

# Imitate and Repurpose: Learning Reusable Robot Movement Skills From Human and Animal Behaviors

Steven Bohez<sup>\*</sup>, Saran Tunyasuvunakool, Philemon Brakel, Fereshteh Sadeghi, Leonard Hasenclever, Yuval Tassa, Emilio Parisotto, Jan Humplik, Tuomas Haarnoja, Roland Hafner, Markus Wulfmeier, Michael Neunert, Ben Moran, Noah Siegel, Andrea Huber, Francesco Romano, Nathan Batchelor, Federico Casarini, Josh Merel<sup>†</sup>, Raia Hadsell and Nicolas Heess

DeepMind, London, UK

We investigate the use of prior knowledge of human and animal movement to learn reusable locomotion skills for real legged robots. Our approach builds upon previous work on imitating human or dog **Motion Capture (MoCap)** data to learn a movement skill module. Once learned, this skill module can be reused for complex downstream tasks. Importantly, due to the prior imposed by the **MoCap** data, our approach does not require extensive reward engineering to produce sensible and natural looking behavior at the time of reuse. This makes it easy to create well-regularized, task-oriented controllers that are suitable for deployment on real robots. We demonstrate how our skill module can be used for imitation, and train controllable walking and ball dribbling policies for both the ANYmal quadruped and OP3 humanoid. These policies are then deployed on hardware via zero-shot simulation-to-reality transfer. Accompanying videos are available at <https://bit.ly/robot-npmp>.

## Introduction

Animals and humans are masters at using their legs to reach even the most remote places on earth, escape predators, chase prey, play sports, or dance. Inspired by the versatility and efficiency of animal and human locomotion, research into legged robots dates back multiple decades. Nevertheless, legged robots continue to fall short of the agile and flexible motions displayed by animals and humans. And while this can, in part, be explained by hardware limitations, current control systems are also limited in their ability to produce a wide range of useful and agile behaviors.

Many existing locomotion controllers employ modular designs in which multiple hand-tuned control modules interact [1, 2]. Such approaches are typically task specific and require significant engineering effort. Their reliance on models of limited accuracy or simplifying assumptions can restrict the modules to regions of the state space where these approximations are valid, thus constraining the pose or movement of the robot.

Trajectory optimization, e.g. combining a motion planner and a tracking controller, can be

used for gait discovery with less manual engineering than modular approaches [3, 4, 5]. However, planning dynamic gaits with discrete contacts is a hard optimization problem, and more advanced formulations quickly become too slow for real-time applications. Performance is also limited by the capabilities of the tracking controller.

Partially in response to these limitations, there has been growing interest in learning-based methods. These methods can amortize the cost of complex optimization strategies at training time into neural-network (and other) controllers that are cheap to evaluate during deployment, and they can find flexible solutions that generalize to novel situations. A number of studies in simulation have demonstrated how **Reinforcement Learning (RL)** techniques can be deployed to generate complex movement strategies including perception-action coupling and object interaction [6, 7, 8, 9]. Since learning directly on the hardware raises concerns related to data efficiency and safety (e.g. [10, 11, 12]), the majority of works with real robots have focused on approaches that learn locomotion skills in simulation, and then transfer the resulting con-

<sup>\*</sup>Corresponding author: [sbohez@deepmind.com](mailto:sbohez@deepmind.com)

<sup>†</sup>Currently at Meta Reality Labs, work done while at DeepMind.

trollers to the hardware [13, 14, 15, 16]. Examples include the traversal of rough terrain with a quadruped [17, 18] and robust walking and stair climbing with a biped [19, 20, 21]. With a sufficiently accurate simulation model (for instance via learned actuator models [13]) and policies that are robust to or can adapt to distribution shift (e.g. due to the use of randomized simulation models [22, 23, 24, 25]) successful transfer can be achieved even without further optimization during deployment (but see, for instance, [26, 27]).

However, generating movements that are both functional and safe for hardware deployment remains a major challenge. Locomotion strategies obtained with RL are often idiosyncratic and unnatural looking, and may exhibit, for instance, high torques and jerk (see e.g. [6, 28]), thus leading to rapid wear and tear or catastrophic hardware failures. Moreover, they are not energy efficient, and may thus rapidly drain a robot’s battery and reduce its autonomy.

Previous studies have therefore devoted significant effort to developing suitable objective functions and regularization strategies, often through cost terms that encourage or discourage particular behavioral features [13, 19, 14]. While such strategies can be effective, they often require extensive manual tuning. Constrained [28] or multi-objective [29] optimization techniques make it easier to trade-off different objectives, but they still require a good understanding of the relevant objectives. A partial remedy is to take inspiration from central pattern generators and restrict the action space of the controller to well-behaved periodic foot trajectories [17, 30]. However, such approaches restrict the flexibility and diversity of the learned motions.

In this work, we introduce a novel approach to learn a diverse set of reusable motor skills for legged robots based on natural human and animal movements. The learned skills are versatile so that they can be used for a variety of different locomotion tasks, and they are robust such that they can be transferred to the real robot while maintaining the desired smooth and natural looking motion styles. Our approach alleviates the need for carefully designed learning objectives or regu-

larization strategies when training task-oriented controllers and constitutes a general strategy for learning useful and functional robot skills.

Our work is inspired by ideas developed in the literature on character animation where human and animal **Motion Capture (MoCap)** data has been used extensively to enable controllable behaviors with smooth *transitions* between different movements, both via kinematic animation [31, 32] and via controllers used in physical simulation models [7, 33, 34, 35]. Going a step further, previous works such as [36] or [37, 38, 39] have investigated strategies to use **MoCap** data as a general purpose prior to constrain movements of downstream **RL** tasks, including complex whole body control and object interaction [8, 9].

We extend previous work on **Neural Probabilistic Motor Primitives (NPMP)** [37, 39] to develop a general purpose movement skill module for legged robots. Like [40] and [26], we use **MoCap** to create controllers for real robots. However, whereas previous work focused on reproducing individual movement trajectories from the data on a robot, we train multi-purpose skill modules which capture general properties of the **MoCap** data. These modules can help to learn or execute new tasks more efficiently, and they can also be used to impose constraints on the final solution. Concurrent work [41] has similarly considered **MoCap** data to regularize behavior in downstream tasks, but learns the task directly with an additional adversarial objective instead of explicitly training a skill module via imitation first.

To show the generality of our approach we create movement skill modules for two very different legged robots: ANYmal B300 [42], a large quadruped, and OP3 [43], a small humanoid. We apply the skill modules to several different downstream tasks and demonstrate how the same module can serve as an inductive bias that encourages functional and natural looking behavior in different contexts. We train the skill modules entirely in simulation but demonstrate that they can be transferred successfully to the real robots.

