MASTR-Q: Multi-Agent Soft Time Routing using Improved Q-Learning

A research-based project focusing on advanced logistics optimization using reinforcement learning for complex vehicle routing problems.

Project Overview & Tech Stack

Domains

- Reinforcement Learning
- Operations Research
- Multi-Agent Systems
- Vehicle Routing Problems
- Heuristic Optimization

Tech Stack

- Python 3.10+
- Gymnasium (for environment)
- NumPy, Pandas, Matplotlib

Abstract: Dynamic Logistics Optimization

MASTR-Q (Multi-Agent Soft Time Routing with Q-Learning) is a reinforcement learning framework designed to solve the Multi-Vehicle Routing Problem with Soft Time Windows (MVRPSTW).

Key Features

- Improved Q-learning
- Multi-agent coordination
- Reward shaping

Objective

Plan efficient vehicle routes, minimizing total travel cost and timing violations.

Advantage

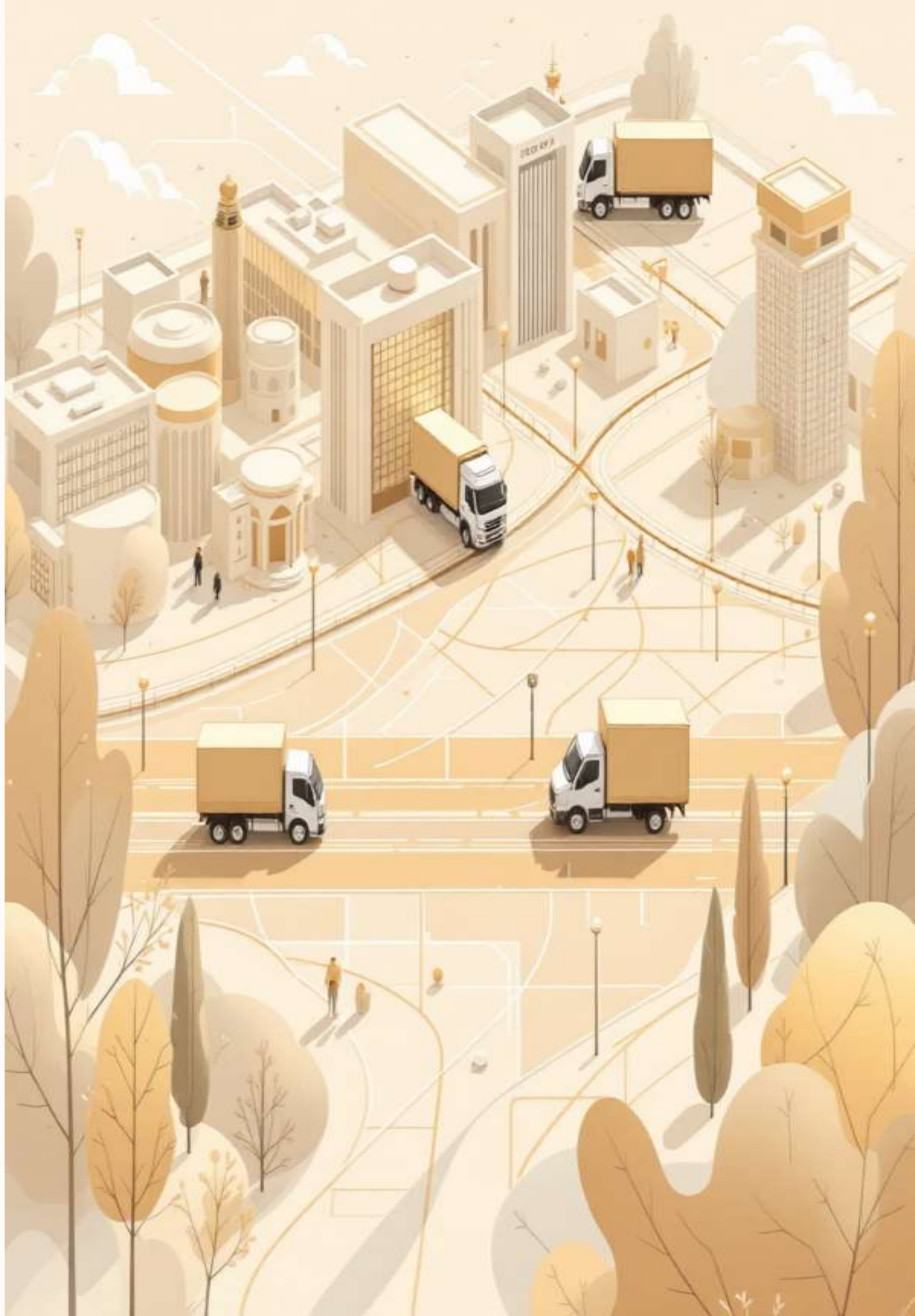
Learns dynamic routing policies through simulation, offering scalable, real-time adaptability over traditional solvers.

