Preference-based Deep Reinforcement Learning for Historical Route Estimation

Boshen Pan[1,4], **Yaoxin Wu**[2], **Zhiguang Cao**[3], **Yaqing Hou**[1,4*], **Guangyu Zou**[1] and **Qiang Zhang**[1,4*]

[1]School of Computer Science and Technology, Dalian University of Technology
[2]Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology
[3]School of Computing and Information Systems, Singapore Management University Singapore
[4]Key Laboratory of Social Computing and Cognitive Intelligence (Dalian University of Technology), Ministry of Education, China
boshenpan@mail.dlut.edu.cn, wyxacc@hotmail.com, zhiguangcao@outlook.com, {houyq, gyzou, zhangq}@dlut.edu.cn

Abstract

Recent Deep Reinforcement Learning (DRL) techniques have advanced solutions to Vehicle Routing Problems (VRPs). However, many of these methods focus exclusively on optimizing distance-oriented objectives (i.e., minimizing route length), often overlooking the implicit drivers' preferences for routes. These preferences, which are crucial in practice, are challenging to model using traditional DRL approaches. To address this gap, we propose a preference-based DRL method characterized by its reward design and optimization objective, which is specialized to learn historical route preferences. Our experiments demonstrate that the method aligns generated solutions more closely with human preferences. Moreover, it exhibits strong generalization performance across a variety of instances, offering a robust solution for different VRP scenarios.

1 Introduction

Vehicle Routing Problems (VRPs) are ubiquitous in the real-world applications, and each problem is unique with its own set of constraints. In addition, the constraints often evolve rapidly due to changes in practical situations.

Classical methods for VRPs can be categorized into exact methods and heuristic methods [Martí and Reinelt, 2022]. The exact methods, while guaranteeing an optimal solution, are computationally expensive. On the other hand, heuristic methods provide good solutions more efficiently depending on problem-specific rules and expert knowledge, but cannot guarantee optimality. Both exact and heuristic methods tend to struggle in dynamic environments, such as those influenced by weather or traffic conditions. Recently, learning-to-optimize techniques have achieved significant success in solving VRPs such as the traveling saleman problem (TSP) and the capacitated vehicle routing problem (CVRP) [Bengio *et al.*, 2021; Kool *et al.*, 2018; Kwon *et al.*, 2020; Zhang *et al.*, 2023; Bello *et al.*, 2016; Wu *et al.*, 2024; Zhou *et al.*, 2023]. They eliminate the need for extensive hand-crafted heuristics and domain-specific expertise, enabling a trained model to perform well across a range of instances.

Existing literature on learning to optimize VRPs can be broadly categorized into two paradigms: constructive solvers and iterative solvers. Constructive solvers learn policies to construct solutions from scratch in an end-to-end autoregressive manner, while iterative solvers start with an initial solution and progressively refine it towards better solutions. Generally, constructive solvers efficiently achieve strong performance, whereas iterative solvers excel in exploring near-optimal solutions within a longer time budget.

However, current learning-to-optimize models for VRPs focus on optimizing metrics such as route distance and travel time. However, these metrics are often impractical in real-world scenarios, and the optimal solutions may fail to satisfy real requirements, such as drivers' preferences for certain paths [Ceikute and Jensen, 2013]. In practice, some navigation software typically suggest the shortest or fastest routes, yet drivers frequently modify these suggestions to retrieve, adjust, and reuse their preferred paths. Through these modifications, drivers are essentially optimizing the paths according to their own preference objectives. The preferences are often subjective and difficult to be formalized into constraints [Toledo *et al.*, 2013]. For example, drivers might prefer routes with a lower probability of traffic congestion, fewer traffic lights, smoother road surfaces, or certain service stations prioritized based on their schedules. The preference-driven objectives pose a key challenge in real-world VRP applications, i.e., the preferences are always subjective and evolve with the dynamics of practical situations. Static learning-to-optimize models or traditional methods often fail to fully capture drivers' preferences and their adjustments to routes. These subjective and ever-changing factors are precisely what make VRP a highly complex and challenging combinatorial optimization problem.

A recent learning-to-optimize approach is proposed to learn drivers' preferences by estimating the transition probabilities between stops based on their historical routes [Mandi

*Yaqing Hou and Qiang Zhang are the Corresponding authors.

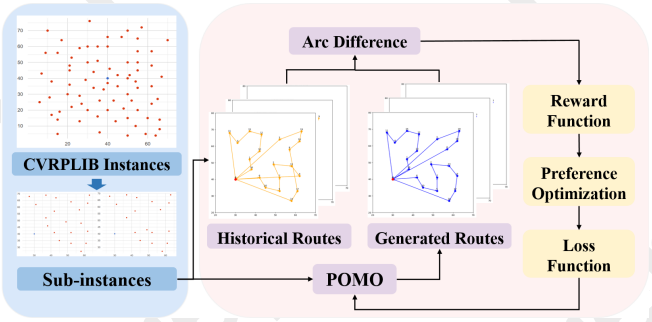

Figure 1: The preference-aware deep reinforcement learning framework.

et al., 2021]. Although this method effectively captures certain aspects of driver preferences, it is heavily based on meticulously labeled historical data, making it challenging to generalize to new and diverse problem instances.

To address the problem, we propose a preference-based deep reinforcement learning framework as illustrated in Figure 1. This framework estimates human preferences from historical route data to generate solutions that align with these preferences. Our main contributions include: 1) We designed a reward signal based on Arc Difference to quantify the similarity between generated routes and historical routes; 2) We proposed a novel optimization objective combining preference loss and reinforcement loss, effectively enhancing the adaptability of the policy network in preference learning; 3) We validated significant advantages of the framework in generating high-quality solutions and favorably generalizing to difference scenarios.

