International Journal of Adaptive Control and Signal Processing

WILEY

RESEARCH ARTICLE

Efficient Pump Scheduling for Large-Scale Multiproduct Pipelines Using Deep Reinforcement Learning

Kai Shao | Xinmin Wang | Min Liu | Aobo Xu | Ling Jian

School of Economics and Management, China University of Petroleum, Qingdao, China

Correspondence: Ling Jian (bebetter@upc.edu.cn)

Received: 24 May 2024 | **Revised:** 6 August 2024 | **Accepted:** 28 August 2024

Funding: This work was supported by the National Key Research and Development Program of China under Grant Nos. 2021YFA1000104 and 2021YFA1000100, Laboratory Project of Higher Education Institutions in Shandong Province-Energy System Intelligent Management and Policy Simulation Laboratory at China University of Petroleum, and Youth Innovation Team of Higher Education Institutions in Shandong Province-Data Intelligence Innovation Team at China University of Petroleum.

Keywords: action space shaping | large-scale multiproduct pipelines | optimal pump scheduling | PPO | reinforcement learning

ABSTRACT

Optimizing pump scheduling in multiproduct pipelines can significantly reduce energy consumption and carbon emissions. For pump scheduling in multiproduct pipelines, to describe hydraulic losses more accurately, the model needs to adopt shorter discrete time intervals, which will lead to longer decision-making time. It combined with the large solution space of large-scale pipelines, will lead to low solution efficiency with dynamic programming methods and poor solution quality with heuristic optimization algorithms. Given that, this article develops a multiproduct refined oil transmission simulation system and employs the enhanced Proximal Policy Optimization (PPO) algorithm with action space shaping trick to optimize pump scheduling for large-scale multiproduct pipelines. The method of converting discrete action space to multi-discrete action space through action shaping can address PPO's low convergence efficiency issue resulting from the large discrete action space challenge common in large-scale multiproduct pipelines. The experimental results indicate that the proposed method, that is, PPO algorithm with multidiscrete action space exhibits significant advantages in terms of efficiency and robustness in large-scale pipelines compared to mainstream methods for pump scheduling such as dynamic programming (DP), genetic algorithms (GA), and ant colony optimization (ACO). Furthermore, we demonstrate the effectiveness of action space shaping in large-scale pipelines from the perspectives of exploration and exploitation.

1 | Introduction

Achieving carbon neutrality is a critical goal in promoting environmental sustainability. It has gained significant attention from researchers in many fields, such as the power industry, thermal industry, and oil and gas industry [1–3]. Refined products account for the majority of energy consumption worldwide [4]. At present, the multiproduct pipelines have been the major mode of transporting refined products from wharves and refineries to each city [5]. Multiproduct pipelines serve as the vital link between upstream refineries and downstream markets. In order to ensure the operation safety of pipelines and the supply of downstream markets, previous research has generally focused on scheduling optimization of multiproduct pipelines with multiple batches in pipeline operation and management [6]. However, multiproduct pipelines incur significant operational expenses, primarily attributed to substantial pumping costs [7]. Under the situation that the demand for refined oil and related

products will continue rising shortly [8], minor enhancements in pump scheduling can lead to substantial reductions in costs and carbon emissions.

The objective of optimal pump scheduling in multiproduct pipelines is deciding when to operate pumps, where the aim is to transport refined products while limiting electric costs and meeting safety constraints. Research on pump scheduling has predominantly focused on water distribution systems [9]. In contrast to a single-product pipeline, transporting diverse products in the same pipeline results in more complicated pressure characteristics [10]. Therefore, the optimal pump scheduling in multiproduct pipelines is a complex problem [11]. To accurately calculate hydraulic losses and represent pump operations within the scheduling horizon, it is imperative to introduce the time representation when establishing the scheduling model. Time expression of scheduling models can be classified into two forms: discrete-time form [12–14] and continuous-time form [15–17]. Using discrete-time expression can simplify the nonlinear coupling relations among each variable, reducing model establishment and solving complexity [18]. The approach partitions the scheduling horizon into discrete packs and intervals, equal [19] or specific lengths [20]. Liang, Li, and Li proposed a discrete-time hydraulic optimization model for multiproduct pipelines and employed dynamic programming (DP) to solve the model [5]. Abbasi and Garousi developed a discrete-time MILP model to optimize the pump operation schedule for oil pipelines with varying electricity tariffs [21]. Xin et al. built a multi-objective mixed-integer linear programming (MOMILP) model based on discrete-time representation for refined product pipelines to reduce the number of pump stops/restarts and the pump running cost simultaneously [22]. However, the discrete-time expression model requires many time slots to describe the model accurately, increasing the model's size [23]. Besides, the growing number of pump stations in large-scale pipelines also contributes to the increase in model size. These significant increases in optimization model scale can considerably reduce computational efficiency if conventional mathematical programming algorithms, such as dynamic and linear programming, are employed. Therefore, Liang et al. introduced a hybrid solution methodology by integrating dynamic programming and the ant colony algorithm, increasing computational efficiency [24]. Zhou et al. applied a Coarse-Grained Simulated Annealing-Genetic Algorithm (CGSAGA) to optimize the multiproduct pipeline's pump scheduling on a large-scale pump scheduling problem [25]. Although the above algorithms can solve large-scale problems to a certain extent, they cannot guarantee the algorithm's high efficiency and the schedule's optimality as the scale continues to increase.

Recent advances in reinforcement learning have made it possible to apply this technique to solve the large-scale pump scheduling problem. Reinforcement learning (RL) is a decision-making process where the agent learns to interact with the environment to maximize returns [26]. RL has been successfully applied to various fields like video games, control, operations research, and so forth [27–29]. Given the issues of pump scheduling problems in different fields, the application of RL is gradually becoming more widespread. In the field of pump scheduling in water distribution systems, Hajgató, Paál, and Gyires-Tóth applied deep reinforcement learning (DRL) to the single-step pump scheduling problem in water distribution systems, utilizing the outputs of the Nelder–Mead method as the reward standard [30]. Hu et al. proposed an exploration-enhanced DRL framework to address real-time pump scheduling problems in water distribution systems [31]. In the domain of residential demand response, Mbuwir et al. proposed the use of fitted Q-iteration (FQI) and policy iteration (PI) with Q-functions to control the battery and heat pump simultaneously in a residential microgrid [32]. Peirelinck, Ruelens, and Decnoninck applied and evaluated a model-free RL agent to a heat pump and residential building model implemented in Modelica [33]. Patyn, Ruelens, and Deconinck discussed the implementation of a heat pump agent in a demand response setting and its cost-effectiveness when implemented with different neural network types [34].

Regarding the pump scheduling problem in multiproduct pipelines, selecting suitable RL algorithms to obtain an ideal pump scheme is theoretically feasible and potent. Reinforcement learning encompasses two primary methods: value-based methods and policy-based methods. The actor-critic algorithm combines aspects of both methods, designed to address each method's limitations when used individually. The Proximal Policy Optimization (PPO) [35] algorithm is a model-free, on-policy reinforcement learning method that employs an Actor-Critic architecture. It improves upon previous methods using a clipped objective function, which helps balance exploration and exploitation, thus ensuring more stable and consistent learning. Considering the high-dimensional, continuous value of input parameters in multiproduct pipelines and the need for algorithm robustness, the PPO method is chosen. It offers improved scalability, data efficiency, and robustness compared to the policy gradient method. In addition, the issue of dealing with huge discrete action space poses a significant and common challenge in RL [36], which we will also encounter in the pump scheduling problem. Therefore, this article establishes an integer linear programming model based on the discrete-time representation and applies the PPO algorithm with the trick of action space shaping [37] to address this problem. The main contributions of this article are threefold and listed as follows.

C1 A multiproduct refined oil transmission simulation system was conducted, providing an interactive environment for reinforcement learning.

C2 Aiming to address decision complexity arising from the need for precise hydraulic loss estimation with shorter discrete time intervals, the paper transforms pump scheduling in multiproduct pipelines as a Markov Decision Process (MDP) and applies the PPO algorithm for its solution.

C3 An enhanced PPO algorithm with an action space shaping trick was proposed to address the low convergence efficiency resulting from the large discrete action space challenge common in large-scale multiproduct pipelines.