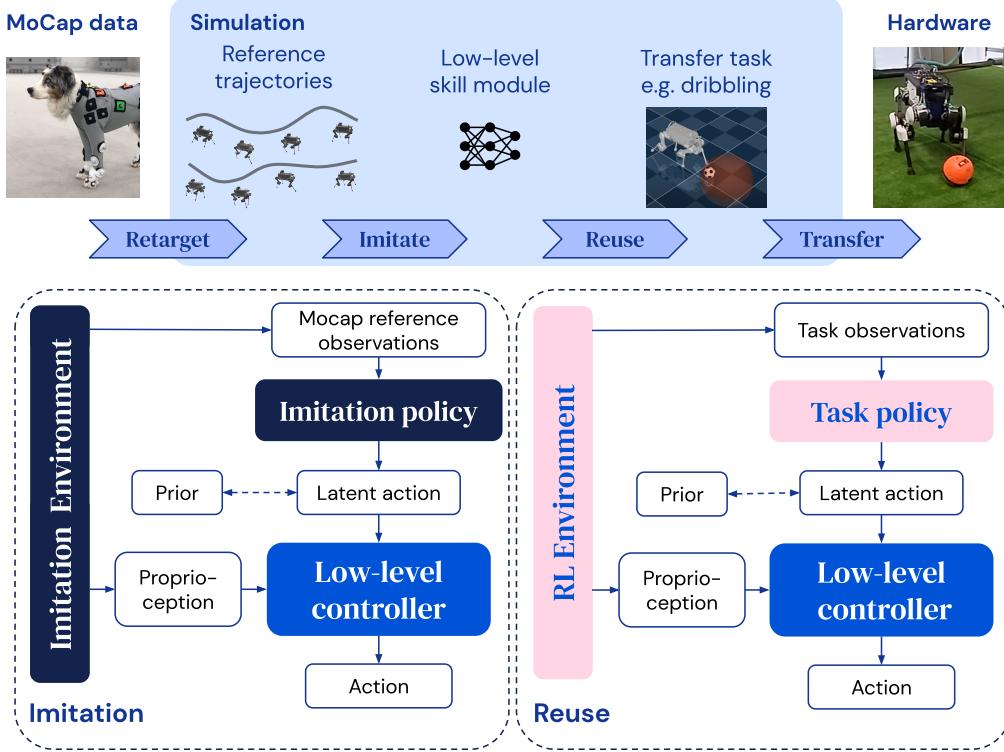


Figure 1 | Our approach consists of four stages: 1) First, we retarget human or dog MoCap data to the ANYmal or OP3 robots. 2) Next, we train a policy to imitate the reference trajectories in simulation. This policy has a hierarchical structure in which a tracking policy encodes the desired reference trajectory into a latent action that subsequently instructs a proprioception-conditioned low-level controller. 3) We can now reuse the low-level controller by training a new task policy to output latent actions to instruct the low-level controller whose parameters are kept fixed. This enables us to solve challenging tasks such as ball dribbling. 4) Finally, we transfer our resulting controllers from simulation to real hardware in zero-shot fashion. This is realized by the use of accurate simulation models as well as dynamics and domain randomization in simulation.

## Method Overview

An overview of our approach is shown in Figure 1. First, we retarget the **MoCap** clips to the respective robot body. For ANYmal we rely on a dataset of dog **MoCap** of mostly walking and turning behaviors [33]; for the OP3 we use a subset of the clips from the CMU dataset [44]. In the second step, we train a universal goal-conditioned policy to imitate different **MoCap** clips. This policy consists of an encoder and a decoder network. The encoder learns to map a sequence of future reference trajectory frames onto a skill embedding in a latent skill space that represents desired future movements relative to the current pose. The decoder, or low-level controller, learns to map this skill embedding to joint actuator commands that achieve the desired future reference poses tak-

ing the current proprioceptive state of the robot into account. Both components are trained together end-to-end, and the low-level controller can subsequently be used as a skill module for downstream tasks by training a task-specific policy to select actions directly in the latent skill space. To ensure that the latent skill space is well-formed and thus facilitates re-use of the skills, we train the imitation policy as an information bottleneck architecture and regularize the induced distribution over latent action sequences by penalizing the divergence to a Gaussian **Order 1 Autoregressive (AR(1))** prior [37, 45].

The result is a general imitation policy that can reproduce a large number of different motion capture trajectories. Using the decoder as a skill module then restricts the search space of all pos-

sible behaviors to a smaller but diverse subset of mostly natural looking and functional ones. This allows more focused exploration for compatible downstream tasks and prevents undesirable solutions. Finally, the decoder network together with the prior can be seen as a policy that randomly generates human-like (or dog-like) movements, i.e. a generative model of diverse human- or dog-like behaviors [46, 45].

We train both the skill module and the task specific policy entirely in simulation but enable the controller to rapidly adapt to the robot hardware during deployment. Similar to [17, 18, 47] we train a history-conditional controller (parameterized with a recurrent network) and randomize various simulation properties during both the imitation and reuse phases. To increase robustness, we apply additional perturbations and add other task-relevant variations to the environment, such as procedurally generated terrain. We evaluate the skill modules on zero-shot trajectory imitation and two types of downstream tasks: controllable walking and ball dribbling. We evaluate our controllers both in simulation and on hardware. Overall, our results suggest that the skill modules derived from MoCap data of humans or animals are a practical component of a general purpose solution strategy for locomotion and whole-body control problems for high-dimensional robots. They can provide an alternative to conventional regularization and shaping strategies, and as reusable components with learned interfaces that can be adapted to a broad range of different, naturalistic target motions, they can be employed as components of non-standard control hierarchies.

## Contributions

The contributions of our work can be summarized as follows:

1. We showcase human and animal locomotion as a suitable prior to control legged robots.
2. We develop an approach for training skill modules in simulation which can subsequently be transferred to real robots. The skill modules enable learning of downstream tasks with minimal additional task specific regularization during training.

3. We provide an extensive experimental evaluation, including both a humanoid and a quadruped platform and multiple movement tasks. The experiments confirm the effectiveness of our approach and highlight that the skill modules produce naturalistic and functional movements in interesting downstream tasks and can effectively transfer to real robots.

## Results

We train two separate movement skill modules following the approach explained in Figure 1 for two different robot platforms: (1) the ANYmal quadruped, based on dog MoCap data, (2) the OP3 humanoid, using human MoCap data. We then use these skill modules to solve several different tasks in simulation and transfer the resulting controllers to real robots. Figure 2 shows some of the highlights of our results. Video A<sup>1</sup> shows a summary of the approach and results. Below we discuss individual results in more detail. We focus our real-world analyses on the ANYmal robot, which suffers less from the simulation-to-reality transfer gap due to its well-engineered design and which therefore allows for more reliable zero-shot evaluation [13]. The OP3 humanoid is a much more low-cost robot which results in a number of properties that complicate simulation-to-reality transfer such as backlash and increased sensitivity to battery charge and health [27]. For OP3 we therefore focus our analysis on results from a realistic simulation to demonstrate that our method produces functional movements for robots with very different morphologies and other properties. We also show that the resulting behavior can transfer to the real robot but do not provide a complete analysis.

## Motion Imitation

First, we examine how well the skill modules imitate specific trajectories. To this end, we use the full goal-conditioned imitation policy including encoder and decoder (skill module). We sequentially encode snippets of MoCap trajectories into

<sup>1</sup>Accompanying videos are available at <https://bit.ly/robot-npmp>.

