Int. J. Simul. Multidisci. Des. Optim. **17**, 1 (2026)

https://doi.org/10.1051/smdo/2025034

IJSMDO
Available online at:
https://www.ijsmdo.org

Research Article Open Access

Rehabilitation robot trajectory planning method for upper limb based on healthy limb motion using multi-objective constrained reinforcement learning

Haotian Xu[1], Bingjing Guo[1,2,3,*], Jianhai Han[1,2,3], Xiangpan Li[1,2,3], and Zhenzhu Li[1]

[1] School of Mechatronics Engineering, Henan University of Science and Technology, Luoyang 471003, PR China
[2] Henan Provincial Key Laboratory of Robotics and Intelligent Systems, Luoyang 471003, PR China
[3] Collaborative Innovation Center of Machinery Equipment Advanced Manufacturing of Henan Province, Luoyang 471003, PR China

Received: 4 August 2025 / Accepted: 5 November 2025

Abstract. Stroke patients with hemiplegia require personalized upper-limb rehabilitation, yet designing safe and effective robot-assisted trajectories that mimic natural human movement remains a significant challenge. This paper proposes a trajectory planning and optimization method to address this need by leveraging multi-objective constrained reinforcement learning. The method involves dynamically capturing motion data from the patient's healthy limb to define personalized Activities of Daily Living (ADL). A reinforcement learning algorithm, guided by a specially designed reward-punishment function, then optimizes the trajectory with objectives for smoothness, jerk minimization, and accurate tracking of key points. The approach was validated on a 4-degree-of-freedom (4-DOF) upper limb rehabilitation robot, which successfully achieved multi-joint coordinated trajectory tracking based on the learned ADL movements. The experiments confirm the method's effectiveness in designing personalized rehabilitation trajectories that improve the continuity and smoothness of robot-assisted movements, offering a promising solution for patient-specific therapy.

Keywords: Upper limb rehabilitation robot / reinforcement learning / healthy limb exercises / multi-objective trajectory optimization

1 Introduction

Stroke is recognized as an acute cerebrovascular disease and is the second leading cause of disability and death worldwide. It not only places a significant economic and psychological burden on patients and their families, but also exerts considerable pressure on social healthcare resources [1,2]. Rehabilitation robots can replace rehabilitation physicians to perform rehabilitation training for patients with movement disorders, which can effectively alleviate the shortage of rehabilitation physicians [3], and thus have received widespread attention.

Rehabilitation treatment should consider not only the specific type and duration of the patient's condition but also factors such as the patient's physical condition, motor ability, and recovery progress to tailor the rehabilitation program [4]. Traditional 'one-size-fits-all' approaches are often inadequate to address individual differences, whereas personalized rehabilitation trajectory design can be developed based on the patient's actual needs, enabling the creation of accurate and effective exercise plans [5]. At the same time, the humanoid motion trajectory enables the robot to mimic the natural movement pattern of the human body, avoiding mechanized or rigid movements, and can carry out movement training in a more physiological way, so as to make the patient's movement closer to normal human movement, and help the patient to restore the normal movement posture and muscle coordination. As the design of the humanoid movement trajectory is more in line with the physiological structure and movement style of the human body, patients can usually reduce discomfort and pain during the training process, and avoid problems such as muscle strains or joint injuries due to unnatural robot movements [6,7].

Task-oriented training designed on the basis of ADL (e.g., dressing, grooming, transferring, eating, etc.) allows patients to actively participate in training in various environments, and makes it easier for patients to master daily living skills [8]. These kinds of movements in clinical rehabilitation are varied, trajectories are complex, and vary from person to person, and it is difficult to satisfy

* e-mail: bingjing@haust.edu.cn

patients' individualized adjustments with the methods based on the optimization of rehabilitation trajectories by traditional geometry and kinetics, but because of the However, due to the different rehabilitation needs, muscle strength and degree of limitation of patients, personalized rehabilitation training is one of the core trends of upper limb rehabilitation robotics [9]. When the rehabilitation trajectory that meets the individualized needs is implemented in the robot, it is also necessary to consider improving the flexibility and smoothness of the rehabilitation process and reducing the degree of impact.

However, designing such personalized and human-like trajectories presents several key challenges. First, traditional trajectory planning methods often struggle to balance multiple conflicting objectives simultaneously, such as ensuring smoothness, minimizing jerk, and maintaining precise tracking of key rehabilitation points. Second, creating trajectories that are truly personalized and adaptive to a patient's specific condition and recovery progress remains a complex task that "one-size-fits-all" approaches cannot adequately address. Finally, there is a need for an intelligent optimization strategy that can learn complex, human-like motion patterns directly from data, rather than relying on pre-defined mathematical models. This study aims to address these specific problems.

Teramae et al. [10] proposed a method to optimize the quaternionic posture matrix and the target EMG signal difference as the target equation, considering that the affected limbs have different muscle co-movement patterns during rehabilitation training. The SPGP model and targeted adjustment of rehabilitation movements were utilized to train the corresponding muscles. Xu Di et al. [11] conducted trajectory planning for a five-degree-of-freedom upper limb rehabilitation robot based on five quasi-uniform B-splines, and then optimized the planned trajectory using particle swarm algorithm to obtain the impact-optimal trajectory profile. Applying intelligent optimization algorithms to robot trajectory and path planning is a significant research direction; for instance, Mohamadwasel and Kurnaz implemented the social spider optimization algorithm for controlling a parallel robot [12], while Basil et al. have explored a hybrid optimization approach for UAV path planning [13] and an accelerated black hole optimization algorithm for mobile robot systems [14].

Li et al. [15] developed a modular upper limb rehabilitation exoskeleton device, based on the theory of cubic polynomial interpolation to design a multi-stage joint motion trajectory planning scheme, this technical program will decompose the mechanical arm motion trajectory into four consecutive stages, each stage uses independent cubic polynomial function for trajectory description, and finally realize the full path planning through the four segments of the cubic curves, and the experimental results show that this multi-stage interpolation Experimental results show that the trajectory curves generated by this multi-segment interpolation algorithm are smooth and continuous, and the joint motion parameters do not have sudden changes, which is in line with the biomechanical characteristics of human movement and helps to control the joint stability in the process of patients' rehabilitation training.

Although the above methods have made progress, specific limitations remain. Trajectory planning based on B-splines and particle swarm optimization can achieve optimization for specific goals like impact reduction, but may struggle to adapt in real-time to the dynamic and personalized needs of patients. Similarly, methods using polynomial interpolation excel at creating smooth, continuous curves but often lack the flexibility to generate complex, non-uniform humanoid trajectories that are truly tailored to an individual's unique motion characteristics. These approaches face significant challenges when confronted with the complex demands of multi-constraint (e.g., smoothness, jerk, accuracy) and personalized collaborative optimization.

To overcome these issues, more adaptive intelligent optimization strategies are required. Reinforcement learning, in particular, is well-suited for this challenge as it transforms the strategy design problem into a sequential decision-making optimization problem. It can learn optimal, adaptive policies directly from interaction, making it a powerful tool for developing truly personalized and complex motion trajectories. For example, Yang Aolei et al. [16] extracted the shoulder angle, elbow angle, and wrist joint motion angle after normality analysis and correlation analysis to obtain the rules of human arm motion characteristics, designed the corresponding reward function according to the different motion characteristic rules, and used reinforcement learning to train the humanoid motion model of the robotic arm to verify the feasibility of the reinforcement learning algorithm. Wei et al. [17] designed the optimal posture of the robotic arm and realized the optimal posture of the robotic arm by optimizing the five objective functions. Duarte et al. [18] designed different motion states and used Gaussian mixture model computation to achieve the control of the humanoid motion of the robot.

In this study, for a 4-DOF upper limb exoskeleton rehabilitation robot, an optical motion capture system is used to collect the ADL training motion trajectory of the patient's healthy limb, extract the motion features during the training process, and use the learning algorithms to simulate and learn to obtain the smooth joint acceleration and additive acceleration, and learn to optimize the trajectory with the comprehensive constraints of impact optimization, having the patient's personalized features and human-like behaviors, so as to provide the patient with a safe and efficient rehabilitation The program has distinctive research characteristics.

The main contributions of this paper can be summarized as follows:

- A novel framework for generating personalized rehabilitation trajectories by capturing and modeling motion data from the patient's own healthy limb.
- The application of a multi-objective constrained reinforcement learning algorithm to simultaneously optimize trajectory smoothness, jerk, and tracking accuracy, addressing the complex, multi-constraint nature of the problem.
- The design of a tailored reward-punishment function that effectively guides the reinforcement learning agent to learn anthropomorphic and safe motion policies.