The rest of the article is organized as follows. Section 2 discusses the problem statement and outlines the model prerequisites. Section 3 introduces the proposed methodology with its mathematical derivation and elaborates on the specifics of the integer linear programming model. Section 4 gives an introduction regarding the designs of important factors for applying PPO

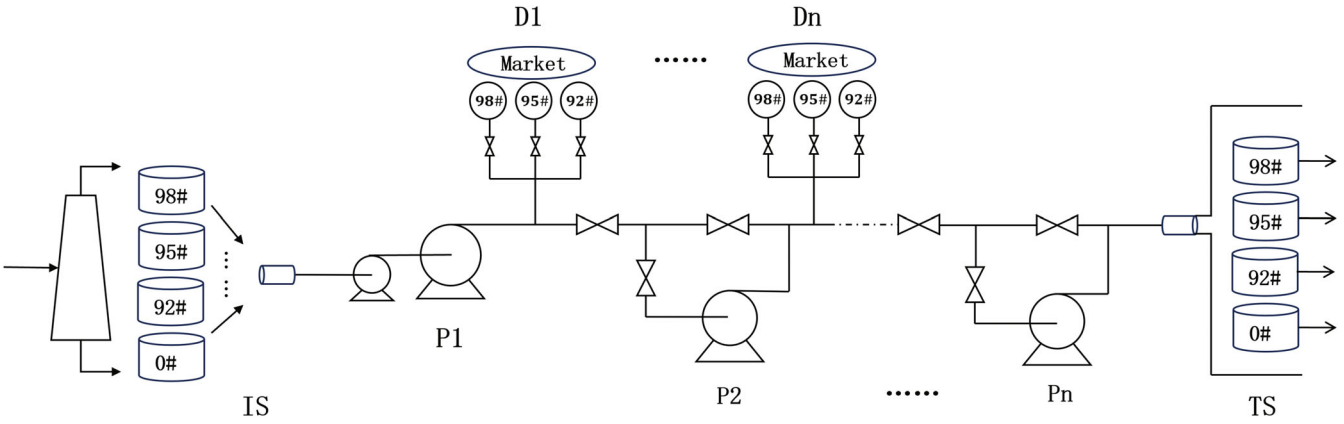

FIGURE 1 | Refined oil transmission pipeline system with multiple pump stations.

with action space shaping in pump scheduling of multiproduct pipelines. Section 5 presents case studies with the proposed algorithm validation and analyses the effects of action space shaping. Finally, the conclusion is drawn in Section 6.

2 | Problem Description

The multiproduct pipeline system transports various refined products from the refinery to downstream markets. Multiple pump stations are strategically placed along the pipeline to provide the necessary energy and ensure the smooth operation of the refined oil within the pipeline route. As shown in Figure 1, the pipeline system is equipped with pump stations, such as P1 to Pn, which boost the products along the pipeline. Delivery stations, including D1 to Dn and a terminal station (TS), are also connected to the local markets. These delivery stations serve as crucial points for the final delivery of the refined products to the respective cities.

Accurately calculating the hydraulic losses in multiproduct pipelines requires a comprehensive analysis of several key parameters. These parameters encompass the pipeline's flow rate, the precise positioning of each batch interface, and constant factors such as the topographical profile of the pipeline, the length and diameter of individual pipeline segments, and the physical properties specific to each refined product being transported. The flow rate fluctuations and the arrangement and distribution of refined products within the pipeline are dynamic and time-dependent, depending on a specific detailed pipeline schedule. With the detailed schedule in place, our focus is pump scheduling to ensure efficient and reliable multiproduct transportation.

To mitigate the complexity arising from time-dependent variables, this study employs mathematical models that utilize a discrete-time representation. The scheduling horizon is divided into distinct time windows of equal or specific lengths, delineated by time nodes. All pump operations are constrained to occur exclusively at these discrete time nodes. To ensure an accurate representation of the model, the discrete presentation mode necessitates a significant number of time slots, leading to an expansion in the model size. While the larger model size gives rise to computational challenges, the proposed method employs the reinforcement learning algorithm which offers a practical solution and overcomes these resource constraints. Unlike traditional optimization algorithms that may struggle with increased model complexity, the proposed approach can efficiently handle such scenarios.

This article applies a reinforcement learning algorithm to solve the optimal pump scheduling in multiproduct pipelines. Once the model is solved, pump scheduling in the single-source multiproduct pipelines can be obtained. The details of pump scheduling in multiproduct pipelines are listed below.

Given:

- production data: physical properties of products including density and viscosity
- station and pipeline data: stations' positions, length, and diameter of each pipeline segment, limits of pressure, and the initial state of batches in the pipeline
- pump data: pump performance parameters such as pumping head, pump efficiency, and pump flow rate limits
- cost data: pumping cost, unit cost caused by pump start/stop
- detailed pipeline schedule: inject flow rate, inject time window, delivery flow rate, and delivery time window

Determine:

- pump schedule: the precise time of each pump's start and stop operations
- energy consumption: total operation costs of the pipeline (i.e., pump running energy and start/stop costs).

Given a detailed pipeline scheduling plan, the discrete-time model aims to develop a pump scheduling with minimum pumping costs, including operational and switching costs. To effectively solve this model, the following reasonable simplifications and assumptions are made:

- The oil products are incompressible and the physical properties of the fluid remain constant, and any variations due to temperature and pressure can be disregarded.

- Within a given time window, the injection flow rate at the initial station and the delivery flow rate at the distribution terminals remain constant.
- The pump start/stop operations are assumed to be instantaneous and do not require any additional time.
- During the sequential movement of batches in a multiproduct pipeline, it is inevitable that contamination will occur between adjacent oils. This results in the formation of a mixed oil section, which is considered as an interface between the two batches.
- The given detailed pipeline schedule satisfies a series of constraints, such as inventory constraints. These constraints are taken into account by the pipeline operator when determining the overall objectives of the short-term plan.

3 | Mathematical Model

In this section, a pump scheduling model of multiproduct pipelines based on discrete-time expressions is established. The model aims to minimize the pumping costs, including pump running energy and start/stop costs. We list the symbols used in the paper first.

3.1 | Objective Function

The objective of this model is to minimize the pumping costs as Equation (1). The pumping costs involve costs of the pump running costs (f_1) and pump start/stop (f_2) costs.

$$\min f = f_1 + f_2 \tag{1}$$

Pump running costs (f_1) are determined by running time, electricity price, pump characteristic, pump efficiency, oil density, and flow rate through the station.

$$f_1 = \sum_t \sum_i \sum_k \frac{B_{t,i,k} \cdot F_i \cdot \rho_{t,i} \cdot g \cdot \left(a_{i,k} - b_{i,k} \cdot Q_{t,i}^{2-m}\right) \cdot Q_{t,i} \cdot \tau_t}{1000 \cdot \eta_{t,i,k}} \quad t \in T, i \in IP, k \in K_i \tag{2}$$

Pump start/stop costs (f_2) are calculated based on the unit cost of each start/stop event and the total number of start/stop events that occur during the pipeline operation.

$$f_2 = \sum_t \sum_i \sum_k E_{i,k} S_{t,i,k} \quad t \in T, i \in IP, k \in K_i \tag{3}$$

3.2 | Pressure Constraints

3.2.1 | The Energy Balance Constraints Along the Pipeline

The pressure generated by a pump station can be described using the following expression.

$$P_{t,i}^H = \sum_k B_{t,i,k} \cdot \rho_{t,i} \cdot g \cdot \left(a_{i,k} - b_{i,k} \cdot Q_{t,i}^{2-m}\right) \quad t \in T, i \in IP, j \in J, k \in K_i \tag{4}$$

The pipeline is divided into I segments based on specific locations, such as pump stations, delivery stations, elevation points, and points with varying diameters. Equation (5) represents the hydraulic loss of the pipe section $(i, i+1)$ at time t, which includes both the gravitational potential loss and the frictional loss. The calculation of friction loss in multiproduct pipelines in China commonly utilizes the Liebenson equation, a widely accepted method for estimating frictional loss.

$$P_{t,i}^F = \sum_j \left(\frac{\lambda \rho_j g v_j^m Q_{t,i}^{2-m} L_{t,i,j}}{d_i^{5-m}} + \rho_j g z_{t,i,j} \right) \quad t \in T, i \in I, j \in J \tag{5}$$

