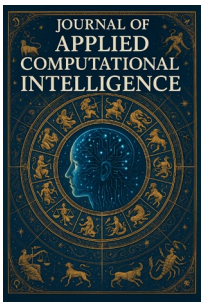
UNIVERSITY OF WEST ATTICA

eISSN: 3136-6937

**Journal of Applied Computational Intelligence**

Vol. 4, No. 2, April, 2026




# Detachment Risk Constrained Soft Actor Critic for Adhesion Force Control of Magnetic Wall Climbing Robots

Živojka Janković[1,*], Branka Čolaković[1], Čedomirka Radošević[1]

[1] *Faculty of Technical Sciences, University of Priština, Priština, 38000, Serbia*

Received 14 January 2026; Revised 05 May 2026; Accepted 10 May 2026; Published 13 May 2026

**Abstract**

Magnetic adhesion control must balance contact maintenance against locomotion resistance. This paper presents a soft actor-critic method with a detachment risk budget and independent action projection for magnetic wall-climbing robots. A reduced contact model links magnetic force, tangential demand, overturning moment and uncertain friction to a normalized attachment margin. The controller uses this margin to construct a conservative risk signal, learns a constrained stochastic policy, and projects candidate commands into a feasible actuator interval prior to execution. The numerical evaluation examines force tracking, route energy consumption, attachment reserve, and execution workload under varying operating conditions. The constrained controller reduces the mean route energy under the five operating conditions from 14.62 to 11.28 Wh compared with a conservative fixed-force baseline and reduces the number of unsafe episodes from 38 to 4 over 500 episodes compared to an unconstrained learning baseline. Its root mean square force tracking error is 11.2 N, compared with 17.8 N for the unconstrained baseline. Based on sensitivity analysis, a reduced risk budget leads to higher energy consumption and intervention frequency; increased latency reduces the attachment margin. Numerical comparisons show the interplay among reserve retention, actuator adjustment and route energy. The framework provides a physically interpretable control specification and an auditable evaluation structure linking mechanical requirements, policy preferences, and command feasibility.

*Keywords:* Magnetic Adhesion, Wall Climbing Robotics, Detachment Risk, Constrained Reinforcement Learning, Soft Actor Critic

## 1. Introduction

Magnetic wall-climbing robots provide inspection and maintenance platforms for ferromagnetic surfaces in environments with restricted access [1]. Research on magnetic adhesion modules has examined how magnetic circuit arrangements and structural parameters affect attachment force and module weight [2]. Experimental work on magneto-rheological contact pads has related static friction to pad properties and countersurface roughness [3]. Mechanical analyses of magnetic climbing robots have examined sliding and overturning during obstacle traversal and evaluated mechanisms for improving stability [4]. Transition-capable magnetic robots have been designed to adjust body attitude on curved surfaces and traverse right-angle and thin-wall transitions [5]. These coupled challenges are particularly consequential when the magnetic force is changed during robot motion. The force request suitable for a uniform vertical plate may have limited reserve after a payload shift, and prolonged high adhesion can increase locomotion resistance on less demanding route sections. Inspection routes combine these operating modes in a single mission and thus impose different requirements on the same actuator and contact estimator. Force regulation therefore requires an estimate of the instantaneous attachment reserve and an actuator response capable of maintaining it. Average tracking accuracy captures

* Corresponding author: jankovic.z@tf.pr.ac.rs

http://dx.doi.org/10.68406/jaci.2026.vol4iss2no8:106-122

➢ This is an open access article under the CC BY-NC-SA 4.0 license (https://creativecommons.org/licenses/by/4.0/).

© Authors retain all copyrights.

only part of this behavior because brief demand peaks can govern contact feasibility despite their limited contribution to the integrated error.

Tian et al. [6] introduced soft magnetic covers to reduce adhesion selectively and improve control during wall-climbing tasks. Model-based control offers an explicit model of the mechanical requirements, actuator limits and contact constraints, and its performance is contingent upon the precision and timeliness of the data provided for this model. Recent reviews of robotic reinforcement learning have examined its use in acquiring control behaviors under the complexity and interaction costs of physical environments [7]. A policy can account for the future energy consequences of increasing adhesion before a disturbance and the tracking consequences of reducing adhesion afterward. Soft actor-critic learns a stochastic policy within a maximum entropy framework, and recent work has extended its policy representation using state-dependent normalizing flows [8]. This formulation is relevant to adjustable adhesion because the force reference varies continuously and influences behavior across multiple control updates. The engineering challenge is to connect the optimization objective to the physical conditions required for contact. An accumulated reward may favor efficient behavior along most of a route while assigning limited influence to brief reserve depletion. Therefore, the scale of the energy term, motion error term or failure penalty affects the resulting preference for continued traversal and additional attachment. This preference requires an operational interpretation that remains meaningful as payload, friction, and observation quality vary.

Safe reinforcement learning research has distinguished safety requirements and constrained objectives from task reward maximization [9]. The budget describes expected exposure across trajectories, while the climbing vehicle must also satisfy an immediate force requirement at its current contact state. Safety filters restrict proposed control inputs using models or certificates of admissible system behavior [10]. Their application depends on the intersection between required attachment and achievable actuator response. When this intersection is narrow, command selection becomes sensitive to force estimation bias, observation delay, and the remaining adjustment rate. When the intersection is empty, the motion request must be revised through deceleration or stopping. These cases require separate reporting because frequent intervention can preserve contact while concealing limited policy anticipation of changing demand. A suitable architecture should retain the original policy command, the accepted command and the corresponding reserve to allow for the observation of each control layer's contribution. It should also differentiate between a warning level and the modelled detachment boundary. A warning excursion indicates declining reserve, whereas crossing the mechanical boundary identifies a contact feasibility violation. These distinctions support consistent evaluation of energy consumption, transient behavior, intervention burden, and unsuccessful route completion.

Given the uncertainty in contact demand, limited magnetic actuation, and continuous policy adaptation, a controller is needed to meet the attachment requirement via both learning and execution. A reduced sliding and overturning model is employed in this formulation to define a conservative reserve, allocate a budget for exposure near the reserve-warning boundary, and verify candidate commands against an achievable force interval. The study addresses adjustable aggregate adhesion on a steel wall with bounded actuator response and estimated tangential contact conditions. The four components of its assessment are condition-dependent energy, force response under disturbance, intervention intensity and the lower tail of attachment reserve. These complementary measures reveal how contact conservatism changes the relationship between task performance and actuator workload. The architecture includes a supervisory response for infeasible attachment demand and logs the interventions required to accept a candidate command. The contribution is the shared mechanical reserve used to connect adhesion demand, learning cost, and command feasibility in an aggregate force controller. The actor-critic architecture and projection principle serve as established building blocks; the proposed formulation concerns their contact-specific coupling. The technical presentation comprises six sections, followed by declarations and references. Section 2 reviews magnetic adhesion regulation, continuous policy learning, and execution constraints. Section 3 presents the contact model, conservative reserve and operating assumptions. Section 4 presents the constrained soft actor-critic objective, replay specification and command projection. Section 5 studies condition-dependent performance, disturbance response and sensitivity via the evaluation. Section 6 presents the contribution, limitations, and future research directions.

## 2. Related Work

### 2.1. Magnetic adhesion and contact regulation

Adaptive magnetic adhesion mechanisms have been developed to support payload transport and traversal of discontinuous surfaces with internal corners [11]. These mechanisms have different speeds of change for force and varying amounts of energy needed to establish or maintain a command. Therefore, a mechanism-specific energy model is required for the controller transfer between coil-driven adhesion and mechanically adjusted permanent magnets. The actuator interface specifies the control input and corresponding energy measurement. Therefore, a useful literature comparison distinguishes between the contact force demanded by the outer controller and the actual physical process generating this force. This separation allows the energy effect of a reduced-force command to be evaluated for the chosen

mechanism. For a permanent magnet system, much of the benefit may arise through reduced rolling resistance or lower adjustment work.

Studies of magnetic climbing on variable-curvature surfaces have modeled passive compliance, wheel contact constraints, and friction differences across motion modes [12]. Aggregate force measurement is the total of module loads, and local peeling assessment also needs module-level load distribution. Tangential traction is also affected by the friction at the contacts, even if there is a large moment reserve. Therefore, even though the final learning state is represented compactly, a sliding condition and an overturning condition should be kept in a practical attachment assessment. This requirement creates a gap between detailed mechanical simulations and deployable learning observations: simulations may contain inaccessible contact variables, whereas deployed observations rely on sensors and estimators available on the vehicle. The control specification addresses this gap through a conservative aggregate margin with explicit operating limits.

### 2.2. Continuous policy learning for robotic control

Distributional actor-critic policies have used different quantiles of the return distribution to represent distinct continuous control preferences [13]. This perspective is relevant when an increase in adhesion avoids the immediate contact issue but results in extra locomotion losses in the subsequent route segment. The policy must distinguish between the temporary reserve requirement and the continuous operating preference. Policy evaluation must account for simulator deficiencies, including unmodeled actuator delays and friction laws that overestimate available traction. Although a reward with many penalties can express these preferences, the numerical scale may obscure the operational significance of a rare failure.