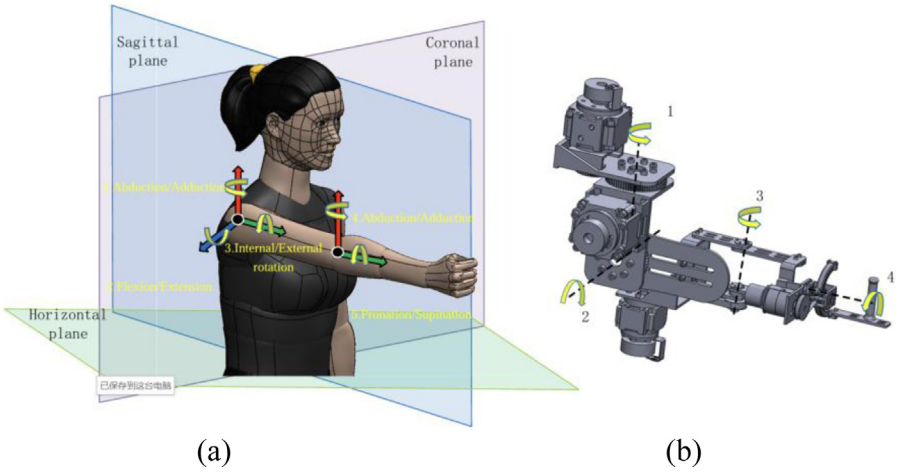

Fig. 1. (a) Human upper limb model; (b) Structural diagram of upper limb rehabilitation robot.

– Experimental validation of the proposed method on a 4-DOF upper limb rehabilitation robot, demonstrating its effectiveness and practical applicability.

The remainder of this paper is organized as follows. Section 2 details the materials and methods, including the kinematic model of the 4-DOF upper limb rehabilitation robot and the process for generating ADL-motion trajectories. Section 3 presents the proposed trajectory optimization strategy based on the Proximal Policy Optimization (PPO) reinforcement learning algorithm and elaborates on the design of the multi-objective reward function. Section 4 describes the simulation and physical prototype experiments, presenting and analyzing the results to validate the effectiveness of our method. Finally, Section 5 concludes the paper and discusses potential future work.

2 Materials and methods

2.1 Upper limb rehabilitation robot and ADL-motion trajectory generation

Based on the anatomical structure of the human upper limb, an equivalent DOF definition of the upper limb is established, as shown in Figure 1a. Due to the limited range of motion of the shoulder joint around the sagittal axis during rehabilitation exercises in the human body, most of the rehabilitation exercise can ignore the shoulder joint around the sagittal axis of the movement, only consider the shoulder joint around the vertical axis and the coronal axis of the movement. Therefore, rehabilitation training of the human shoulder joint complex movement is achieved through a dual degree of freedom drive mechanism consisting of coronal flexion and extension of the shoulder joint and sagittal abduction and adduction (as shown in Fig. 1b, the shoulder joints of the rehabilitation robot we developed).

Each joint of the upper limb rehabilitation robot is driven by the swing cylinder controlled by the proportional valve, as shown in Figure 1b. The pneumatic driving method increases the flexibility of the robot mechanism and improves the comfort of patients' training [19]. The first and second joints of the robot realize the abduction/adduction of the shoulder in the horizontal plane and the flexion/extension in the sagittal plane. The third joint of the robot realizes the abduction/adduction of the elbow joint in the horizontal plane, and the fourth joint of the robot realizes the rotation of the forearm. During the rotation of the patient's forearm, due to the integration of the human arm and the coordination of the upper arm and the forearm, the patient's upper arm can rotate with it. Therefore, the designed robot's DOF is consistent with the range of motion (ROM) required by the human upper limb rehabilitation training.

In rehabilitation training, it is essential to ensure the safety and comfort of patients during exercise. Therefore, when adopting upper limb rehabilitation robots to assist patients in rehabilitation training, improving the personalization and human-likeness of the robot's movements can achieve better rehabilitation experiences and outcomes. The rehabilitation trajectory is first planned and optimized for human-like characteristics in a simulation environment. To enhance the realism of the simulation environment and improve the training effectiveness of control strategies, it is necessary to acquire human motion trajectories that replicate upper limb rehabilitation training movements.

Once the affected limb achieves an appropriate joint range of motion (ROM), Activities of Daily Living (ADL) training is typically initiated in clinical practice. ADL refers to the routine tasks individuals perform in daily life. Through functional exercises, patients can enhance their ADL capabilities and improve quality of life [20,21].

The first step involves acquiring human motion data. The experiments were conducted with three healthy subjects. Data was collected using a NOKOV optical motion capture system Mars2H (Beijing Nokov Science & Technology Co., Ltd.) at a sampling rate of 60 Hz. Before each session, the system was calibrated according to the manufacturer's protocol, and data collection proceeded only after the system's software confirmed a successful calibration. For this study, we focused on the representative ADL task of drinking water, and each subject performed five trials. The marker layout, attached to the shoulder, upper arm, and forearm, is shown in Figure 2.

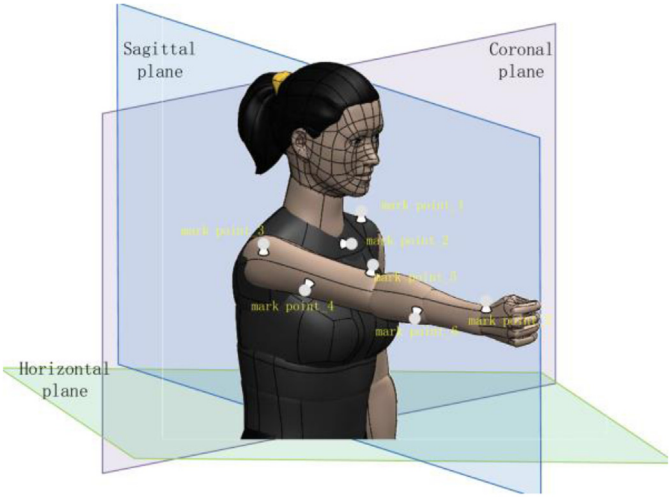

Fig. 2. Motion capture marker points.

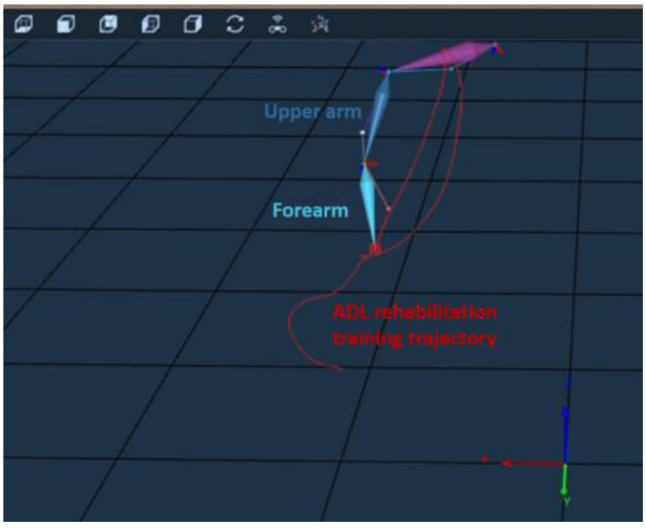

Fig. 3. Motion trajectories captured by the optical motion capture system.

The raw marker data was processed and filtered using the proprietary software bundled with the system to generate the joint motion trajectories. Figure 3 illustrates a sample trajectory from one of these trials.

Next, the end position, velocity, acceleration, rigid body distance, and orientation of the ADL rehabilitation training motion are selectively exported. Joint angles are calculated based on the collected kinematic data. Through a mapping algorithm, the acquired joint angles are mapped to the robotic model in the simulation environment to derive equivalent joint space trajectories. Equidistant key nodes are then extracted from these trajectories to form the training dataset.

Directly mapping the ADL training trajectory of the healthy limb as the desired motion trajectory of the affected limb will make the rehabilitation training more difficult to achieve the desired effect due to the natural jerks generated during the natural movement of the healthy limb. Therefore, it is necessary to extract motion features through isometric extraction and use reinforcement learning algorithms to train better performance motion trajectories, which can enhance the impact and smoothness performance of the upper limb rehabilitation robot during motion while preserving the patient's individualized training needs.

2.2 Reinforcement learning algorithms

Reinforcement learning can optimize the robot's trajectory by continuously interacting with the environment, adapting to the needs of individual patients and automatically adjusting the rehabilitation training program. This approach trains on the specific needs of each patient and generates a personalized multi-objective optimization strategy for the rehabilitation trajectory.

We adopt the proximal policy optimization PPO (Proximal Policy Optimization) algorithm framework and enhance the policy exploration ability by introducing the entropy regularization mechanism to construct the trajectory optimization control strategy for upper limb rehabilitation robots. As a deep reinforcement learning algorithm based on the policy gradient [22], PPO is used to construct the upper limb rehabilitation robot trajectory optimization control strategy by introducing the importance sampling and the policy updating magnitude of the Clip constraint mechanism to realize efficient policy search while ensuring training stability [23]. Compared with the traditional policy gradient method, the PPO algorithm can avoid policy collapse due to over-optimization by restricting the maximum step size of the policy update, and it can use the probability ratio between the old policy and the new policy to achieve sample reuse and improve data efficiency.