The pressure along the entire pipeline must adhere to the energy balance constraints, as expressed by Equations (6) and (7). If point i is not a pump station, such as a delivery station or an elevation point, its provided pressure $P_{t,i}^H$ is zero.

$$P_{t,i}^{OUT} = P_{t,i}^{IN} + P_{t,i}^H \quad t \in T, i \in I \tag{6}$$

$$P_{t,i+1}^{IN} = P_{t,i}^{OUT} - P_{t,i}^F \quad t \in T, i \in I \tag{7}$$

3.2.2 | Station Inlet and Outlet Pressures Constraints

Each station's inlet and outlet pressures must comply with upper and lower limits, as shown by Equations (8) and (9). These limits are determined by considering downstream equipment specifications, and the allowable suction pressure of the pumps.

$$P_i^{IN,MIN} \le P_{t,i}^{IN} \le P_i^{IN,MAX} \quad t \in T, i \in IP \tag{8}$$

$$P_i^{OUT,MIN} \le P_{t,i}^{OUT} \le P_i^{OUT,MAX} \quad t \in T, i \in IP \tag{9}$$

3.2.3 | Pipeline Pressure Constraints

In the pipeline system, the high points are subject to a lower pressure constraint to prevent the pressure from dropping below the saturation vapor pressure of the transported fluid, which could lead to vaporization. Conversely, the low points in the pipeline must not exceed the maximum pressure capacity to avoid the risk of pipeline rupture caused by excessive pressure. Equations (10) and (11) show the pressure limit of each pipeline's pressure.

$$P_{t,i} \ge P_i^{HIGH,MIN} \quad t \in T, i \in IH \tag{10}$$

$$P_{t,i} \le P_i^{LOW,MAX} \quad t \in T, i \in IL \tag{11}$$

3.3 | Pump Characteristic Constraints

For constant-speed pumps, the characteristics of the pump unit can be determined using the least squares regression method. The relationships between the pumping head, efficiency, and flow rate can be approximated as Equations (12) and (13). In the case of a pump station consisting of multiple pumps connected in series, the total head provided by the pump station is equal to the sum of the heads of each pump.

$$H_{t,i,k} = a_{i,k} - b_{i,k} Q_{t,i}^{2-m} \quad t \in T, i \in IP, k \in K_i \tag{12}$$

$$\eta_{t,i,k} = c_{i,k} Q_{t,i}^2 + d_{i,k} Q_{t,i} + e_{i,k} \quad t \in T, i \in IP, k \in K_i \tag{13}$$

3.4 | Batch Transportation Constraints

Obtaining the coordinate of each batch is known during the research horizon, making it possible to determine the injected or delivered products during each time. Based on the principle of volume conservation, the volume injected at the initial station should be equivalent to the cumulative volume of deliveries in discrete time intervals expressed by Equation (14).

$$V_t^I = \sum_i V_{t,i}^D \quad t \in T, i \in ID \tag{14}$$

4 | Method

4.1 | Reinforcement Learning Factors

The reinforcement learning (RL) algorithm addresses sequential decision-making challenges and can be mathematically expressed as a MDP. The learning process in RL involves continuous interactions between the agent and its environment, allowing the agent to adapt and make decisions based on its observations. This section provides a detailed discussion of RL's primary factors, state space, action space, and reward function, and their application to achieve optimal pump scheduling in multiproduct pipelines.

4.1.1 | State Space

In reinforcement learning, the state space is the set of all possible states the agent may encounter. It is essential to balance providing enough practical information and avoiding excess irrelevant information that benefits the agent to acquire effective policy through learning. For this work, the pump scheduling information in the optimization of the multiproduct pipeline includes the pump schedules at time $t-1$, the oil product physical properties (Flow rate, density, and viscosity) for the oil passing through each pump station at time t and the volume coordinates of each oil batch interface at time t.

4.1.2 | Action Space

Action can be represented as the state of each pump. For small-scale pipelines, we use discrete action space, where each action a_t is $\left[B_{t,1,1}, B_{t,1,2}, \ldots, B_{t,i,k}, \ldots\right]$, with $B_{t,i,k}$ as a binary variable indicating whether the pump is turned on or off. As the number of pumps increases, the action space grows exponentially. Therefore, we adopt multi-discrete action spaces through the trick of action space shaping for large-scale pipelines to make learning easier. The action a_t is $\left[a_t^1, a_t^2, \ldots, a_t^i, \ldots\right]$ for large-scale pipelines, where $a_t^i = \left[B_{t,i,1}, B_{t,i,2}, \ldots, B_{t,i,k}, \ldots\right]$ represents the operation state of the ith individual pump station. The trained agent selects the optimal action from the action space during each time interval of the scheduling period based on the learned policy.

4.1.3 | Reward Function

The reward function is designed to incentivize the agent to accomplish its objective, with the reward value reflecting the efficacy of the chosen action. In this study, the reward function encompasses three critical components: the energy consumption cost of the pumps, the cost associated with the operation of the pump switches, and the penalty associated with pressure constraints.

- *Energy cost of pump*: The optimization objective of pump scheduling is to minimize energy consumption while satisfying certain constraints. The critical aspect of energy cost is that the lower the energy cost of the pumps, the higher the designed reward.
- *Pump switches cost*: In cases where the pump's operational state changes between time steps $t-1$ and t, corresponding cost penalties should be applied. Considering the hydraulic constraints, pump start/stop operations should be minimized as much as possible. Therefore, more frequent pump switching results in a lower reward in the reward design for pump start/stop operations.
- *Penalty of pressure constraints*: When the energy provided by the pump schedule determined at t cannot meet the pressure constraints, a penalty based on comparing the pressure and upper or lower limits is added to the reward function. After being penalized, the agent learns from this negative feedback and adjusts its policy. This adaptation in the policy is driven by the recognition that the penalized action is unfavorable and should be minimized in future decision-making.

Therefore, with the fundamental demands fulfilled, the reward function can be designed as the following equation.

$$r_t = -I + (1 - I)\text{sigmoid}(-(f_1 + f_2)) \tag{15}$$

where I is an indicator function that monitors whether an agent's action meets the pressure constraint. If the action complies with the constraint, I is set to 0. If the action does not satisfy the constraint, I is set to 1. Furthermore, the elements f_1 and f_2 are derived from Equation (1).

4.2 | Reinforcement Learning Model

In this section, we applied the PPO algorithm to obtain the pump schedule during the scheduling horizon. Figure 2 illustrates the interaction between the PPO algorithm with action space shaping trick and the environment of the multiproduct pipeline. For the multiproduct pipeline pump scheduling problem, which varies in scale, discrete action space is employed for small-scale pipelines, where a single actor is used to output the action distribution. On the other hand, multi-discrete action space is utilized for large-scale pipelines, requiring multiple actors corresponding to the number of pump stations to output the action distribution. The Policy Gradient (PG) algorithm involves computing the gradient of the actor and using stochastic gradient ascent to update it toward the optimal value iteratively. The most commonly used gradient estimator has the form:

$$L^{PG}(\theta^i) = \widehat{E}_t\left[\log \pi_{\theta^i}\left(a_t^i \mid s_t\right)\widehat{A}_t\right] \tag{16}$$

where π_{θ^i} denotes a stochastic policy implemented by actor i and a_t^i represents the action of the ith individual pump station.

