263-3300-10L Data Science Lab: Develop meta-RL policy for varying morphologies using proxy and task training

David Scherrer*
Department of Computer Science
ETH Zurich, Switzerland

Constantin Pinkl*
Department of Computer Science
ETH Zurich, Switzerland

Fatemeh Zargarbashi
ETH Computational Robotics Lab, Switzerland

Arnout Devos
ETH AI Center, ETH Zurich, Switzerland

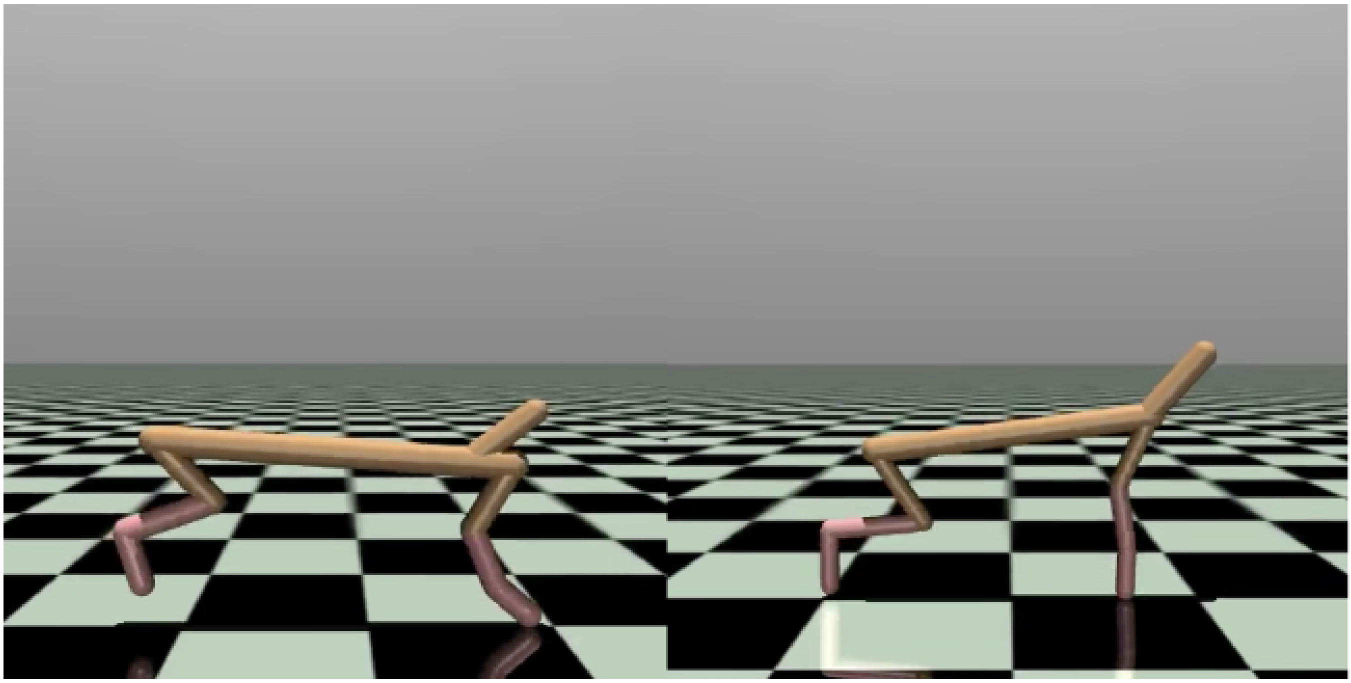

Figure 1: Different Cheetah robots with various body variations [2]

Abstract

We present a recurrent reinforcement learning approach for morphology-adaptive locomotion that leverages a sinusoidal height-tracking proxy task to bootstrap representation learning. To address the challenge of inferring embodiment parameters from sparse rewards, we employ a GRU-based policy within a two-phase episode structure: agents first solve a dense-feedback proxy task to encode morphology-dependent dynamics, then transfer this hidden state to the main locomotion objective. Experiments on randomized HalfCheetah environments demonstrate that this structured pretraining significantly outperforms standard recurrent and feed-forward baselines, achieving higher final returns and improved generalization to unseen morphologies.

Keywords

Reinforcement Learning, Robot Locomotion, Morphology Adaptation, Recurrent Policies, Meta-Reinforcement Learning

*Equal contribution.

