actuators

MDPI

Article

Gated Reward Weighting for Single-Policy Reinforcement Learning Control of a Valve-Controlled Hydraulic Inverted Pendulum

Penghui Li [1], Bobo Helian [1,*] and Litong Lyu [2]

[1] Institute of Mobile Machines, Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany; penghui.li@student.kit.edu
[2] School of Mechanical Engineering, Shijiazhuang Tiedao University, Shijiazhuang 050043, China; litong_lyu@stdu.edu.cn
* Correspondence: bobo.helian@kit.edu

Abstract

Proportional valve-controlled cylinders are the most widely used and most representative electro-hydraulic drives. This paper considers such a drive moving an inverted pendulum on a cart: one controller must swing the pendulum up, stabilize it, and track a cart reference, while the valve flow is nonlinear and the three tasks compete for the same valve. A single reinforcement learning (RL) policy based on deep deterministic policy gradient (DDPG) performs all three tasks. During training, a gated, state-dependent reward weighting shifts the emphasis from pendulum stabilization to cart tracking once the pendulum has been captured. In simulation, the controller is compared with a staged energy-based swing-up and linear quadratic regulator (LQR) baseline and with an RL controller trained with fixed reward weights. The proposed controller swings the pendulum up in 2.0 s instead of 4.7 s and completes a 0.5 m/s reference at which both comparison controllers reach the end stop. It also tracks more accurately than the fixed-weight controller, and without the gate the training does not balance the pendulum at all. The costs are a constant tracking offset of 10 mm, more valve motion, and a larger pendulum excursion than the baseline under an external force.

Keywords: valve-controlled hydraulic system; inverted pendulum; deep deterministic policy gradient; state-dependent reward weighting; multi-objective control; trajectory tracking

1. Introduction

Hydraulic actuators offer high power density and large output forces and are widely used in construction and agricultural machinery and in industrial presses. Their most widely used and most representative form is valve control, in which a proportional directional valve meters the flow from a pump supply into a cylinder. This paper studiesan inverted pendulum driven by such a valve-controlled actuator: the valve feeds an asymmetric cylinder from an 18 MPa supply, and the cylinder moves the cart carrying the pendulum over a travel of ± 0.3 m. One controller, updated at 50 Hz, has to swing the pendulum up from the hanging position, hold it upright, and move the cart along point-to-point references at up to 0.5 m/s. Three properties make this difficult. The cylinder force depends on the valve flow, the chamber pressures, and the load motion, not on the valve command alone. The pendulum is open-loop unstable and underactuated, and the cart position is non-minimum phase with respect to the actuator force. Finally, the three tasks compete for the same valve, because the cart motion that balances the pendulum also produces tracking error.

Check for updates

Academic Editor: Guanghong Yang

Received: 23 October 2026
Revised: 12 November 2026
Accepted: 15 November 2026
Published: 18 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Valve-controlled cylinders are the standard drive of mobile machinery, where the conventional directional valve is being extended to independent metering of the inlet and outlet flows [1]. Their motion control is complicated by the nonlinear flow–pressure characteristics of proportional valves, fluid compressibility, friction, leakage, and parameter uncertainties [2–4], and the non-minimum-phase cart dynamics of the inverted pendulum constrain position tracking [5]. For the pendulum, model-based designs treat the tasks in stages: an energy-based law swings the pendulum up [6], a linear quadratic regulator (LQR) stabilizes it locally [7], and a switching or blending rule joins the two laws [8]. For hydraulic actuators, adaptive, robust, and other nonlinear controllers achieve accurate tracking under modeling uncertainties and disturbances [9–11], also for the synchronized motion of several cylinders [12]. Reinforcement learning (RL) instead learns a mapping from states to actions that maximizes the cumulative reward [13], and deep deterministic policy gradient (DDPG) is an actor–critic algorithm for continuous actions [14]. Deep RL has been applied to inverted-pendulum swing-up and stabilization [15–17], in one case with separate swing-up and stabilization policies joined by switching logic [18]. It has also been applied to electro-hydraulic servo systems [19]: for direct position or speed control [20–24], for tuning or augmenting model-based controllers [25–27], for position tracking of valve-controlled actuators [28,29], and for the autonomous operation of a hydraulic excavator through its proportional valves [30]. When several objectives are combined, fixed reward weights fix the trade-off at training time [31,32], and reward curricula change the reward over the course of training [33]; neither changes the priority between the objectives within an episode. The closest precedent is the authors' earlier work on an inverted-pendulum testbench driven by a pump-controlled electro-hydrostatic actuator [34], in which a DDPG agent trained with reward shaping achieved swing-up and balancing in experiments and was compared with an LQR. Cart trajectory tracking was not part of that task. Adding it creates a conflict of priorities: pendulum regulation must dominate during swing-up and capture, and tracking must dominate once the pendulum is upright. It remains open whether one RL policy can cover all three tasks of a hydraulically driven pendulum, and how such a policy compares with a staged model-based design on the same plant, including the actuator-level cost.

This paper addresses this problem with a single RL policy for the proportional valve-controlled cylinder, the most common configuration in practice. The work proceeds in three steps. First, a Simscape model of the system is built that contains the nonlinear valve flow, the valve dynamics and limits, and the multibody cart–pendulum dynamics (Section 2). Second, one DDPG policy is trained for swing-up, stabilization, and tracking, with the tasks coordinated through the reward (Section 3). The pendulum weight is high when the pendulum deviates from the upright position and low near it, where the tracking weight takes over. A gate releases this angle-dependent weight only after the pendulum has stayed within angle and angular-velocity bounds for a hold interval, so that the capture of the pendulum is rewarded before tracking; first-order smoothing limits the rate of change in the weight. As the weighting acts during training only, the deployed controller is a single network without switching between control laws. Third, the policy is compared with a staged model-based baseline that also tracks, with an RL controller trained with fixed weights, and with a policy retrained without the gate, and the valve and force signals are reported beside the task performance (Section 4). The novelty lies in the gated, state-dependent reward weighting, which makes the three tasks trainable in one policy. Compared with the authors' earlier pump-controlled study, the present work moves to the valve-controlled drive, adds cart trajectory tracking as a third objective, and reports the actuator-level cost of the learned policy. Sections 5 and 6 present the discussion and conclusions.

2. System Description and Modeling

As shown in Figure 1, a proportional 4/3 directional valve, supplied by a pump and a pressure-relief valve, drives a double-acting asymmetric cylinder that moves a cart carrying the pendulum.

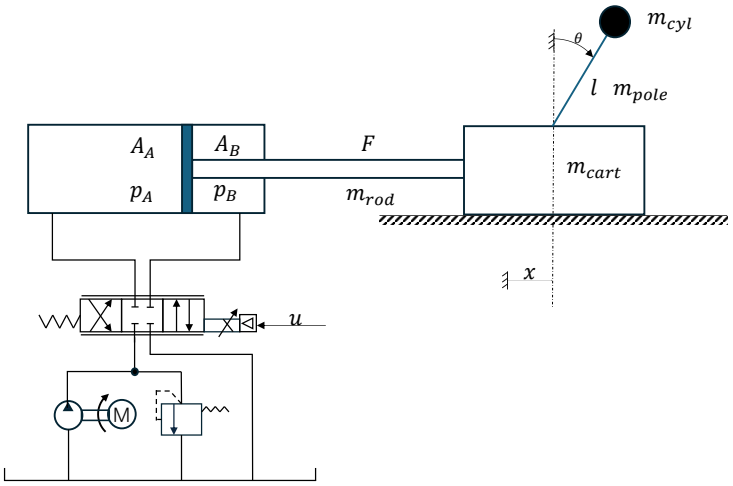


Figure 1. Configuration of the valve-controlled electro-hydraulic inverted pendulum system.

2.1. Mechanical System

The moving mass $M = m_{\text{cart}} + m_{\text{rod}}$ comprises the cart and the piston rod. The pendulum consists of a pole of length l and mass m_{pole} and a cylindrical tip mass m_{cyl}; $m = m_{\text{pole}} + m_{\text{cyl}}$ is its total mass, l_c the distance from the pivot to its center of mass, and I its moment of inertia about the center of mass. The cart displacement x and the hydraulic force F are positive to the right, and the pendulum angle θ is measured clockwise from the upright position, so that $\theta > 0$ tilts the pendulum toward positive x. With friction and joint damping neglected, Lagrange's equations give [6,7]

$$\begin{aligned} (I + ml_c^2)\ddot{\theta} + ml_c \cos\theta\, \ddot{x} - mgl_c \sin\theta = 0, \\ (M + m)\ddot{x} + ml_c \cos\theta\, \ddot{\theta} - ml_c \sin\theta\, \dot{\theta}^2 = F, \end{aligned} \tag{1}$$

where g is the gravitational acceleration.

2.2. Hydraulic System

The hydraulic force on the cart is [4]

$$F = A_A p_A - A_B p_B, \tag{2}$$

where A_A and A_B are the piston areas and p_A and p_B the chamber gauge pressures. The flow through each metering orifice of the valve is [2]

$$Q_o = C_d A_o(x_v)\, \mathrm{sgn}(\Delta p_o) \sqrt{\frac{2|\Delta p_o|}{\rho}}, \tag{3}$$

where x_v is the spool displacement, C_d the discharge coefficient, ρ the fluid density, and A_o and Δp_o the orifice area and pressure drop. The spool displacement follows the valve command u through the spool dynamics of Table 1. For $x_v > 0$, the valve connects the supply port P to chamber A and chamber B to the tank port T; for $x_v < 0$, it connects P to B and A to T. The opening area increases linearly from the leakage area A_{leak} at the zero position to $A_{\max}$ at the spool displacement limit $x_{v,\max}$. The net inflows into the chambers

are $Q_A = Q_{PA} - Q_{AT}$ and $Q_B = Q_{PB} - Q_{BT}$. With the leakage across the piston neglected, mass conservation in the two chambers gives the pressure dynamics [4,9]

$$\dot{p}_A = \frac{\beta_e}{V_A(x)}(Q_A - A_A\dot{x}), \qquad \dot{p}_B = \frac{\beta_e}{V_B(x)}(Q_B + A_B\dot{x}), \tag{4}$$

where β_e is the effective bulk modulus and $V_A(x) = V_{10} + A_A x$ and $V_B(x) = V_{20} - A_B x$ are the chamber volumes, including the dead volumes, with V_{10} and V_{20} at mid-stroke ($x = 0$).