We adopt the framework of proximal policy optimization algorithm and enhance the policy exploration ability by introducing the entropy regularization mechanism. The core objective function of the PPO algorithm can be expressed as:

$$L^{CLIP}(\theta) = \mathbb{E}_t[min(r_t(\theta)A_t, clip(r_t(\theta), 1-\varepsilon, 1+\varepsilon)A_t)] \tag{1}$$

where θ is the strategy network parameter and $r_t(\theta)$ denotes the ratio of old to new strategy probabilities:

$$r_t(\theta) = \pi_\theta(a_t|s_t)/\pi_{\theta_{old}}(a_t|s_t) \tag{2}$$

s_t is the state space of the joint at moment t, a_t obeys the distribution of actions generated by the upper limb rehabilitation robot based on the state of the environment s_t, A_t is the dominance function, ε is the shear threshold, and the probability ratio shear term:

$$clip((r_t(\theta), 1-\varepsilon, 1+\varepsilon) \tag{3}$$

The design limits the KL scatter of the policy update to a preset confidence interval by constraining the deviation amplitude of the importance sampling weight

$r_t(\theta) = \pi_\theta / \pi_{\theta_{old}}$. Its geometric significance lies in constructing the trust domain in the strategy parameter space, so that the difference between the strategy distribution π_θ updated in each iteration and the old strategy $\pi_{\theta_{old}}$ is controlled, thus avoiding the strategy collapse phenomenon due to over-optimization. This design avoids premature convergence to the local optimal solution by restricting the magnitude of the strategy update while ensuring the stability of training.

In order to improve the diversity of strategies, this paper introduces the entropy term:

$$\mathcal{H}(\pi_\theta) = -\mathbb{E}_a[log\pi_\theta(a|s)] \tag{4}$$

As a quantitative indicator of the stochasticity of the strategy, the dynamic regulation of the exploration-utilization trade-off is achieved through the coefficient β:

$$\beta_{t+1} = \beta_t \cdot \exp(-\alpha \cdot \text{sign}(\mathcal{H}_t - \mathcal{H}_{target})) \tag{5}$$

where α is the adaptive regulation rate and $\mathscr{H}_{target}$ is the target entropy value. When the strategy entropy is higher than the target value, β decays to inhibit over-exploration; conversely, it enhances the entropy regularization effect. This mechanism keeps the algorithm highly exploratory at the beginning of training and gradually converges to a deterministic strategy as the strategy optimization progresses.

The strategy network outputs a diagonal Gaussian distribution with parameter (μ, σ), which is constructed as:

$$\begin{cases} \mu = f_\mu(s;\theta) \\ log\sigma = f_\sigma(s;\theta) + log\sigma_0 \end{cases} \tag{6}$$

where f_μ and f_σ are the neural network mapping functions and $\log \sigma_0$ is the trainable base log standard deviation. This explicit parameterization is designed so that the action variance is initially large at σ to ensure sufficient exploration, the variance scale is automatically adjusted by gradient backpropagation during training, and σ converges to the minimum safe threshold at final convergence to ensure control accuracy.

2.3 Reinforcement learning based trajectory optimization strategy

The proposed method formulates the trajectory optimization task as a reinforcement learning problem. The process begins with a baseline reference trajectory generated from the patient's mirrored healthy limb motion. The PPO agent then learns to refine this trajectory through interaction with the simulation environment. In each step, the agent observes the current state (comprising joint angles, angular velocities, and normalized time) and selects an action. Through this iterative process of observation and correction, the agent is trained via a multi-objective reward function to produce a final trajectory that has minimal jerk, high smoothness, and meets the patient's personalized needs. The overall training block diagram is shown in Figure 4.

2.4 Multi-objective trajectory optimization reward function design

The design of reward function is based on the theory of collaborative optimization, aiming to achieve the multi-objective optimization task between personalization, smoothness and safety in the rehabilitation trajectory tracking of upper limb rehabilitation robot. The reward function $R(s, a)$ adopts a hierarchical fusion architecture, which realizes the dynamic priority assignment of different optimization objectives through a nonlinear coupling mechanism, defined as:

$$R(s,a) = \sum_{i=1}^{4} w_i \cdot \phi_i(f_i(s,a)) \tag{7}$$

where w_i is the dynamic weight coefficient, $\phi_i(\cdot)$ is the nonlinear transformation function, and $f_i(s, a)$ characterizes the original evaluation index of each optimization objective.

To preserve the personalized characteristics of patients, it is necessary to establish trajectory accuracy metrics. The desired trajectory tracking accuracy R_{track} is realized by an improved L1 deviation metric function:

$$R_{track} = \alpha_1 \left(1 - \frac{1}{n}\sum_{i=1}^{n} |q_i^{act} - q_i^{ref}|\right) \tag{8}$$

where q_i^{act} and q_i^{ref} denote the actual and reference joint angles, respectively, and α_1 is the accuracy weighting coefficient. Compared with the traditional L2 paradigm, the L1 metric has higher sensitivity to abnormal deviations, and its derivative discontinuity can effectively enhance the fast response ability of the strategy to trajectory deviation. Theoretical analysis shows that this design makes the strategy gradient show exponential growth when the deviation increases, which significantly improves the convergence efficiency.

In order to ensure the smoothness of the motion during the rehabilitation process, the motion smoothness optimization reward function R_{smooth} is implemented by means of second-order differential constraints:

$$R_{smooth} = \alpha_2 \exp(-\lambda \cdot \frac{1}{n}\sum_{i=1}^{n} (\ddot{q}_i^{\,2} + \eta |\dddot{q}_i|)) \tag{9}$$

where $\ddot{q}_i$ and $\dddot{q}_i$ denote the joint angular acceleration and impact degree, respectively, λ is the sensitivity adjustment factor, and η is the impact degree penalty weight. The exponential decay form transforms the higher-order dynamics constraints into a smooth reward surface, which ensures the consistency of the optimization direction through the positivity of the Hessian matrix, and the introduction of the impact term suppresses the torque mutation phenomenon to prolong the life of mechanical components.

The velocity safety barrier function $R_{velocity}$ is designed as a segmented linear form:

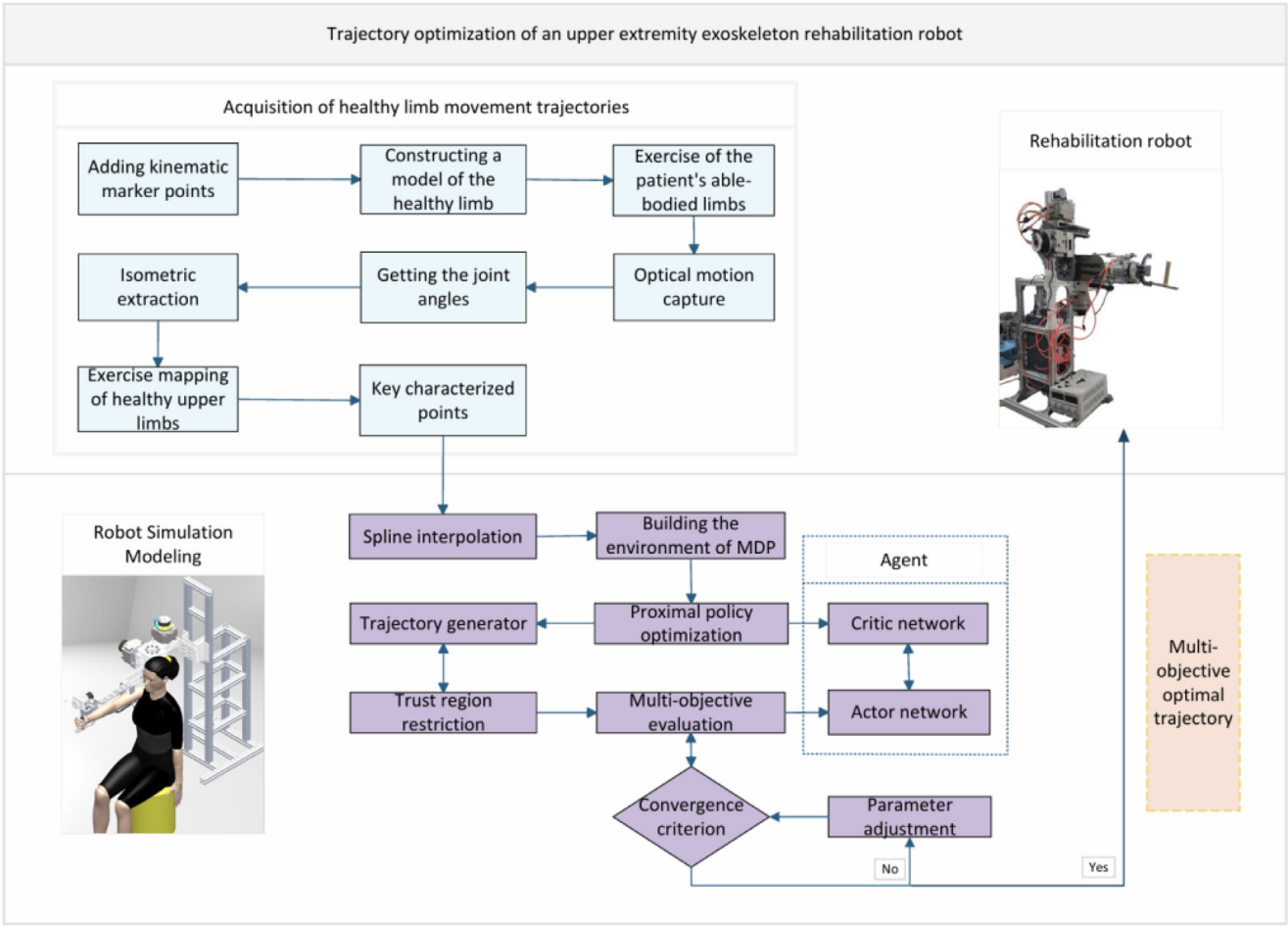

Fig. 4. Flowchart of the trajectory optimization strategy.

$$R_{velocity} = \begin{cases} \alpha_3, & \overline{v} \le V_{safe} \\ \alpha_3 - \beta(\overline{v} - v_{safe}), & V_{safe} < \overline{v} \le V_{max} \\ -\infty, & \overline{v} > V_{max} \end{cases} \tag{10}$$

where $\overline{v}$ is the average joint velocity, V_{Safe} is the safety threshold, and β is the penalty slope. The hard constraint term $-\infty$ realizes the non-violability of the safety specification through the reward collapse mechanism, the gradient direction of the linear penalty interval is clear, which guides the strategy to find the optimal within the safety boundary, and the environmental adaptability is enhanced through the threshold parameter v_{safe} which can be dynamically adjusted according to the load.

