# MODRAL: Two-Phase PPO-Guided ALNS for Multi-Center Home Healthcare Routing with Time-Window Satisfaction

Yongqi Liu
*Faculty of Computer Science and Technology*
*Qilu University of Technology*
Jinan, China
10431240168@stu.qlu.edu.cn

Ying Li*
*School of Data and Computer Science*
*Shandong Women's University*
Jinan, China
liying@sdwu.edu.cn

Jie Tian*
*School of Data and Computer Science*
*Shandong Women's University*
Jinan, China
29052@sdwu.edu.cn

Hui Liu
*School of Data and Computer Science*
*Shandong Women's University*
Jinan, China
27069@sdwu.edu.cn

Shengpeng Yu
*School of Data and Computer Science*
*Shandong Women's University*
Jinan, China
ysp@sdwu.edu.cn

Zhen Cui
*School of Data and Computer Science*
*Shandong Women's University*
Jinan, China
202401063@sdwu.edu.cn

***Abstract*—Multi-center home healthcare routing under tight and heterogeneous time windows poses significant challenges: the feasible region is sparse and fragmented, and the effectiveness of destroy–repair operators is highly state-dependent, limiting static or weakly adaptive policies. We study a bi-objective variant where patient satisfaction is defined by arrival times relative to both hard and preferred time windows. To address these challenges, we propose MODRAL, a two-phase framework that decouples structure-aware initialization from learning-guided improvement. In Phase-1, a diverse pool of feasible solutions is constructed using time-window-prioritized heuristics that consider constraint tightness and spatial distribution. In Phase-2, Adaptive Large Neighborhood Search (ALNS) is enhanced with a Proximal Policy Optimization (PPO) agent that learns a state-aware operator scheduling policy, selecting destroy–repair pairs based on solution quality, search progress, and constraint satisfaction. A global external archive with dominance filtering preserves non-dominated solutions. Experiments on adapted Cordeau MDVRPTW instances show that the proposed initialization improves Pareto-front coverage and robustness, while the learned policy yields additional gains on tightly constrained instances by balancing exploration and exploitation. These results demonstrate the effectiveness of combining structured initialization with state-aware operator control for strongly constrained multi-objective routing.**

***Index Terms*—Multi-objective optimization, home healthcare routing, adaptive large neighborhood search, PPO, operator scheduling.**

## I. Introduction

The multi-center home healthcare routing problem requires optimizing operational cost and patient satisfaction under strict and highly heterogeneous service time windows. Compared with classical vehicle routing problems, time windows in this setting are often tighter and unevenly distributed across patients, while uncertainty in travel times further compresses the feasible solution space. In practice, when route structures are poorly aligned with time window distributions, the search process frequently encounters time-window violations, causing many candidate solutions to be discarded early. Consequently, obtaining high-quality feasible solutions is itself challenging [1], [2].

Various solution frameworks have been proposed for complex vehicle routing problems. Adaptive Large Neighborhood Search (ALNS) iteratively improves solutions through diverse destroy and repair operators [3], while multi-objective evolutionary and decomposition-based algorithms, such as NSGA-II and MOEA/D, explore the objective space by maintaining solution diversity [4], [5]. More recently, deep reinforcement learning methods have been applied to directly construct routing solutions [6], [7]. However, when strict and highly unbalanced time window constraints are present, the performance of these approaches often becomes unstable [8], [9].

In addition, recent studies have extended home healthcare routing to more realistic and complex settings, including stochastic demand, human behavior modeling, multi-center coordination, and integrated decision-making [10]–[16].

This instability mainly arises from two factors. First, strong time window constraints rapidly shrink the local improvement neighborhood, making the search highly sensitive to the structure of initial solutions. Poor time allocation or ordering in initial routes often limits subsequent improvement without violating feasibility. Second, the effectiveness of destroy–repair operators strongly depends on the current solution state. Aggressive operators may introduce hard-to-repair conflicts when time windows are saturated, whereas conservative operators may yield limited gains when the solution is relatively loose. Operator selection strategies based on fixed rules or historical statistics generally fail to capture this state dependence, leading to search stagnation. Motivated by these observations, this paper integrates Proximal Policy Optimization (PPO) [17] into the ALNS framework to learn solution-state-aware operator scheduling strategies.

From a methodological perspective, existing approaches still exhibit limitations for strongly constrained multi-objective routing problems. Traditional metaheuristics rely on coarse-grained feedback for operator selection, which becomes ineffective in highly fragmented feasible regions. Although multi-objective evolutionary algorithms provide strong global exploration capabilities, their generic crossover and mutation operators lack problem-specific guidance, generating many infeasible solutions under strict time window constraints and reducing search efficiency. While reinforcement learning-based combinatorial optimization has advanced rapidly, most studies focus on constructive heuristics, with limited attention to learning the internal decision-making processes of improvement-based heuristics such as operator scheduling in ALNS.