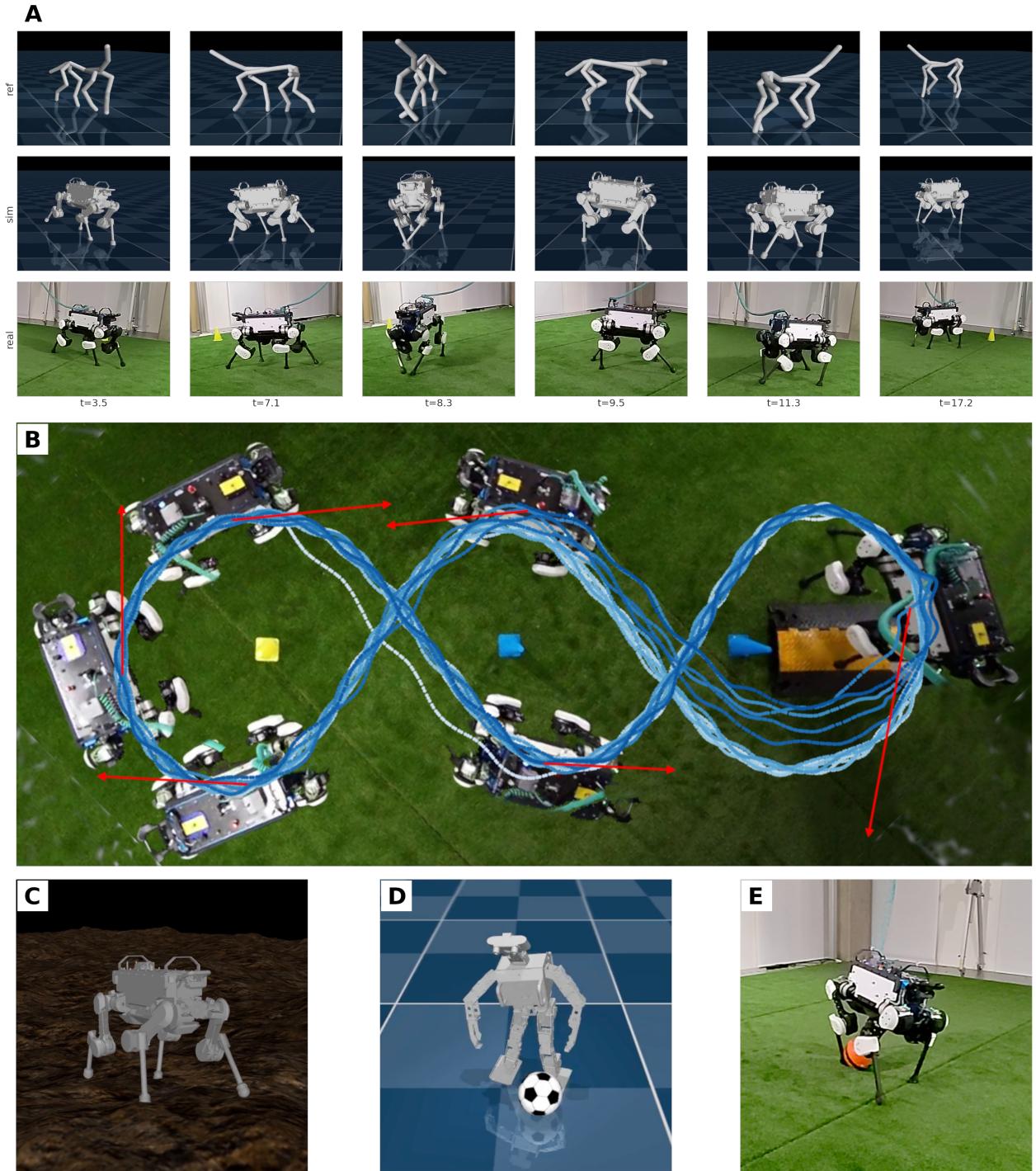


Figure 2 | Result highlights. **(A)** Imitation of dog MoCap by ANYmal. The top row shows a visualization of the original reference, the middle row imitation in simulation and the bottom row imitation in reality. **(B)** Top-down view of ANYmal following a slalom trajectory using a trajectory-following controller on top of a controllable-walking policy. The color gradient (light to dark) indicates position over time. Arrows indicate the target velocity of the blended keyframes. The figure highlights the accuracy and consistency with which the policy follows the instructions. An obstacle (right-hand side of the figure) is introduced during the trial. This forces the controller to locally deviate from the instructed trajectory but it recovers within a half-turn. **(C)** Reusing the low-level controllers for controllable walking on hilly terrain with ANYmal as well as dribbling with OP3 **(D)** and ANYmal **(E)** in simulation and real, respectively.

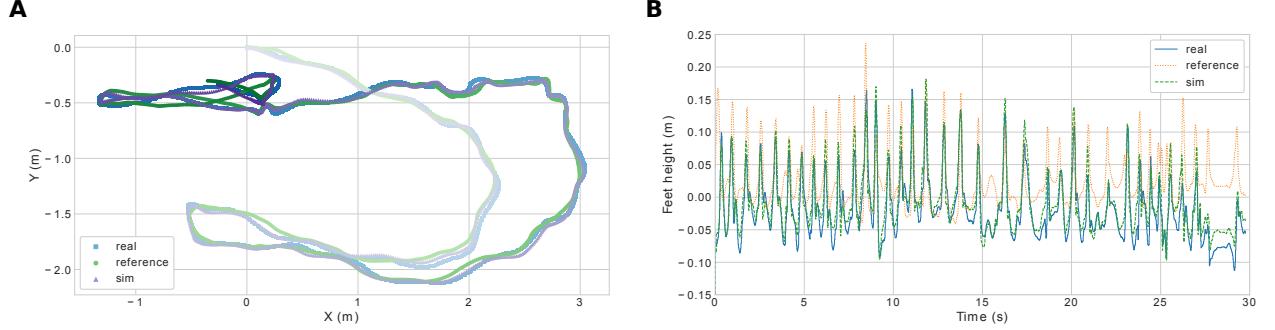


Figure 3 | Zero-shot imitation on ANYmal. (A) Top-down view of the path traversed by the base of the robot while following a MoCap reference clip, in simulation and reality. Color darkness is proportional to time. (B) Comparison of the height of the left-front foot over time for the MoCap reference, the simulated robot and the real robot.

latent commands for the skill modules. As explained above, each of these latent commands effectively describes the desired change in the robot pose. The latent commands are then executed by the skill module which acts as a feedback controller and generates the corresponding joint-level commands that produce the desired change in pose as closely as possible.

We first test zero-shot imitation in simulation. Videos B and C show results for ANYmal and OP3 respectively. The skill modules allow the simulated robots to faithfully track the occasionally rather rapid and agile dog and human movements. We then test whether zero-shot imitation of dog MoCap trajectories is also feasible with the real ANYmal robot. As Video D demonstrates, the dog movements are faithfully imitated by the robot. Figure 2A shows a number of corresponding keyframes from the reference trajectory, imitation in simulation and imitation on hardware respectively, indicating the close correspondence with the reference and between simulation and reality.

A quantitative analysis of the imitation performance on ANYmal both in simulation and on hardware is provided in Figure 3. The robot’s spatial base movement and the height of the feet relative to the ground, closely follow the tracked trajectory. We observe a maximum base position deviation of 0.23m over a trajectory of 30m both in simulation and on hardware. There is an appreciable gap in the foot height compared to the reference, which we attribute to the dynamically-

feasible approximation to a purely kinematic reference trajectory (e.g. slight difference in actual vs. expected contact point). Overall, our results bear similarity to [48] but with the important difference that we do not train separate controllers for each individual MoCap clip but a single controller that can imitate a wide range of different movements<sup>2</sup>.

## Skill Reuse

We verify whether the skill module can produce naturalistic and functional robot behavior in a sufficiently general fashion to solve different locomotion tasks. We consider a controllable walking task for both robots, as well as a ball dribbling task. While the former requires general locomotion behavior including agile and rapid turns, the latter tests specific goal directed leg movements and object interaction.

