An Empirical Study on Curriculum Learning for Reinforcement Learning-based Biomechanical Arm Control

Asmae OUHSSAIN[1], Mohamed-Harith IBRAHIM[1], Sébastien HARISPE[1], Joseph DIAB[1], Denis MOTTET[2], Jacky MONTMAIN[1]

***Abstract*— Controlling high-dimensional musculoskeletal systems is challenging due to the large number of muscle actuators and the need for coordinated motor behavior. Recent work has shown that combining reinforcement learning with curriculum learning can improve performance on such tasks, yet the design of effective curricula remains an open question. In this work, we develop two curriculum strategies, Large-to-Small and Small-to-Large, against a no-curriculum baseline in order to control a musculoskeletal model of the arm during a task involving the tracing of geometric shapes. Both curriculum strategies significantly outperform the baseline, with success rates of 94% and 88%, in contrast to 42% without a curriculum. The best results are presented using the Large-to-Small strategy, which suggests that the agent learns better when it begins with larger geometrical shapes and gradually progresses on to smaller ones, where the tracing becomes more precise. Furthermore, we evaluate generalization to unseen instances, where the results demonstrate that curriculum-trained agents effectively generalize to novel shapes and scale variations, illustrating the robustness of the learned motor skills beyond the training conditions.**

The code for reproducing our experiments is accessible at https://github.com/AsmaeOUHSSAIN/CuRL-BioArm.git

I. INTRODUCTION

Reinforcement Learning (RL) has emerged as a powerful paradigm for learning control policies in complex domains where explicit programming or precise mathematical models are unavailable. RL has achieved remarkable success in areas ranging from robotic manipulation to autonomous systems, by discovering control strategies through interaction with an environment and reward signals, without requiring manually designed controllers. Recently, RL has demonstrated promise for learning musculoskeletal control, where agents must coordinate a large number of muscle actuators to produce desired movements [1] [2]. This growing interest is reflected in dedicated research challenges such as Learn to Move and MyoChallenge, which have shown that RL, combined with realistic simulation environments like MyoSuite, can tackle complex musculoskeletal tasks including locomotion, reaching, and dexterous manipulation [3] [4].

Applying RL to musculoskeletal systems introduces distinct challenges. Unlike torque-controlled robots that apply joint torques directly, musculoskeletal control must work within muscle actuation constraints. Musculoskeletal systems exhibit high dimensionality, complex nonlinear dynamics, and high redundancy [5] [6]. A human arm contains more than 20 degrees of joint freedom actuated by dozens of muscle–tendon units, all contributing to the same movements. This redundancy makes it difficult for RL agents to discover useful behaviors through trial-and-error, since random muscle activations rarely produce coordinated movements.

To address these learning challenges, Curriculum Learning (CL) has been proposed as a complementary strategy. Rather than training on the full task from the start, CL structures training by exposing the agent to a carefully designed sequence of increasingly difficult tasks. By starting with easier versions of the problem, agents can learn more efficiently before tackling full task complexity. In RL, CL can be implemented through task parameter variations, experience orderings, or task sequences designed to progressively increase difficulty. Empirical evidence supports CL's effectiveness: the top-performing solutions in the NeurIPS MyoChallenge 2022 competition combined RL with CL strategies, demonstrating practical benefits for musculoskeletal control [7].

Despite growing adoption of CL in musculoskeletal RL, a comprehensive analysis of curriculum design choices remains absent from the literature. It is unclear which specific curriculum strategies are most effective for muscle-driven control and how they compare in learning efficiency and final performance. This paper contributes to addressing this gap through an empirical study comparing multiple curriculum strategies for musculoskeletal arm control. In this work, we make three contributions. First, we design a 3D shape tracing task on the MyoArm musculoskeletal model where difficulty varies with the shape size and orientation of the target, enabling analysis of how difficulty affects learning. Second, we evaluate different curriculum designs using identical RL algorithms, hyperparameters, and evaluation protocols, to ensure performance differences arise purely from curriculum design, and to identify the strategies that are most effective for musculoskeletal arm control, thereby providing practical guidance for future work. Third, we assess the generalization ability of the trained policies across unseen scales, shapes, and traversal directions, highlighting both their strengths and limitations beyond the training conditions.