Based on these considerations, this paper proposes a two-phase solution framework named MODRAL, which explicitly separates structured construction from adaptive improvement. In the first phase, a time-window-tightness-oriented heuristic incorporates time window characteristics and patient spatial distribution to generate a diverse set of feasible initial solutions with reasonable structures. In the second phase, a PPO-driven agent is embedded into the ALNS framework to dynamically select suitable destroy–repair operator combinations according to solution quality, search progress, and constraint satisfaction. An external archive is maintained to preserve the non-dominated solutions obtained during the search.

The main contributions of this paper are summarized as follows:

- A principled two-phase optimization framework (MODRAL) that decouples feasible structure construction from learning-guided improvement. By separating the generation of diverse, high-quality initial solutions (Phase-1) from adaptive iterative refinement (Phase-2), the framework reduces search complexity in highly constrained spaces and ensures that the learning component operates from a strong structural foundation.
- A time-window-prioritized heuristic initialization strategy that explicitly leverages domain knowledge—including constraint tightness, temporal urgency, and spatial distribution—to construct a structurally diverse candidate pool. Unlike conventional random or purely cost-driven initialization, this strategy generates solutions spanning different trade-off regions in the objective space, thereby stabilizing subsequent multi-objective search under tight time-window regimes.
- A PPO-based operator scheduling mechanism that models destroy–repair operator selection as a sequential decision process. We design a compact yet informative state representation encoding solution quality, iteration progress, and constraint violation intensity. The learned policy dynamically selects operator pairs based on the current search state, transitioning ALNS from heuristic-driven to intelligence-driven scheduling. Empirical analysis reveals that the policy adapts its exploration–exploitation balance across search stages, yielding pronounced advantages on tightly constrained instances.

The remainder of the paper is organized as follows. Section II defines the problem and benchmark. Section III presents the MODRAL framework, including Phase-1 construction and Phase-2 DRL-ALNS. Section IV reports experimental results and ablations. Section V concludes with discussions and future work.

## II. Problem Description and Benchmark

### A. Problem Formulation

We consider a multi-depot home healthcare routing problem with time windows (MDHHRP-TW). Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a complete graph where $\mathcal{V} = \mathcal{C} \cup \mathcal{P}$. The vertex set consists of a set of service centers (depots) $\mathcal{C} = \{0, \ldots, M-1\}$ and a set of patients $\mathcal{P} = \{M, \ldots, M+N-1\}$. Each patient $i \in \mathcal{P}$ has a service duration and a benchmark-provided hard time window $[a_i, b_i]$. To model service quality, we further define a preferred (soft) time window $[\ell_i, r_i]$ with $a_i \le \ell_i \le r_i \le b_i$ [18].

A solution is represented as a collection of routes $\mathcal{R} = \{R_1, \ldots, R_K\}$. Each route $R_k$ originates and terminates at a specific center. In our implementation, we use a "one center = one route" representation: each center maintains at most one (possibly empty) route, and $K$ equals the number of non-empty center routes. A solution is considered **feasible** if it satisfies the following constraints:

**Full Coverage**: Every patient must be served exactly once. Formally,

$$\bigcup_{k=1}^{K} R_k = \mathcal{P} \quad \text{and} \quad R_k \cap R_{k'} = \emptyset \quad \text{for all } k \neq k'.$$

**Working Time Constraint**: Each route must respect the maximum allowable working time per shift. Let $T_k$ denote the total duration of route $k$, including travel, service, and waiting time. The constraint is defined as $T_k \le H$, where $H$ is the maximum working time per shift (e.g., 480 minutes).

Time window violations do not render a solution infeasible; instead, they are penalized through the satisfaction objective.

### B. Objectives and Evaluation

The problem is modeled as a bi-objective optimization minimizing operational cost ($f_1$) and patient dissatisfaction ($f_2$).

*1) Operational Cost ($f_1$):* The first objective minimizes the total operational expenditure, combining travel distance and fixed infrastructure usage. Let $k$ denote the number of non-empty center routes (i.e., activated centers). The cost is defined as:

$$f_1 = \sum_{R \in \mathcal{R}} \sum_{(u,v) \in R} d_{uv} + k \cdot C_{\text{fixed}} \tag{1}$$

where $d_{uv}$ denotes the Euclidean distance between nodes $u$ and $v$, and $C_{\text{fixed}}$ represents the fixed activation cost per center (set to 100 in our experiments).

*2) Patient Dissatisfaction ($f_2$):* The second objective focuses on service quality, quantified by the vehicle's arrival time $t_i$ at patient $i$. We define a satisfaction function $s(t_i) \in [0, 10]$ as a trapezoidal measure based on the preferred window:

$$s(t_i) = \begin{cases} 10, & \ell_i \le t_i \le r_i, \\ 10 \cdot \frac{t_i - a_i}{\ell_i - a_i}, & a_i \le t_i < \ell_i, \\ 10 \cdot \frac{b_i - t_i}{b_i - r_i}, & r_i < t_i \le b_i, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

If a vehicle arrives early ($t_i < \ell_i$), it waits until $\ell_i$ to start service, but satisfaction $s(t_i)$ is computed from arrival time $t_i$, penalizing earliness.