Table 1. Parameters and modeling settings of the Simscape model.

Parameter	Value
Mechanical system	
Moving mass M (cart and piston rod)	14.70 kg
Pole mass m_{pole}, length l	0.113 kg, 0.20 m
Tip mass m_{cyl}, distance from pivot	0.261 kg, 0.20 m
Center of mass l_c, inertia I	0.170 m, 1.17×10^{-3} kg m^2
Cart end stops	±0.30 m; 10^6 N/m, 10^3 N s/m
Hydraulic supply and fluid	
Fixed-displacement pump	60 cm^3/rev, 1800 rpm; efficiencies 0.92 (volumetric), 0.88 (mechanical)
Relief-valve setting p_s; tank pressure	18.1 MPa; 0.1 MPa
Fluid	Skydrol LD-4 at 20 °C, 0.5% entrained air
Proportional 4/3 valve	
Spool displacement limit $x_{v,\max}$	6 mm
Orifice area $A_{\max}$; leakage area A_{leak}	12.0×10^{-6} m^2; 3.6×10^{-8} m^2
Discharge coefficient C_d; critical Reynolds number	0.64; 150
Flow at full opening and 9 MPa per orifice	62 L/min
Spool dynamics	second order, 70 Hz, damping ratio 0.82; rate limit 0.25 m/s
Cylinder	
Piston areas A_A, A_B; stroke	9.62×10^{-4}, 4.71×10^{-4} m^2; 0.60 m
Modeling settings	
Valve flow	Equation (3), laminar near zero pressure drop; pressure recovery
Chamber compressibility, cylinder friction	not included
Cart and pendulum-joint damping	not included
Piston leakage	clearance 2×10^{-11} m (negligible)

For the LQR design of the baseline (Section 3.2), Equation (4) is linearized at mid-stroke and at the zero spool position ($x = \dot{x} = 0$, $x_v = 0$), and the spool dynamics are neglected ($x_v = u$). To first order, the chamber volumes are then constant at V_{10} and V_{20}, and the net inflows become $Q_A \approx K_{q1}u$ and $Q_B \approx K_{q2}u$ with the flow gains

$$K_{q1} = -K_{q2} = C_d k_A \sqrt{\frac{2\Delta p_n}{\rho_n}}, \qquad k_A = \frac{A_{\max} - A_{\text{leak}}}{x_{v,\max}}, \tag{5}$$

obtained from Equation (3) with a nominal density ρ_n and a nominal pressure drop Δp_n across each open orifice, taken as half the supply-to-return pressure difference. The flow–pressure coefficients of the linearized orifice equation [4] are neglected; they are proportional to the orifice area and therefore small near the zero position. Differentiating Equation (2) and substituting the linearized Equation (4) gives the force dynamics used in this work,

$$\begin{aligned} \dot{F} &= b_{Fu}u + b_{Fx}\dot{x}, \\ b_{Fu} &= \beta_e\left(\frac{A_A K_{q1}}{V_{10}} - \frac{A_B K_{q2}}{V_{20}}\right), \qquad b_{Fx} = -\beta_e\left(\frac{A_A^2}{V_{10}} + \frac{A_B^2}{V_{20}}\right). \end{aligned} \tag{6}$$

Linearizing Equation (1) around the upright position ($\sin\theta \approx \theta$, $\cos\theta \approx 1$, and $\dot{\theta}^2 \sin\theta \approx 0$) and combining it with Equation (6) gives the design model with the state $\mathbf{z} = (\theta, \dot{\theta}, x, \dot{x}, F)^{\mathrm{T}}$,

$$\dot{\mathbf{z}} = \mathbf{A}\mathbf{z} + \mathbf{B}u, \qquad \mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ \frac{(M+m)mgl_c}{d} & 0 & 0 & 0 & -\frac{ml_c}{d} \\ 0 & 0 & 0 & 1 & 0 \\ -\frac{m^2 g l_c^2}{d} & 0 & 0 & 0 & \frac{I+ml_c^2}{d} \\ 0 & 0 & 0 & b_{Fx} & 0 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ b_{Fu} \end{bmatrix}, \tag{7}$$

where $d = (I + ml_c^2)(M + m) - (ml_c)^2$. The reduced model retains an effective hydraulic compliance through β_e, so that the force F is a state that can be fed back. It is used only to compute the LQR gains; its parameters are listed in Section 3.2. The baseline results on the nonlinear simulation model (Section 4.2) show that it is adequate for this purpose.

2.3. Simulation Model and Parameters

All simulations use a Simscape model in MATLAB/Simulink R2024b (MathWorks, Natick, MA, USA) (Figure 2). The hydraulic system, consisting of the pump, the relief valve, the 4/3 valve, and the cylinder, is modeled in Simscape Fluids. The cart and the pendulum are modeled in Simscape Multibody, with a prismatic joint for the cart and a revolute joint for the pendulum. The two subsystems are coupled through a translational multibody interface at the piston rod: the cylinder force acts on the cart, and the cart motion drives the piston. Table 1 lists the parameters and modeling settings. The pump and the relief valve hold the supply pressure at 18.1 MPa. The valve has identical linear orifices following Equation (3) and a leakage area at the zero position. Its spool displacement x_v follows the command u through second-order dynamics with a rate limit and is limited to ± 6 mm for every controller. The cylinder chambers are modeled as incompressible and frictionless, so the piston velocity is set by the valve flow and the cylinder acts as a velocity source on the cart. Compliant end stops limit the cart travel to ± 0.30 m. The model therefore excludes the hydraulic resonance and the friction of a real actuator; Section 5 discusses the consequences. The same model is used to train the RL agent and to evaluate all controllers.

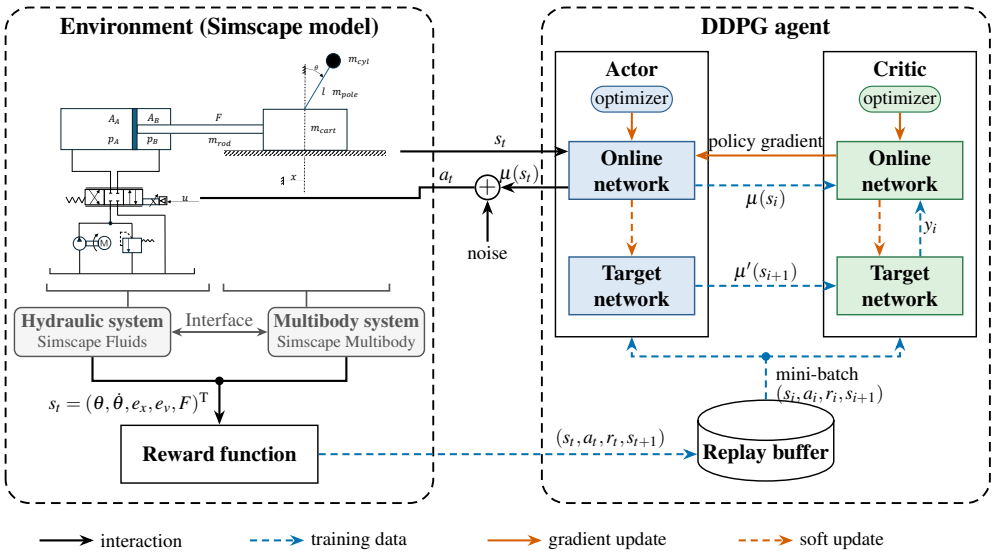


Figure 2. DDPG-based control framework with the Simscape environment.

3. Control Design

3.1. Reinforcement Learning Control Using DDPG

3.1.1. Control Framework

Figure 2 shows the control framework. The actor represents the policy $\mu(s)$, and the critic estimates the action-value function $Q(s,a)$ [14]. The policy is trained in the Simscape model of Section 2.3. At step t, the observation is

$$s_t = (\theta, \dot{\theta}, e_x, e_v, F)^{\mathrm{T}}, \tag{8}$$

where $e_x = x - x_{\mathrm{ref}}$ and $e_v = \dot{x} - \dot{x}_{\mathrm{ref}}$ are the cart position and velocity errors, and F is the hydraulic force calculated from the measured chamber pressures. During training, exploration noise n_t is added to the actor output, $a_t = \mu(s_t) + n_t$; after training, $a_t = \mu(s_t)$. The action $a_t \in [-1, 1]$ is scaled to the spool command $u_t = 10\ \mathrm{mm} \cdot a_t$, which is limited to ± 6 mm.

3.1.2. Reward Function Design

The reward function, Equations (9) to (15), is the design of this work. It consists of a goal reward, a dense reward, and a cart-position penalty,

$$r = r_{\mathrm{goal}} + r_{\mathrm{dense}} + r_{\mathrm{penalty}}. \tag{9}$$

The goal reward is

$$r_{\mathrm{goal}} = \begin{cases} \dfrac{C_g}{\sqrt{0.1 + |e_x|}}, & s \in \mathcal{G}, \\ 0, & \text{otherwise}, \end{cases} \tag{10}$$

with e_x in meters, where the goal set $\mathcal{G}$ requires $|\theta| \le \theta_g$, $|\dot{\theta}| \le \omega_g$, $|e_x| \le e_{x,g}$, and $|e_v| \le e_{v,g}$. The dense reward weights a cart penalty against a pendulum penalty,

$$\begin{aligned} r_{\mathrm{dense}} &= -3[(1-w)r_{\mathrm{cart}} + w r_{\mathrm{pend}}], \\ r_{\mathrm{cart}} &= \alpha_x \Big(\frac{e_x}{e_{x,s}}\Big)^2 + (1-\alpha_x)\Big(\frac{e_v}{e_{v,s}}\Big)^2, \qquad r_{\mathrm{pend}} = \alpha_\theta \Big(\frac{\theta}{\theta_s}\Big)^2 + (1-\alpha_\theta)\Big(\frac{\dot{\theta}}{\omega_s}\Big)^2, \end{aligned} \tag{11}$$

where the denominators are normalization scales and $\alpha_x, \alpha_\theta \in [0, 1]$. The fixed-weight RL controller uses $w = 0.7$. The cart-position penalty is $r_{\mathrm{penalty}} = -C_p$ for $|x| > x_p$ and zero otherwise, and an episode terminates when $|x| > x_{\mathrm{term}}$. Table 2 lists the values.