263-3300-10L Data Science Lab, December 19, 2025, Zurich, Switzerland

DS: Methodology, Software, Visualization, Investigation, Writing - Original Draft
CP: Methodology, Software, Visualization, Investigation, Writing - Original Draft
FZ: Conceptualization, Supervision
AD: Supervision, Methodology, Project Administration

1 Introduction

Reinforcement learning has enabled impressive results in robotic control [7], ranging from simple balancing tasks to complex legged locomotion. However, a major limitation of many RL-based approaches is their lack of transferability: policies are typically trained from scratch for a specific robot morphology, dynamics configuration, and task. Even small changes in body proportions or physical parameters often require extensive retraining, leading to high computational cost and limited scalability.

To address this issue, recent research has focused on learning universal or morphology-general policies that can control a family of robots using a single shared controller. By training across diverse embodiments, such policies aim to amortize training cost and enable faster adaptation to new robots. This line of work is particularly relevant for legged locomotion, where morphology variations naturally arise due to differences in limb lengths, masses, or joint configurations.

Existing approaches to multi-morphology control largely fall into two categories. One class of methods explicitly provides morphological information - such as joint dimensions or physical parameters - as part of the policy input, often combined with attention mechanisms or specialized encoders. These approaches enable direct conditioning on robot structure but require access to accurate morphology descriptors at test time. A complementary line of work adopts a meta-reinforcement learning perspective, using recurrent architectures to implicitly infer morphology from interaction history. In this setting, the policy adapts online by encoding environment- and body-specific dynamics into its hidden state.

Despite these advances, learning effective locomotion policies across morphologies remains challenging when task rewards are sparse or difficult to optimize from scratch. In particular, recurrent policies may struggle to rapidly infer morphology if early interaction signals are weak or noisy. In this work, we propose to address this limitation by introducing a structured proxy task pretraining phase. Before optimizing the main forward-locomotion objective, the agent first solves a simpler proxy task designed to expose morphology-dependent dynamics through interaction.

We demonstrate this idea on a family of HalfCheetah morphologies, where a recurrent meta-RL policy first learns a sinusoidal height-tracking task and subsequently transfers its internal representation to the main locomotion task. Our results show that proxy task training leads to improved stability and higher final performance compared to training without proxy supervision, highlighting the benefits of structured interaction for morphology inference.

2 Related Work

Meta-learning aims to train models that can rapidly adapt to new tasks by acquiring general learning strategies. A prominent approach is Model-Agnostic Meta-Learning (MAML) [4], which optimizes a model initialization such that a small number of gradient updates yields strong performance on unseen tasks. While MAML is domain-agnostic and effective across supervised and reinforcement-learning settings, its reliance on gradient-based adaptation restricts the class of task-solving strategies it can represent.

An alternative paradigm treats meta-learning as a sequence modeling problem, where adaptation emerges implicitly from interaction history rather than explicit parameter updates. In this direction, Mishra et al. [9] propose the Simple Neural Attentive Learner (SNAIL), a domain-agnostic architecture that combines temporal convolutions with causal self-attention to overcome the memory limitations of recurrent meta-learners. Temporal convolutions provide high-bandwidth access to recent experience, while attention enables selective retrieval from long interaction histories. Unlike gradient-based methods such as MAML, SNAIL learns task-specific algorithms implicitly from interaction sequences and achieves state-of-the-art performance across a wide range of supervised and reinforcement-learning benchmarks without relying on hand-designed algorithmic priors.

Beyond architectural choices, auxiliary or proxy tasks have been widely used in reinforcement learning to shape representations and accelerate learning. Jaderberg et al. [6] introduce unsupervised auxiliary tasks that are optimized jointly with the main reinforcement-learning objective, demonstrating improved data efficiency and stability through easier proxy objectives. Similarly, Pathak et al. [10] propose a self-supervised prediction task as an intrinsic reward signal, using forward-dynamics prediction error to encourage exploration. These works highlight how proxy tasks can provide structured learning signals that induce useful inductive biases prior to or alongside optimization of the main task.

In the context of legged locomotion, meta-reinforcement learning has been explored as a mechanism for generalization across tasks and robot morphologies. Zargarbashi et al. [14] introduce MetaLoco, a GRU-based meta-RL framework for universal quadrupedal locomotion. By conditioning the policy on interaction history, the recurrent hidden state implicitly encodes morphology- and task-specific information, enabling a single controller to adapt online to different robot embodiments and motion objectives. MetaLoco primarily focuses on motion imitation and locomotion generalization, demonstrating that recurrent policies can infer latent system properties without explicit morphology inputs.

Building on this line of work, we extend recurrent meta-RL by introducing a proxy task pretraining phase that explicitly encourages morphology-dependent structure in the GRU [3] hidden state prior to optimizing the downstream locomotion objective.