Duan et al. [14] combined expected-value targets, twin value distributions, and variance-based critic gradient adjustment to improve value estimation and learning stability. Off-policy learning allows the reuse of stored transitions and is thus favorable when generating contact events is costly. Transition reuse also introduces a distribution mismatch when the action sent to the environment is different from the action proposed by the actor. If a safety filter modifies the command, the replay record must save the modified command and its results. Safe learning studies have examined unobserved environmental contexts, including payload changes and differing surface conditions, when assessing controller safety [15]. Variability for smooth traversal, abrupt disturbances and cases without a feasible adhesion command should be reported separately in magnetic climbing. Averaging these cases can obscure the distinction between poor policy selection and insufficient hardware capability.

### 2.3. Risk constraints and execution filters

Probabilistically constrained reinforcement learning has combined explicit safety policy gradients with primal-dual optimization to balance expected return and constraint satisfaction [16]. Cost selection is consequential: a binary detachment indicator provides sparse feedback, whereas a continuous margin penalty supplies earlier information but requires a defensible relationship to contact failure. A margin deficit cost serves as a risk surrogate, with its interpretation established through calibration against observed detachments. Statistical calibration of the surrogate to the observed detachment events is required for probability interpretation. The same distinction applies when comparing controllers with different estimators or confidence assumptions.

Data-driven safety certifiers have combined reinforcement learning actions with a learned backup controller to preserve safety and stability [17]. Their effectiveness depends on the availability of actuators, model uncertainty and the time lag between observation and execution. The filter restricts command acceptance according to the force capability and contact state available at execution. CVaR-constrained policy optimization has explicitly controlled the upper tail of cumulative constraint costs rather than relying solely on expected cost [18]. Combining tail evaluation with an interpretable margin exposes reserve depletion on difficult routes that may have limited influence on mean energy consumption. An unresolved engineering requirement is a transparent link among the mechanical margin, the budget employed in policy learning, and the realized command accepted by the actuator. This connection must be observable in cases of uncertainty about the risk estimate or an empty feasible action interval. The reviewed studies address magnetic mechanism design, continuous policy representation, and constrained execution at different levels. The present focus is the interface between an aggregate contact reserve and both policy cost and command acceptance; a comprehensive historical survey of reinforcement learning lies outside this focused comparison.

## 3. Robot Dynamics and Detachment Risk Representation

### 3.1. Contact mechanics and controllable adhesion

The proposed plant consists of a rigid vehicle with four magnetic contact modules arranged around a rectangular support region. The outer adhesion loop requests the total normal force, while a lower-level actuator loop distributes that request using a calibrated allocation rule. The reduced model assumes approximately equal module loading on a

flat wall. A separate moment reserve accounts for the principal peeling direction; individual wheel unloading requires module level resolution. The controlled quantity is the total effective adhesion force, measured in newtons after filtering the module force signals. Vehicle mass includes the payload, and tangential demand incorporates gravity resolved along the wall, commanded acceleration, and an external disturbance. Locomotion speed serves as a supervisory input that can be reduced when adhesion capability becomes inadequate.

Mechanical demand is represented by the larger of the forces required to resist sliding and overturning. This construction retains the governing contact requirement under the assumed friction lower bound and effective support geometry. The required adhesion force follows from the equation

$$F_{req} = \max\left\{\frac{mg\sin\theta + ma_t + d_t}{\mu_L}, \frac{|M_o|}{b_{eff}}\right\}. \quad (1)$$

Here, m denotes vehicle mass, g gravitational acceleration, θ wall inclination, $a_t$ tangential acceleration, and $d_t$ the tangential disturbance force. The tangential terms are expressed in the adverse climbing direction, and their combined demand is bounded below by zero when evaluating the sliding condition. The friction lower bound $\mu_L$ remains positive and decreases as estimation uncertainty increases. The effective lever arm $b_{eff}$ represents the available restoring moment per unit aggregate adhesion force. Its value depends on the support geometry and contact load distribution. The overturning moment $M_o$ includes contributions from payload offset and cable loading. The equation applies to traversal with an unchanged support arrangement; external corner transitions require a contact description that represents the changing support topology. The adverse-direction convention treats the resolved terms as demand contributions along the limiting slip direction. For bidirectional disturbances, the limiting direction must be evaluated for each candidate direction before selecting the larger sliding requirement.

Figure 1 presents the sensing and contact assessment flow from module force measurements and body motion to the conservative attachment requirement and supervisory speed request. Its branches distinguish tangential loading from overturning demand before combining them into the required force estimate. Friction uncertainty and actuator capability enter through separate paths, allowing their influence on reserve estimation and command feasibility to remain identifiable.

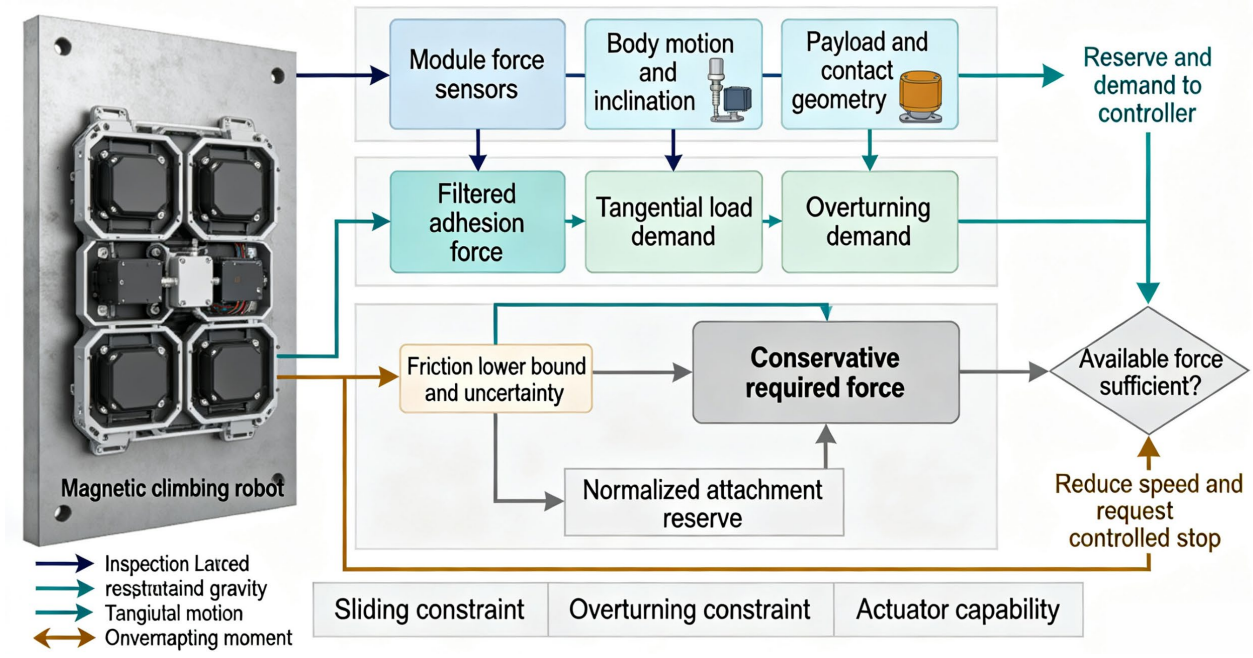



**Figure 1. Mechanical sensing and conservative attachment assessment for adjustable magnetic adhesion**

### 3.2. Normalized reserve and risk surrogate

The attachment reserve compares the lower credible value of the available adhesion force with the required force. Measurement uncertainty enters this calculation as an explicit force deduction, preserving the physical scales of energy and tracking performance when confidence conservatism changes. A positive force floor stabilizes normalization under lightly loaded conditions. The uncertainty deduction coefficient is calibrated using held out sensor residuals, with its value reflecting the desired conservatism of the force estimate. Reproducing this allowance requires the selected coefficient, the residual collection window, and empirical coverage on held-out measurements. The present specification describes the calibration role without supplying those numerical calibration records, so coverage remains an unresolved deployment requirement. These considerations lead to the normalized reserve equation:

$$h_t = \frac{F_t - k_\sigma \sigma_F - F_{req,t}}{max\left(F_{req,t}, F_{floor}\right)}. \quad (2)$$

The measured aggregate force is $F_t$, the force estimation uncertainty is $\sigma_F$, and $k_\sigma$ controls the uncertainty allowance. The denominator uses the larger of the instantaneous required force and the normalization floor $F_{floor}$. A positive reserve indicates that the conservative available force exceeds the modeled requirement. A negative reserve identifies a violation of the reduced contact condition. Physical separation additionally depends on compliance and local contact evolution: a brief force deficit may be recovered through contact compliance, while a positive aggregate reserve may coexist with local peeling. A warning margin of 0.20 therefore provides additional operating headroom above the mechanical boundary. Its selection depends on the intended operating envelope and the uncertainty represented by the estimator. The value 0.20 is an operating choice for the reported comparisons. It should be tuned against delay and uncertainty within the intended application, while the zero-reserve boundary retains its definition as the reduced mechanical feasibility limit.

Risk exposure increases continuously as the reserve falls below the warning margin. Squaring the normalized deficit emphasizes deeper incursions, while bounding the cost limits the influence of an individual extreme observation on learning updates. The resulting cost is expressed by the equation

$$c_t = min\left\{1, \left[\frac{max(0, h_{warn} - h_t)}{h_{warn}}\right]^2\right\} \quad (3)$$

In this equation, $h_{warn}$ denotes the warning margin. The cost remains zero when the reserve exceeds that margin, rises as the reserve decreases, and reaches its upper bound at the modeled contact boundary. It represents normalized reserve depletion; a probability interpretation requires calibration against observed detachment events. A budget of 0.02 constrains normalized discounted surrogate exposure under the stated objective. Converting this value into a detachment probability or a certified failure bound requires additional event calibration and assumptions about contact dynamics.

