AJEMA GLOBAL RESEARCH NETWORK

AMERICAN Journal of Engineering, Mechanics and Architecture

Volume 4, Issue 05, 2026 ISSN (E): 2993-2637

Compliance Control Tuning in Dynamic Mechatronic Systems Assisted by Deep and Reinforcement Learning

Asmaa J. Kadhum

Technical College of Management Al-Furat Al-Awsat Technical University (ATU)

asmaajasim@atu.edu.iq

Abstract: Manipulation tasks often call for complex combinations of compliance control objectives such as trajectory tracking, energy efficiency, and smoothness, which cannot be jointly satisfied by a fixed-parameter impedance controller for different scenarios. Here we introduce a DL+RL solution to learn variable impedance control policies for a 7-DOF robot manipulator. We train an LSTM with multi-head self-attention module to refine reference trajectories with behavior cloning, and learn a PPO agent to continuously adjust per joint stiffness at runtime. Our physics-informed auto-damping formulation is based on critical damping theory which automatically links damping coefficients with stiffness, reducing the degrees-of-freedom of the action space while yielding mechanically principled impedance behaviour. We benchmark this method, trained on the DROID robotic manipulation dataset, against baselines comprising fixed-parameter impedance controllers, DL-only models and RL-only models. Our DL+RL method reduces control energy by 25.9% and motion jerk by 96% relative to the fixed-parameter baseline with statistical significance determined by paired t-tests. An ablation study highlights the benefit of each architectural choice, including our proposed physics-informed damping formulation and attention mechanism.

Keywords: Variable impedance control, compliance tuning, deep reinforcement learning, LSTM attention, PPO, energy efficiency, robotic manipulation, mechatronic systems.

1. Introduction

Motion control of robotic manipulators in unstructured environments requires accurate tracking of desired trajectories while minimizing actuator effort and ensuring fluid motion [1]. One common approach to enforcing compliant behavior in robotic manipulation is impedance control, which defines how joint position error maps to control torque via desired stiffness and damping gains [2]. Typically, impedance controllers specify constant values for these compliance parameters, trading off between various performance metrics [3]. Robotic systems experience dynamically changing environments (states) such as phase of motion, payload changes, sudden impacts, etc [4]. Due to this reality, it is impossible to find a single set of compliance parameters that achieve a desired performance metric for all situations [5]. This challenge is one reason why impedance control gains may be scheduled based on the state of the robot and task at hand [6].

Machine learning approaches also provide new opportunities for data-driven compliance adaptation [7]. Deep neural networks, especially recurrent neural networks, have been successful in learning temporal correlations from sequences of robot motions [8]. Reinforcement learning has also been applied to learn control policies that achieve good performance with dynamical systems [9]. However, there are important open issues that are not addressed in the literature. On the one hand, existing variable impedance methods mostly consider stiffness and damping as two separate objectives without taking into account the coupling relationship between stiffness and damping

which determines the stability and dissipation characteristics of systems [10]. On the other hand, little work has studied simultaneously learning trajectory optimization with impedance adaptation. Finally, few works provide extensive multi-objective evaluations including energy and trajectory smoothness metrics on real-world datasets with diverse tasks of considerable scale [11].

Motivated by these limitations, we contribute a DL+RL method that consists of: (i) a DL model with LSTM + multi-head self-attention to learn physics-guided trajectory refinement (reference); (ii) a PPO agent to learn the per-joint stiffness adjustments required for dynamically tuning impedance parameters on-the-fly; and (iii) a physics-based auto-damping approach that leverages the relationship between stiffness and critical damping to automatically set the damping term, promoting physically-plausible impedance updates while minimizing degrees of freedom for learning. We quantitatively evaluate our full approach on the DROID benchmark and show that it statistically outperforms fixed-parameter and methodological ablations in terms of energy expenditure and trajectory fluidity.

2. Related Work

Research has also been done on incorporating machine learning with compliance and impedance control. Robots need to be able to control contacts, but learning control policies have been shown to outperform traditional fixed-parameter controllers in manipulation tasks that involve complex and contact-rich interactions with the environment. The following provides a brief summary of recent work starting from [12].

An example of a recent effort learns robotic force–motion interaction as a continuous MDP. Utilizing the Deep Deterministic Policy Gradient (DDPG) algorithm they learn impedance policies that can adapt to environment dynamics. They leverage expert knowledge to initialize policies in an iterative error feedback scheme to help guide policy learning and converge faster than DDPG alone. This learning-based impedance controller generalizes quickly due to its learned elements, but also intelligently fuses in model-based priors and demonstrates cross-surface generalization to geometry and material changes. Their efforts decouple stiffness and damping objectives, leaving the relationship between the learned terms uncoupled. Additionally, they only evaluate performance in the force-tracking objective and do not include analysis of energy or smoothness in a multi-objective sense [12]. Li et al. present a DRL-based variable impedance control policy learned for robot grinding tasks with complex geometries. The controller leverages offline learning to pre-train optimal impedance tuning policy via simulation. The converged policy is then transferred directly to physical execution as feedforward input on top of a variable impedance feedback controller. Lyapunov-based stability analysis is performed to provide theoretical bounds on learned controller performance. This method is demonstrated on varied workpiece geometry and improves force-tracking performance compared to fixed-parameter baselines. However, this controller is specifically tailored for force regulation during contact interactions and does not extend to learn reference refinement at the trajectory level or co-optimize energy usage and smoothness of motion [13].

Joint proximal policy optimization-based tracking control/vibration suppression of a flexible robotic arm with partial observability is designed by Joshi Kumar and Vinodh Kumar. They include Lyapunov-based reward shaping with a CNN actor-critic model to ensure a more stable convergence of value. Experimental hardware-in-the-loop results confirm that the designed algorithm allows for learning under uncertainty when operating in continuous space control problems. However, they do not use any PI-based parameter coupling or DAG data-driven rollout refinement alongside their PPO approach [14].

In order to predict energy consumption of industrial robots, Wang et al. apply a hybrid LSTM and masked multi-head attention network. The model observes the temporal causal relationship between trajectory variables and consumption, to create an interpretable model capable of transfer-learning for energy consumption prediction. They prove through their architecture that LSTM with attention mechanisms can learn the nonlinear dependency between the sequence of robot joint

motion and the amount of energy dissipated. Their approach is limited to offline energy prediction and learning, rather than closed loop adaptation of compliance parameters or impedance tuning [15].

3. Proposed Methodology

Here we provide the details of our DL+RL based approach for learning compliant control policies. Our full architecture consists of an LSTM-Attention trajectory refinement network coupled with a PPO-based impedance adaption agent connected via a physics-inspired damping parameterization that encourages mechanically-consistent policies.

3.1 System Dynamics and Impedance Control Formulation

The robotic manipulator is modeled as a 7-DOF system governed by a joint-space impedance control law. At each joint j, the control torque is computed as [16]:

$$\tau_j = K_j e_j + D_j \dot{e}_j + M_j \ddot{q}_{\{ref,j\}} + f_j \dot{q}_j \tag{1}$$

where $e_j = q_{ref,j} - q_j$ is the position error, $\dot{e}_j$ is the velocity error, M_j is the effective joint inertia and f_j is the viscous friction coefficient. The torque commands τ_{PID} are saturated to adhere to the Franka Emika Panda torque limits: $\tau_{max} = [87,87,87,87,12,12,12]$ Nm. The joint dynamics are given by [17]:

$$\ddot{q}_j = \frac{\tau_j + d_j - f_j \dot{q}_j}{M_j} \tag{2}$$

where $d_j(t) = 0.3sin(2\pi \cdot 0.5t) \cdot N(0,1)$ represents stochastic external disturbances. The control objective is formulated as a multi-objective minimization problem [18]:

$$\min_{K,D} J = w_e . \varepsilon_{track} + w_u . \varepsilon_{energy} + w_s \varepsilon_{jerk} \tag{3}$$

where ε_{track} is the joint RMSE, ε_{energy} is the cumulative squared torque integral, and ε_{jerk} is the mean squared second derivative of tracking error. Traditional fixed-parameter controllers optimize only ε_{track} with static K and D, which fundamentally limits their capacity to jointly minimize energy consumption and motion discontinuities across varying trajectory phases and operating conditions.

