

Multi-Agent Pointer Transformer: Seq-to-Seq Reinforcement Learning for Multi-Vehicle Dynamic Pickup-Delivery Problems

Zengyu Zou^{1, 4}, Jingyuan Wang^{1, 2, 3, 4, *}, Yixuan Huang^{1, 4}, Junjie Wu^{2, 3}

¹School of Computer Science and Engineering, Beihang University, Beijing, China

²School of Economics and Management, Beihang University, Beijing, China

³MIIT Key Laboratory of Data Intelligence and Management, Beihang University, Beijing, China

⁴MOE Engineering Research Center of Advanced Computer Application Technology, Beihang University, China
{zyy0001213, jywang, yixuanhuang, wujj}@buaa.edu.cn

Abstract

This paper addresses the cooperative Multi-Vehicle Dynamic Pickup and Delivery Problem with Stochastic Requests (MVDPPSR) and proposes an end-to-end centralized decision-making framework based on sequence-to-sequence, named Multi-Agent Pointer Transformer (MAPT). MVDPPSR is an extension of the vehicle routing problem and a spatio-temporal system optimization problem, widely applied in scenarios such as on-demand delivery. Classical operations research methods face bottlenecks in computational complexity and time efficiency when handling large-scale dynamic problems. Although existing reinforcement learning methods have achieved some progress, they still encounter several challenges: 1) Independent decoding across multiple vehicles fails to model joint action distributions; 2) The feature extraction network struggles to capture inter-entity relationships; 3) The joint action space is exponentially large. To address these issues, we designed the MAPT framework, which employs a Transformer Encoder to extract entity representations, combines a Transformer Decoder with a Pointer Network to generate joint action sequences in an AutoRegressive manner, and introduces a Relation-Aware Attention module to capture inter-entity relationships. Additionally, we guide the model's decision-making using informative priors to facilitate effective exploration. Experiments on 8 datasets demonstrate that MAPT significantly outperforms existing baseline methods in terms of performance and exhibits substantial computational time advantages compared to classical operations research methods.

Code — <https://github.com/wszyer/MAPT>

1 Introduction

The Pickup and Delivery Problem is a type of vehicle routing problem that has demonstrated its importance in many real-world applications, such as online food delivery. In real-world scenarios, the arrival time, origin, and destination of requests are often unpredictable, and there are strict requirements for service response times. Therefore, efficiently planning vehicle routes based on real-time updated information is particularly important in solving this problem. This

work focuses on solving the cooperative Multi-Vehicle Dynamic Pickup and Delivery Problem with Stochastic Requests (MVDPPSR). Unlike the traditional Pickup and Delivery Problem, in this problem, we need to route multiple vehicles, and the occurrence times of all requests are unknown in advance. Moreover, the origin and destination of each request are only revealed after the request emerges.

Existing methods for addressing dynamic vehicle routing problems fall into three categories: **heuristic methods**, **classical operations research methods**, and **reinforcement learning-based algorithms**. **Heuristic methods** have shown some success (Sheridan et al. 2013; Fikar 2018; Andersson 2021) by routing vehicles through manually designed heuristic rules, such as the nearest-distance rule. Their strengths are simplicity and efficiency, fitting real-time decision scenarios, but they have clear limits as they depend heavily on manually designed rules and struggle to ensure optimality. **Classical operations research methods** typically adopt the Rolling-Horizon paradigm, which transforms dynamic problems into static ones. These static subproblems are then solved using either exact algorithms (Lu and Dessouky 2004; Liu et al. 2018; Savelsbergh and Sol 1998) or metaheuristics (Schilde, Doerner, and Hartl 2011; Geiser, Hanne, and Dornberger 2020; Cai et al. 2022). This paradigm requires accumulating a sufficient number of requests before initiating planning, which fails to meet the needs of real-time dynamic decision-making. Also, both exact and metaheuristic algorithms have high computational complexity, greatly affecting their time efficiency in practice. With technology development, **reinforcement learning-based algorithms** have been used for multi-vehicle routing. Multi-agent reinforcement learning methods have been proposed for the static multi-vehicle pickup and delivery problem (Zong et al. 2022; Berto et al. 2024). They encode the current state with an Encoder, decode independently for each vehicle with a Decoder, and use a conflict handler in post-processing to solve vehicle conflicts.

However, these multi-agent methods still face the following issues: (1) Existing multi-agent reinforcement learning methods decode each agent independently, failing to model the joint probability distribution of agents' actions. Furthermore, the decisions generated by these methods may conflict among agents (e.g., competing for the same request). (2) Existing feature extraction networks fail to capture rela-

*Corresponding author

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

tionships between entities (e.g., stations, vehicles, requests), which are key components of the MVDPDPSR. (3) The joint action space is often exponentially large, which prevents reinforcement learning from exploring useful actions.

To tackle these remaining issues, we propose the Multi-Agent Pointer Transformer (MAPT), an end-to-end centralized decision-making framework based on sequence-to-sequence to solve MVDPDPSR. Specifically, we model the MVDPDPSR problem as a Markov Decision Process, use a Transformer Encoder to extract entity representations, and then combine a Transformer Decoder with a Pointer Network to convert the embedding sequence into a joint action sequence in an AutoRegressive manner. To extract relationships between entities, we propose a Relation-Aware Attention module that embeds entity relationships and adds them to the attention matrix. To address the challenge of reinforcement learning exploring a vast joint action space, we design informative priors based on load balancing and station distance, and fuse these priors with the action distribution decoded by the Decoder, enabling the model to select actions with higher potential rewards.

In summary, the main contributions of our work are as follows:

- We formalize the MVDPDPSR problem as a Markov Decision Process and design the Multi-Agent Pointer Transformer (MAPT) framework to decode joint actions and model the joint probability.
- We design a Relation-Aware Attention module to capture various relationships between entities.
- We design informative priors and integrate them into the model’s decision distribution to facilitate effective exploration.
- We validate the effectiveness of MAPT on 8 datasets, and experiments show that MAPT outperforms various baselines and demonstrates significant computational time advantages over classical operations research methods.

2 Preliminaries

2.1 Problem Formulation

In this section, we formally define the Multi-Vehicle Dynamic Pickup and Delivery Problem with Stochastic Requests(MVDPDPSR). A typical MVDPDPSR scenario involves a fleet of vehicles (Multi-Vehicle) tasked with picking up and delivering cargos among a set of stations. The pickup and delivery requests arise dynamically over time, reflecting the “Dynamic” nature of the problem. Accordingly, stations, vehicles, and requests constitute the three fundamental components of a MVDPDPSR. We provide their formal definitions below.

Definition 1 (Stations). The stations in MVDPDPSR are represented as a fully connected weighted graph $\mathcal{G} = \{N, \mathbf{E}\}$. Here, $N = \{n_1, \dots, n_i, \dots, n_I\}$ denotes the set of stations, where n_i represents the i -th station. $\mathbf{E} \in \mathbb{R}^{|N| \times |N|}$ is the distance matrix, where each element e_{n_i, n_j} indicates the travel time from station n_i to station n_j . In the station graph \mathcal{G} , each station can serve as either an origin (pickup station) or a destination (delivery station) for cargos.

Definition 2 (Vehicles). A MVDPDPSR scenario involves K vehicles, where each vehicle v_k is represented by a tuple of four state variables:

$$v_k = \langle cap_k, spa_k, tok_k, dist_k \rangle. \quad (1)$$

The meanings of these variables are as follows:

- cap_k (capacity): the total capacity of vehicle v_k .
- spa_k^t (space): the remaining available space of vehicle v_k at time t . For notational brevity, we omit the superscript time index t when unambiguous (e.g. spa_k).
- tok_k (destination): the current destination station of the vehicle.
- $dist_k$ (distance): the remaining distance (travel time) to reach the current destination.

If $dist_k \neq 0$, the vehicle is en route; otherwise, it is located at a station. Notably, the destination of a vehicle cannot be changed while it is en route.

Definition 3 (Requests). The objective of MVDPDPSR is to fulfill M cargo delivery requests, where each request r_m is represented by a tuple of six state variables:

$$r_m = \langle from_m, to_m, val_m, vol_m, time_m, state_m \rangle. \quad (2)$$

The meanings of these variables are as follows:

- $from_m$ (origin): the origin station of request r_m .
- to_m (destination): the destination station of request r_m .
- val_m (value): the profit associated with fulfilling the request.
- vol_m (volume): the cargo volume required for the request, which consumes vehicle space.
- $time_m$ (appearance time): the time at which the request becomes visible to the system.
- $state_m$ (state): the current state of request r_m .

A request becomes visible only when the current time reaches its appearance time $time_m$. The state variable $state_m \in \{\text{unassigned}, \text{picked}, \text{delivered}\}$ indicates whether the request has not been picked up, has been picked up by a vehicle, or has been delivered.

In the MVDPDPSR scenario, the planning horizon is discretized into time slices $t = 0, \dots, T$. At the initial time slice $t = 0$, all vehicles are located at their initial origin stations, and their available space equals their full capacity. At each time slice, a scheduler performs the following actions:

1. For every visible request in the unassigned state at time t (i.e. $time_m \leq t$), assign a vehicle that is currently at the same station and has sufficient space ($spa_k \geq vol_m$) to load the request’s cargo. Update the state of the request to picked and reduce the vehicle’s available space accordingly: $spa_k := spa_k - vol_m$.
2. For each vehicle at a station ($dist_k = 0$), select its next destination station and dispatch it. Set the vehicle’s travel distance $dist_k$ to the distance between the current station and the selected destination.
3. For every vehicle in transit ($dist_k \neq 0$), decrease its remaining travel distance by one unit: $dist_k := dist_k - 1$.

- For each vehicle arriving at its destination ($dist_k = 0$), unload all cargos whose destination matches the current station. Update the state of each delivered request to delivered and restore the vehicle's available space: $spa_k := spa_k + vol_m$ for each delivered request.