The remainder of this paper is organized as follows. Section II reviews related work on curriculum learning and musculoskeletal RL. Section III describes the task design and curriculum strategies. Section IV presents experimental results and analysis. Section V concludes with discussion

[1] SyCoIA, IMT Mines Ales, 30100 Ales, France asmae.ouhssain@mines-ales.fr, mohamed-harith.ibrahim@mines-ales.fr, sebastien.harispe@mines-ales.fr, joseph.diab@mines-ales.fr, jacky.montmain@mines-ales.fr

[2] EuroMov DHM, Université de Montpellier, 34000 Montpellier, France denis.mottet@umontpellier.fr

and future work.

II. Related Works

CL refers to a training strategy in which learning samples or tasks are presented in a structured order rather than being sampled randomly [8]. Although the notion is based on long-standing principles from education and developmental psychology, it was formalized within machine learning, and it has then been applied to a variety of contexts including RL frameworks. Recent studies indicates CL is crucial for complex biomechanical tasks. In the NeurIPS MyoChallenge 2022 [7], competitors trained controllers for hand manipulation using a 39-muscle hand model. Over 340 solutions that were submitted, the top solutions all combined RL with some form of CL. Among research works, a simulated hand was trained to rotate two balls using an Static-to-Dynamic Stabilization (SDS) curriculum, which gradually increases task difficulty in a structured way [10]. The agent first learned to hold the balls still, then to rotate them slowly, and over multiple stages the difficulty was increased until fast continuous rotation was achieved. The SDS curriculum enabled the agent to achieve 100% success. More generally, in RL, CL typically takes the form of a sequence of tasks or environment configurations designed to train the agent on easier tasks before moving to harder ones. One common approach is to control the initial states from which the agent begins training. Reverse curriculum is one such method, where training starts from states close to the target and is gradually extended to more difficult ones [11]. Similarly, CL integrated in RL controllers was applied for functional electrical stimulation of a 2-DOF arm with six muscles [12]. As the agent improves its performance, the set of starting states is gradually extended farther from the goal, so that complexity increases progressively. This idea has been further developed through backward reachability curricula, where an approximate dynamics model is used to expand the distribution of initial states in a physically consistent manner [13]. Another line of research on CL focuses on task generation and sequencing. Instead of modifying initial conditions, tasks are selected based on their expected contribution to learning. For instance, a skill-environment Bayesian network can be used to link skills, goals, and task features, allowing the system to predict which task will most improve performance and to select it accordingly [14]. A third group of approaches addresses the exploration problem rather than the task structure. In high-dimensional musculoskeletal systems, standard random exploration is often ineffective. To overcome this, self-organization methods have been proposed to generate coordinated movements before RL training, providing a more structured starting point for learning [15]. While many CL approaches focus on controlling initial states or improving exploration, other works define curricula through variations of task parameters, such as target positions, task geometry, or environmental conditions. These methods progressively increase task difficulty by expanding the complexity of the task [9]. Building on this idea, our work studies CL over task parameters, with a particular focus to how the ordering of tasks affects learning performance in musculoskeletal control problems.

III. Materials and Methods

A. Musculoskeletal arm model and task definition

To investigate how curriculum design and task difficulty affect performance in musculoskeletal control, we use the MyoSuite simulation framework, which is built on the MuJoCo physics engine and designed for musculoskeletal control research [4]. Specifically, we employ the MyoArm model, a detailed upper-limb musculoskeletal model with 20 joints and 32 muscle–tendon units. These muscle–tendon units actuate the shoulder, elbow, and wrist joints. We extend the standard one-point reach task into shape tracing task, in which the agent must sequentially reach a sequence of waypoints defining a closed square trajectory in three-dimensional space. A square generates a set of waypoints; given the square's center coordinates, half-side length, and horizontal orientation, we calculate the square's four vertices. Each edge is then split into a given number of uniformly spaced intermediate waypoints. In each step, we compute the distance between the fingertip and the current goal waypoint. When this distance goes below the tolerance, the waypoint is indicated as reached and the next waypoint becomes the target. The task is considered successful if all waypoints are reached in the correct sequence and within the allowed number of steps.