The minimization objective $f_2$ is defined as the average dissatisfaction across all patients:

$$f_2 = 10 - \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} s(t_i) \tag{3}$$

### *C. Benchmark Instances*

We evaluate the proposed framework using adapted instances from the **Cordeau MDVRPTW benchmark suite** [19]. Experiments are conducted on two representative instances, **pr01** and **pr11**, selected to illustrate performance under wider and tighter time-window regimes respectively. These instances characterize a diverse range of problem complexities:

- **Scale**: Service centers $M$ range from 4 to 10; patients $N$ from 48 to 288.
- **Constraints**: Instances pr01–pr10 feature wider time windows, while pr11–pr20 impose stricter scheduling limitations.

We quantify constraint level using the *Time-Window Tightness* metric $\tau$, defined as the average hard time window width:

$$\tau = \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} (b_i - a_i) \tag{4}$$

Table I illustrates this contrast: pr11 exhibits substantially lower $\tau$ than pr01, indicating a narrower feasible search space and higher optimization difficulty.

TABLE I: Time-window tightness statistics (hard windows). Smaller width indicates tighter constraints.

| Instance | $\lvert\mathcal{P}\rvert$ | Mean Width ($\tau$) | Median Width | Min Width |
|---|---|---|---|---|
| pr01 | 48 | 266.52 | 262.00 | 64.00 |
| pr11 | 48 | 172.12 | 168.50 | 54.00 |

## III. Algorithm Design

### *A. Overall Framework*

To address the multi-center home healthcare routing problem with strong time window constraints, we propose a two-phase optimization framework named MODRAL. The framework follows a *"structure-first, adaptive-optimization"* principle. In Phase-1, a diverse pool of feasible initial solutions is generated by exploiting problem-specific structural characteristics. In Phase-2, a deep reinforcement learning driven adaptive large neighborhood search (DRL-ALNS) is conducted on top of this pool to iteratively expand the global Pareto front.

### *B. Phase-1: Time-Window-Oriented Heuristic Initialization*

*1) Design Motivation:* Phase-1 implements our contribution (2): a time-window-prioritized heuristic initialization that explicitly leverages domain signals (constraint tightness, temporal urgency, spatial distribution) to construct a structurally diverse and feasible candidate pool. Rather than attempting to approximate the Pareto front, Phase-1 focuses on producing multi-start inputs with varied center allocations and temporal patterns so that the learning-guided Phase-2 starts from high-quality, feasible foundations.

*2) Heuristic Construction Strategies:* To realize this, Phase-1 combines three complementary constructors: (i) global time-window ordering (satisfaction-first insertion), (ii) clustering with targeted insertion (spatial and temporal grouping), and (iii) randomized greedy diversification. Controlled perturbations and seed variation increase structural coverage across different trade-off regions. All Phase-1 solutions are re-evaluated under a unified cost/capacity setting (including fixed center cost $C_{\text{fixed}}$ and nominal $Q_{\max}$) and only full-coverage feasible candidates are forwarded to Phase-2.

*a) Clustering and Insertion-Based Constructions:* These heuristics complement the global ordering strategy by covering other regions of the trade-off space:

- **Geographical clustering:** patients are grouped based on spatial proximity and assigned to nearby centers to reduce travel cost.
- **Time-window clustering:** patients are clustered according to time-window similarity to reduce temporal conflicts.
- **Greedy insertion:** patients are inserted based on minimum cost increase, providing a stable baseline structure.
- **Time-window-first insertion:** locally prioritizes patients with tight time windows while enforcing hard feasibility checks to reduce severe violations.

*b) Randomized Greedy Diversification:* To avoid over-concentration on a limited set of structural patterns, controlled randomness is introduced into certain insertion decisions. This randomized greedy mechanism significantly improves diversity in the candidate pool and enhances coverage of the solution space.

*3) Multi-start Output and Candidate Pool Design:* Phase-1 outputs the entire set of feasible candidate solutions after coverage and feasibility screening, instead of retaining only non-dominated solutions. All feasible solutions are passed to Phase-2 as multiple starting points. We denote the Phase-1 candidate pool by $\mathcal{P}_{init}$. For archive seeding we may compute the non-dominated subset $\mathcal{A}_{init} = \text{ND}(\mathcal{P}_{init})$, but Phase-2 receives the full pool $\mathcal{P}_{init}$ as multi-start inputs; the global