Complementary approaches pursue morphology generalization through explicit representations. Bohlinger et al. [1] train a single feed-forward policy across diverse robot embodiments by explicitly encoding joint positions and physical attributes, which are aggregated via an attention mechanism. While effective, this approach relies on explicit morphology representations and does not maintain an interaction history. In contrast, our method learns morphology implicitly through recurrent interaction dynamics, without requiring explicit morphology encoding or attention-based aggregation.

Similarly, Luo et al. [8] propose a morphologically adaptive locomotion framework that uses a dedicated morphology network and privileged training signals to infer body and velocity parameters from recent interactions. Our approach instead avoids privileged information and explicit morphology supervision, relying on proxy task rewards to induce morphology-aware behavior within the recurrent hidden state.

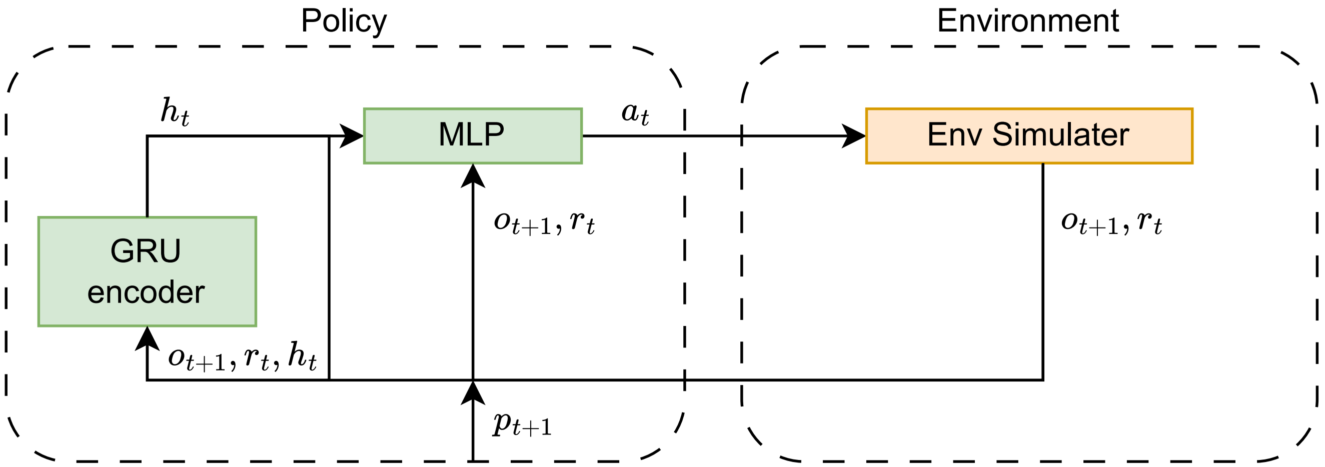

Figure 2: Policy architecture. A GRU integrates interaction history into a hidden state, which together with the current observation and task phase is mapped to actions by a feed-forward MLP.

3 Methodology

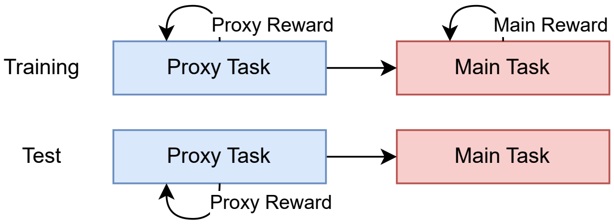

Figure 3: Training and evaluation procedure. The agent first solves a proxy task and subsequently transitions to the main task while maintaining the GRU hidden state. At test time, the same sequence is executed and proxy rewards remain available during the main task.

3.1 Problem setting.

Each episode samples an unobserved morphology $m \sim p(m)$ (e.g., link lengths and masses) that remains fixed and induces different dynamics. While the environment is Markov given m, the agent observes only standard HalfCheetah proprioception, resulting in partial observability. Our goal is to learn a single policy that generalizes to unseen morphologies by adapting online from interaction history.

3.2 Architecture

As illustrated in Fig. 2, the policy consists of a GRU [3] encoder followed by a feed-forward MLP, forming a recurrent meta-RL architecture. At each time step, the GRU receives the current observation, reward, previous action, and a phase indicator specifying whether the agent is in the proxy or main task. The resulting hidden state summarizes interaction history and captures morphology-dependent dynamics. This hidden state, together with the current observation and phase, is passed to the MLP to generate actions. The policy is trained using PPO [11] in an actor–critic framework.

3.3 Design Considerations

The proxy task is designed to bootstrap representation learning in settings where the main task reward may be sparse or difficult to optimize directly. By first solving a structured, morphology-sensitive task, the recurrent policy acquires a hidden state that accelerates and stabilizes learning on the main task across varying morphologies (Fig. 3).