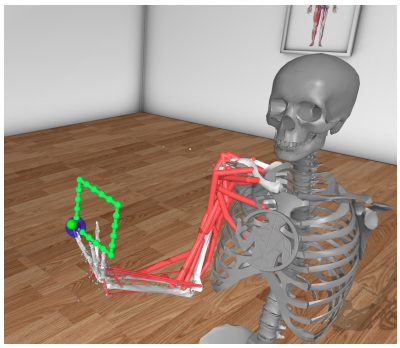


Fig. 1: Visualization of the MyoArm model performing the shape-tracing task in MyoSuite. The green path indicates the waypoints already traced by the fingertip, and the red lines represent the muscle-tendon units.

B. Reinforcement Learning formulation

We formulate the arm control problem as a Markov Decision Process (MDP) defined by the tuple $(\mathcal{S}, \mathcal{A}, T, R, \gamma)$ [16] [17]. At each timestep t, the agent observes a state $s_t \in \mathcal{S}$ comprising proprioceptive signals (joint angles, angular velocities, muscle-tendon excitations), the position of the current waypoint, the distance between the fingertip and the target, and a binary vector of length 21 indicating reached waypoints. Based on this state, the agent selects an action $a_t \in \mathcal{A} \subseteq \mathbb{R}^{32}$ representing muscle excitation signals for the 32 MyoArm actuators. The transition function

TABLE I: Curriculum levels for both training strategies. Scale and center position are expressed in meters (m).

Strategy	Level	Scale	Rot.	Center pos.	Thr.
NoCL	–	[0.02, 0.08]	[±45°]	[−0.30, −0.12]	–
Large → Small	Large (L0)	[0.06, 0.08]	[±10°]	[−0.24, −0.16]	70%
	Large-Med. (L1)	[0.04, 0.08]	[±25°]	[−0.24, −0.12]	60%
	Lrg-Med-Sml (L2)	[0.02, 0.08]	[±45°]	[−0.30, −0.12]	Terminal
Small → Large	Small (L0)	[0.02, 0.04]	[±10°]	[−0.24, −0.16]	70%
	Small-Med. (L1)	[0.02, 0.06]	[±25°]	[−0.24, −0.12]	60%
	Sml-Med-Lrg (L2)	[0.02, 0.08]	[±45°]	[−0.30, −0.12]	Terminal

$T : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ generates the next state using the MuJoCo physics engine and the muscle activation dynamics. The agent's goal is to maximize the expected value of discounted return according to the reward function:

$$r_t = -\|\mathbf{p}_t^{\text{finger}} - \mathbf{p}_t^{\text{wp}}\| + b_{\text{wp}}\,\mathbf{1}_{\text{reached}} + b_{\text{final}}\,\mathbf{1}_{\text{done}}, \quad (1)$$

where $\mathbf{p}_t^{\text{finger}}$ is the position of the fingertip and $\mathbf{p}_t^{\text{wp}}$ is the current waypoint position. The first term penalizes the Euclidean distance to the waypoint, $b_{\text{wp}} = 1.0$ is a bonus awarded each time a waypoint is reached, and $b_{\text{final}} = 50$ is a terminal bonus for completing the entire square trajectory. The indicator functions $\mathbf{1}_{\text{reached}}$ and $\mathbf{1}_{\text{done}}$ equal 1 when a waypoint is reached or the episode succeeds, respectively. An episode corresponds to a single attempt to complete the shape tracing task. It starts from an initial position where the arm is initialized and the first waypoint is set as the target, and it terminates either when all waypoints are reached in the correct sequence or when a maximum number of steps is exceeded (see Fig. 1).

C. Experimental design

To examine the effect of CL on shape-tracing skill acquisition, we compare three training strategies that differ in how the task parameters are presented to the agent during training. The complexity of each episode is defined by three parameters, which are the square's half-side length (scale), rotation around the vertical axis, and vertical position of the square center. The first strategy serves as a baseline and doesn't include any curriculum. In each episode, all three parameters are randomly sampled from their entire ranges. The agent is exposed to the whole task distribution from the beginning of training. The second one involves a Large-to-Small curriculum, where the agent is first trained on large squares and then progressively introduced to smaller ones. Training starts with large squares, small rotation ranges, and centered vertical positions. The agent is evaluated periodically on tasks sampled from the current level distribution. When the success rate exceeds the threshold associated with the current level, the scale range expands to include medium and then small squares, the rotation range increases, and the vertical range extends. At the final stage, all parameter ranges are included. The third strategy follows a Small-to-Large curriculum, which reverses the ordering of the scale parameter. Training begins with small squares and progressively includes medium and then large ones. The rotation and vertical position ranges expand at each level following the same progression thresholds as in the Large-to-Small curriculum. Table I provides a detailed level description, including parameter ranges and progression levels for the three training strategies. All strategies use the same RL algorithm, more specifically Proximal Policy Optimization (PPO) [18], and share an identical network architecture consisting of two hidden layers of 256 neurons each and the same set of hyperparameters. Each experiment is trained for 10^7 timesteps over 8 parallel environments.
The overall training framework, including the interaction between the agent, the environment, and the CL strategy, is illustrated in Fig. 2.