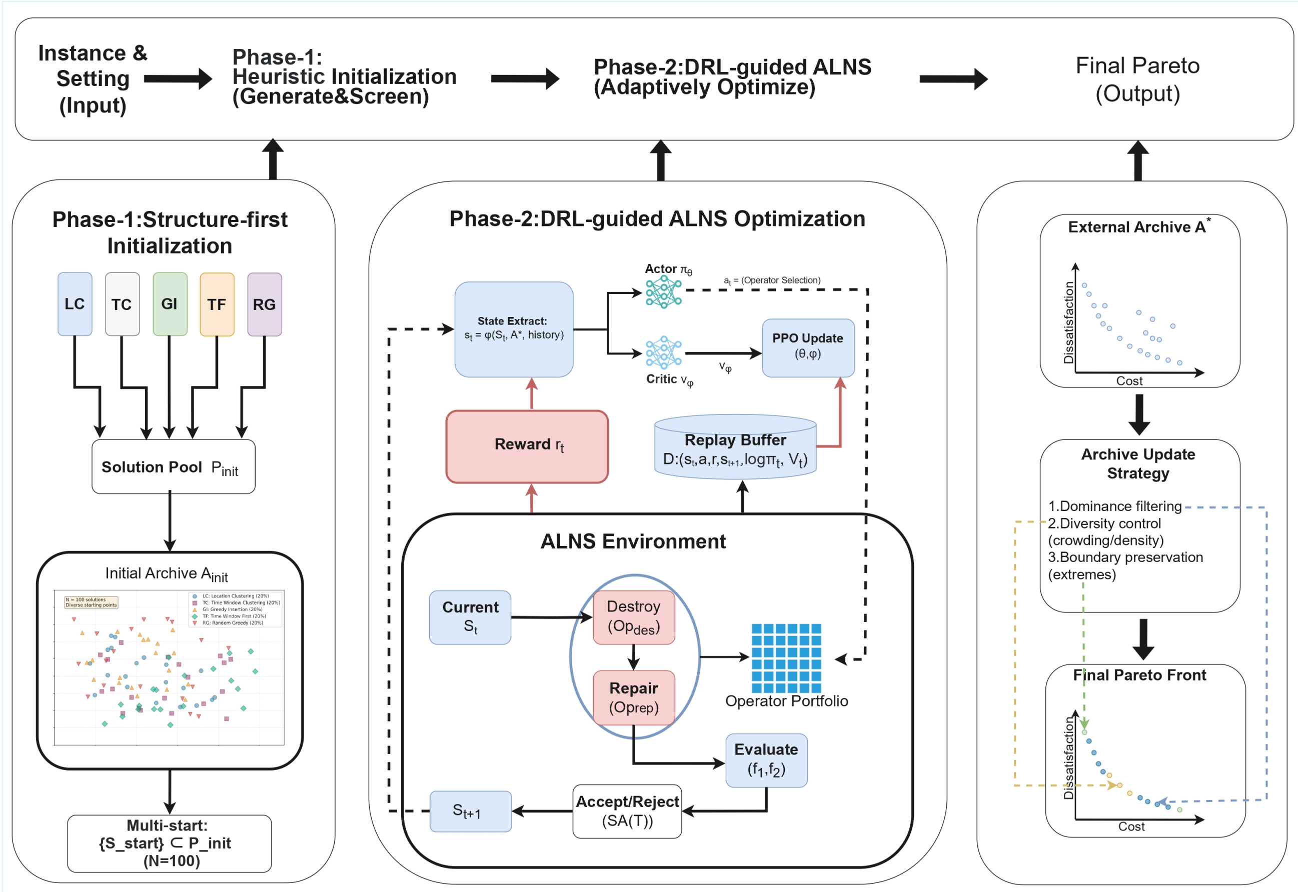



Fig. 1: Overall Framework: Two-Phase PPO-Guided ALNS for Multi-Center Home Healthcare Routing.

external archive used during Phase-2 is denoted $\mathcal{A}^*$ and is initialized (optionally) from $\mathcal{A}_{init}$.

This design differs from common elite-only initialization strategies. Some solutions that are dominated at the end of Phase-1 may still contain structural features that can be amplified by local search in Phase-2, such as center allocation patterns or service orderings. Preserving these solutions allows Phase-2 to explore multiple promising directions simultaneously and improves the robustness of Pareto front expansion.

In addition, Phase-1 supports an adaptive center-number mechanism. Candidate solutions are generated under multiple center configurations and merged into a unified pool, enabling natural competition and selection among different cost-satisfaction-allocation trade-offs during Phase-2.

### C. Phase-2: DRL-ALNS Multi-objective Optimization

*1) DRL-Guided Optimization Architecture:* Phase-2 runs ALNS as the environment and a PPO agent as the decision maker. Each ALNS iteration corresponds to a DRL step: the agent observes a low-dimensional state (solution quality, iteration progress, violation intensity, recent operator outcomes, archive statistics) and selects one of the 36 destroy–repair pairs. The learned policy receives reward signals biased toward archive contribution and immediate improvement, while simulated annealing acceptance preserves controlled exploration. An external Pareto archive is maintained online to record non-dominated solutions and to provide long-term training signals for the agent.

*a) State Perception and Feature Extraction:* To enable intelligent decisions, the system first transforms the raw solution space into a learnable format. At each iteration $t$, a 20-dimensional feature vector $s_t$ is extracted from the current solution $S_t$. This vector explicitly encodes:

- **Search Status:** Current temperature, iteration progress, and stagnation counters.
- **Objective Features:** Total travel cost, route smoothness, and average load factors.
- **Constraint Features:** Time-window violation intensity and unserved patient ratios.