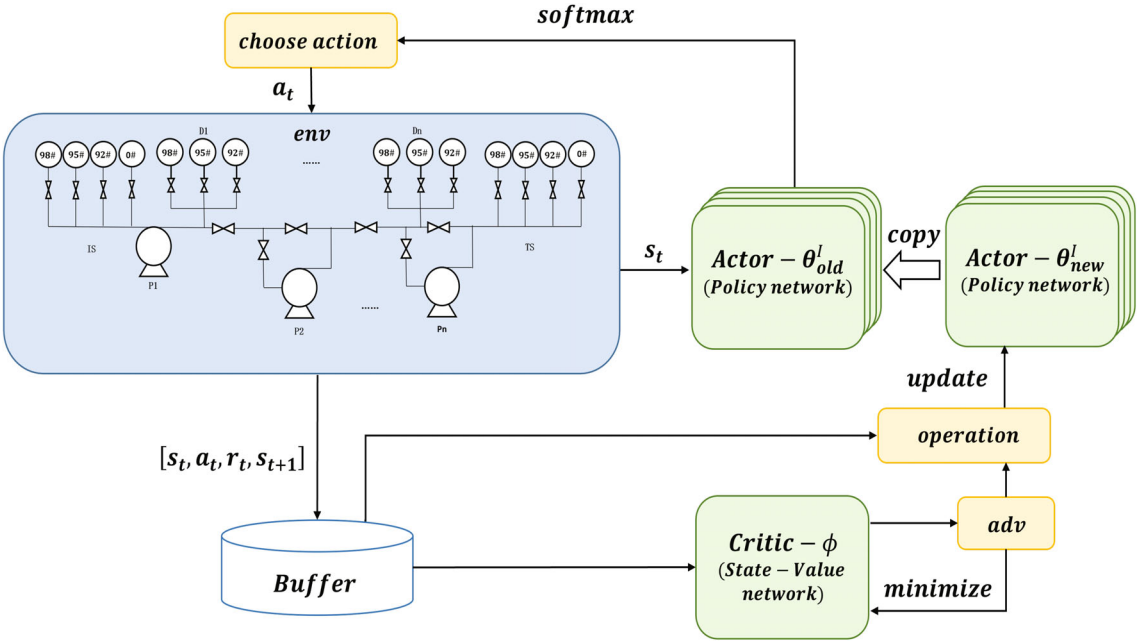

FIGURE 2 | PPO algorithm with action space shaping interacts with the refined oil transmission pipeline system.

In particular, concerning the discrete action space in small-scale pipelines, action a_t^i is turned into action a_t mentioned in Section 4.1.2. $\widehat{A}_t$ is an estimator of the advantage function at t time and $\widehat{E}_t$ represents the empirical mean for a limited number of samples. The PPO algorithm employs the probability ratio $r_t(\theta^i)$ to maximize the use of sampled data and enable significant improvements to the policy function implemented by actor i. The Equation (17) provides the expression for the definition of $r_t(\theta^i)$.

$$r_t(\theta^i) = \frac{\pi_{\theta^i}(a_t^i \mid s_t)}{\pi_{\theta^i{}_{old}}(a_t^i \mid s_t)} \tag{17}$$

In order to constrain the gradient and prevent too large policy updates, the PPO algorithm employs a clipping technique, which is reflected in its objective function, as shown in Equations (18) and (19).

$$L^{CLIP}(\theta^i) = \widehat{E}_t\left[\min\left(r_t(\theta^i)\hat{A}_t, \mathrm{clip}\left(r_t(\theta^i), 1-\varepsilon, 1+\varepsilon\right)\hat{A}_t\right)\right] \tag{18}$$

$$\mathrm{clip}\left(r_t(\theta^i), 1-\varepsilon, 1+\varepsilon\right) = \begin{cases} 1+\varepsilon, & \text{if } r_t(\theta^i) \geq 1+\varepsilon \\ r_t(\theta^i), & \text{if } 1-\varepsilon < r_t(\theta^i) < 1+\varepsilon \\ 1-\varepsilon, & \text{if } r_t(\theta^i) \leq 1-\varepsilon \end{cases} \tag{19}$$

where ε is a parameter to ensure the similarity of policy π_{θ^i} and policy $\pi_{\theta^i{}_{old}}$. When the new policy $\pi_{\theta^i}(a_t^i \mid s_t)$ is significantly different from the old policy $\pi_{\theta^i{}_{old}}(a_t^i \mid s_t)$, the PPO algorithm employs a clipping mechanism to prevent the probability ratio $r_t(\theta^i)$ from exceeding the range of $[1-\varepsilon, 1+\varepsilon]$. The objective is to select the minimum between the clipped and the unclipped objectives, thus establishing a lower bound (i.e., a pessimistic bound) on the unclipped objective. This method ignores changes in the probability ratio when they would improve the objective and includes them when they would worsen it. Applying PPO in pump scheduling of various scale multiproduct pipelines is expressed as Algorithm 1.

5 | Result and Discussion

In this section, a real-world multiproduct pipeline was selected as the test case to demonstrate the applicability of the proposed method. Simulations were carried out on a computer with Intel core i5-12450H CPU @2.00GHZ. The Reinforcement learning models were written in Python version 3.9, and the environment of the pipeline system was written in C#. RL models can interact with the environment by dynamically-link library.

5.1 | Pipeline Basic Data

Taking a Chinese multiproduct pipeline as an example, the pipeline has a total length of 855.2 km and is equipped with eight stations, including one initial station (IS), seven download stations (DS1 to DS7), and five pump stations (P1 to P5). Specially, stations IS, DS1, DS4, DS5, and DS6 also serve as pump station P1, P2, P3, P4, and P5, respectively. This large-scale pipeline transports three kinds of oil products (95# gasoline, 92# gasoline, and 0# diesel). The fundamental pipeline data are listed in Table 1, while Table 2 displays each pump station's pump configurations and each pump's corresponding parameters. The pressure constraints for each station are enumerated in Table 3. Table 4 details each refined product's physical characteristics. The initial oil in the pipeline is filled with gasoline 92# gasoline. Based on practical operational practices at the site, the unit cost for starting/stopping the pump operation is set as 4000 CNY, and the electricity charge at each station is determined to be 0.8 CNY/(kW · h).

5.2 | Case Studies

The operation modes of the pipeline can be categorized into two distinct scenarios: (1) The pipeline experiences a reduction in

ALGORITHM 1 | PPO based pump scheduling algorithm.

Initialize: the policy parameters $\theta_0^I = \left[\theta_0^1, \theta_0^2, \cdots, \theta_0^I\right]$, the value function parameters ϕ_0, the pump stations number M.
  if The pipeline is large-scale **then**
    $I = M$
  else
    $I = 1$
  end if
  for $i = 0, 1, \ldots, I$ **do**
    $\theta_{old}^i = \theta_0^i$
  end for
  for $k = 0, 1, \ldots,$ **do**
    Save N trajectories $\left(s_t, a_t, r_t, s_{t+1}\right)$ by running policy $\pi_{\theta_{old}^i}$ in the environment of pipeline system
    Compute advantage estimation $\hat{A}_t$ based on the current value function V_{ϕ_k}
    for $i = 0, 1, \ldots, I$ **do**
      Update the policy by a gradient method:
$$\theta_{k+1}^i = \text{argmax}\frac{1}{NT}\sum^N\sum_{t=0}^T \min\left(r_t(\theta^i)\hat{A}_t,\ \text{clip}\left(r_t(\theta^i), 1-\varepsilon, 1+\varepsilon\right)\hat{A}_t\right)$$
    end for
    Update the value function parameters by regression on mean-squared error:
$$\phi_{k+1} = \text{argmin}\frac{1}{NT}\sum^N\sum_{t=0}^T\left(V_{\phi_k}(s_t) - R_t\right)^2$$
    $\theta_{old}^i = \theta_{k+1}^i$
  end for

TABLE 1 | Basic data of pipeline.

Pipeline segment	Length (km)	Inner diameter (mm)	Terminal elevation (m)
IS(P1)-DS1	130.4	492.2	7
DS1(P2)-DS2	139.5	492.2	111
DS2-DS3	66.9	441.2	53
DS3-DS4	38.7	441.2	51
DS4(P3)-DS5	166.6	392.2	105
DS5(P4)-DS6	131.1	392.2	43
DS6(P5)-DS7	91.8	392.2	50
DS7-TS	90.2	392.2	28

scale as the segments spanning from DS4 to TS are temporarily shut down. In this mode, the pipeline retains a configuration with one injection station, four delivery stations, and two pump stations. (2) The pipeline functions normally, with each delivery station requiring various refined products, making it a conventional large-scale multiproduct pipeline. The pipeline's operational modes correspond to small-scale and large-scale multiproduct pipeline pump scheduling cases. The two cases were used to verify the practicality and effectiveness of the PPO algorithm in addressing the pump scheduling problem for pipelines of different sizes and complexities. To illustrate the advantage of the PPO algorithm, the algorithm is compared against DP [5] (as an exact algorithm), genetic algorithm [25], and ant colony algorithm [24].