To ensure the individualized needs of patients' rehabilitation exercise, i.e., rehabilitation comfort, a three-stage reward enhancement mechanism R_{key} is proposed to constrain the key points of the trajectory:

$$R_{key} = \begin{cases} \gamma_1 \cdot \exp(-k_1 \Delta q^2), & \Delta q \le \delta_1 \\ \gamma_2 \cdot \text{sign}(\delta_2 - \Delta q), & \delta_1 < \Delta q \le \delta_2 \\ -\gamma_3 \cdot \Delta q^2, & \Delta q > \delta_2 \end{cases} \tag{11}$$

where Δq is the attitude deviation at the critical point and δ_1, δ_2 are the deviation thresholds. The structure contains a dual optimization mechanism:

In the $\Delta q \le \delta_1$ Gaussian reward zone, a local concave optimization surface is formed by exponential decay to improve the positioning accuracy; in the $\delta_1 < \Delta q \le \delta_2$ linear transition zone, the gradient direction is kept constant to avoid the strategy from falling into the local extremes; and in the $\Delta q > \delta_2$ quadratic penalty zone, a gradual penalty is imposed to ensure that the key points cannot be deviated from each other too much.

The multi-objective optimization strategy with low impact, high smoothness and good interaction comfort is realized by separating velocity, acceleration and trajectory accuracy to reduce the coupling interference between objectives.

3 Algorithm training and validation based on robot simulation models

In order to verify the effectiveness of the trajectory optimization algorithm based on improved PPO in the control of rehabilitation exoskeleton, this study constructed a 4-DOF upper limb exoskeleton rehabilitation robot simulation model based on MuJoCo platform, as shown in Figure 5.

The experimenter entered the optical motion capture experiment space and marked the left arm with a reflective sphere, followed by selecting the marking point in the

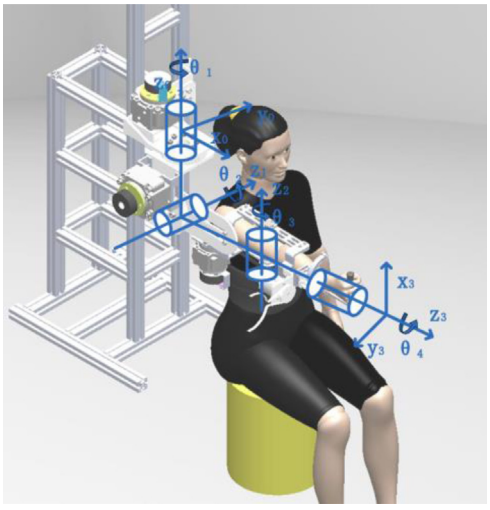

Fig. 5. Simulation model of upper limb exoskeleton rehabilitation robot.

accompanying software to build a model of the left arm. The experimenter performed the specified ADL training maneuvers in the optical motion capture device, and the joint angles were solved and captured by the software in real time. Since the upper-limb exoskeleton robot of this experiment lacks a degree of freedom at the shoulder joint, to prevent deformation of the robot motion, the mapping algorithm is utilized to map the collected joint angles into the upper-limb exoskeleton rehabilitation robot simulation system. While a formal error analysis is beyond the scope of this algorithm-focused study, this mapping process is designed to minimize kinematic discrepancies, and the robot's workspace was verified to be sufficient for the tested ADL tasks. The equivalent joint angles after mapping are shown in Figure 6.

The original joint angle data matrix $\mathbb{Q} \in \mathbb{R}^{20\times4}$ is normalized and the initial reference trajectory is constructed using spline interpolation. Its state space $\mathbf{s}$:

$$\boldsymbol{s}_t = [t, \theta_1, \theta_2, \theta_3, \theta_4] \in \mathbb{R}^5 \tag{12}$$

where $t \in [0,\ 1]$ is the normalized time parameter. The action space $\boldsymbol{a}$:

$$\boldsymbol{a}_t = [\Delta t, \Delta\theta_1, \Delta\theta_2, \Delta\theta_3, \Delta\theta_4] \in \mathbb{R}^5 \tag{13}$$

each joint trajectory $\boldsymbol{q}_i(t)$ satisfies the boundary conditions:

$$\begin{cases} \dot{q}_i(0) = \dot{q}_i(1) = 0 \\ \ddot{q}_i(0) = \ddot{q}_i(1) = 0 \end{cases} \tag{14}$$

A uniform and continuous trajectory $Q_{ref}(t)$ is generated as a reference based on the joint data sampled and processed by the optical motion capture device. The trajectory optimization problem is modelled as a strategy search task in continuous state-action space, and the intelligent body generates the action distribution through the strategy network $\pi_\theta(\boldsymbol{a}|\boldsymbol{s})$. The importance sampling mechanism is used in the training process by minimizing the strategy gradient loss function with constraints:

$$\mathcal{L}^{CLIP}(\theta) = \mathbb{E}_t[min(r_t(\theta)A_t, clip(r_t(\theta), 1-\varepsilon, 1+\varepsilon)A_t)] - \eta\mathcal{H}(\pi_\theta) \tag{15}$$

where $\hat{A}_t$ is the generalized dominance estimator. When $\hat{A}_t > 0$, the strategy gradient rises, increasing the probability of the action that leads to that payoff; when $\hat{A}_t < 0$suppresses the corresponding action, preventing the joint angle from deviating. $\mathscr{H}$ is the strategy entropy regular term, forcing to maintain a certain degree of exploration, avoiding limiting the overall performance due to the local optimum caused by the premature convergence of a joint. The algorithm ensures the stability of the policy update through the trust region constraints, and at the same time adopts dynamic gradient trimming to prevent parameter mutation and gradient explosion. The reward function integrates multi-physics constraints:

$$R(\boldsymbol{s}, \boldsymbol{a}) = \Phi_{base} - \sum_{i=1}^{5} \lambda_i C_i(\boldsymbol{s}, \boldsymbol{a}) + B(\boldsymbol{s}) \tag{16}$$

where C_i characterizes the trajectory deviation (MAE), shock (Jerk), acceleration energy (L2 paradigm), velocity overrun penalty and must-pass point constraints, respectively, and $B(s)$ is the precise arrival reward function.

The algorithm is presented as Algorithm 1. The inputs are keypoint data, maximum training epochs, and batch size, while the output is the optimal trajectory. Initially, a continuous trajectory model is obtained through spline interpolation to serve as the reference trajectory. An interactive environment with a multi-objective reward function is constructed, where the initial input consists of normalized time and four joint angles, and the output is the robot's optimal trajectory. The PPO agent is encapsulated, and after configuring the learning rate, discount factor, and clipping coefficient, an empty experience buffer is created to store interaction data.

In line 5, the main loop begins. Line 6 initializes the cumulative reward for the current episode, episodic_reward. In line 7, batch_size interactions are performed within a single episode. In line 8, normalized time points are randomly sampled. Line 9 constructs the state vector using env.GetState(t). In line 10, ppo_agent.SelectAction(state) is called to sample an action based on a normal distribution and record the log-probability of the action, log_prob. Line 11 applies constraints to the time correction. Lines 12-17 iterate through the four degrees of freedom for the joints: first, the reference angle is obtained using the spline function splines[i], then the joint correction for the action is limited. In line 18, CalculateReward is called to merge tracking error, impact, acceleration, and other objectives to generate the reward value. In line 19, the state, action, log-probability, and reward are stored in the experience buffer. In line 20, the cumulative reward for the current episode is updated. After a single batch interaction, ppo_agent.Update(memory_buffer) is called to perform the core PPO