Table 2. Reward parameters.

Parameter	Symbol	Value
Goal-reward coefficient	C_g	50
Goal angle and angular-velocity bounds	θ_g, ω_g	5°, 10°/s
Goal tracking-error bounds	$e_{x,g}, e_{v,g}$	0.05 m, 0.10 m/s
Cart-error normalization scales	$e_{x,s}, e_{v,s}$	0.30 m, 1.00 m/s
Pendulum normalization scales	θ_s, ω_s	180°, 3500°/s
Internal penalty weights	α_x, α_θ	0.7, 0.6
Position-penalty magnitude	C_p	100
Penalty and termination thresholds	x_p, x_{term}	0.298 m, 0.30 m

A fixed weight cannot follow the change in priority from the capture of the pendulum to tracking. Two mechanisms therefore make w depend on the pendulum state.

State-dependent weighting. The angle-dependent weight is

$$w(\theta) = 0.3 + 0.5\left[1 - \exp\left(-\frac{|\theta|}{5^\circ}\right)\right]. \tag{12}$$

The pendulum weight increases with the angular deviation, and the tracking weight $1 - w$ rises to 0.7 at the upright position (Figure 3a).

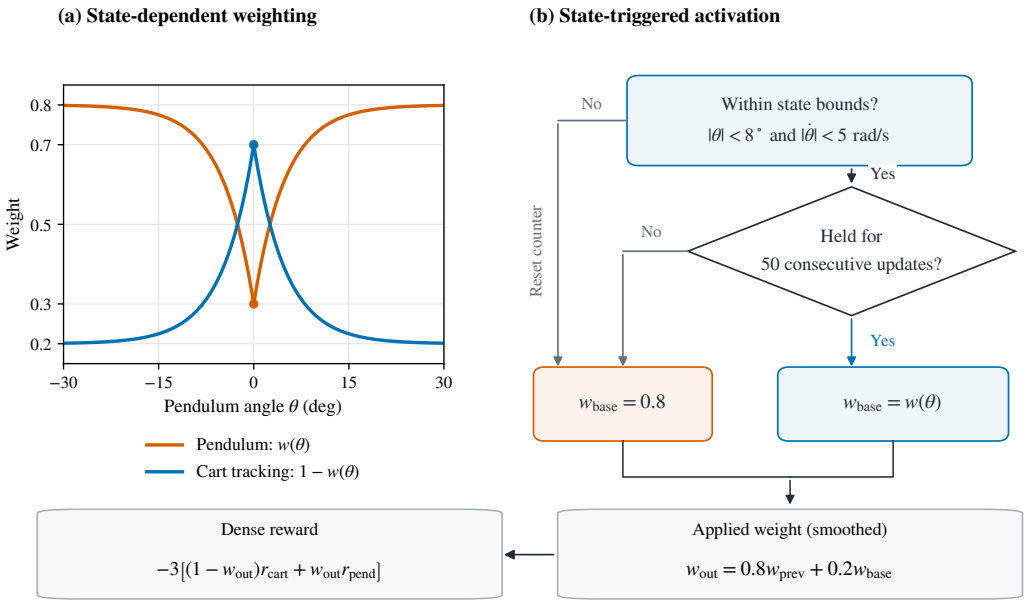


Figure 3. Gated state-dependent reward weighting.

State-triggered weight activation. Used alone, $w(\theta)$ lowers the pendulum weight at every passage through the upright position during swing-up, even at a high angular velocity. Training can then settle on a policy that keeps the pendulum rotating (Section 4.4). The gate therefore enables $w(\theta)$ only after the pendulum has stayed within state bounds for 50 consecutive weight updates. The weight is updated every 0.005 s, four times per agent step of 0.02 s, so that the 50 updates last 0.25 s. With k denoting the weight update, the state bounds are met when $c_k = 1$, where

$$c_k = \begin{cases} 1, & |\theta_k| < 8^\circ \text{ and } |\dot{\theta}_k| < 5 \text{ rad/s}, \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

The counter $n_{\text{ok},k}$ counts the consecutive updates up to k with $c_k = 1$ and resets to zero when $c_k = 0$. The base weight is

$$w_{\text{base},k} = \begin{cases} w(\theta_k), & n_{\text{ok},k} \geq 50, \\ 0.8, & \text{otherwise,} \end{cases} \tag{14}$$

and the applied weight is smoothed by a first-order filter,

$$w_{\text{out},k} = 0.8 w_{\text{prev}} + 0.2 w_{\text{base},k}, \tag{15}$$

where $w_{\text{prev}} = w_{\text{out},k-1}$ is the previous applied weight. The filter has a time constant of about 0.02 s; it only limits the rate of change in the applied weight and does not change its steady-state values. The dense reward of Equation (11) uses $w = w_{\text{out}}$ (Figure 3b).

Both mechanisms act on the reward only: they schedule the training objective according to the pendulum state, not the controller gains. Gain scheduling, in contrast, adjusts the controller gains during operation according to a scheduling variable. Here, the deployed controller is the actor network alone, without a scheduling variable or switching logic, because the schedule is absorbed into the policy during training. Owing to the gate, the schedule also follows the actual capture of the pendulum in each episode rather than a fixed time.

3.1.3. Network Structure and Training

Table 3 lists the network architecture and training parameters; the hidden layers are fully connected with rectified linear unit (ReLU) activations. These values were adopted from the authors' earlier work on a hydraulically driven inverted pendulum [34] and are common DDPG settings. The critic is larger than the actor because it represents the action value over the observations and the action, whereas the actor is kept small for fast inference. The critic learning rate is three times the actor learning rate, and the target networks are updated slowly; both choices stabilize the updates. The discount factor of 0.99 gives an effective horizon of about 2 s, longer than the capture of the pendulum, and the 20 s episode covers swing-up and several reference moves.

Training uses the MATLAB Reinforcement Learning Toolbox. Each transition (s_t, a_t, r_t, s_{t+1}), with the reward r_t of Equation (9), is stored in the replay buffer. For a mini-batch of stored transitions (s_i, a_i, r_i, s_{i+1}), the critic is trained toward the target $y_i = r_i + \gamma Q'(s_{i+1}, \mu'(s_{i+1}))$, where μ' and Q' are the target actor and critic, and the actor is updated along the policy gradient obtained from the critic. The target networks follow the online networks through soft updates with the factor τ [14]. An episode ends after 20 s or when the cart-position limit is exceeded. Training stopped when the 100-episode average return first exceeded 1.2×10^5, about 76% of the maximum return of a 20 s episode (Section 4.4).

Table 3. Network architecture and DDPG training parameters.

Parameter	Value
Actor hidden layers; output	250 and 250 neurons, ReLU; tanh
Critic hidden layers; output	1000 and 500 neurons, ReLU; 1 neuron, linear
Agent sampling interval T_s; weight-update interval	0.02 s; 0.005 s
Maximum episode duration	20 s
Replay buffer capacity; mini-batch size	10^6 transitions; 256
Discount factor γ; target-network update factor τ	0.99; 5×10^{-3}
Actor and critic learning rates	5×10^{-5}; 1.5×10^{-4}

3.2. Model-Based Baseline Controller

The baseline is a staged controller: an energy-based law swings the pendulum up, an LQR stabilizes it and tracks the cart reference, and the two commands are blended according to the pendulum angle and angular velocity. One parameter set (Table 4) is used for all tasks.

3.2.1. Energy-Based Swing-Up Control

With a point mass at l_c, the pendulum energy and its error to the upright value are [6]

$$E = \frac{1}{2} m l_c^2 \dot{\theta}^2 + m g l_c (1 + \cos\theta), \qquad \Delta E = E - E_0, \quad E_0 = 2 m g l_c, \tag{16}$$

where θ is wrapped to $(-\pi, \pi]$, as in all parts of the baseline. The desired cart acceleration is

$$a_{\mathrm{sw}} = k_E \Delta E_{\mathrm{eff}} \tanh\left(k_\omega \dot{\theta} \cos\theta\right) - k_d \dot{\theta}\, [\, |\theta| < 45^\circ \,] + a_{\mathrm{c}} + a_0, \tag{17}$$

where $[\,\cdot\,]$ is the indicator of the condition and ΔE_{eff} is ΔE with a deadband and a limit. For a point-mass pendulum, Equation (1) gives $\dot{E} = -m l_c \dot{\theta} \cos\theta\, \ddot{x}$, so the first term adds energy for $\Delta E_{\mathrm{eff}} < 0$ and removes it for $\Delta E_{\mathrm{eff}} > 0$ [6]. The second term shortens the final approach to the upright position, where the LQR takes over. The centering term $a_{\mathrm{c}} = -k_w(|x| - x_w)\,\mathrm{sgn}(x) - k_v \dot{x}$ acts only for $|x| > x_w$, its damping part only while the cart moves outward, and keeps the cart away from the end stops. The startup term

$a_0 = 0.5\ \mathrm{m/s^2}\,\mathrm{sgn}(\theta)$ starts the motion from rest ($|\dot{\theta}| < 0.05$ rad/s, $|\theta| > 20°$), where the energy term is zero. Because the valve flow sets the cart velocity, $\dot{x} \approx K_{q1}u/A_A$ for the incompressible chambers, a_{sw} is integrated into a velocity command,

$$\dot{v}_{\mathrm{sw}} = a_{\mathrm{sw}} - \frac{v_{\mathrm{sw}}}{\tau}, \qquad |v_{\mathrm{sw}}| \le v_{\mathrm{lim}}, \tag{18}$$

which gives the valve command $u_{\mathrm{sw}} = (A_A/K_{q1})\,v_{\mathrm{sw}}$. Applying a_{sw} directly to the valve would produce a velocity step at every sign change in $\dot{\theta}$ and excite a limit cycle between the valve and the pendulum.