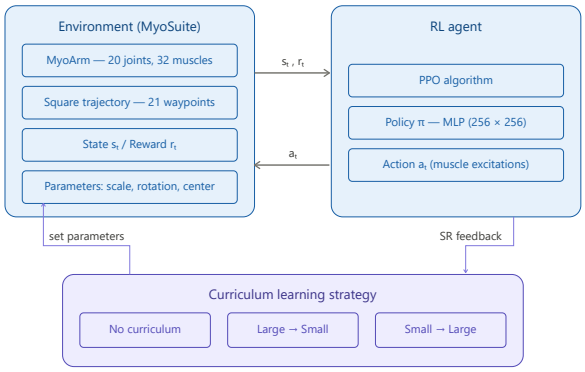


Fig. 2: Overview of the proposed training framework.

D. Evaluation metrics

We evaluate each training strategy's performance using metrics that indicate how effectively the agent learns the task and how effective the final policy is.

i) The *cumulative episodic reward* is defined as the sum of all rewards received by the agent over a single episode. This metric is recorded to evaluate the learning speed and convergence behavior.

ii) The *waypoint progress metric* evaluates the agent's progress during an episode. It is defined as:

$$\text{WP}_{\text{prog}} = \frac{k}{K},$$

where k is the number of waypoints reached and K is the total number of waypoints, which is 21. A WP_{prog} value of a 1 indicates that the agent completed the entire square.

iii) The *progress-weighted tracking error* evaluates tracing precision while accounting for task progress. It is defined as:

$$e_{\text{pw}} = \bar{e}\,(1 - WP_{\text{prog}})\,,$$

where $\bar{e}$ is the mean tracing error, defined as the average Euclidean distance between the fingertip and the waypoints over the episode:

$$\bar{e} = \frac{1}{T}\sum_{t=1}^{T} \|\mathbf{p}_t^{\text{finger}} - \mathbf{p}_t^{\text{wp}}\|.$$

The mean error alone can be misleading when agents reach different numbers of waypoints. An agent that

makes little progress may obtain a low error by remaining near the initial part of the trajectory, while an agent that progresses further may accumulate a higher error. To address this issue, the error is weighted by the remaining fraction of the trajectory. As a result, an agent that completes the full trajectory obtains a value of zero, enabling a fair comparison between agents with different levels of progress.

iv) The *success rate* is defined as the fraction of evaluation episodes in which the agent reaches all waypoints within the allowed simulation steps.

$$\mathrm{SR} = \frac{n_{\mathrm{success}}}{n_{\mathrm{episodes}}},$$

where n_{success} is the number of successfully completed episodes and n_{episodes} is the total number of evaluation episodes. A success rate of zero does not necessarily indicate that the agent has not achieved any significant progress. An agent may frequently achieve a number of waypoints without fully completing the trajectory; yet, this progress remains fully ignored in the success rate.