Furthermore, we conducted additional experiments to demonstrate that the PPO algorithm performs better with multi-discrete action spaces than with discrete action spaces when addressing the large-scale pipeline pump scheduling problem. The detailed settings of PPO in the experiments are listed in Table 5.

5.2.1 | Performance Comparison

To validate the performance and superiority of the PPO algorithm compared with other algorithms in solving pump scheduling problems of multiproduct pipelines, we conducted the detailed pipeline schedule for two operation cases for testing purposes. All test cases were performed with a discrete time interval of 1 h. Based on the detailed pipeline schedule presented in Figure 3 for Case 1, the fluctuations in segment flow rate and hydraulic loss can be determined. The total scheduling horizon for the pipeline spans 250 h. At the start of each hour, the agent receives the pipeline's current state and performs the corresponding action online until the end of the total scheduling horizon. Figure 4 shows the experimental results by interacting with the environment. The minimum energy cost in Figure 4 reached approximately 290 thousand CNY. Table 6 compares the optimization results of PPO with DP, GA, and ACO. This table displays the energy and time costs of pump scheduling for Case 1 under each algorithm. According to Table 6, PPO exhibits higher accuracy than the other two heuristics, almost matching the precision of DP. Moreover, PPO's model training time holds an advantage over the model solving time of the other two heuristic algorithms. Figure 5 illustrates the pump schedules generated by the PPO algorithm and other comparison methods.

To further demonstrate the effectiveness of the PPO algorithm on large-scale pipelines, we selected a detailed pipeline schedule for Case 2 and performed the pump scheduling based on this plan. Figure 6 shows the detailed pipeline schedule with a scheduling horizon of 360 h. In addressing the exponentially growing search space in large-scale pipeline pump scheduling, the PPO algorithm converts discrete action space to multi-discrete action space. The PPO algorithm's effectiveness is demonstrated in Figure 7, where it achieved a minimal energy cost of nearly 1.26 million CNY. Similarly, Table 7 describes the optimization results of each algorithm. While the DP algorithm offers the optimal

TABLE 2 | Basic data of pump station.

Pump station	Pump	Pump head		Pump efficiency		
		a	$b(\times 10^{-4})$	$c(\times 10^{-7})$	$d(\times 10^{-3})$	$e(\times 10^{-2})$
P1	P1-1	438.98	5.07	−6.60	1.40	8.36
	P1-2	439.47	5.09	−6.18	1.34	9.20
	P1-3	213.59	2.13	−6.67	1.41	9.29
P2	P2-1	439.87	4.48	−7.44	1.51	5.46
	P2-2	445.10	4.44	−7.17	1.47	5.64
	P2-3	211.89	1.93	−7.82	1.52	8.96
P3	P3-1	466.78	5.34	−11.30	1.72	16.10
	P3-2	466.77	5.13	−13.50	1.91	12.70
	P3-3	243.45	2.56	−15.10	2.06	11.40
P4	P4-1	262.57	2.92	−25.20	2.81	4.72
	P4-2	262.26	2.91	−25.00	2.79	5.05
	P4-3	143.86	3.45	−28.30	3.01	5.84
P5	P5-1	259.51	3.57	−30.82	3.07	4.79
	P5-2	259.47	3.64	−30.80	3.08	4.45
	P5-3	137.93	3.38	−35.10	3.37	4.90

TABLE 3 | Pressure limits of each stations.

Station	Inlet lower/upper limits (MPa)	Outlet upper limits (MPa)
IS(P1)	/	7.9
DS1(P2)	0.1 or 0.5 (if any pump is on)/7.9	7.9
DS2	0.6 (high-elevation point)/7.9	7.9
DS3	0.1/7.9	7.9
DS4(P3)	0.1 or 0.5 (if any pump is on)/7.9 or 2.0 (if it is the terminal station)	7.9
DS5(P4)	0.6 (high-elevation point)/7.9	7.9
DS6(P5)	0.1 or 0.5 (if any pump is on)/7.9	7.9
DS7	0.1/7.9	7.9
TS	0.1/5	7.9

TABLE 4 | Physical characteristics of refined products.

Product	Density $(\mathrm{kg/m^3})$	Viscosity $(\times 10^{-6}\ \mathrm{m^2/s})$
0# diesel	845	6.11
92# gasoline	740	2.05
95# gasoline	750	1.07

solution, its extensive time consumption produces low algorithmic efficiency. Conversely, the GA algorithm is time-efficient but yields high pump switching costs, making it impractical for operational use. On the other hand, the ACO algorithm hardly provides feasible solutions due to inefficient computational performance in large-scale pipeline pump scheduling. Overall, the PPO algorithm demonstrates energy and time cost advantages, making it well-suited for addressing large-scale pipeline pump scheduling problems. Figure 8 shows the pump schedules generated by the PPO and DP algorithms.

TABLE 5 | Detailed settings of PPO.

Symbol	Hyperparameter	Value
N_{layer}	Structure and number of neurons in layers of the actor network	[256, 256]
N_{layer}	Structure and number of neurons in layers of the critic network	[256, 256]
α_A	Learning rate of the actor network	3×10^{-4}
α_C	Learning rate of the critic network	3×10^{-4}
γ	Discounter factor	0.99
K	Clip range	0.2

5.2.2 | Effects of Multi-Discrete Action Space

In this study, we conducted a comparative analysis and validation of discrete action space versus multi-discrete action space in large-scale pipeline pump scheduling. We demonstrated the superiority of multi-discrete actions over discrete actions by focusing on two fundamental perspectives: exploration and exploitation. Concerning exploration, Figure 9 illustrates the probability of choosing various actions within these two types of action space under different iteration steps. As depicted in Figure 9a, the exploration within discrete action space tends to be uniformly distributed across the entire action space during

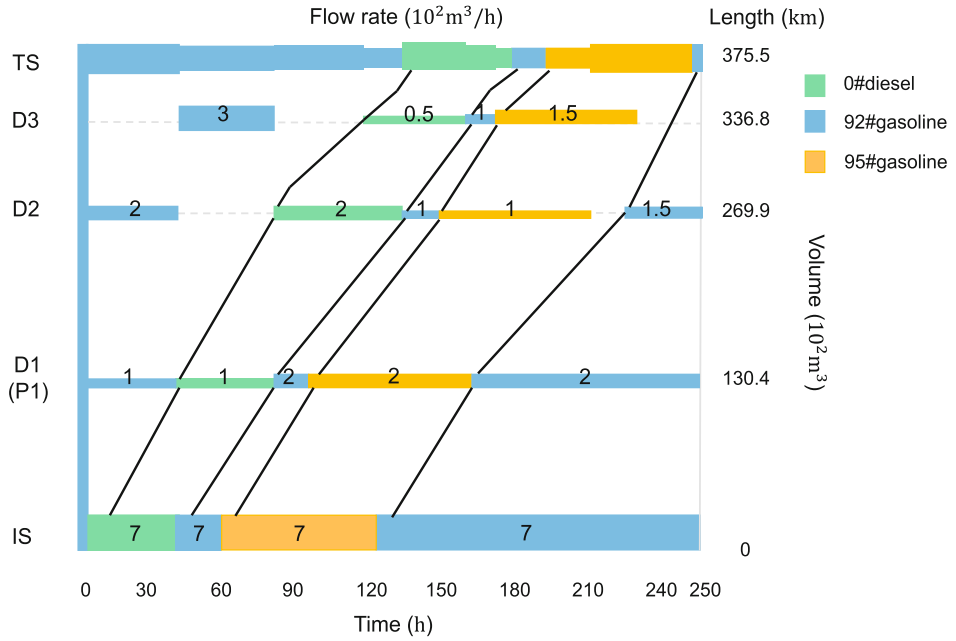

FIGURE 3 | Batch transportation diagram for Case 1.

