, L. -M. Nguyen and Q. Ma, "RPMTD: A Route Planning Model with Consideration of Tourists' Distribution," in IEEE Access, vol. 12, pp. 69488-69504, 2024.

[2] Yuntao, K., Chen, P., Le, N. & Qiang, M. Dual Congestion-Aware Route Planning for Tourists by Multi-agent Reinforcement Learning. International Conference On Database And Expert Systems Applications. pp. 331-336 (2023)

[3] Yuntao, K., Phuong, N.M., Racharak, T., Le, T. and Le Minh Nguyen 0001, 2022. An Effective Method to Answer Multi-hop Questions by Single-hop QA System. In ICAART (2) (pp. 244-253).

[4] Yi, K., Maekawa, T., Kong, Y., Bai, Z., Jin, X., Ma, Q. (2024). U-KyotoTrip: A Travel Planning System for User Experience Oriented Trips. In: Berezina, K., Nixon, L., Tuomi, A. (eds) Information and Communication Technologies in Tourism 2024. ENTER 2024. Springer Proceedings in Business and Economics.

[5] Yi, K., Jin, X., Bai, Z., Kong, Y., Ma, Q. (2024). An Empirical User Study on Congestion-Aware Route Recommendation. In: Berezina, K., Nixon, L., Tuomi, A. (eds) Information and Communication Technologies in Tourism 2024. ENTER 2024. Springer Proceedings in Business and Economics.