This perception module ensures the agent can distinguish between different search phases (e.g., exploration vs. exploitation).

*b) Dual-Network Decision Mechanism:* The core intelligence is embodied in an Actor-Critic dual-network structure, as illustrated in the central block of Fig. 1:

- **Actor Network (Performer):** Taking state $s_t$ as input, the Actor network outputs a probability distribution $\pi(a_t|s_t)$ over all possible destroy-repair operator pairs.

The action $a_t$ is sampled from this distribution, balancing learned preferences with stochastic exploration.

- **Critic Network (Judge):** Operating in parallel, the Critic network estimates the value function $V(s_t)$ of the current state. This value serves as a baseline to evaluate the quality of the selected action, reducing the variance of the policy gradient updates.

*c) Environment Interaction and Feedback:* The selected composite action $a_t = (Op_{des}, Op_{rep})$ is executed in the ALNS environment. The destroy operator removes a subset of patients, and the repair operator reinserts them to generate a candidate solution $S_{new}$. A reward $r_t$ is computed using a tiered piecewise scheme. At a conceptual level,

$$r_t = w_c \Delta_{cost} + w_s \Delta_{sat} + \beta \mathbb{I}_{Pareto},$$

where $\Delta_{cost}$ and $\Delta_{sat}$ are objective improvements, and $\mathbb{I}_{Pareto}$ indicates Pareto-archive contribution. Implementation uses explicit reward tiers: infeasible solutions receive $-5$; new archive minimum cost receives $+5$; new archive maximum satisfaction receives $+6$ (with $+1$ extra if triggered by satisfaction-oriented repair); Pareto-dominating improvement receives $+3$; accepted moves receive $+1$ (or $+1.5$ near archive boundaries); rejected but boundary-near moves receive $+0.5$; otherwise $-0.5$.

*2) ALNS Operator Pool Design:* The operator pool consists of six destroy and six repair operators, yielding $6 \times 6 = 36$ action combinations.

*a) Destroy Operators:* Random Removal (diversity), Worst-Cost Removal (cost reduction), Worst-Satisfaction Removal (quality focus), Time-Window Violation Removal (feasibility recovery), High-Impact Center Removal (workload adjustment), Center-Close Removal (center reduction).

*b) Repair Operators:* Greedy Cost Repair (cost efficiency), Regret-2 Repair (look-ahead), Satisfaction-Aware Repair (quality focus), Time-Window First Repair (feasibility), Center-Aware Regret Repair (balancing), Center-Activation Repair (capacity expansion; activates an additional center only if the resulting route satisfies the working time constraint $T_k \leq H$).

This structured pool enables high-level strategies, e.g., pairing Center-Close Removal with Greedy Cost Repair for center reduction, or Time-Window Violation Removal with Satisfaction-Aware Repair for service quality restoration.

*3) Policy Optimization and Loop Closure:* A distinct feature of Phase-2 is its self-evolving capability: the DRL agent improves its policy via a rigorous feedback loop. Trajectory tuples $(s_t, a_t, r_t, s_{t+1})$ are stored in an on-policy buffer; periodic PPO updates reinforce operator pairs that yield consistent improvements under specific search states. The policy is trained online for each instance independently, as instance-specific features limit direct transferability.

Finally, the loop is closed by the Metropolis acceptance criterion and the external Pareto archive update. The archive maintains non-dominated solutions (size limit 100) with dominance filtering; solutions with incomplete coverage or satisfaction below 6.0 are rejected. When exceeding capacity, diversity preservation employs crowding distance [4] in normalized 2D objective space, with boundary and Phase-1 protected solutions retained. This closed-loop interaction is illustrated in Fig. 1.

**Algorithm 1** MODRAL Two-Phase Optimization Framework

**Require:** Instance $\mathcal{I}$, population size $N$, max iterations $T_{max}$
**Ensure:** Final Pareto archive $\mathcal{A}^*$

**// — Phase 1: Structured Initialization —**
$\mathcal{P}_{init} \leftarrow \emptyset$
**for** $i = 1$ **to** $N$ **do**
  $S_i \leftarrow$ ConstructSolution($\mathcal{I}$) {Mix: 40% global ordering, 30% clustering, 30% randomized greedy}
  **if** Feasible($S_i$) **then**
    $\mathcal{P}_{init} \leftarrow \mathcal{P}_{init} \cup \{S_i\}$
  **end if**
**end for**
$\mathcal{A}^* \leftarrow$ NonDominatedSort($\mathcal{P}_{init}$)