E. Generalization evaluation

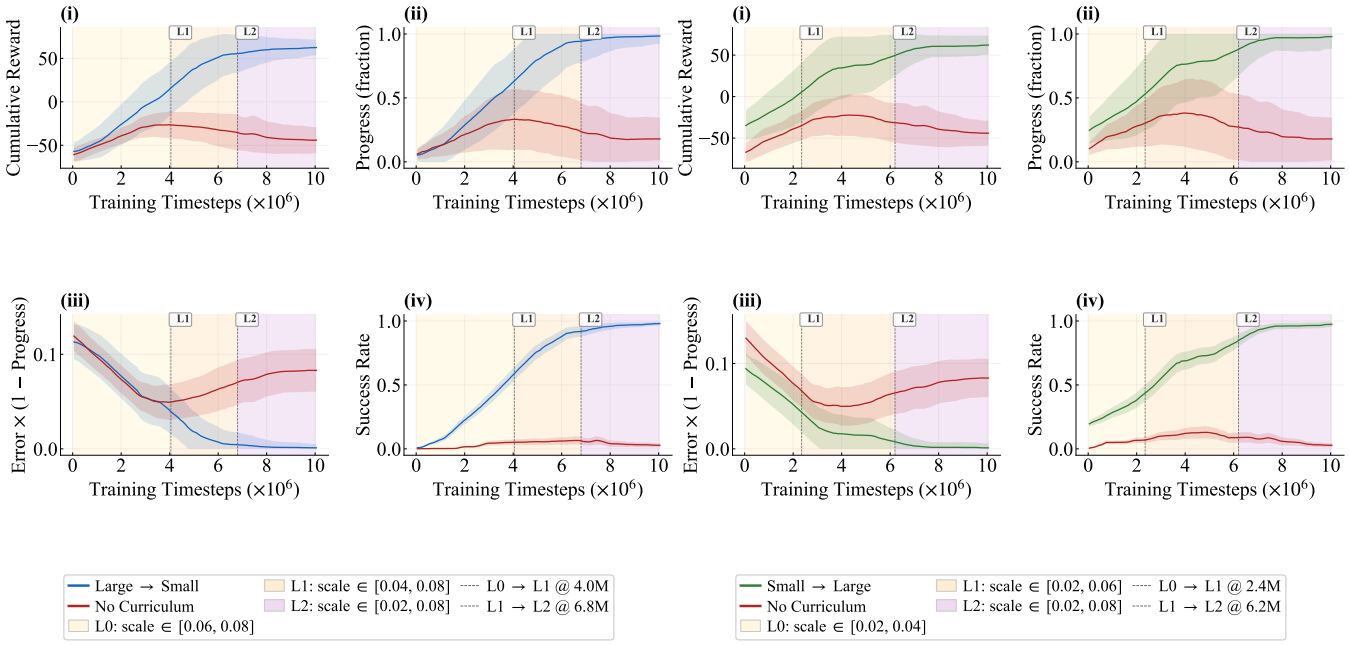
Beyond comparing agent performance with and without CL, we evaluate the generalization capacity of each learned policy to conditions unseen during training. This evaluates whether the learned behavior is robust and transferable beyond the training distribution. We consider three types of generalization tests. First, in the scale test, the agent is evaluated on trajectories with different spatial scales. We evaluate each model on the original training range $[0.02,\, 0.08]$ as a baseline, on a smaller-than-training range $[0.005,\, 0.02]$ that produces more closely spaced waypoints, and on a larger-than-training range $[0.08,\, 0.14]$. This test examines whether the learned policy can adapt to both reduced and expanded versions of the task that were never encountered during training. Second, in the shape test, we evaluate generalization to unseen geometries: a wide rectangle (aspect ratio 2:1), a tall rectangle (aspect ratio 1:2), a star, and an equilateral triangle, replacing the square used in training. This evaluation determines whether the agent has learned a general tracing strategy, rather than just memorizing a specific path. Finally, in the direction test, each model is evaluated using both the standard waypoint order and a reversed waypoint sequence. This determines whether the policy captures a general tracing behavior or depends on a specific directed pattern.

IV. Results and Discussion

To track the evolution of learning performance, we save a policy checkpoint every $50k$ timesteps. Each checkpoint is then evaluated on a fixed set of test episodes sampled from the distribution associated with the curriculum level at which it was saved. In the Large-to-Small strategy, policies learned at level $L0$ are examined based on large configurations. At level $L1$, the evaluation set extends to both large and medium scales, whereas at the final level $L2$, it covers the whole parameter range. A similar approach is used for the Small-to-Large strategy, whereby evaluations at level $L0$ are executed on small-scale configurations, with progressively larger sizes included at following levels. To ensure a fair comparison, the baseline noCL agent, which is trained on the full parameter range, is evaluated on the same level-specific test configurations as the curriculum-based agents. This means that, at each stage, the baseline is tested on the same subset of task parameters corresponding to the current curriculum level. Figures 3a and 3b present the resulting training curves for the Large-to-Small and Small-to-Large curricula, respectively, each compared against the noCL strategy, which is used as the baseline. The four evaluation metrics introduced in Section III-D are reported over the full 10^7 training timesteps. Background shading indicates the active curriculum level, and dashed vertical lines denote the timesteps at which level transitions occurred.