2 Related Work

2.1 Traditional VRPs

Since Dantzig and Ramser [Dantzig and Ramser, 1959] first introduced the VRP to optimize fuel delivery routes for truck fleets, VRP has become one of the most economically significant combinatorial optimization problems, widely applied among distributors and logistics companies. Traditional VRP focuses on minimizing operational costs [Hu *et al.*, 2009], e.g., travel time [Lecluyse *et al.*, 2009], fuel consumption, or carbon emission [Xiao *et al.*, 2012; Peng and Wang, 2009]. Early research concentrated on CVRP, which assumes fixed vehicle capacities and aims to minimize total travel distance, providing a foundation for more complex real-world routing scenarios [Laporte, 2007].

With advances in optimization methods, research has shifted towards Rich VRP variants [Caceres-Cruz *et al.*, 2014; Drexl, 2012], which take into account multiple objectives, uncertainty, and real-world constraints such as inventory, environment, energy, and driver-specific requirements [Mor and Speranza, 2022]. These additional constraints and objectives significantly increase the complexity of VRP, posing challenges to the development of more effective optimization algorithms.

As a human-centered objective, driver preferences further complicate the VRP owing to its subjective and dynamic properties, encompassing diverse preferences such as avoiding high-traffic areas, reducing the number of stops at traffic lights, or prioritizing smoother road surfaces. Unlike explicit constraints like capacity or time windows, these preferences are difficult to handle, or if possible, it is hard to formalize them into weights or decision rules. As Potvin et al. found, it is often easier to collect examples and historical solutions than to derive explicit rules from route planners [Potvin *et al.*, 1993]. This makes preference-based optimization an emerging and crucial aspect of VRP research.

2.2 Preference Learning in VRPs

Driver preferences can be integrated into the optimization process by incorporating them into the objective function, typically formalized as a multi-objective VRP [Jozefowiez *et al.*, 2008] and solved through weighted sums or multi-objective evolutionary algorithms to find Pareto optimal solutions [Schaffer, 2014]. However, the optimization with implicit preferences of drivers is a significant challenge, since they cannot be explicitly formalized in practice.

To address the challenge, Canoy et al. [2019] proposed a different perspective by introducing a Markov model to learn drivers' preferences. This approach eliminates the need for explicitly specifying preference constraints or sub-objectives. Instead, the Markovian model learns preferences directly from historical routes (i.e., plans) that were manually adjusted by human planners to better meet real-world requirements after being generated by off-the-shelf solvers.

Previous research on learning driver preferences primarily focused on scenarios with a single origin and destination. For example, TRIP [Letchner *et al.*, 2006] leverages historical GPS data to infer drivers' preferences by comparing their travel time ratios to average travel times, thereby generating routes that closely mimic those preferred by drivers. Similarly, Funke et al. [2016] deduced drivers' preferences from GPS traces and encoded them into the weights of a linear program, which is subsequently optimized to provide route suggestions. Guo et al. [2020] improved solution quality by considering diverse routing preferences that vary depending on contextual factors. While these methods explicitly represent driver preferences, Mandi et al. [2021] assume that preferences can be expressed probabilistically as utilities or likelihoods of arcs in the graph, providing a more flexible framework for capturing implicit preferences. They built upon the maximum likelihood routing framework proposed by Canoy et al. [2019], which models transition probabilities between stops as explicit preferences of drivers or planners to identify route with the maximum utility. To incorporate contextual features, Mandi et al. extended the framework by introducing a neural network model that leverages both historical and contextual information for more accurate transition probability estimation. Their approach achieved higher solution quality compared to the method of [Canoy and Guns, 2019].

Although the approach by Mandi et al. achieved notable success, it relies heavily on meticulously labeled historical data to provide the model with diverse information. Furthermore, the trained model is effective only for the specific instance, making it difficult to generalize to broader instances.

3 Preliminaries

3.1 Problem Description

Capacitated Vehicle Routing Problem (CVRP) is defined on a graph $G = (V, A)$, where $V = \{0, 1, ..., N\}$ is the vertex set and $A = \{(i, j) : i, j \in V, i \neq j\}$ represents the set of directed arcs. Vertex 0 denotes the depot, while vertices $\{1, ..., N\}$ represent customer nodes. Each arc (i, j) has an associated cost c_{ij}, which could represent distance, travel time, or a combination of relevant metrics. Each customer $i \in V$ has a non-negative demand q_i, while the deopt 0 has no demand, i.e., $q_0 = 0$. A fleet of m identical vehicles, each with a capacity Q, is stationed at the depot. The objective of CVRP is to determine a set of least-cost routes for the fleet, such that the following constraints are satisfied:

1. Each customer $i \in V$ is visited exactly once by exactly one vehicle.
2. The route of each vehicle starts and ends at the depot.
3. The total demand of customers assigned to a route does not exceed the vehicle capacity Q.

The solution to the CVRP is represented as a set of routes $\tau = \{\tau_1, ..., \tau_m\}$, where each route τ_k is a sequence of nodes $\{v_0, v_{k1}, ..., v_{kp}, v_0\}$ visited by a vehicle, starting and ending at the depot. The cost function $c(\tau)$ computes the total cost of all routes. The objective is to find the optimal set of routes τ^* that minimizes the total cost [Lau and Liang, 2002; Munari *et al.*, 2016; Yu *et al.*, 2017], i.e., $\arg\min_{\tau \in \Phi} c(\tau)$. Φ denotes the set of all feasible solutions satisfying the problem constraints.