**// — Phase 2: DRL-guided ALNS —**
Initialize PPO agent $(\pi_\theta, V_\phi)$ and replay buffer $\mathcal{D}$
**for** each $S_{start} \in \mathcal{P}_{init}$ **do**
  $S \leftarrow S_{start}$, $T \leftarrow T_0$
  **for** $t = 1$ **to** $T_{max}$ **do**
    $\mathbf{s} \leftarrow$ StateVector($S$) {Perception}
    $a \sim \pi_\theta(\mathbf{s})$ {Operator selection}
    $S' \leftarrow$ ApplyOperators($S, a$) {Destroy-repair}
    $r \leftarrow$ Reward($S', S, \mathcal{A}^*$) {Improvement & archive contribution}
    Store transition $(\mathbf{s}, a, r, \mathbf{s}')$ in $\mathcal{D}$
    **if** MetropolisAccept($S', S, T$) **then**
      $S \leftarrow S'$
      UpdateArchive($\mathcal{A}^*, S'$)
    **end if**
    Periodically update $(\theta, \phi)$ via PPO using $\mathcal{D}$
    $T \leftarrow T \cdot \gamma$ {Cooling}
  **end for**
**end for**
**return** $\mathcal{A}^*$

## D. Summary

Phase-1 employs time-window-prioritized heuristics to construct a diverse feasible candidate pool; Phase-2 utilizes DRL-guided ALNS with adaptive operator scheduling and external archive management to expand the Pareto front. This separation of structural construction from adaptive search effectively combines domain knowledge and learning for robust multi-objective optimization, as summarized in Algorithm 1.

# IV. Experiment

## A. Experimental Overview

We evaluate MODRAL on multi-center routing with strong time-window constraints through three stages: (1) overall performance comparison, (2) Phase-1 ablation analysis, and (3) Phase-2 DRL benefits evaluation.

TABLE II: Compared configurations.

| Method | Phase-1 initialization | Phase-2 operator scheduling |
|---|---|---|
| MODRAL | heuristic | DRL-guided (PPO) |
| R-DRL | random | DRL-guided (PPO) |
| H-ALNS | heuristic | random scheduling |
| R-ALNS | random | random scheduling |

### *B. Experimental Setup*

*1) Benchmark and Implementation:* We conduct experiments on the Cordeau MDVRPTW benchmark [19], focusing on two representative instances: **pr01** (wider windows, $\tau = 266.52$) and **pr11** (tighter windows, $\tau = 172.12$). All experiments are implemented in Python with PyTorch; each stochastic run is assigned a unique `--run-id` with full logging for reproducibility.

*2) Evaluation Metrics:* We use Hypervolume (HV) and Inverted Generational Distance (IGD) on min–max normalized objectives:

$$\tilde{f}_j(x) = \frac{f_j(x) - f_j^{\min}}{f_j^{\max} - f_j^{\min}}, \quad j \in \{1, 2\}.$$

IGD measures the average distance from a reference set $V^*$ to the obtained set $V$, while HV measures the dominated volume relative to $r = (1, 1)$. $V^*$ is the non-dominated union of all solutions across 20 runs.

*3) Parameter Settings:*

*a) Phase-1.:* Phase-1 generates $N = 100$ diverse feasible initial solutions using a mixed set of constructive heuristics (global ordering, clustering/insertion, and randomized greedy diversification; 40/30/30 split). All feasible candidates are passed to Phase-2 as multi-start inputs.

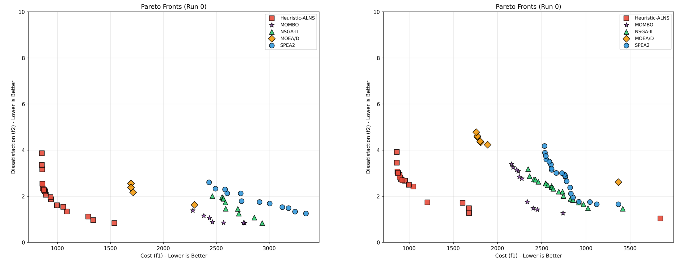
*b) Phase-2.:* Each candidate solution is optimized for 500 ALNS iterations. The simulated-annealing acceptance uses $T_0 = 100.0$ and cooling rate $\gamma = 0.995$ [20]. Six destroy and six repair operators yield 36 composite actions. When DRL is active, operator selection is guided by a policy network with mixing coefficient $\alpha = 0.4$.

*c) Working Time Constraint.:* The maximum working time per route is set to $H = 480$ minutes (8 hours) by default, reflecting a standard shift duration. This constraint is enforced as a hard constraint throughout the optimization process.

*4) Compared Methods:* Table II summarizes the compared configurations.

All variants share the same operator set and acceptance criterion. For fairness, each run generates $N = 100$ initial solutions and optimizes each start for $T_{\max} = 500$ ALNS iterations (total budget $N \times T_{\max}$). In the non-DRL baselines (H-ALNS and R-ALNS), destroy and repair operators are selected uniformly at random from the same $6 \times 6$ operator pool.

In addition to the ALNS-based variants, we compare against classic multi-objective evolutionary baselines (NSGA-II, SPEA2, MOEA/D, MOMBO). These baselines use the same solution evaluator as ALNS and are run with population size 100 and $n_{gen} = 500$.