**Controllable Walking** The controllable walking task requires the robot to turn and walk in different directions following directional velocity commands (forward, lateral and yaw rate) provided in egocentric coordinates. The single-term task reward penalizes the difference between the commanded and achieved velocity but provides no further regularization for the gait or other properties of the movement itself. To provide

<sup>2</sup>In contrast to [48] we also did not find it necessary to perform additional adaptation on ANYmal beyond the implicit adaptation provided by a history-conditioned policy.

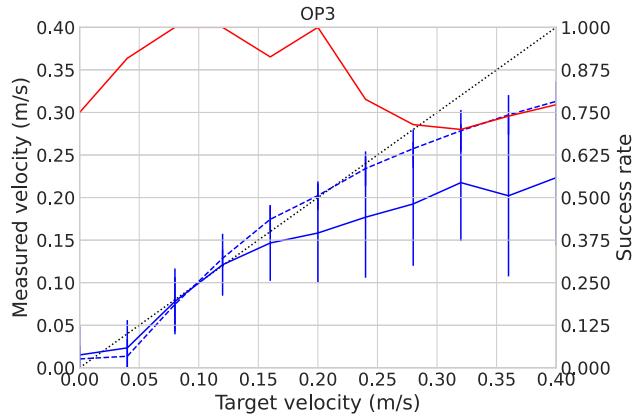
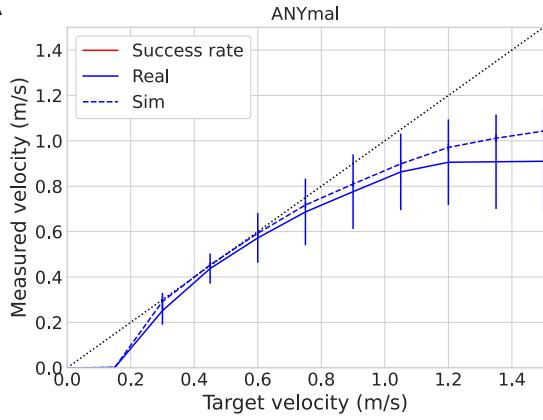
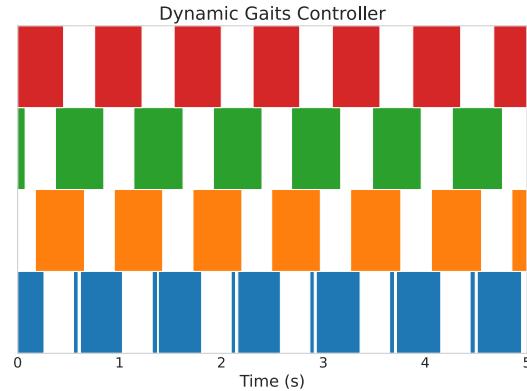
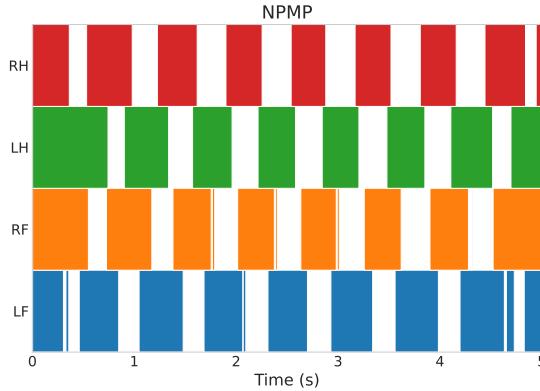
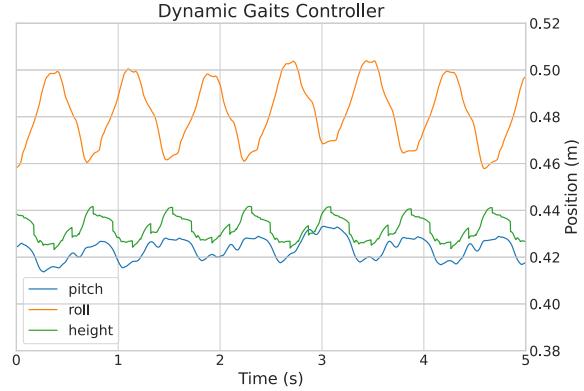
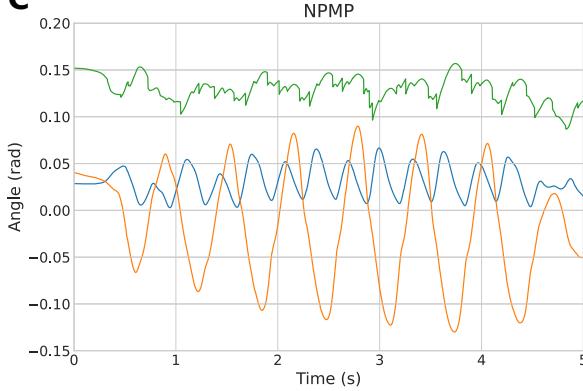
**A****B****C**

Figure 4 | Analysis of the controllable walking results. (A) Accuracy at which the ANYmal (left) and OP3 (right) controllers follow commanded forward velocities in simulation and reality. For OP3, the real-world success rate is also plotted. Commanded velocities are held fixed throughout each trial and values shown are mean and standard error across all trials. (B) Contact patterns of the feet for both the learned and the baseline walking controller on ANYmal as estimated on the real robot during walking at fixed velocity. (C) Corresponding pitch, roll and height measures of the base. See main text for discussion.

some robustness to small obstacles for ANYmal specifically, we use a procedurally generated terrain during training, as seen in Figure 2C.

We evaluate the resulting policies in simulation and on hardware. We test the policies for a range of forward commands that are kept fixed during each trial. In the hardware experiments, a single trial ends when the robot reaches the border of the available space (as detected via a MoCap system) or when it falls over. See Video E for an example on OP3. Figure 4A shows the velocity tracking accuracy for ANYMal and OP3 in response to different commanded forward velocities both in simulation and on hardware. Overall the policies track reasonably accurately over a large range of velocity commands for both OP3 and ANYmal. Deviations at very slow speeds are expected as maintaining balance is more challenging. Deviations at high speeds can partially be explained by limitations of the MoCap datasets, which have comparatively few high-speed running clips. We also observe that the policies are slightly slower on hardware than in simulation due to the simulation-to-reality gap, which is larger for OP3. For the latter we also show the success rate of the trials in real, indicating how many trials the robot finished successfully without prematurely falling over. For OP3 in simulation and ANYmal overall the success rate is 100% over the tested velocities.

We further test the policies’ ability to follow velocity commands from a human user with a joystick. Results can be seen in Videos F and G. Overall ANYmal is very responsive to user commands and can rapidly accelerate, decelerate and turn. Although the movement quality on the real OP3 is generally good, it shows occasional instabilities and is less responsive and more sensitive to rapid changes in the user’s control input, presumably due to unmodelled system properties or other limitations of the underlying control stack. For ANYmal we also implemented a simple feedback controller that provides velocity commands to our trained policy to keep the robot on a designated trajectory around several cones. The results are shown in Figure 2B and Video H. Using our approach, ANYmal can reliably and accurately follow the trajectory and at the requested veloci-

ties guided by the high-level controller, including tight turns and high reference velocities. Finally, we also probe the policies’ robustness by placing an obstacle in the robot’s path. The policy is able to overcome small bumps and perturbations. However, since the simulated training environment only contains smooth terrain variations, the controller does not learn a foot-trapping reflex, and it therefore fails to move when the obstacles cause one of the feet to get stuck, though it does maintain its balance. In Figure 4B we visualize the gait pattern that is produced by our policy deployed on ANYmal and compare it to a baseline classical controller [1] based on [Model Predictive Control \(MPC\)](#) that is configured to produce a similar dynamic walking gait. Figure 4C also visualizes the roll, pitch and height of the base for the respective controllers. While they both realize a walking gait, the base tends to sway significantly more with the learned policy than with the classical controller, which has an explicit objective to keep the base steady. This illustrates that there is no strict need for the base of a quadruped robot to be kept steady to generate stable locomotion, and this also reflects some of the characteristics of quadruped locomotion in animals.