The estimator combines filtered force measurements, body acceleration, and a slowly varying friction lower bound. A short history of these estimates enters the observation to expose trends that an individual sample may conceal. During an abrupt air gap increase, the force history can reveal a declining attachment reserve before the instantaneous estimate crosses the warning threshold. Partial observability persists because surface contamination and rapid support changes become available through sampled measurements. Estimator delay consequently enters the uncertainty allowance and the assessment of command feasibility.

### 3.3. Actuation limits and operating assumptions

The actuator is represented by a bounded first order force response with a maximum adjustment rate. The aggregate force range is 180-650 N, and the force rate limit is 900 N per second. The outer controller operates at 50 Hz, with a faster inner force loop. At each outer update, the previous command and force rate bound define an achievable force interval. A candidate target is evaluated against that interval before execution. Observation delay reduces the useful reserve because mechanical demand can change while a command is being calculated and delivered. The latency evaluation therefore examines reserve sensitivity to the age of the information used for command selection. At 50 Hz, the outer command interval is 20 ms, corresponding to an 18 N command change under the stated rate bound. This is an admissible command increment; predicting realized force additionally requires the inner-loop response time and delay.

The operating envelope includes wall inclinations from 70 to 110 degrees, payloads from zero to 4 kg, and friction conditions ranging from clean to degraded contact. The 110-degree condition represents a modest overhang. Steel thickness, magnetic saturation, and module temperature remain fixed in the reduced model. These simplifications limit the range over which the force capability description can be transferred, particularly for thin plates where flux leakage changes the attainable adhesion. The aggregate force range is treated as available across the evaluated air gaps. Deployment requires a capability map that relates attainable force to gap, material properties, and actuator state. The gap comparison therefore tests reserve sensitivity within the declared force-capability assumption. A measured gap-dependent reduction in maximum force would further restrict the feasible command interval and could change the attainable energy and reserve tradeoff.

Energy accounting separates actuator electrical work from locomotion work associated with normal loading. The energy comparisons in Section 5 use a common route and duration. Mechanism specific efficiency assessment requires separate records of actuator voltage and current, drive motor power, and adjustment mechanism motion. Force tracking error is evaluated against the accepted reference after feasibility filtering, whereas task error refers to the original motion request. Retaining both quantities exposes the relationship between tracking accuracy and task relaxation. When adhesion demand becomes infeasible, the supervisor reduces tangential demand and requests a controlled stop. Episodes terminated through this response are recorded separately from completed traversals. Reported route energy should be

read as the combined task-energy quantity defined here, with attachment actuation and locomotion included. Comparisons involving interrupted routes require completion status and delivered route length alongside energy, because task duration changes the accumulated consumption.

## 4. Detachment Risk Constrained Soft Actor Critic Control

### 4.1. Observation action and learning objective

The observation contains filtered adhesion force, requested force, tangential speed, body acceleration, estimated wall inclination, payload estimate, friction lower bound, overturning demand, normalized reserve, previous action, and a short history of force residuals. Each channel is normalized using fixed engineering ranges established before training. These fixed ranges preserve a consistent observation scale across training and evaluation conditions. The action specifies a bounded increment in the aggregate force reference. This representation exposes the actuator rate limit to the policy and discourages unnecessary command reversals. The policy output is mapped to its admissible numerical range before undergoing mechanical feasibility projection. A reproducible observation specification must additionally list the history length, channel ordering, normalization limits, and treatment of missing history samples. These details determine the network input dimension and remain implementation parameters beyond the channel list given here.

The reward penalizes locomotion energy, deviation from the requested speed, and abrupt force command changes. A separate constraint represents attachment risk exposure. Reference scales normalize the reward components so that their relative influence remains consistent across physical units. The risk budget limits normalized discounted exposure, while unsafe episode frequency describes the occurrence of contact condition violations. These quantities distinguish accumulated reserve depletion from the distribution of boundary crossings across trajectories. Their relationship to policy optimization is specified in the constrained objective equation:

$$\begin{aligned} &\underset{\pi}{maximize} \quad \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t\left(r_t+\alpha\mathcal{H}\left(\pi(\cdot\mid s_t)\right)\right)\right] \\ &subject\ to \quad (1-\gamma)\mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t c_t\right]\le\varepsilon. \end{aligned} \tag{4}$$

The policy is denoted by $\pi$, the state by $s_t$, and the instantaneous reward by $r_t$. The discount factor $\gamma$ determines the weighting of future outcomes, while $\alpha$ controls the contribution of policy entropy $\mathcal{H}$. The factor $1-\gamma$ places discounted cost exposure on a normalized scale, and $\varepsilon$ sets the corresponding budget. The objective equation therefore retains energy and motion performance in the reward while assigning reserve depletion an explicit constraint.

The entropy temperature can be adjusted toward a target action entropy. The training specification uses a discount factor of 0.99 and a risk budget of 0.02. Episodes terminate upon modeled detachment, an infeasible emergency stop, route completion, or a fixed time limit. Terminal failure receives an additional cost so that route completion and attachment preservation remain relevant when trajectories have different durations. This treatment prevents low energy consumption caused by early termination from dominating the evaluation of useful task performance. The replay interface should distinguish absorbing contact failure and route completion from a time-limit truncation. Bootstrapping at a truncation depends on the continuing-task convention; the terminal penalty magnitude and that convention must be specified before numerical reproduction.

### 4.2. Critic updates and risk multiplier

Two reward critics estimate the soft action value, while a cost critic estimates discounted future risk exposure. Slowly updated target networks limit abrupt changes in learning targets. The actor favors actions that retain reward value and entropy while reducing predicted cost. Taking the minimum reward critic estimate limits optimistic action evaluation. The cost critic is calibrated on the same normalized exposure scale as the risk budget, preserving consistency between policy improvement and multiplier adaptation. The actor loss takes the form given by the equation

$$\mathcal{L}_\pi=\mathbb{E}\left[\alpha log\,\pi(a\mid s)-\min_{i\in\{1,2\}}Q_{r,i}(s,a)+\lambda(1-\gamma)Q_c(s,a)\right]. \tag{5}$$

Here, $Q_{r,i}$ denotes a reward critic, $Q_c$ the cost critic, and $\lambda$ the nonnegative risk multiplier. The expectation is evaluated using replay states and policy actions. The reward critics and entropy term retain the task objective, while the scaled cost term adjusts the preference for actions associated with future reserve depletion. The equation uses the same discount normalization as the exposure constraint, allowing the multiplier to respond to a consistently defined budget. Implementation must use one cost convention throughout: either estimate normalized discounted exposure directly or

apply the objective normalization consistently to an unnormalized discounted cost critic. The budget and multiplier estimate must follow the same convention.

The multiplier increases when estimated exposure exceeds the budget and decreases when exposure falls below it. Its adaptation proceeds more slowly than critic learning to limit policy oscillations caused by transient value estimation errors. A finite upper bound supports numerical stability, and reaching that bound triggers a diagnostic of unresolved constraint exposure. Saturation at the multiplier bound indicates that the constrained optimization may remain unresolved. A bounded multiplier supports stable updates but cannot by itself establish that the exposure budget is satisfied; that conclusion requires a separately reported exposure estimate. Multiplier adaptation is governed by the equation

$$\lambda_{k+1} = \mathrm{clip}\left(\lambda_k + \eta_\lambda\left(\hat{C}_k - \varepsilon\right), 0, \lambda_{\max}\right). \tag{6}$$

The update index is k, the multiplier learning rate is $\eta_\lambda$, and $\hat{C}_k$ estimates normalized discounted exposure. The upper limit is $\lambda_{max}$. Through this equation, persistent budget violations strengthen the influence of predicted risk on policy improvement. Entropy adaptation and multiplier adaptation retain different responsibilities: entropy regulates action dispersion, while the multiplier regulates the preference for risk exposure.

Replay stores the raw observation, candidate action, executed action, reward, cost, terminal flag, and intervention magnitude. Critic transitions use the executed action because that command produces the subsequent state. Actor updates evaluate a differentiable approximation of projection where feasible, and emergency stop transitions receive separate labels. These labels expose policy dependence on intervention and support comparison across operating envelopes. A balanced replay sampler retains rare low reserve transitions and records sampling probabilities for correction. The training specification uses two hidden layers with 256 units, batches of 256 transitions, a replay capacity of one million transitions, and independent seeds for initialization and environment variation. The differentiable approximation is a training device for nonempty feasible intervals. Clipping boundaries and emergency fallback require explicit gradient handling, and the present description leaves that implementation choice open; executed actions remain the inputs used for critic transitions.

