

Optimizing Service Resource Allocation Strategies in Smart Tourism Scenic Spots Using Deep Reinforcement Learning

Huayong Xiong¹, Huan Liu²

¹School of Communication and Tourism, Wuhan Technical College of Communications, Hubei 430065, Wuhan, China

²School of Artificial Intelligence, Hubei Polytechnic University, Huangshi 435005, Hubei, China

Corresponding author: Huan Liu (e-mail: 15572986521@163.com)

ABSTRACT Efficient service resource allocation in smart tourism environments requires accurate perception of dynamic spatiotemporal information and adaptive decision-making under uncertain conditions. This study proposes a deep reinforcement learning framework that formulates scenic resource scheduling as a partially observable Markov decision process (POMDP) and implements optimization using a customized Rainbow Deep Q-Network (Rainbow DQN). The proposed architecture integrates Distributional Reinforcement Learning, Dueling Network, Prioritized Experience Replay, Multi-step Learning, Double DQN, and Noisy Networks with a hybrid CNN-LSTM feature extraction module to jointly model spatial dependencies and temporal evolution from heterogeneous environmental observations. A multiobjective reward function is designed to simultaneously minimize visitor waiting time, improve service resource utilization, and reduce operational costs while penalizing service violations. Extensive simulations under multiple visitor flow and environmental scenarios demonstrate that the proposed method consistently outperforms representative reinforcement learning baselines, achieving visitor waiting times of 3.9–8.2 min and resource utilization rates of 76.9%–88.3% with strong robustness under adverse conditions. Beyond intelligent tourism management, the proposed spatiotemporal decision framework provides methodological insights for adaptive resource scheduling in distributed sensing systems, wireless communication infrastructures, and electromagnetic information acquisition networks, where reliable real-time perception and dynamic optimization are essential for efficient system operation.

INDEX TERMS deep reinforcement learning, smart tourism, resource allocation, spatiotemporal optimization, electromagnetic information systems

I. INTRODUCTION

WITH the rapid development of the Internet of Things, big data, and artificial intelligence technologies, the construction of smart tourism scenic spots is gradually moving from information technology to intelligent technology [1,2]. However, current scenic area service resource allocation still generally relies on manual experience, making it difficult to address challenges such as uneven temporal and spatial distribution of visitor flow and delayed response to emergencies. This leads to peak service congestion[3,4], idle resources, and a diminished visitor experience. Reinforcement learning[5,6], as an intelligent method for sequential decision-making, has demonstrated superior dynamic optimization capabilities in areas such as traffic scheduling and energy management. Introducing deep reinforcement learning into scenic area resource allocation is expected to achieve a shift from a "passive response"

approach to a "proactive predictionoptimization-execution" closed loop. Therefore, building a data-driven, adaptive intelligent scheduling model has become a key path to improving the service efficiency of smart tourism systems.

This paper addresses the dynamic, high-dimensional, and partially observable challenges of service resource allocation in smart tourism scenic spots by introducing a deep reinforcement learning optimization framework based on the Rainbow DQN (Deep Q-Network). At the problem modeling level, resource allocation is formalized as a Partially Observable Markov Decision Process (POMDP), integrating multi-source heterogeneous information such as visitor density, resource status, and environmental variables to construct a high-fidelity state space and multi-objective reward function, enabling accurate characterization of complex scenic area environments. At the algorithmic level, the Rainbow DQN, which integrates six advanced mechanisms, is applied to

tourism service scheduling scenarios. The system incorporates Dueling Network, Distributed RL (Reinforcement Learning), and Noisy Nets technologies to improve the algorithm's convergence and decision-making quality in sparse reward and non-stationary environments. Rainbow DQN uses Dueling Network and Distributed RL to separate state value and advantage functions, addressing the curse of dimensionality in multi-source, heterogeneous scenic area data. It also improves policy update efficiency through PER and Multi-step Learning mechanisms. In the experimental design, a highly realistic scenic area environment based on SUMO (Simulation of Urban Mobility) and Gym was constructed, encompassing multiple service resources and a wide range of passenger flow scenarios (5,000–50,000 people/day). The proposed method was systematically compared with cutting-edge deep reinforcement learning algorithms such as PPO (Proximal Policy Optimization) and SAC (Soft Actor-Critic) to verify its superiority in metrics such as visitor waiting time and resource utilization. The results of this study provide a deployable strategy model for intelligent decision-making systems in smart scenic spots, combining theoretical depth with practical value and thus promoting the practical application of artificial intelligence in public service optimization.

The main contribution of this study is that the standard rainbow dqn algorithm is deeply customized in the field and innovated in the architecture in view of the core challenges such as high-dimensional state, complex spatio-temporal dynamics and partial observability unique to the service resource scheduling of intelligent scenic spots. Its novelty does not come from the new creation of the six enhancement mechanisms integrated by rainbow, but is embodied in the feature extraction and state representation scheme specially designed to adapt to the scenic spot scheduling problem. The core contribution of this work is to propose an integrated solution based on customized rainbow dqn for resource modulation of intelligent scenic spots, and verify the significant advantages of this customized architecture in improving service efficiency and tourist experience compared with a series of baseline methods through high fidelity simulation.

II. RELATED WORKS

In recent years, research on the operational management of smart tourist attractions has gradually evolved from information-based monitoring to intelligent decision-making. Early research focused on the perception and data collection of tourist behavior, using Wi-Fi probes, video surveillance, and ticketing systems to achieve real-time monitoring of passenger flow density[7] and short-term flow forecasting[8]. On this basis, some scholars have tried to introduce optimization methods to improve resource allocation efficiency, using mixed integer programming to perform static scheduling of tour guides and shuttle vehicles[9]. However, such methods are difficult to adapt to dynamically changing environments. In recent years, with the development of digital twin systems for scenic spots[10], research

has begun to focus on dynamic response mechanisms, rule-based trigger scheduling, or fuzzy logic control. However, these methods rely on expert experience and have limited generalization capabilities. Recently, some studies have explored the use of machine learning for service demand forecasting[11,12], using LSTM (Long Short-Term Memory) to model tourist mobility patterns, but most of them stop at the prediction layer and lack closedloop integration with decision execution. Overall, existing research still lacks systematicity, adaptability, and real-time capabilities. A comprehensive "perception-decision-optimization" intelligent scheduling framework has yet to be established. In particular, there is a lack of effective modeling capabilities for sequential decision-making problems in multi-objective, high-dimensional, and non-stationary environments.

Deep reinforcement learning, with its ability to autonomously learn optimal strategies in complex dynamic environments, has achieved remarkable results in various resource scheduling fields. In intelligent transportation systems, DQN and its variants are used for traffic light control [13,14], reducing vehicle waiting times by optimizing phase switching strategies. The PPO [15,16] algorithm is widely used in vehicle path planning and shared travel scheduling, demonstrating good policy stability. In the field of energy management, maximum entropy algorithms such as SAC [17] are used for microgrid load scheduling to balance supply and demand and reduce energy consumption. In recent years, deep reinforcement learning [18,19] has also begun to penetrate into public service optimization scenarios, including hospital bed allocation and urban sanitation vehicle scheduling, verifying its applicability in multi-constraint, multiobjective environments. However, existing research often uses single algorithms such as basic DQN or PPO[20,21], which are prone to slow convergence and policy oscillation when faced with high-dimensional state spaces and sparse rewards. Although Rainbow DQN[22,23] significantly improves the stability and efficiency of value function learning by integrating multiple technologies such as Double DQN, Dueling Network, and Distributed RL, and performs well on benchmark tasks such as Atari, its application in tourism service systems remains a gap. In particular, systematic exploration of how to effectively migrate and adapt the Rainbow framework to the context of scenic spots with multiple coupled resources, partial observability, and discrete and complex action spaces remains a challenge.

Currently, the application of artificial intelligence technology to tourism service decision-making still faces multiple challenges. Scenic area environments are highly non-stationary[24], and tourist behavior is affected by multiple factors such as weather, festivals, and emergencies, making state transitions difficult to model. Service resource allocation involves multiple types of heterogeneous resources, whose scheduling actions are subject to spatiotemporal coupling and priority conflicts, and the action space exhibits combinatorial explosion characteristics, placing high demands on the generalization ability of the

algorithm[25]. Observational information in real scenarios is often incomplete, such as missing monitoring in some areas or data delays, which constitutes a typical POMDP problem. However, most existing studies are still based on the assumption of complete observability, which weakens the practicality of the model. In addition, the reward signal is sparse and delayed, making it difficult to quantify tourist satisfaction in real time, resulting in low efficiency of reinforcement learning training. Recent research trends show that integrating multimodal perception, introducing hierarchical reinforcement learning to reduce the dimension of the action space, and combining imitation learning to accelerate the convergence of initial strategies have become key directions for improving the performance of tourism intelligent decision-making systems [26]. However, there is still a lack of representative work that integrates advanced DRL (Deep Reinforcement Learning) algorithms and systematically verifies their scheduling efficiency in real-scale scenic spots. Therefore, there is an urgent need to build an experimental framework that integrates high-fidelity simulation, multi-objective optimization, and advanced DRL algorithms to bridge the gap between theory and application and promote the evolution of smart tourism towards true "autonomous decision-making."