Table 4. Baseline controller parameters.

Parameter	Value
Swing-up gains k_E, k_ω, k_d	1.6, 5 s/rad, 0.8
Energy deadband, limit	$0.05E_0$, $0.8E_0$
Velocity command τ, v_{lim}	2 s, 0.6 m/s
Centering x_w, k_w, k_v	0.18 m, 200 $\mathrm{s^{-2}}$, 5 $\mathrm{s^{-1}}$
Diagonal of $\mathbf{S}_z$	(8°, 80°/s, 0.15 m, 0.5 m/s, 25 kN)
Diagonal of $\mathbf{Q}_s$; input scale s_u; input weight R_s	(6300, 1200, 10800, 6300, 2.5); 3 mm; 8000
Design bulk modulus β_e	7×10^8 Pa
Chamber volumes at mid-stroke V_{10}, V_{20}	3.59×10^{-4}, 2.11×10^{-4} $\mathrm{m^3}$
Nominal density ρ_n, orifice pressure drop Δp_n	1000 $\mathrm{kg/m^3}$, 9 MPa
Valve flow gain K_{q1}	0.1712 $\mathrm{m^2/s}$

3.2.2. LQR Stabilization and Controller Blending

The LQR uses the design model of Equation (7) with the valve area of the Simscape model. With the scaled state $\bar{\mathbf{z}} = \mathbf{S}_z^{-1}\mathbf{z}$ and input $\bar{u} = u/s_u$, the gain minimizes [35]

$$J_{\mathrm{LQR}} = \int_0^\infty \left(\bar{\mathbf{z}}^{\mathrm{T}}\mathbf{Q}_s\bar{\mathbf{z}} + R_s\bar{u}^2\right)\mathrm{d}t. \tag{19}$$

The LQR law is the standard state feedback [35], applied to the tracking error,

$$u_{\mathrm{LQR}} = -\mathbf{K}\mathbf{e}, \qquad \mathbf{e} = \mathbf{z} - \mathbf{z}_{\mathrm{ref}}, \qquad \mathbf{z}_{\mathrm{ref}} = (\theta_{\mathrm{ref}},\ \dot{\theta}_{\mathrm{ref}},\ x_{\mathrm{ref}},\ \dot{x}_{\mathrm{ref}},\ F_{\mathrm{ref}})^{\mathrm{T}}, \tag{20}$$

where $\mathbf{K} = (K_1, \ldots, K_5)$ is the gain vector for the five states and $\mathbf{z}_{\mathrm{ref}}$ is the reference state. The law is pure error feedback: no reference feedforward, integral action, or gain scheduling is added, and the same $\mathbf{K}$ is used for stabilization and for all tracking references. The force F is calculated from the chamber gauge pressures with Equation (2). On the design model, the LQR places the dominant closed-loop poles at -2.8 ± 1.4j and $-15\ \mathrm{s^{-1}}$. The design model omits the valve dynamics, whereas the simulation model has incompressible chambers. To check that the gains carry over, the closed loop was also linearized for the simulation model with the second-order valve dynamics of Table 1. Its dominant poles stay close to those of the design, and it is stable for any velocity gain $\dot{x}/u$ above 62 $\mathrm{s^{-1}}$, whereas the asymmetric cylinder has about 170 $\mathrm{s^{-1}}$ in retraction and 240 $\mathrm{s^{-1}}$ in extension.

The LQR weight λ is 1 for $|\theta| \le 10°$ and 0 for $|\theta| \ge 25°$; in between, $\lambda = 1 - 3\xi^2 + 2\xi^3$ with $\xi = (|\theta| - 10°)/15°$. When $|\dot{\theta}| > 2$ rad/s, λ is limited to 0.3 to reduce the LQR contribution during fast rotation. The valve command $u_{\mathrm{cmd}} = (1-\lambda)u_{\mathrm{sw}} + \lambda u_{\mathrm{LQR}}$ then passes through the valve dynamics and the ±6 mm limit.

4. Simulation Results

This section evaluates whether the gated reward weighting enables one policy to perform swing-up, stabilization, and tracking. Section 4.2 compares the proposed controller with the staged baseline, including the valve and force signals, and Sections 4.3 and 4.4

assess the effects of the state-dependent weight and the gate. All results are simulation runs on the model of Section 2.3, whose incompressible, frictionless chambers make the cylinder a velocity source; hardware validation is the next stage.

4.1. Simulation Setup and Measures

Three controllers are compared on the same Simscape model, with the same ±6 mm spool limit and one parameter set each: the adaptive-weight RL controller of Section 3.1 (hereafter the proposed controller), a fixed-weight RL controller trained with $w = 0.7$, and the baseline of Section 3.2. Both RL controllers are trained with the training reference only and are evaluated on the validation references without retraining. The runs of the proposed controller and the baseline start with the pendulum at rest near the hanging position, those of the fixed-weight controller near the upright position, and the cart starts at rest at the origin. The references are repeated point-to-point moves with velocity and acceleration limits (Table 5). The swing-up comparison uses the first 9 s of the high-speed runs, during which the reference is held at −0.20 m, and the mass and force tests use the low-speed reference starting at 9 s.

Table 5. Reference trajectories.

Reference	End Positions (m)	Max. Velocity (m/s)	Max. Acceleration (m/s^2)	Moves Start (s)
Training	±0.025	0.01	0.005	3
Low-speed validation	±0.20	0.10	0.05	3
High-speed validation	±0.20	0.50	0.50	9

The settling time t_s is the time after which $|\theta| \leq 5^\circ$ and $|\dot{\theta}| \leq 10^\circ/\text{s}$ remain satisfied, and the pendulum is counted as fallen when $|\theta|$ exceeds 25° after it has been captured. Tracking is evaluated over 15–50 s, when every completed run has settled and the first reference transition is over, by

$$\text{RMSE}_x = \sqrt{\frac{1}{t_b - t_a} \int_{t_a}^{t_b} e_x^2(t)\, \mathrm{d}t}, \qquad e_{x,\max} = \max_{t \in [t_a, t_b]} |e_x(t)|, \tag{21}$$

together with the mean error $\bar{e}$ and the standard deviation $\sigma_e = \sqrt{\text{RMSE}_x^2 - \bar{e}^2}$, which measures the error during the reference transitions with the steady-state offset removed. At the actuator level, the time fractions at the spool displacement limit ($|x_v| \geq 5.9$ mm) and at the spool rate limit ($|\dot{x}_v| \geq 0.24$ m/s), the number of spool reversals (sign changes with 0.2 mm hysteresis), and the hydraulic work $W_h = \int |F\dot{x}|\, \mathrm{d}t$ exchanged between the cylinder and the cart are reported. The first 0.5 s of each run, with the pressure build-up of the initialization, are excluded from these measures.

4.2. Comparison with the Model-Based Baseline

Table 6 collects the valve and force measures of the proposed controller and the baseline in all test phases; the following subsections refer to its rows.

4.2.1. Swing-Up

Figure 4 shows the swing-up of both controllers with the reference held at −0.20 m. The proposed controller settles at 2.0 s and the baseline at 4.7 s. The proposed controller reaches a higher peak angular velocity (17.2 against 14.4 rad/s) and a larger cart excursion (0.26 against 0.21 m), within the ±0.30 m stroke, and after settling both hold the pendulum within 0.25° of the upright position.

Table 6. Valve and force measures per test phase.

Case, Phase [1]	Controller	$x_{v,\mathrm{rms}}$ (mm)	$\lvert x_v \rvert_{\max}$ (mm)	Displ. Limit (%)	Rate Limit (%)	$\dot{x}_{v,\mathrm{rms}}$ (mm/s)	$\lvert F \rvert_{\max}$ (N)	W_h (J)
Swing-up	Adaptive-weight RL	1.35	6.02	1.0	9.8	92.3	997	198.8
	Model-based baseline	1.75	3.37	0.0	0.0	14.9	643	27.6
Training, tracking	Adaptive-weight RL	0.11	0.23	0.0	0.0	0.4	10	0.2
	Model-based baseline	0.10	0.15	0.0	0.0	0.1	1	<0.1
Low speed, tracking	Adaptive-weight RL	0.35	0.79	0.0	0.0	1.2	60	3.2
	Model-based baseline	0.35	0.80	0.0	0.0	0.4	3	1.1
High speed, moves	Adaptive-weight RL	0.80	6.02	0.1	0.6	27.3	955	267.7
	Model-based baseline [2]	1.44	6.00	1.3	3.6	50.5	18219	1334.3
1000 N force, 12–18 s	Adaptive-weight RL	3.90	6.02	14.9	58.8	208.6	1907	4098.4
	Model-based baseline	0.34	0.76	0.0	0.0	3.1	1032	273.4

[1] Windows: swing-up 0.5–5 s, tracking 15–50 s, high-speed moves 9–50 s, force test 12–18 s. [2] Includes the end-stop impact and the fallen pendulum after 10.7 s.