### 4.3. Command projection and fallback behavior

The execution layer converts the current force requirement and warning margin into a lower acceptable command, then intersects this requirement with the actuator range and rate interval. Projection preserves the candidate command as closely as feasibility permits. Since the aggregate command is scalar, the projection reduces to clipping when the feasible set is a nonempty interval. Module level allocation subsequently checks each distributed command against the capability of its actuator. Uneven support additionally requires local force balance assessment. The projection and its force constraint are represented together in the equation

$$\begin{aligned} u_{exec} &= \underset{u \in \mathcal{U}_t}{arg\,min}(u - u_\pi)^2 \\ \mathcal{U}_t &= \left\{u \in \mathcal{U}_{act,t} : F_{pred}(u) \geq (1 + h_{warn})F_{req,t} + k_\sigma \sigma_F\right\}. \end{aligned} \tag{7}$$

The candidate command is $u_\pi$, the executed command is $u_{exec}$, and $\mathcal{U}_{act,t}$ contains commands satisfying the actuator magnitude and rate limits. The predicted force $F_{pred}(u)$ accounts for actuator response over the command interval. Including the actuator admissible set in the equation makes the intersection between mechanical demand and achievable response explicit. The uncertainty allowance increases the required predicted force, while the warning margin preserves operating headroom above the reduced contact boundary. This check addresses the predicted force over the next command interval. Continued feasibility depends on the subsequent demand, estimator error, and actuator response; an accepted command alone does not establish forward invariance of the contact condition.

Feasibility is assessed before projection. An empty interval invokes supervisory deceleration, records the cause, and classifies the episode as infeasible. The supervisor may request maximum available adhesion during deceleration, with the event retained in the infeasibility record. This classification distinguishes completed traversals from routes interrupted by insufficient attachment capability. Intervention frequency, intervention duration, and unsuccessful stops are therefore evaluated alongside energy consumption and attachment reserve.

Figure 2 shows the learning and execution flow. Candidate actions pass through the feasibility check before reaching the plant, and executed transitions return to replay. The reserve estimator feeds both the cost critic and the immediate projection layer, while multiplier adaptation remains within the training branch. A separate dashed path identifies emergency deceleration when the feasible interval is empty. The diagram distinguishes policy improvement from command acceptance and identifies the mechanical checks applied during execution.

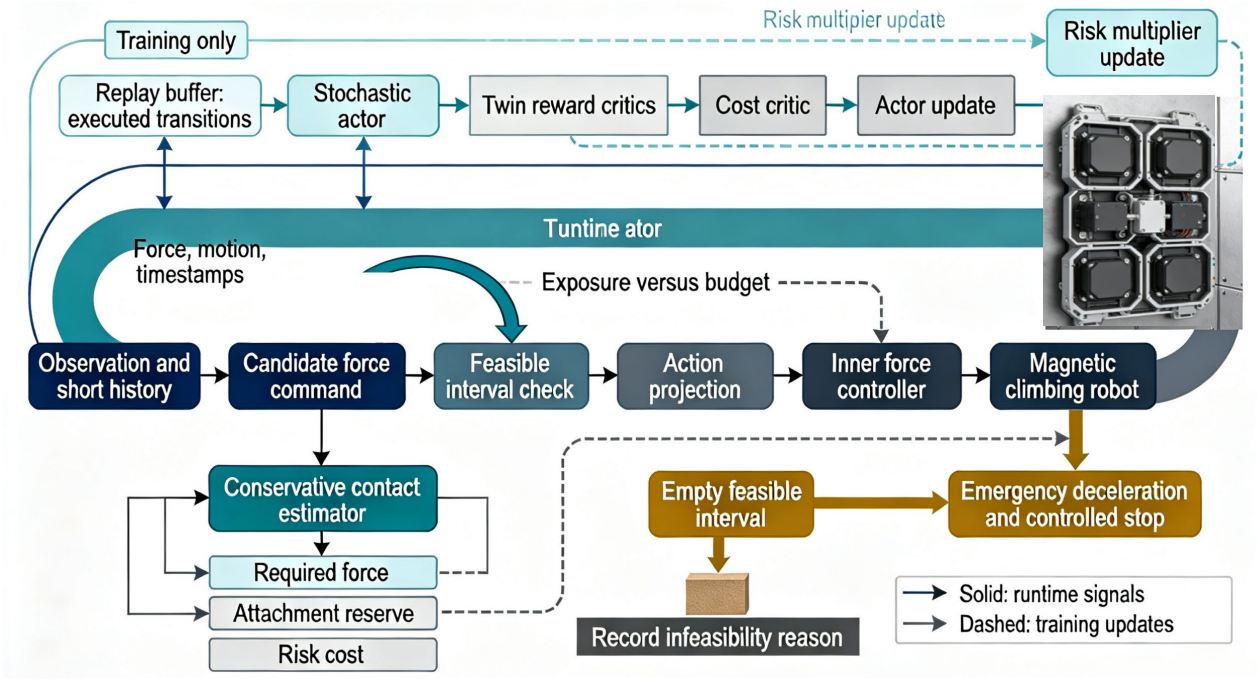



**Figure 2. Constrained actor critic learning and command projection with explicit infeasibility handling**

The deployment policy uses the actor mean to issue a deterministic candidate command. During estimator initialization, a conservative force request remains active until sufficient observation history is available. Sensor dropout, stale timestamps, or inconsistent force readings trigger the supervisory response. Runtime monitoring records observation age and intervention duration because occasional deadline misses can substantially reduce the available reserve. Timing assessment includes acquisition, communication, inference, and command delivery, providing the complete sensing to actuation delay used in feasibility evaluation. The latency analysis in Section 5 examines how this delay changes the attachment reserve available during disturbances.

## 5. Performance Evaluation and Numerical Analysis

### 5.1. Evaluation protocol and condition dependent performance

The numerical evaluation covers five condition categories: smooth steel, a weld transition, a local gap change, reduced friction, and a payload shift. Condition means, transient trajectories, and reserve distributions provide complementary views of controller behavior. Standardized robot control benchmarks have provided common obstacle-crossing tasks and baseline learning algorithms for comparative evaluation [19]. The comparison uses a common denominator of five hundred episodes per method, with one hundred episodes allocated to each condition. An implementation of the training protocol would retain ten independent initialization seeds to distinguish optimization variability from condition variability. Table 1 records the common scenario settings, including a vehicle mass of 12 kg, a route duration of 120 s, and a nominal speed of 0.08 m per second. The five method labels identify fixed force, proportional integral derivative regulation, unconstrained soft actor critic, a penalty variant, and the proposed projected configuration. Wall-climbing trajectory studies have compared predicted and measured climbing times when validating optimized gait trajectories [20]. The condition categories combine surface and load effects, while a factorial extension would resolve their individual contributions. An unsafe episode contains a negative modeled reserve interval, and intervention share uses the number of control updates as its denominator. This separation preserves the distinction between episode outcomes and execution workload. The ten-seed statement specifies a replication protocol; the displayed aggregates do not establish between-seed variability. The 500-episode denominator specifies the reported episode population and should be interpreted separately from the number of independently trained policies.

**Table 1. Operating conditions and evaluation specification**

| Quantity | Value | Interpretation |
|---|---|---|
| Vehicle mass | 12 kg | Vehicle configuration |
| Payload | 0–4 kg | Five-condition envelope |
| Inclination | 70–110 degrees | Modest overhang included |

| Quantity | Value | Interpretation |
|---|---|---|
| Adhesion range | 180–650 N | Assumed available range |
| Force rate limit | 900 N/s | Assumed actuator bound |
| Outer update | 50 Hz | Proposed control rate |
| Route | 120 s at 0.08 m/s | Common nominal task |
| Unsafe event | Any negative modeled reserve | Contact condition boundary |
| Nominal episodes | 500 per method | Episode denominator |

Figure 3(a) compares route energy across the five conditions. The proposed series increases from 9.8 Wh on smooth steel to 12.8 Wh during the payload shift, while the fixed force series increases from 13.4 to 15.9 Wh. The corresponding means are 11.28 and 14.62 Wh, giving a relative reduction of 22.85 percent. The absolute difference narrows from 3.6 Wh on smooth steel to 3.1 Wh under the payload shift. This narrowing expresses the assumed loss of scope for reducing adhesion as mechanical demand rises. Low-power magnetic climbing designs have used permanent magnetic adhesion to reduce the energy demand associated with conventional electromagnetic attachment [21]. The aggregate route values used here combine those contributions, so their interpretation concerns the total task energy of the declared scenario. Figure 3(b) examines thirty force error samples per method, centered at 25, 20, 17, 14, and 11 N. The spread complements the aggregate root mean square errors in Table 2 by exposing variability around the assigned operating levels. For the proposed configuration, the aggregate error is 11.2 N against 17.8 N for the unconstrained configuration, a 37.08 percent reduction. The distribution and aggregate metric describe complementary aspects of tracking performance: the distribution exposes concentrated and extreme errors, while the aggregate metric emphasizes sustained or large deviations from the accepted reference. The thirty displayed absolute-error values describe a separate error sample, whereas the tabulated root mean square metric refers to the aggregate tracking record. Reconstructing the latter requires signed residuals and their sampling weights from that record.