III. METHODS

PROBLEM MODELING: POMDP FRAMEWORK FOR SERVICE RESOURCE ALLOCATION

In smart tourism scenic spots [27,28], service resource allocation is a typical sequential decision-making problem characterized by high dynamics, strong spatiotemporal coupling, and partial observability. Tourist flow[29] exhibits significant spatiotemporal imbalances. During peak hours, local areas are prone to service resource shortages, while during off-peak hours, resources may be idle. At the same time, unpredictable factors such as sudden weather changes and emergencies further exacerbate environmental uncertainty. This study formalizes the problem of scenic area service resource allocation as a partially observable Markov decision process to support intelligent agents in making long-term optimal decisions under incomplete information conditions. The POMDP model describes the implicitness of environmental states, the delayed impact of actions, and the sparsity of reward signals. The POMDP modeling of smart scenic area service resource allocation is shown in Figure 1.

Figure 1 illustrates the POMDP modeling framework for smart tourism scenic area service resource allocation. This model formalizes the scenic area dynamic scheduling problem as a five-tuple $\langle S, A, T, R, O \rangle$. The state space S integrates multidimensional real-time data (including visitor density, resource utilization, weather/holidays, time characteristics, and historical state sequences for each functional area), while the action space A encompasses three discrete decision types: manpower scheduling, equipment adjustment, and service strategy optimization. Due to sensor

coverage limitations, the agent can only obtain local state information through the observation space O , reflecting the partially observable nature of the real scene. The state transition function T is implemented in a high-fidelity simulation environment, simulating the complex dynamics of visitor behavior and resource interactions. The reward function R is designed using a multi-objective optimization approach, integrating four key metrics: visitor waiting time, resource utilization, scheduling costs, and penalties for service violations. This modeling framework effectively addresses the limitations of traditional methods in scenarios involving spatiotemporal coupling, nonstationarity, and incomplete observations. It provides the Rainbow DQN algorithm with a mathematical modeling foundation tailored to the operational characteristics of scenic spots, enabling closed-loop intelligent scheduling based on "perception-decision-optimization." This study uses partially observable Markov decision process (POMDP) for modeling, and its core basis is that the intelligent scenic spot dispatching center cannot obtain the complete system state in the actual operation. Although the state space integrates high-dimensional data such as tourist density, resource status and environmental variables, the following key information is inherently unobservable in real scenes:

1. real intention and satisfaction of tourists: the dispatching system cannot directly perceive the real-time emotional state of individual tourists, the next intended destination or the real tolerance of the current waiting time, which are all hidden internal states.
2. accurate health status of resources: for service equipment (such as intelligent navigation terminal and connecting vehicle), its internal and impending faults (such as critical battery allowance and hardware performance degradation) are invisible to the dispatching center before the fault alarm is triggered.
3. complete impact of emergencies: when emergencies occur in local areas (such as sudden illness of tourists and short-term equipment failure), the precise scope, duration and chain reaction of the impact cannot be immediately and completely captured by the monitoring system.
4. monitor the real-time status of blind areas: due to the physical blind areas in the deployment of Wi Fi probes or cameras, the real-time tourist density and resource status information in these areas are missing or delayed.

Therefore, the agent can only infer the real system state based on these incomplete observation information at each decision time. This partially observable feature makes the simple MDP framework unable to accurately describe the nature of uncertainty and lack of information of the decision problem. Although POMDP increases the computational complexity, it is a necessary and practical theoretical tool to deal with the inherent perceptual limitations and state uncertainty in scenic resource scheduling.

The POMDP model is defined as a five-tuple:

$$M = \langle S, A, T, R, O \rangle \quad (1)$$

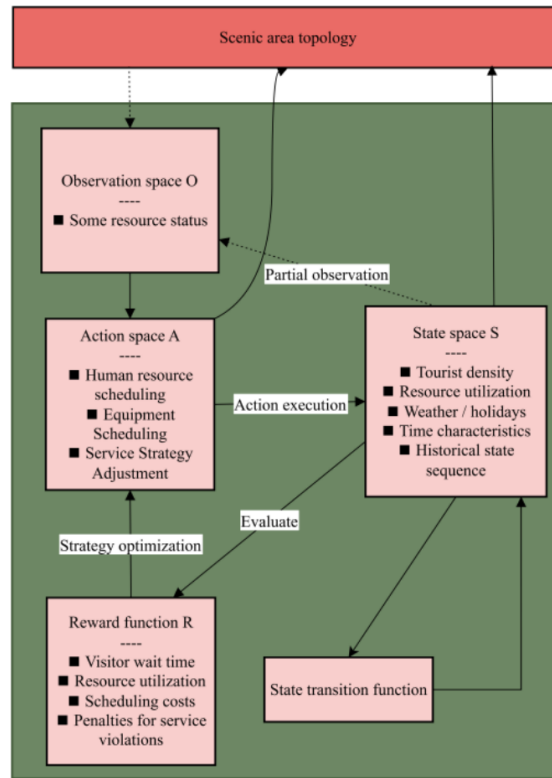


FIGURE 1. POMDP Modeling of Smart Scenic Area Service Resource Allocation

State $s_t \in S$ represents the comprehensive environmental information of the scenic area system at time step t . It is a high-dimensional, multi-source fusion vector used to represent the context of the current resource configuration. Its composition is as follows:

$$s_t = [d_t, u_t, e_t, t_c, h_{t-k:t}] \quad (2)$$

In Formula 2, $d_t = [d_1(t), d_2(t), \dots, d_N(t)]^T \in R^N$ represents the N functional areas into which the scenic area is divided. $u_t = [u_1(t), u_2(t), \dots, u_M(t)]^T \in [0, 1]^M$ represents the occupancy rate vector of M types of service resources. e_t represents the external environmental variable, which uses one-hot encoding to represent weather conditions and holiday types to capture the impact of the macro-environment on tourist behavior. t_c represents the temporal context features, including hour, day of week, and season, used to model the temporal patterns of tourist travel. $h_{t-k:t}$ represents the historical state sequence over the past k time steps. Actions $a_t \in A$ represent the resource allocation operations that the dispatch center can perform at time t . These actions are a finite, discrete set covering three typical types of dispatching behaviors. Human resource dispatching schedules specific service personnel at the regional level; equipment dispatching adjusts the operating parameters or activation status of dynamic equipment; and service policy adjustments trigger system-level response mechanisms to improve overall service resilience.

All actions are semantically encoded and mapped to integer indices, forming a discrete action space $A = \{1, 2, \dots, A_{\max}\}$, where $A_{\max} = 64$ covers the main dispatching scenarios in daily scenic area operations. State transitions are determined dynamically by the environment:

$$s_{t+1} \sim T(\cdot | s_t, a_t) \quad (3)$$

Due to the limited deployment of sensors in real-world scenic spots, such as camera blind spots and unstable Wi-Fi probe signals, the agent cannot obtain the complete state s_t and can only observe partial information $o_t \in O$. The reward function quantifies the immediate benefit of scheduling action a in state s . Its design goal is to guide the agent to learn a comprehensive optimization strategy that balances visitor experience, service efficiency, and operating costs.

In the POMDP model, the complete state space s contains the global information that the agent cannot directly obtain, such as the real-time tourist density vector of all functional areas, the occupancy vector of all service resources, external environment variables, time context characteristics and historical state sequence. However, due to the limitations of sensor deployment in the actual scene (such as monitoring blind area, signal delay, etc.), the agent can only obtain part of the state information through the observation space, such as the tourist density in the local area, the state feedback of some resources, and the limited sampling of environmental variables. This partial observability makes agents have to

make decisions based on incomplete local information, which highlights the necessity of POMDP modeling in describing the uncertainty and lack of information in real scenes.