(a) pr01 (b) pr11

Fig. 2: Pareto fronts comparison.

### *C. Results*

*1) Pareto Front Quality:* Figs. 2(a) and 2(b) compare the Pareto fronts obtained by the ALNS-based variants on pr01 and pr11. MODRAL achieves broader Pareto front coverage on both instances, with particularly pronounced advantages on the tightly constrained pr11.

*2) Comparison with MOEA Baselines:* To contextualize our method against classical multi-objective evolutionary solvers, we compare **MODRAL** with NSGA-II, SPEA2, MOEA/D, and MOMBO under a comparable evaluation budget. Tables III and IV report HV (higher is better) and IGD (lower is better) on pr01 and pr11. Wilcoxon signed-rank tests with Holm correction show that MODRAL significantly outperforms all baselines across all 16 pairwise comparisons (2 instances $\times$ 2 metrics $\times$ 4 baselines), with $p_{\text{Holm}} = 1.53 \times 10^{-5}$.

TABLE III: Hypervolume (HV) results on pr01 and pr11.

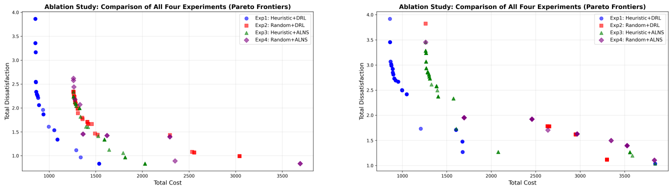
| Method | pr01 HV | pr11 HV | Avg. HV |
|---|---|---|---|
| H-ALNS | $0.9738 \pm 0.0031$ | $0.9276 \pm 0.0112$ | 0.9507 |
| MOEA/D | $0.2114 \pm 0.0333$ | $0.2902 \pm 0.0438$ | 0.2508 |
| SPEA2 | $0.0788 \pm 0.0317$ | $0.1357 \pm 0.0297$ | 0.1072 |
| MOMBO | $0.0610 \pm 0.0442$ | $0.1533 \pm 0.0503$ | 0.1072 |
| NSGA-II | $0.0811 \pm 0.0240$ | $0.1172 \pm 0.0300$ | 0.0992 |

TABLE IV: IGD results on pr01 and pr11.

| Method | pr01 IGD | pr11 IGD | Avg. IGD |
|---|---|---|---|
| H-ALNS | $0.0171 \pm 0.0047$ | $0.0435 \pm 0.0111$ | 0.0303 |
| MOEA/D | $0.5943 \pm 0.0416$ | $0.4929 \pm 0.0485$ | 0.5436 |
| MOMBO | $0.8193 \pm 0.0693$ | $0.6720 \pm 0.0704$ | 0.7456 |
| SPEA2 | $0.8284 \pm 0.0536$ | $0.7166 \pm 0.0475$ | 0.7725 |
| NSGA-II | $0.8288 \pm 0.0422$ | $0.7466 \pm 0.0480$ | 0.7877 |

*3) Ablation Study: Phase-1:* Figs. 3(a) and 3(b) show the ablation comparison of Pareto fronts in different ablation settings, isolating the effect of Phase-1 initialization. Heuristic initialization substantially improves solution quality. On pr01, it lowers the best cost from 1261.32 to 852.46; on pr11, it reduces the cost from 1261.32 to 860.06. The Pareto front size also increases from 54 to 95 solutions on pr11, indicating better diversity.

*4) Ablation Study: Phase-2:* A striking observation is that all three ablated variants (R-DRL, H-ALNS, R-ALNS) converge to the same minimum cost of 1261.32 on both instances,




(a) pr01 (b) pr11

Fig. 3: Ablation comparison of Pareto fronts on pr01 and pr11.

while MODRAL achieves significantly lower costs (852.46 on pr01, 860.06 on pr11). This suggests that the search space contains a prominent local optimum around 1261.32, which baseline methods struggle to escape.

The full MODRAL framework, combining heuristic initialization with DRL scheduling, is the only configuration that consistently breaks this barrier. On pr01, it nearly doubles the number of Pareto solutions (81 vs. 41 compared to H-ALNS) and extends the front toward higher satisfaction ($\max f_2$: 3.86 vs. 2.34). On pr11, it pushes the low-cost extreme from 1261.32 to 860.06 while maintaining solution diversity (95 vs. 93).

This demonstrates that neither heuristic initialization nor DRL scheduling alone is sufficient to overcome deep local optima; their synergy is essential for discovering superior regions of the solution space.

*5) Effect of Constraint Tightness:* DRL's advantage grows with constraint tightness. On pr01, where feasible regions are abundant, heuristic ALNS already performs well; DRL provides a moderate improvement. On pr11, the search space is heavily constrained, and poor operator choices easily lead to infeasibility or local traps. Here, DRL-guided selection significantly improves exploration of extreme Pareto solutions, especially low-cost ones.