Figure 3(c) displays cumulative margin distributions built from eighty samples per controller. A rightward displacement indicates greater reserve at a given cumulative fraction, while the intersection with zero identifies the portion of the population lying beyond the mechanical boundary. The warning line at 0.20 supplies a second interpretation: reserve can remain positive while providing limited room for delayed actuation. Figure 3(d) presents unsafe episode proportions of 3.6, 5.2, 7.6, 2.4, and 0.8 percent. The proposed and unconstrained counts are four and thirty-eight, producing a relative count reduction of 89.47 percent. Hyperparameter adaptation has been investigated as a consequential part of reinforcement learning performance optimization across benchmark tasks [22]. The ordering is evaluated jointly with energy and intervention burden. The count intervals express the effect of the episode denominator, whereas the cumulative distributions describe the spread of reserve values. Each measure retains its own population and denominator. Table 2 also exposes the incremental energy associated with projection: the proposed configuration uses 0.24 Wh more than the penalty variant, with eight fewer unsafe episodes. The paired metrics show the engineering preference being tested, namely whether a modest energy increase accompanies a meaningful reduction in boundary crossings under matched tasks. These aggregate differences compare complete controller configurations. Isolating the effect of projection requires holding the learned policy, estimator, and evaluation episodes fixed while changing only the execution filter. The displayed count bars use the normal approximation based on each observed episode proportion and a denominator of 500. At four events, this approximation is fragile; its bars indicate binomial sampling uncertainty under independent episodes and exclude policy-training variability.

**Table 2. Aggregate controller performance**

| Method | Energy Wh | RMSE N | Unsafe /500 | Intervention % |
|---|---|---|---|---|
| Fixed | 14.62 | 26.1 | 18 | 0 |
| PID | 13.24 | 21.4 | 26 | 0 |
| SAC | 10.50 | 17.8 | 38 | 0 |
| Penalty | 11.04 | 14.3 | 12 | 0 |
| Proposed | 11.28 | 11.2 | 4 | 20 |

Figure 3(e) examines five air gaps from 0.5 to 2.5 mm. The initial margins are 0.51, 0.46, 0.40, 0.47, and 0.50; the final margins are 0.13, 0.09, 0.04, 0.19, and 0.25. The proposed configuration consequently retains 0.05 reserve above the warning threshold at the largest gap, while the penalty variant falls 0.01 below it. The linear trend places the threshold crossing between neighboring gap settings; curvature and hysteresis in the force map would shift that location. Magnetic adhesion calculations have explicitly specified the magnet–wall air gap and wall material when comparing the force

generated by different magnet grades [23]. An empirical gap sweep should therefore refine the sampling near threshold crossings and distinguish increasing from decreasing gap. Figure 3(f) presents descriptive normalized scores for energy, tracking, reserve, completion, and smoothness. The proposed scores are 0.85, 0.90, 0.90, 0.95, and 0.90, whereas the unconstrained configuration combines a higher energy score of 0.90 with a reserve score of 0.40. This profile visualizes the relative preference among competing criteria. Its interpretation depends on consistent normalization across all five criteria. A high reserve score could otherwise reflect either increased force throughout the route or well-timed adjustment around the demanding segments, two behaviors with different consequences for locomotion and actuator wear. The profile is descriptive because the present record supplies normalized scores without the raw-to-score mappings. Its values support visual comparison within this profile; quantitative ranking should use the reported physical metrics until those mappings and criterion directions are specified.

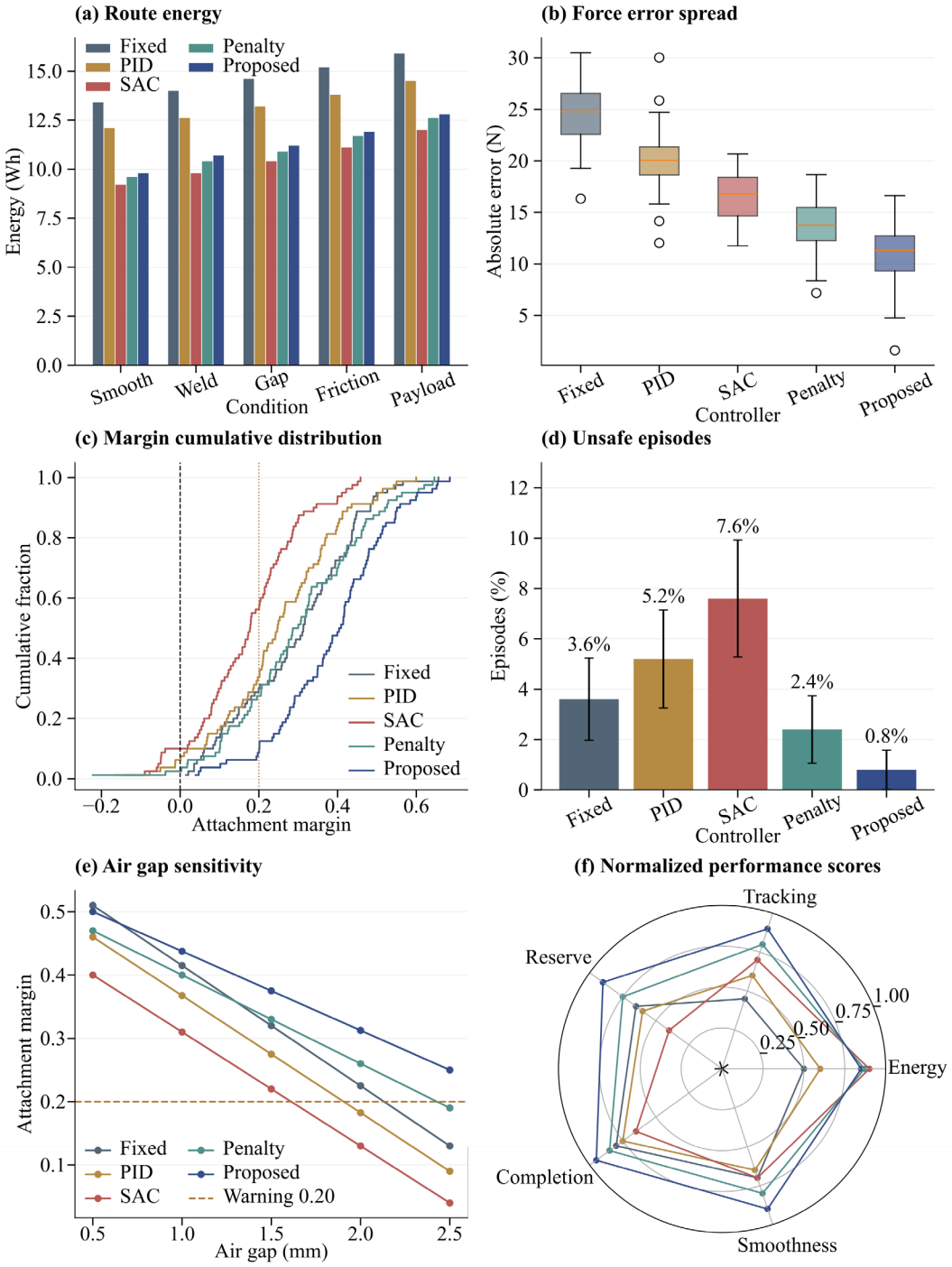



**Figure 3. Condition dependent comparisons: (a) adhesion energy; (b) force error distributions; (c) margin cumulative distributions; (d) unsafe episode proportions; (e) margin sensitivity to air gap; (f) normalized descriptive performance profiles**

### 5.2. Disturbance response and execution burden

The transient scenario contains 241 samples over twelve seconds at 0.05 s spacing, with disturbances centered at four and eight seconds. Figure 4(a) shows a reference that changes from 380 to 450 N and then to 410 N, together with the proposed, unconstrained, and penalty responses. Immediately before the upward reference change, the localized deficit is greatest for the unconstrained response, exposing a difficult combination of rising force demand and reduced available force. After the second event, the reference decreases while a smaller transient deficit remains visible. Evaluations of compliant wall-adhesion modules have reported both force retention and response time across different surface curvatures [24]. This methodological choice is particularly relevant to the asymmetric demands of increasing and decreasing adhesion. Increasing force must provide sufficient reserve before the disturbance reaches its peak, whereas decreasing force can usually proceed more gradually if adequate contact remains. In the transient record, force oscillations continue after each event and therefore contribute to tracking variability even when the reserve remains positive. The samples retain their temporal correlation throughout interpretation. The trajectory describes transient structure, while repeatability assessment requires independently repeated routes and policy initializations.

Figure 4(b) relates the same event times to the reserve histories. The proposed curve reaches approximately 0.18 at the first disturbance, crossing the 0.20 warning threshold while remaining above the modeled violation boundary. The unconstrained curve reaches approximately −0.08, and the penalty curve approaches 0.095. These values distinguish three responses: retaining a positive reserve with a warning excursion, crossing the mechanical boundary, and preserving contact with a deeper reduction in operating headroom. Force and reserve provide complementary response profiles. Their mechanical relationship depends on the uncertainty deduction and contemporaneous tangential and overturning demands. Such computation would show whether a force deficit is caused by an underestimated requirement, delayed actuation, or a surface dependent loss of capability. The second event produces a smaller reserve reduction for the proposed series, reflecting its lower disturbance amplitude. The time required to return above the warning threshold should be measured separately for each event because a similar minimum can arise from either a brief sharp excursion or a prolonged low reserve interval. Recovery duration also determines how long the controller remains exposed to an additional disturbance arriving before the first response has settled.