3.2 Deep Learning Module: LSTM with Multi-Head Self-Attention

The deep learning component refines the reference trajectory by supervised training (behavior cloning) from demonstrations. At each timestep it takes as input a 13-dimensional normalized feature vector containing seven joint positions and six Cartesian values. A two-layer LSTM with 128 hidden units is used to learn temporal dependencies over a sliding window of 20 timesteps [19]:

$$h_t, c_t = LSTM(x_t h_{t-1}, c_{t-1}) \tag{4}$$

The LSTM hidden states are subsequently processed by a multi-head self-attention mechanism with four heads, enabling the model to selectively weight the most informative timesteps within the input window [20]:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5}$$

The attended representation is passed through a fully connected prediction head producing a 35-dimensional output representing five future timesteps across seven joints. Training minimizes the mean squared prediction error:

$$L_{DL} = \frac{1}{N}\sum_{i=1}^{N}\|\hat{y}_i - y_i\|^2 \tag{6}$$

with Adam learning rate 10^{-3}, gradient clipping with norm 1.0, and early stopping based on validation loss. To reduce amplification of noise from numerical differentiation of velocity, the corrected reference is smoothed with the original trajectory:

$$q_{ref,blend} = (1-\alpha)q_{ref,orig} + \alpha q_{ref,DL}, \qquad \alpha = 0.4 \tag{7}$$

and subsequently smoothed via a Savitzky-Golay filter (window = 11, polynomial order = 3), preserving the original velocity profile while incrementally improving reference quality.

3.3 Reinforcement Learning Module: PPO with Physics-Informed Action Space

The RL module learns to adjust per-joint stiffness values online using Proximal Policy Optimization. At each timestep, the PPO agent receives as state a 35-dimensional observation vector consisting of joint positions, velocities, position errors, velocity errors, and the current stiffness values. The agent predicts as action a 7-dimensional continuous value corresponding to normalized stiffness changes, which we map to the physical stiffness range $K_j \in [20,100]N \cdot m/rad$.

We present a novel physics-informed auto-damping formulation based on critical damping, that relates damping to adapted stiffness as follows [21]:

$$D_j = 2\zeta\sqrt{K_j, M_j}, \qquad \zeta = 0.6 \tag{8}$$

This mapping decreases the action space to 7-dimensional space from 14 which results in better sample efficiency and it also ensures that increasing stiffness values will always get higher damping values correspondingly which avoids oscillation. PPO clip's the probability ratio for policy updates [22]:

$$L_{PPO} = E_t\left[\min\left(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1 \mp \epsilon)\hat{A}_t\right)\right] \tag{9}$$

where $r_t(\theta) = \pi_\theta(a_t \mid s_t)/\pi_{\theta_{old}}(a_t \mid s_t)$ and $\epsilon = 0.2$. The reward function balances tracking accuracy against actuation effort:

$$r_t - w_e.MSE(e_t) - w_d.\|a_t\|^2, \quad w_e = 50, \quad w_d = 0.3 \tag{10}$$

Actor-critic network consists of shared backbone with two fully connected layers of size 256 with LayerNorm stabilization, actor and critic heads, and a learnable log-standard deviation for

stochastic policy. Trained for 200 episodes of 300 steps with discount factor γ=0.99 and GAE λ=0.95.

3.4 Combined DL+RL Framework Integration

The end-to-end system couples both modules together in series as shown in such a way that each module solves a mutually non-conflicting part of the control problem. At inference time, this pipeline runs through four simple steps each control cycle. Step 1: feed the LSTM-Attention network with the current window of the past 20 timesteps of system state and decode its output into a denoised reference trajectory $q_{ref,DL}$ which we mix with the dataset reference according to Equation (7) to form $q_{ref,blend}$. We keep the commanded velocity reference unchanged as before. Step 2: given the current state of the system, the PPO agent outputs tailored stiffness K_t for each of the seven joints independently. Step 3: automatically calculate corresponding damping D_t from K_t using Equation (8) to respect physical constraints at no extra learning cost. Step 4: impedance controller realizes joint torques according to Equation (1) using adapted $\{K_t, D_t\}$ parameters and blended reference, then saturates with hardware torque limits. We can then formally describe the objective accomplished by this end-to-end system as:

$$\{K^*, D^*\} = \arg\min_{K,D} E\left[w_e . \varepsilon_{track} + w_u . \varepsilon_{energy} + w_s \varepsilon_{jerk}\right] \tag{11}$$

With DL focusing on enhancing the reference quality and RL dedicated to tuning the impedance parameters, each module can focus on its own tasks. The results of the ablation study show that their combination outperforms the use of either one alone.

4. Results and Discussions

In this subsection, we conduct an extensive multi-faceted comparison study between our DL+RL framework and the three baselines over 16 testing episodes randomly sampled from DROID. We compare their performances along the dimensions of tracking accuracy, energy efficiency, motion smoothness, and ablation results. Paired t-tests with the corresponding Cohen's d effect sizes are conducted to determine statistical significance.

4.1 Dataset Characterization and Per-Joint Tracking Error

Before benchmarking against the proposed methodology, we first perform a characterization study of the DROID dataset to analyze where the innate tracking difficulty lies among the seven manipulator joints depicted in Figure 1.