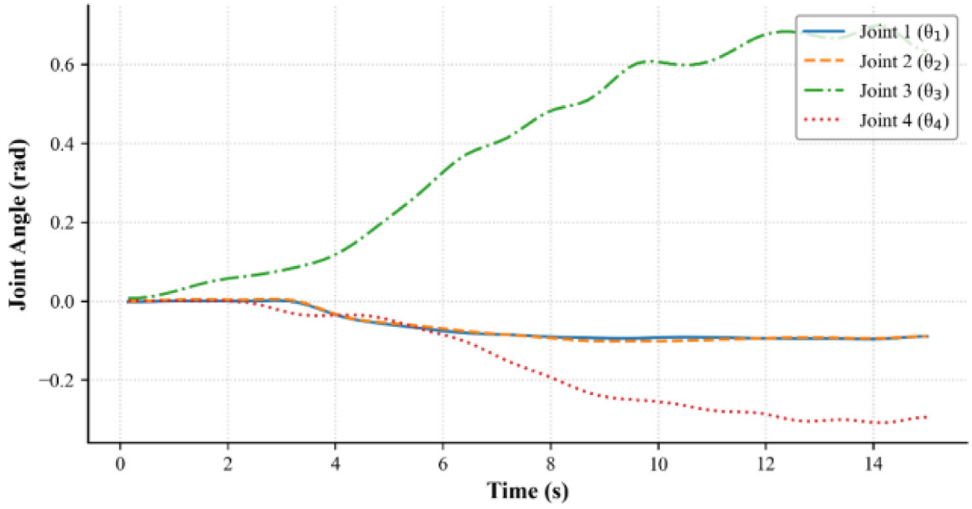

Fig. 6. Equivalent joint angles after mapping.

update: discount cumulative rewards are calculated, multi-round gradient descent is used to optimize the policy, the update magnitude is constrained by the ratio clipping, and an entropy regularization term is introduced to balance exploration and exploitation. Line 23 clears the experience buffer. Every 200 epochs, the learning rate is decayed by 5%, and in the later stages of training, the exploration intensity is reduced to stabilize convergence.

Algorithm 1

Multi-objective Trajectory Optimization of Robot.

Input:Keypoints data from Excel, max_episodes=1000, batch_size=128
Output: Optimized trajectory
1: env ← InitializeEnvironment(splines, original_traj)
2: policy_net ← InitializePolicyNetwork(input_dim=5, output_dim=5)
3: ppo_agent ← PPO(policy_net, lr=3e-4, γ=0.98, clip_ε=0.1)
4: memory_buffer ← EmptyBuffer()
5: **for** episode = 1 to max_episodes **do**
6: episodic_reward ← 0
7: **for** step = 1 to batch_size **do**
8: t ← SampleUniform(0, 1)
9: state ← env.GetState(t)
10: action, log_prob ← ppo_agent.SelectAction(state)
11: new_t ← Clip(t + action[0]×Amp_time, 0, 1)
12: new_angles ← []
13: **for** each joint i **do**
14: base_angle ← splines[i](new_t)
15: delta ← Clip(action[i+1]×Amp_angle, −0.087, 0.087)
16: constrained_angle ← Clip(base_angle + delta, $-\pi/2$, $\pi/2$)
17: new_angles.append(constrained_angle)
18: reward ← CalculateReward(env, new_angles, new_t)
19: StoreTransition(memory_buffer, state, action, log_prob, reward)
20: episodic_reward ← episodic_reward + reward
21: **end for**
22: ppo_agent.Update(memory_buffer)
23: memory_buffer.Clear()
24: if episode % 200 == 0 then
25: ppo_agent.LearningRate ← ppo_agent.LearningRate × 0.95
26: **end for**

Table 1. PPO hyperparameter configuration.

Parameter	Value
Network Architecture	2 hidden layers, 256 neurons each, ReLU
Optimizer	Adam
Learning Rate (α)	3.0×10^{-4} (with 5% decay every 200 epochs)
Discount Factor (γ)	0.98
GAE Lambda (λ)	0.95
Clipping Epsilon (ε)	0.1
Entropy Coefficient	0.01
Batch Size	128
Max Episodes	1000
Random Seeds	3

The key hyperparameters used for the PPO agent implementation are summarized in Table 1.

The simulation scenario is shown in Figure 7. During the exploration process, the overall reward trend of the control strategy in the training episodes is depicted in Figure 8. The reward function gradually increases and stabilizes, indicating that the algorithm exhibits good convergence.

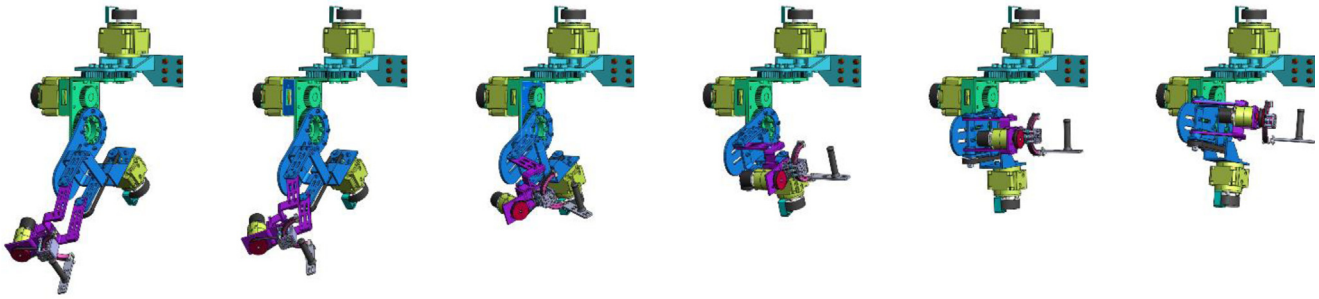

Fig. 7. Upper limb rehabilitation robot simulation scene diagram.

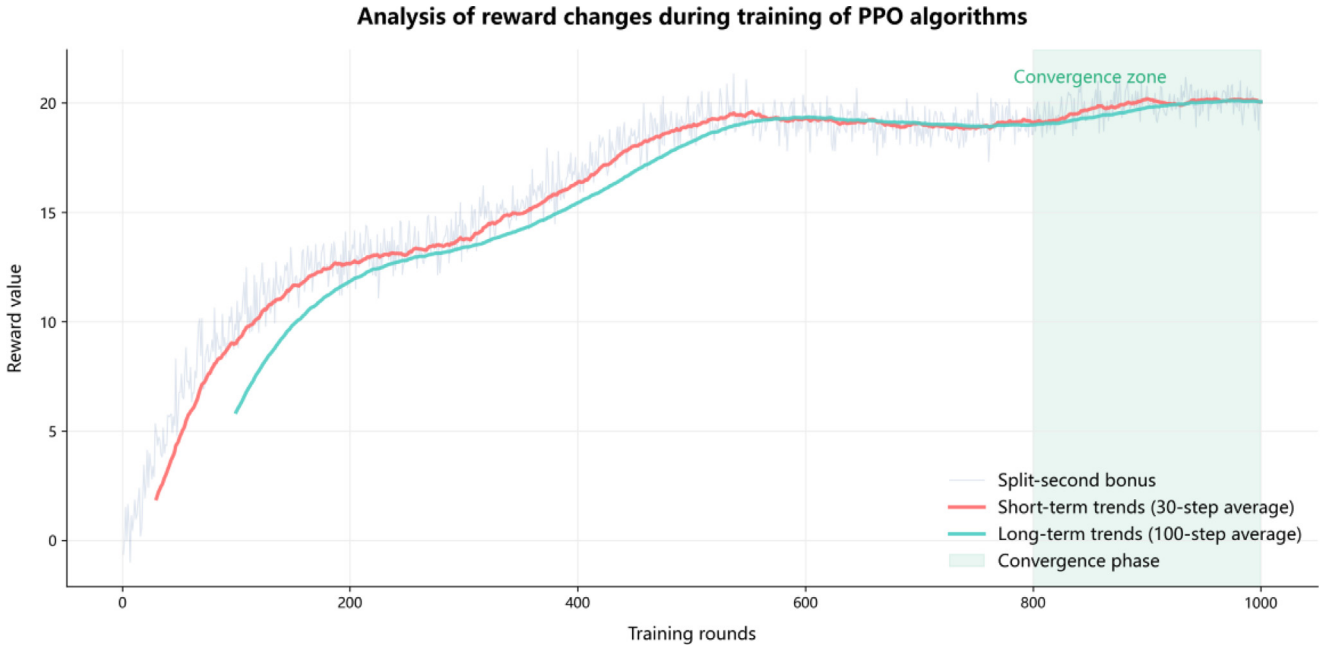

Fig. 8. Reward change analysis during the training process.