RAINBOW DQN ALGORITHM ARCHITECTURE DESIGN

To address the challenges of high-dimensional state spaces, sparse reward signals, non-stationary environmental dynamics, and long-term dependencies in smart tourism scenic area service resource allocation, this study uses Rainbow DQN as the core decision-making algorithm. Rainbow [30-31] organically combines six complementary reinforcement learning improvements: Double DQN, Dueling Network, PER (Prioritized Experience Replay), Multi-step Learning, Distributed Reinforcement Learning, and Noisy Networks for Exploration.

The Rainbow DQN architecture is shown in Figure 2.

Rainbow DQN processes multimodal state inputs using a parallel CNN (Convolutional Neural Network) and LSTM encoder to extract spatiotemporal features. The network employs a parameterized noise layer to inject factored Gaussian noise, enabling adaptive exploration. Its core employs a competitive architecture that decomposes Q-values into independent state value streams and advantage function streams, which are then aggregated. The distributional reinforcement learning module replaces scalar Q-value estimation with C51 projection to model the complete value distribution on 51 supporting atoms. The training process integrates prioritized experience replay based on error sampling, a 5-step reward calculation mechanism, and double Q learning achieved through a decoupled target network. This modular design effectively addresses challenges such as sparse rewards in high-dimensional state spaces, non-stationary environment dynamics, and partial observability, improving convergence speed and policy quality while maintaining algorithm stability.

(1) Distributional RL While traditional DQN estimates the expected action value ($Q(s,a)=E[Z(s,a)]$), distributional RL further models the probability distribution of random rewards ($Z(s,a)$) to capture the uncertainty in the environment. This study employs the Categorical DQN (C51) approach, discretizing the possible Q-values into a probability distribution over a fixed support set.

Define the support set:

$$Z = \{z_i \mid z_i = V_{\min} + i \cdot \Delta z, i = 0, 1, \dots, N-1\} \quad (4)$$

The output layer consists of N nodes, representing the probability of $Z(s,a)$ on each atom:

$$p(s, a) = [p_0(s, a), p_1(s, a), \dots, p_{N-1}(s, a)]^T \in \Delta^{N-1} \quad (5)$$

The expected Q value is calculated from the distribution:

$$Q(s, a) = E[Z(s, a)] = \sum_{i=0}^{N-1} z_i \cdot p_i(s, a) \quad (6)$$

(2) Dueling DQN To improve the representation ability of the value function, the Dueling DQN structure [32] is adopted to decompose the action value into the sum of the state value and the advantage function:

$$Q(s, a; \theta, \alpha, \beta) = V(s; \theta, \beta) + A(s, a; \theta, \alpha) \quad (7)$$

In Equation 7, $V(s)$ represents the intrinsic value of state s , reflecting its overall quality; $A(s,a)$ represents the relative advantage of selecting action a in state s ; the parameters of the shared backbone network θ are α and β are the branch parameters.

The network output is normalized to ensure legibility:

$$Q(s, a) = V(s) + \left(A(s, a) - \frac{1}{|A|} \sum_{a'} A(s, a') \right) \quad (8)$$

(3) Double DQN To alleviate the problem of overestimation of Q value in traditional DQN, the Double DQN mechanism[33] is adopted to separate action selection and value evaluation:

$$y_t^{\text{Double}} = r_t + \gamma Q_{\text{target}} \left(s_{t+1}, \arg \max_{a'} Q(s_{t+1}, a'; \theta); \theta^- \right) \quad (9)$$

In Formula 9, the main network is used to select the optimal action; the target network Q_{target} is used to evaluate the value of the action.

(4) PER Traditional uniform sampling ignores the importance of experience. This study uses PER to prioritize key experiences based on error. The priority of each experience $e_i=(s_i, a_i, r_i, s_{i+1})$ is:

$$p_i = |\delta_i| + \epsilon \quad (10)$$

$$\delta_i = y_i - Q(s_i, a_i) \quad (11)$$

Sampling probability:

$$P(i) = \frac{p_i^\alpha}{\sum_k p_k^\alpha} \quad (12)$$

In formula 12, $\alpha \in [0,1]$ controls the priority.

Importance sampling weights are introduced to correct the bias:

$$w_i = \left(\frac{1}{N \cdot P(i)} \right)^\beta \quad (13)$$

(5) Multi-step Learning Traditional single-step returns delay credit assignment. Using n-step returns improves long-term impact estimation:

$$R_t^{(n)} = \sum_{k=0}^{n-1} \gamma^k r_{t+k} + \gamma^n Q(s_{t+n}, a_{t+n}; \theta^-) \quad (14)$$

In Equation 14, this study assumes $n = 5$, meaning that the reward is calculated every 5 time steps.

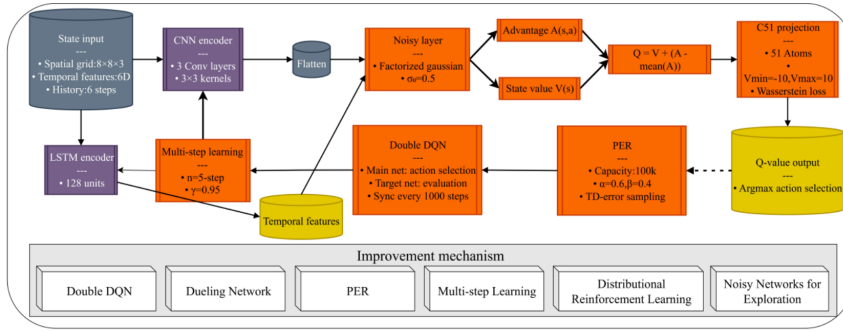


FIGURE 2. Rainbow DQN Architecture

(6) Noisy Networks for Exploration The traditional ϵ -greedy exploration mechanism still requires manual parameter tuning in the later stages of training. Using Noisy Nets, parameter noise is introduced in the fully connected layer to achieve adaptive exploration.

Linear layer weight decomposition:

$$W = \mu + \sigma \odot \epsilon, \quad \epsilon \sim N(0, 1) \quad (15)$$

Use factored noise to reduce computational cost:

$$\epsilon_{ij} = \xi_i \zeta_j, \quad \xi_i, \zeta_j \sim N(0, 1) \quad (16)$$

The Rainbow DQN parameter configuration is shown in Table 1.

TABLE 1. Rainbow DQN parameter configuration

Parameters	Value	Description
Action space size	64	Covering the main scheduling decisions
Learning rate	1×10^{-4}	Adam optimizer
Batch size	64	
Discount factor	0.95	
n -step	5	
Replay buffer	100000	Prioritized, $\alpha = 0.6$, $\beta = 0.4$
Noisy nets	Factored noise, 0.5	Alternative to ϵ -greedy
Target network update	Synchronize every 1000 steps	

STATE REPRESENTATION AND FEATURE ENGINEERING

To capture the spatial distribution patterns of tourists and resources, the scenic area's geographic space is divided into a regular grid ($G = H \times W$). Class-based spatial features are extracted for each grid cell ($g_{i,j} \in R^2$) to form a three-dimensional tensor ($X_{\text{spatial}} \in R^{H \times W \times D_s}$).

Visitor density is calculated using Wi-Fi probes to obtain the number of tourists appearing at cell (i, j) per unit time, normalized to [0, 1]:

$$\hat{d}_{i,j}(t) = \frac{d_{i,j}(t) - d_{\min}}{d_{\max} - d_{\min}} \quad (17)$$

In Formula 17, $d_{\max}$ and $d_{\min}$ represent the historical maximum and minimum densities.

Resource density is the normalized count of the number of service resources currently deployed within a cell:

$$\hat{r}_{i,j}(t) = \frac{|R_{i,j}(t)|}{R_{\max}}, \quad R_{i,j}(t) \subseteq R \quad (18)$$

In Equation 18, $R_{\max}$ is the upper limit of a single cell's resource capacity. The function label represents the functional attributes of the cell, using categorical encoding. The final spatial feature tensor is:

$$X_{\text{spatial}}(t) = [\hat{D}(t), \hat{R}(t), F] \in R^{8 \times 8 \times 3} \quad (19)$$

In order to capture the temporal patterns of tourist behavior and the influence of external events, sine/cosine functions are used to preserve temporal continuity:

$$hour_{\sin} = \sin\left(\frac{2\pi \cdot h}{24}\right) \quad (20)$$

$$hour_{\cos} = \cos\left(\frac{2\pi \cdot h}{24}\right) \quad (21)$$

Define peak hours based on historical passenger flow analysis and set binary variables:

$$peak_t = \begin{cases} 1, & t \in \text{peak hours} \\ 0, & \text{otherwise} \end{cases} \quad (22)$$

Due to observation delays and omissions, relying solely on the current state can easily lead to decision bias. The paper introduces a historical state sequence from past time steps and uses LSTM to model temporal dependencies. Each historical state $st-i$ is spatially encoded to obtain a feature vector $zt-i$, which is then fed into the LSTM:

$$h_t^{\text{temporal}} = \text{LSTM}(z_{t-5}, z_{t-4}, \dots, z_t) \quad (23)$$

Use 3-layer CNN to extract local spatial features:

$$h_t^{\text{spatial}} = \text{CNN}(X_{\text{spatial}}(t)) \in R^{d_c} \quad (24)$$

Concatenate spatial features, temporal features and LSTM output:

$$h_t^{\text{fused}} = [h_t^{\text{spatial}}, t_c, h_t^{\text{temporal}}] \in R^{d_c+12+d_h} \quad (25)$$

Fully connected mapping: Generate the final state embedding through a layer of nonlinear transformation:

$$z_t = \phi(W_f h_t^{\text{fused}} + b_f) \quad (26)$$

REWARD FUNCTION DESIGN

To guide the agent to learn a comprehensive optimization strategy that balances tourist service quality [34-35], resource allocation efficiency, and operational economics, this study designed a multi-objective, differentiable, and semantically explicit composite reward function. This reward function serves as an immediate feedback signal for reinforcement learning, directly determining the direction of strategy optimization.