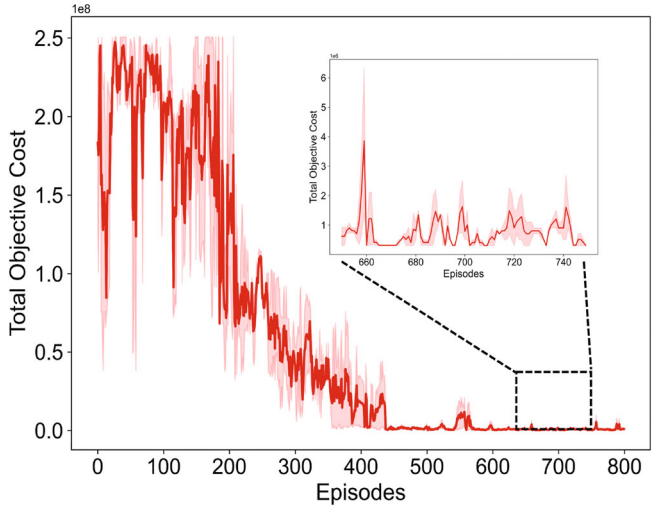

FIGURE 4 | Episode total objective cost during training in Case 1.

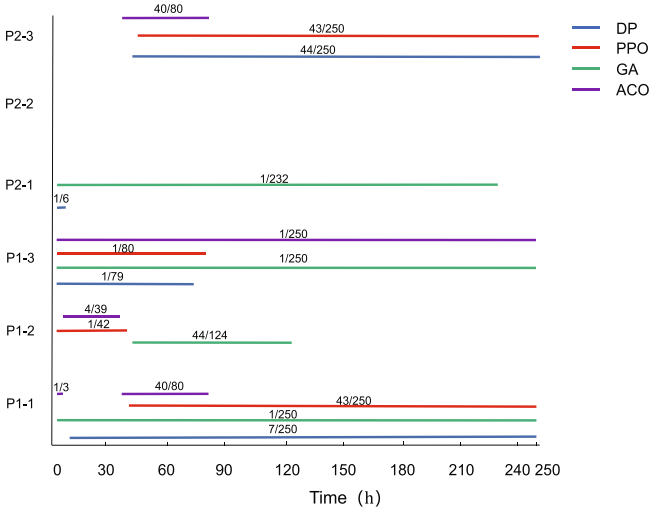

FIGURE 5 | Pump schedules generated by all algorithms for Case 1.

TABLE 6 | Optimization results of PPO, DP, GA, and ACO for Case 1.

Model	Pump switch costs (CNY)	Pump running costs (CNY)	Model training time (s)	Model solving time (s)
PPO	24,000	266,757	4103.40	0.41
DP	24,000	264,869	/	171.88
GA	24,000	267,432	/	3120.32
ACO	36,000	275,468	/	10,800.67

pre-training. In contrast, exploration within multi-discrete action space exhibits a distinct periodicity. The periodicity may stem from the training process, where the current policy exhibits a preference for a specific pumping scheme at one pump station. Consequently, other pump stations explore additional actions based on the unchanged scheme, potentially resulting in the observed periodic pattern. This approach is more conducive to exploring large discrete action space. Furthermore, Figure 9b reveals that in the mid-training phase, probability functions for exploring actions in discrete action space are unimodal, while in multi-discrete action space, they are multimodal, indicating a higher efficiency in exploration.

Regarding exploitation, Figure 10 illustrates that multi-discrete action space allow the agent to find better pump schemes in each episode that result in higher average reward than discrete action space. Conversely, agent often selects suboptimal or infeasible pump schemes in discrete action space.

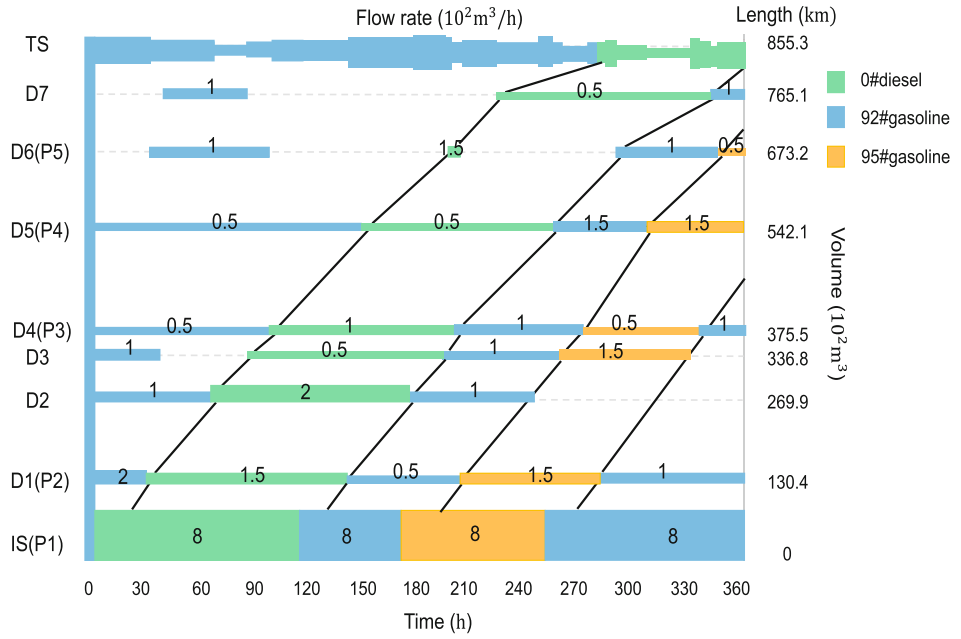

FIGURE 6 | Batch transportation diagram for Case 2.

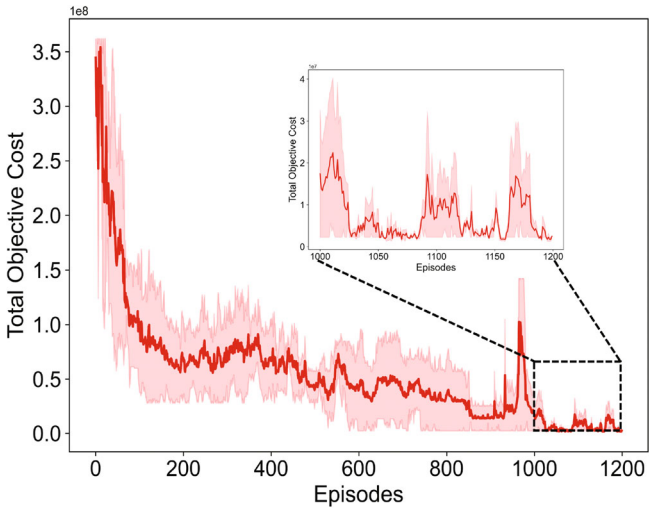

FIGURE 7 | Episode total objective cost during training in Case 2.

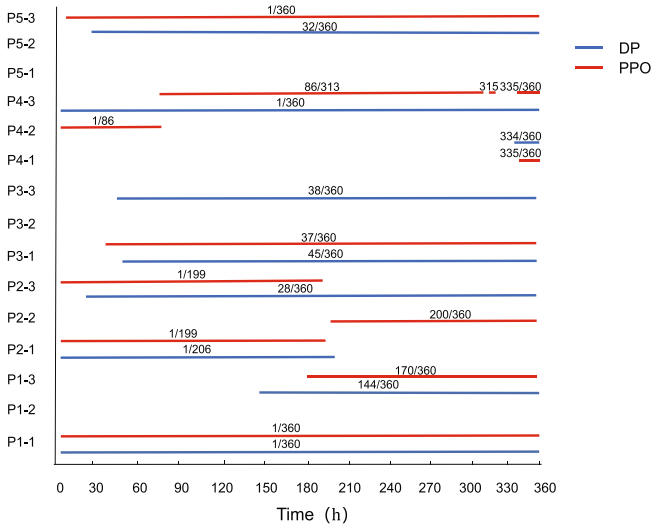

FIGURE 8 | Pump schedules generated by the PPO and DP algorithms for Case 2.

TABLE 7 | Optimization results of PPO, DP, GA, and ACO for Case 2.