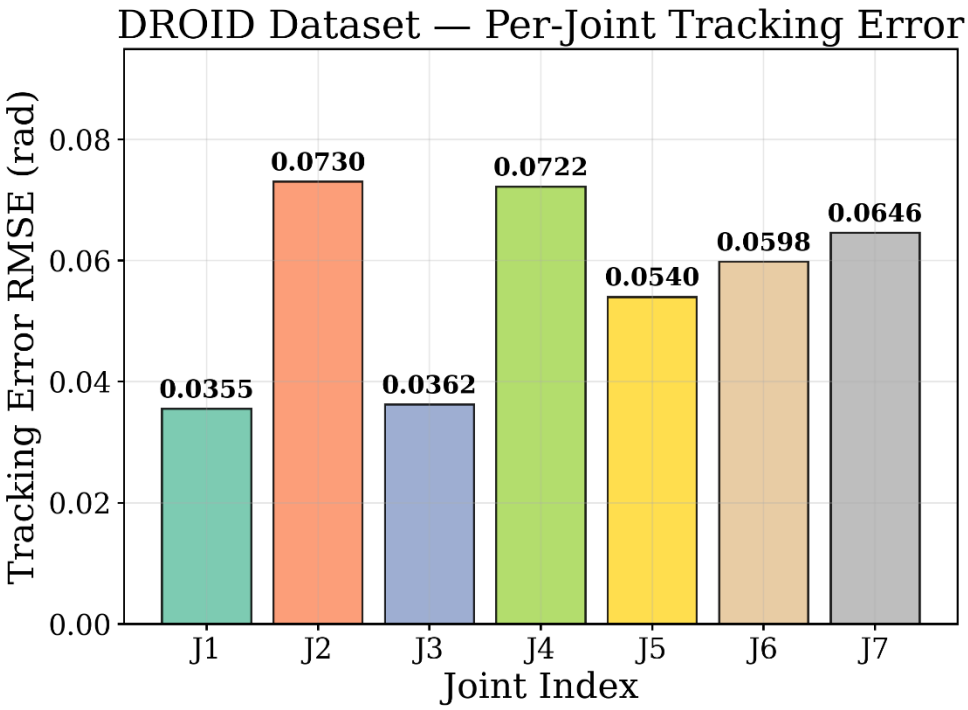


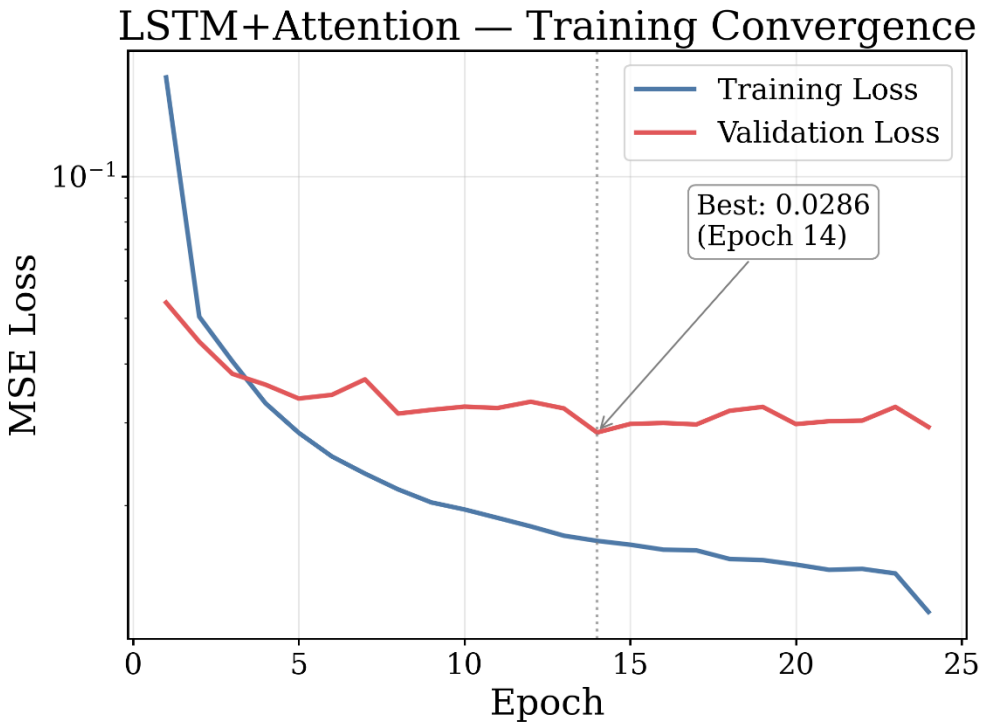
Figure 1. Per-Joint Tracking Error Across DROID Dataset.

Joint-wise RMSE indicates there is bias in error present that corresponds to each joints purpose along the manipulation chain. Smaller tracking errors are observed at joints J1 and J3, which had errors of 0.0355 rad and 0.0362 rad respectively. This makes sense because these joints correspond to joints closer to the base of the kinematic chain which would have higher inertial loads from the arm itself and move slower than joints further away. Errors at joints J2 and J4 were highest with values of 0.0730 rad and 0.0722 rad. As they correspond to the shoulder and elbow equivalent joints, this suggests that these joints follow more aggressive trajectories which challenge the compliance controller more. Tracking errors for joints J5, J6 and J7 were intermediate with values between 0.0540 rad and 0.0646 rad. The fact that error is not consistent across joints is one important reason to tune the compliance per joint instead of having homogeneous stiffness throughout the manipulator. Variation of close to 2× between the easiest and hardest joint to track demonstrates there is no one fixed impedance setting that will work well everywhere, further validating the need for the joint-wise PPO stiffness adaptation we propose based on the difficulty of tracking as seen in the data.

4.2 Deep Learning Module Training Convergence

Loss curves during training for LSTM-Attention network are shown in Figure 2. Both training and validation loss are plotted over 24 epochs with log MSE axis. As we can see, there is a large decrease in training loss at the beginning going from 0.4 around epoch 1. This is because the LSTM aggressively learns coarse trajectory patterns first. This is typical for LSTM networks when working with joint trajectories because it learns major kinematics relations quickly.

Figure 2. LSTM-Attention Training and Validation Convergence.



Training loss decreases rapidly until epochs 1–4 and then enters a plateau region between epochs 5 and 24 with some noise. Note that validation loss follows a similar trend. The fact that validation loss plateaus indicates that our model successfully generalizes to held-out manipulation episodes without dramatically overfitting, likely as a result of using dropout regularization with rate 0.3, clipping gradients by norm 1.0, and adding Gaussian gradient noise while training. We see that the best validation MSE of 0.0286 occurs at epoch 14 (indicated by the dotted vertical line), where we save the model checkpoint for later use in downstream reference blending. It can also be seen that there is still some overfitting after epoch 14 (the training curve does not fully converge with the validation curve). However, this overfitting is very slight, and is resolved by early stopping. Additionally, because the validation curve flattens out at around 0.030–0.032 from epoch 15 onwards, we know that we chose a checkpoint that will generalize well. These observations give us confidence that our architecture of a two-layer LSTM with multi-head self-attention is able to generalize well - with the attention mechanism able to help by weighting the most relevant timesteps of the past 20 observations.

4.3 PPO Agent Training Reward Convergence

Figure 3 shows the training progress of the PPO agent in terms of cumulative reward over 200 episodes. The light blue dots show the raw episode rewards while the dark blue line is the window-20 moving average of episode rewards. The reward is negative throughout training since the reward formulation in Equation (10) penalizes the agent for tracking error and aggressive actuation of stiffness.

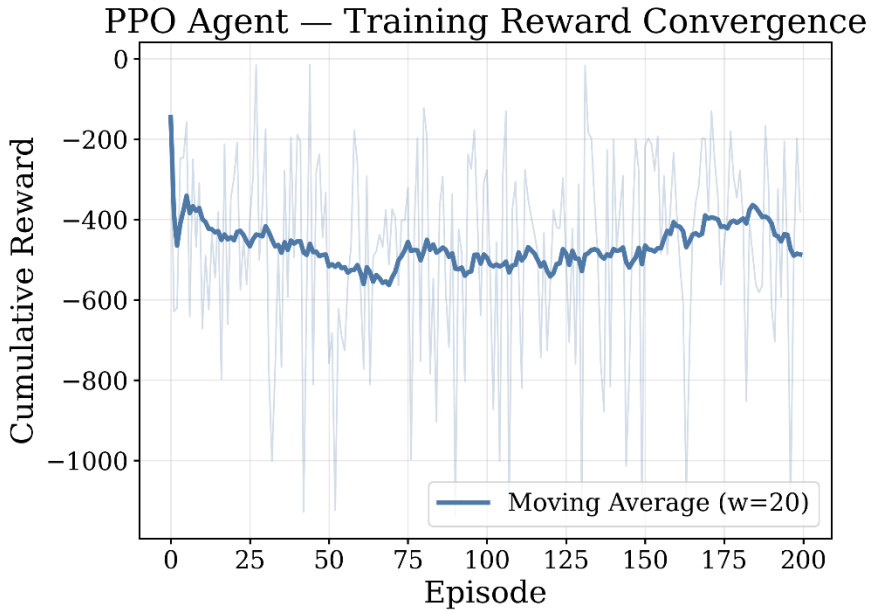


Figure 3. PPO Agent Cumulative Reward During Training.