In practical applications, CVRP solutions often need to consider real-world complexities, such as time windows, driver preferences, and dynamic constraints. These factors further complicate the optimization process, making CVRP a critical combinatorial optimization problem in logistics, transportation, and supply chain management.

3.2 Deep Reinforcement Learning for VRP

Deep Reinforcement Learning (DRL) trains an agent to maximize the cumulative reward by interacting with an environment and receiving reward signals. In VRPs, state transitions are typically modeled as a deterministic process. One commonly used policy gradient method is REINFORCE [Sutton, 2018], whose update rule is given by:

$$\begin{aligned}\nabla_\theta J(\theta) &= E_{x\sim D, \tau\sim\pi_\theta(\tau|x)}[(r(x,\tau) \\ &\qquad - b(x))\nabla_\theta \log \pi_\theta(\tau|x)] \\ &\approx \frac{1}{|D|}\sum_{x\in D}\frac{1}{|S_x|}\sum_{\tau\in S_x}[(r(x,\tau) \\ &\qquad - b(x))\nabla_\theta \log \pi_\theta(\tau|x)]\end{aligned}$$

where D denotes the dataset of problem instances, $x \in D$ represents an instance, and S_x is the set of sampled solutions (or trajectories) for x. The reward function $r(x, \tau)$ is defined by objective function, and $b(x)$ is the baseline function used to calculate the advantage function $A(x, \tau) = r(x, \tau) - b(x)$, which helps reduce the variance of the gradient estimator. The policy $\pi_\theta(\tau|x)$ defines a distribution over trajectories $\tau = (a_0, a_1, \ldots, a_T)$, where each trajectory is a sequence of actions generated by the policy as $\pi_\theta(\tau|x) = \prod_{t=0}^{T} \pi_\theta(a_t|s_t)$. The initial state s_0 is determined by x, and the state s_t at time step t depends on the previous state and action (e.g., $s_t = f(s_{t-1}, a_{t-1})$). The action a_t is selected by the policy based on the state s_t.

Unlike popular RL environments such as Atari [Bellemare *et al.*, 2013] and Mujoco [Todorov *et al.*, 2012], which provide diverse and strong reward signals, VRPs present unique challenges. As the policy improves, the differences in reward signals between solutions become increasingly subtle. Specifically, the agent often encounters solutions with minimal reward differences, i.e., $|r(x, \tau) - b(x)| < \epsilon$, where ϵ is a small constant. This results in negligible updates to the policy objective $J(\theta)$, which heavily relies on the advantage function $A(x, \tau) = r(x, \tau) - b(x)$. Consequently, the policy struggles to escape local optima, particularly during the later stages of training.

Moreover, DRL for VRPs aims to maximize the expected optimal reward during inference. There is an inconsistency between the training objective (which optimizes expected values of rewards) and the inference objective (which seeks to maximize the best possible reward) can degrade performance.

$$E_{x\sim D}\left[\max_{\tau\sim\pi_\theta(\tau|x)} r(x,\tau)\right] \neq E_{x\sim D}\left[E_{\tau\sim\pi_\theta(\tau|x)} r(x,\tau)\right]$$

Since inference only considers the best solutions, when the advantage function $A(x, \tau)$ approaches zero, RL methods struggle to differentiate among solutions and fail to emphasize optimality. Therefore, it is essential to construct a more stable reward signal that highlights the optimality of solutions during training.

4 Methodology

In this section, we first introduce Preference-based Reinforcement Learning (PbRL) and explain how utilize this technique to extract human drivers' preferences from historical trajectory data. By learning these preferences, our method guides the reinforcement learning policy network to generate solutions that align with human preferences.

4.1 Preference-Based Reinforcement Learning

In the PbRL [Wirth *et al.*, 2017] framework, the agent's optimization objective is to learn a reward function from a preference dataset rather than receive a reward signal directly by interacting with the environment. Our approach extracts human preference information from historical route data and uses it to train the policy network such that the generated routes conform to these preferences.

Suppose that we have access to a preference dataset $D_p = \{(\tau_1, \tau_2, y)\}$, where each triplet consists of two trajectories τ_1 and τ_2 along with a preference label $y \in \{0, 1\}$. When τ_1 is preferred over τ_2, $y = 1$; otherwise, $y = 0$. These preferences are assumed to be generated by an underlying reward function $\hat{r}(x, \tau)$, which encodes the rewards between trajectories.

To map the reward difference to preference probability, we employ classical preference models such as the Bradley-Terry model [Bradley and Terry, 1952] or the Thurstone

model [David, 1963]. These models transform the reward difference between trajectories into a pairwise preference probability distribution:

$$p^*(\tau_1 \succ \tau_2 \mid x) = f(\hat{r}(x, \tau_1) - \hat{r}(x, \tau_2)) \quad (1)$$

In the BT model, the sigmoid function is utilized as the mapping function, defined by

$$\sigma(x) = \frac{1}{1 + e^{-x}},$$

whereas the Thurstone model adopts the cumulative distribution function $\Phi(x)$ of the standard normal distribution for $f(\cdot)$. By defining this relationship, the task of learning the reward function $\hat{r}_\phi(x, \tau)$ can be approached as a binary classification problem. The goal is to optimize the parameters by maximizing the likelihood of the observed preference data, which is mathematically represented as:

$$\max_\phi E_{(\tau_1, \tau_2, y) \sim D_p} [y \log p_\phi(\tau_1 \succ \tau_2)]$$

This formulation effectively transforms the reward learning process into a supervised learning task, leveraging the binary nature of preference judgments to guide the optimization of the reward function.