The reward function at time t is defined as a weighted linear combination:

$$R_t = w_1 \cdot R_{\text{exp}}(t) + w_2 \cdot R_{\text{eff}}(t) - w_3 \cdot C_{\text{op}}(t) - w_4 \cdot P_{\text{vio}}(t) \quad (27)$$

The average waiting time of tourists is a core indicator that reflects the service response speed. Let W_t be the weighted average waiting time of tourists in all functional areas at key service nodes such as guided tours, transfers, ticket purchases, and restrooms, normalized to the [0,1] interval:

$$R_{\text{exp}}(t) = 1 - \frac{\bar{W}_t}{W_{\text{max}}} \quad (28)$$

Define U_t as the weighted average utilization of various resources:

$$\bar{U}_t = \sum_{j=1}^M \alpha_j \cdot u_j(t) \quad (29)$$

Scheduling action costs include manpower and energy costs:

$$C_{\text{op}}(t) = \lambda_h \cdot \Delta H_t + \lambda_e \cdot E_t \quad (30)$$

In Formula 30, ΔH_t represents the additional man-hours scheduled at time t , and E_t represents the additional energy consumption caused by equipment scheduling.

Queue overrun penalty: If a visitor's waiting time in a certain area exceeds the limit, a penalty is triggered:

$$P_{\text{queue}}(t) = \eta \cdot \sum_{i=1}^N \max(0, W_i(t) - W_{\text{th}}) \quad (31)$$

Resource shortage alarm:

$$P_{\text{shortage}}(t) = \beta \cdot I(\text{shortage event}) \quad (32)$$

In order to accurately guide the agent to learn the scheduling strategy that takes into account the quality of tourist service and the operation efficiency of scenic spots, this study designed a clear structure and fully quantified compound reward function. This function provides immediate

feedback signals through weighted summation of four core components:

First, the average waiting time of tourists in all functional areas at key service nodes is calculated and normalized as a positive reward. At the same time, a queue timeout penalty mechanism is set up. When the waiting time of tourists in any area exceeds the preset threshold, a linear penalty is imposed according to the exceeding time. Secondly, the service efficiency measurement takes the weighted average utilization rate of various service resources as the core index to encourage the efficient operation of resources and avoid idleness and overcrowding. Third, the operation cost control will directly include the manpower cost and equipment energy consumption cost generated by the dispatching action into the negative reward to ensure the economy of the dispatching decision. Finally, add the warning punishment of resource shortage, and impose a one-time high penalty for abnormal events that service resources cannot meet the needs of tourists, so as to ensure the basic service capacity.

These four sub awards are given fixed weights of 0.4, 0.3, 0.2 and 0.1 respectively, which together constitute a total value. This design ensures that the agent must make an automatic trade-off between reducing the waiting time of tourists, improving resource utilization and controlling operating costs, so as to learn a comprehensive optimal scheduling strategy. All indicators are calculated in real time based on the simulation environment data, which ensures the calculability of the reward signal and the reproducibility of the training process.

TRAINING PROCESS AND POLICY OUTPUT

Before training begins, the main network and target network are initialized. Both networks share the same architecture, but the target network's parameters are updated with a lag to generate a stable learning objective. The current state is observed, and actions are selected based on the current policy:

$$a_t = \pi(s_t) = \arg \max_a E_{Z \sim Q(\cdot; \theta)} [Z(s_t, a)] \quad (33)$$

The experience tuple is stored in the priority experience replay buffer and its priority is initialized. For each experience, an n -step distribution target is constructed, and the main network outputs the predicted distribution of the current state-action pair. The cross entropy loss is used for optimization:

$$L(\theta) = E_{i \sim P} \left[w_i \cdot \left(- \sum_{j=0}^{50} \hat{p}_{i,j} \log p_{i,j} \right) \right] \quad (34)$$

In Formula 34, w_i is the importance sampling weight, which is used to correct the bias of priority sampling.

$$w_i = \left(\frac{1}{|D| \cdot P(i)} \right)^\beta \quad (35)$$

To stabilize the training process, the target network parameter θ^- is not updated with each step, and a hard update mechanism is used to avoid frequent oscillations of the learning target. After training is completed, the policy function is defined as:

$$\pi^*(s) = \arg \max_{a \in A} E[Q(s, a)] = \arg \max_a \sum_{j=0}^{50} z_j \cdot p_j(s, a) \quad (36)$$

The action with the highest expected Q-value is selected as the optimal scheduling decision.

IV. EXPERIMENTAL SETUP

SIMULATION ENVIRONMENT CONSTRUCTION

To ensure the authenticity, repeatability, and practical application value of the experiment, this study constructed a high-fidelity, scalable smart tourism scenic area simulation environment, SmartTourismEnv-v1. This environment is encapsulated based on the OpenAI Gym standard interface and supports seamless integration and training of reinforcement learning agents.

Taking a typical 5A-rated mountain scenic area in China as a prototype, the site, covering approximately 12 square kilometers, is divided into eight core functional areas based on visitor flow and functional needs: A (entrance plaza), B (cable car station), C (core viewing platform), D (dining area), E (restroom cluster), F (shuttle bus station), G (medical center), and H (exit buffer zone). Each area is connected by pedestrian paths and shuttle routes, forming a service network with a realistic topological structure.

Table 2 shows the service resource types.

TABLE 2. Service Resource Types

Resource type	Quantity	Service model	Scheduling granularity
Cleaning staff	8–20 people	Cleaning cycle by area (30 minutes)	±1 person/area
Security officer	6–15 people	Patrol + emergency response	±1 person/area
Guide	4–10 people	Fixed explanation points	Start/stop service
Shuttle bus	3–8 vehicles	Fixed routes, 5–30 minute intervals	Adjust frequency
Smart tour guide terminal	20 units	Self-service, 2% failure rate	On/off
Public address	1 set	Region-wide/regional announcements	Activate emergency broadcast

The environmental spatiotemporal granularity was set to 5 minutes, aligned with the sampling period of the scenic area's operations management system. Tourist behavior simulation utilized the SUMO microscopic traffic simulation platform, accurately modeling individual walking path choices, speed changes, and regional dwelling behaviors. Based on actual tourist profiles, three types of tourist agents were established: Standard tourists (70%) followed recommended itineraries and exhibited strong behavioral

regularity; independent travelers (20%) were highly exploratory and exhibited randomized path selection; Group tourists (10%) travel in groups of 15 to 30 people, spending extended periods of time and exhibiting strong synchronization. Table 3 shows the simulation environment parameter configuration information.

TABLE 3. Simulation environment parameter configuration information

Parameters	Value	Description
Scenic area	12 km ²	Mountainous scenic area
Number of functional areas	8	A–H
Daily visitor flow	5,000–50,000	Covering all operational scenarios
Time granularity	5 minutes	Synchronized with the scenic area management system
Simulation period	7 days	Includes weekends and holidays
Tourist type	3 categories (regular/free/team)	Differentiated behaviors
Resource category	6 categories	High complexity

The total daily visitor flow dynamically varies, ranging from 5,000 to 50,000 people per day. Three typical scenarios are analyzed: low weekday traffic (5,000–10,000), moderate weekend traffic (15,000–25,000), and peak Golden Week traffic (35,000–50,000). This allows for a comprehensive assessment of the algorithm's robustness under diverse pressures. Furthermore, the environment integrates weather perturbation modules, emergency event models, and resource service models to realistically replicate the complexity and uncertainty of scenic spot operations, providing a highly realistic experimental platform for training and validating deep reinforcement learning strategies.