4 Experimental Setup

4.1 Environment Configuration

Setup. Training follows a two-phase structure within each episode, shown in Fig. 3. The agent first solves a proxy task and then transitions to the main locomotion task while preserving the GRU hidden state. This design allows information acquired during the proxy phase to directly influence the downstream task. To encourage generalization, the HalfCheetah morphology is periodically randomized during training, forcing the recurrent policy to adapt across diverse embodiments rather than overfitting to a single configuration.

Proxy task: sinusoidal height tracking. During the proxy phase, the agent tracks a sinusoidal target torso height by raising and lowering its body while remaining approximately stationary. This task is fully observable and provides dense reward signals, in contrast to the potentially sparse or delayed rewards of the main locomotion objective. As a result, it promotes morphology-aware control and encourages the GRU to infer body proportions and dynamics through interaction.

The proxy reward is defined as:

$$r_{\text{proxy}}(t) = w_1 \exp\left(-\frac{(h(t) - h_{\text{target}}(t))^2}{\sigma_h}\right) + w_2 \exp\left(-\frac{v(t)^2}{\sigma_v}\right) \quad (1)$$

$$h_{\text{target}}(t) = h_{\text{base}} + A \sin\left(\tfrac{2\pi t}{T}\right) \quad (2)$$

The target height is defined relative to the robot's initial torso height, which varies across morphologies due to changes in leg length. Empirically, we set the base height to 80% of the initial torso height, reflecting the observation that the robot's legs are initially fully extended and that stable behavior occurs below this maximum height. The proxy reward combines a height-tracking term with a small velocity penalty to discourage forward and backward motion during this phase.

Main task: forward locomotion. In the main task, the agent is rewarded for fast and stable forward locomotion. The reward encourages forward velocity and includes a small uprightness bonus to promote stability. Episodes are terminated early when the robot falls, preventing learning from unstable or uninformative states. No explicit morphology information is provided, requiring the agent to rely on the representation formed during proxy task interaction.

$$r_{\text{main}}(t) = w_3 \exp\left(\tfrac{v_x(t)}{\sigma_f}\right) + w_4 \exp\left(\frac{\max(0,\ z_{\text{torso}}(t) \cdot z_{\text{world}} - m)}{(1 - m)\sigma_u}\right). \quad (3)$$

The exact hyperparameters we used in our experiments can be viewed in Appendix A.

Observation space. At each time step, the policy receives the default MuJoCo HalfCheetah [2, 13] observation vector, augmented with three additional signals:

(1) A binary task-phase indicator specifying whether the agent is currently in the proxy or main task.
(2) The current target torso height used in the proxy task.
(3) The reward obtained at the previous time step.

These additional signals allow the recurrent policy to associate interaction patterns with task objectives and reward structure, facilitating adaptation when transitioning between proxy and main tasks.

Environment and tasks. All experiments are conducted in a MuJoCo-based HalfCheetah environment. To evaluate generalization across embodiments, the HalfCheetah morphology is randomized during training by independently scaling torso and limb dimensions and masses. Morphologies are resampled periodically, ensuring that the policy cannot overfit to a single body configuration and must instead infer morphology from interaction history.

4.2 Training & Evaluation

Training protocol. Policies are trained using PPO [11] with a recurrent actor–critic architecture [5]. Each episode consists of a proxy phase followed by a main task phase, with the GRU hidden state preserved across phases. All results are averaged over multiple random seeds. To successfully train the architecture, we reset the robot every time it fails, even when it fails in the proxy phase. This encourages the agent to perform such that it actually survives the proxy phase. The hidden state we of the GRU we carry over to the main task phase.

Evaluation protocol. We evaluate the learned policies on a held-out set of HalfCheetah embodiments that are not seen during training. These evaluation morphologies vary torso length, radius, and mass, as well as front and back leg segment lengths and masses, including asymmetric configurations. All scaling factors are sampled relative to the default HalfCheetah model, typically within the range [0.7, 1.3], resulting in substantial variation in body proportions and dynamics.

Policies are evaluated on a fixed set of 17 held-out morphologies, with performance measured over multiple episodes per morphology. We report mean and variance of episode returns across scenarios to assess robustness and generalization beyond the training distribution. Detailed morphology parameters and evaluation settings are provided in the appendix.

In contrast to the training protocol, we do not end the episode if the agent fails in the proxy phase, but just reset the HalfCheetah to its initial position, maintain the hidden state and directly move on to the main task. This yielded much higher rewards as it ensures that we always reach the main task.