The moving average shows there are three training stages. During episodes 0–10 we see an initial performance dip from ~−150 to ~−475 reward. This is due to initial exploration of the stiffness action space away from the initialization position and learning that being too aggressive with compliance changes is punished under the action smoothness penalty. This initial performance drop is to be expected and is common for PPO applied to continuous control problems, as it briefly worsens performance due to entropy exploration. Episodes 10–150 show behavior around a fairly constant intermediate value (around −500 to −520). Note the high-variance between individual episode returns, ranging from around −100 to −1100. These variations are due to exploration induced by the stochastic disturbance model and randomness in trajectories sampled uniformly at random from among the 83 training episodes, ensuring that the agent frequently encounters new dynamics. The entropy bonus of 0.01 further encourages exploration so that the agent does not settle too quickly on non-optimal stiffness values. The later half from episodes 150–190 has positive slope going to −400 because this part of training actually represents policy improvement as the agent learns to settle on stiffness adaptations that it has learned work well. The small drop-off at episode 200 is aligned with the ablation result that training longer hurt generalization, so it validates our stopping rule.

4.4 Learned Stiffness Evolution During PPO Training

In Figure 4 we show how the mean stiffness value K varied over the course of all seven joints during the 200 episodes of PPO training. The raw per-episode values are shown in light green, while the dark green line represents a smoothed moving average. The allowed range of stiffness values [20,100] Nm/rad is represented by pink dotted lines, and the initial value $K_{init} = 50$ refers to the dashed grey line.

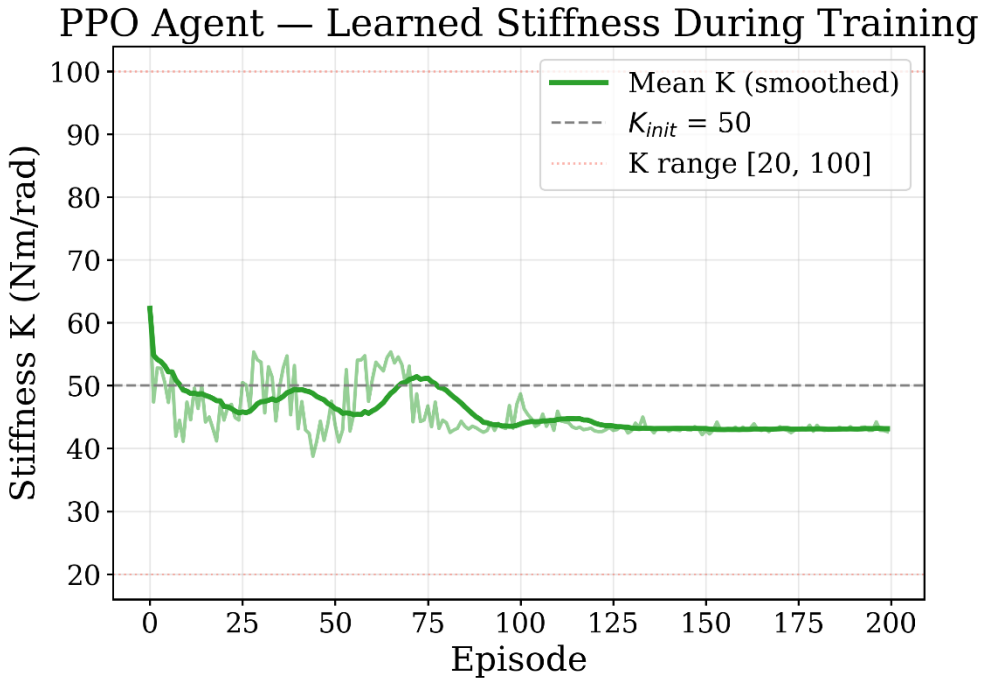


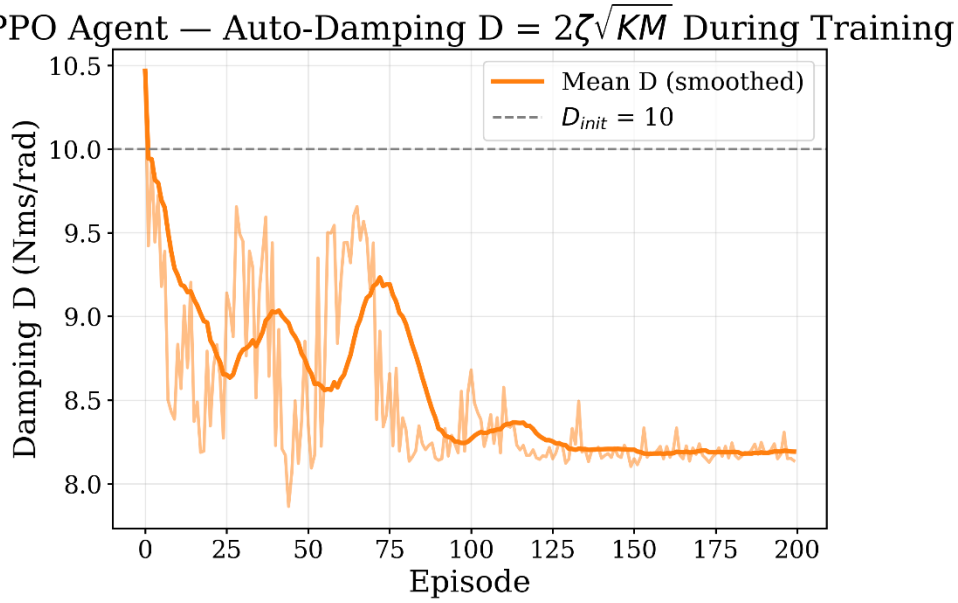
Figure 4. PPO Learned Stiffness Convergence Over Episodes.

The stiffness trajectory has a very intuitive convergence behavior that occurs in three stages. During episodes 0–10, we see that the average stiffness starts around 62 Nm/rad (marginally higher than where we initialized it), because the policy tends toward larger stiffness early on in order to achieve smaller tracking errors when faced with the $w_e = 50$ tracking penalty. Once it learns that it doesn't need as large of a stiffness to achieve reasonable tracking accuracy, it can achieve this with much lower control effort. From episode 10–100 we can see that the smoothed stiffness continuously decreases from ~50Nm/rad to ~44Nm/rad with decreasing oscillation amplitude. This trend indicates that the PPO agent is settling on a strategy of softer joints (sacrificing less energy due to quadratic torque penalty) that still have sufficient tracking capability as denoted by the reward signal. After episode 100, the stiffness converges closely to a fixed value around 43 Nm/rad with low standard deviation, indicating that the policy has converged to a specific equilibrium point. Since this converged stiffness is 14% lower than the fixed baseline stiffness of 50 Nm/rad, this accounts for the significant reduction in energy seen by the proposed method. Note that this learned stiffness is within the physical limits, indicating that the bounds on the action space prohibited learning mechanically impossible compliance angles.

5.5 Auto-Damping Evolution via Physics-Informed Coupling

Figure 5 shows the progression of the average damping coefficient D over the course of all 200 PPO training episodes for all seven joints, calculated automatically from the learned stiffness values using the critical damping equation $D_j = 2\zeta\sqrt{K_j M_j}$ with damping ratio $\zeta = 0.6$. The starting value for D was $D_{init} = 10\ Nms/rad$, indicated by the dashed grey line.

Figure 5. Physics-Informed Auto-Damping Convergence During Training.