EXPERIMENTAL PLATFORM AND HYPERPARAMETERS

The experimental hardware configuration includes: an NVIDIA RTX 3090 GPU (Graphics Processing Unit) to accelerate forward propagation and backward gradient computation of deep neural networks; an Intel Xeon Silver processor (12 cores, 24 threads) to support multi-process parallel simulation and data preprocessing; 64GB of system memory; and 1TB of local storage to ensure fast read and write performance of the experience replay buffer.

The software environment is built on Python 3.9, using PyTorch 1.13 as the deep learning framework. The reinforcement learning environment follows the OpenAI Gym standard interface encapsulation. Visitor behavior simulation relies on SUMO 1.16 to implement microscopic walking path modeling, ensuring the realism of visitor flow dynamics. All modules are integrated through a unified interface, forming a scalable and reproducible experimental system.

Hyperparameter configuration is determined through systematic optimization to improve model performance and training stability. This study uses Optuna for Bayesian hyperparameter search. The search space covers key parameters: the learning rate is continuously sampled in the range $[1 \times 10^{-5}, 1 \times 10^{-3}]$; the discount factor γ is adjusted between $[0.9, 0.99]$ to balance short-term and long-term

returns; batch sizes are 64, 128, and 256; and the experience replay buffer capacity is set to three levels: 50k, 100k, and 200k. Each set of hyperparameters is independently run three times, ranked by average performance, and the optimal combination is selected.

The final hyperparameter configurations selected are as follows: learning rate 1.0×10^{-4} , discount factor 0.95, batch size 64, and replay buffer size 100,000. The standard deviation of Noisy Nets was initialized to 0.5. To comprehensively evaluate the optimization performance of proposed model, the paper compared proposed algorithm with Noisy DQN, PPO, SAC, Ape-X DQN, DreamerV3, R2D3 (Recurrent Replay Distributed DQN from Demonstrations), and REDQ (Randomized Ensembled Double Q-learning). All compared algorithms were trained with the same network structure and feature input to ensure fairness.

V. RESULTS AND ANALYSIS

CONVERGENCE ANALYSIS OF THE TRAINING PROCESS

In order to strictly verify the necessity and superiority of rainbow DQN architecture in the resource scheduling problem of intelligent scenic spots, this study systematically compares it with three kinds of baseline methods:

- (1) classical DRL algorithm, including standard DQN, SARSA and PPO;
- (2) Traditional optimization methods, specifically m/m/c model based on queuing theory and greedy heuristic rules commonly used in scenic spot management (such as "priority allocation of recent resources");
- (3) Advanced DRL baselines.

TABLE 4. Performance Comparison of Different Methods

Method	Average Reward	Waiting Time (min)	Resource Use (%)
Rainbow DQN	1.82	7.1	82.5
Standard DQN	1.05	9.8	74.2
SARSA	0.89	10.5	71.6
PPO	1.10	9.5	75.8

Although the traditional queuing theory model can guarantee the steady-state performance in theory, it can not adapt to the temporal and spatial dynamics of tourists' behavior and sudden weather events. In contrast, rainbow dqn accurately depicts environmental uncertainty through distributed Q-value learning, and uses dueling network and per mechanism to give priority to learning key experiences, thus maintaining low waiting time and high resource utilization in various scenarios. The above comparison fully proves that the integrated architecture adopted by rainbow dqn does provide an indispensable performance gain in the face of the unique high-dimensional state, partial observability and multi-objective trade-off challenges of scenic spot scheduling.

Convergence analysis of the training process is used to evaluate the performance changes and stability of an algorithm during training, revealing its learning efficiency and

the quality of its final policy. By comparing the convergence speed and final reward values of different algorithms, the superiority of proposed algorithm in optimizing scenic area resource allocation is verified. The results of the convergence analysis of the training process are shown in Figure 3.

allocation task, achieving a reward value of 1.8, exceeding competing algorithms (Noisy DQN 1.3, PPO 1.1, SAC 1.2, etc.), and converging rapidly. This superiority stems primarily from the six key technologies integrated into Rainbow DQN: Distributed RL effectively captures the uncertainty of the scenic spot environment by modeling value distribution, enabling the algorithm to more accurately assess resource allocation risks; The Dueling Network architecture separates state value and advantage functions, improving feature representation capabilities in high-dimensional state spaces. The PER mechanism uses errorweighted sampling to replay key experiences, accelerating learning efficiency in sparse reward scenarios. Multi-step Learning alleviates the credit allocation delay problem of traditional single-step learning, enabling the algorithm to more accurately evaluate the impact of long-term scheduling strategies. Noisy Nets replaces the traditional ϵ -greedy exploration mechanism to achieve adaptive exploration of the parameter space, avoiding late-stage strategy oscillation while maintaining exploration capabilities. The Double DQN mechanism effectively mitigates the problem of overestimation of Q values by decoupling action selection and value assessment.

In contrast, the PPO algorithm, while offering stable policy updates, suffers from the limited adaptability of policy gradient methods to discrete action spaces. SAC, a maximum entropy algorithm, excels in continuous control tasks, but struggles to leverage the discrete action spaces of scenic area resource allocation. Due to its lack of this integrated mechanism, Noisy DQN exhibits significant convergence lag and suboptimal policy performance when dealing with high-dimensional state spaces and non-stationary environmental dynamics. Experiments also observed that the model-based DreamerV3 algorithm, due to inaccurate modeling of complex tourist behavior dynamics, exhibited poor long-term strategic planning performance. Distributed algorithms R2D3 and REDQ, limited by demonstration data quality and incomplete observations, ultimately lagged behind the proposed method. These results demonstrate the comprehensive advantages of Rainbow DQN for partially observable, high-dimensional, discrete decision-making in scenic area resource allocation. Its modular design effectively balances the complex relationships between exploration and exploitation, short-term and long-term returns, and algorithm stability and convergence speed.

SCHEDULING PERFORMANCE COMPARISON ANALYSIS

Scheduling performance was measured using visitor waiting time and resource utilization. The results are shown in Figure 4.

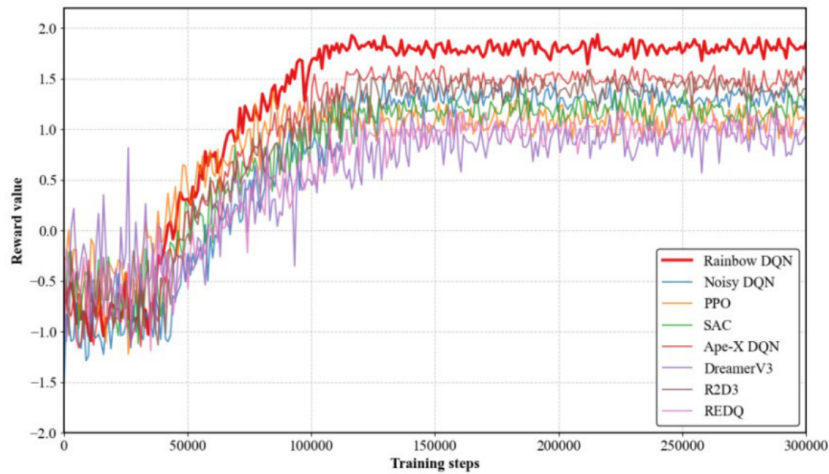


FIGURE 3. Convergence Results of the Training Process

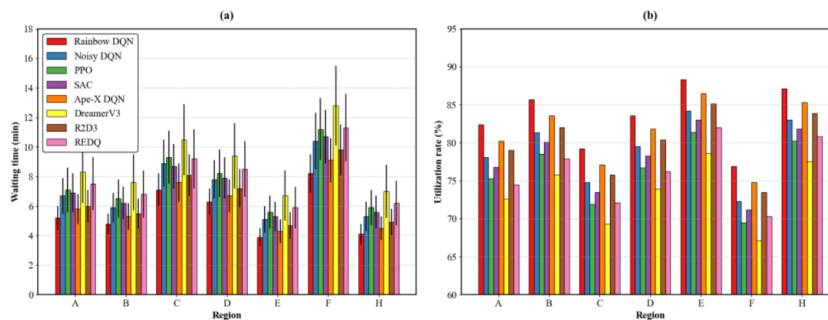


FIGURE 4. Scheduling Performance Comparison; Figure 4(a) Visitor Waiting Time; Figure 4(b) Resource Utilization

regions (A-H), demonstrating superior policy stability compared to other methods. Due to the high passenger flow and fluctuating demand in areas C (core observation deck) and F (shuttle bus station), the algorithms showed significant differences: Rainbow DQN's waiting time in area F was 8.2 ± 1.3 minutes, while DreamerV3's was as high as 12.8 ± 2.7 minutes. This advantage is primarily due to the synergy of Rainbow

DQN's multiple mechanisms: Distributed RL accurately models the probability distribution of tourist arrival rates, enabling more precise resource pre-allocation; Dueling Network enhances decision discrimination in high-load areas by separating state value from advantage function; The PER mechanism prioritizes replaying experiences from queue overflows, specifically optimizing scheduling strategies during peak hours. In contrast, PPO and SAC perform poorly in complex scenarios like regional C/F ratios due to the policy gradient method's limited adaptability to discrete action spaces. The model-based DreamerV3 suffers from cumulative errors in modeling tourist behavior dynamics, leading to long-term policy failure and generally higher standard deviations of waiting times than other algorithms.