5 Results

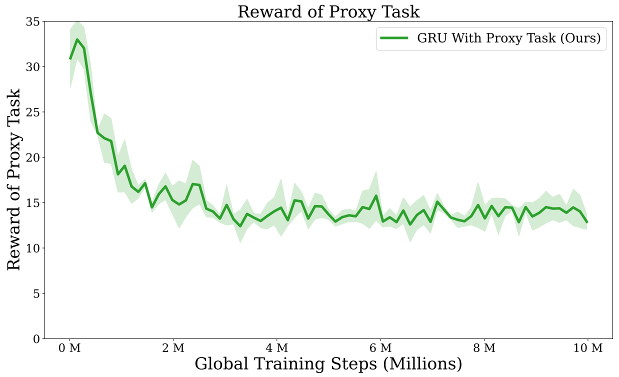

Figure 4: Proxy task reward over training, averaged across all evaluation scenarios. Only *GRU With Proxy Task* achieves non-zero proxy reward, which decreases as the agent shifts focus toward the main task. The experiment was conducted over 3 seeds.

5.1 Proxy Task Analysis

Fig. 4 shows the evolution of the proxy task reward during training, averaged across all evaluation scenarios. As expected, only the *GRU With Proxy Task* model achieves non-zero proxy reward, confirming that the proxy objective is active during training and that task-phase conditioning functions as intended.

At the beginning of training, the proxy reward is relatively high. This is largely due to the conservative dynamics of the HalfCheetah early in training: the agent remains mostly upright with limited forward motion. By staying stable and within a reasonable range of the sinusoidal target height, the agent receives reward from the height-tracking term while incurring only a small velocity penalty. This behavior alone does not imply that the agent actively tracks or optimizes the full sinusoidal height trajectory.

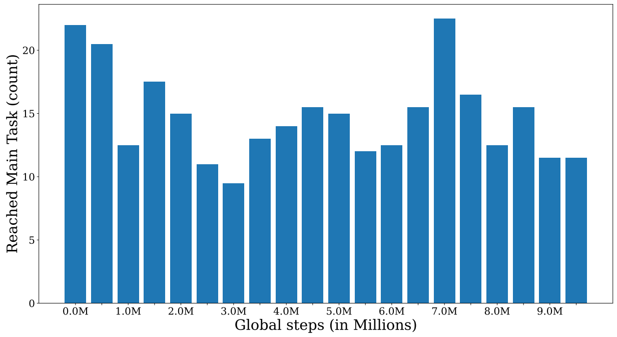

Figure 5: Frequency of successful transitions from the proxy phase to the main task over training, aggregated in one-million–step bins.

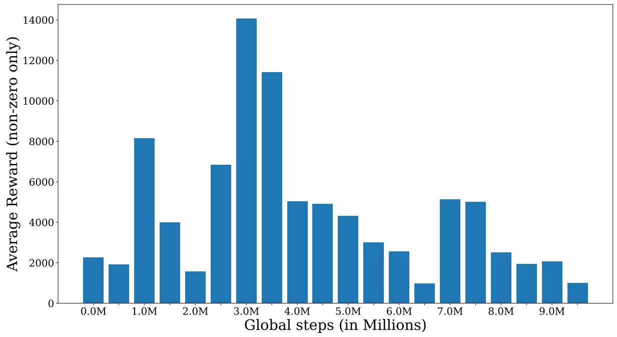

Figure 6: Average main task reward during training for episodes in which the agent reaches the main task. Although rewards do not consistently increase over time, the agent exhibits stable returns in later training stages.

As training progresses, the proxy reward decreases and stabilizes at a lower level. This decline coincides with the emergence of forward running behavior during the proxy phase. As forward velocity increases, the velocity penalty in the proxy reward becomes more prominent, leading to reduced proxy returns. In this later regime, the agent no longer consistently optimizes the proxy objective.

Overall, these results indicate that the proxy task is not optimized as a persistent control objective throughout training. Instead, proxy reward is highest early on and diminishes as the policy increasingly prioritizes behaviors that are beneficial for the downstream locomotion task.

Further, we analyze how often the agent transitions from the proxy phase to the main task. Fig. 5 reports the frequency of successful transitions over training. We observe no clear increasing trend in how often the agent reaches the main task as training progresses.

5.2 Main Task Analysis

Although the agent does not frequently reach the main task during training, as shown in Fig. 5, these transitions are nevertheless sufficient for learning to solve the main task. In Fig. 6, we report the reward obtained after reaching the main task across all runs. We observe high variance early in training, which decreases substantially over time. While absolute reward values around 3.5 million steps are higher, returns toward the end of training exhibit reduced variance and more closely match those observed during evaluation.