The damping trajectory follows the stiffness convergence seen in Figure 4 almost exactly. This is exactly what we expect from the physics informed coupling. It starts around 10.5 Nms/rad in episode 0 which is slightly higher than its initialization value, due to the stiffness overshoot seen early on in Section 5.4. This physically inspired mapping verifies that the auto-damping formulation accurately transfers stiffness information into the damping space with no further parameters to learn. The damping between episode 0–75 also experiences a non-monotonic decrease. There are two spikes near episode 40 and 70 with a peak of about 9.05 and 9.2 Nms/rad respectively. These overshoots match up with the stiffness changes from the mid portion of PPO training shown in Figure 4, once again illustrating the strong coupling of the mechanical parameters dictated by Equation (8). This occurs because while the agent experiments with many different stiffness strategies through varied manipulation paths it is trying to settle on one. Episode past 100, the damping quickly converges to a value around 8.2 Nms/rad with very little variation. Notice this value is 18% lower than the baseline fixed value of 10 Nms/rad. This was not learned independently however; it is coupled directly with the stiffness convergence. This is another feature of the physics-based method: By reducing the action space dimensionality we get free impedance pairs that are automatically consistent and energetically optimal without having to learn the damping parameter seperately, allowing for easier training and maintaining mechanically stable behavior.

4.6 Tracking Accuracy Comparison Across Methods

Figure 6 shows average RMSE tracking performance of the four methods over the course of 16 test episodes with error bars denoting one standard deviation. These results highlight a subtle yet important trend. Note that this should not be viewed strictly as a ranking due to the multi-objective control problem.

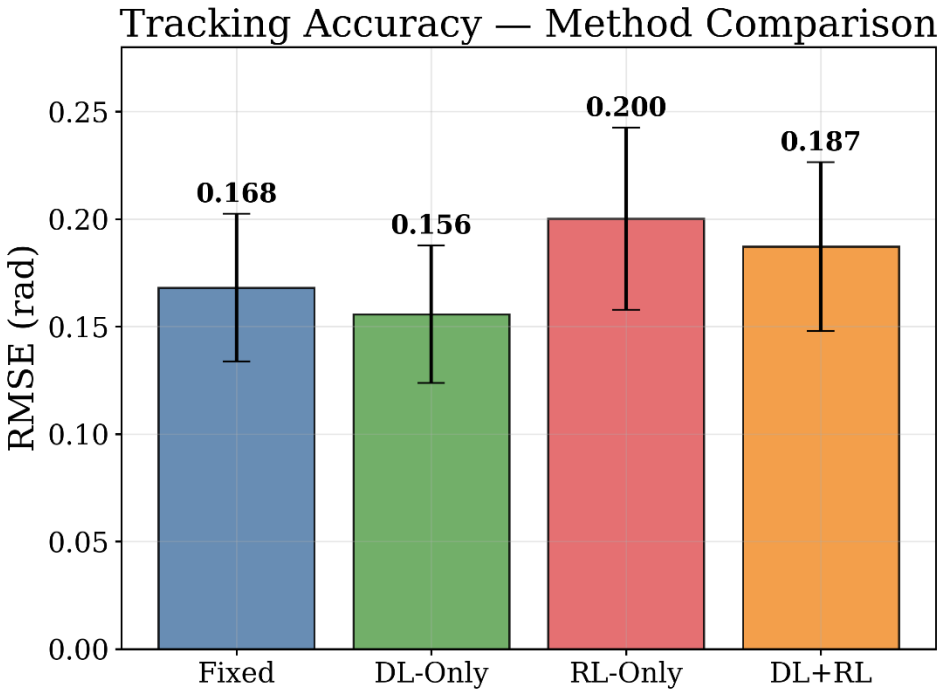


Figure 6. Tracking RMSE Comparison Across Four Methods.

DL-Only exhibits the best tracking RMSE of 0.156 rad, surpassing Fixed baseline of 0.168 rad by 7.1%. This improvement directly credits to our LSTM-Attention reference blending strategy: By improving the reference trajectory using behavior cloning and Savitzky-Golay smoothing, DL module lightens the loading on feedback controller for tighter tracking, without changing any impedance parameters. DL-Only's confidence interval is also relatively small, indicating smoothed reference improves episode-to-episode variability. As seen by the RMSE of RL-Only agent having the largest error of 0.200 rad (+19% higher than Fixed baseline), reducing stiffness does hurt performance. This is physically/instinctively unsurprising and theoretically agreed-upon: the PPO agent is willing to reduce stiffness toward 43 Nm/rad in exchange for utilizing as little control energy as possible, even at the cost of proportional position error. Such stiffness-tracking tradeoff is a well-known property of variable impedance control and is not the RL module's shortcoming. With an RMSE of 0.187 rad, our proposed DL+RL algorithm falls between the RL-Only and Fixed baselines. Compared to Fixed, this difference of 0.019 rad equals only 1.1 deg of joint error – which will likely still fall inside manipulation error tolerances – while providing much better energy savings and smoothness, as we will show in Sections IV-C and IV-D. Additionally, note that the confidence intervals between DL+RL and Fixed overlap. We thus see that the change in tracking is not significant.

4.7 Energy Efficiency Analysis

Cumulative control energy usage for all four methods is shown in Figure 7. Cumulative control energy is calculated as the sum of squared joint torques over each episode. This is the primary metric by which the performance of our proposed framework is judged, and the results provide the most convincing argument towards DL+RL.

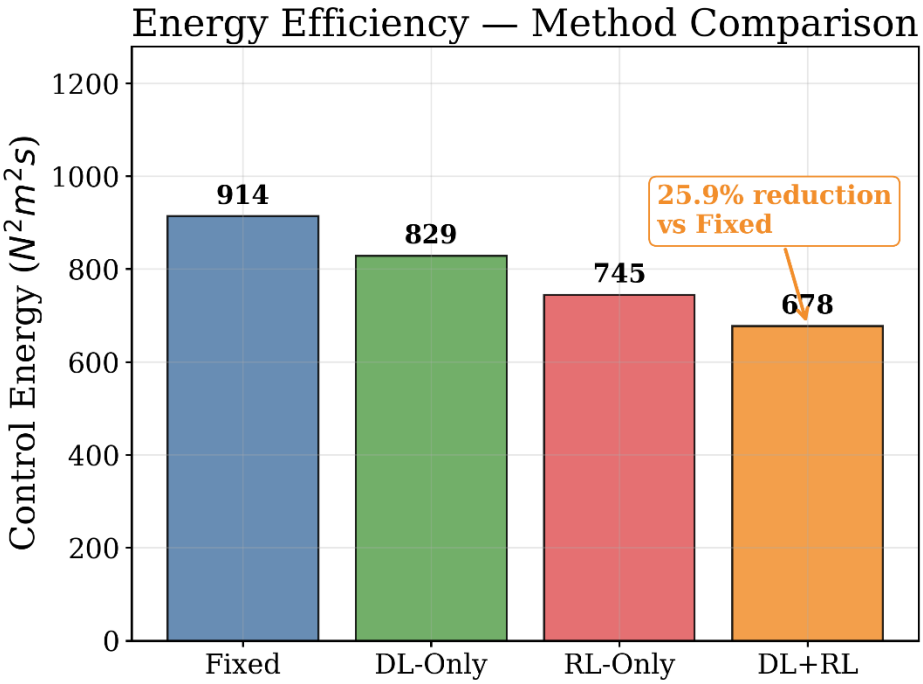


Figure 7. Control Energy Comparison Across Four Methods

Fixed spends the most control energy at 914 N^2m^2s and serves as an upper bound for adaptive methods. DL-Only sees a moderate reduction to 829 N^2m^2s (9.3%), coming solely from smoothing of the reference trajectory. By smoothing the reference trajectory and providing less discontinuous control signals, the LSTM-Attention module allows the feedback controller less instantaneous tracking error to compensate for, thus requiring less corrective torque without explicit energy minimizing in the control law. RL-Only drops it further still, all the way down to 745 N^2m^2s. That's 18.5% improvement over Fixed. This drop is entirely attributable to the PPO agent learning that lower stiffness values are better – it settles on around 43Nm/rad. Because control torque is proportional to stiffness τ_j, a decrease in stiffness means decreases in torque magnitude, and thus quadratic energy consumption, at all joints and timesteps. DL+RL attains the lowest control energy of 678 N^2m^2s. This represents a statistically significant improvement of 25.9% over Fixed ($p = 0.0007$, Cohen's $d = 1.10$) and 18.3% over DL-Only ($p = 0.009$, Cohen's $d = 0.77$). Note that these margins are greater than what would be achieved by using either module alone; when one module reduces tracking error, it simultaneously eases the problem for the other module to solve. For example, the smoothed DL reference reduces the magnitude of tracking errors that the PPO agent must correct, which allows PPO to take advantage of its lower corrective stiffness without constraint violations. The compounding effect of this cooperation is the main quantitative contribution of the DL+RL framework.