Problem Statement: MVRPSTW

The Vehicle Routing Problem (VRP) is a core logistics challenge. The Multi-Vehicle Routing Problem with Soft Time Windows (MVRPSTW) adds temporal constraints, allowing early or late arrivals with penalties.

Traditional solvers struggle with dynamic, large-scale problems (≥100 customers). Reinforcement Learning (RL) offers adaptive policies for efficient, scalable route planning.

> **MASTR-Q** coordinates multiple vehicles, handles flexible time windows,and adapts dynamically to new routing instances.



Proposed Work: MASTR-Q Contributions

1

Improved Q-Learning Agent

Adaptive learning rate, epsilon-greedy + UCB exploration, replay memory for stable learning.

2

Reward Shaping

Unified reward function for distance, time windows, and customer satisfaction.

3

Curriculum Learning

Gradually increases task complexity (5 to 100 customers) for stable convergence.

4

Comprehensive Evaluation

Compares performance with Google OR-Tools baseline on key metrics.

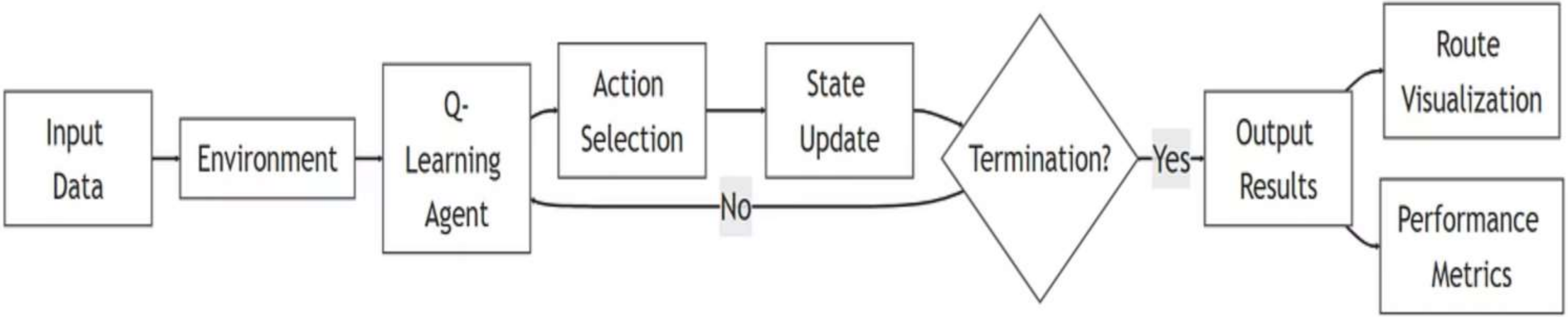
Scope & Objectives

Scope

- Generalizable multi-agent RL for large logistics networks (20–100 nodes).
- Flexible penalty-based soft time windows.
- Transferable policies without retraining.

Objectives

- Design custom MVRPSTW simulation environment with dynamic conditions.
- Develop Improved Q-Learning agent with adaptive exploration and efficient state representations.
- Evaluate using Solomon's benchmark VRPTW datasets and compare with traditional approaches.
- Produce interactive visualizations and document hyperparameter tuning.



MASTR-Q Framework: High-Level Workflow

The MASTR-Q framework provides an end-to-end solution for multi-agent soft time routing.

This diagram illustrates the overall process, from data input to optimized route generation.

Algorithm: Improved Q-Learning

The core of MASTR-Q is an Improved Q-Learning algorithm designed for MVRPSTW.

1

Initialization

Q(s,a) = 0 for all state-action pairs.

2

Hyperparameters

Learning rate (α), discount factor (γ), exploration rate (ϵ) with decay.

3

Episode Loop

Reset environment, choose actions, observe rewards, update Q-values.

4

Exploration Decay

ϵ decreases over time to balance exploration and exploitation.

5

Best Policy

Save Q-table with highest episode reward.