Method	Reward (Mean)	Reward (Std)
Proxy + 448 steps (ours)	1262.80	170.24
Proxy + 512 steps (ours)	1773.93	173.45
512 steps (MetaLoco)	535.81	70.16
512 steps (MLP)	162.03	25.34

Table 1: Final main task performance across HalfCheetah morphologies. We report the mean and standard deviation of episode returns over multiple evaluation scenarios and random seeds. proxy task pretraining substantially improves both final performance and robustness compared to MetaLoco and feed-forward baselines, with longer proxy horizons yielding the highest returns.

Fig. 7 reports the main task reward during training for all considered methods. The *MLP No Proxy Task* baseline exhibits slow learning and achieves limited performance, highlighting the difficulty of multi-morphology locomotion without recurrence or structured supervision.

The *GRU No Proxy Task* baseline initially improves rapidly and reaches a high reward early in training. However, its performance peaks and subsequently declines, suggesting unstable adaptation across varying morphologies.

In contrast, the *GRU With Proxy Task* model shows slower initial improvement but continues to improve steadily throughout training, ultimately achieving the highest final performance across evaluation scenarios.

Across all evaluation scenarios, proxy-based training consistently outperforms both feed-forward and recurrent baselines without proxy supervision, as summarized in Tab. 1.

6 Discussion

The results show that proxy task training substantially improves final performance and robustness of recurrent policies across morphologies, despite the proxy task itself not being optimized as a persistent control objective. In particular, proxy reward peaks early in training and declines as the policy increasingly prioritizes forward locomotion. This behavior should not be interpreted as failure or forgetting of the proxy task, but rather as a shift in objective weighting during training.

We argue that the benefit of the proxy task arises not from learning to solve the proxy objective itself, but from early exposure to a dense, structured interaction regime. During the proxy phase, the agent experiences morphology-sensitive feedback related to vertical stability and body configuration. This exposure encourages the recurrent policy to internalize information relevant for morphology-dependent control, even though proxy reward optimization later diminishes.

The persistence of improved main-task performance, even when proxy reward is no longer maximized, suggests that the proxy phase primarily influences how recurrent representations are formed early in training. In contrast, recurrent policies trained without proxy

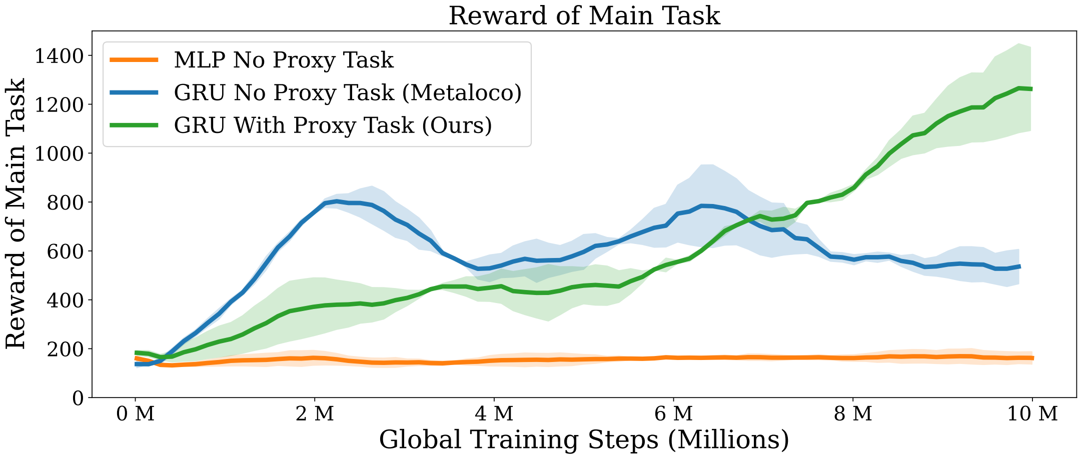

Figure 7: Main task reward during evaluation, averaged across all evaluation scenarios. *GRU With Proxy Task* achieves the highest final performance, surpassing *GRU No Proxy Task*, which peaks early and then declines. *MLP No Proxy Task* shows slow and limited improvement, highlighting the advantage of recurrent policies and proxy-based training. The experiment was conducted over 3 seeds.

supervision rely solely on sparse locomotion rewards, which may encourage premature specialization and lead to unstable adaptation across morphologies.

Additionally, we isolate the effect of the proxy task by evaluating agents trained with proxy task pretraining on the main task, both with and without executing the proxy phase at evaluation time. We find that performing the proxy task prior to the main task consistently yields higher returns and lower variance across random seeds (Appendix B).

A key limitation of the current study is that the proxy task is manually designed and tailored to HalfCheetah. Future work should investigate alternative proxy objectives—such as pitch stabilization, energy-efficient standing, or controlled leaning—to assess how different interaction primitives influence representation learning. Extending the approach to full quadrupedal robots would further validate the generality of proxy-based meta-reinforcement learning.