**Ball Dribbling** To assess whether the same skill module that generates good locomotion behavior can also produce precise, goal directed movements with individual limbs and object interaction, we train policies to dribble a ball to a shifting target position. Details of the task are provided in [Materials and Methods](#). In short, the robot perceives both the ball and target positions in egocentric coordinates. During training we randomly change the target position. The policy is trained with a simple reward based on the distance between the ball and the target, without any task-specific regularization.

We evaluate the learned dribble policy for both robots in simulation and on the real ANYmal robot. Results for ANYmal are shown in Figure 5 and in Videos I and J. As Figure 5 demonstrates, the learned policy for ANYmal moves the ball to the shifting target with a high accuracy not just in simulation but also on the real robot. The robot has learned to use its legs to interact with the ball in a goal-directed precise fashion, and this

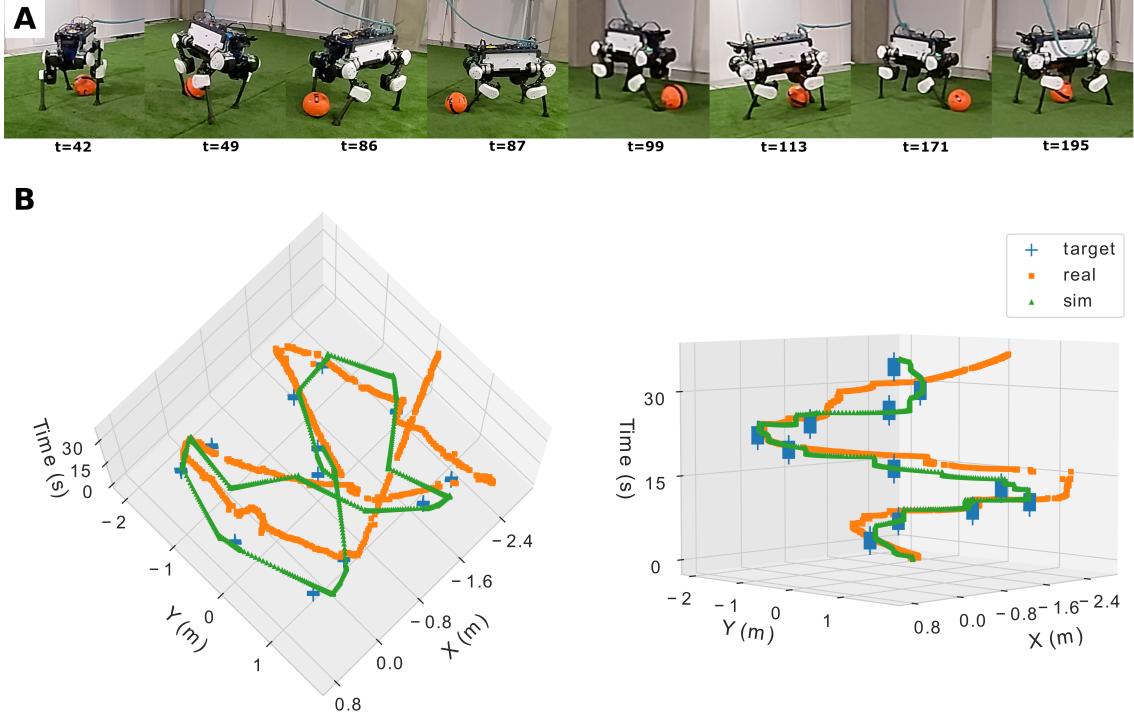


Figure 5 | ANYmal dribbling a ball. (A) A series of keyframes that illustrate the dribbling controller using all of its legs to interact with the ball. (B) Two views of a 3D plot of the top-down trajectory followed by the dribbling controller both in simulation (green) and on the real robot (orange). Blue markers show the position of the target over time. The vertical axis represents time. Both the orange and the green trace closely follow the target positions. Note that the targets are moved after fixed time intervals and the robot sometimes has to wait for the next target to appear.

behavior transfers well to the hardware. Interestingly, the agent has learned to control the ball with both its front and hind legs. This result is especially remarkable since minimal effort was made to identify and randomize ball and ground properties for accurate sim-to-real transfer. We observe a small number of situations where the real robot seemingly loses control of the ball. An analysis of the corresponding data reveals that these are not necessarily failures of the controller but can be explained by a failure of the MoCap system to track the ball, presumably due to occlusion. Video K shows OP3 dribbling in simulation. The policy successfully navigates to the ball and has even learned to carefully positions itself relative to the ball before kicking it in the appropriate direction. A challenge in testing this policy on hardware is the difficulty in tracking an appropriately-sized ball with MoCap, which resulted in very noisy position estimates and erratic behavior from the policy.

Overall, these results demonstrate that the same skill module that produces agile and dynamic locomotion can also be used to produce precise goal directed movements with individual limbs, for tasks that are semantically quite different from the set of behaviors seen in the MoCap dataset which consist mostly of walking and turning.

### Analysis of Skill Space

The nature of the skill space determines the effectiveness of the skill module both in terms of the quality of the movements it produces and the ease with which it can be applied to downstream tasks. As explained in *Materials and Methods* (Equation (5) and (6)) regularization towards an AR(1) prior on the latent actions is applied both during training and reuse, controlled by a coefficient  $\beta$ . This prior encourages the latent commands to change more slowly over time.

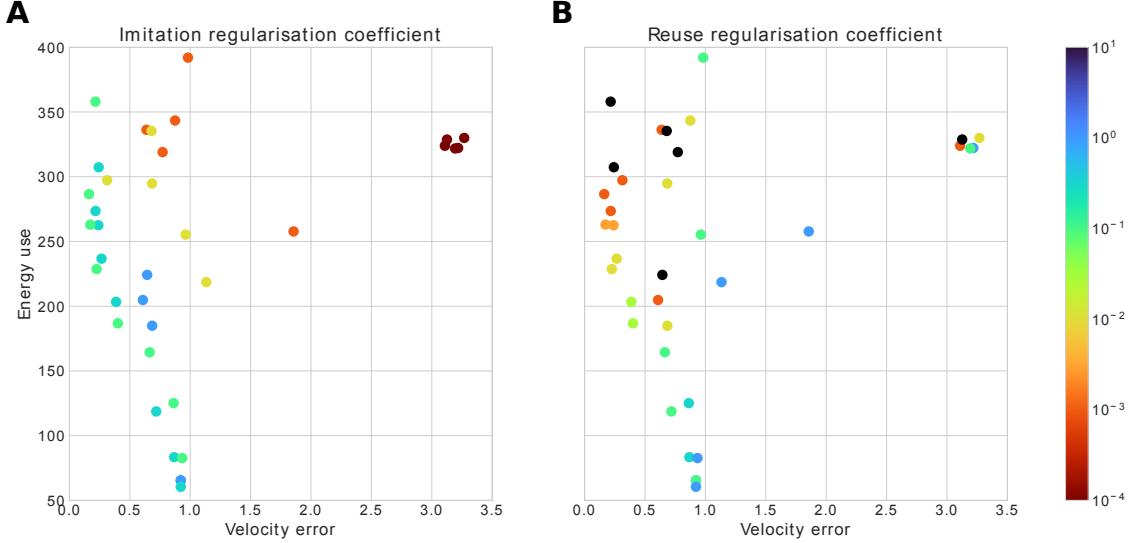


Figure 6 | Multi-objective plots showing the effect of the KL regularization strength on velocity error and energy use (estimated by the sum of squared currents) for the controllable walking task for ANYmal, evaluated on procedural terrain in simulation. The color signifies the strength of the KL regularization, in (A) while learning the skill module via imitation, in (B) while learning the reuse controllable walking itself. Black corresponds to  $\beta = 0$ . Note that both plots show the same data points, but indicate regularization strength in different phases.