For the Large-to-Small curriculum (Figure 3a), the agent transitions from level $L0$ to $L1$ at approximately 4.0×10^6 timesteps and from $L1$ to $L2$ at 6.8×10^6 timesteps. During level $L0$, where only large squares are presented, the curriculum agent exhibits a continuous increase in mean cumulative episodic reward (i), rising from approximately -60 to 0, while the baseline remains around -35 over the same period. The waypoint progress (ii) shows a more significant distinction; the curriculum agent reaches over 0.6 by the first level transition, whereas the baseline rises to approximately 0.3 before declining back to 0.2. The progress-weighted tracking error (iii) decreases consistently for the curriculum agent, dropping below 0.04 by the $L0{-}L1$ transition and approaching zero by the end of training. In contrast, the baseline error initially decreases but then rises after 4×10^6 timesteps, stabilizing around 0.08. The success rate (iv) captures the overall contrast most clearly; the curriculum agent begins improving around 2×10^6 timesteps and reaches approximately 95% by the end of training, while the baseline remains below 10% throughout. Thus, despite the fact that each level transition increases task variability, the curriculum agent continues to improve without any degradation, indicating that the skills acquired at earlier levels transfer effectively to the expanded task distributions. The Small-to-Large curriculum (Figure 3b) achieves qualitatively similar results. The agent advances from $L0$ to $L1$ in 2.4×10^6 timesteps and from $L1$ to $L2$ in 6.2×10^6 timesteps, completing the first level earlier than the Large-to-Small agent. The mean reward (i) starts lower, at -75, reflecting the higher initial difficulty of small squares, but gradually increases and reaches values similar to the Large-to-Small agent by the end of training. The waypoint progress (ii) displays a similar increasing trend, and the progress-weighted tracking error (iii) approaches zero, while the success rate (iv) reaches around 95%, closely aligning with the final performance of the Large-to-Small strategy.

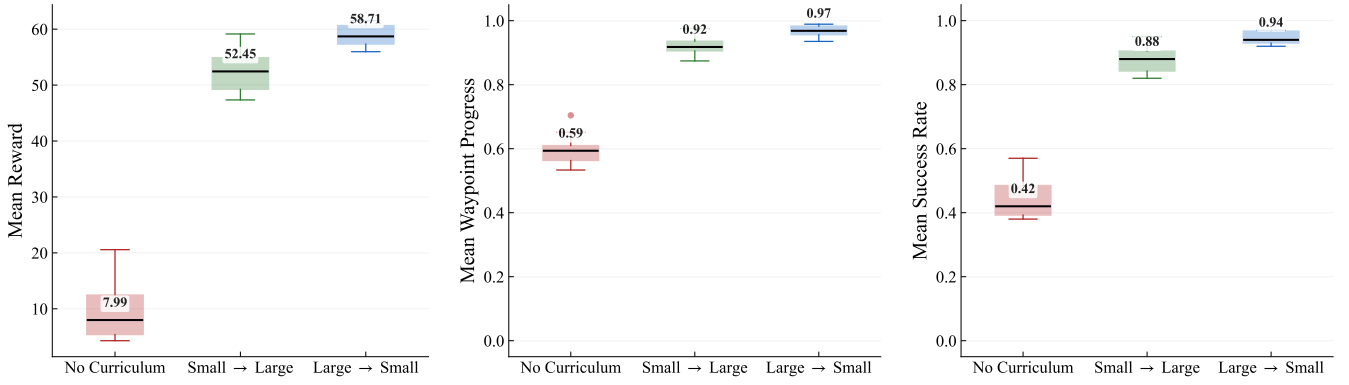
To compare the three strategies after training, each final policy is evaluated on 1000 identical test episodes with parameters sampled uniformly from the full range. Figure 4 summarizes the results. The Large-to-Small agent achieves the highest performance across all three metrics, with a



(a) Large → Small vs No Curriculum

(b) Small → Large vs No Curriculum

Fig. 3: Comparison of curriculum learning strategies against the no-curriculum baseline across all evaluation metrics.



(a) Episode Reward

(b) Waypoint Progress

(c) Success Rate

Fig. 4: The final model performance comparison across CL and no Curriculum Learning (noCL) strategies.

median success rate of 94%, a mean reward of 58.71, and a mean waypoint progress of 97%. The Small-to-Large agent follows closely, reaching a median success rate of 88%, a mean reward of 52.45, and a mean waypoint progress of 92%. The baseline agent, by contrast, achieves a median success rate of only 42%, a mean reward of 7.99, and a mean waypoint progress of 59%, with significantly more variation.