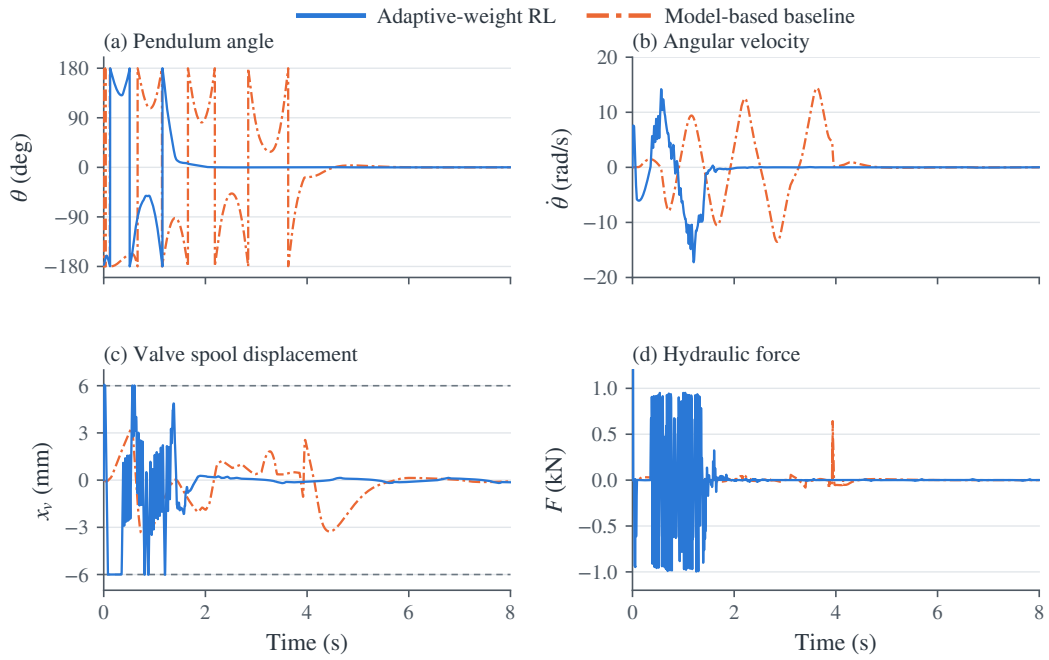


Figure 4. Swing-up of the proposed controller and the baseline: (**a**) pendulum angle; (**b**) angular velocity; (**c**) valve spool displacement; (**d**) hydraulic force.

The valve signals explain the faster swing-up (Figure 4c). The proposed controller drives the spool between its displacement limits, with 32 reversals in 0.5–5 s. It is at the rate limit for 9.8% of this window but at the displacement limit for only 1.0%, so its valve is limited by the spool rate rather than by the spool displacement. The baseline commands a smooth cart velocity (Equation (18)) and keeps the spool within 3.4 mm with five reversals. In all runs of Table 6, the chamber pressures remain between 1.2 and 16.7 MPa, within the tank and relief limits, except during the end-stop impact of the baseline (Section 4.2.2). The hydraulic force peaks at 1.0 kN against 0.64 kN (Figure 4d), and W_h is 199 J against 28 J. Only 1.2 J remains as potential energy of the upright pendulum; the remainder is exchanged between the cylinder and the cart at each reversal. W_h therefore measures the loading of the valve and the cylinder, not the energy consumption, which is set by the fixed-displacement pump. Because the reward contains no actuator-effort term, the policy is optimized for speed only.

Because the policy and the plant model are deterministic, a repeated run gives the same result. The sensitivity to the initial condition was therefore tested on a grid of 25 runs under the training reference, combining five initial angles within the training range (±175.5°, ±177.5°, and 180°) with five cart start positions (−0.20 to 0.20 m in steps of 0.10 m). The pendulum was captured and held upright in all 25 runs, with a settling time of 2.21 ± 0.67 s (mean ± standard deviation; range 1.37–4.38 s), without end-stop contact, and with the same tracking RMSE of 11.14 mm.

4.2.2. Tracking and Operating Range

Figure 5 shows the cart position of all three controllers under the three references, and Table 7 lists the errors; the fixed-weight controller is discussed in Section 4.3. At the training and low-speed references, the baseline is more accurate, with an RMSE of 3.98 and 15.77 mm against 11.14 and 17.63 mm. The difference is almost entirely the mean error. The proposed controller settles with a constant offset of about +10 mm under all three references, the baseline with about −1 mm. With the mean removed, both follow the reference transitions equally well (σ_e of 4.6 against 3.8 mm and 14.5 against 15.7 mm).

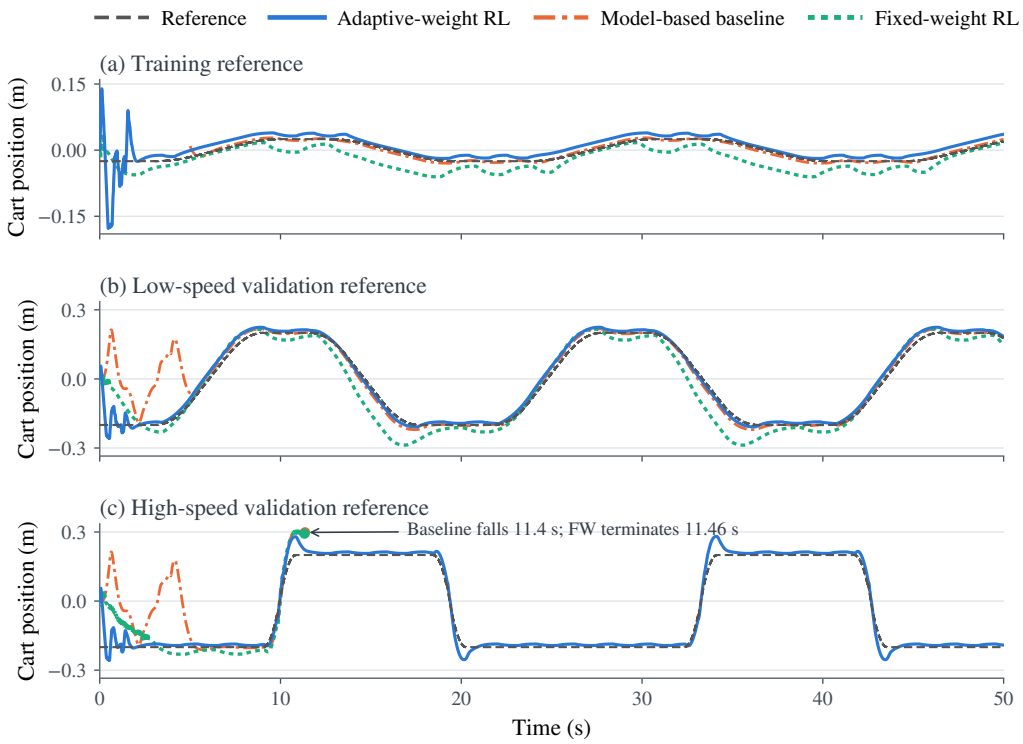


Figure 5. Cart position of the three controllers under the three references: (**a**) training reference; (**b**) low-speed validation reference; (**c**) high-speed validation reference.

The offset is a property of the policy. At rest the valve must be closed, so the cart stops where the policy output crosses zero, here 7–13 mm ahead of the reference. Nothing in the training fixes this crossing at $e_x = 0$: inside the goal set, the goal reward of Equation (10) changes by less than 5% between $e_x = 0$ and 10 mm, and the policy has no integral action. The LQR law of Equation (20), in contrast, is zero at $\mathbf{e} = 0$ by construction.

Table 7. Tracking errors in mm over 15–50 s.

Reference	Controller	RMSE	$e_{x,\max}$	$\bar{e}$	σ_e
Training	Model-based baseline	3.98	6.42	−1.08	3.83
	Fixed-weight RL	25.25	39.52	−21.87	12.62
	Adaptive-weight RL	11.14	16.29	+10.13	4.64
Low-speed validation	Model-based baseline	15.77	32.78	−1.26	15.72
	Fixed-weight RL	53.01	123.57	−27.18	45.51
	Adaptive-weight RL	17.63	33.56	+10.00	14.52
High-speed validation	Model-based baseline	pendulum falls at 11.4 s			
	Fixed-weight RL	pendulum falls at 11.3 s			
	Adaptive-weight RL	20.96	85.35	+8.64	19.10

At the actuator level, both controllers use the same valve amplitude during low-speed tracking (spool RMS 0.35 mm, peak 0.8 mm, no valve limit reached; Table 6), but the proposed controller moves the spool more. Its output is held for 20 ms and changes stepwise, so its spool moves in small steps, with three times the spool-rate RMS of the baseline. Because the chambers are incompressible, each spool step is a step in cart velocity. These steps appear in the hydraulic force, with peaks of about 20 N in the move of Figure 6c and of 60 N over the whole tracking window, against 3 N for the baseline (Table 6). They also cause the small angular-velocity bursts of the pendulum during the moves (Figure 7g).

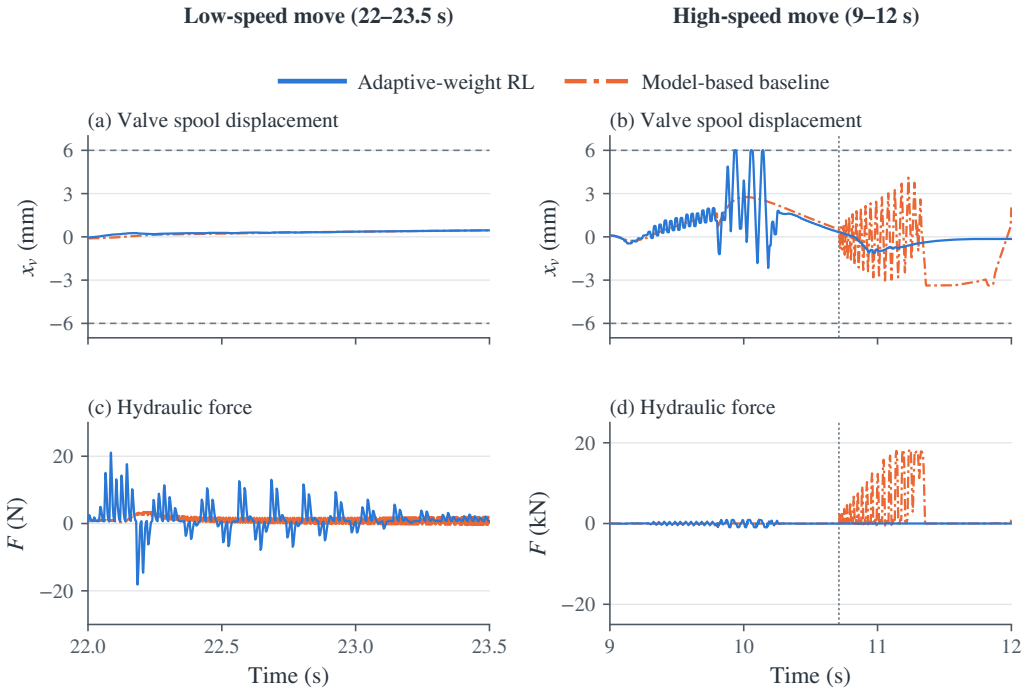


Figure 6. Valve spool displacement and hydraulic force during tracking: (**a**,**c**) low-speed move; (**b**,**d**) high-speed move.