In this scenario, cargo delivery requests emerge dynamically over time, characterizing MVDPDPSR as a *dynamic* pickup-and-delivery problem.

A MVDPDPSR instance terminates when the planning horizon reaches the upper limit T . The goal of scheduling is to maximize the total profit of requests completed within the time limit while minimizing the aggregate travel cost.

2.2 Markov Decision Process (MDP)

MVDPDPSR is a sequential decision-making problem over time, and we model its decision process as a Markov decision process, defined as follows:

Definition 4 (Observation/State). Since we have a centralized decision system, the system can fully observe all state information at the current time step. The state at time step t includes the distance between stations e_{n_i, n_j} , all vehicle information v_k , and all appeared requests r_m where $time_m \leq t$. Notably, within a centralized decision-making system, the relationship $rel_{v_k, r_m} \in \{\text{unassigned}, \text{picked}, \text{delivered}\}$ between each vehicle v_k and request r_m at time t must be explicitly tracked. This relationship indicates the request has not been picked up, has been picked up by the vehicle, or has been delivered by the vehicle.

Definition 5 (Action). The decision includes decisions for requests and decisions for vehicles.

- Request decisions. For the set of unassigned requests $\mathcal{R}^t = \{r_m \mid state_m = \text{unassigned} \wedge time_m \leq t\}$, we need to decide which vehicle to assign them to. The assigned vehicle's current location must be the same as the request's origin, i.e.,

$$A_{r_m}^t \in \{v_k \mid dist_k = 0 \wedge to_k = from_m\} \cup \{\tau\}, \quad (3)$$

where τ denotes the action of temporarily deferring the request assignment. The assignment must ensure that the vehicle's capacity meets the requirements.

- Vehicle decisions. For the set of vehicles that have reached their destinations $\mathcal{V}^t = \{v_k \mid dist_k = 0\}$, we need to decide their next destination station, i.e.,

$$A_{v_k}^t \in \{n_1, \dots, n_I\}. \quad (4)$$

The final joint action space is the Cartesian product of all sub-action spaces:

$$A^t = \prod (\{A_{r_m}^t \mid r_m \in \mathcal{R}^t\} \cup \{A_{v_k}^t \mid v_k \in \mathcal{V}^t\}) \quad (5)$$

Definition 6 (Transition). For vehicles $v_k \in \mathcal{V}^t$, we need to update their destination station to_k and remaining time to reach the new destination station $dist_k$ based on $A_{v_k}^t$, and update their remaining space based on unloaded and newly loaded goods. For vehicles $v_k \notin \mathcal{V}^t$, the remaining time to reach their destination stations decreases by one unit. For all requests assigned to vehicles, their $state_m$ and rel_{v_k, r_m} changes to picked, and for all requests that have reached their destinations, their $state_m$ and rel_{v_k, r_m} changes to delivered.

Definition 7 (Objective/Reward). The objective is to optimize the overall routing solution quality by maximizing the profit from completed requests while accounting for vehicle travel costs. Formally, we define the objective at time T as:

$$obj_T = \sum_{t < T} \sum_k \left(\sum_{m \in D_k^t} val_m - cost \cdot e_{to_k^t, to_k^{t+1}} \right), \quad (6)$$

where D_k^t denotes the set of requests delivered by vehicle v_k by time t , and $cost$ represents the cost per unit distance. The single-step reward at each decision point is then defined as the incremental change in the objective value:

$$rwd_t = obj_t - obj_{t-1}. \quad (7)$$

3 Methodology

3.1 System Overview

We propose the Multi-Agent Pointer Transformer (MAPT), which is built on the Transformer (Vaswani et al. 2017) Encoder-Decoder architecture. At each time step of the MDP, the current state is fed into MAPT for decision-making. The Encoder processes the states of all entities and incorporates a Relation-Aware Attention module to capture inter-entity relationships. The Decoder generates actions for each entity in an AutoRegressive manner. To enhance exploration during reinforcement learning, we integrate informative priors into the action selection process. The overall architecture is illustrated in Figure 1.

3.2 Encoder

Feature Representation with Dense Vectors First, we transform the raw input into dense vectors for unified representation. We compute some augmented inputs to enhance the model's perception capabilities, namely ori_i and $dest_i$, which represent the number of requests originating from station n_i , and the number of requests destined for station n_i , respectively. Next, we concatenate the raw inputs of stations, requests, and vehicles and pass them through a linear layer to transform them into dense vectors:

$$\mathbf{d}_{n_i} = [ori_i \parallel dest_i] \mathbf{W}^n, \quad \mathbf{d}_{n_i} \in \mathbb{R}^{hs} \quad (8)$$

$$\mathbf{d}_{v_k} = [cap_k \parallel spa_k \parallel dist_k] \mathbf{W}^v, \quad \mathbf{d}_{v_k} \in \mathbb{R}^{hs} \quad (9)$$

$$\mathbf{d}_{r_m} = [val_m \parallel vol_m] \mathbf{W}^r, \quad \mathbf{d}_{r_m} \in \mathbb{R}^{hs} \quad (10)$$

where \parallel denotes the concatenation operation for scalars, and hs is the model's hidden size. Additionally, we transform global information into a dense vector, where m_t represents the current number of requests, as follows:

$$\mathbf{d}_g = [t \parallel m_t] \mathbf{W}^g, \quad \mathbf{d}_g \in \mathbb{R}^{hs}. \quad (11)$$

Relation-Aware Attention Although there are some neural network models for solving the Vehicle Routing Problem, these models cannot effectively capture the relationships between entities. To overcome these limitations, this study designs a Relation-Aware Attention module aimed at incorporating relationship information between entities into the embeddings. For any two entities (stations, requests,

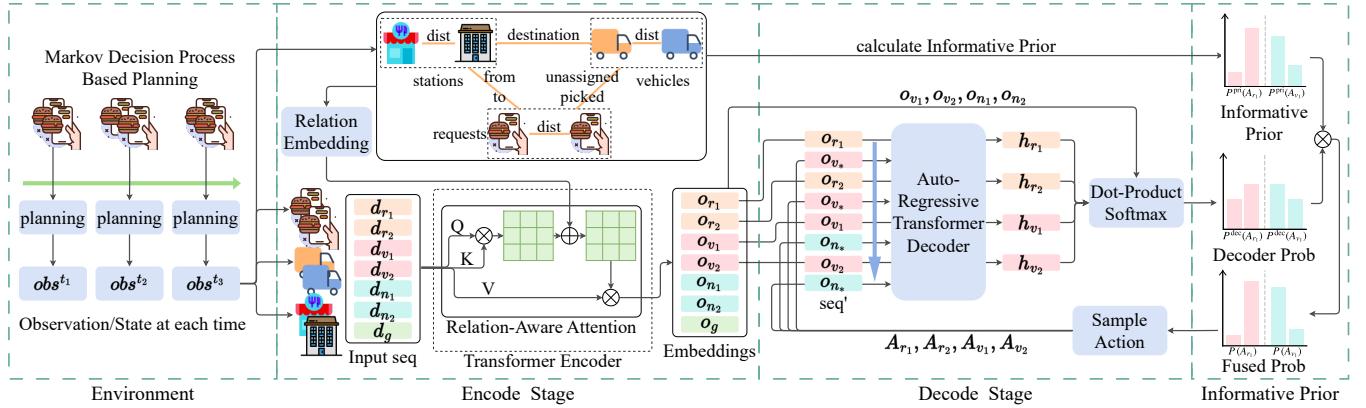


Figure 1: The framework of MAPT. The blue arrow indicates that the elements in the sequence are generated in an AutoRegressive manner. The elements marked with * are the actions that need to be decoded.

Relation	Vehicle	Request	Station
Vehicle	distance	unassigned picked delivered	is destination isn't destination
Request	unassigned picked delivered	distance	from to
Station	is destination isn't destination	from to	distance

Table 1: Relation Types between Vehicle, Request and Station.

vehicles), there are some relationships between them, as shown in Table 1. We design a relation encoding network Relation-Embedding(u, v) $\in \mathbb{R}$. If the relationship between entities u and v is a distance relationship, we use a linear layer to project it; for other relationships, we use a learnable parameter as their embedding. We denote the relationship matrix between entities as $R \in \mathbb{R}^{|Q| \times |K|}$, which is calculated as follows:

$$R_{u,v} = \text{Relation-Embedding}(q_u, k_v), \quad (12)$$

where q_u represents the u -th entity in Query sequence (Q), and k_v represents the v -th entity in Key sequence (K). Following the definition of the standard Transformer (Vaswani et al. 2017), we represent the Query sequence as Q , the Key sequence as K , the Value sequence as V and the head dimension as d_k . The Relation-Aware Attention is expressed as:

$$\text{Rel-Aware-Attn}(Q, K, V, R) = \text{softmax} \left(\frac{QK^T + R}{\sqrt{d_k}} \right) V. \quad (13)$$

We incorporate relationship information into the attention mechanism, enabling the model to capture relationships between entities when generating entity embeddings.

Encode with Relation-Aware Attention We use a standard Transformer Encoder and replace its attention layer

with Relation-Aware Attention. We concatenate the dense vectors of all stations, requests, vehicles, and global information into a sequence and input them into the Encoder to obtain their respective embeddings:

$$[\mathbf{o}_{r_1}, \dots, \mathbf{o}_{r_M}, \mathbf{o}_{v_1}, \dots, \mathbf{o}_{v_K}, \mathbf{o}_{n_1}, \dots, \mathbf{o}_{n_I}, \mathbf{o}_g] = \text{Encoder}([\mathbf{d}_{r_1}, \dots, \mathbf{d}_{r_M}, \mathbf{d}_{v_1}, \dots, \mathbf{d}_{v_K}, \mathbf{d}_{n_1}, \dots, \mathbf{d}_{n_I}, \mathbf{d}_g]). \quad (14)$$

3.3 Decoder

Decoder Model We use a standard Transformer Decoder and replace its attention layer with Relation-Aware Attention. The cross-attention layer needs to perform cross-attention on the output of the Encoder to perceive all information about the current problem. Additionally, we add learnable positional encoding before the Decoder.