4.2 Arc Difference and Reward Function

To quantify the similarity between a generated route and historical routes, and to connect this similarity to human preferences, we adopt the concept of arc difference [Mandi *et al.*, 2021]. The arc difference measures the structural disparities between the generated route and historical routes, using arc sets as the basis for comparison. A smaller arc difference indicates that the generated route is more similar to historical route. Since humans typically prefer routes resembling those traveled historically, the arc difference serves as an important indicator of route generation quality and reflect the human preference to certain degree.

Based on the arc difference, we define a reward function that quantifies the similarity between the generated routes and historical routes. Given an instance x, the reward function for route τ is defined as:

$$reward = -(1 - \beta) * distance - \beta * ad(x, \tau) \quad (2)$$

where $distance$ represents the route length optimized in traditional reinforcement learning methods. Parameter β is a hyperparameter that controls the influence of the arc difference on the reward function. Specifically, when $\beta = 0$, the reward function focuses solely on minimizing the distance, aligning with the traditional reinforcement learning objective without considering similarity to historical routes. And $ad(x, \tau)$ denotes the arc difference between τ and the historical routes τ^*, which is defined as:

$$ad(x, \tau) = \frac{1}{N} \sum_{i=1}^{N} |E(\tau) \setminus E(\tau_i^*)| \quad (3)$$

The goal of designing this reward signal is to make the generated route similar to historical routes of a driver, i.e., aligning them with the objective of optimizing human preferences. The reward function can be used to a policy network for learning routes that are similar to historical routes biased towards human preference.

4.3 Policy Optimization with Preference

A key aspect of our method is to convert quantitative reward signals into qualitative preference information. This conversion not only stabilizes the learning process but also reduces reliance on numerical reward signals, meanwhile emphasizing the relative superiority among solutions. In our approach, we train the policy network with the preference information extracted from historical routes to generate routes that conform to human preferences. Unlike traditional reinforcement learning methods that focus solely on operational metrics, such as travel time or distance, our objective is to capture relative preferences among solutions, thereby making the generated routes better reflect human-centered decision-making processes.

Similar to common practices in reinforcement learning, applying the arc difference as a reward signal encounters a significant challenge: the state and action spaces grow exponentially with the problem size, leading to inefficient exploration. A common solution is to include an entropy regularization term $H(\pi_\theta)$ during optimization to balance exploitation and exploration, encouraging the network to explore a wider range of route choices:

$$J(\theta) = E_{x \sim D, \tau \sim \pi_\theta(\tau|x)}[r(x, \tau)] + \alpha H(\pi_\theta(\tau \mid x)) \quad (4)$$

where α controls the strength of entropy regularization term, and $H(\pi_\theta(\tau \mid x)) = -\sum_\tau \pi_\theta(\tau \mid x) \log \pi_\theta(\tau \mid x)$ is the entropy of the policy, designed to encourage policy diversity and exploration. Based on previous work [Ziebart *et al.*, 2008; Haarnoja *et al.*, 2017] , the analytical form of the optimal policy in Eq. 4 is:

$$\pi(\tau \mid x) = \frac{1}{Z(x)} \exp\left(\frac{1}{\alpha} r(x, \tau)\right) \quad (5)$$

where $Z(x)$ is the normalization factor defined as:

$$Z(x) = \sum_\tau \exp\left(\frac{1}{\alpha} r(x, \tau)\right)$$

The normalization factor $Z(x)$ represents the weighted sum of all possible trajectories τ, ensuring that $\pi(\tau \mid x)$ is a valid probability distribution and assigns appropriate probability weights to each route. This analytical form indicates that the probability assigned to each trajectory by the optimal policy is determined by its reward $r(x, \tau)$; the higher the reward, the higher the probability of generating that trajectory. This expression not only provides a solid theoretical foundation for policy optimization but also reveals the close relationship between reward signals and policy distribution. Similar methods appear in other related studies, such as the reparameterization approach in a KL-regularized objective by Rafailov et al. [2024], the application in an inverse reinforcement learning framework by Hejna & Sadigh [2024]. These works all demonstrate that explicitly associating reward signals with the policy probabilities is an effective means of designing reinforcement learning objectives.

Preference-Based Reward

Building on the above analysis, this form further supports the reparameterization of the reward function by expressing it as

a function of the policy probability, thereby guiding the design of the optimization objective more directly [Lin *et al.*, 2025]. Specifically:

$$\hat{r}(x,\tau) = \alpha \log \pi(\tau \mid x) + \alpha \log Z(x) \tag{6}$$

The direct association between the reward function and the policy probability provides theoretical support for preference-based route optimization, while making the optimization process more tractable. From Eq. 6, we can directly relate the optimal policy π^* to the reward function r, thereby revealing the preference relationships among trajectories. Specifically, the preference between two routes τ_1 and τ_2 can be mapped into a pairwise preference distribution via their reward difference:

$$p^*(\tau_1 \succ \tau_2 \mid x) = f\Big(\alpha\big[\log \pi(\tau_1 \mid x) - \log \pi(\tau_2 \mid x)\big]\Big) \tag{7}$$

A key feature of this representation is that it naturally avoids the computational complexity of calculating the normalization factor $Z(x)$. When considering reward differences, $Z(x)$ is a constant term with respect to a particular pair of routes and is canceled out when computing the preference probability.

By leveraging this relationship, we successfully transform the numerical reward signal into a qualitative preference representation based on the policy probabilities. This transformation not only effectively captures the relative merits of different routes but also better reflects human preferences in route selection, thus further guiding the policy network to generate solutions more similar to historical routes.