Model	Pump switch costs (CNY)	Pump running costs (CNY)	Model training time (s)	Model solving time (s)
PPO	68,000	1,200,811	14,010.57	0.42
DP	40,000	1,131,651	/	136,083.60
GA	660,000	1,353,416	/	22,482.03
ACO	/	/	/	/

6 | Conclusion

To address the pump scheduling problem in multiproduct pipelines, this article establishes an integer linear programming model based on the discrete-time representation, with various technical and operational constraints, including pressure limits, pump characteristics, and avoidance of frequent pump on/off switching operations. The paper mathematically expresses the pump scheduling in multiproduct pipelines as a MDP and solves it using PPO. In addition, multi-discrete action space was introduced into the PPO applied in large-scale pipelines to enhance optimization efficiency. The experimental results support the effectiveness and applicability of the PPO algorithm in solving pump scheduling problems through a real-world multiproduct pipeline with nine stations (including five pump stations) and two operation modes. The PPO algorithm can learn suboptimal pump scheduling policies for various operation modes of multiproduct pipelines. The superiority and accuracy of the enhanced PPO algorithm with action space shaping in

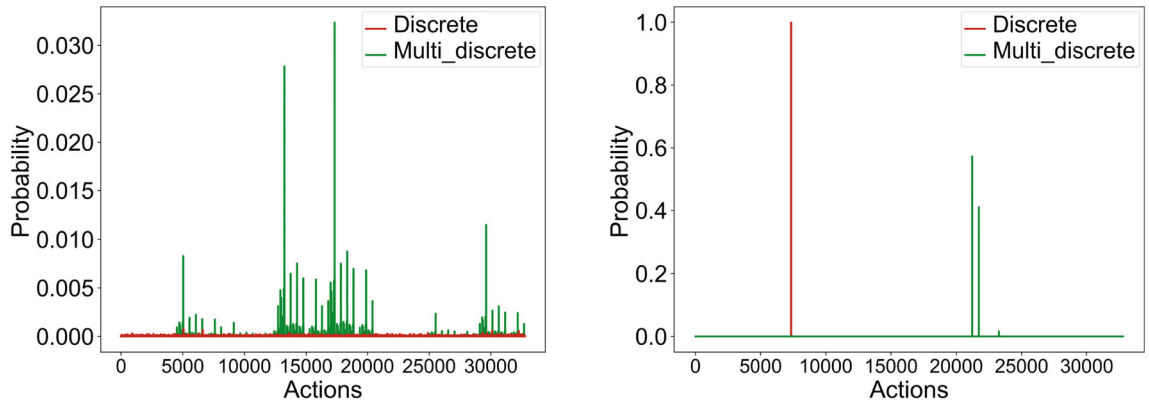

FIGURE 9 | The action probabilities on multi-discrete action spaces and discrete action spaces vary across different training episodes. Here (a) episode = 100 and (b) episode = 600.

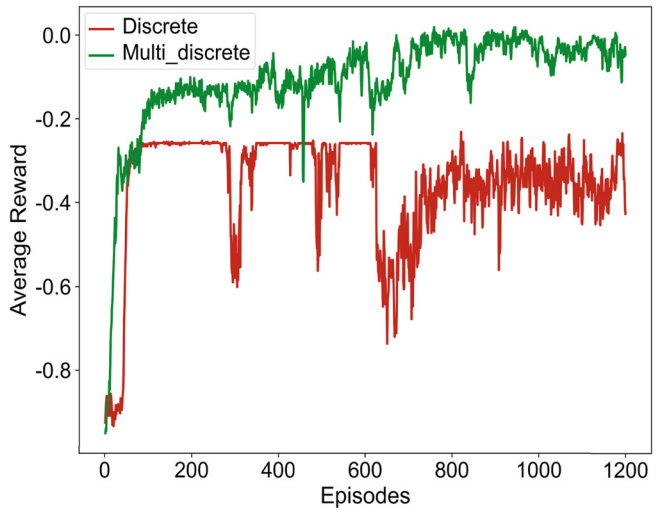

FIGURE 10 | The average reward in each episode comparison of agent in two types of action space.

large-scale pipelines are validated through a comparison with the results obtained by dynamic programming (DP), genetic algorithms (GA), and ant colony optimization (ACO). Moreover, the experiment shows that multi-discrete action space exhibits conducive exploration and more effective exploitation in large-scale pipelines.

However, the proposed method determines the pump scheduling based on a given detailed pipeline schedule, which results in operational flow rates determined by the detailed pipeline schedule not necessarily falling into the pump's efficient operating region. Future work will focus on the joint optimization of detailed scheduling and pump scheduling in multiproduct pipelines.

Nomenclature

Sets and Indices

$t \in T$	set of discrete time
τ_t	discrete time interval
$i \in I$	set of special points along the pipeline (including stations and special elevation points)
$ID \subset I$	set of delivery stations
$IH \subset I$	set of high elevation points
$IL \subset I$	set of low elevation points
$IP \subset I$	set of pump stations
$k \in K_i$	set of pumps in pump station i
$j \in J$	set of batches

Parameters

$\eta_{t,i,k}$	efficiency of pump k at pump station i at time t
λ, m	constant coefficients used to compute friction loss, dependent on the flow regime where $(0 < m < 1)$
ρ_j	product's density of batch j $\left(\mathrm{kg/m^3}\right)$
$\rho_{t,i}$	product's density in point i at time t $\left(\mathrm{kg/m^3}\right)$
$a_{i,k}, b_{i,k}$	constant coefficient of pump k at pump station i, determined by pump characteristics
$c_{i,k}, d_{i,k}, e_{i,k}$	efficiency equation coefficients of pump k at pump station i
d_i	internal diameter of segment $(i, i+1)$ (m)
$E_{i,k}$	cost of switch of pump k at pump station i (CNY/time)
F_i	electricity price at pump station i (CNY/kW · h)
g	gravitational acceleration $\left(\mathrm{m/s^2}\right)$
$H_{t,i,k}$	head of pump k at pump station i at time t
$L_{t,i,j}$	length of batch j in segment $(i, i+1)$ at time t (m)
$P_i^{HIGH,MIN}$	minimum pressure of high elevation points i (MPa)
$P_i^{IN,MAX}$	maximum inlet pressure of point i (MPa)
$P_i^{IN,MIN}$	minimum inlet pressure of point i (MPa)
$P_i^{LOW,MAX}$	maximum pressure of high elevation points i (MPa)
$P_i^{OUT,MAX}$	maximum outlet pressure of point i (MPa)
$P_i^{OUT,MIN}$	minimum outlet pressure of point i (MPa)
$P_{t,i}^{F}$	frictional loss in segment $(i, i+1)$ at time t (MPa)
$P_{t,i}^{H}$	pressure provided by pump station i at time t (MPa)
$P_{t,i}^{IN}$	inlet pressure of point i at time t (MPa)
$P_{t,i}^{OUT}$	outlet pressure of point i at time t (MPa)

$Q_{t,i}$ flow rate of segment $(i, i+1)$ at time t $\left(\mathrm{m}^3/\mathrm{h}\right)$

v_j product's kinematic viscosity of batch j $\left(\mathrm{m}^2/\mathrm{s}\right)$

V_t^I injection volume at initial station I in discrete time interval $(t, t+1)$ (m^3)

$V_{t,i}^D$ delivery volume at delivery station i in discrete time interval $(t, t+1)$ (m^3)

$Z_{t,i,j}$ elevation drop of batch j in segment $(i, i+1)$ at time t (m)

Binary Variables

$B_{t,i,k}$ pump k at pump station i at time t is on/off if $B_{t,i,k} = 1/0$

$S_{t,i,k}$ pump k at pump station i at time t is switched/nonswitched if $S_{t,i,k} = 1/0$

Acknowledgments

This work was supported by the National Key Research and Development Program of China under Grant Nos. 2021YFA1000104 and 2021YFA1000100, the Laboratory Project of Higher Education Institutions in Shandong Province-Energy System Intelligent Management and Policy Simulation Laboratory at China University of Petroleum, and the Youth Innovation Team of Higher Education Institutions in Shandong Province-Data Intelligence Innovation Team at China University of Petroleum.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. X. Xue, X. Ai, J. Fang, W. Yao, and J. Wen, "Real-Time Schedule of Integrated Heat and Power System: A Multi-Dimensional Stochastic Approximate Dynamic Programming Approach," *International Journal of Electrical Power & Energy Systems* 134 (2022): 107427.

2. S. Song, T. Li, P. Liu, and Z. Li, "The Transition Pathway of Energy Supply Systems Towards Carbon Neutrality Based on a Multi-Regional Energy Infrastructure Planning Approach: A Case Study of China," *Energy* 238 (2022): 122037.