We vary  $\beta$  for both phases of training and measure both task performance (velocity error) and efficiency (energy use) during the controllable walking task for ANYmal in simulation. Figure 6 shows the trade-off between these two objectives for different values of  $\beta$  as applied during skill module training (6A), and during reuse itself (6B). In both cases we plot the mean sum of squared currents to approximate energy use, and mean squared velocity error to indicate task performance on the controllable walking task. Note that we use different values for  $\beta$  during the imitation and reuse phases, and that the regularization strength in either has a different effect on energy-error trade-off.

We find that the regularization strength  $\beta$  during the imitation training phase is crucial for preserving the style and smoothness of locomotion in the reuse phase. Little to no regularization generally leads to poor reuse and the higher  $\beta$ , the closer one gets to the Pareto front of solutions. There is a cut-off point after which increasing  $\beta$  will prevent successful imitation and subsequent reuse. In this case information cannot flow from the encoder to the decoder before the

policy has learned to imitate the MoCap trajectories. A schedule on  $\beta$  proves to be very effective in maximizing regularization while retaining successful imitation. Learning then occurs in two stages: first the policy learns to effectively imitate the reference trajectories, and subsequently the regularization encourages the latent commands to follow the prior and change more slowly over time. Figure 7 shows the learning curves for various combinations of domain randomization, regularization strength and schedule. For very high regularization strengths ( $\beta = 0.3$ ) we see a small performance drop during imitation compared to minimal regularization, but still significantly increased performance during reuse, as shown in Figure 6A. Without a schedule, learning with such a high regularization strength fails, even without domain randomization. Finally, we only pay a small penalty for training with domain randomization compared to without.

To retain the stylistic qualities of the MoCap motions, we employ the same Kullback–Leibler (KL) regularization towards the prior during reuse. Figure 6B shows that the regularization strength during reuse now directly controls the

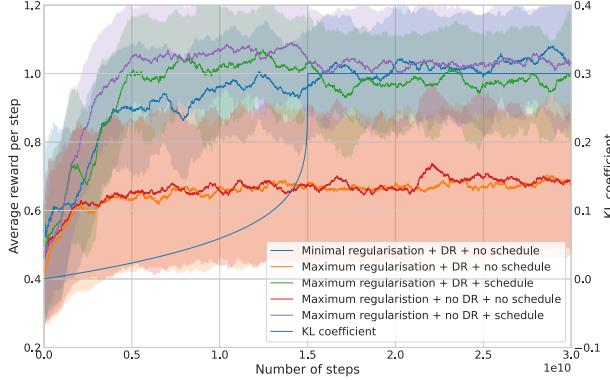


Figure 7 | Training curves for the ANYmal robot on the imitation task showing the effect on convergence of domain randomization and various regularization settings.

trade-off between energy use and velocity error. A stronger regularization forces the high-level policy’s latent actions to change more slowly over time, meaning that it cannot respond as quickly to changes in commanded velocity or match the *instantaneous* target velocity exactly. On the upside, higher regularization results in behaviors that stay closer to the [MoCap](#) motions, are hence smoother and draw less current. If the regularization is too high however, the robot will stop responding to the commands altogether. In the results reported we use  $\beta = 0.3$  during [MoCap](#) imitation and  $\beta = 0.01$  when training downstream tasks.

## Discussion

Our approach for learning and transferring skill modules allowed both a quadruped and a biped to perform various tasks with dynamic, natural looking motion styles derived from [MoCap](#) data. These tasks included controllable locomotion, motion imitation and ball dribbling. The skill modules could be deployed to solve several quite different tasks (including object interaction), while also being able to imitate motion clips accurately. This demonstrates that they capture the data effectively while still enabling generalization towards new tasks. Similar controllers, e.g. for walking, have been learned before [13] but with qualitatively different looking motion styles and a need for more task specific reward tuning.

The same methodology achieved good results for both a biped and a quadruped, suggesting that it should be easily adaptable to other platforms.

The skill module simplifies task design in two ways: firstly it constrains robot’s movements for a given task to the manifold of naturalistic movements for the respective body, thus alleviating the need for the often complex reward shaping strategies required by most prior works. This avoids the potentially complex interplay between behavior shaping reward terms and task reward; we only need to tune the strength of the regularization towards the prior. During imitation, regularization should be maximized without significant impact on the imitation performance, while during reuse the regularization strength controls the trade-off between optimizing the downstream task and preserving naturalness. Empirically we find this easy to tune and we used the same coefficients across robot platforms and downstream tasks. Secondly, the skill module simplifies the design of high-level tasks. The default behavior of the low-level controller induces more natural and effective exploration (Video L). This reduces the need for task-specific shaping, for instance in the dribbling task.

All in all, the skill module demonstrates how learned low-level controllers provide interfaces that can be more flexible than, for instance, more conventional [MPC](#) controllers which have been used as the starting point in successful simulation-to-real locomotion approaches [49, 50, 51]. Our skill module provides access both to regular gaits, but also to more non-standard, asymmetric, and goal directed whole-body movements.

## Limitations

While we believe that our controllers generate more natural looking behavior than existing ones, the concept of “natural looking behavior” is, of course, subjective; we encourage readers to judge for themselves and watch the accompanying videos. Furthermore, there is also no guarantee that movements derived from a biological body would be optimal for a particular robot platform according to criteria such as energy efficiency or minimization of wear and tear. Nevertheless, ex-

perimentally we found that our approach led to smooth movements, that qualitatively were well-regularized and functional, and performed well on the hardware.

**MoCap** recordings can of course only aid controller learning for platforms for which suitable data is available, taking both morphology and dynamic feasibility into account. Still, many platforms *are* modeled after animals and can thus benefit from our approach. Furthermore, our results on two different robots demonstrate that even a loose similarity allows for flexible data usage without excessive effort during retargeting. There are limits to how much the desired downstream behaviors can differ from the data and collecting additional **MoCap** data to fill in the gaps can be time consuming and costly. However, our skill modules were constructed using datasets that were relatively small in size and diversity, but nevertheless enabled a diverse set of movements and reuse scenarios.

We argue that our method does not require extensive effort in designing reward and regularization strategies in the reuse phase, as we can use simple rewards and only need to tune the strength of the regularization towards the prior. During the imitation phase we do have to take care when tuning the reward (involving several terms and coefficients, see *Supplementary Materials*) and regularization (**KL** schedule) strategies. The effort of this additional phase is, however, easily amortized when targeting multiple downstream tasks, which would otherwise require dedicated tuning.