4.4 Generating Conflict-Free Preference Labels

To produce conflict-free preference labels, we use the baseline reward function $r(x,\tau)$, which is defined in Eq. 2, as a physical measure of route quality and generate preference labels $y = 1[\cdot] : R \to \{0,1\}$ through pairwise comparisons. This reward-based preference label generation ensures consistency and transitivity of preference relationships across the dataset.

It is worth noting that, unlike common approaches in previous DRL methods that adjust the reward signal via affine transformations, our method fully exploits the affine-invariance of preference labels. Specifically, the indicator function remains invariant under the following affine transformation:

$$1\big[k \cdot r(x,\tau_1) + b > k \cdot r(x,\tau_2) + b\big] = 1\big[r(x,\tau_1) > r(x,\tau_2)\big] \tag{8}$$

where $k > 0$ is any positive constant and b is any real value. This property indicates that preference labels remain independent of the scale and shift of the reward function, consistently focusing on the relative superiority of solutions. This approach not only improves the stability of the learning process but also reinforces the consistency of policy preferences in the optimization objective.

4.5 Preference-Based Learning Objective

We combine the arc difference with the policy probability to design a preference-based learning objective:

$$\begin{aligned} J(\theta) = & E_{x\sim D,(\tau_1,\tau_2)\sim\pi_\theta(\cdot|x)} \\ & [1[r(x,\tau_1) < r(x,\tau_2)] \cdot \log p_\theta(\tau_1 \succ \tau_2 \mid x)] \end{aligned} \tag{9}$$

where $r(x,\tau_1)$ and $r(x,\tau_2)$ are defined as above, representing the arc difference between the generated routes τ_1, τ_2 and the historical routes, respectively. Meanwhile, $p_\theta(\tau_1 \succ \tau_2 \mid x)$ is the preference probability assigned by the policy network π_θ for routes τ_1 and τ_2. By emphasizing the relative preference among routes, this objective function guides the policy network to generate solutions that more closely resemble the historical routes, thereby maximizing consistency with human preferences.

5 Experiments

5.1 Setup

We evaluated the performance of the proposed preference-driven deep reinforcement learning framework on the classic CVRP. The instances used in the experiments were sourced from CVRPLIB. The experiments were conducted on a computer equipped with an Intel(R) Core(TM) i5-13400 2.5GHz CPU, 32.0GB RAM, and an NVIDIA GeForce RTX 4090 GPU, with model training and inference carried out using POMO [Kwon *et al.*, 2020], which is widely regarded as a classic benchmark algorithm in the VRP field. Our code is publicly available.[1]

Benchmark Algorithms

To comprehensively assess the effectiveness of the proposed method, we first validated its performance advantage in generating solutions that are more aligned with human preferences by comparing it with the benchmark algorithms proposed by Mandi et al. [2021](denoted as Neural Net(NN) in the results table) and Canoy et al. [2019](denoted as Markov in the results table), as well as the POMO [Kwon *et al.*, 2020], which serves as a widely regarded baseline in solving VRP. Our policy network is also based on the POMO. Furthermore, we evaluated the impact of key components, such as the reward function and optimization objectives, on model performance by gradually removing these elements. Finally, we tested the performance of the trained model on different types of instances, exploring its adaptability to unseen instances, thereby validating the generalization and stability of the model.

Generation of Simulated Historical Routes

Training deep reinforcement learning models typically requires large amounts of data, but acquiring real historical routes data is expensive and difficult to cover all possible scenarios. To address this issue, we adopted a method for simulating historical route generation by randomly sampling CVRP instances and perturbing the routes, thereby constructing a diverse and representative training dataset.

Specifically, we selected a benchmark instance containing 73 customer nodes (excluding the depot) from CVRPLIB as the initial dataset. From this instance, we randomly sampled sub-instances containing 20 customer nodes and randomly assigned demand values to each customer node. Each training batch consisted of 50 sub-instances, ensuring the diversity of the data. For each sub-instance, we first calculated

[1]https://github.com/pandarking/Preference-based-DRL

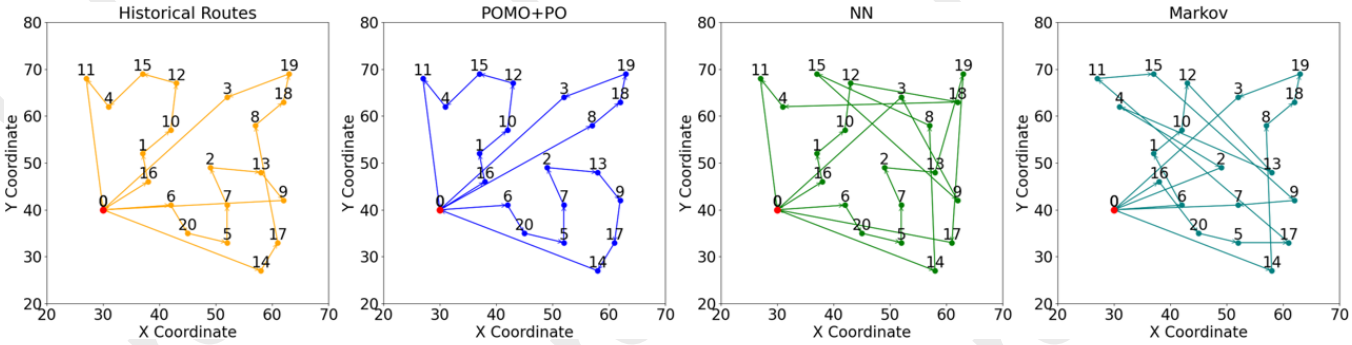

Figure 2: Route Comparison. From left to right: historical route, route generated by POMO+PO, route generated by the neural network (NN) [Mandi *et al.*, 2021], and route generated by the Markov method [Canoy and Guns, 2019].