AutoRegressive Decoding We model the joint probability distribution A^t using chain rule decomposition, which is order-agnostic:

$$P(A^t) = \prod_m^M P(A_{r_m}^t | A_{r_1 \dots r_{m-1}}^t) \times \prod_k^K P(A_{v_k}^t | A_{v_1 \dots v_{k-1}}^t, A_{r_1 \dots r_M}^t). \quad (15)$$

From this formula, we can see that each request needs to refer to the decision results of previous requests when making decisions; each vehicle needs to refer to the decision results of previous vehicles and all requests when making decisions. Due to this sequential decision-making process, we adopt an AutoRegressive approach to generate the decision sequence and model it in a sequence-to-sequence framework, as also observed by Wen et al. (2022). We combine the Transformer’s ability to capture long-distance dependencies with the Pointer Network’s (Vinyals, Fortunato, and Jaitly 2015) ability to handle entity selection, using the Transformer Decoder to decode actions.

We detail our decoding process, for notational convenience, we omit the superscript time index t when unambiguous. Let our sequence be seq , initially empty. When de-

coding the action for the m -th request r_m , we first append it into the sequence, which then becomes:

$$\text{seq} = [r_1, A_{r_1}, \dots, r_m]. \quad (16)$$

We first convert entities in the sequence to their corresponding embeddings to obtain seq' (e.g., converting request r_m to its embedding \mathbf{o}_{r_m} which is output of the Transformer Encoder). The Decoder(seq') produces a hidden sequence of the same length as seq' . We take the last item of the hidden sequence, \mathbf{h}_{r_m} , as the query vector and use the embeddings of all vehicles as the key vectors. By calculating the attention scores between them, we can quantify the request's preference for each vehicle:

$$P^{\text{dec}}(A_{r_m} = v_k) = \text{softmax}_k(\mathbf{h}_{r_m}^\top [\mathbf{o}_{v_1}, \dots, \mathbf{o}_{v_K}]), \quad (17)$$

where $\text{softmax}_k(\cdot)$ denotes the k -th element of the softmax output. We determine A_{r_m} by sampling from $P^{\text{dec}}(A_{r_m})$ or selecting the maximum probability vehicle. This decoding result is then placed at the end of seq , providing critical reference information for subsequent decisions:

$$\text{seq} = [r_1, A_{r_1}, \dots, r_m, A_{r_m}]. \quad (18)$$

This process can be repeated to decode all requests. We decode all vehicles in the same manner:

$$P^{\text{dec}}(A_{v_k} = n_i) = \text{softmax}_i(\mathbf{h}_{v_k}^\top [\mathbf{o}_{n_1}, \dots, \mathbf{o}_{n_I}]). \quad (19)$$

After all decoding is completed, the elements in the sequence are:

$$\text{seq} = [r_1, A_{r_1}, \dots, v_1, A_{v_1}, \dots, v_K, A_{v_K}]. \quad (20)$$

3.4 Informative Priors

The joint action space in MVDPDPSR is often exponentially large because of multiple vehicles, requests, and stations, which prevents reinforcement learning from exploring useful actions. To address this, we designed manually-computed informative priors for vehicle selection and destination selection. By multiplying the probabilities output by the Decoder with informative priors, we guide the agent to prioritize actions with higher potential rewards.

Informative Priors for Vehicle Selection In a multi-vehicle scenario, the balance of vehicle load is crucial for overall delivery efficiency and resource utilization. Based on this, we design a load-balancing informative prior, which is inversely proportional to the current vehicle load. The smaller the current load of a vehicle, the higher its informative prior weight in vehicle selection, thereby guiding the assignment of requests toward load balancing:

$$P^{\text{pri}}(A_{r_m} = v_k) = \frac{s_{\text{pa}_k}}{c_{\text{ap}_k}}. \quad (21)$$

Although temporarily deferring request assignments may yield positive benefits (e.g., when the current vehicle load is high), we aim to avoid excessive occurrences of such cases, as this would severely affect the exploration efficiency during reinforcement learning. Therefore, we set a small probability for action τ (defined in Eq. (3)):

$$P^{\text{pri}}(A_{r_m} = \tau) = \beta. \quad (22)$$

The vehicle selection probability after informative priors' guidance is:

$$P(A_{r_m} = v_k) = P^{\text{dec}}(A_{r_m} = v_k) \cdot P^{\text{pri}}(A_{r_m} = v_k). \quad (23)$$

Informative Priors for Destination Selection For the destination station selection of vehicle v_k , the informative prior $P^{\text{pri}}(A_{v_k} = n_i)$ is determined as follows:

- If station n_i has requests to be delivered, $P^{\text{pri}}(A_{v_k} = n_i) = 1$.
- If station n_i has requests to be picked up, $P^{\text{pri}}(A_{v_k} = n_i) = \frac{0.1 \times \bar{E}}{e_{\text{to}_k, n_i}}$, where \bar{E} denotes the average value of all elements in the distance matrix E . This formulation prioritizes stations that are closer in distance. Since the number of stations with pickup requests is significantly larger than those with delivery requests, we use a smaller coefficient (0.1) to balance the vehicle's priorities between pickup and delivery tasks.
- For stations n_i that do not fall into the above two categories, $P^{\text{pri}}(A_{v_k} = n_i) = 0$.

The destination selection probability after informative priors' guidance is:

$$P(A_{v_k} = n_i) = P^{\text{dec}}(A_{v_k} = n_i) \cdot P^{\text{pri}}(A_{v_k} = n_i). \quad (24)$$

3.5 Optimization via PPO

We use Proximal Policy Optimization (Schulman et al. 2017) to train MAPT. Our value function is defined as $V(s_t) = \text{MLP}(\mathbf{o}_g)$, where MLP transforms global information into a scalar. We use Generalized Advantage Estimation (Schulman et al. 2015) to balance the bias and variance of advantage estimation:

$$\hat{A}_t^{\text{GAE}} = \sum_{l=0}^{\infty} (\gamma \lambda)^l (rwd_{t+l} + \gamma V(s_{t+l+1}) - V(s_{t+l})). \quad (25)$$

The Actor loss is defined as

$$\rho_t(\theta) = \frac{\pi_\theta(A^t | s_t)}{\pi_{\theta_{\text{old}}}(A^t | s_t)}, \quad (26)$$

$$\mathcal{L}^{\text{CLIP}}(\theta) = \mathbb{E}_t \left[\min \left(\rho_t(\theta) \hat{A}_t^{\text{GAE}}, \text{CLIP}(\rho_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t^{\text{GAE}} \right) \right], \quad (27)$$

where $\pi_\theta(A^t | s_t)$ is the joint probability defined in Eq. (15). The Critic loss is defined as

$$\mathcal{L}^{\text{Critic}} = (V(s_t) - (r_{t+1} + \gamma V(s_{t+1})))^2. \quad (28)$$

The final total loss is

$$\mathcal{L} = \mathcal{L}^{\text{CLIP}} + \alpha \mathcal{L}^{\text{Critic}}. \quad (29)$$

4 Experiments

4.1 Experimental Scenarios

To validate the effectiveness of the MAPT framework, we conducted experiments on 8 datasets, including both synthetic and real-world datasets, for comprehensive evaluation. A brief description of the datasets is provided below.

- *Synthetic Dataset*: We generated synthetic datasets with different scales of stations $I \in \{20, 50, 300\}$, corresponding to dataset suffixes *-S*, *-L*, and *-XL*, respectively. For these datasets, the number of requests and vehicles were

Scenario Metric	synth-S		synth-S-cost		synth-L		synth-L-cost		dhrd-tpe		dhrd-sg		dhrd-se		synth-XL	
	Obj.↑	Comp.↑	Obj.↑	Comp.↑	Obj.↑	Comp.↑	Obj.↑	Comp.↑	Obj.↑	Comp.↑	Obj.↑	Comp.↑	Obj.↑	Comp.↑	Obj.↑	Comp.↑
OR Tools	182.1	0.54	117.4	0.52	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	55.3	0.45	N/A	N/A
SA	162.5	0.52	108.1	0.52	641.1	0.28	339.3	0.29	244.9	0.31	136.7	0.30	50.4	0.41	1531.9	0.31
GA	108.7	0.30	58.3	0.29	329.4	0.14	89.5	0.12	64.1	0.08	40.9	0.13	21.3	0.17	863.8	0.17
SA*	133.0	0.42	81.2	0.42	563.1	0.25	284.8	0.24	223.3	0.28	139.6	0.30	56.6	0.46	1650.9	0.34
GA*	75.2	0.21	39.4	0.19	219.9	0.09	88.0	0.09	53.8	0.07	39.0	0.13	18.4	0.15	543.2	0.11
MAPDP*	174.9	0.53	82.0	0.49	177.6	0.08	65.2	0.08	287.5	0.36	247.2	0.35	72.9	0.56	807.0	0.16
PARCO*	178.8	0.53	86.6	0.52	177.6	0.08	65.2	0.08	391.2	0.49	267.5	0.46	64.2	0.49	1834.7	0.36
Nearest	189.2	0.57	124.1	0.57	962.6	0.41	592.3	0.41	309.2	0.39	194.5	0.40	39.8	0.32	2641.1	0.53
MAPDP	157.5	0.41	70.0	0.42	958.1	0.37	385.2	0.37	135.4	0.17	64.9	0.18	19.3	0.16	2543.1	0.50
MAPT	275.2	0.83	192.1	0.84	1875.1	0.80	1348.6	0.81	697.2	0.87	376.1	0.71	78.9	0.64	3227.5	0.65

Table 2: Overall performance comparison on synthetic datasets and real-world datasets. The best result in each column is bolded. "N/A" indicates that the algorithm cannot provide results within an acceptable time frame.