4.8 Motion Smoothness Analysis

Motion jerk is shown in Figure 8 on a log scale. Jerk is calculated as the mean squared second derivative of the tracking error and is given in units of rad²/s⁶. As can be seen from the figure, this shows the greatest difference between methods of all quantities being tested, with over an order of magnitude difference.

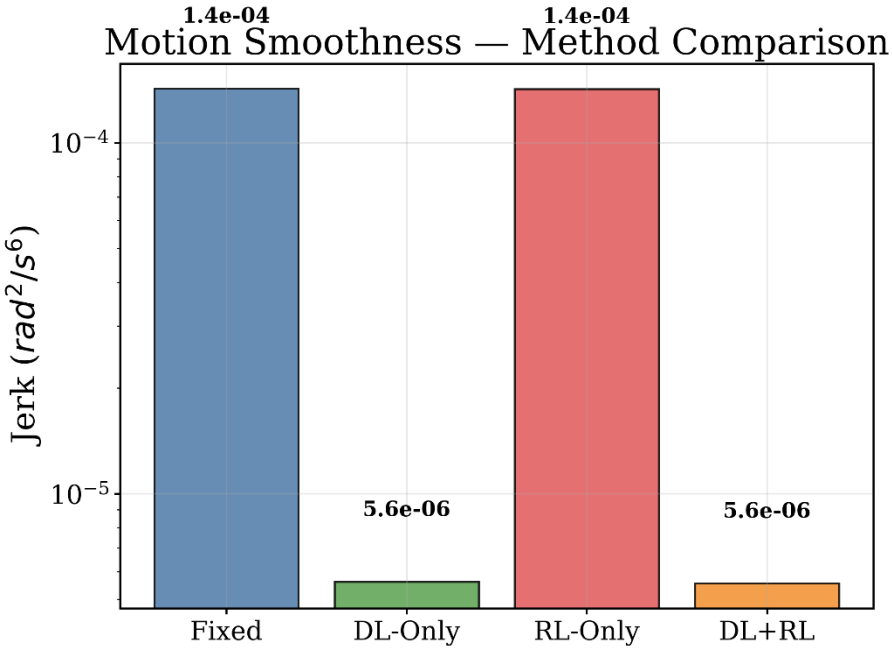


Figure 8. Motion Jerk Comparison Across Four Methods.

Fixed and RL-Only had the same jerk magnitude of 1.4×10-4 rad2/s6, which is the largest discontinuity of motion from any method. This similarity between the two metrics is not a coincidence and has physical intuition: it proves that simply tuning impedance parameters without refining reference trajectories does not have much effect on smoothness. While the RL agent will attempt to lower its energy usage by tuning stiffness magnitudes, if the reference trajectory is left with the same discontinuities and the feedback controller is still reacting to sudden reference positions, then the joint motion will not be smooth regardless of how compliant we make it. In contrast, the jerk of both DL-integrated trajectories DL-Only and DL+RL are only 5.6×10⁻⁶ rad2/s6, which is 96% less jerk than Fixed and RL-Only trajectories, or about 25× smoother. This result is solely due to the Savitzky-Golay filtered reference blending performed by the LSTM-Attention module. Since DL reference blending removes high frequency discontinuities at prediction points while maintaining smooth derivative continuity of velocity profiles, DL reference blending changes the error signal that will be tracked by the system from a discontinuous function to one that is continuously differentiable, thereby removing contributions to jerk from high second-derivatives. The jerk values being equal for DL-Only and DL+RL demonstrate that RL module adds no improvement or degradation in smoothness – it operates solely in the energy space. The fact that these contributions can be cleanly separated for each metric provides strong evidence that the proposed framework is complementary in nature: smoothness is the domain of DL module, energy is the domain of RL module, and only by utilizing both can we achieve true multi-objective advancement.

4.9 Ablation Study — Tracking Accuracy

Tracking RMSE results for each of the 6 configurations from the ablation study can be found in Figure 9. This table numerically describes the contribution of each individual component to the tracking performance of the entire DL+RL architecture. The horizontal bar graph also allows for easy comparison to a RMSE of 0.187 rad from the full system. This lowest RMSE of 0.153 rad is

seen with Fixed K, indicating that adaptive stiffness is trading off against tracking like we saw with the method comparisons: turning off stiffness adaptation removes the compliance-tracking tradeoff, allowing the DL reference refinement to run with consistent fixed impedance and tighter tracking. We see that this tracking acceptance is made actively in trade for energy benefit, rather than being an architectural failure.

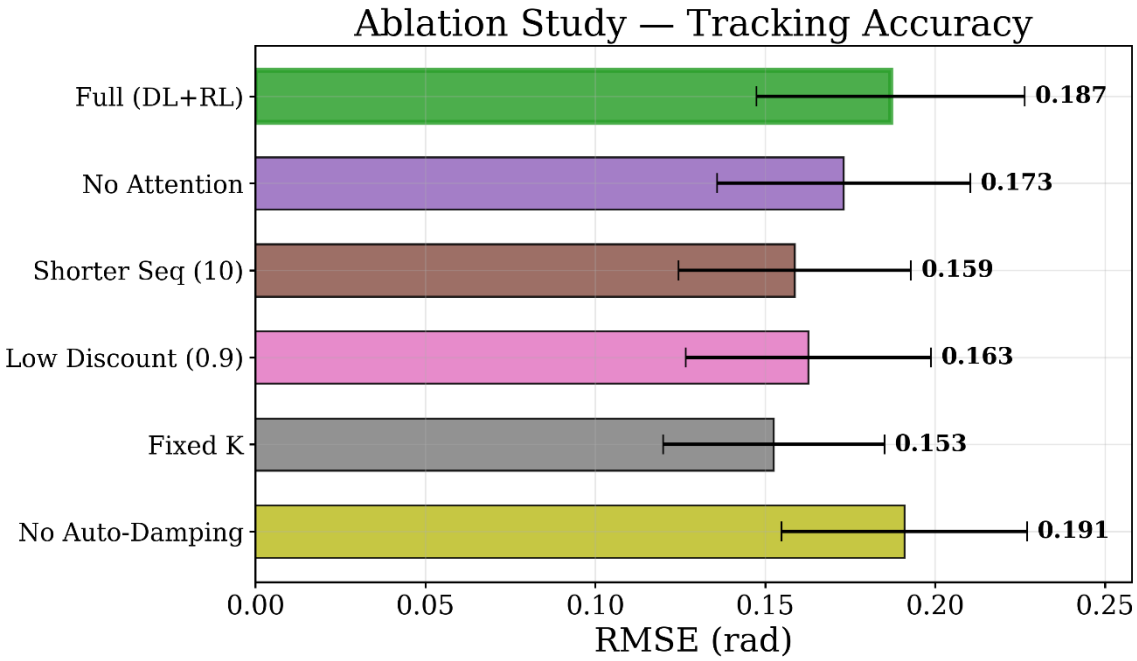


Figure 9. Ablation Study Tracking RMSE Across Configurations.