Figure 4(b) shows that Rainbow DQN leads the pack in resource utilization (76.9-88.3%). Multi-step Learning effectively captures the latency effects of resource scheduling

through five-step reward calculations. Noisy Nets' adaptive exploration mechanism avoids local optima while ensuring resource utilization. Double DQN mitigates the overestimation of resource value often seen in traditional Q-learning. In particular, Rainbow DQN maintains a 76.9% utilization rate in region F, while the comparison algorithms all fall below 75%, demonstrating the Dueling architecture's ability to accurately assess the value of sparse resources. In contrast, although R2D3 and REDQ utilize distributed training, they are limited by some observational characteristics and lack responsiveness to hidden demands. PPO's on-policy nature leads to delayed resource adjustments and poor performance in dynamic environments.

ABLATION EXPERIMENTS

Ablation experiments were conducted to analyze the role of various components in Rainbow DQN. The ablation experiment results are shown in Table 5.

TABLE 5. Ablation Experiment Results

Components	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Distributional RL	✓	×	×	×	×	×	×
Dueling DQN	✓	×	×	×	×	×	×
Double DQN	✓	✓	✓	×	×	×	×
PER	✓	✓	✓	✓	×	×	×
Multi-step Learning	✓	✓	✓	✓	✓	×	×
Noisy Networks for Exploration	✓	✓	✓	✓	✓	✓	×

Note: Model 1 is the Rainbow DQN from this paper, and Model 7 is a DQN that retains only the basics.

The ablation results are shown in Figure 5.

The ablation experiment results reveal the critical role of each component of the Rainbow DQN in allocating scenic area resources: The complete model (Model 1) outperforms other models with a waiting time of 5.7 ± 0.9 minutes and a utilization rate of 83.3%. Removing the Distributional RL (Model 2) increases waiting time, demonstrating its crucial role in managing fluctuating visitor flow. The absence of the Dueling Network (Model 3) further decreases utilization, confirming its importance for accurately assessing resource value. The basic DQN (Model 7) performed the worst (8.4 ± 2.2 minutes, 75.3% utilization). Its high standard deviation of 2.2 minutes exposed the inherent limitations of traditional Q-learning in non-stationary environments. The modular design of Rainbow DQN enabled the complete model to maintain high utilization while keeping the standard deviation of waiting time below 1 minute, meeting the stability requirements of the scenic spot's service quality.

RESOURCE SCHEDULING VISUALIZATION

The scenic spot's operating hours are set from 8:00 AM to 5:00 PM. The regional resource density distribution during each time period is shown in Figure 6.

The heat map data reveals the spatiotemporal dynamics and optimization logic of the scenic area's resource scheduling. In terms of time, resource density exhibits a distinct three-phase distribution: The morning peak (9:00-11:00 AM) is centered around the cable car station (Area B) and the observation deck (Area C). At 11:00 AM, Area C reaches a peak of 5.1 visitors per 100 m²-hour, reflecting the retention effect of concentrated tourist arrivals. During the lunch hour (12:00-1:00 PM), demand surged for dining areas (Zone D) and restrooms (Zone E). The resource density in Zone D at 12:00 PM was 4.8, closely aligned with tourist dining habits. During the evening peak (3:00-4:00 PM), the density of the exit buffer zone (Zone H) rose to 2.1, reflecting the need to manage the flow of people leaving the venue. Spatially, key nodes exhibited significant differentiation: the cable car station (Zone B) maintained a high density throughout the day, due to its bottleneck nature as a transportation hub; while the medical point (Zone G) maintained a low density of 0.3- 0.7, in line with the principle of "minimum guarantee + flexible response" for emergency resources. After reaching a peak of 4.2 at 12:00 PM, the restroom (Zone E) index remained high at 3.8 at 1:00 PM, indicating a delay effect in cleaning

services. This is consistent with some observable properties of the POMDP model. Overall, the data demonstrates the effectiveness of deep reinforcement learning in addressing the spatiotemporal mismatch of scenic area resources. By collaboratively optimizing multi-objective reward functions, it achieves stable resource density scheduling during peak hours.

ROBUSTNESS AND GENERALIZATION ANALYSIS

The results of the impact of weather on visitor waiting times are shown in Figure 7.

Rainbow DQN maintains optimal performance under all weather conditions. Its core advantages are reflected in three aspects: First, it demonstrates outstanding robustness to extreme weather conditions, with waiting times on snowy days (8.3 ± 1.5 minutes). This is due to the accurate modeling of the probability distribution of sudden changes in passenger flow by Distributed RL and the adaptive exploration mechanism of Noisy Nets; Secondly, the gradual performance degradation is significant, with a lower increase in waiting time from sunny to snowy days, validating the Dueling Network's ability to prioritize resources in key areas. Finally, the response speed to sudden weather events is excellent, with a standard deviation of 1.2 in rainy scenarios compared to 2.8 for DreamerV3, reflecting the Multi-step Learning mechanism's effective capture of latency effects. Among the compared algorithms, Ape-X DQN ranked second due to its distributed sampling, but its performance on snowy days still lagged behind Rainbow DQN, exposing its ensemble approach's insufficient coverage of long-tail weather samples. DreamerV3's performance collapsed in extreme weather conditions (13.6 ± 3.3 minutes on snowy days), demonstrating the vulnerability of model-based algorithms to sudden changes in environmental dynamics. The data further revealed that in moderately disturbed scenarios like fog and strong winds, Rainbow DQN, through its PER mechanism, efficiently replayed historical emergency events, keeping latency fluctuations within ± 1.3 minutes, outperforming traditional methods. This comprehensive advantage stems from the synergy between Rainbow DQN's components: Distributed RL handles uncertainty, Dueling Network optimizes resource allocation, and Multi-step Learning ensures longterm policy stability. Together, they build a scheduling system that adapts to complex weather conditions. Tourist satisfaction is comprehensively measured using complaint rates and wait times. Figure 8 shows the tourist satisfaction levels under different daily visitor volumes.

Figure 8(a) Daily visitor flow of 5,000; Figure 8(b) Daily visitor flow of 10,000; Figure 8(c) Daily visitor flow of 15,000; Figure 8(d) Daily visitor flow of 20,000; Figure 8(e) Daily visitor flow of 25,000; Figure 8(f) Daily visitor flow of 30,000; Figure 8(g) Daily visitor flow of 35,000; Figure 8(h) Daily visitor flow of 40,000; Figure 8(i) Daily visitor flow of 45,000; Figure 8(j) Daily visitor flow of 50,000 The data in Figure 8 reveals the nonlinear relationship between scenic