Table II summarizes the generalization performance across scale, shape, and direction conditions. For the scale test, both curriculum agents generalize well to scales that are smaller than training, with success rates of 89% for Large-to-Small and 88% for Small-to-Large, closely matching their in-distribution performance. Meanwhile, the larger-than-training scales tend to be more challenging; the Large-to-Small agent decreases to 48% and the Small-to-Large agent falls to 14%. For the shape test, both curriculum agents generalize well to unseen geometries, with the Large-to-Small agent achieving 69%–90% success across all novel shapes and the Small-to-Large agent reaching 70%–77%. In the other hand, the baseline remains below 40% in all cases.

TABLE II: Results across Direction, Shape, and Scale conditions.

		noCL		S-to-L		L-to-S	
		SR	WP_{prog}	SR	WP_{prog}	SR	WP_{prog}
Scale	[.02,.08] (base.)	43%	60.7%	89%	90.8%	92%	93.5%
	[.005,.02]	26%	39.2%	88%	92.9%	89%	92.2%
	[.08,.14]	2%	22%	14%	39.3%	48%	66.8%
Shape	Square (base.)	43%	60.7%	89%	90.8%	92%	93.5%
	Rect. wide	29%	48.7%	75%	81.1%	90%	91.6%
	Rect. tall	40%	58.7%	77%	86.3%	90%	95.7%
	Star	27%	43.6%	70%	78.8%	70%	82.5%
	Triangle	18%	33.8%	71%	79.4%	69%	84.7%
Direction	Forward (base.)	43%	60.7%	89%	90.8%	92%	93.5%
	Reversed	0%	13.1%	1%	17.3%	5%	21.6%

Rectangles yield the highest transfer due to their similarity to the training square, while the triangle produces the largest drop. In the direction test where the order is reversed, success rates for the Large-to-Small, Small-to-Large, and baseline agents significantly drop to 5%, 1%, and 0%, respectively. This indicates that all policies acquired a direction-specific tracing pattern.

A key observation across all experiments is that the Large-to-Small curriculum consistently outperforms the Small-to-Large strategy. We attribute this to the nature of the initial training phase. Starting with large squares allows the agent to learn basic arm movements first, since the waypoints are far apart. Once this foundation is established, the agent can adapt to smaller squares by refining its precision. This interpretation is further supported by the scale generalization results, where the Large-to-Small agent transfers substantially better to larger-than-training scales (48% vs. 14%), suggesting that early exposure to larger movements develops a more transferable motor capacity. These results appear to align with the progressive development of motor skills in children, characterized by a transition from gross to fine motor control. Children's learning to trace geometric shapes is part of the gradual development of motor skills, following a progression from gross motor skills to fine motor skills. Findings from developmental psychology and motor neuroscience show that proximal control (shoulder, arm) precedes distal control (wrist, fingers). In this context, starting with large shapes is recommended, as this involves broad movements, promotes visuomotor coordination, and allows children to build a general understanding of shapes without excessive pressure to be precise [19].

V. Conclusion

This work investigated the effect of CL on RL-based control of a musculoskeletal arm model for a shape-tracing task. We compared two curriculum orderings, Large-to-Small and Small-to-Large, against a no-curriculum baseline under identical training and evaluation conditions. Both curriculum strategies significantly outperformed the baseline in terms of success rate, waypoint progress, and reward, while also producing more consistent policies. The Large-to-Small ordering achieved the best overall performance, suggesting that acquiring broad motor patterns before refining precision is a more effective learning progression for musculoskeletal control. Generalization tests showed that curriculum-trained agents transfer well to unseen shapes and smaller scales, though all strategies remain sensitive to reversed traversal directions and scales beyond the training range.

Several directions remain for future work. First, the sensitivity to reversed direction suggests that the current design may encourage direction-specific strategies. We hypothesize that augmenting the observation with direction information about upcoming waypoints, or training with randomized traversal directions, could help the agent to acquire direction-invariant policies. Second, the curriculum used in this work relies on fixed thresholds. As a result, the agent cannot return to earlier levels if performance drops or skills are lost. An adaptive curriculum that adjusts task difficulty based on performance could address this limitation and improve robustness.