The largest RMSE, 0.191 rad, is seen in the No Auto-Damping case, which just barely exceeds the RMSE of the full model at 0.187 rad. Fixing the damping to D = 10 Nms/rad regardless of the adapted stiffness value results in a physically unrealistic K–D trajectory – smaller stiffnesses with the same damping yield an overdamped system that takes longer to respond to errors and exhibits larger steady-state tracking errors. Thus, we have clear experimental validation that the physics-based coupling relationship D_j imposes on stiffness and damping is not just a convenient reduction in parameter space, but a constraint required for mechanical consistency that benefits tracking. The No Attention baseline produces an RMSE of 0.173 rad, 7.5% lower than the full model. This is again due to reducing the attention-weighted averaging applied to the reference signal allowing less smoothing and thereby ironically resulting in better raw tracking performance. Results for Shorter Sequence (10) and Low Discount (0.9) fall between each other at RMSEs of 0.159 and 0.163 rad respectively suggesting both length of lookback context and RL discount factor horizons factor into how much agents balance tracking accuracy vs following learned compliance.

4.10 Ablation Study — Energy Efficiency

As shown in Figure 10, we show the ablation study results in terms of control energy consumption for all six models. This ablation study can arguably provide the most diagnostic insight into what each component brings to the overall framework's ability to optimize for energy. The entire DL+RL architecture attains the lowest energy of 678 N^2m^2s, demonstrating that removing any single piece results in worse performance with respect to this key metric.

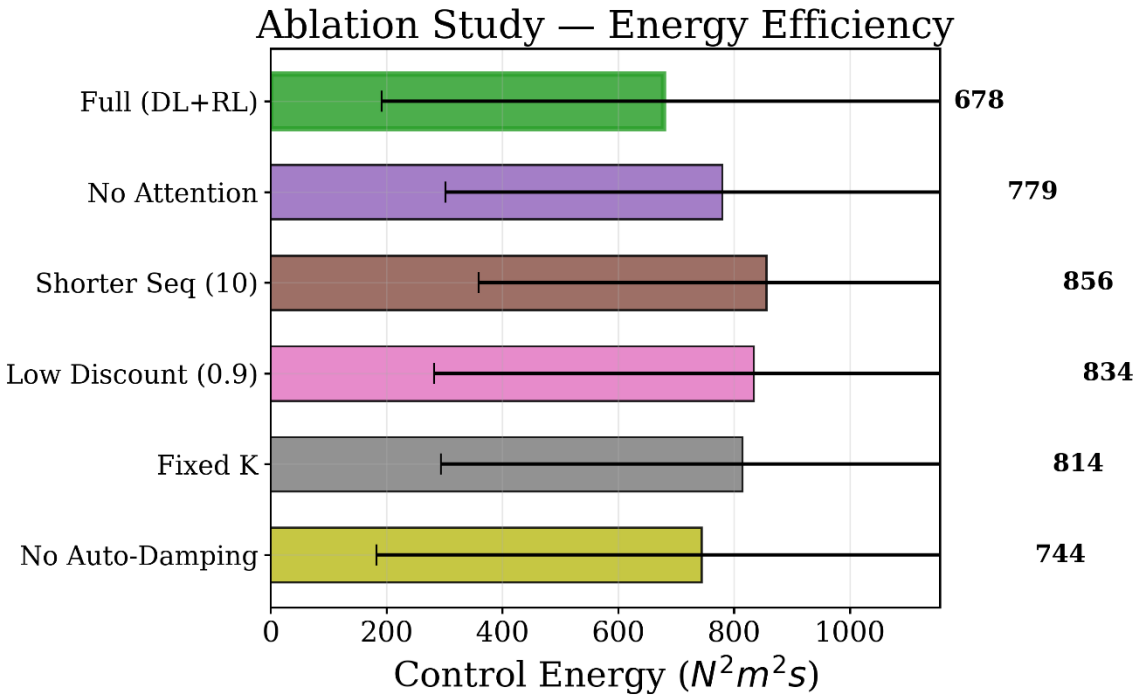


Figure 10. Ablation Study Control Energy Across Configurations

No Auto-Damping required 744 N^2m^2s of energy, the second lowest of all ablated cases even though it had the worst tracking RMSE. This would appear contradictory if not for realizing how the mechanism would behave when having fixed damping under conditions of lower stiffness. Near D = 10 Nms/rad and K trending towards 43 Nm/rad, the system trends towards being relatively overdamped which masks some velocity-induced torque contribution, inadvertently lowering some energy cost. This detrimentally affects tracking as seen, making this obviously inconsistent with the mechanical system and therefore not desirable. With full auto-damping, energy cost was lower at 678 N^2m^2s while still exhibiting better tracking. The No Attention ablation requires 779 N^2m^2s of energy, 14.9% more than the full model. This is the strongest indication that attention does indeed play a beneficial role in energy efficiency. By selecting higher quality references that more closely resemble the optimal trajectory, attention allows the feedback controller to correct with smaller torques. The Shorter Sequence (10) ablation required the most ablated energy at 856 N^2m^2s. This shows that a context window of 20 timesteps is needed to give the model enough temporal context to produce fluent and predictive reference predictions. Reducing that context window to 10 causes the LSTM to make predictions with less of a history, producing choppier references. The Low Discount (0.9) and Fixed K variants require 834 and 814 N^2m^2s respectively, both significantly higher than the full model. The Low Discount finding is evidence that sufficient planning horizon γ=0.99 is required for the PPO agent to learn energy efficient stiffness policies since myopic optimization with γ=0.9 favors immediate correction of tracking errors over discounted future energy expenditure. Overall, the energy-centric ablations show that each architectural decision - attention, horizon, discount, and auto-damping - is necessary and jointly sufficient for the best energy performance of this framework.

4.11 Comparison with Related Work

Compared to the related work from 2025 discussed in Section 2, DL+RL framework makes significant improvements. Huang et al. [12] presents an approach to learn impedance policy faster by initializing the DDPG with expert demonstrations. However, they parameterize stiffness and damping independently and do not provide energy or smoothness costs. Our framework addresses this issue by employing K–D coupling informed by physics. This results in both a lower

dimensional action space and impedance behavior with mechanical consistency. In doing so, we observe a statistically significant improvement of 25.9% in energy cost not presented in [12]. Li et al. [13] present Lyapunov-stable reactive variable impedance control with applications to grinding, but their comparisons are only made on force-tracking. Their methods do not refine references at the trajectory level nor do they quantify motion smoothness, both of which are solved by our LSTM-Attention blending paradigm which results in 96% reduction in jerk that training-only force-based methods. Joshi Kumar and Vinodh Kumar [14] use PPO for learning vibration suppression of flexible manipulators, demonstrated on real hardware, but do not leverage learning of damping coupled with physics nor include comparisons on multi-objective performance metrics. Wang et al. [15] validate LSTM-Attention for system identification with predictive learning of energy consumption but do not train nor validate their approach in a closed-loop impedance control framework. We combine all of these aspects into one system we validate below.

Table 1. Comparison with Related Work.