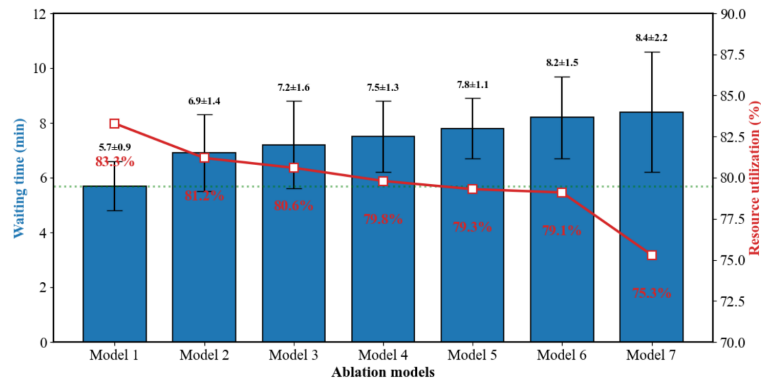


FIGURE 5. Ablation Experiment Results

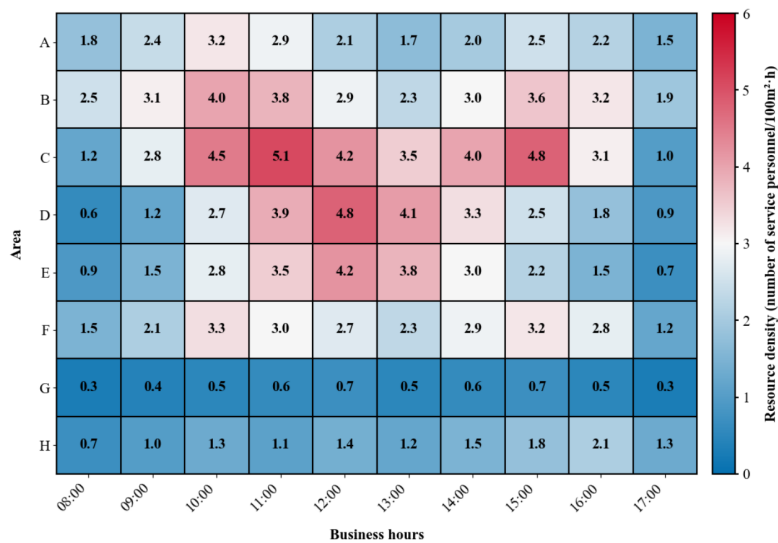


FIGURE 6. Resource Scheduling Visualization

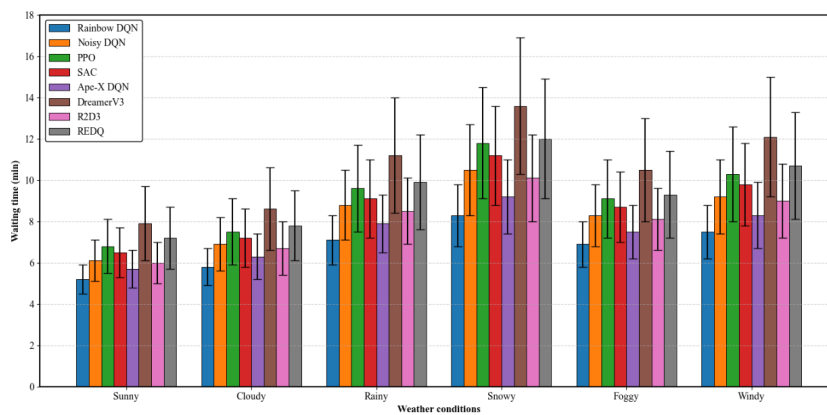


FIGURE 7. Impact of Weather on Visitor Waiting Times

spot service performance and visitor flow, exhibiting three typical characteristics: 1) Key bottleneck identification: The cable car station (Zone B) and the core observation deck (Zone C) exhibited service degradation when passenger

traffic exceeded 25,000. With a passenger traffic of 50,000, waiting times reached 21.0 and 26.0 minutes, respectively, and complaint rates reached 2.25% and 1.96%, respectively. This is directly related to the physical capacity limitations

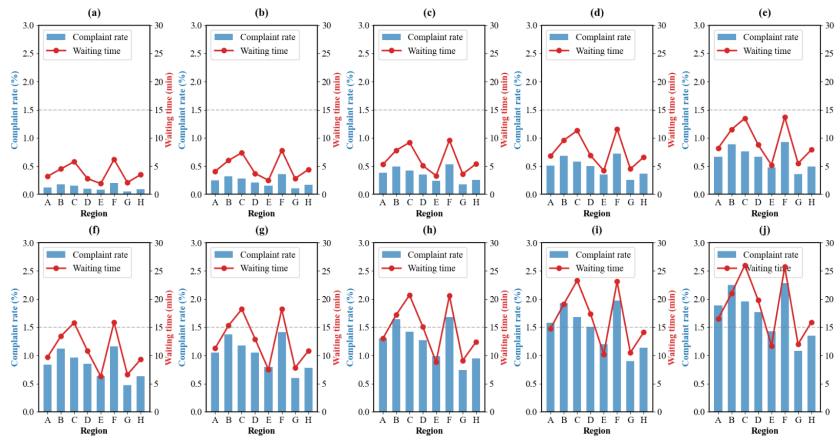


FIGURE 8. Tourist Satisfaction

of the facilities. 2) Differences in service resilience: Zones E and G performed the best, with waiting times (11.7 and 12.0 minutes) and complaint rates (1.43% and 1.08%) significantly lower than those of other zones when passenger traffic reached 50,000. This benefited from the flexible resource scheduling strategy.

3) Nonlinear Growth Pattern: When passenger volume increases from 5,000 to 20,000, the increase in waiting time in each area is relatively low, while the increase is higher when the passenger volume ranges from 20,000 to 50,000, consistent with the critical congestion effect in queuing theory. Through dynamic resource scheduling, this model maintains a complaint rate of 0.05% to 2.28% in each area within a daily passenger volume range of 5,000 to 50,000, validating the model's effectiveness in ensuring visitor satisfaction.

VI. CONCLUSION

This study formalizes the resource allocation problem of smart tourism scenic spots as a Partially Observable Markov Decision Process (POMDP) and constructs a high-dimensional state representation that incorporates spatiotemporal features, mirroring the complex modeling required to optimize manufacturing logistics and quality control across a dispersed supply chain. By integrating six key technologies, including Rainbow DQN (Distributed RL) and Dueling Network, the algorithm effectively addresses the dynamics of sparse rewards and non-stationary environments in high-dimensional state spaces. Experiments, based on the SUMO simulation platform, covering multiple service resource types and complex passenger flow scenarios, validated the algorithm's superiority and generalization capabilities. The research results demonstrate that Rainbow DQN reduces visitor waiting times, outperforming existing methods. It maintains stable performance under extreme weather conditions (such as snow) and peak visitor flow (50,000 people per day). Ablation experiments confirm the key role of each component in improving performance, offering validation similar to component-level testing for intelligent quality control systems in production. This re-

search provides a deployable intelligent decision-making model for smart tourism and promotes the application of artificial intelligence in public services.

AUTHOR CONTRIBUTIONS

Conceptualization – Huayong Xiong and Huan Liu; methodology – Huayong Xiong and Huan Liu; investigation – Huayong Xiong and Huan Liu; writing-original draft preparation – Huayong Xiong and Huan Liu. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

- [1] X. L. Ma, Z. N. Le, J. X. Guo, and Y. F. Yu, "Research on tourist classification of smart scenic spots based on tourist perception: a case study of Tengwangge scenic spot," *Journal of Central China Normal University (Natural Science Edition)*, vol. 59, no. 3, pp. 462-470, 2025, doi: 10.19603/j.cnki.1000-1190.2025.03.017.
- [2] L. Z. Yu, R. Z. Liu, Y. Chen, C. X. Liu, X. R. Hu, Y. Hu, et al., "Exploration of the application of BIM technology in smart tourism: A case study of Taizishan Ecological Tourism Area in Hubei Province," *Journal of Wildland Fire Science*, vol. 43, no. 1, pp. 155, 2025, doi: 10.3969/j.issn.1002-2511.2025.01.036.
- [3] H. H. Zhang, L. Y. Zhao, and S. Y. Zheng, "Research on strategies to improve tourism transportation in existing areas: A case study of Shenzhen Phoenix District," *Urbanism & Architecture/Chengshi Jianzhu*, vol. 22, no. 499, pp. 48, 2025.
- [4] T. Hu, C. Z. Zhang, Y. Zhao, H. G. Xu, and X. D. Cui, "The impact of time distribution of tourist flow in mountain cultural tourism destinations: Big data analysis based on Mount Tai's tourist flow," *Mountain Research* (10082786), vol. 41, no. 3, pp. 435-445, 2023.
- [5] N. Kumar, S. Mittal, V. Garg, and N. Kumar, "Deep reinforcement learning-based traffic light scheduling framework for SDN-enabled smart transportation system," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 3, pp. 2411-2421, 2021, doi: 10.1109/ITITS.2021.3095161.
- [6] C. Shyalika, T. Silva, and A. Karunananda, "Reinforcement learning in dynamic task scheduling: A review," *SN Computer Science*, vol. 1, no. 6, pp. 306, 2020, doi: 10.1007/s42979-020-00326-5.