Scenario Metric	synth-S		synth-S-cost		synth-L		synth-L-cost		dhrd-tpe		dhrd-sg		dhrd-se		synth-XL	
	Obj.↑	Comp.↑	Obj.↑	Comp.↑	Obj.↑	Comp.↑	Obj.↑	Comp.↑	Obj.↑	Comp.↑	Obj.↑	Comp.↑	Obj.↑	Comp.↑	Obj.↑	Comp.↑
MAPT	275.2	0.83	192.1	0.84	1875.1	0.80	1348.6	0.81	697.2	0.87	376.1	0.71	78.9	0.64	3227.5	0.65
w/o Rel	270.9	0.82	190.7	0.83	1869.9	0.80	1338.4	0.81	692.1	0.87	365.3	0.69	77.1	0.63	3136.8	0.63
w/o AR	217.2	0.65	171.1	0.77	1604.0	0.68	1074.5	0.69	611.4	0.76	258.1	0.50	49.4	0.40	1905.8	0.38
w/o Priors	175.2	0.54	92.7	0.54	1089.5	0.47	529.1	0.48	511.6	0.64	250.9	0.49	56.6	0.46	1375.0	0.27

Table 3: Results of ablation studies. The best result in each column is bolded.

set as $(M, K) = (110, 5)$, $(550, 15)$, and $(550, 50)$, respectively. Vehicle capacity was fixed at 3. Travel cost per unit distance is set to 0 or 0.3, with the latter case marked by an additional suffix `-cost`. Requests were uniformly sampled from stations, with appearance times sampled from $\text{Uniform}(1, T)$, where T was set to 58 or 128. The profit for each request was the distance between its origin and destination.

- **Real-World Dataset:** We used the DHRD dataset (Assylbekov et al. 2023), which contains food delivery requests from three cities: Taipei (`tpe`), Singapore (`sg`), and Stockholm (`se`). Each city was treated as an independent dataset with $I = 36$ stations. The corresponding numbers of requests and vehicles were $(M, K) = (800, 20)$, $(700, 15)$, and $(200, 3)$ for the three cities, respectively. Vehicle capacity was set to 6. The dataset was divided into 76 days for training/validation and 14 days for testing.

4.2 Performance Evaluation

Baselines We evaluate our MAPT against several baselines, including Rolling-Horizon algorithms, static algorithms, rule-based algorithms, and MDP-based algorithms.

- **Rolling-Horizon Algorithms:** The Rolling-Horizon paradigm addresses dynamic problems by decomposing them into static subproblems within a sliding time window. For solving these static subproblems, we employ **OR Tools**, **Simulated Annealing (SA)**, and **Genetic Algorithm (GA)**.

- **Static Algorithms:** These methods assume complete foreknowledge of all requests (though unrealistic), solving the problem statically without Rolling-Horizon. **MAPDP*** (Zong et al. 2022) is a MARL approach with Encoder-Decoder for static problems, adapted to handle distance matrix inputs. **PARCO*** (Berto et al. 2024) is a Transformer-based RL framework with parallel decision-making, modified to support distance matrix inputs. **SA*/GA*** are static versions of our Rolling-Horizon SA and GA with full request information.

- **Rule-Based Algorithms:** These methods use predefined rules for scheduling decisions. **Nearest** is a greedy approach that always selects the closest available request for pickup or delivery.

- **MDP-Based Algorithms:** At each step of the MDP, we solve the problem statically using **MAPDP** (Zong et al. 2022) as described in the Static Algorithms section, and only execute the first step of the solution to ensure applicability to dynamic problems.

Overall Comparison The performance results on synthetic and real-world datasets are summarized in Table 2, using the objective (Obj.) and request completion rate (Comp.) as metrics (higher is better). MAPT significantly outperforms all baselines across all datasets. MAPT also surpasses the state-of-the-art MAPDP, which fails to model multi-vehicle joint probabilities and lacks our informative priors. As can be seen from Table 4, the decision time of MAPT is significantly shorter than that of exact and meta-heuristic algorithms.

Scenario	synth-S	synth-S-cost	synth-L	synth-L-cost	synth-XL
OR Tools	401.727	459.143	N/A	N/A	N/A
SA	23.684	23.255	166.660	164.102	613.920
GA	16.654	16.448	103.147	99.864	354.789
Nearest	0.059	0.060	0.117	0.119	0.148
MAPDP	0.794	0.827	31.981	32.544	33.217
MAPT	1.420	1.416	8.300	7.888	27.434

Table 4: Comparison of inference times on the synthetic dataset (in seconds), which is the average running time of each instance in the dataset.

4.3 Ablation Studies

We conducted ablation studies on all datasets to validate the effectiveness of key components:

- **w/o Relation-Aware Attention (Rel):** Replaces the Relation-Aware Attention module with the original attention mechanism.
- **w/o AutoRegressive Decoding (AR):** Decodes each action independently and employs a conflict handler to resolve conflicts where request assignments exceed vehicle capacity.
- **w/o Informative Priors (Priors):** Removes informative priors guidance for vehicle and destination selection.

The results are shown in Table 3. As can be seen from the table, AutoRegressive Decoding and Informative Priors significantly improve performance across all datasets, and Relation-Aware Attention also contributes to some extent. The ablation experiments demonstrate that all three proposed components enhance the final performance, validating their effectiveness.

4.4 Sensitivity Analysis

We conduct a sensitivity analysis on the hyperparameter β as defined in Eq. (22). We vary β from 0 to 0.3 in increments of 0.02 and observe the resulting changes in the final objective. The results on the 4 representative datasets are illustrated in Figure 2. As shown in the figure, when β is small, the performance is better and more stable, while larger values of β lead to a gradual decline in performance. However, in the dhrd-tpe and dhrd-se datasets, using a small but non-zero β improves performance. This indicates that slightly deferring request allocation can be beneficial in these cases.

5 Related Work

5.1 Dynamic Pickup and Delivery Problems

Dynamic Pickup and Delivery Problems have been studied through exact, heuristic, metaheuristic, and learning-based methods (Cai et al. 2023). Exact methods such as MILP can yield optimal results but are restricted to small instances (Liu et al. 2018; Savelsbergh and Sol 1998). Heuristic methods (Fikar 2018; Fikar, Hirsch, and Gronalt 2018; Andersson 2021) are efficient for large-scale dynamic problems but cannot ensure optimality. Metaheuristics (Schilde, Doerner, and Hartl 2011; Tirado and Hvattum 2017; Novaes et al.

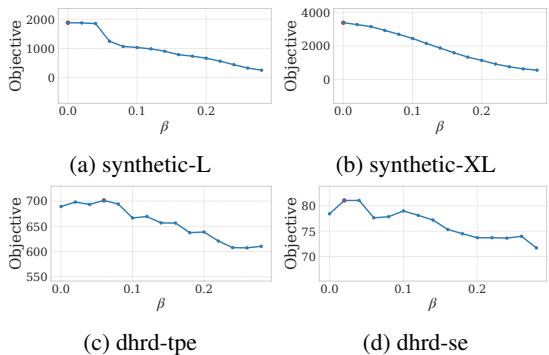


Figure 2: Sensitivity analysis of β across datasets.

2015) improve scalability but still struggle in highly dynamic settings. Reinforcement learning methods have been explored (Li et al. 2021; Yu et al. 2025), but they mainly focus on single-vehicle operations. Hybrid approaches first learn edge costs using neural networks and then apply traditional optimization techniques (Jiang et al. 2023b; Wang, Wu, and Zhao 2021; Wang et al. 2019; Wang, Lin, and Li 2025; Wu et al. 2019). Other studies have addressed ride-on-demand problems with dynamic pricing (Guo et al. 2019, 2020, 2023), as well as dynamic metro scheduling (Wang et al. 2022). There are also entity representation learning methods that use GNNs, hyper-GNNs, contrastive learning, transfer learning, or LLM enhancement and can be applied to spatio-temporal system optimization (Wu et al. 2020; Yang et al. 2025; Han et al. 2025; Jiang et al. 2023a; Li et al. 2025; Cheng et al. 2025; Zhang et al. 2024).

5.2 RL-based Methods for Vehicle Routing Problems

Reinforcement Learning has been applied to both single and multiple Vehicle Routing Problems. For single-vehicle problems, attention-based models (Vinyals, Fortunato, and Jaitly 2015; Bello et al. 2016; Nazari et al. 2018; Kool, Van Hoof, and Welling 2018) and GNN-based variants (Fellek et al. 2023; Heydaribeni et al. 2023) have been proposed. For multi-vehicle problems, multi-agent RL methods (Zhang et al. 2020; Zong et al. 2022; Zhang, Qi, and Guan 2023) have been developed but mainly for static settings. Recent RL studies for dynamic settings (Anuar et al. 2022; Xiang et al. 2024) have emerged, but they simplify the original problem by introducing additional assumptions.

6 Conclusion

This paper presents MAPT to solve the cooperative Multi-Vehicle Dynamic Pickup and Delivery Problem with Stochastic Requests. MAPT utilizes the Transformer’s Encoder-Decoder architecture combining with Pointer Network to generate joint action sequences in an AutoRegressive manner. By incorporating the Relation-Aware Attention module and informative priors, the framework achieves a better performance. Experiments show that MAPT outperforms existing baselines in solution quality across synthetic datasets and real-world datasets.

7 Acknowledgments

This work is supported by the National Key R&D Program of China (2023YFC3304700). Prof. Jingyuan Wang's work was partially supported by the National Natural Science Foundation of China (No. 72222022, 72171013) and the Fundamental Research Funds for the Central Universities (JKF-2025017226182). Prof. Junjie Wu's work was partially supported by the National Natural Science Foundation of China (72242101, 72031001), the Outstanding Young Scientist Program of Beijing Universities (JWZQ20240201002), and the Artificial Intelligence Technology R&D Center for Exploration and Development of China National Petroleum Corporation (2024-KFKT-22).