The high-speed reference separates the operating ranges. All three controllers balance the pendulum before the reference starts at 9 s. The proposed controller completes the run with an RMSE of 20.96 mm. The baseline has no feedforward of the reference acceleration and, in the first deceleration from 0.5 m/s, overshoots the end position by 105 mm, more than the 100 mm between the end position and the stroke limit. The cart reaches the end stop at 10.71 s (dotted line in Figure 6b,d), p_A peaks at 22.8 MPa, the force reaches 18.2 kN, and the pendulum falls at 11.4 s. The valve limits are not the cause: up to the contact, the baseline spool stays within 2.8 mm, the opening that 0.5 m/s requires, and its limit

occupancy in Table 6 occurs after the contact. The proposed controller keeps a margin: its peak error is 85 mm, its cart stays at least 19 mm from the stroke limit, and its spool reaches the displacement limit only in three brief bursts between 9.93 and 10.15 s (Figure 6b).

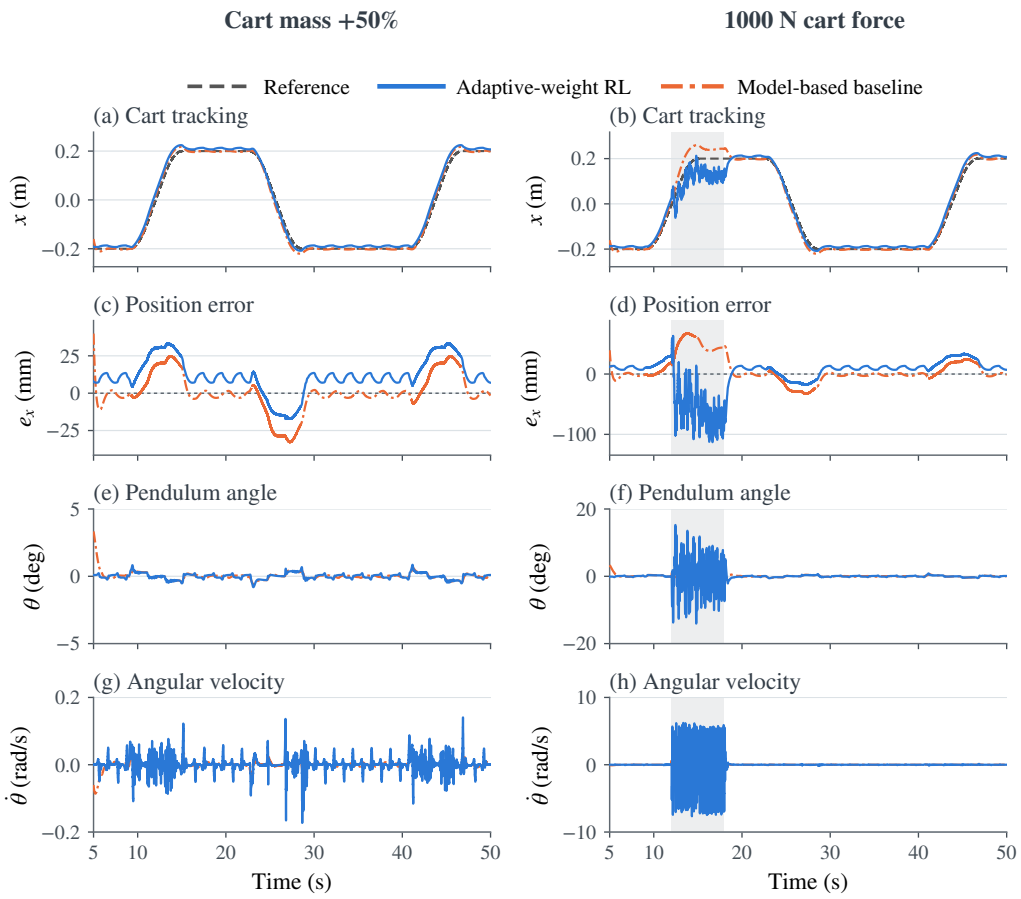


Figure 7. Responses to a 50% cart-mass increase and a 1000 N cart force: (**a**,**b**) cart position; (**c**,**d**) position error; (**e**,**f**) pendulum angle; (**g**,**h**) angular velocity. The shaded area marks the interval in which the force is applied.

4.2.3. Cart Mass and External Force

A 50% increase in cart mass and a 1000 N horizontal force on the cart during 12–18 s were simulated separately with the low-speed reference. Figure 7 shows the responses, Table 8 the measures, and Figure 8 the valve and force signals of the force test. The mass increase leaves the tracking of both controllers practically unchanged (Figure 7a,c,e): over 5–12 s, the RMSE differs by less than 0.4 mm from that of the force test before the force is applied. At the peak reference acceleration of 0.05 m/s^2, the inertial force rises only from 0.8 to 1.1 N, which is negligible against the force capability of about 17 kN and is absorbed by the chamber pressures.

Both controllers keep the pendulum upright under the 1000 N force and recover after it. The force is carried by the chamber pressures, and both controllers see it as a 1000 N step in the measured hydraulic force (Figure 8c,d). The baseline responds through its force feedback: its valve closes briefly, the cart settles 42–49 mm ahead of the reference (Figure 7b,d), and the pendulum stays within 1.3° (Figure 7f). The proposed controller, for which no external force was present during training, responds with valve reversals between the displacement limits throughout the interval (Figure 8a,b): 108 reversals, 59% of the time at the rate limit, and a force swinging between 0 and 1.9 kN. These reversals swing the pendulum to 15.2° and raise W_h to 4098 J, against 273 J for the baseline (Table 6). After

the force is removed, the pendulum is back within the angular bounds after 0.43 s, and the RMSE returns to 16.08 mm, close to the 15.15 mm before the disturbance.

Table 8. Responses to the cart-mass increase and the external force.

Case/Phase	Controller	Interval (s)	RMSE_x (mm)	$e_{x,\max}$ (mm)	$\max\lvert\theta\rvert$ (deg)	$\max\lvert\dot{\theta}\rvert$ (rad/s)
Cart mass +50%	Adaptive-weight RL	5–50	16.12	33.51	0.85	0.17
	Model-based baseline	5–50	12.21	39.88	3.33	0.09
Before force	Adaptive-weight RL	5–12	15.15	30.32	0.41	0.06
	Model-based baseline	5–12	7.63	39.37	3.31	0.09
Force applied	Adaptive-weight RL	12–18	64.99	112.88	15.24	7.66
	Model-based baseline	12–18	50.60	67.05	1.31	0.55
After removal	Adaptive-weight RL	18–50	16.08	80.51	7.38	5.50
	Model-based baseline	18–50	13.25	45.72	1.30	0.50

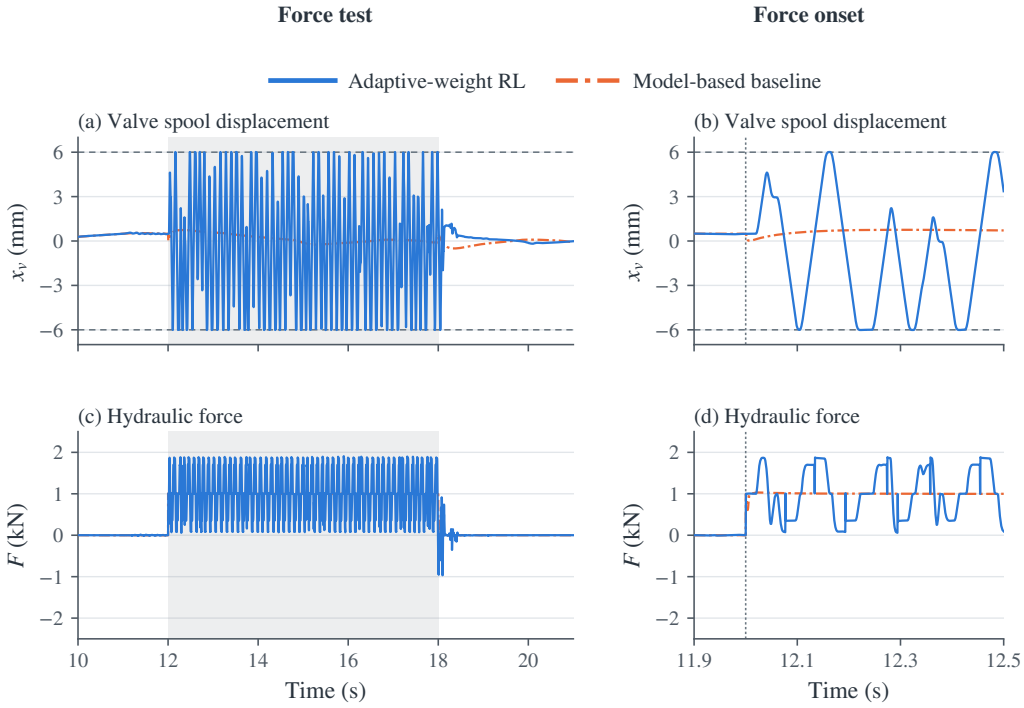


Figure 8. Valve spool displacement and hydraulic force in the 1000 N force test: (**a**,**c**) whole test; (**b**,**d**) force onset. The shaded area marks the interval in which the force is applied.

4.3. Effect of the State-Dependent Weight

The fixed-weight RL controller uses $w = 0.7$ throughout. Its runs start near the upright position, so the comparison covers stabilization and tracking but not swing-up. Its RMSE under the training and low-speed references is 25.25 and 53.01 mm, so the state-dependent weight lowers the RMSE by 55.9% and 66.7% (Table 7). The fixed-weight policy settles with a negative offset of 12–30 mm and lags further during the transitions (σ_e of 12.6 and 45.5 mm). Under the high-speed reference, it reaches the end stop at 10.75 s, in the same transition as the baseline, and the pendulum falls at 11.3 s; the run stops at 11.46 s, when $|\theta|$ exceeds the 1 rad termination bound of this environment (Figure 5c). With $w = 0.7$, the cart error carries 30% of the dense reward in every state; the state-dependent weight raises this share to 70% near the upright position and lowers it when the pendulum deviates.