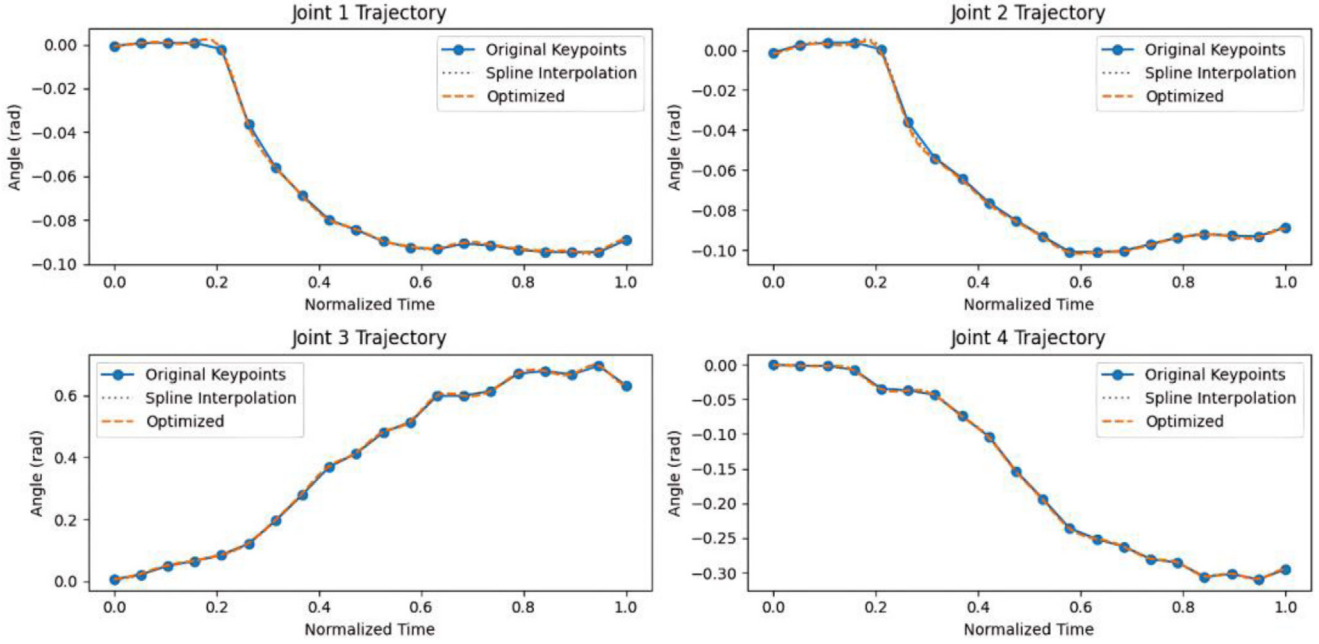

Fig. 9. Joint angles' curves after training.

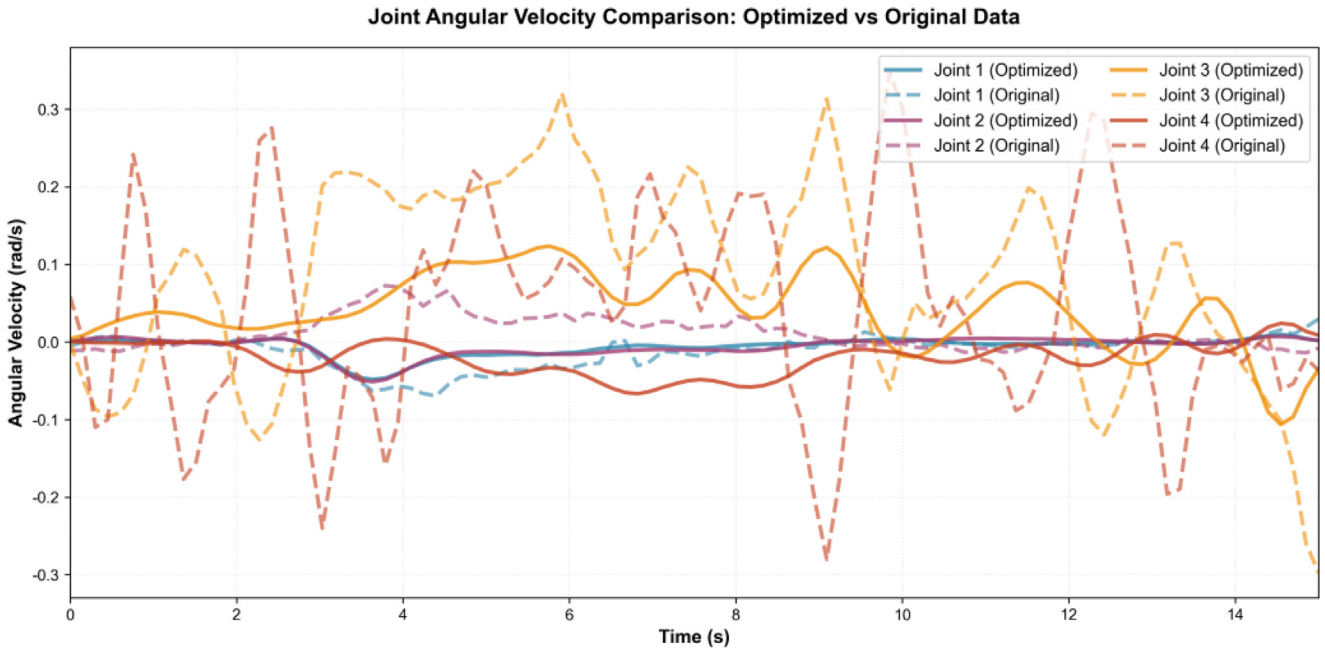

Fig. 10. The comparison of velocity between the optimal trajectory and the original trajectory.

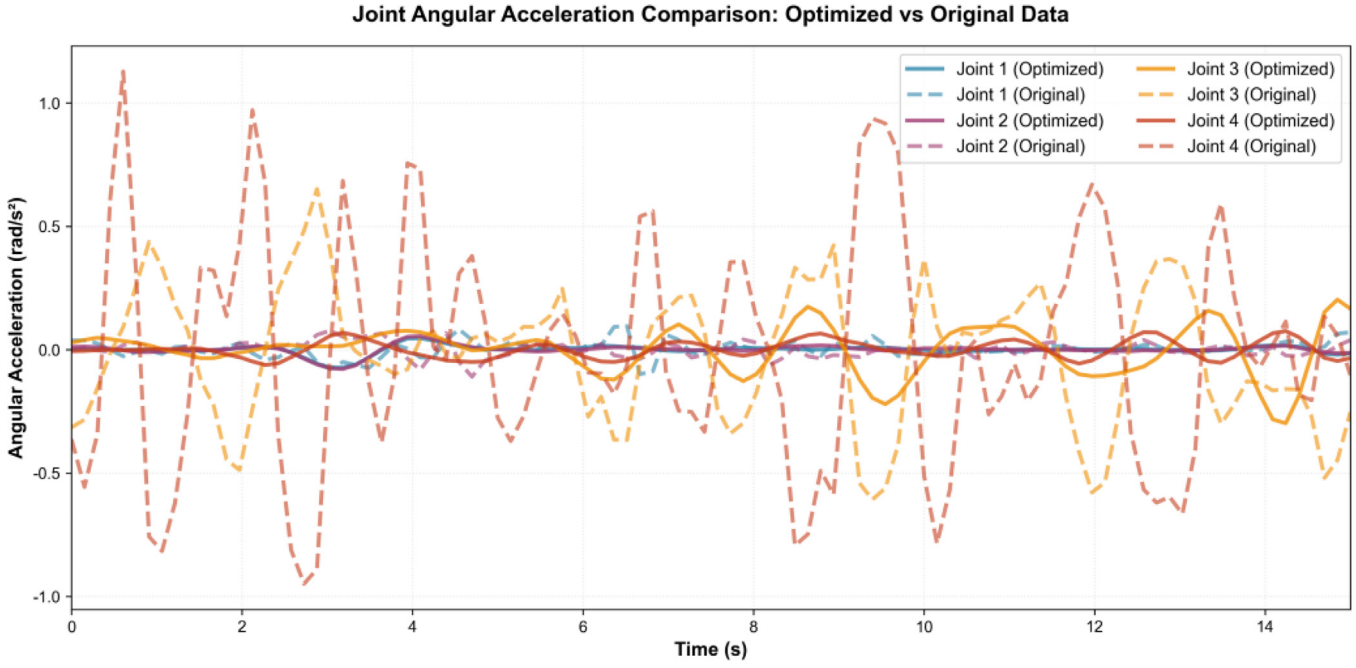

Fig. 11. The comparison of acceleration between the optimal trajectory and the original trajectory.

The trajectory after multi-objective constrained optimization, the initial must-pass point position, and the four joint angle changes of the trajectory planning of the spline interpolation after the completion of the drinking training are shown in Figure 9. The comparison of velocity and acceleration between the optimal trajectory and the original trajectory is shown in Figures 10 and 11.

As shown in the velocity and acceleration comparison plots of the optimal and original trajectories in Figures 10 and 11, the average angular velocity and average angular acceleration planned for the rehabilitation robot are significantly reduced through trajectory optimization. This method enhances the compliance and safety during the rehabilitation process, providing patients with a more comfortable and safer training environment.

4 Experimentation and analysis of trajectory planning based on robot prototypes

The platform of the upper extremity exoskeleton rehabilitation robot system mainly contains a four-degree-of-freedom

Fig. 12. Upper limb exoskeleton rehabilitation robot system platform.