References

- Andersson, T. 2021. A Comparative Study on a Dynamic Pickup and Delivery Problem: Improving routing and order assignment in same-day courier operations.
- Anuar, W. K.; Lee, L. S.; Seow, H.-V.; and Pickl, S. 2022. A multi-depot dynamic vehicle routing problem with stochastic road capacity: An MDP model and dynamic policy for post-decision state rollout algorithm in reinforcement learning. *Mathematics*, 10(15): 2699.
- Assylbekov, Y.; Bali, R.; Bovard, L.; and Klaue, C. 2023. Delivery Hero Recommendation Dataset: A Novel Dataset for Benchmarking Recommendation Algorithms. In *Proceedings of the 17th ACM Conference on Recommender Systems*, 1042–1044.
- Bello, I.; Pham, H.; Le, Q. V.; Norouzi, M.; and Bengio, S. 2016. Neural combinatorial optimization with reinforcement learning. *arXiv preprint arXiv:1611.09940*.
- Berto, F.; Hua, C.; Luttmann, L.; Son, J.; Park, J.; Ahn, K.; Kwon, C.; Xie, L.; and Park, J. 2024. PARCO: Learning Parallel Autoregressive Policies for Efficient Multi-Agent Combinatorial Optimization. *arXiv preprint arXiv:2409.03811*.
- Cai, J.; Zhu, Q.; Lin, Q.; Li, J.; Chen, J.; and Ming, Z. 2022. An efficient multi-objective evolutionary algorithm for a practical dynamic pickup and delivery problem. In *International conference on intelligent computing*, 27–40. Springer.
- Cai, J.; Zhu, Q.; Lin, Q.; Ma, L.; Li, J.; and Ming, Z. 2023. A survey of dynamic pickup and delivery problems. *Neurocomputing*, 126631.
- Cheng, J.; Wang, J.; Zhang, Y.; Ji, J.; Zhu, Y.; Zhang, Z.; and Zhao, X. 2025. Poi-enhancer: An llm-based semantic enhancement framework for poi representation learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, 11509–11517.
- Fellek, G.; Farid, A.; Gebreyesus, G.; Fujimura, S.; and Yoshie, O. 2023. Deep Graph Representation Learning to Solve Vehicle Routing Problem. In *2023 International Conference on Machine Learning and Cybernetics (ICMLC)*, 172–180. IEEE.
- Fikar, C. 2018. A decision support system to investigate food losses in e-grocery deliveries. *Computers & Industrial Engineering*, 117: 282–290.
- Fikar, C.; Hirsch, P.; and Gronalt, M. 2018. A decision support system to investigate dynamic last-mile distribution facilitating cargo-bikes. *International Journal of Logistics Research and Applications*, 21(3): 300–317.
- Geiser, T.; Hanne, T.; and Dornberger, R. 2020. Best-match in a set of single-vehicle dynamic pickup and delivery problem using ant colony optimization. In *Proceedings of the 2020 the 3rd International Conference on Computers in Management and Business*, 126–131.
- Guo, S.; Chen, C.; Wang, J.; Ding, Y.; Liu, Y.; Xu, K.; Yu, Z.; and Zhang, D. 2020. A force-directed approach to seeking route recommendation in ride-on-demand service using multi-source urban data. *IEEE Transactions on Mobile Computing*, 21(6): 1909–1926.
- Guo, S.; Chen, C.; Wang, J.; Liu, Y.; Xu, K.; Yu, Z.; Zhang, D.; and Chiu, D. M. 2019. Rod-revenue: Seeking strategies analysis and revenue prediction in ride-on-demand service using multi-source urban data. *IEEE Transactions on Mobile Computing*, 19(9): 2202–2220.
- Guo, S.; Shen, Q.; Liu, Z.; Chen, C.; Chen, C.; Wang, J.; Li, Z.; and Xu, K. 2023. Seeking based on dynamic prices: Higher earnings and better strategies in ride-on-demand services. *IEEE Transactions on Intelligent Transportation Systems*, 24(5): 5527–5542.
- Han, C.; Wang, J.; Wang, Y.; Yu, X.; Lin, H.; Li, C.; and Wu, J. 2025. Bridging traffic state and trajectory for dynamic road network and trajectory representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 11763–11771.
- Heydaribeni, N.; Zhan, X.; Zhang, R.; Eliassi-Rad, T.; and Koushanfar, F. 2023. HypOp: Distributed Constrained Combinatorial Optimization leveraging Hypergraph Neural Networks. *arXiv preprint arXiv:2311.09375*.
- Jiang, J.; Pan, D.; Ren, H.; Jiang, X.; Li, C.; and Wang, J. 2023a. Self-supervised trajectory representation learning with temporal regularities and travel semantics. In *2023 IEEE 39th international conference on data engineering (ICDE)*, 843–855. IEEE.
- Jiang, W.; Zhao, W. X.; Wang, J.; and Jiang, J. 2023b. Continuous trajectory generation based on two-stage GAN. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 4374–4382.
- Kool, W.; Van Hoof, H.; and Welling, M. 2018. Attention, learn to solve routing problems! *arXiv preprint arXiv:1803.08475*.
- Li, X.; Luo, W.; Yuan, M.; Wang, J.; Lu, J.; Wang, J.; Lü, J.; and Zeng, J. 2021. Learning to optimize industry-scale dynamic pickup and delivery problems. In *2021 IEEE 37th international conference on data engineering (ICDE)*, 2511–2522. IEEE.
- Li, Y.; Wang, J.; Yu, X.; Wang, P.; and Huang, Q. 2025. Cross City Traffic Flow Generation via Retrieval Augmented Diffusion Model. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Liu, S.; Tan, P. H.; Kurniawan, E.; Zhang, P.; and Sun, S. 2018. Dynamic scheduling for pickup and delivery with time

- windows. In *2018 IEEE 4th World Forum on Internet of Things (WF-IoT)*, 767–770. IEEE.
- Lu, Q.; and Dessouky, M. 2004. An exact algorithm for the multiple vehicle pickup and delivery problem. *Transportation Science*, 38(4): 503–514.
- Nazari, M.; Oroojlooy, A.; Snyder, L.; and Takac, M. 2018. Reinforcement Learning for Solving the Vehicle Routing Problem. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.
- Novaes, A. G.; Bez, E. T.; Burin, P. J.; and Aragão Jr, D. P. 2015. Dynamic milk-run OEM operations in over-congested traffic conditions. *Computers & Industrial Engineering*, 88: 326–340.
- Savelsbergh, M.; and Sol, M. 1998. Drive: Dynamic routing of independent vehicles. *Operations Research*, 46(4): 474–490.
- Schilde, M.; Doerner, K. F.; and Hartl, R. F. 2011. Meta-heuristics for the dynamic stochastic dial-a-ride problem with expected return transports. *Computers & operations research*, 38(12): 1719–1730.
- Schulman, J.; Moritz, P.; Levine, S.; Jordan, M.; and Abbeel, P. 2015. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*.
- Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Sheridan, P. K.; Gluck, E.; Guan, Q.; Pickles, T.; Balciog, B.; Benhabib, B.; et al. 2013. The dynamic nearest neighbor policy for the multi-vehicle pick-up and delivery problem. *Transportation Research Part A: Policy and Practice*, 49: 178–194.
- Tirado, G.; and Hvattum, L. M. 2017. Improved solutions to dynamic and stochastic maritime pick-up and delivery problems using local search. *Annals of Operations Research*, 253: 825–843.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Vinyals, O.; Fortunato, M.; and Jaitly, N. 2015. Pointer networks. *Advances in neural information processing systems*, 28.
- Wang, J.; Lin, Y.; and Li, Y. 2025. GTG: Generalizable Trajectory Generation Model for Urban Mobility. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 834–842.
- Wang, J.; Wu, N.; and Zhao, W. X. 2021. Personalized route recommendation with neural network enhanced search algorithm. *IEEE Transactions on Knowledge and Data Engineering*, 34(12): 5910–5924.
- Wang, J.; Wu, N.; Zhao, W. X.; Peng, F.; and Lin, X. 2019. Empowering A* search algorithms with neural networks for personalized route recommendation. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 539–547.
- Wang, Z.; Pan, Z.; Chen, S.; Ji, S.; Yi, X.; Zhang, J.; Wang, J.; Gong, Z.; Li, T.; and Zheng, Y. 2022. Shortening passengers’ travel time: A dynamic metro train scheduling approach using deep reinforcement learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(5): 5282–5295.
- Wen, M.; Kuba, J.; Lin, R.; Zhang, W.; Wen, Y.; Wang, J.; and Yang, Y. 2022. Multi-Agent Reinforcement Learning is a Sequence Modeling Problem. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 16509–16521. Curran Associates, Inc.
- Wu, N.; Wang, J.; Zhao, W. X.; and Jin, Y. 2019. Learning to effectively estimate the travel time for fastest route recommendation. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 1923–1932.
- Wu, N.; Zhao, X. W.; Wang, J.; and Pan, D. 2020. Learning effective road network representation with hierarchical graph neural networks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 6–14.
- Xiang, C.; Wu, Z.; Tu, J.; and Huang, J. 2024. Centralized Deep Reinforcement Learning Method for Dynamic Multi-Vehicle Pickup and Delivery Problem With Crowdshippers. *IEEE Transactions on Intelligent Transportation Systems*.
- Yang, Y.; Wang, J.; Yu, X.; and Tang, Y. 2025. HygMap: Representing All Types of Map Entities via Heterogeneous Hypergraph. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 9438–9446.
- Yu, X.; Wang, J.; Yang, Y.; Huang, Q.; and Qu, K. 2025. BIGCity: A universal spatiotemporal model for unified trajectory and traffic state data analysis. In *2025 IEEE 41st International Conference on Data Engineering (ICDE)*, 4455–4469. IEEE.
- Zhang, K.; He, F.; Zhang, Z.; Lin, X.; and Li, M. 2020. Multi-vehicle routing problems with soft time windows: A multi-agent reinforcement learning approach. *Transportation Research Part C: Emerging Technologies*, 121: 102861.
- Zhang, W.; Wang, J.; Yang, Y.; et al. 2024. VecCity: A taxonomy-guided library for map entity representation learning. *arXiv preprint arXiv:2411.00874*.
- Zhang, Z.; Qi, G.; and Guan, W. 2023. Coordinated multi-agent hierarchical deep reinforcement learning to solve multi-trip vehicle routing problems with soft time windows. *IET Intelligent Transport Systems*, 17(10): 2034–2051.
- Zong, Z.; Zheng, M.; Li, Y.; and Jin, D. 2022. Mapdp: Cooperative multi-agent reinforcement learning to solve pickup and delivery problems. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 9980–9988.