Method	AD mean	AD min	AD% mean
Neural Net(NN)	8.81	2	36.9
NN(without week)	8.82	2	37.0
NN(LSTM)	12.80	2	53.6
NN(different layers for stops)	9.51	2	39.8
NN(without past)	13.39	2	56.1
NN(without dist)	8.71	2	36.5
NN(without Markov)	8.81	2	36.9
NN(without stop)	8.75	2	36.6
Markov	13.21	2	55.3
POMO	9.15	5	28.6
POMO+PO(ours)	**3.15**	**0**	**12.5**

Table 1: Experiment results on CVRP

its absolute distance matrix $D = [d_{ij}]$, consisting of pairwise distances between all available nodes. A random matrix $E = [e_{ij}]$ was then introduced, with elements e_{ij} sampled from a uniform distribution $U(0.8, 1.2)$. By element-wise multiplying the distance matrix D with E, we generated a preference matrix $P = D \odot E$ to simulate human preferences in route selection.

After obtaining the preference matrix P, we used a heuristic algorithm to solve an initial historical route for each sub-instance. To further simulate the diversity of human route selections in real-world scenarios, we introduced random perturbations to each initial route solution, thereby generating additional historical routes. Specifically, each sub-instance generated 30 historical routes, including the initial solution and its 29 randomly perturbed versions. These perturbed routes reflect possible changes in human preferences by adjusting the order of nodes and the structure of the route.

Through the generation process, we constructed a training dataset with diverse and dynamic preferences, ensuring both the effectiveness of model training and a more realistic representation of the operational characteristics of CVRP instances.

Evaluation Metrics

To quantitatively assess the effectiveness of the proposed method, we utilize the following evaluation metrics:

Method	AD mean	AD min	AD% mean
POMO+PO β=0	7.96	2	31.3
POMO+PO β=0.1	6.78	2	26.7
POMO+PO β=0.9	5.77	1	22.9
POMO+PO β=1	**3.15**	**0**	**12.5**

Table 2: Experiment results on different β

Method	AD mean	AD min	AD% mean
POMO-RL	7.54	2	29.8
POMO+PO	**3.15**	**0**	**12.5**

Table 3: Experiment results on different optimization objective

- **AD mean**: The average arc difference between generated and historical routes, reflecting overall alignment with human preferences.
- **AD min**: The minimum arc difference observed, indicating the model's best-case performance in matching historical routes.
- **AD% mean**: The normalized average arc difference as a percentage of total arcs, enabling fair comparisons across instances of varying sizes.

5.2 Comparison with Preference-Based Routing Methods

Table 1 summarizes the performance of several algorithms across key evaluation metrics, including average arc difference (AD mean), minimum arc difference (AD min), and the average arc difference ratio (AD% mean). In this comparison, we involve the current best preference-based learning method for VRPs [Mandi *et al.*, 2021], which includes different architectural variations such as LSTM-based models and models trained without specific contextual features (e.g., weekday or distance information), and models where each stop has a different network architecture. We also compare our method with the preference-based learning method [Canoy and Guns, 2019] as well as POMO [Kwon *et al.*, 2020] which is a typical DRL based method for VRPs. The results demonstrate that our proposed method (POMO+PO) significantly outperforms

Method	AD mean	AD min	AD% mean	Method	AD mean	AD min	AD% mean
Neural Net(NN)	8.81	2	36.9	NN(without distance)	8.71	2	36.5
NN*	15.74	10	64.3	NN(without distance)*	16.01	12	65.4
NN(without weekday)	8.82	2	37.0	NN(without Markov)	8.81	2	36.9
NN(without weekday)*	15.90	11	64.9	NN(without Markov)*	15.85	10	64.8
NN(LSTM)	12.80	2	53.6	NN(without stop info)	8.75	2	36.6
NN(LSTM)*	17.18	12	70.2	NN(without stop info)*	15.96	10	65.2
NN(different layers for stops)	9.51	2	39.8	Markov	13.21	2	55.3
NN(different layers for stops)*	17.03	11	69.6	Markov*	17.54	13	71.6
NN(without past)	13.39	2	56.1	**POMO+PO**	**3.15**	**0**	**12.51**
NN(without past)*	17.34	12	70.8	**POMO+PO***	**6.68**	**1**	**26.5**

Table 4: Generalization results

Method	AD mean	AD min
NN(without dist)	15.97	9
NN(without stop)	15.93	9
POMO+PO	**8.82**	**4**

Table 5: Generalization results with unseen 100-nodes instances

all baselines. Notably, POMO+PO achieves an AD min of 0, indicating its ability to perfectly align with historical route for certain instance.

Figure 2 illustrates the comparison between the routes generated by different methods and the historical route, highlighting that our approach produces route most similar to the historical route. These findings underline the effectiveness of our reward function and preference optimization strategy in generating solutions that better align with human preferences, surpassing both the DRL baseline POMO and preference-based learning methods.

5.3 Evaluation of Reward Function and Policy Learning Objectives

The ablation study investigates the impact of the designed reward function and optimization objectives on model performance. As described in Eq. 2, to evaluate the effect of the parameter β, we set its values to 0, 0.1, 0.9, and 1 while keeping the other configurations unchanged. The results are summarized in Table 2. It is evident that as β decreases, the model's performance degrades. Specifically, when $\beta = 0$, where only the traditional route distance is optimized, the performance reaches its lowest point. This demonstrates that models solely optimized for distance fail to effectively generate solutions similar to historical paths.