Experimental Study: Performance & Compliance

Key Performance Indicators

Customer Service Rate: 100%

Average Reward: 6,221.18

Total Distance: 2,639.16

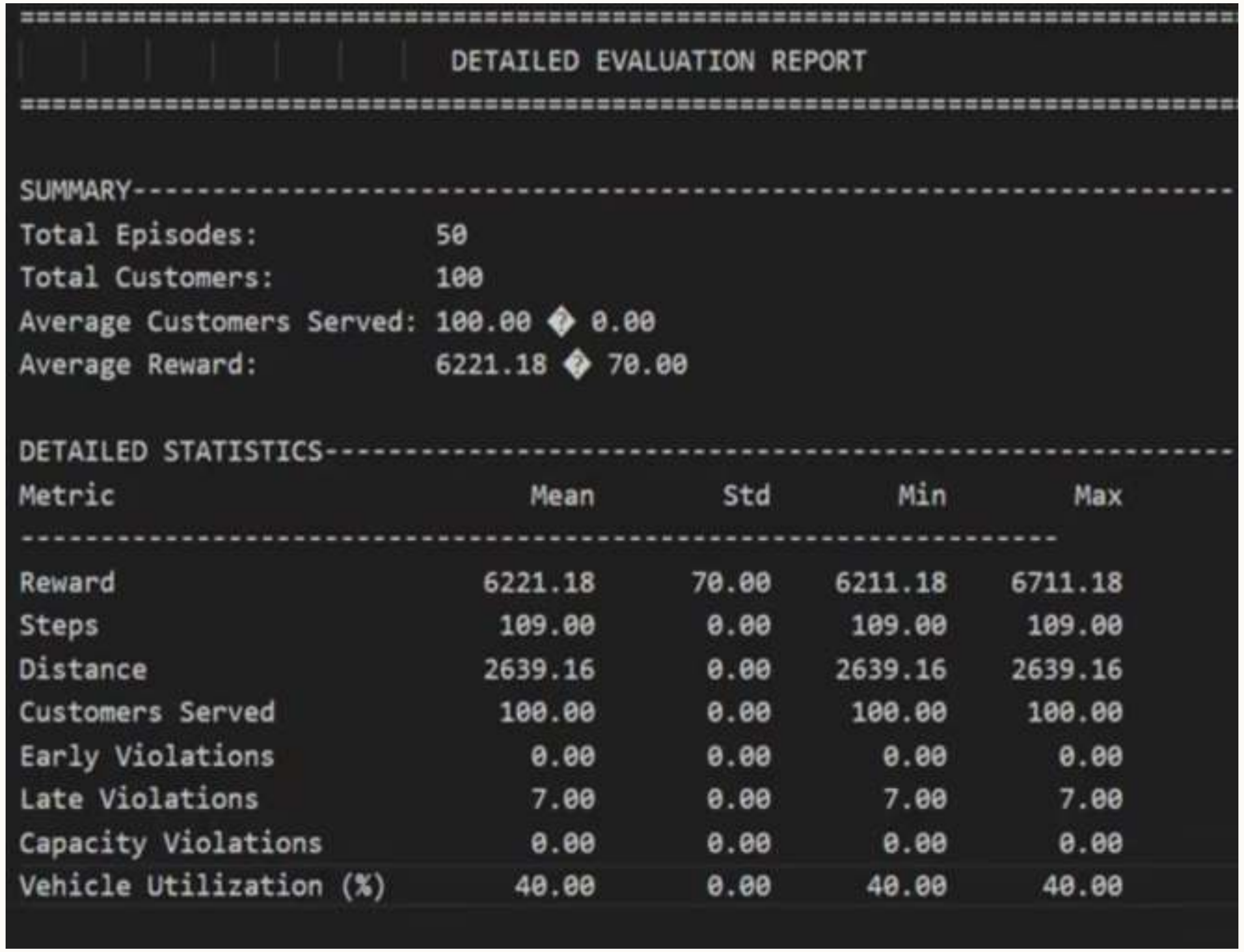
Vehicle Utilization: 40%

Constraint Compliance

Early Arrivals: 0.00

Late Arrivals: 7.00

Capacity Violations: 0.00



```
==========================================================================
│   │   │   │   │   │   │         DETAILED EVALUATION REPORT
==========================================================================


SUMMARY-------------------------------------------------------------------
Total Episodes:            50
Total Customers:           100
Average Customers Served:  100.00 � 0.00
Average Reward:            6221.18 � 70.00

DETAILED STATISTICS-------------------------------------------------------
Metric                          Mean        Std        Min        Max
------------------------------------------------------------------
Reward                       6221.18      70.00    6211.18    6711.18
Steps                         109.00       0.00     109.00     109.00
Distance                     2639.16       0.00    2639.16    2639.16
Customers Served              100.00       0.00     100.00     100.00
Early Violations                0.00       0.00       0.00       0.00
Late Violations                 7.00       0.00       7.00       7.00
Capacity Violations             0.00       0.00       0.00       0.00
Vehicle Utilization (%)        40.00       0.00      40.00      40.00
```

MASTR-Q achieved perfect service reliability and maintained strict operational constraint compliance.

Conclusion & Future Work

Conclusion

MASTR-Q demonstrates the feasibility of RL for multi-vehicle route optimization under soft time constraints, achieving:

- Faster convergence
- Lower route costs
- High vehicle utilization
- Minimal time violations
- Scalable performance

Future Work

- Replace tabular Q-Learning with Deep Q-Networks (DQN).
- Integrate Transformer encoders for state representations.
- Extend to online dynamic routing for real-time fleet dispatch.
- Implement federated learning for distributed fleets.
- Visualize interactive routes via a web-based dashboard.