7 Conclusion

We presented a recurrent meta-RL approach for morphology-adaptive locomotion that uses a structured proxy task phase to bootstrap representation learning before optimizing a downstream locomotion objective. Our method trains a GRU-based PPO policy in a two-phase episode structure: a sinusoidal height-tracking proxy task followed by forward locomotion, while preserving the GRU hidden state across phases. The proxy task provides dense, morphology-sensitive feedback without requiring explicit embodiment descriptors or privileged information. Experiments on a morphologically randomized HalfCheetah suite show that proxy task training improves learning stability and yields higher final returns compared to both a feed-forward baseline and a recurrent baseline trained without proxy supervision, with improved robustness on held-out and asymmetric morphologies.

7.1 Limitations.

Our training protocol resets the environment upon failure during the proxy phase, whereas evaluation proceeds by resetting state and continuing to the main task while preserving the hidden state. This mismatch improves evaluation stability but may inflate performance relative to a fully continuous episode setting. Additionally, the proxy task is manually designed and may not transfer to environments where torso height is less informative. Finally, while we observe reduced variance later in training, early learning remains unstable across seeds, suggesting sensitivity to PPO hyperparameters and exploration during the proxy phase.

7.2 Future Work

A first direction is to transfer this framework from the planar HalfCheetah setting to full quadruped locomotion (e.g., Unitree/A1-style simulations), where morphology variation and stability challenges are more realistic and the benefits of recurrent adaptation may be stronger. Second, we plan to evaluate alternative proxy tasks and study which proxy properties are most predictive for downstream transfer. Beyond sinusoidal height tracking, promising candidates include the following:

1. *Pitch/lean control* leaning forward / backward and recovering
2. *Stance stabilization* under external pushes or uneven ground
3. *Gait primitives* such as stepping-in-place or controlled foot-lifting, which directly probe morphology-dependent contact dynamics.

Third, we will conduct more systematic ablations on design choices, including proxy duration, whether proxy rewards remain active during the main task, and the role of phase indicators, to better isolate when proxy pretraining helps or harms. Finally, an important extension is to test sim-to-real practicality by removing assumptions that may not hold on hardware (e.g., accurate state estimates) and by assessing robustness under sensor noise and actuation delays.

References

[1] Nico Bohlinger, Grzegorz Czechmanowski, Maciej Krupka, Piotr Kicki, Krzysztof Walas, Jan Peters, and Davide Tateo. One policy to run them all: an end-to-end learning approach to multi-embodiment locomotion. *arXiv preprint arXiv:2409.06366*, 2024.

[2] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[3] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[4] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[5] Nicolas Heess, Jonathan J Hunt, Timothy P Lillicrap, and David Silver. Memory-based control with recurrent neural networks, 2015. URL https://arxiv.org/abs/1512.04455.

[6] Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki, Tom Schaul, Joel Z Leibo, David Silver, and Koray Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. *arXiv preprint arXiv:1611.05397*, 2016.

[7] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[8] Zeren Luo, Yinzhao Dong, Xinqi Li, Rui Huang, Zhengjie Shu, Erdong Xiao, and Peng Lu. Moral: Learning morphologically adaptive locomotion controller for quadrupedal robots on challenging terrains. *IEEE Robotics and Automation Letters*, 9:4019–4026, 2024. URL https://api.semanticscholar.org/CorpusID:268409280.

[9] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. *arXiv preprint arXiv:1707.03141*, 2017.

[10] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pages 2778–2787. PMLR, 2017.

[11] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[12] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE, 2017.

[13] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 5026–5033. IEEE, 2012.

[14] Fatemeh Zargarbashi, Fabrizio Di Giuro, Jin Cheng, Dongho Kang, Bhavya Sukhija, and Stelian Coros. Metaloco: Universal quadrupedal locomotion with meta-reinforcement learning and motion imitation, 2024. URL https://arxiv.org/abs/2407.17502.

Appendices

A Additional Implementation Details

A.1 Observation Space Augmentation

The base observation vector provided by the MuJoCo HalfCheetah environment is augmented with additional task-specific signals to support recurrent representation learning. At each time step t, the observation is defined as

$$\mathbf{o}_t = \left[\mathbf{o}_t^{\text{env}},\ \phi_t,\ h_{\text{target}}(t),\ r_{t-1}\right],$$

where $\mathbf{o}_t^{\text{env}}$ denotes the original environment observation, $\phi_t \in \{0, 1\}$ is a phase indicator specifying whether the agent is in the proxy task (0) or the main task (1), $h_{\text{target}}(t)$ is the current target torso height during the proxy phase, and r_{t-1} is the reward obtained at the previous time step.