Our results on hardware differ in quality between ANYmal and OP3. This is primarily due to a difference in the quality of simulation-to-real transfer which is significantly easier for the well-engineered and more faithfully-simulated ANYmal robot. Several approaches, e.g. for fine tuning or adaptation during deployment [27, 26], could naturally be combined with the proposed skill module to achieve even better results for more challenging platforms like OP3. The question of improving simulation-to-real transfer is, however, outside the scope of this work.

## Future Work

In future work, we want to extend our datasets with a larger variety of behaviors and further explore the range of downstream tasks that the skill module enables. It should be possible to merge and grow datasets over time with recordings that target behaviors that the skill modules struggle with. Datasets could also be enriched with other sources than **MoCap** such as trajectory optimization over procedural terrains [52, 53]. In particular it will be interesting to explore tasks that include more dynamic movements and obstacle traversal, or advanced object interaction [8, 9]. Hierarchical controllers like in this work could also be particularly suitable for enabling more efficient learning of perceptive policies as has already been shown in simulation [8].

More broadly we expect that easy-to-use but flexible representations of movement related prior knowledge (including from demonstrations) will play an important role in future training pipelines for legged robots, especially as the field moves its focus from core movement skills towards more complex, higher-level and goal-oriented whole-body control.

## Materials and Methods

### Preliminaries

We use **RL** to train our controllers. Each task is defined as a discrete time **Partially Observed MDP (POMDP)** represented by the tuple  $(S, O, A, P, P_0, r, p_0, \gamma)$ . Here  $S$ ,  $O$ , and  $A$  denote the state, action and observation spaces;  $p_0$ ,  $P$  and  $P_0$  denote the initial state, state transition and observation probability distributions;  $r$  is the reward function; and  $\gamma$  a discount parameter. Our goal is to learn a policy conditioned on a history of observations  $\pi(a_t | o_{\leq t})$  that maximizes the expected discounted reward

$$\pi^* = \arg \max_{\pi} \mathbb{E}_{\tau(\pi)} \left( \sum_{t=0}^{\infty} \gamma^t r_t \right), \quad (1)$$

where  $\tau(\pi)$  denotes the distribution over trajectories induced by policy  $\pi$ . The imitation stage constitutes a multi-objective problem and we use

the approach of [29] to optimize a finite set of  $K$  reward functions  $r_k$  simultaneously.

## Simulation

We use MuJoCo [54] for rigid body simulation of both robot platforms and dm\_control [55] to provide the RL environment logic. We create accurate MuJoCo models of the robots building on existing URDF models of ANYmal and OP3.

A key element of successful simulation-to-reality transfer for locomotion is accurate modeling of the actuator dynamics [14, 13]. Similar to [13, 56], we train an actuator network that models the complexities of the *Series Elastic Actuators (SEAs)* used in ANYmal, but we split the model into an analytical part that models the top-level PID control and a learned low-level part that models the torque transfer function. The former allows us to change PID gains or even control mode without retraining the actuator model. Besides joint torque, the model also predicts current draw, which allows us to estimate and optimize the efficiency of our controllers in simulation. See *Supplementary Materials* for more details. OP3 has arguably simpler, high-gear servo actuators which can be modeled sufficiently accurately by MuJoCo’s built-in actuators after system identification.

We control the ANYmal and OP3 at 50 and 33Hz, respectively. For both platforms the control inputs are joint target positions that are tracked by a PD controller. We center the actions around the stable standing pose but use no additional action filtering, clipping or scaling. For ANYmal specifically, we use delayed first-order hold interpolation at 400Hz (in accordance with the main control stack) for the setpoint [57]. While this introduces an additional delay between sensing and acting, it avoids exciting internal drive dynamics and increases the efficiency of the controller.

To further bridge the simulation-to-reality gap, we employed dynamics/domain randomization techniques [22, 23, 24, 25]. We varied both kinematic and dynamic properties of the model: the friction at the foot contacts; body masses and center-of-mass locations; and the joint positions, offsets, damping and friction losses. We also

added perturbations to the robots using random forces applied to the torso, and we added noise and delays to the simulated sensor readings. A full list of randomization and noise settings for both platforms is provided in the *Supplementary Materials*.

## Motion Capture Imitation

**Motion Capture Data** For ANYmal, we use the dog MoCap dataset from [33] and a similar point-cloud retargeting procedure as [48, 58]. The data consists of unstructured behavior and contains walking, running, turning and jumping among others. Due to the symmetries of the robot, we can augment the data by mirroring the reference motions left-to-right and / or front-to-back, which allows e.g. ANYmal to walk backwards as no backwards walking is otherwise present in the dataset. In total we train on roughly 2.5 hours of reference motion. For OP3, we start from roughly 1.5 hours of walking and running trajectories from [37] that have already been retargeted to the CMUHumanoid model from the dm\_control [55] environment and adapt it to the model of the robot. This involved accounting for the lack of degrees of freedom in OP3’s torso.

See *Supplementary Materials* for more details.

**Task Definition** In the imitation task, the goal is to train a universal goal-conditioned policy that can imitate the complete dataset of reference motions. We largely follow the procedure in [39]: at the start of each episode, a random reference MoCap clip is sampled such that there is an approximately uniform distribution over velocities, to prevent overfitting to either extreme of the velocity range. Subsequently, the initial state is sampled uniformly from the clip frames excluding the last 15. At every time step, reward terms are computed that compare the current state  $s_t$  with the corresponding state in the reference trajectory  $s_t^{\text{ref}}$ . For ANYmal we add an additional reward term that minimizes the actuators’ current draw, which we found helpful to further suppress higher-frequency excitations. See *Supplementary Materials* for more details.

The episode terminates either when all reference frames have been exhausted or when the robot’s pose deviates too much from the reference frame according to the following metric:

$$\delta = \frac{1}{3|\mathcal{B}|} \sum_{i \in \mathcal{B}} \left\| \mathbf{p}_i - \mathbf{p}_i^{\text{ref}} \right\|_1 + \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} \left\| q_j - q_j^{\text{ref}} \right\|_1, \quad (2)$$

where  $\mathbf{p}_i$  are the position vectors of all the bodies with indices  $\mathcal{B} \subset \mathbb{N}$  and  $q_j$  are the joint positions with indices  $\mathcal{J} \subset \mathbb{N}$ . Episodes terminate when  $\delta > \eta$ , with  $\eta = 0.3$ .

**Architecture** Our skill modules are based on the NPMP framework [37, 39] and transfer skills by learning a two-level *encoder-decoder* architecture of which only the encoder is replaced when learning new tasks. Figure 8 shows an overview. The *high-level* encoder policy  $\pi_{\text{HL}}(z_t | z_{t-1}, x_t)$ , is conditioned on context information  $x_t$  and produces a stochastic latent command  $z_t$ . The *low-level* latent command conditioned decoder policy  $\pi_{\text{LL}}(a_t | o_{\leq t}, z_t)$ , subsequently produces the action  $a_t$  to be executed by the robot.

The complete learning architecture consists therefore of multiple networks: the encoder, the decoder, and a value function network. These networks have access to different, potentially overlapping, sources of information, some of which are privileged and/or specific to the MoCap imitation task.

The context information  $x_t$  provided to the high-level encoder  $\pi_{\text{HL}}(z_t | z_{t-1}, x_t)$  describes the reference motion clip to imitate, specifically the body positions and orientations at the subsequent 5 *future* time steps, encoded relative to the current pose of the robot<sup>3</sup>.