Acknowledgment

This work is supported by the European Union's HORIZON Research and Innovation Programme, grant agreement No 101120657, project ENFIELD (European Lighthouse to Manifest Trustworthy and Green AI). This work was granted access to the HPC resources of IDRIS under the allocation 2026-AD011011309R5 made by GENCI.

References

[1] M. Simos, A. S. Chiappa, and A. Mathis, "Reinforcement learning-based motion imitation for physiologically plausible musculoskeletal motor control," arXiv:2503.14637, 2025.

[2] V. Caggiano, S. Dasari, and V. Kumar, "MyoDex: A generalizable prior for dexterous manipulation," in *Proc. Int. Conf. Machine Learning (ICML)*, 2023, pp. 3327–3346.

[3] S. Song, Ł. Kidziński, X. B. Peng, et al., "Deep reinforcement learning for modeling human locomotion control in neuromechanical simulation," *Journal of NeuroEngineering and Rehabilitation*, vol. 18, no. 1, p. 126, 2021.

[4] V. Caggiano, H. Wang, G. Durandau, et al., "MyoSuite: A contact-rich simulation suite for musculoskeletal motor control," arXiv:2205.13600, 2022.

[5] H. Park and D. M. Durand, "Motion control of musculoskeletal systems with redundancy," *Biological Cybernetics*, vol. 99, no. 6, pp. 503–516, 2008.

[6] E. Todorov and M. I. Jordan, "Optimal feedback control as a theory of motor coordination," *Nature Neuroscience*, vol. 5, no. 11, pp. 1226–1235, 2002.

[7] V. Caggiano, G. Durandau, H. Wang, A. Chiappa, A. Mathis, P. Tano, et al., "MyoChallenge 2022: Learning contact-rich manipulation using a musculoskeletal hand," in *NeurIPS 2022 Competition Track*, 2023, pp. 233–250.

[8] Bengio, Y., Louradour, J., Collobert, R., Weston, J. (2009, June). Curriculum learning. In Proceedings of the 26th annual international conference on machine learning (pp. 41-48).

[9] S. Narvekar, B. Peng, M. Leonetti, et al., "Curriculum learning for reinforcement learning domains: A framework and survey," *Journal of Machine Learning Research*, vol. 21, no. 181, pp. 1–50, 2020.

[10] A. S. Chiappa, P. Tano, N. Patel, A. Ingster, A. Pouget, and A. Mathis, "Acquiring musculoskeletal skills with curriculum-based reinforcement learning," *Neuron*, vol. 112, no. 23, pp. 3969–3983, 2024.

[11] C. Florensa, D. Held, M. Wulfmeier, M. Zhang, and P. Abbeel, "Reverse curriculum generation for reinforcement learning," in *Conf. Robot Learning*, 2017, pp. 482–495.

[12] D. C. Crowder, J. Abreu, and R. F. Kirsch, "Improving the learning rate, accuracy, and workspace of reinforcement learning controllers for a musculoskeletal model of the human arm," *IEEE Trans. Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 30–39, 2021.

[13] B. Ivanovic, J. Harrison, A. Sharma, M. Chen, and M. Pavone, "BARC: Backward reachability curriculum for robotic reinforcement learning," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2019, pp. 15–21.

[14] V. Hsiao, M. Roberts, L. M. Hiatt, G. Konidaris, and D. Nau, "Automating curriculum learning for reinforcement learning using a skill-based Bayesian network," arXiv:2502.15662, 2025.

[15] P. Schumacher, D. Häufle, D. Büchler, S. Schmitt, and G. Martius, "DEP-RL: Embodied exploration for reinforcement learning in over-actuated and musculoskeletal systems," arXiv:2206.00484, 2022.

[16] W. B. Knox and P. Stone, "Reinforcement learning from simultaneous human and MDP reward," in *Proc. AAMAS*, 2012, pp. 475–482.

[17] R. S. Sutton and A. G. Barto, "Reinforcement learning: An introduction," MIT Press, 2018.

[18] J. Schulman, F. Wolski, P. Dhariwal, et al., "Proximal policy optimization algorithms," arXiv:1707.06347, 2017.

[19] M. Hadders-Algra, "Development of postural control during the first 18 months of life," *Neural Plasticity*, vol. 12, no. 2–3, pp. 99–108, 2005.