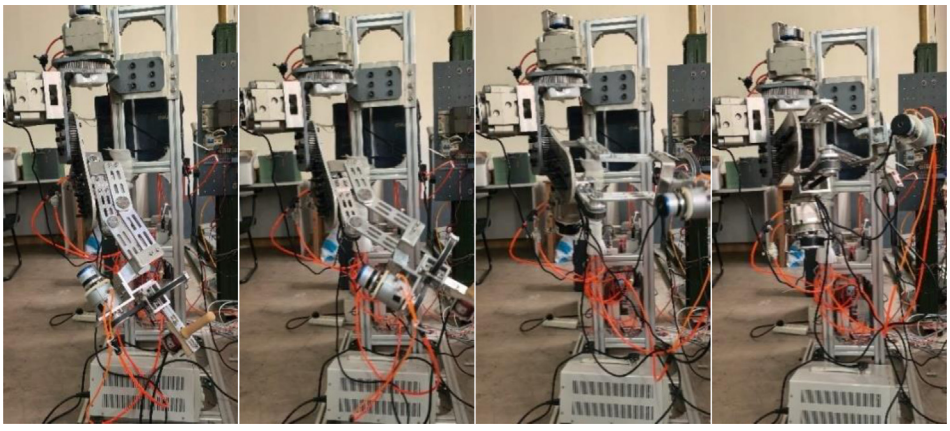

Fig. 13. The movement of the experimental prototype.

experimental prototype of the upper extremity exoskeleton rehabilitation robot, a LINKS semi-physical simulator (Beijing LinksTech co., Ltd), a NOKOV optical motion capture system Mars2H (Beijing Nokov Science & Technology Co., Ltd.), and a computer platform, as shown in Figure 12.

To validate the good compliance and impact characteristics of the personalized rehabilitation trajectory based on the multi-objective optimization strategy, this study designs the following verification scheme based on the upper-limb exoskeleton rehabilitation robot system platform. The trained rehabilitation trajectory is deployed onto the rehabilitation robot experimental prototype. In the experimental setup, the subject's upper limb follows the motion of the rehabilitation robot, and the shoulder joint angles $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, elbow joint angle $\boldsymbol{\theta}_3$, and wrist joint angle $\boldsymbol{\theta}_4$, collected by the optoelectronic encoders of the experimental prototype, are recorded and analyzed. Figure 13 shows the motion of the optimal trajectory deployed to the experimental prototype, while the joint angle errors of the upper-limb rehabilitation robot during the motion are shown in Figure 14. To further demonstrate that the trajectory trained by the PPO learning algorithm exhibits better smoothness and lower impact, a fifth-order polynomial interpolation curve is deployed in the upper-limb rehabilitation robot for a comparative experiment. The joint angle errors are shown in Figure 15.

The standard deviation of angle, standard deviation of angular velocity, standard deviation of angular acceleration, standard deviation of angular plus acceleration for the four joints of the analyzed optimal trajectory and the fifth-degree polynomial interpolated trajectory are shown in Tables 2 and 3.

As can be seen from Tables 1 and 2, the trajectory after multi-objective constrained optimization trained by the learning algorithm has an average reduction of 70.65% in the standard deviation of angular plus acceleration for all joints compared to the five-polynomial interpolation, with lower impact. The angular standard deviation of the optimal trajectory is reduced by 58.1% on average compared to the five polynomial interpolated trajectory, so the trajectory tracking stability is better, and the angular velocity standard deviation is reduced by 70.35% on average, so the motion continuity is better. The multi-objective optimal trajectory trained by the learning algorithm has better trajectory tracking accuracy, motion smoothness and impact degree than the quintuple polynomial interpolation trajectory, and can accomplish the rehabilitation training task more safely and effectively.

To validate the contribution of each component in our multi-objective reward function and to provide a rationale for the weight selection, we conducted a concise ablation study. The results, summarized in Table 4, demonstrate a clear performance improvement as key components are

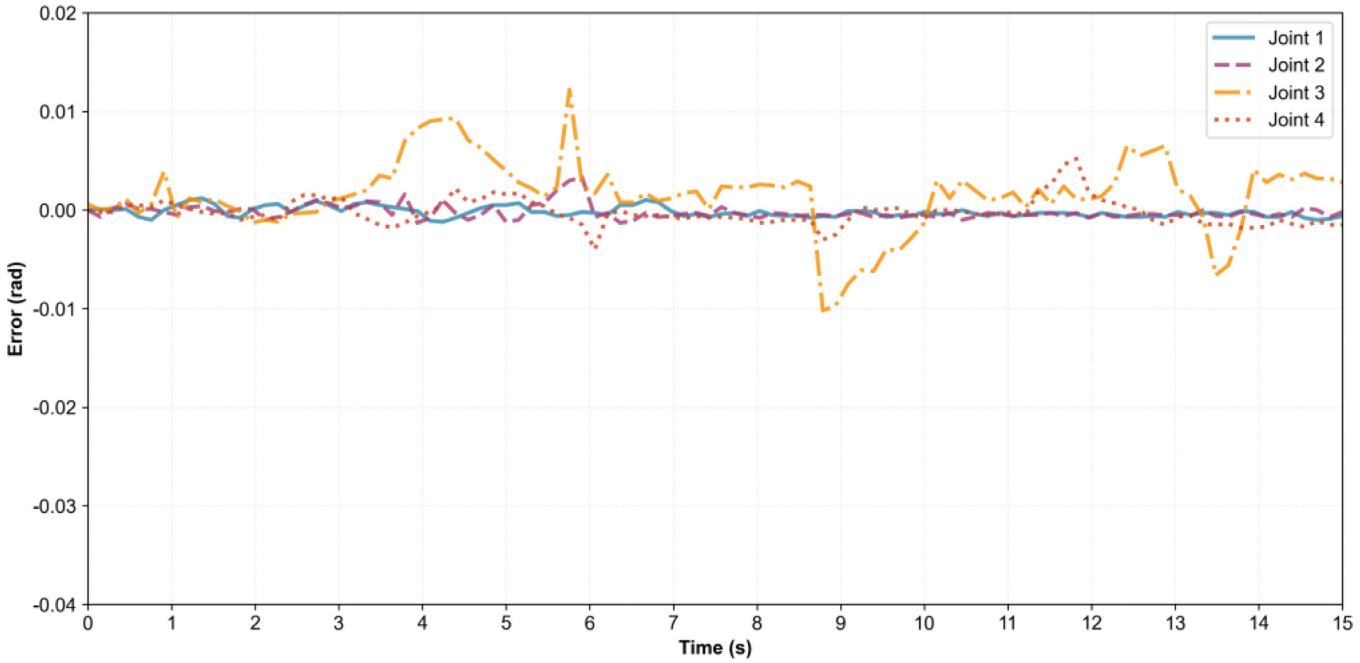

Fig. 14. The motion error of the optimal trajectory deployed to the experimental prototype.

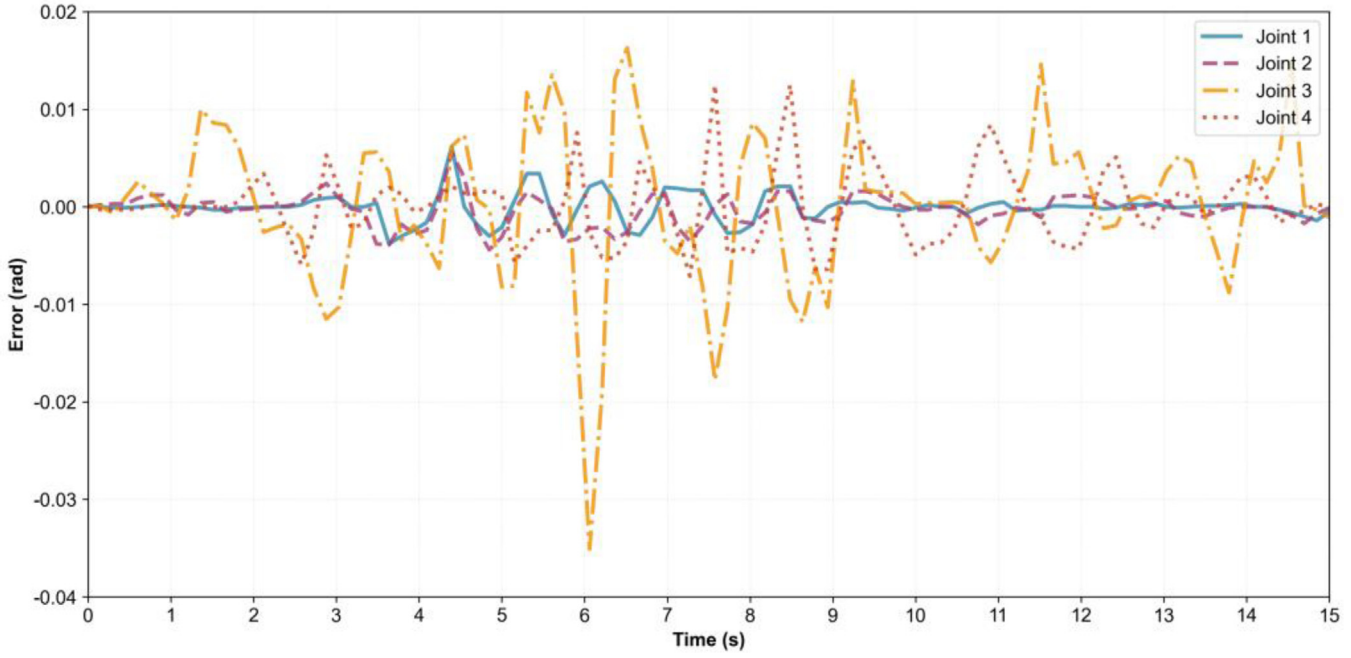

Fig. 15. The motion error of the fifth-order polynomial interpolation trajectory deployed to the experimental prototype.