A Example Problem Instance

Figure 3 presents a representative problem instance in our study. This instance consists of 2 pickup stations and 4 delivery stations, where each request is associated with one pickup station and one delivery station. It also includes 2 vehicles (yellow and blue) and 5 requests — each request is identified by a pair of pickup (p_n) and delivery (d_n) stations (with $n = 1, 2, 3, 4, 5$), and has timestamps like 9:00, 9:10, etc. The yellow vehicle follows a sequential route: $p_1 \rightarrow p_2 \rightarrow d_1 \rightarrow d_2 \rightarrow p_3 \rightarrow d_3$ — as visualized by the orange arrows connecting these stations in the figure. It first handles early requests p_1 (pickup at 9:00) and p_2 (pickup at 9:10), then completes their deliveries (d_1 at 9:20, d_2 at 9:30). When p_3 (pickup at 9:40) dynamically arrives later, the yellow vehicle is scheduled to *re-route back to the top-left pickup station* to fulfill this new request — highlighting how our problem requires adapting to **dynamically arriving requests**. Meanwhile, the blue vehicle executes another sequence: $p_4 \rightarrow d_4 \rightarrow p_5 \rightarrow d_5$ — as shown by the blue arrows. Here, p_4 (pickup at 9:00) and d_4 (delivery at 9:10) are processed first, followed by p_5 (pickup at 9:20) and d_5 (delivery at 9:30).

When the blue vehicle reaches the upper-middle pickup station, Request 4 (e.g., p_4 , the pickup for Request 4) is assigned to it, while Request 2 (e.g., p_2 , the pickup for Request 2) is deferred and then assigned to the yellow vehicle upon its subsequent arrival — demonstrating the core challenge of **vehicle-request assignment**: we need to decide which vehicle each request should be assigned to.

B Difference between Rolling Horizon and Markov Decision Process

Rolling Horizon is often used in classical operations research methods. It transforms dynamic problems into static subproblems by accumulating a sufficient number of requests over a time window before planning begins (top of Fig. 4). During this accumulation already-arrived requests stay unprocessed causing unavoidable time waste. These static subproblems are then solved with exact algorithms or metaheuristics but their high computational complexity further hinders time efficiency. In contrast Markov Decision Process (MDP) triggers planning at each time step (bottom of Fig. 4) and makes decisions right as requests arrive. The former fails to meet real-time dynamic decision-making needs due to its batch-based approach while the latter prioritizes real-time adaptation through frequent smaller-scale planning.

C Dataset Details

In the MVDPDPSR problem addressed in this paper, the problem scale is not solely determined by the number of stations but also depends on the number of requests, as we need to determine the assignment relationship between requests and vehicles. In the largest dataset, the number of requests reaches 800, which already exceeds the problem scale of most prior studies on static pickup-and-delivery problems, such as (Zong et al. 2022). Below, we provide details of the

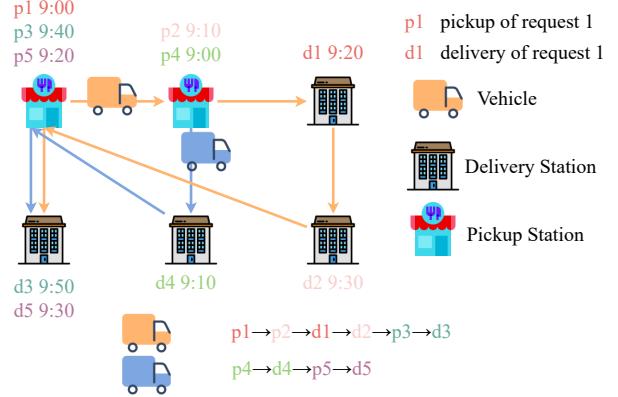


Figure 3: Schematic illustration of the MVDPDPSR problem.

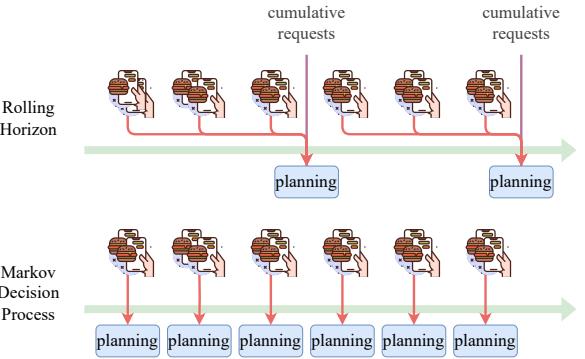


Figure 4: Comparison between Rolling Horizon and Markov Decision Process paradigms. The Rolling Horizon paradigm requires accumulating a sufficient number of requests before invoking a static solver for vehicle routing optimization, while the Markov Decision Process paradigm enables real-time decision-making as requests arrive.

synthetic datasets and the real-world datasets. Their statistics are summarized in Table 5.

- *Synthetic Dataset:* We first synthesize some datasets to validate the effectiveness of the algorithm. We generated $I \in \{20, 50, 300\}$ stations, with $I = 20, 50$, and 300 corresponding to dataset suffixes $-S$, $-L$, and $-XL$, respectively. The distance between stations is sampled from $\text{Uniform}\{0, 1, \dots, 10\}$ for $-S$ datasets, from $\text{Uniform}\{0, 1, \dots, 30\}$ for $-L$ datasets and from $\text{Uniform}\{0, 1, \dots, 20\}$ for $-XL$ datasets. Then, the shortest path algorithm is applied. The capacity of each vehicle was set to 3, and the number of vehicles was $K \in \{5, 15, 50\}$. Travel cost per unit distance is set to 0 or 0.3, with the latter case marked by an additional suffix $-\text{cost}$. The origin and destination of each request were uniformly sampled from the I stations, and the request's appearance time was sampled from $\text{Uniform}\{1, 2, \dots, T\}$, where T is set to 58 for $-S$ datasets and 128 for $-L$ and $-XL$ datasets.

The profit for each request was the distance between its origin and destination.

- *Real-World Dataset:* We used the DHRD(Assylbekov et al. 2023) dataset, which consists of millions of food delivery requests from three cities: Taipei (tpe), Singapore (sg), and Stockholm (se). We treated each city as an independent dataset. In this dataset, we treated each geographical area as a station, with areas divided by 5-character geohash granularity. After excluding stations with almost no requests, each dataset had $I = 36$ stations. The distance between two areas was defined as the minimum number of areas that needed to be traversed. The capacity of each vehicle was set to 6, and the number of vehicles for the three cities was set to $K \in \{20, 15, 3\}$. In this dataset, we did not consider the *cost* of vehicle travel, only the total profit of completed requests. The origin and destination of each request were the stations corresponding to their respective geographical areas. We divided a day into $T = 48$ time slots, corresponding to T time steps in the simulation environment. The time step of a request’s appearance was determined by its real-time slot. The profit for each request is set to 1. We divided the dataset into 76 days for training and validation, and 14 days for testing.

D Computational Complexity Analysis

The computational efficiency of MAPT is crucial for its practical application in dynamic vehicle routing scenarios, where real-time decision-making is essential. Our architecture maintains the fundamental structure of Transformer models while introducing minimal additional overhead through the Relation-Aware Attention and Informative Prior modules. The encoder stage processes all entities (stations, requests, and vehicles) as a single sequence. For each encoder block with hidden dimension H , the complexity is dominated by the self-attention mechanism and feed-forward network operations. The self-attention computation scales as $O((I + M + K)^2 H)$ where I , M , and K represent the number of stations, requests, and vehicles respectively, while the feed-forward network contributes $O((I + M + K)H^2)$ per layer.

During decoding, we employ the KV-Cache technique commonly used in large language models to optimize autoregressive generation. This approach caches previously computed key-value pairs, significantly reducing redundant computations when processing sequential decisions. Each decoder block involves three main components: self-attention over the partially constructed solution (complexity $O((M + K)^2 H)$), cross-attention between the decoder inputs and encoder outputs ($O((M + K)(I + M + K)H)$), and the feed-forward network operations ($O((M + K)H^2)$).

So the overall computational complexity of MAPT is $O(L[(I + M + K)^2 H + (I + M + K)H^2])$, where L represents the number of transformer layers. While this demonstrates quadratic scaling with respect to the number of entities, the parallel processing capabilities of modern GPUs enable efficient computation even for large problem instances. This efficiency is particularly important for our target application, where the system must respond to dynamic requests

Scenio	Stations	Reqs.	Veh.	T	Cost	Profit
synthetic-S	20	110	5	58	0	distance
synthetic-S-cost	20	110	5	58	0.3	distance
synthetic-L	50	550	15	128	0	distance
synthetic-L-cost	50	550	15	128	0.3	distance
synthetic-XL	300	550	50	128	0	distance
dhrd-tpe	36	800	20	48	0	1
dhrd-sg	36	700	15	48	0	1
dhrd-se	36	200	3	48	0	1

Table 5: Statistics of experimental datasets. *cost* is the cost per unit distance; *profit* is the profit per request (either distance-based or fixed).

in real-time while maintaining solution quality.

E Training Algorithm

The main part of our training algorithm follows the original Proximal Policy Optimization (PPO) algorithm (Schulman et al. 2017), and we have incorporated some learning rate schedule strategies. We summarize our training process in Algorithm 1.