*6) Discussion:* The experimental results lead to three key insights. First, heuristic initialization provides a diverse and high-quality starting pool, which proves especially valuable in tightly constrained instances. Second, DRL-based operator scheduling dynamically adapts to the search state, demonstrating clear advantages over static strategies. Most importantly, their synergy enables escape from deep local optima, yielding significantly better Pareto front coverage and convergence on strongly constrained problems.

## V. Conclusion

The proposed MODRAL framework effectively addresses strongly constrained multi-center home healthcare routing by combining structure-aware initialization with learning-guided improvement. The experimental results demonstrate three key insights: heuristic initialization provides a diverse, high-quality starting pool essential for tightly constrained instances; DRL-based operator scheduling dynamically adapts to the search state, outperforming static strategies; and critically, their synergy enables escape from deep local optima that neither component can overcome alone, yielding superior Pareto front coverage and convergence on strongly constrained problems.

## Acknowledgment

This research is supported, in part, by the National Natural Science Foundation of China under Grant 62006143; the Natural Science Foundation of Shandong Province under Grants ZR2025MS1012 and ZR2022MA091.

## References

[1] C. Fikar and P. Hirsch, "Home health care routing and scheduling: A review," *Computers & Operations Research*, vol. 77, pp. 86–95, 2017.

[2] K. Haripriya and V. K. Ganesan, "Solving large scale vehicle routing problems with hard time windows under travel time uncertainty," *IFAC-PapersOnLine*, vol. 55, no. 10, pp. 233–238, 2022.

[3] S. Røpke and D. Pisinger, "An adaptive large neighborhood search heuristic for the pickup and delivery problem with time windows," *Transportation Science*, vol. 40, no. 4, pp. 455–472, 2006.

[4] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002.

[5] Q. Zhang and H. Li, "MOEA/D: A multiobjective evolutionary algorithm based on decomposition," *IEEE Transactions on Evolutionary Computation*, vol. 11, no. 6, pp. 712–731, 2007.

[6] M. Nazari, A. Oroojlooy, L. V. Snyder, and M. Takáč, "Reinforcement learning for solving the vehicle routing problem," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[7] W. Kool, H. van Hoof, and M. Welling, "Attention, learn to solve routing problems!" in *International Conference on Learning Representations (ICLR)*, 2019.

[8] M. M. Solomon, "Algorithms for the vehicle routing and scheduling problems with time window constraints," *Operations Research*, vol. 35, no. 2, pp. 254–265, 1987.

[9] K. Braekers, K. Ramaekers, and I. Van Nieuwenhuyse, "Vehicle routing problems with time windows: State of the art and future directions," *Annals of Operations Research*, vol. 259, no. 1-2, pp. 89–158, 2016.

[10] J. A. Nasir and Y. H. Kuo, "Stochastic home care transportation with dynamically prioritized patients: an integrated facility location, fleet sizing, and routing approach," *Transportation Research Part B: Methodological*, vol. 184, p. 102949, 2024.

[11] T. Zhang, Y. Liu, X. Yang, J. Chen, and J. Huang, "Home health care routing and scheduling in densely populated communities considering complex human behaviours," *Computers & Industrial Engineering*, vol. 182, p. 109332, 2023.

[12] X. M. Ma, Y. P. Fu, K. Z. Gao, H. Zhang, and J. H. Mou, "A knowledge-based multi-objective evolutionary algorithm for solving home health care routing and scheduling problems with multiple centers," *Applied Soft Computing*, vol. 144, p. 110491, 2023.

[13] G. Bektur and D. Nenbhard, "A multi-objective approach to home health care routing problem with team formation," *International Journal of Industrial Engineering*, vol. 30, no. 5, pp. 1150–1168, 2023.

[14] N. Yadav and A. Tanksale, "An integrated routing and scheduling problem for home healthcare delivery with limited person-to-person contact," *European Journal of Operational Research*, vol. 303, no. 3, pp. 1100–1125, 2022.

[15] R. Qi, J. Q. Li, J. Wang, H. Jin, and Y. Y. Han, "Qmoea: A q-learning-based multiobjective evolutionary algorithm for solving time-dependent green vehicle routing problems with time windows," *Information Sciences*, vol. 608, pp. 178–201, 2022.

[16] E. Nikzad, M. Bashiri, and B. Abbasi, "A matheuristic algorithm for stochastic home health care planning," *European Journal of Operational Research*, vol. 288, no. 3, pp. 753–774, 2021.

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

[18] T. Xiang, Y. F. Li, and W. Y. Szeto, "The daily routing and scheduling problem of home health care: based on costs and participants' preference satisfaction," *International Transactions in Operational Research*, vol. 30, no. 1, pp. 39–69, 2023.

[19] J.-F. Cordeau, G. Laporte, and A. Mercier, "A unified tabu search heuristic for vehicle routing problems with time windows," *Journal of the Operational Research Society*, vol. 52, no. 8, pp. 928–936, 2001.

[20] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, pp. 671–680, 1983.