Furthermore, Table 3 examines the impact of different optimization objectives on the performance of the policy network. We compare two objectives: the traditional reinforcement learning objective (POMO-RL) and an objective incorporating preference optimization (POMO+PO). The results reveal that POMO+PO significantly outperforms POMO-RL. This indicates that relying solely on traditional reinforcement learning objectives is insufficient to fully capture human preferences. By introducing preference optimization objectives, POMO+PO markedly enhances the policy network's ability to learn human preferences, resulting in routes that better align with them.

5.4 Generalization Analysis

To evaluate the generalizability of our proposed method, we tested the trained model on previously unseen CVRPLIB instances. As shown in Table 4, our method (POMO+PO) maintained strong performance even on unseen instances, consistently generating routes that closely resemble historical routes. In contrast, the methods proposed by Mandi et al. [2021] and Canoy et al. [2019] experienced significant performance degradation when applied to other instances, with the average AD ratio of generated routes exceeding 64%. Additionally, the methods tested on unseen instances are marked with a * in the table, while those on the original instances are not. We also conducted additional experiments on larger-scale CVRPLIB instances and compared our method against previously well-performing models(Table 5). These results demonstrate that our method, by incorporating preference-based reward functions and optimization frameworks, exhibits superior generalization capability in producing human-preference-aligned solutions. It remains robust and adaptable even in dynamic and diverse VRP scenarios.

6 Conclusion

This paper introduced a preference-based deep reinforcement learning framework for solving VRPs, emphasizing alignment with human preferences derived from historical routes. Unlike traditional methods focused solely on minimizing distance or time, our approach incorporates a novel reward function using arc difference and preference-based optimization objective to guide the policy network. Through extensive experiments, including ablation studies and generalization tests, we demonstrated that our method outperforms existing preference-based routing methods. Our model excelled in generating routes matching historical routes while maintaining adaptability to unseen instances.

The proposed framework integrates human preferences to enhance the practicality of routing solutions, offering a new perspective for tackling complex VRPs. Future work will focus on refining the framework for higher performance and exploring its application to broader optimization challenges, such as time-dependent, large-scale VRPs and real-time dynamic scenarios, enhance its practical relevance in logistics and transportation systems.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62372081, the Young Elite Scientists Sponsorship Program by CAST under Grant 2022QNRC001, the Liaoning Provincial Natural Science Foundation Program under Grant 2024010785-JH3/107, the Dalian Science and Technology Innovation Fund under Grant 2024JJ12GX020, the Dalian Major Projects of Basic Research under Grant 2023JJ11CG002 and the 111 Project under Grant D23006. This research was supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant.

References

[Bellemare *et al.*, 2013] Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

[Bello *et al.*, 2016] Irwan Bello, Hieu Pham, Quoc V Le, Mohammad Norouzi, and Samy Bengio. Neural combinatorial optimization with reinforcement learning. *arXiv preprint arXiv:1611.09940*, 2016.

[Bengio *et al.*, 2021] Yoshua Bengio, Andrea Lodi, and Antoine Prouvost. Machine learning for combinatorial optimization: a methodological tour d'horizon. *European Journal of Operational Research*, 290(2):405–421, 2021.

[Bradley and Terry, 1952] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[Caceres-Cruz *et al.*, 2014] Jose Caceres-Cruz, Pol Arias, Daniel Guimarans, Daniel Riera, and Angel A Juan. Rich vehicle routing problem: Survey. *ACM Computing Surveys (CSUR)*, 47(2):1–28, 2014.

[Canoy and Guns, 2019] Rocsildes Canoy and Tias Guns. Vehicle routing by learning from historical solutions. In *Principles and Practice of Constraint Programming: 25th International Conference, CP 2019, Stamford, CT, USA, September 30–October 4, 2019, Proceedings 25*, pages 54–70. Springer, 2019.

[Ceikute and Jensen, 2013] Vaida Ceikute and Christian S Jensen. Routing service quality–local driver behavior versus routing services. In *2013 IEEE 14th international conference on mobile data management*, volume 1, pages 97–106. IEEE, 2013.

[Dantzig and Ramser, 1959] George B Dantzig and John H Ramser. The truck dispatching problem. *Management science*, 6(1):80–91, 1959.

[David, 1963] Herbert Aron David. *The method of paired comparisons*, volume 12. London, 1963.

[Drexl, 2012] Michael Drexl. Rich vehicle routing in theory and practice. *Logistics Research*, 5:47–63, 2012.

[Funke *et al.*, 2016] Stefan Funke, Sören Laue, and Sabine Storandt. Deducing individual driving preferences for user-aware navigation. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 1–9, 2016.

[Guo *et al.*, 2020] Chenjuan Guo, Bin Yang, Jilin Hu, Christian S Jensen, and Lu Chen. Context-aware, preference-based vehicle routing. *The VLDB Journal*, 29:1149–1170, 2020.

[Haarnoja *et al.*, 2017] Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *International conference on machine learning*, pages 1352–1361. PMLR, 2017.

[Hejna and Sadigh, 2024] Joey Hejna and Dorsa Sadigh. Inverse preference learning: Preference-based rl without a reward function. *Advances in Neural Information Processing Systems*, 36, 2024.

[Hu *et al.*, 2009] Xiangpei Hu, Zheng Wang, Minfang Huang, and Amy Z Zeng. A computer-enabled solution procedure for food wholesalers' distribution decision in cities with a circular transportation infrastructure. *Computers & Operations Research*, 36(7):2201–2209, 2009.