Figure 4(c) examines correction magnitudes for five disturbance severity levels using thirty-six samples per level. Their specified centers increase from 8 through 14, 19, and 25 to 31 N. The violin widths indicate where the corrections concentrate, and individual points expose the sample density behind the smoothed distributions. Increasing correction magnitude represents greater separation between the candidate and accepted actions, which matters even when the accepted force maintains a positive reserve. A policy that anticipates demand would ideally reduce this separation by increasing force before the feasibility check becomes restrictive. A policy that depends on projection can achieve a similar accepted command while imposing more frequent or larger corrections. Studies of safety-filtered reinforcement learning have compared action filtering, penalties for corrections, and safe resets to assess their effects on performance, safety, and chattering [25]. Applied to execution behavior, that principle requires storing candidate actions, executed actions, intervention duration, and estimator state for the same episode. The correction distributions also motivate a distinction between a single large adjustment and repeated small adjustments. Both can yield comparable cumulative correction magnitude, yet repeated adjustment may create different actuator duty and thermal demands. The spreads remain similar across severity levels, making the shift in operating level the dominant visible feature. Changes in the upper tail would indicate additional sensitivity to disturbance variability. Correction magnitude denotes the absolute difference, in newtons, between candidate and executed aggregate force commands at one update. Its distribution describes adjustment size; cumulative actuator effort additionally depends on update frequency, intervention duration, and the actuator energy model.

Figure 4(d) maps intervention share over five disturbance levels and observation delays of 0, 20, 40, 60, and 80 ms. The values span 4 to 42 percent. At the lowest severity, delay increases intervention share from 4 to 13 percent; at the highest severity, it increases from 19 to 42 percent. The delay related increment therefore grows from nine to twenty-three percentage points across the severity range. This pattern expresses the intended interaction between a faster changing demand and older contact information. Within the middle severity row, the progression from 10 through 14, 18, and 21 to 25 percent shows that even moderate disturbances can consume a substantial part of the execution filter's available workload. Each cell is a control update fraction, while unsafe outcomes remain episode level quantities. An empirical implementation should log the age of the observation used by each intervention and distinguish sensor filtering delay from communication and inference delay. These sources act at different points in the loop and can produce different estimation errors. Persistent high intervention share may also reveal a mismatch between the actor's observation history and the contact estimator used by projection. Inspecting the residual between those two information paths would help determine whether additional training, estimator alignment, or reduced motion demand is the relevant engineering response. A cell therefore describes the percentage of updates requiring command correction in its specified condition. Comparing cells requires the same update-counting rule; comparing them with unsafe proportions requires preserving the distinct update-level and episode-level denominators.

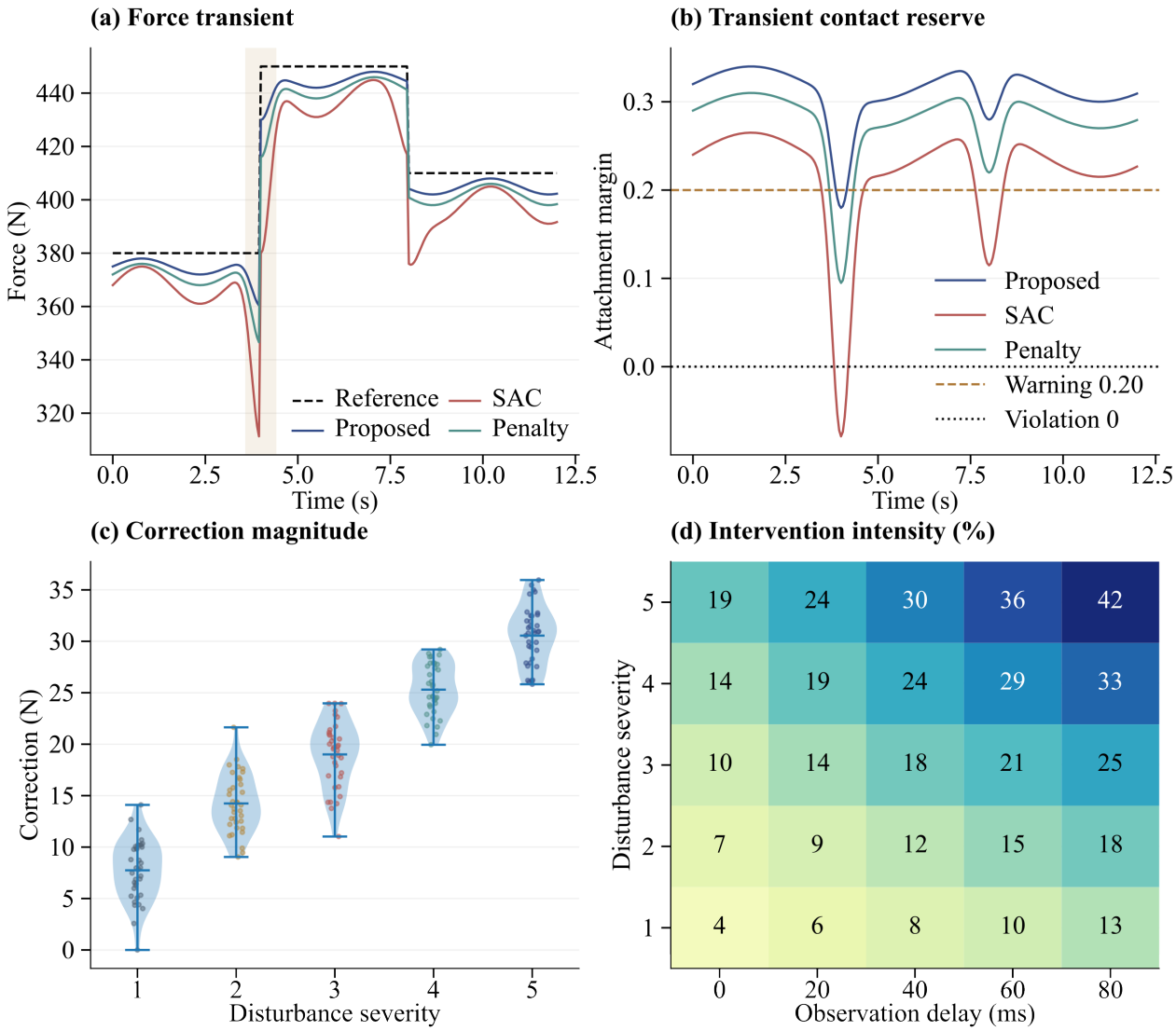



**Figure 4. Disturbance response: (a) force reference and responses; (b) attachment reserve histories; (c) correction magnitude distributions; (d) intervention intensity across disturbance and delay**

### 5.3. Component sensitivity and reserve distributions

Figure 5(a) compares five component settings through unsafe counts of 14 without budget adaptation, 12 without projection, 19 without the uncertainty deduction, 9 without history, and 4 for the full configuration. Each paired dot segment begins at the full configuration reference so that the additional count is directly visible. The uncertainty ablation produces fifteen additional events, compared with ten for the budget ablation, eight for projection, and five for history. The ordering associates optimistic contact estimation with the largest increase in unsafe episodes within the selected envelope. Component effects remain coupled because the estimated reserve enters both the cost signal and the feasible command interval. Removing uncertainty conservatism can therefore affect learning preferences and command acceptance simultaneously. Figure 5(b) places the same configurations in energy and unsafe proportion space. The uncertainty ablation combines 10.3 Wh with 3.8 percent unsafe episodes, while the full configuration combines 11.28 Wh with 0.8 percent. The 0.98 Wh difference describes the energy cost of the more conservative configuration under these assumptions. The history ablation uses 11.1 Wh with 1.8 percent unsafe episodes, locating it closer to the full configuration in energy than in event frequency. Future matched experiments should retain the same raw observations and separate the effects of information removal, network capacity, and altered optimization difficulty. These differences rank the evaluated configurations under the stated envelope. Independent component attribution requires matched observations and training conditions because uncertainty, history, and projection can alter both the learned policy and the accepted command.

Figure 5(c) shows risk budgets of 0.005, 0.010, 0.020, 0.030, 0.040, and 0.060. Route energy decreases from 12.6 to 10.5 Wh, while intervention share decreases from 32 to 8 percent. At the nominal budget of 0.020, energy is 11.28 Wh and intervention share is 20 percent. Tightening the budget to 0.010 adds 0.72 Wh and seven percentage points of intervention; relaxing it to 0.030 removes 0.38 Wh and five percentage points. These unequal increments express a

diminishing energy response across the sweep. Time-consistent risk-sensitive learning has formulated actor-critic updates using dynamic convex risk measures [26]. The warning threshold and unsafe episode definition are consequently held fixed across this comparison. The decrease in interventions with a relaxed budget depends on the resulting candidate actions relative to the fixed execution requirement. If a relaxed actor proposed more low force actions, the projection workload could instead rise. Logging candidate action deficits and multiplier trajectories would resolve that mechanism. The budget sweep therefore evaluates a joint energy and workload hypothesis, with each quantity retaining its own denominator and interpretation. The sweep establishes a joint trend in the displayed energy and workload values. Its mechanism remains unresolved without paired candidate-command deficits, feasible interval widths, and multiplier histories across the budget settings.

Figure 5(d) examines observation latency from zero to 100 ms in 20 ms increments. The central minimum reserves are 0.31, 0.28, 0.24, 0.19, 0.14, and 0.08. The central series crosses the warning level between 40 and 60 ms, while the lower scenario bound reaches 0.20 at 40 ms. This difference shows why a latency decision based on the central response can leave limited allowance for adverse estimation conditions. At 100 ms, the lower bound is 0.01, placing the condition close to the modeled violation boundary. The asymmetric intervals are declared scenario ranges. Their engineering role is to expose how uncertainty can consume the reserve remaining after a delay related loss. Hardware assessment should measure sensing to actuation age at the command actually used by the inner loop, including deadline misses and buffered observations. A processor's average inference time describes only one contribution to that age. The proposed benchmark retains a fixed force capability across the latency sweep, allowing observation age to serve as its designated stressor. A later coupled sweep could vary both delay and force rate, since a delayed demand estimate becomes more difficult to recover when the actuator also approaches its adjustment limit. The crossing brackets a warning transition for this fixed-capability scenario; it does not define a certified maximum allowable delay. Such a limit would require measured timing tails, calibrated force prediction error, and a bound on demand growth during the observation age.