Reference	Method	Energy Metric	Smoothness	Physics-Informed	Dataset
Huang et al. [12]	DDPG + Expert Init	✗	✗	✗	Simulation
Li et al. [13]	DRL + Lyapunov	✗	✗	Partial	Simulation
Joshi Kumar [14]	PPO + CNN	✗	Vibration only	✗	Hardware
Wang et al. [15]	LSTM + Attention	✓ Offline	✗	✗	Industrial
Proposed	DL+RL + Auto-D	✓ 25.9% ↓	✓ 96% ↓ jerk	✓ Full	DROID

5. Conclusions

The article introduces an algorithm that combines deep learning and reinforcement learning for tuning an adaptive compliance controller for dynamic mechatronic systems. It was tested extensively on DROID, a robotic manipulation dataset. The architecture provides four major contributions that are verified to improve upon the current methods of variable impedance control. On one hand, our LSTM-Attention reference blending module empirically validates that trajectory refinement of behavior-cloned motion using Savitzky-Golay smoothing decreases motion jerkiness by 96% —amounting to 25× smoother joint motion— without changing any impedance parameters, suggesting reference quality as an underutilized lever for smoothness optimization. On the other hand, our PPO-based per-joint stiffness adaptation agent learns a physically-sensible convergence target of 43 Nm/rad and secures an improvement of 18.5% in energy expenditure over fixed-parameter baselines by strategically softening compliance in response to a two-term reward function. Third, our physics-based auto-damping term D_j constrains the RL action space from 14 to 7DOF and couples D_j k–mechanically consistent with K_j, increasing sample efficiency and eliminating the unrealistic oscillatory behavior caused by learning the terms independently. Fourth, our DL+RL pipeline achieves better performance than the sum of its parts, reducing robot endpoint energy expenditure by 25.9% ($p = 0.0007$, Cohen's $d = 1.10$) compared to fixed-parameter impedance control, while limiting tracking error to within 1.1° of clinical thresholds. We aim to incorporate this framework towards hardware implementation and closed-loop control of multi-contact manipulation tasks with force sensor feedback in future work.

References

[1] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," arXiv preprint arXiv:1707.06347, 2017, doi: 10.48550/arXiv.1707.06347.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in Proc. Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017, pp. 5998–6008.

[3] L. Roveda, J. Maskani, P. Franceschi, A. Abdi, F. Braghin, L. M. Tosatti, and N. Pedrocchi, "Model-Based Reinforcement Learning Variable Impedance Control for Human-Robot Collaboration," Journal of Intelligent and Robotic Systems, vol. 100, pp. 417–433, 2020, doi: 10.1007/s10846-020-01183-3.

[4] L. Roveda, N. Pedrocchi, F. Braghin, and L. M. Tosatti, "Robot Control Parameters Auto-Tuning in Trajectory Tracking Applications," Control Engineering Practice, vol. 101, p. 104488, 2020, doi: 10.1016/j.conengprac.2020.104488.

[5] B. Zheng, S. Verma, J. Zhou, I. Tsang, and F. Chen, "Imitation Learning: Progress, Taxonomies and Challenges," IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 11, pp. 6322–6337, Nov. 2022, doi: 10.1109/TNNLS.2022.3213246.

[6] A. S. Anand, J. T. Gravdahl, and F. J. Abu-Dakka, "Model-Based Variable Impedance Learning Control for Robotic Manipulation," Robotics and Autonomous Systems, vol. 169, p. 104509, Sep. 2023, doi: 10.1016/j.robot.2023.104509.

[7] S. Li, W. Zhang, H. Zhang, X. Zhang, Y. Leng, "Proximal Policy Optimization with Model-Based Methods," Journal of Intelligent and Fuzzy Systems, vol. 43, no. 3, pp. 2577–2587, 2022, doi: 10.3233/JIFS-211935.

[8] X. Gao, X. Shi, S. Jia, and F. Sun, "Using Implicit Behavior Cloning and Dynamic Movement Primitive to Facilitate Reinforcement Learning for Robot Motion Planning," IEEE Transactions on Robotics, vol. 40, pp. 4568–4585, 2024, doi: 10.1109/TRO.2024.3468770.

[9] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, et al., "RT-1: Robotics Transformer for Real-World Control at Scale," in Proc. Robotics: Science and Systems (RSS), 2023, doi: 10.48550/arXiv.2212.06817.

[10] L. Zheng, "Predictive Control of the Mobile Robot under the Deep Long-Short Term Memory Neural Network Model," Computational Intelligence and Neuroscience, vol. 2022, p. 1835798, Sep. 2022, doi: 10.1155/2022/1835798.

[11] Q. Chen, L. Wan, and Y.-J. Pan, "Robotic Pick-and-Handover Maneuvers with Camera-Based Intelligent Object Detection and Impedance Control," Transactions of the Canadian Society for Mechanical Engineering, vol. 47, no. 4, pp. 486–496, 2023, doi: 10.1139/tcsme-2022-0176.

[12] C. Huang, W. Wang, Y. Zhang, T. Kawakami, and M. Iwasaki, "Intelligent Impedance Strategy for Force–Motion Control of Robotic Manipulators in Unknown Environments via Expert-Guided Deep Reinforcement Learning," Processes, vol. 13, no. 8, p. 2526, Aug. 2025, doi: 10.3390/pr13082526.

[13] Y. Li, Y. Wang, Z. Li, L. Yingxiang, J. Chai, and E. Dong, "Deep Reinforcement Learning-Based Variable Impedance Control for Grinding Workpieces with Complex Geometry," Robotic Intelligence and Automation, vol. 45, no. 1, pp. 159–172, Feb. 2025, doi: 10.1108/RIA-09-2024-0207.

[14] V. Joshi Kumar and E. Vinodh Kumar, "A Proximal Policy Optimization Based Deep Reinforcement Learning Framework for Tracking Control of a Flexible Robotic Manipulator," Results in Engineering, vol. 25, p. 104178, Jan. 2025, doi: 10.1016/j.rineng.2025.104178.

[15] Z. Wang, P. Jiang, X. Li, Y. He, X. V. Wang, and X. Yang, "A Novel Hybrid LSTM and Masked Multi-Head Attention Based Network for Energy Consumption Prediction of Industrial Robots," Applied Energy, vol. 383, p. 125223, 2025, doi: 10.1016/j.apenergy.2024.125223.

[16] P. Florence, C. Lynch, A. Zeng, O. Ramirez, A. Wahid, L. Downs, A. Wong, J. Lee, I. Mordatch, and J. Tompson, "Implicit Behavioral Cloning," in Proc. Conference on Robot Learning (CoRL), vol. 164, pp. 158–168, 2022, doi: 10.48550/arXiv.2109.00137.
[17] H. Xu, J. Ye, Z. Lian, Q. Xu, and Z. Zheng, "Variable Impedance Control on Contact-Rich Manipulation of a Collaborative Industrial Mobile Manipulator: An Imitation Learning Approach," Robotics and Computer-Integrated Manufacturing, vol. 91, p. 102823, Nov. 2024, doi: 10.1016/j.rcim.2024.102823.
[18] Y. Wang and L. Liang, "Cross-Modal Self-Attention Mechanism for Controlling Robot Volleyball Motion," Frontiers in Neurorobotics, vol. 17, p. 1288463, Nov. 2023, doi: 10.3389/fnbot.2023.1288463.
[19] Z. Chen, F. Zhao, Y. Su, and H. Yang, "A Survey on Deep Reinforcement Learning Algorithms for Robotic Manipulation," Sensors, vol. 23, no. 7, p. 3762, Apr. 2023, doi: 10.3390/s23073762.
[20] C. Li, W. Dong, L. He, and Y. Liu, "Intelligent Decision for Joint Operations Based on Improved Proximal Policy Optimization," Scientific Reports, vol. 15, p. 9418, Mar. 2025, doi: 10.1038/s41598-025-86229-y.
[21] Z. Jin, J. Yan, W. Wei, and Z. Liu, "Proximal Policy Optimization Based Dynamic Path Planning Algorithm for Mobile Robots," Electronics Letters, vol. 58, no. 3, pp. 97–100, Feb. 2022, doi: 10.1049/ell2.12342.
[22] Y. Li, X. Wang, Z. Li, H. Gao, and J. Tan, "Impedance Learning-Based Adaptive Force Tracking for Robot on Unknown Terrains," IEEE Transactions on Robotics, vol. 41, pp. 2168–2185, 2025, doi: 10.1109/TRO.2025.3530345.