Figure 9 shows the pendulum angle and the applied weight under the high-speed reference, from the same initial condition as Figure 4. During swing-up, the gate holds

the weight at 0.8 until the state bounds have held for 0.25 s, here from 1.54 to 1.79 s. The weight then follows Equation (12), delayed only by the filter, and is close to 0.3 from 2.2 s on. During the moves, the angle deviates by up to 10°; the weight rises with it and returns to 0.3 between moves. In the first move, $|\theta|$ exceeds 8° for 0.03 s at 9.86 s, so the gate resets and holds the weight at 0.8 for 0.25 s. In the other three moves, $|\theta|$ stays below 7°, and the weight follows Equation (12) up to 0.62–0.66. The gate thus acts mainly during the capture of the pendulum, and the angle-dependent weight during the moves.

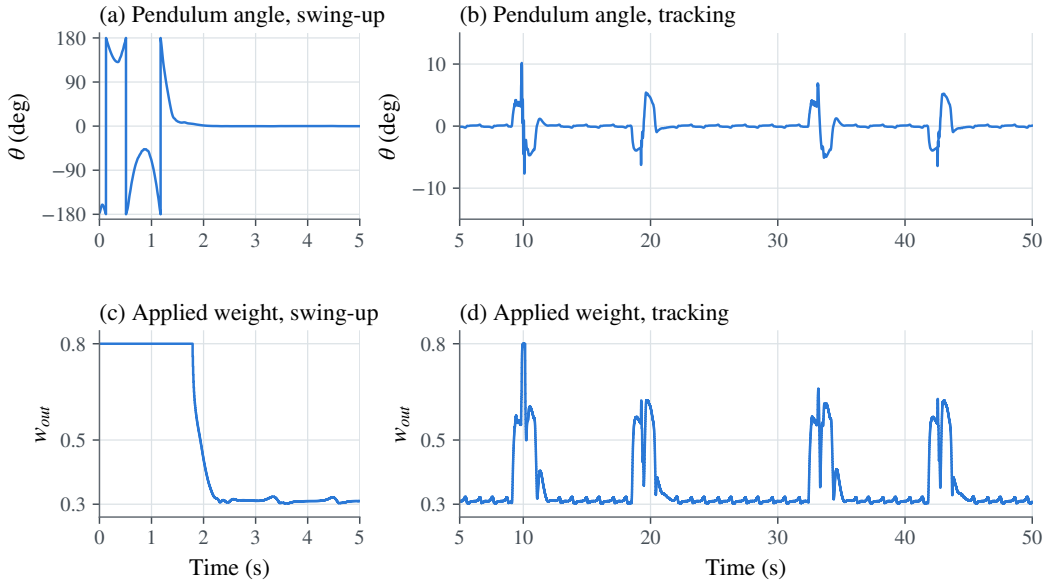


Figure 9. Pendulum angle and applied weight of the proposed controller: (**a**) pendulum angle during swing-up; (**b**) pendulum angle during tracking; (**c**) applied weight during swing-up; (**d**) applied weight during tracking.

4.4. Effect of the Activation Gate

To assess the effect of the gate, an agent was trained with $w_{\text{base},k} = w(\theta_k)$ at every update. The training settings of this run differed in part from those described above, so the comparison is indicative only. Figure 10a shows the 100-episode average return of both runs and, as markers, the returns of the evaluation episodes of the gated run. Without the gate, none of the 3000 episodes returned a positive reward: 1925 episodes ended at the end stop, and the others ran for 20 s with negative returns. The gated run shown, an earlier run than that of the evaluated policy, reached the stopping threshold. From the initial condition of Figure 9 and without exploration noise, the gated policy settles in 2.0 s, whereas the ungated policy saved at episode 3012, in a 20-episode continuation of the run, never captures the pendulum (Figure 10b). It rotates the pendulum through the upright position, 15.7 turns in 20 s, while keeping the cart within 0.28 m of the center. The other policies saved at the end of the ungated training behave in the same way or drive the cart into the end stop.

The cause is the timing of the weight release. Without the gate, every passage through the upright position lowers the pendulum weight and raises the cart-error weight, at the moment when the capture of the pendulum requires a large cart motion. The capture is thus penalized, whereas rotation is not. With the gate, the weight stays at 0.8 until the pendulum has remained within the state bounds for the hold interval, so the capture is rewarded before the tracking weight takes over.

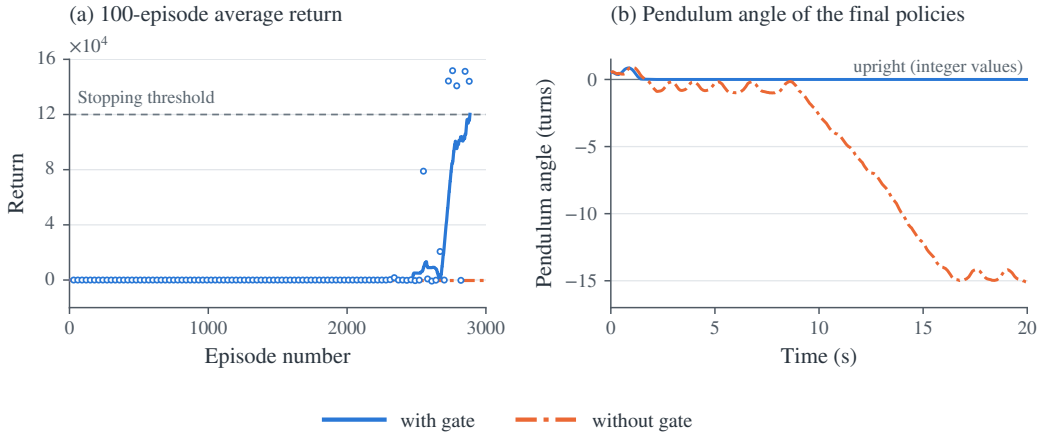


Figure 10. Gate ablation: training with and without the gate. (**a**) 100-episode average return, where the circles are the returns of the evaluation episodes of the gated run; (**b**) pendulum angle of the final policies.

5. Discussion

Against the staged baseline, the proposed controller performs swing-up, balancing, and tracking with one policy, swings up in less than half the time, and completes the 0.5 m/s reference at which the baseline overshoots into the end stop. The baseline is more accurate at low speed, almost entirely because of the 10 mm offset of the proposed controller, and it rejects the 1000 N force with a much smaller excursion (1.3° against 15°). The proposed controller also uses more valve motion and hydraulic work. Its reward contains no actuator-effort term, so the policy is optimized for speed only; an effort penalty would trade speed for valve motion. Reference feedforward or gain scheduling would remove the overshoot of the baseline at 0.5 m/s, but it would make the baseline specific to the reference, so the comparison uses the baseline without them. Against the fixed-weight RL controller, the state-dependent weight lowers the tracking error by more than half and extends the operating range to the high-speed reference, and the training run without the gate did not converge.

The gated weighting schedules the training objective according to the pendulum state, not the controller gains. Its model-based counterpart is the blending of the baseline, which switches between two control laws at run time; in the proposed controller, the schedule is absorbed into one network during training. Compared with the earlier pump-controlled study of the authors' group [34], which addressed swing-up and balancing in experiments, the present work keeps the DDPG algorithm, uses a valve-controlled cylinder, and adds cart trajectory tracking as a third objective, which the gated weighting makes trainable in one policy.

The results are obtained in simulation. The plant model has incompressible chambers and no friction, so the piston velocity is set by the valve flow, and inertial or external loads are taken up by the chamber pressures rather than by a displacement of the cart. On a real actuator with compliant chambers, the mass and force tests would also excite the hydraulic resonance and would be more demanding for both controllers. Two items should be addressed before hardware tests. The 10 mm offset could be removed by a goal reward that is sharper near zero error or by an integrated error in the observation, at the cost of retraining. The valve motion, in particular the reversals under the external force, would wear a real valve and should be penalized or filtered.

Future work will validate the controller experimentally. Sim-to-real transfer of RL policies has been achieved on hydraulic machines with domain randomization, modeled

actuation delays, and action-smoothness penalties [36,37]; these measures, together with a compressible-chamber model, define the path to hardware tests. The effects of control delay, measurement noise, and the control frequency will also be examined.

6. Conclusions

A DDPG controller with gated, state-dependent reward weighting was developed for swing-up, stabilization, and cart trajectory tracking with a proportional valve-controlled cylinder, the most widely used electro-hydraulic drive. The simulation results lead to three conclusions.

First, one policy performs all three tasks. It swung the pendulum up in 2.0 s instead of 4.7 s for a staged energy-based swing-up and LQR baseline, and it captured the pendulum in all 25 runs of the initial-condition grid, with a settling time of 2.21 ± 0.67 s and without end-stop contact. It completed the 0.5 m/s reference with an RMSE of 20.96 mm and a stroke margin of at least 19 mm, whereas the baseline and a fixed-weight RL controller reached the end stop.

Second, the weighting improves the tracking, and training without the gate did not converge. Compared with the fixed-weight controller ($w = 0.7$, started near the upright position), the RMSE under the training and low-speed references was 55.9% and 66.7% lower. The weight trace confirms the intended roles: the gate acts mainly during the capture of the pendulum, and the angle-dependent weight during the moves. Without the gate, none of the 3000 training episodes returned a positive reward, because the weight shifted toward the cart error at each passage through the upright position, when the capture required a large cart motion.

Third, these gains have costs. The baseline was more accurate at low speed, almost entirely because of the constant 10 mm offset of the policy, which has no integral action; with the mean error removed, both controllers followed the reference transitions equally well. Under a 1000 N external force, the policy reversed the valve 108 times, and the pendulum deviated by 15.2° against 1.3° for the baseline.