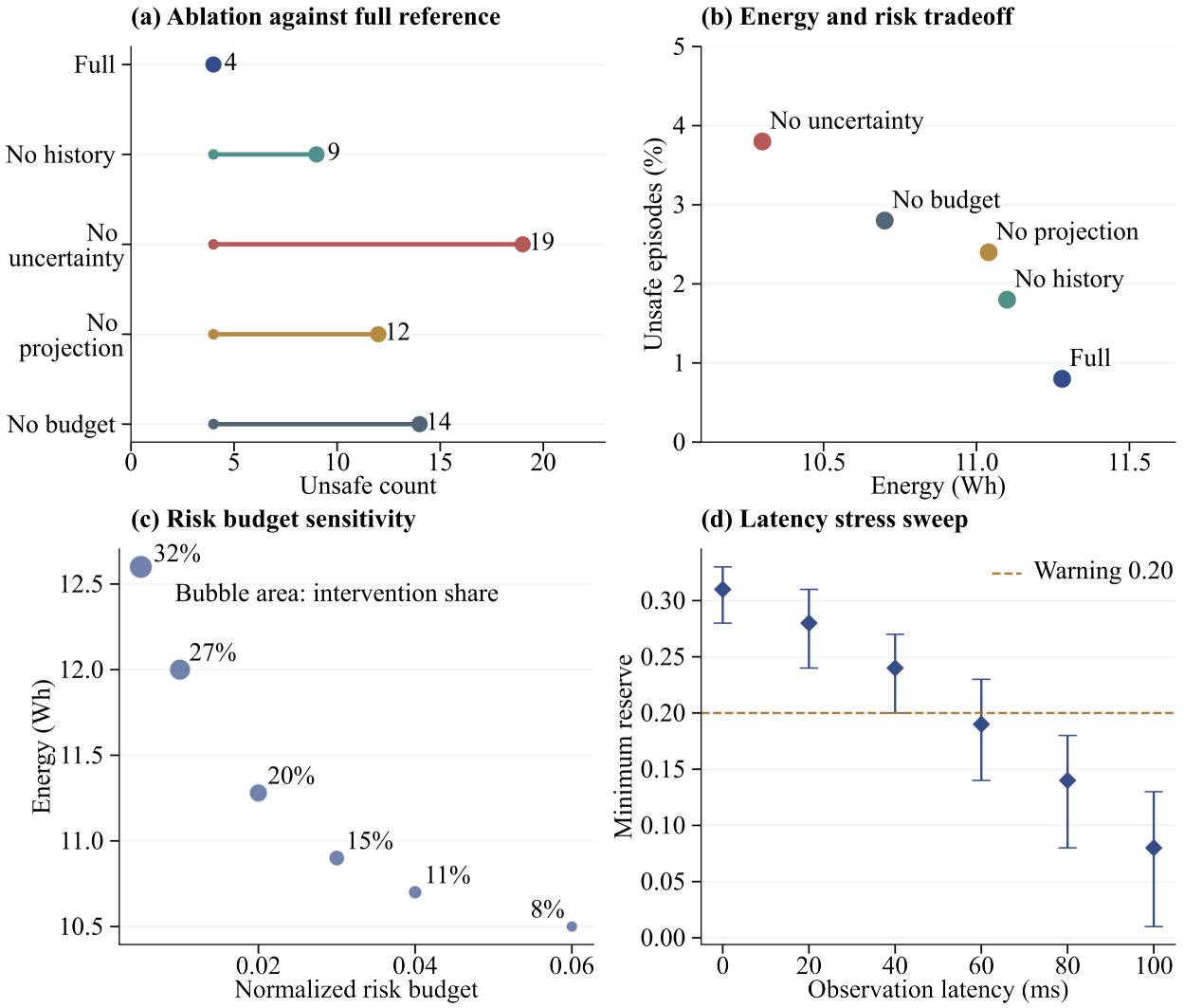



**Figure 5. Ablation and sensitivity studies: (a) paired unsafe counts; (b) energy and unsafe proportion tradeoff; (c) risk budget response; (d) latency and minimum reserve**

Figure 6 presents quantiles from 0.05 to 0.95 for five separate populations of 120-episode minimum reserve values. Risk-conditioned policies have used weighted value-at-risk representations to adapt behavior across different risk preferences [27]. The proposed population is centered at 0.36 with a specified spread of 0.17, compared with centers of 0.23, 0.19, 0.13, and 0.29 for the other labels. Its low quantiles remain higher in the profile, while the unconstrained population extends further into negative reserve. These distributional populations and the five hundred episode count populations support different statistical summaries. The light vertical strokes mark a fixed display range of 0.012 on either side of each quantile estimate; the shaded bands identify the warning and violation regions. Bootstrap procedures for hierarchical data must preserve dependence between data levels to support valid interval estimation [28]. Empirical replication should therefore derive event counts and quantile estimates from a common episode record, bootstrap within each training seed, and retain variation between seeds. Near the lower tail, the number of observations supporting a quantile decreases, making the result sensitive to a small set of difficult contact histories. Those histories should be inspected against individual module forces, payload motion, and air gap estimates to identify local peeling or demand peaks that the aggregate reserve model may conceal. The fixed strokes are display markers rather than estimated confidence intervals. The separate 120-episode populations characterize distribution shape and cannot be pooled directly with the 500-episode populations to estimate a common event rate or a between-method significance level.

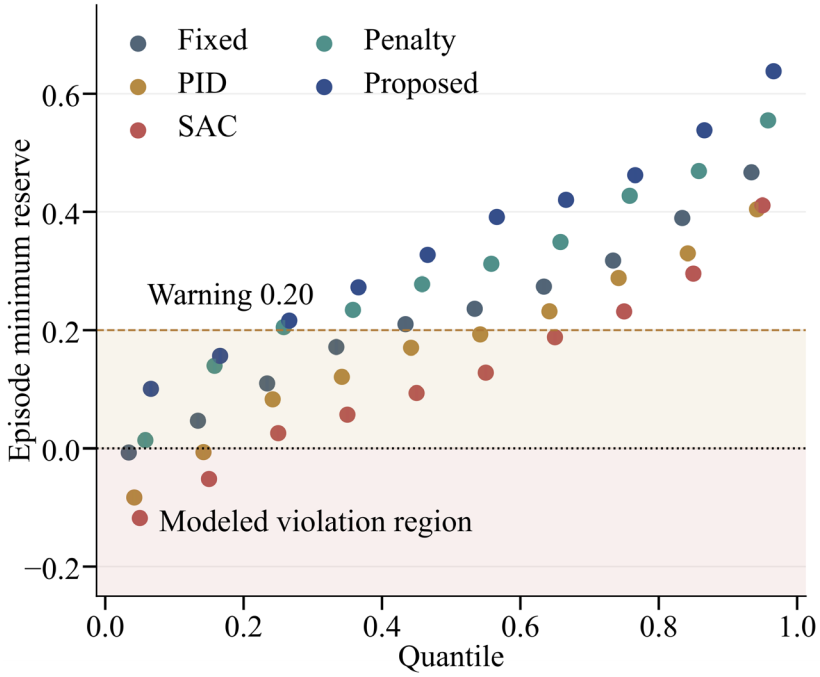



**Figure 6. Distribution of episode minimum attachment reserve**

## 6. Conclusion

This study developed a detachment risk constrained soft actor critic framework that combines conservative attachment assessment, risk budget adaptation, and feasible command projection to balance adhesion reliability and energy consumption in magnetic wall climbing robots.

1. Energy consumption and force tracking. Across the five operating conditions, the proposed configuration achieved a mean route energy of 11.28 Wh, compared with 14.62 Wh for fixed force control, corresponding to a reduction of 22.85%. Its root mean square force tracking error was 11.2 N, representing a 37.08% reduction relative to the unconstrained soft actor critic configuration.

2. Risk exposure and operating tradeoffs. The unsafe episode count decreased from 38 to 4 over 500 episodes relative to unconstrained soft actor critic, corresponding to an 89.47% reduction. Compared with the penalty configuration, the proposed configuration recorded 4 rather than 12 unsafe episodes and used 0.24 Wh more route energy. Relaxing the risk budget reduced energy consumption and intervention frequency, whereas increasing observation latency progressively depleted the minimum attachment reserve. These relationships identify risk budget selection and sensing delay as consequential factors in adhesion regulation.

3. Contribution and applicability. The framework connects a mechanically interpretable attachment reserve to both policy optimization and command acceptance. Its applicability depends on the accuracy of friction estimation, actuator response prediction, and the aggregate contact representation. Local module unloading, uneven support, and surface dependent magnetic capability require additional contact resolution and capability calibration. The reported comparisons support numerical assessment within the declared envelope. Application readiness additionally requires a calibrated

module capability map, independently repeated controller training, and hardware tests that record local contact forces and complete sensing-to-actuation timing.