F Implementation Details

For the model architecture, we employ a Transformer with 6 encoder layers and 2 decoder layers. The hidden size is set to 128, and the number of attention heads is 2. Regarding the optimizer configuration, we utilize the Adam optimizer with a learning rate schedule that includes a warm-up phase and a decay phase. Specifically, the learning rate increases linearly during the first quarter of the total training steps, reaching the target learning rate, and subsequently undergoes linear decay for the remaining steps. For the Proximal Policy Optimization (PPO) algorithm, we set the discount factor $\gamma = 0.99$, the clipping coefficient $\epsilon = 0.2$, and the Generalized Advantage Estimation (GAE) parameter $\lambda = 0.99$. Critic loss coefficient α is set to 1. The number of PPO epochs is fixed at 1. The hyperparameter β is set to 0.03. During training, we monitor the model’s performance on the validation set and retain the best-performing model as the final model. Additional dataset-specific hyperparameters are provided in Table 6. All experiments are conducted using Python 3.10 and PyTorch 2.5.

G Difference between Non-AutoRegressive and AutoRegressive Decoding

We use a simple VRP example to illustrate the difference between the Non-AutoRegressive method with independent decoding and our AutoRegressive method in terms of decoding, as shown in Figure 5. Suppose two vehicles, v_1 and v_2 , are at the same location and need to collaboratively visit stations $n_1 \dots n_4$. The optimal solution is achieved when both vehicles move left and right respectively, while a suboptimal solution is obtained when they move up and down respectively. Let the actions of the two vehicles be A_{v_1} and A_{v_2} ,

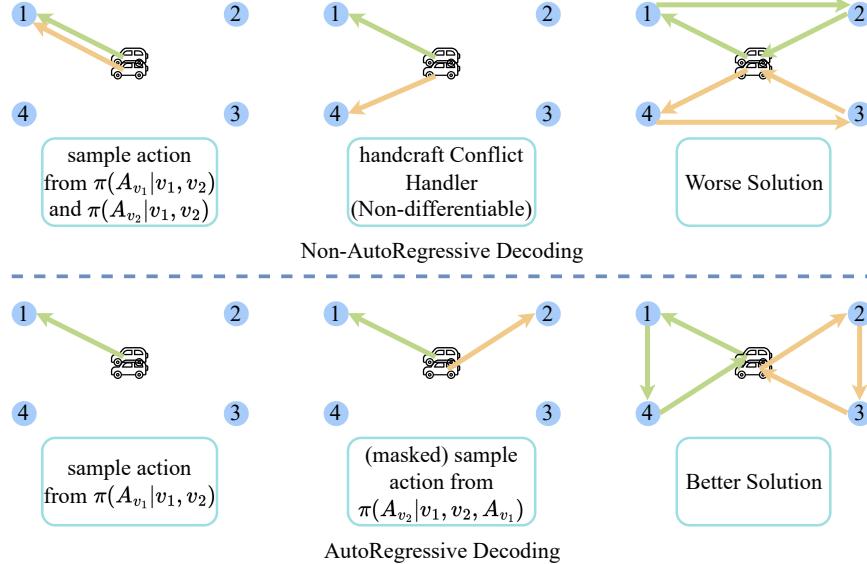


Figure 5: Comparison between Non-AutoRegressive Decoding(up) and AutoRegressive Decoding(down).

Dataset	Total	Rollout	MiniBatch	LR
synth-S	16K	256	256	1e-4
synth-S-cost	16K	256	256	1e-4
synth-L	4K	64	64	2e-5
synth-L-cost	4K	64	64	5e-5
synth-XL	4K	64	64	2e-5
dhrd-se	8K	256	256	2e-5
dhrd-tpe	2K	64	64	2e-5
dhrd-sg	2K	64	64	1e-5

Table 6: Dataset Specific Hyper-parameters. *Total* indicates the overall training budget. *Rollout* specifies the experience collection per iteration. *Minibatch* determines the SGD batch size. *LR* controls the policy update step size.

representing the next station they move to at the first step, and π represent a scheduling policy. However, for the Non-AutoRegressive independent decoding method, it can only model the probability distribution of each vehicle's individual actions $\pi(A_{v_1} | v_1, v_2)$ and $\pi(A_{v_2} | v_1, v_2)$, but fails to model the joint probability distribution of the two vehicles' actions $\pi(A_{v_1}, A_{v_2} | v_1, v_2)$. Consequently, it cannot guarantee that both vehicles move left and right respectively. Moreover, since a conflict handler is applied afterward, its gradient cannot be backpropagated, making it impossible to optimize the neural network parameters using conflict information when conflicts occur, which limits the capability of such methods. For AutoRegressive decoding, we first decode the action A_{v_1} of v_1 according to $\pi(A_{v_1} = n_1 \dots n_4 | v_1, v_2)$, then use A_{v_1} as an additional condition to decode the action A_{v_2} of v_2 , i.e., $\pi(A_{v_2} = n_1 \dots n_4 | v_1, v_2, A_{v_1})$. At this point, the policy can determine the direction of v_2 based

on whether A_{v_1} is left or right, thus achieving the optimal solution. Additionally, we can mask out A_{v_1} when decoding A_{v_2} to avoid conflicts.

H Does MAPT Overly Rely on Informative Prior?

To verify whether MAPT excessively relies on the Informative Prior, we compare the experimental results between MAPT and using only the Informative Prior as the decision-making algorithm, as shown in Table 7. In synthetic datasets, problem instances are randomly generated, making it difficult to form complex and realistic distributions. As a result, the generated problem instances tend to be simpler, leading to cases where the difference between the Informative Prior and MAPT is small. However, for real-world datasets, our model (MAPT) shows significant improvements over the Informative Prior, with enhancement rates of 14% for dhrd-tpe, 16% for dhrd-sg, and 19% for dhrd-se. In Figure 6, we show the distribution of request origins and destinations in the dhrd-tpe dataset, which differs significantly from the Uniform distribution of synthetic data. Reinforcement learning can learn more complex distributions based on the Informative Prior, which is why its improvement on real-world datasets is more significant than on synthetic datasets. Moreover, even for the large-scale synthetic-XL dataset, our model achieves a notable 37% improvement over the Informative Prior. These results demonstrate that our model does not overly rely on Informative Priors.

I Generalization

Since our decision-making does not strictly depend on the number of vehicles and requests, we expect our model to

Scenario	synth-S		synth-S-cost		synth-L		synth-L-cost		dhrd-tpe		dhrd-sg		dhrd-se		synth-XL	
Metric	Obj. ↑ Comp.	↑ Comp.	Obj. ↑ Comp.	↑ Comp.	Obj. ↑ Comp.	↑ Comp.	Obj. ↑ Comp.	↑ Comp.	Obj. ↑ Comp.	↑ Comp.	Obj. ↑ Comp.	↑ Comp.	Obj. ↑ Comp.	↑ Comp.	Obj. ↑ Comp.	↑ Comp.
Infor Prior	267.1	0.81	185.0	0.81	1858.9	0.79	1305.5	0.79	614.1	0.77	323.7	0.62	66.1	0.54	2340.7	0.47
MAPT	275.2	0.83	192.1	0.84	1875.1	0.80	1348.6	0.81	697.2	0.87	376.1	0.71	78.9	0.64	3227.5	0.65

Table 7: Performance comparison between Informative Prior and MAPT.

Algorithm 1: PPO Training Algorithm for MAPT

```

1: Initialize:
2:   Policy network  $\pi_\theta$  with parameters  $\theta$ .
3:   Replay buffer  $\mathcal{B}$ .
4:   Environment  $\mathcal{E}$ , episode length  $T$ .
5:   Learning rate scheduler with warmup and linear de-
   cay.
6:   Rollout iterations  $K$ , rollout threads  $N$ .
7: for iteration  $k = 1, 2, \dots, K$  do
8:   for thread  $n = 1, 2, \dots, N$  do
9:     Reset environment  $\mathcal{E}$  and initialize observations  $s_0$ .
10:    for step  $t = 0, 1, 2, \dots, T - 1$  do
11:      Collect Data:
12:        Sample actions  $a_t$  from policy  $\pi_\theta(a_t|s_t)$ .
13:        Execute actions  $a_t$  in environment  $\mathcal{E}$ .
14:        Observe next state  $s_{t+1}$  and reward  $r_t$ .
15:        Store transition  $(s_t, a_t, r_t, s_{t+1})$  in buffer  $\mathcal{B}$ .
16:      end for
17:    end for
18:    Compute Advantages:
19:      Compute advantages using Generalized Advantage
         Estimation (25).
20:    for epoch  $e = 1, 2, \dots, E$  do
21:      Optimize Surrogate Objective:
22:        Sample mini-batch of transitions from buffer  $\mathcal{B}$ .
23:        Compute critic loss by (28) and policy loss by
         (27).
24:        Update policy  $\pi_\theta$  using gradient descent on joint
         loss (29).
25:    end for
26:    Update Old Policy:
27:     $\theta_{\text{old}} \leftarrow \theta$ 
28:    update learning rate through scheduler.
29: end for

```

have some generalization ability. To this end, we conducted a series of experiments:

- **synthetic-S → synthetic-XL** represents the validation of whether a model trained on small-scale data can be used on large-scale problems.
- **synthetic-S → dhrd-se** represents the validation of whether a model trained on data from one distribution can generalize to other distributions.

The generalization results are shown in Table 8. The experiments show that models trained on small synthetic dataset can generalize to real-world dataset and large-scale dataset to some extent. This is due to two reasons: first, the use of Transformer, a model structure with good generalization ca-

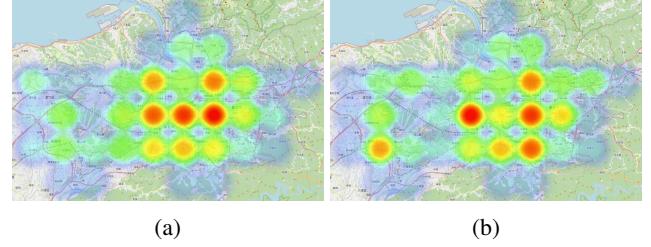


Figure 6: Heatmap of the distribution of request origins(a) and destinations(b) in the dhrd-tpe dataset, where darker colors indicate higher quantities.