Author Contributions: Investigation, P.L.; supervision, B.H. and L.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (52675080), the Science and Technology Project of Hebei Education Department (BJ2025200), and the Youth Talent Program Supported by China Railway Society.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting the simulation results reported in this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

DDPG	Deep deterministic policy gradient
LQR	Linear quadratic regulator
ReLU	Rectified linear unit
RL	Reinforcement learning
RMSE	Root mean square error

References

1. Ding, R.; Sun, G.; Zhang, J.; Peng, L.; Cheng, M.; Xu, B.; Yang, H. A Review of Independent Metering Control System for Mobile Machinery. *Int. J. Hydromechatronics* **2025**, *8*, 1–39. [CrossRef]
2. Watton, J. *Fundamentals of Fluid Power Control*; Cambridge University Press: Cambridge, UK, 2009. [CrossRef]
3. Yao, J. Model-based Nonlinear Control of Hydraulic Servo Systems: Challenges, Developments and Perspectives. *Front. Mech. Eng.* **2018**, *13*, 179–210. [CrossRef]
4. Jelali, M.; Kroll, A. *Hydraulic Servo-Systems: Modelling, Identification and Control*; Advances in Industrial Control; Springer: London, UK, 2003. [CrossRef]
5. Xie, B.; Yao, B. Adaptive Robust Control for a Class of Non-Minimum Phase Nonlinear System. In Proceedings of the 2006 ASME International Mechanical Engineering Congress and Exposition, Chicago, IL, USA, 5–10 November 2006; Paper IMECE2006-15025. Available online: https://engineering.purdue.edu/~byao/Papers/IMECE2006-15025_NMP.pdf (accessed on 22 October 2026).
6. Åström, K.J.; Furuta, K. Swinging up a Pendulum by Energy Control. *Automatica* **2000**, *36*, 287–295. [CrossRef]
7. Donatoni, M.M.; Chavarette, F.R.; Preto, E.; Karmouche, D.C.J. Numerical and Experimental Investigation for Swing-Up Control of an Inverted Pendulum Using Arduino Microcontroller. *J. Mech. Eng. Sci.* **2024**, *18*, 10017–10025. [CrossRef]
8. Muškinja, N.; Tovornik, B. Swinging up and Stabilization of a Real Inverted Pendulum. *IEEE Trans. Ind. Electron.* **2006**, *53*, 631–639. [CrossRef]
9. Yao, B.; Bu, F.; Reedy, J.; Chiu, G.T.C. Adaptive Robust Motion Control of Single-Rod Hydraulic Actuators: Theory and Experiments. *IEEE/ASME Trans. Mechatron.* **2000**, *5*, 79–91. [CrossRef]
10. Yao, J.; Deng, W. Active Disturbance Rejection Adaptive Control of Hydraulic Servo Systems. *IEEE Trans. Ind. Electron.* **2017**, *64*, 8023–8032. [CrossRef]
11. Nguyen, M.H.; Dao, H.V.; Ahn, K.K. Active Disturbance Rejection Control for Position Tracking of Electro-Hydraulic Servo Systems under Modeling Uncertainty and External Load. *Actuators* **2021**, *10*, 20. [CrossRef]
12. Chen, X.; Tang, Y.; Huang, X.; Wang, W.; Guo, J.; Lin, F.; Jia, L.; Yang, H.; Zhu, Y. Automation and Synchronisation on Electro-Hydraulic Lifting System of Tunnel Boring Machine Segment Assembly. *Int. J. Hydromechatronics* **2025**, *8*, 94–120. [CrossRef]
13. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*, 2nd ed.; MIT Press: Cambridge, MA, USA, 2018.
14. Lillicrap, T.P.; Hunt, J.J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; Wierstra, D. Continuous Control with Deep Reinforcement Learning. In Proceedings of the International Conference on Learning Representations (ICLR), San Juan, Puerto Rico, 2–4 May 2016. Available online: https://arxiv.org/abs/1509.02971 (accessed on 22 October 2026).
15. Bhourji, R.S.; Mozaffari, S.; Alirezaee, S. Reinforcement Learning DDPG–PPO Agent-Based Control System for Rotary Inverted Pendulum. *Arab. J. Sci. Eng.* **2024**, *49*, 1683–1696. [CrossRef]
16. Baek, J.; Lee, C.; Lee, Y.S.; Jeon, S.; Han, S. Reinforcement Learning to Achieve Real-Time Control of Triple Inverted Pendulum. *Eng. Appl. Artif. Intell.* **2024**, *128*, 107518. [CrossRef]
17. Bajrami, X.; Kaçiu, F.; Shala, E.; Likaj, R. Real-Time Swing-Up of a Linear Inverted Pendulum Using Reinforcement Learning. *Mechanika* **2025**, *31*, 123–135. [CrossRef]
18. Xu, N.; Tran, H. Zero-Shot Transfer of Reinforcement Learning Control Policies for the Swing-Up and Stabilization of a Cart-Pole System. *arXiv* **2026**, arXiv:2606.22145 . [CrossRef]
19. Brumand-Poor, F.; Sharma, B.; Duensing, Y.; Logothetis, P.; Atali, E.; Schmitz, K. Machine Learning in Fluid Power—Applications, Trends, and Challenges. *Int. J. Fluid Power* **2026**, *27*, 359–434. [CrossRef]
20. Brumand-Poor, F.; Kauderer, L.; Matthiesen, G.; Schmitz, K. Application of Deep Reinforcement Learning Control of an Inverted Hydraulic Pendulum. *Int. J. Fluid Power* **2023**, *24*, 393–418. [CrossRef]
21. Gao, L.; Zhao, J.; Duan, J.; Liu, S.; Ma, F. Reinforcement Learning Based Motion Control Method for Electrical-hydraulic Valve-Controlled System. *IEEE/ASME Trans. Mechatron.* **2025**, *30*, 751–762. [CrossRef]
22. Coşkun, M.Y.; Itik, M. Position Control of a Digital Electrohydraulic System with Limited Sensory Data Using Double Deep Q-Network Controller. *Expert Syst. Appl.* **2024**, *252*, 124275. [CrossRef]
23. Ren, Z.; Tang, P.; Zheng, W.; Zhang, B. A Deep Reinforcement Learning Approach to Injection Speed Control in Injection Molding Machines with Servomotor-Driven Constant Pump Hydraulic System. *Actuators* **2024**, *13*, 376. [CrossRef]
24. Yuan, X.; Wang, Y.; Zhang, R.; Gao, Q.; Zhou, Z.; Zhou, R.; Yin, F. Reinforcement Learning Control of Hydraulic Servo System Based on TD3 Algorithm. *Machines* **2022**, *10*, 1244. [CrossRef]
25. Khater, A.A.; Fekry, M.; El-Bardini, M.; El-Nagar, A.M. Deep Reinforcement Learning-Based Adaptive Fuzzy Control for Electro-Hydraulic Servo System. *Neural Comput. Appl.* **2025**, *37*, 24607–24624. [CrossRef]
26. He, J.; Zhou, L.; Li, C.; Li, T.; Huang, J.; Su, S. Control Strategy of Hydraulic Servo Control Systems Based on the Integration of Soft Actor-Critic and Adaptive Robust Control. *IEEE Access* **2024**, *12*, 63629–63643. [CrossRef]
27. Li, Y.; Qi, X. Research on Motion Control of Hydraulic Manipulator Based on Prescribed Performance and Reinforcement Learning. *Actuators* **2026**, *15*, 39. [CrossRef]

28. Rout, R.; Kumawat, A.K. Reinforcement Learning-Based Position Tracking Control for Proportional Directional Control Valve-Based Electro-Hydraulic System. *IEEE Access* **2025**, *13*, 159597–159609. [CrossRef]
29. Han, T.; Nie, X.; Que, N.; Lu, J.; Yao, J.; Yu, X. Predefined-Time Tracking Control of Servo Hydraulic Cylinder Based on Reinforcement Learning. *Actuators* **2026**, *15*, 9. [CrossRef]
30. Helian, B.; Liu, X.; Liu, Z.; Geimer, M. Reinforcement Learning-Based Autonomous Control Methodology of Hydraulic Excavators. In Proceedings of the 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hangzhou, China, 19–25 October 2025; pp. 4881–4888. [CrossRef]
31. Roijers, D.M.; Vamplew, P.; Whiteson, S.; Dazeley, R. A Survey of Multi-Objective Sequential Decision-Making. *J. Artif. Intell. Res.* **2013**, *48*, 67–113. [CrossRef]
32. Hayes, C.F.; Rădulescu, R.; Bargiacchi, E.; Källström, J.; Macfarlane, M.; Reymond, M.; Verstraeten, T.; Zintgraf, L.M.; Dazeley, R.; Heintz, F.; et al. A Practical Guide to Multi-objective Reinforcement Learning and Planning. *Auton. Agents Multi-Agent Syst.* **2022**, *36*, 26. [CrossRef]
33. Freitag, K.; Ceder, K.; Laezza, R.; Åkesson, K.; Haghir Chehreghani, M. Curriculum Reinforcement Learning for Complex Reward Functions. *arXiv* **2024**, arXiv:2410.16790 . [CrossRef]
34. Helian, B.; Schmitt, G.; Yang, M.; Bian, Y.; Geimer, M. Reinforcement Learning-Based Control for Electrohydraulic Actuators: A Case Study on an Inverted Pendulum Testbench. *IEEE Trans. Ind. Electron.* **2025**, *72*, 14700–14711. [CrossRef]
35. Lewis, F.L.; Vrabie, D.L.; Syrmos, V.L. *Optimal Control*, 3rd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2012. [CrossRef]
36. Wiberg, V.; Wallin, E.; Fälldin, A.; Semberg, T.; Rossander, M.; Wadbro, E.; Servin, M. Sim-to-Real Transfer of Active Suspension Control Using Deep Reinforcement Learning. *Robot. Auton. Syst.* **2024**, *179*, 104731. [CrossRef]
37. Gruetter, J.; Terenzi, L.; Egli, P.; Hutter, M. Towards Learning Boulder Excavation with Hydraulic Excavators. In Proceedings of the 2026 IEEE International Conference on Robotics and Automation (ICRA), Vienna, Austria, 1–5 June 2026; pp. 21903–21910. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.