3. B. Zhou, J. Fang, X. Ai, C. Yang, W. Yao, and J. Wen, "Dynamic Var Reserve-Constrained Coordinated Scheduling of LCC-HVDC Receiving-End System Considering Contingencies and Wind Uncertainties," *IEEE Transactions on Sustainable Energy* 12, no. 1 (2020): 469–481.

4. Z. Li, Y. Liang, Q. Liao, B. Zhang, and H. Zhang, "A Review of Multiproduct Pipeline Scheduling: From Bibliometric Analysis to Research Framework and Future Research Directions," *Journal of Pipeline Science and Engineering* 1, no. 4 (2021): 395–406.

5. Y. Liang, M. Li, and J. Li, "Hydraulic Model Optimization of a Multi-Product Pipeline," *Petroleum Science* 9 (2012): 521–526.

6. S. MirHassani, M. Abbasi, and S. Moradi, "Operational Scheduling of Refined Product Pipeline With Dual Purpose Depots," *Applied Mathematical Modelling* 37, no. 8 (2013): 5723–5742.

7. C. Zeng, C. Wu, L. Zuo, B. Zhang, and X. Hu, "Predicting Energy Consumption of Multiproduct Pipeline Using Artificial Neural Networks," *Energy* 66 (2014): 791–798.

8. M. Yuan, H. Zhang, Y. Long, R. Shen, B. Wang, and Y. Liang, "Economic, Energy-Saving and Carbon-Abatement Potential Forecast of Multiproduct Pipelines: A Case Study in China," *Journal of Cleaner Production* 211 (2019): 1209–1227.

9. D. Fooladivanda and J. A. Taylor, "Energy-Optimal Pump Scheduling and Water Flow," *IEEE Transactions on Control of Network Systems* 5, no. 3 (2017): 1016–1026.

10. L. Yongtu, L. Ming, and Z. Ni, "A Study on Optimizing Delivering Scheduling for a Multiproduct Pipeline," *Computers and Chemical Engineering* 44 (2012): 127–140.

11. X. Zhou, H. Zhang, R. Qiu, et al., "A Hybrid Time MILP Model for the Pump Scheduling of Multi-Product Pipelines Based on the Rigorous Description of the Pipeline Hydraulic Loss Changes," *Computers and Chemical Engineering* 121 (2019): 174–199.

12. R. Rejowski, Jr. and J. M. Pinto, "Scheduling of a Multiproduct Pipeline System," *Computers and Chemical Engineering* 27, no. 8–9 (2003): 1229–1246.

13. H. Chen, L. Zuo, C. Wu, et al., "Optimizing Detailed Schedules of a Multiproduct Pipeline by a Monolithic MILP Formulation," *Journal of Petroleum Science and Engineering* 159 (2017): 148–163.

14. A. Herrán, J. M. de la Cruz, and B. de Andrés, "A Mathematical Model for Planning Transportation of Multiple Petroleum Products in a Multi-Pipeline System," *Computers and Chemical Engineering* 34, no. 3 (2010): 401–413.

15. D. C. Cafaro and J. Cerdá, "Optimal Scheduling of Multiproduct Pipeline Systems Using a Non-Discrete MILP Formulation," *Computers and Chemical Engineering* 28, no. 10 (2004): 2053–2068.

16. P. M. Castro and H. Mostafaei, "Product-Centric Continuous-Time Formulation for Pipeline Scheduling," *Computers and Chemical Engineering* 104 (2017): 283–295.

17. R. Rejowski, Jr. and J. M. Pinto, "A Novel Continuous Time Representation for the Scheduling of Pipeline Systems With Pumping Yield Rate Constraints," *Computers and Chemical Engineering* 32, no. 4–5 (2008): 1042–1066.

18. Y. Zhang, C. Yu, Y. Xu, and K. L. Teo, "Minimizing Control Variation in Discrete-Time Optimal Control Problems," *Journal of Computational and Applied Mathematics* 292 (2016): 292–306.

19. L. Magatão, L. V. Arruda, and F. Neves, Jr., "A Mixed Integer Programming Approach for Scheduling Commodities in a Pipeline," *Computers and Chemical Engineering* 28, no. 1–2 (2004): 171–185.

20. D. Zyngier and J. D. Kelly, "Multi-Product Inventory Logistics Modeling in the Process Industries," *Optimization and Logistics Challenges in the Enterprise* 30 (2009): 61–95.

21. E. Abbasi and V. Garousi, "An MILP-Based Formulation for Minimizing Pumping Energy Costs of Oil Pipelines: Beneficial to Both the Environment and Pipeline Companies," *Energy Systems* 1 (2010): 393–416.

22. S. Xin, Y. Liang, X. Zhou, et al., "A Two-Stage Strategy for the Pump Optimal Scheduling of Refined Products Pipelines," *Chemical Engineering Research and Design* 152 (2019): 1–19.

23. S. Relvas, H. A. Matos, A. P. F. Barbosa-Póvoa, J. Fialho, and A. S. Pinheiro, "Pipeline Scheduling and Inventory Management of a Multiproduct Distribution Oil System," *Industrial and Engineering Chemistry Research* 45, no. 23 (2006): 7841–7855.

24. Y. Liang, X. Zhou, X. Yan, H. Zhang, and J. Liu, "Optimization of Pump Start-Up Schemes for Large-Scale Multiproduct Pipelines," *Journal of China University of Petroleum* 41, no. 5 (2017): 130–138.

25. X. Zhou, Y. Liang, Q. Liao, J. Ma, and H. Zhang, "Optimization on the Pump Start-Up Scheme of Products Pipelines Based on Parallel SAGA," *Transportation & Storage* 4 (2018): 395–402.

26. R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction* (Cambridge, MA: MIT Press, 1998).

27. O. Vinyals, I. Babuschkin, W. M. Czarnecki, et al., "Grandmaster Level in StarCraft II Using Multi-Agent Reinforcement Learning," *Nature* 575, no. 7782 (2019): 350–354.

28. D. Bertsekas, *Dynamic Programming and Optimal Control*, vol. I, 4th ed. (Belmont, MA: Athena Scientific, 2012).

29. F. L. Lewis and D. Vrabie, "Reinforcement Learning and Adaptive Dynamic Programming for Feedback Control," *IEEE Circuits and Systems Magazine* 9, no. 3 (2009): 32–50.

30. G. Hajgató, G. Paál, and B. Gyires-Tóth, "Deep Reinforcement Learning for Real-Time Optimization of Pumps in Water Distribution Systems," *Journal of Water Resources Planning and Management* 146, no. 11 (2020): 04020079.

31. S. Hu, J. Gao, D. Zhong, R. Wu, and L. Liu, "Real-Time Scheduling of Pumps in Water Distribution Systems Based on Exploration-Enhanced Deep Reinforcement Learning," *System* 11, no. 2 (2023): 56.

32. B. V. Mbuwir, D. Geysen, F. Spiessens, and G. Deconinck, "Reinforcement Learning for Control of Flexibility Providers in a Residential Microgrid," *IET Smart Grid* 3, no. 1 (2020): 98–107.

33. T. Peirelinck, F. Ruelens, and G. Decnoninck, "Using Reinforcement Learning for Optimizing Heat Pump Control in a Building Model in Modelica," in *2018 IEEE International Energy Conference (ENERGYCON)* (Limassol, Cyprus: IEEE, 2018), 1–6.

34. C. Patyn, F. Ruelens, and G. Deconinck, "Comparing Neural Architectures for Demand Response Through Model-Free Reinforcement Learning for Heat Pump Control," in *2018 IEEE International Energy Conference (ENERGYCON)* (Limassol, Cyprus: IEEE, 2018), 1–6.

35. J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," arXiv preprint arXiv:1707.06347 (2017).

36. G. Dulac-Arnold, R. Evans, V. H. Hasselt, et al., "Deep Reinforcement Learning in Large Discrete Action Spaces," arXiv preprint, arXiv:1512.07679 (2015).

37. A. Kanervisto, C. Scheller, and V. Hautamäki, "Action Space Shaping in Deep Reinforcement Learning," in *2020 IEEE Conference on Games (CoG)* (Osaka, Japan: IEEE, 2020), 479–486.