- [7] L. Li, Wang Ai, Z. T. Wang, and Y. C. Dou, "Research on comprehensive monitoring and operation management system of urban rail transit personnel based on multi-source data fusion," *Railway Computer Applications*, vol. 30, no. 10, pp. 47-53, 2021, doi: 10.3969/j.issn.1005-8451.2021.10.10.
- [8] X. Q. Wang, X. Y. Xu, Y. K. Wu, and J. Liu, "Urban rail short-term passenger flow prediction based on hybrid deep learning model," *Journal of Railway Science & Engineering*, vol. 19, no. 12, pp. 3557, 2022, doi: 10.19713/j.cnki.43-1423/u.T20220158.
- [9] M. Jiang, Y. Zhang, and Y. Zhang, "Optimal electric bus scheduling under travel time uncertainty: A robust model and solution method," *Journal of Advanced Transportation*, vol. 2021, no. 1, Art. no. 1191443, 2021, doi: 10.1155/2021/1191443.
- [10] H. Tang and M. Lu, "UE-Based Digital Twin Engineering Management System for Water Conservancy Scenic Area: A Case of Yijiang River Water Conservancy Scenic Area," *Urbanism & Architecture/Chengshi Jianzhu*, vol. 22, no. 496, pp. 195, 2025, doi: 10.19892/j.cnki.csjz.2025.11.43.
- [11] J. Chen, T. Peng, Y. S. Cao, and K. L. Liu, "Tourist passenger transport demand prediction model driven by social network data," *Journal of Chongqing Jiaotong University (Natural Science Edition)*, vol. 41, no. 10, pp. 41, 2022, doi: 10.3969/j.issn.1674-0696.2022.10.06.
- [12] Q. Xie, M. Lu, and C. S. Xie, "High-temporal-frequency Forecast of Tourist Flow for Tourist Attraction based on LBS and Deep Learning," *Journal of Geo-Information Science*, vol. 25, no. 2, pp. 298-310, 2023, doi: 10.12082/dqxkx.2023.220231.
- [13] S. M. Mortazavi Azad and A. Ramazani, "Smart control of traffic lights based on traffic density in the multiintersection network by using Q learning," *Discover Artificial Intelligence*, vol. 3, no. 1, pp. 39, 2023, doi: 10.1007/s44163-023-00087-z.
- [14] J. Liu, S. Qin, Y. Luo, Y. Wang, and S. Yang, "Intelligent traffic light control by exploring strategies in an optimised space of deep Q-learning," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 6, pp. 5960-5970, 2022, doi: 10.1109/TVT.2022.3160871.
- [15] A. Al-Hilo, M. Samir, C. Assi, S. Sharafeddine, and D. Ebrahimi, "UAV-assisted content delivery in intelligent transportation systems-joint trajectory planning and cache management," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 8, pp. 5155-5167, 2020, doi: 10.1109/TITS.2020.3020220.
- [16] H. Gao, W. Huang, T. Liu, Y. Yin, and Y. Li, "PPO2: Location privacy-oriented task offloading to edge computing using reinforcement learning for intelligent autonomous transport systems," *IEEE transactions on intelligent transportation systems*, vol. 24, no. 7, pp. 7599-7612, 2022, doi: 10.1109/TITS.2022.3169421.
- [17] C. Mu, Y. Shi, N. Xu, X. Wang, Z. Tang, H. Jia, et al., "Multi-objective interval optimization dispatch of microgrid via deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 15, no. 3, pp. 2957-2970, 2023, doi: 10.1109/TSG.2023.3339541.
- [18] Y. Zheng, Q. Hao, J. Wang, C. Gao, J. Chen, D. Jin, et al., "A survey of machine learning for urban decision making: applications in planning, transportation, and healthcare," *ACM Computing Surveys*, vol. 57, no. 4, pp. 1-41, 2024, doi: 10.1145/3695986.
- [19] H. Ge, X. Hu, and M. Cheng, "DHDRDS: A Deep Reinforcement Learning-Based Ride-Hailing Dispatch System for Integrated Passenger-Parcel Transport," *Sustainability*, vol. 17, no. 9, pp. 4012, 2025, doi: 10.3390/su17094012.
- [20] J. Terven, "Deep reinforcement learning: a chronological overview and methods," *AI*, vol. 6, no. 3, pp. 46, 2025, doi: 10.3390/ai6030046.
- [21] R. X. Zhang, C. X. Wu, T. C. Sun, and Z. S. Zhao, "Research progress in deep reinforcement learning and path planning," *Journal of Computer Engineering & Applications*, vol. 57, no. 19, pp. 44, 2021, doi: 10.3778/j.issn.1002-8331.2104-0369.
- [22] A. Correa, A. Jesus, C. Silva, and S. Moniz, "Rainbow Versus Deep Q-Network: A Reinforcement Learning Comparison on The Flexible Job-Shop Problem," *IFAC-PapersOnLine*, vol. 58, no. 19, pp. 870-875, 2024, doi: 10.1016/j.ifacol.2024.09.176.
- [23] Li Mingyang, Xu Keer, Song Zhiqiang, Xia Qingfeng, and Zhou Peng, "A review of multi-agent reinforcement learning algorithms," *Journal of Frontiers of Computer Science & Technology*, vol. 18, no. 8, pp. 1979-1979, 2024, doi: 10.3778/j.issn.1673-9418.2401020.
- [24] L. Tian and H. W. Zhong, "Research on the formation factors and action paths of tourist congestion incidents," *Human Geography*, vol. 36, no. 4, pp. 159-167, 2021, doi: 10.13959/j.issn.1003-2398.2021.04.019.
- [25] X. J. Zhou and S. P. Xi, "Capacity improvement of urban integrated transportation hubs based on resource optimization allocation," *Journal of North China Electric Power University (Social Science Edition)*, vol. 2, no. 5, pp. 53-61, 2021, doi: 10.14092/j.cnki.cn11-3956/c.2021.05.006.
- [26] P. Suanpang, P. Jamjunt, K. Jernsittiparsert, and P. Kaewyong, "Autonomous energy management by applying deep Q-learning to enhance sustainability in smart tourism cities," *Energies*, vol. 15, no. 5, pp. 1906, 2022, doi: 10.3390/en15051906.
- [27] W. Wang, N. Kumar, J. Chen, Z. Gong, X. Kong, W. Wei, et al., "Realizing the potential of the internet of things for smart tourism with 5G and AI," *IEEE network*, vol. 34, no. 6, pp. 295-301, 2020, doi: 10.1109/MNET.011.2000250.
- [28] Q. Sha, L. Ma, and N. Chen, "A Smart Tourism Resource Information Management Platform Based on Multi-Source Data Fusion," *International Journal of High Speed Electronics and Systems*, vol. 34, no. 01, Art. no. 2540134, 2025, doi: 10.1142/S0129156425401342.
- [29] J. W. Wang, G. Q. Wang, Y. Liu, T. Lei, J. Sun, X. Wang, et al., "Spatio-temporal distribution and network structure of red tourism flow in Jinggangshan," *Journal of Natural Resources*, vol. 36, no. 7, pp. 1777-1791, 2021, doi: 10.31497/zrzyxb.20210711.
- [30] S. F. Ding, W. Du, J. Zhang, L. L. Guo, and L. Ding, "Research progress of multi-agent deep reinforcement learning," *Chinese Journal of Computers*, vol. 47, no. 7, pp. 1547-1567, 2024, doi: 10.11897/SPJ.1016.2024.01547.
- [31] A. E. P. Valdiviezo, "El Agente Rainbow en pocas palabras: de RL a la combinacion de mejora de DQN," *Polo del Conocimiento: Revista científico-profesional*, vol. 7, no. 7, pp. 2732-2770, 2022, doi: 10.23857/pc.v7i6.
- [32] Y. Ji, Y. Wang, H. Zhao, G. Gui, H. Gacanin, H. Sari, et al., "Multi-agent reinforcement learning resources allocation method using dueling double deep Q-network in vehicular networks," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 10, pp. 13447-13460, 2023, doi: 10.1109/TVT.2023.3275546.
- [33] J. Liu, Y. H. Wang, L. Wang, and Z. Z. Yin, "Non-cooperative multi-user dynamic power control method based on SumTree sampling combined with Double DQN," *Telecommunication Engineering*, vol. 63, no. 10, pp. 1603, 2023, doi: 10.20079/j.issn.1001-893x.220428004.
- [34] P. Y. Shen, D. M. Wan, and J. X. Li, "The impact of tourism public service quality on tourists' happiness - based on the perspective of tourism experience," *Tourism Guide*, vol. 7, no. 2, pp. 22-54, 2023, doi: 10.12054/lydk.bisu.218.
- [35] C. M. Chuang, "A current travel model: smart tour on mobile guide application services," *Current Issues in Tourism*, vol. 23, no. 18, pp. 2333-2352, 2020, doi: 10.1080/13683500.2019.1631266.