Table 2. Table of evaluation indexes of kinematic properties of optimal trajectories.

Item	Angular standard deviation(rad)	Angular velocity standard deviation(rad/s)	Angular acceleration standard deviation(rad/s^2)	Angular Jerk standard deviation(rad/s^3)
Joint_1	0.03	0.11	0.61	3.47
Joint_2	0.04	0.18	1.00	5.75
Joint_3	0.21	0.63	3.16	18.23
Joint_4	0.08	0.24	1.17	6.65

Table 3. Table of evaluation metrics for kinematic properties of fifth-degree polynomial interpolated trajectories.

Item	Angular standard deviation(rad)	Angular velocity standard deviation(rad/s)	Angular acceleration standard deviation(rad/s^2)	Angular Jerk standard deviation(rad/s^3)
Joint_1	0.08	0.43	2.38	13.51
Joint_2	0.09	0.43	2.42	14.37
Joint_3	0.44	2.15	11.71	67.42
Joint_4	0.21	1.10	6.25	36.70

Table 4. Ablation study of reward function components and final weight configuration.

Stage / Reward component added	Final Avg. reward	SD of Jerk (rad/s^3) (Joint 3 avg.)[1]	Tracking error (rad) (Avg. RMSE)
Baseline (Tracking only)	15.2	45.8	0.008
+ Smoothness & Jerk	18.5	19.5	0.012
+ Velocity Barrier	19.1	18.9	0.011
Full Model (+ Keypoints, etc.)	19.8	18.2	0.007

progressively added. The study confirms that the inclusion of smoothness and jerk terms is critical for reducing kinematic volatility. The weights for the full model were determined through a grid search. We prioritized three main categories: (1) Core Task Weights for tracking and keypoints (w_track, w_hard), (2) Kinematic Quality Weights for smoothness and jerk (w_smooth, w_jerk), and (3) Safety & State Weights for velocity limits and manipulability (w_vel, w_manip). The final weights, detailed below Table 4, reflect a primary focus on hard constraints and task accuracy, followed by kinematic quality.

5 Conclusion

In this paper, a trajectory planning and optimization method based on reinforcement learning algorithm is proposed for the upper limb rehabilitation robot. Typical ADL rehabilitation actions are captured by an optical motion capture system, and four-joint angular trajectory datasets of the shoulder, elbow, and forearm are obtained by rigid-body alignment. A simulation platform is built, and a multi-objective reward function containing personalized trajectory tracking accuracy, motion smoothness and safety constraints is constructed for the problem of sudden change in impact degree of traditional interpolated trajectory planning algorithms, and the trajectory after multi-objective constrained optimization is steadily learned by using a dynamic trust domain mechanism to optimize the strategy search process. The personalized rehabilitation training based on the ADL movement of the subject's own healthy limb is realized by deploying it into the robot prototype. The simulation and prototype experiments show that the proposed trajectory planning and optimization method can effectively achieve the task-oriented rehabilitation goal by ensuring the personalized and human-like rehabilitation needs of patients while combining the features of smooth robot motion and low impact.

Funding

This study was supported by the "Research on Key Technologies for Embodied Intelligent Collaborative Control of Upper Limb Exoskeleton Robots" project granted by the "Project of science and technology of the Henan Province" (242102220116).

Conflicts of interest

The authors declare that there is no competing financial interest or personal relationship that could have appeared to influence the work reported in this paper.

Data availability statement

Article data can be obtained from the corresponding author upon reasonable request.

Author contribution statement

Conceptualization, Bingjing Guo; methodology, Haotian Xu and Zhenzhu Li; writing-original draft preparation, Bingjing Guo, Haotian Xu and Zhenzhu Li; writing-review and editing, Bingjing Guo, Haotian Xu, Zhenzhu Li, Xiangpan Li, Jianhai Han; project administration, Bingjing Guo.; funding acquisition, Bingjing Guo. All authors have read and agreed to the published version of the manuscript.

References

1. H. Cheng, R. Huang, J. Qiu et al., Rehabilitation robots and their clinical applications: a review, Robot, **43**, 606–619 (2021)
2. T. Mao, Y. Li, M. Li et al., Survey on deep learning-based lesion segmentation and detection in acute ischemic stroke, Comput. Syst. Appl. **34**, 11–25 (2025)

3. S. Luo, L. Meng, H. Yu, An overview of the research and application of rehabilitation robot technology in China, China Acad. J. Electron. Publ. House **38**, 1762–1768 (2023)
4. C. He, C. Xiong, W. Chen. Brain injury upper limb rehabilitation robot and its clinical application research, J. Mech. Eng. **59**, 65–80 (2023)
5. L. Gao, D. Zhang, X. Zhao, Upper limb rehabilitation robot direct teaching technology adapted to individual patient differences, Chin. J. Rehabil. Theory Pract. **28**, 1231–1240 (2022)
6. L. Li, R. Zhang, G. Cheng et al., Trajectory tracking control of upper limb rehabilitation robot based on optimal discrete sliding mode control, Meas. Control **56**, 1142–1155 (2023)
7. S. Guo, Y. Song, X. Wang et al., Learning and transfer method of active rehabilitation strategy for upper limb rehabilitation robot, Robot **46**, 562–575 (2024)
8. J. Zhang, X. Fu, Y. Wang et al., Task-oriented training based on activities of daily living analysis for stroke patients, Chin. J. Phys. Med. Rehabil. **44**, 595–599 (2022)
9. Y. Yao, S. Pei, J. Guo et al., Upper limb rehabilitation robot research review, J. Mech. Eng. **60**, 115–134 (2024)
10. T. Terama, T. Matsubara, T. Noda et al., Quaternion-based trajectory optimization of human postures for inducing target muscle activation patterns, IEEE Robot. Autom. Lett. **5**, 6607–6614 (2020)
11. D. Xu, Z. Wang. Trajectory planning and optimization research of upper limb rehabilitation robot based on 5th quasi-uniform B-spline, Indust. Control Comput. **36**, 59–60 +63 (2023)
12. N.B. Mohamadwasel, S. Kurnaz, Implementation of the parallel robot using FOPID with fuzzy type-2 in use social spider optimization algorithm, Appl. Nanosci. **13**, (2023). https://doi.org/10.1007/s13204-021-02034-9
13. N. Basil, A.F. Mohammed, B.M. Sabbar et al., Performance analysis of hybrid optimization approach for UAV path planning control using FOPID-TID controller and HAOAROA algorithm, Sci. Rep. **4840**, (2025). https://doi.org/10.1038/s41598-025-86803-4
14. N. Basil, H.M. Marhoon, D.F. Sahib et al., Accelerated black hole optimization algorithm with enhanced FOPID controller for omni-wheel drive mobile robot system, Neural Comput. Appl. **37**, 16983–17014 (2025)
15. G. Li, Q. Fang, T. Xu et al., Inverse kinematic analysis and trajectory planning of a modular upper limb rehabilitation exoskeleton, Technol. Health Care **27**, 123–132 (2019)
16. A. Lei, Y. Chen, Y. Xu, Human-like motion planning method for robot arm based on reinforcement learning, Chin. J. Sci. Instrum. **42**, 136–145 (2021)
17. Y. Wei, W. Jiang, A. Rahmani et al., Motion planning for a humanoid mobile manipulator system, Int. J. Hum. Robot. **16**, 1950006 (2019)
18. N.F. Durate, M. Raković, J. Santos-Victor, Biologically inspired controller of human action behaviour for a humanoid robot in a dyadic scenario, in: *IEEE EUROCON 2019-18th International Conference on Smart Technologies*, 2019, pp. 1–6
19. Y. Zhu, X. Tong, R. Yang et al., A survey on modeling mechanism and control strategy of rehabilitation robots: recent trends, current challenges, and future developments, Int. J. Control Autom. Syst. **20**, 2724–2748 (2022)
20. H. Zhang, T. Zhang, Research progress in stroke rehabilitation techniques, Chin. J. Rehabil. **37**, 371–375 (2022)
21. N. Mangalabarathi, B. Devi, K Chinnathambi et al., Effectiveness of nurse-led stroke rehabilitation on awareness, activities of daily living and coping in stroke patients at a tertiary care hospital in India, Cureus, e72843 (2024). https://doi.org/10.7759/cureus.72843
22. R. Dizor, A. Raj, B. Gonzalez et al., Deep reinforcement learning to assess lower extremity movement intention and assist a rehabilitation exoskeleton, in: *Proc. SPIE 13058, Disruptive Technologies in Information Sciences VIII*, 1305805 (6 June 2024). https://doi.org/10.1117/12.3013039
23. Y. Liu, Z. Chen, Y. Li et al., Robot search path planning method based on prioritized deep reinforcement learning, Control Autom. Syst. **20**, 2669–2680 (2022)

Cite this article as: Haotian Xu, Bingjing Guo, Jianhai Han, Xiangpan Li, Zhenzhu Li, Rehabilitation robot trajectory planning method for upper limb based on healthy limb motion using multi-objective constrained reinforcement learning, Int. J. Simul. Multidisci. Des. Optim. **17**, 1 (2026), https://doi.org/10.1051/smdo/2025034