Including the phase indicator and previous reward allows the recurrent policy to condition its hidden state on task context and reward feedback, facilitating implicit inference of task structure and morphology-dependent dynamics.

A.2 Reward Functions and Termination Criteria

Proxy task. The proxy task reward consists of a height-tracking term and a velocity penalty,

$$r_{\text{proxy}}(t) = w_1 \exp\left(-\frac{(h(t) - h_{\text{target}}(t))^2}{\sigma_h}\right) + w_2 \exp\left(-\frac{v_x(t)^2}{\sigma_v}\right),$$

with weights $w_1 = 1.0$ and $w_2 = 0.2$. The target height follows a sinusoidal trajectory

$$h_{\text{target}}(t) = h_{\text{base}} + A \sin\left(\tfrac{2\pi t}{T}\right),$$

where h_{base} is set to 0.8 times the initial torso height and A is scaled proportionally to the morphology-dependent torso height.

Episodes are terminated early during the proxy phase if the torso deviates more than a fixed threshold from the origin along the forward axis, preventing degenerate behaviors.

Main task. The main task reward encourages forward locomotion and upright posture,

$$r_{\text{main}}(t) = w_3 \exp\left(\tfrac{v_x(t)}{\sigma_f}\right) + w_4 \exp\left(\frac{\max(0,\ z_{\text{torso}}(t) \cdot z_{\text{world}} - m)}{(1 - m)\sigma_u}\right),$$

with $w_3 = 1.0$ and $w_4 = 0.2$. Episodes terminate early when the robot falls or collapses, defined by torso height and orientation thresholds.

A.3 Morphology Randomization and Evaluation Scenarios

During training, morphology parameters are randomized periodically to encourage generalization [12]. Torso length, radius, and mass, as well as front and back leg segment lengths, radii, and masses, are independently scaled relative to the default HalfCheetah morphology.

For evaluation, we consider a fixed suite of 17 held-out embodiments, including symmetric and asymmetric configurations. Scaling factors typically lie in the range $[0.7, 1.3]$ of the default morphology. These scenarios include variations such as elongated torsos, heavier front or back legs, asymmetric limb lengths, and altered segment radii.

In addition to the fixed suite, we evaluate on randomly sampled morphologies drawn from truncated normal distributions centered at the default embodiment. This provides a complementary assessment of robustness to unseen morphologies.

A.4 Optimization Details

All policies are trained using Proximal Policy Optimization (PPO). We use a learning rate of 2.5×10^{-4}, a discount factor $\gamma = 0.99$, and GAE parameter $\lambda = 0.95$. Policy updates are performed over 10 epochs with a clipping coefficient of 0.2. The entropy and value loss coefficients are set to 0.001 and 0.5, respectively.

Training is performed with 32 parallel environments and rollout lengths of 512 steps. GRU hidden states are reset on episode termination.

B Approach Validation

We validate our approach by comparing agent performance with and without the proposed proxy task. In both settings, the final policy is evaluated on the main task using six independent random seeds.

As shown in Figure 8, executing the the proxy task leads to consistently higher returns and substantially reduced variance across seeds. This indicates that the proxy task provides a more structured and stable initialization, enabling faster convergence and improved robustness during downstream training on the main task.

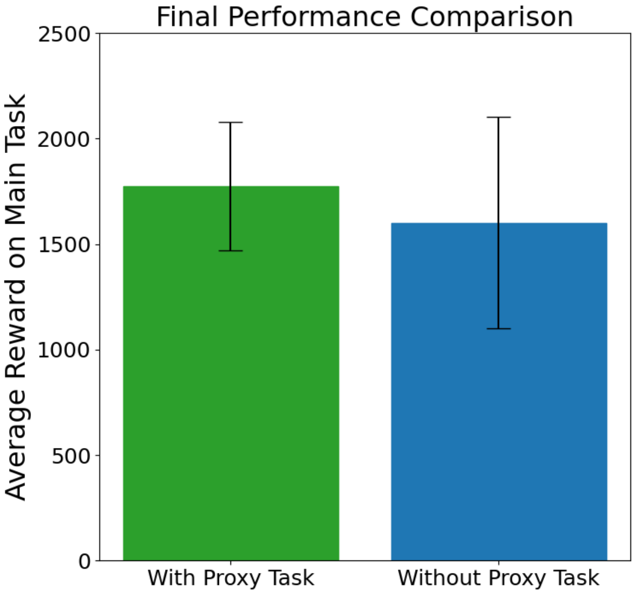

Figure 8: Performance comparison of our best agent trained with proxy task versus directly solving the the main task. proxy task initialization yields higher average returns and lower variance across six random seeds.