Scenario	synth-S		dhrd-se		synth-XL	
Metric	Obj. ↑ Comp.	↑ Comp.	Obj. ↑ Comp.	↑ Comp.	Obj. ↑ Comp.	↑ Comp.
synth-S	275.2	0.83	74.4	0.61	2470.2	0.50
dhrd-se	117.4	0.36	78.9	0.64	1480.0	0.30
synth-XL	109.8	0.34	52.6	0.42	3227.5	0.65

Table 8: Generalization of MAPT. The row index indicates that the model is trained on the corresponding dataset, and the column index indicates that the model is evaluated on the corresponding dataset.

pabilities, and second, the inclusion of informative priors in the model, which provides guidance in unseen distributions. It is worth noting that when tested on real-world datasets and large-scale datasets, our method still outperforms the pure Informative Priors approach (see Tables 7 and 8), proving that our deep model has also learned some transferable knowledge, rather than relying solely on the Informative Priors for generalization. Although the model trained on the dhrd-se dataset performed best on this dataset, it performed poorly when generalized to synthetic datasets. We speculate that this is because the request OD distribution and station embeddings learned from the dhrd-se dataset are not suitable for synthetic datasets. However, there is still much to explore in the field of generalization as part of our future research.

J Baselines Implementation

Rolling Horizon The Rolling Horizon Policy addresses the dynamic nature of the Multi-Vehicle Dynamic Pickup and Delivery Problem by breaking it into smaller, static subproblems within a sliding time window, enabling the use of traditional optimization methods. At each decision point, the policy determines a planning horizon based on the remaining time frames and the maximum allowable planning horizon, ensuring computational tractability. The maximum

Dataset	Horizon	Replan
synth-S	20	10
synth-S-cost	20	10
synth-L	60	30
synth-L-cost	60	30
synth-XL	40	20
dhrd-se	30	15
dhrd-tpe	30	15
dhrd-sg	30	15

Table 9: Rolling Horizon Hyper-parameters. *Horizon* indicates the maximum allowable planning horizon. *Replan* specifies the re-planning interval.

allowable planning horizon used for each dataset is shown in Table 9, column *Horizon*. The problem within this window is formulated as a static instance, where vehicles, requests, and constraints (e.g., vehicle capacity, request appearance times, and preloaded requests) are fixed for the duration of the window. A solver, chosen from a set of available methods (e.g., **OR-Tools**, **Genetic Algorithm(GA)**, or **Simulated Annealing(SA)**), is then applied to optimize the assignment of requests to vehicles and their routes within the window. The solver’s solution is translated into actionable decisions, such as assigning requests to vehicles and determining the next hop for each vehicle, while ensuring feasibility with respect to constraints like vehicle capacity and request uniqueness. The policy periodically re-evaluates and re-plans at predefined intervals to adapt to new information and changing conditions. The re-planning interval used for each dataset is shown in Table 9, column *Replan*. By iteratively solving these static subproblems and updating the solution as the horizon shifts, the Rolling Horizon Policy effectively handles the dynamic and stochastic nature of the problem, balancing computational efficiency with solution quality.

OR-Tools We address the Multi-Vehicle Dynamic Pickup and Delivery Problem with Stochastic Requests using the Google OR-Tools CP-SAT solver. The problem involves a fleet of vehicles K , a set of requests M , and a set of stations N over a time horizon T . Each vehicle $k \in K$ has a starting location $start_k$, a capacity $capacity_k$, and a join time $jointime_k$ indicating when it becomes available for the first decision-making. Requests $m \in M$ appear at time $appear_m$ and must be picked up from $from_m$ and delivered to to_m . Some requests may be pre-loaded on vehicles, denoted by $preload_{k,m}$, and vehicles are en route until their join time, during which they cannot perform any actions. For the convenience of modeling, we assume that the volume of each request is 1.

Decision Variables:

- $pickup_{k,m,t}$: Binary variable indicating whether vehicle k picks up request m at time t .
- $delivery_{k,m,t}$: Binary variable indicating whether vehicle k delivers request m at time t .
- $load_{k,t}$: Integer variable representing the load of vehicle k

at time t .

- $loc_{k,t}$: Integer variable representing the location of vehicle k at time t .
- $noact_{k,t}$: (No Action) Binary variable indicating whether vehicle k is not performing any action at time t , e.g. the vehicle is en route.

Constraints:

- **Pre-Loaded Requests:** For pre-loaded requests, the pickup action is enforced at $t = -1$:

$$pickup_{k,m,-1} = preload_{k,m} \quad \forall k \in K, m \in M \quad (30)$$

$$load_{k,-1} = \sum_m preload_{k,m} \quad \forall k \in K \quad (31)$$

- **Load Dynamics:** The load of each vehicle is updated based on pickup and delivery actions:

$$load_{k,t} = load_{k,t-1} + \sum_{m \in M} pickup_{k,m,t} - \sum_{m \in M} delivery_{k,m,t} \quad \forall k \in K, t \in [0, T] \quad (32)$$

- **Location Dynamics:** The location of each vehicle is updated based on pickup, delivery, and no-action variables:

$$loc_{k,t} = \begin{cases} start_k & \text{if } t \leq jointime_k \\ loc_{k,t-1} & \text{if } noact_{k,t-1} = 1 \\ to_m & \text{if } delivery_{k,m,t} = 1 \\ from_m & \text{if } pickup_{k,m,t} = 1 \end{cases} \quad \forall k \in K, t \in [0, T], m \in M \quad (33)$$

- **No-Action Constraints:** If no pickup or delivery occurs at time t , the no-action variable is set to 1:

$$noact_{k,t} = 1 \iff \sum_{m \in M} pickup_{k,m,t} + \sum_{m \in M} delivery_{k,m,t} = 0 \quad \forall k \in K, t \in [0, T] \quad (34)$$

- **Request Assignment:** Each request must be picked up and delivered by at most one vehicle:

$$\sum_{k \in K} \sum_{t \in [-1, T]} pickup_{k,m,t} \leq 1 \quad \forall m \in M \quad (35)$$

$$\sum_{k \in K} \sum_{t \in [-1, T]} delivery_{k,m,t} \leq 1 \quad \forall m \in M \quad (36)$$

- **Pickup-Delivery Sequence:** A request must be picked up before it can be delivered:

$$\sum_{t' \in [-1, t]} pickup_{k,m,t'} = 1 \implies delivery_{k,m,t} = 1 \quad \forall k \in K, m \in M, t \in [-1, T] \quad (37)$$

- **Capacity Constraints:** The load of each vehicle must not exceed its capacity:

$$load_{k,t} \leq capacity_k \quad \forall k \in K, t \in [0, T] \quad (38)$$

- **Time-Distance Constraints:** The time between consecutive actions must be sufficient to travel between locations:

$$t_1 - t_0 \geq dist_{loc_k, t_0, loc_k, t_1} \quad (39)$$

$$\forall k \in K, t_0, t_1 \in [0, T], t_0 < t_1$$

- **Join Time Constraints:** Vehicles cannot perform any actions before their join time:

$$noact_{k,t} = 1 \quad \forall k \in K, t \in [0, jointime_k - 1] \quad (40)$$

Objective Function: The objective is to maximize the total profit of delivered requests while minimizing travel costs:

$$\begin{aligned} \text{Maximize} \quad & \sum_{m \in M} \sum_{k \in K} \sum_{t \in [0, T]} delivery_{k,m,t} \cdot value_m \\ & - \sum_{k \in K} \sum_{t \in [0, T-1]} cost_{loc_k, t, loc_k, t+1} \end{aligned} \quad (41)$$

Simulated Annealing The Simulated Annealing (SA) algorithm for solving the Multi-Vehicle Dynamic Pickup and Delivery Problem with Stochastic Requests starts by generating an initial solution through random request assignments to vehicles, ensuring constraints such as initial positions, displacement, location consistency, request uniqueness, capacity, request appearance times, and vehicle start times are satisfied. To construct a neighbor solution, the algorithm randomly selects a request and removes its current assignment from the vehicle's path. Then the selected request is reassigned to a new vehicle, and both pickup and delivery times are randomly selected within their respective feasible windows, ensuring the pickup occurs after the request's appearance time and the delivery occurs after the pickup plus the travel time. The annealing process iteratively explores the solution space by accepting better solutions or probabilistically accepting worse solutions based on the current temperature, which gradually decreases according to a predefined cooling rate. The initial temperature is set to 1000, the final temperature is set to 1, the cooling rate is set to 0.99, and the maximum number of iterations is set to 5000.

Genetic Algorithm The Genetic Algorithm (GA) for solving the Multi-Vehicle Dynamic Pickup and Delivery Problem with Stochastic Requests begins by generating an initial population of feasible solutions, where requests are randomly assigned to vehicles while ensuring constraints such as initial positions, displacement, location consistency, request uniqueness, capacity, request appearance times, and vehicle start times are satisfied. Each individual in the population represents a solution, and its fitness is evaluated using an exponential transformation of the objective function, which combines the total profit of served requests and the cost of vehicle movements. The algorithm evolves the population through selection, crossover, and mutation operations. During selection, individuals are chosen probabilistically based on their fitness, favoring better solutions. The crossover operation randomly selects a request and swaps its assignment between two parent solutions, ensuring feasibility by updating vehicle locations and loads. The mutation operation randomly selects a request and reallocates it

to a different vehicle or adjusts its pickup and delivery times, maintaining feasibility through constraint checks. The population iteratively evolves over a predefined number of generations, with the best solution being tracked and updated based on fitness. The population size is set to 10, and the number of generations is set to 500.