[Jozefowiez *et al.*, 2008] Nicolas Jozefowiez, Frédéric Semet, and El-Ghazali Talbi. Multi-objective vehicle routing problems. *European journal of operational research*, 189(2):293–309, 2008.

[Kool *et al.*, 2018] Wouter Kool, Herke Van Hoof, and Max Welling. Attention, learn to solve routing problems! *arXiv preprint arXiv:1803.08475*, 2018.

[Kwon *et al.*, 2020] Yeong-Dae Kwon, Jinho Choo, Byoungjip Kim, Iljoo Yoon, Youngjune Gwon, and Seungjai Min. Pomo: Policy optimization with multiple optima for reinforcement learning. *Advances in Neural Information Processing Systems*, 33:21188–21198, 2020.

[Laporte, 2007] Gilbert Laporte. What you should know about the vehicle routing problem. *Naval Research Logistics (NRL)*, 54(8):811–819, 2007.

[Lau and Liang, 2002] Hoong Chuin Lau and Zhe Liang. Pickup and delivery with time windows: Algorithms and test case generation. *International Journal on Artificial Intelligence Tools*, 11(03):455–472, 2002.

[Lecluyse *et al.*, 2009] Christophe Lecluyse, Tom Van Woensel, and Herbert Peremans. Vehicle routing with stochastic time-dependent travel times. *4or*, 7:363–377, 2009.

[Letchner *et al.*, 2006] Julia Letchner, John Krumm, and Eric Horvitz. Trip router with individualized preferences (trip): Incorporating personalization into route planning. In *AAAI*, pages 1795–1800, 2006.

[Lin *et al.*, 2025] Guanquan Lin, Mingjun Pan, You-Wei Luo, Zhien Dai, Bin Zhu, Lijun Sun, and Chun Yuan. Preference optimization for combinatorial optimization problems, 2025.

[Mandi *et al.*, 2021] Jayanta Mandi, Rocsildes Canoy, Víctor Bucarey, and Tias Guns. Data driven vrp: A neural network model to learn hidden preferences for vrp. *arXiv preprint arXiv:2108.04578*, 2021.

[Martí and Reinelt, 2022] Rafael Martí and Gerhard Reinelt. Exact and heuristic methods in combinatorial optimization. *Applied Mathematical Sciences*, 2022.

[Mor and Speranza, 2022] Andrea Mor and Maria Grazia Speranza. Vehicle routing problems over time: a survey. *Annals of Operations Research*, 314(1):255–275, 2022.

[Munari *et al.*, 2016] Pedro Munari, Twan Dollevoet, and Remy Spliet. A generalized formulation for vehicle routing problems. *arXiv preprint arXiv:1606.01935*, 2016.

[Peng and Wang, 2009] Yong Peng and Xiaofeng Wang. Research on a vehicle routing schedule to reduce fuel consumption. In *2009 International Conference on Measuring Technology and Mechatronics Automation*, volume 3, pages 825–827. IEEE, 2009.

[Potvin *et al.*, 1993] Jean-Yves Potvin, Gina Dufour, and Jean-Marc Rousseau. Learning vehicle dispatching with linear programming models. *Computers & operations research*, 20(4):371–380, 1993.

[Rafailov *et al.*, 2024] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

[Schaffer, 2014] J David Schaffer. Multiple objective optimization with vector evaluated genetic algorithms. In *Proceedings of the first international conference on genetic algorithms and their applications*, pages 93–100. Psychology Press, 2014.

[Sutton, 2018] Richard S Sutton. Reinforcement learning: An introduction. *A Bradford Book*, 2018.

[Todorov *et al.*, 2012] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 5026–5033. IEEE, 2012.

[Toledo *et al.*, 2013] Tomer Toledo, Yichen Sun, Katherine Rosa, Moshe Ben-Akiva, Kate Flanagan, Ricardo Sanchez, and Erika Spissu. Decision-making process and factors affecting truck routing. In *Freight Transport Modelling*, pages 233–249. Emerald Group Publishing Limited, 2013.

[Wirth *et al.*, 2017] Christian Wirth, Riad Akrour, Gerhard Neumann, and Johannes Fürnkranz. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18(136):1–46, 2017.

[Wu *et al.*, 2024] Yaoxin Wu, Mingfeng Fan, Zhiguang Cao, Ruobin Gao, Yaqing Hou, and Guillaume Sartoretti. Collaborative deep reinforcement learning for solving multi-objective vehicle routing problems. In *23rd International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2024*, pages 1956–1965, 2024.

[Xiao *et al.*, 2012] Yiyong Xiao, Qiuhong Zhao, Ikou Kaku, and Yuchun Xu. Development of a fuel consumption optimization model for the capacitated vehicle routing problem. *Computers & operations research*, 39(7):1419–1431, 2012.

[Yu *et al.*, 2017] Miao Yu, Viswanath Nagarajan, and Siqian Shen. Minimum makespan vehicle routing problem with compatibility constraints. In *International Conference on AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems*, pages 244–253. Springer, 2017.

[Zhang *et al.*, 2023] Cong Zhang, Yaoxin Wu, Yining Ma, Wen Song, Zhang Le, Zhiguang Cao, and Jie Zhang. A review on learning to solve combinatorial optimisation problems in manufacturing. *IET Collaborative Intelligent Manufacturing*, 5(1):e12072, 2023.

[Zhou *et al.*, 2023] Jianan Zhou, Yaoxin Wu, Wen Song, Zhiguang Cao, and Jie Zhang. Towards omni-generalizable neural methods for vehicle routing problems. In *International Conference on Machine Learning*, pages 42769–42789, 2023.

[Ziebart *et al.*, 2008] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.