Future work should integrate module level force allocation and delay aware reserve prediction, followed by independent controller training and hardware evaluation across representative inspection surfaces.

## 7. Declarations

### 7.1 Author Contributions

Živojka Janković contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. Branka Čolaković contributes to validation, analysis, investigation, data collection, draft preparation. Čedomirka Radošević contributes to conceptualization, validation, analysis. All authors have read and agreed with the manuscript before its submission and publication.

### 7.2 Funding and Acknowledgments

This research received no specific financial support from any funding agency.

### 7.3 Institutional Review Board Statement

Not applicable.

### 7.4 Informed Consent Statement

Not applicable.

### 7.5 Date Availability

Not applicable.

### 7.6 Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### 7.7 Declaration of generative AI and AI-assisted technologies

During the preparation of this manuscript, the authors utilized Grammarly for linguistic polishing, logical organization and draft revision. The generative AI tool was not involved in research design, data collection, data analysis, or the equation of core conclusions. All outputs generated by AI were reviewed, revised and verified manually by the authors. The authors take full responsibility for all content of this manuscript.

## 8. References

[1] Kong, J., Ji, A., Han, Q., Shen, H., Liu, S., Xiang, W., & Zhang, Q. (2025). Advances in research of wall-climbing robots: From biology to bionics—A review. Journal of Bionic Engineering, 22(3), 945–981. https://doi.org/10.1007/s42235-025-00689-6

[2] Yang, P., Sun, L., Zhang, M., & Chen, H. (2024). A lightweight optimal design method for magnetic adhesion module of wall-climbing robot based on surrogate model and DBO algorithm. Journal of Mechanical Science and Technology, 38(4), 2041–2053. https://doi.org/10.1007/s12206-024-0334-3

[3] Yun, S., Um, Y., Park, H.-W., & Kim, S. (2024). Static friction of magneto-rheological elastomer pads in wall-climbing robots. CIRP Annals, 73(1), 441–444. https://doi.org/10.1016/j.cirp.2024.04.100

[4] Wang, G., Li, W., & Che, H. (2024). Design and analysis of anti-overturning mechanism for magnetic wall-climbing robot. Journal of Mechanical Science and Technology, 38(1), 379–387. https://doi.org/10.1007/s12206-023-1231-x

[5] Duan, Z., Shi, G., Deng, C., & Yu, C. (2025). Design and implementation of a split-flexible wall-climbing robot with transition ability. Journal of Mechanical Science and Technology, 39(4), 2163–2180. https://doi.org/10.1007/s12206-025-0339-6

[6] Tian, Y., Jitsukawa, H., Ma, S., & Zhang, G. (2024). Magnetic wall-climbing wheels with controllable adhesion reduction via soft magnetic material. IEEE Robotics and Automation Letters, 9(12), 11553–11560. https://doi.org/10.1109/LRA.2024.3496315

[7] Tang, C., Abbatematteo, B., Hu, J., Chandra, R., Martín-Martín, R., & Stone, P. (2025). Deep reinforcement learning for robotics: A survey of real-world successes. Annual Review of Control, Robotics, and Autonomous Systems, 8(1), 153–188. https://doi.org/10.1146/annurev-control-030323-022510

[8] Hwang, H.-S., Kim, Y., & Seok, J. (2025). Generative adversarial soft actor–critic. IEEE Transactions on Neural Networks and Learning Systems, 36(7), 11917–11927. https://doi.org/10.1109/TNNLS.2024.3493113

[9] Gu, S., Yang, L., Du, Y., Chen, G., Walter, F., Wang, J., & Knoll, A. (2024). A review of safe reinforcement learning: Methods, theories, and applications. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(12), 11216–11235. https://doi.org/10.1109/TPAMI.2024.3457538

[10] Hsu, K.-C., Hu, H., & Fisac, J. F. (2024). The safety filter: A unified view of safety-critical control in autonomous systems. Annual Review of Control, Robotics, and Autonomous Systems, 7(1), 47–72. https://doi.org/10.1146/annurev-control-071723-102940

[11] Wang, B., Li, P., Li, P., Zhang, L., Guan, E., Liu, X., Hu, X., & Zhao, Y. (2025). Development of a wheeled wall-climbing robot with an internal corner wall adaptive magnetic adhesion mechanism. Journal of Field Robotics, 42(1), 97–114. https://doi.org/10.1002/rob.22402

[12] Jia, P., Jia, J., Li, M., Zhang, M., & Zhao, J. (2025). Development of an omnidirectional mobile passive-compliant magnetic-wheeled wall-climbing robot for variable curvature facades. Journal of Field Robotics, 42(7), 3877–3901. https://doi.org/10.1002/rob.22601

[13] Xiao, L., Lyu, Y., Zhang, F., Chen, L., Yu, G., Li, S. E., Ma, F., & Duan, J. (2025). Multi-style distributional soft actor-critic: Learning a unified policy for diverse control behaviors. IEEE Transactions on Intelligent Vehicles, 10(3), 1759–1770. https://doi.org/10.1109/TIV.2024.3432891

[14] Duan, J., Wang, W., Xiao, L., Gao, J., Li, S. E., Liu, C., Zhang, Y.-Q., Cheng, B., & Li, K. (2025). Distributional soft actor-critic with three refinements. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(5), 3935–3946. https://doi.org/10.1109/TPAMI.2025.3537087

[15] Baumann, D., & Schön, T. B. (2024). Safe reinforcement learning in uncertain contexts. IEEE Transactions on Robotics, 40, 1828–1841. https://doi.org/10.1109/TRO.2024.3354176

[16] Chen, W., Subramanian, D., & Paternain, S. (2024). Probabilistic constraint for safety-critical reinforcement learning. IEEE Transactions on Automatic Control, 69(10), 6789–6804. https://doi.org/10.1109/TAC.2024.3379246

[17] Modares, A., Sadati, N., Esmaeili, B., Adib Yaghmaie, F., & Modares, H. (2024). Safe reinforcement learning via a model-free safety certifier. IEEE Transactions on Neural Networks and Learning Systems, 35(3), 3302–3311. https://doi.org/10.1109/TNNLS.2023.3264815

[18] Zhang, Q., Leng, S., Ma, X., Liu, Q., Wang, X., Liang, B., Liu, Y., & Yang, J. (2025). CVaR-constrained policy optimization for safe reinforcement learning. IEEE Transactions on Neural Networks and Learning Systems, 36(1), 830–841. https://doi.org/10.1109/TNNLS.2023.3331304

[19] Zhang, H., Ren, J., Xiao, J., Pan, H., Lu, H., & Xu, X. (2025). FTR-Bench: Benchmarking deep reinforcement learning for flipper-track robot control. Journal of Field Robotics, 42(5), 2375–2389. https://doi.org/10.1002/rob.22528

[20] Li, J., Shi, X., Liang, P., Li, Y., Lv, Y., Zhong, M., & Han, Z. (2024). Research on gait trajectory planning of wall-climbing robot based on improved PSO algorithm. Journal of Bionic Engineering, 21(4), 1747–1760. https://doi.org/10.1007/s42235-024-00538-y

[21] Li, W., Lu, T., Ren, H., Luo, J., Wu, L., Tan, C., & Hu, M. (2025). Design and analysis of a quadruped wall climbing robot with low power consumption. Industrial Robot: The International Journal of Robotics Research and Application, 52(6), 821–832. https://doi.org/10.1108/IR-12-2024-0554

[22] Bai, H., & Cheng, R. (2024). Generalized population-based training for hyperparameter optimization in reinforcement learning. IEEE Transactions on Emerging Topics in Computational Intelligence, 8(5), 3450–3462. https://doi.org/10.1109/TETCI.2024.3389777

[23] Zhong, H., Wang, D., Gong, J., Ban, C., Wei, W., & Long, Y. (2024). Design and force performance analysis of a climbing robot based on Halbach magnetic array. Jordan Journal of Mechanical and Industrial Engineering, 18(4), 811–822. https://doi.org/10.59038/jjmie/180413

[24] Yin, Y., Li, Z., Zhang, W., Zhang, Y., & Xu, Q. (2025). Design of a flexible passive adaptive suction module for wall-climbing robots. IEEE Robotics and Automation Letters, 10(12), 12429–12436. https://doi.org/10.1109/LRA.2025.3621974

[25] Pizarro Bejarano, F., Brunke, L., & Schoellig, A. P. (2025). Safety filtering while training: Improving the performance and sample efficiency of reinforcement learning agents. IEEE Robotics and Automation Letters, 10(1), 788–795. https://doi.org/10.1109/LRA.2024.3512374

[26] Coache, A., & Jaimungal, S. (2024). Reinforcement learning with dynamic convex risk measures. Mathematical Finance, 34(2), 557–587. https://doi.org/10.1111/mafi.12388

[27] Yoo, G., Park, J., & Woo, H. (2024). Risk-conditioned reinforcement learning: A generalized approach for adapting to varying risk measures. Proceedings of the AAAI Conference on Artificial Intelligence, 38(15), 16513–16521. https://doi.org/10.1609/aaai.v38i15.29589

[28] Fatoretto, M. B., Demétrio, C. G. B., Brophy, C., & de Andrade Moral, R. (2026). Bootstrap confidence intervals for generalized linear mixed models in dose-response analysis. Journal of Statistical Computation and Simulation, 96(14), 3630–3653. https://doi.org/10.1080/00949655.2026.2677023