The output of the encoder is parameterized as

$$\begin{aligned} \pi_{\text{HL}}(z_t | z_{t-1}, x_t) = \\ \mathcal{N}(\mu_{\text{HL}}(z_{t-1}, x_t) + \alpha \cdot z_{t-1}, \Sigma_{\text{HL}}(z_{t-1}, x_t)), \end{aligned} \quad (3)$$

with  $\mu_{\text{HL}}$  and  $\Sigma_{\text{HL}}$  the output of a two-layer Multi-layer Perceptron (MLP) and  $\alpha$  the time constant of the AR(1) prior. In this work we choose  $\alpha = 0.95$ .

<sup>3</sup>For experiments with the real robots we use a MoCap system to obtain the input to the reference encoder by tracking the robot’s global pose and computing its difference to the reference trajectory.

As the decoder will be reused and deployed on hardware, it only observes noisy, raw sensor readings: the joint positions and position setpoints, angular velocity, linear acceleration and roll & pitch estimates from the IMU. It also receives a latent command  $z_t$  from the encoder which instructs the movement to execute. Given just instantaneous observations, however, the environment state would be partially observed. We enable the low-level decoder to infer more of the state by adding memory in the form of an Long Short-Term Memory (LSTM) [59] layer. This has the additional benefit that the decoder can learn to identify and implicitly adapt to the different dynamics variations mentioned in *Simulation*. Providing memory to the low-level could cause it to overfit to the latent command sequences seen during training, thus becoming overly or insufficiently sensitive to the latent commands during reuse. To mitigate this, the decoder is split in two branches after an initial input normalization layer that only receives the proprioception, as seen in Figure 8. The first branch contains two fully connected and one LSTM layer. The second branch takes the output of the first, concatenates it with the normalized proprioception and the latent command and feeds it through two more fully-connected layers. Finally a linear combination of the outputs of both branches is used to produce the primitive action distribution as

$$\begin{aligned} \pi_{\text{LL}}(a_t | o_t, z_t, h_{t-1}) = \\ \mathcal{N}(\mu_{\text{LL}}(o_t, z_t, h_{t-1}), \Sigma_{\text{LL}}(o_t, z_t, h_{t-1})). \end{aligned} \quad (4)$$

Since the value function network is only used for training in simulation, it can observe the full simulation state, including true joint positions and velocities, base orientation and twist, and egocentric feet positions. It also receives the future reference frames, clip identity, and values of randomized simulation parameters. These features are concatenated and passed through a three-layer MLP to produce value estimates.

**Training** To train the MoCap imitation policy, we used MO-VMPO [29], a multi-objective variant of the VMPO algorithm [60]. MO-VMPO has been shown to be effective at learning MoCap imitation for simulated humanoids [29]. The algorithm

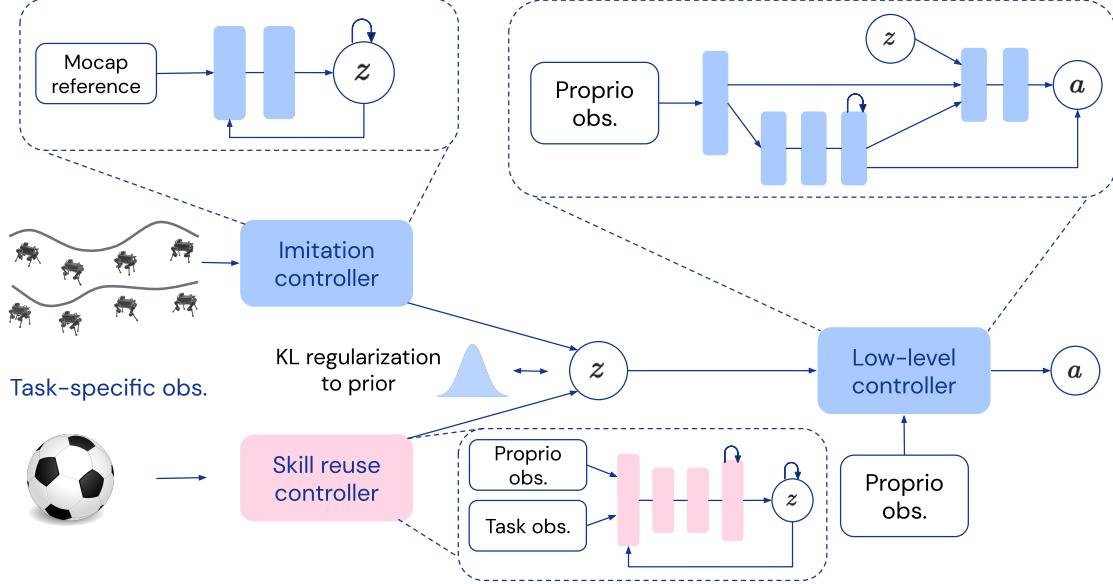


Figure 8 | Architecture of the different controller components. During imitation, we train the imitation and low-level controller end-to-end. The imitation controller is a feed-forward network, while the low-level controller consist of two branches: one with memory and one conditioned on the latent command. During reuse we only train the reuse controller while keeping the low-level fixed. The reuse controller takes in proprioceptive and task-specific observations and also contains memory. The latent commands  $z$  are regularized towards the AR(1) prior in both phases.

optimizes for  $K$  rewards  $r_k$  and balances their relative importance via corresponding constraint parameters  $c_k > 0$ . The value function network outputs separate value predictions  $V(s_t)_k$  for each reward and is trained with the conventional  $n$ -step temporal difference loss. Since the full MO-VMPO objective is implicit, we report the sum of the reward functions  $\sum_{k=1}^K r_k$  in our results.

We train the policy end-to-end, using the re-parameterization trick to train the latent space and encoder, similar to a **Variational Auto-Encoder (VAE)** [61]. The latent space  $\mathcal{Z} \subset \mathbb{R}^d$  should not only represent skills effectively, but also have a structure that allows a new high-level policy to efficiently explore this space and learn to transition between skills smoothly. Therefore, we minimize the **KL** divergence between the encoder and some prior  $p(z)$  by adding it to the losses of the policy optimization problem. To promote temporally *consistent* behaviors, we use an **AR(1)** prior

$$p(z_t | z_{t-1}) = \mathcal{N}\left(\alpha \cdot z_{t-1}, (1 - \alpha^2) \cdot \mathbf{I}\right), \quad (5)$$

in which  $\alpha \in [0, 1]$  (which we set to 0.95) scales

the correlation with the previous time step. The **KL** divergence takes the following form:

$$\beta \mathbb{E}_{z^* \sim \pi_{\text{HL}}} \sum_t \text{KL} [\pi_{\text{HL}}(z_t | z_{t-1}^*, x_t) \| p(z_t | z_{t-1}^*)], \quad (6)$$

where  $\beta > 0$  is a scaling parameter. This regularization can also be interpreted as an information bottleneck [62, 63] in which the prior  $p(z)$  functions as the (unoptimized) variational distribution [64, 65, 45]. Recall that this regularization to the prior happens both during skill module learning and during reuse.

## Downstream Tasks

**General Setup** To reuse the low-level controller for downstream tasks, we replace the encoder trained during imitation with a new task-specific high-level controller  $\hat{\pi}_{\text{HL}}(z_t | z_{t-1}, o_{\leq t}, y_t)$  that acts and is trained directly in the latent command space. The low-level policy's parameters  $\pi_{\text{LL}}$  are kept fixed, effectively making it part of the high-level policy's environment. As we will evaluate these downstream tasks on hardware, the new high-level controllers can, like the low-level, only

observe raw, noisy sensor data  $o_{\leq t}$ , besides task-specific input  $y_t$ . As shown in Figure 8, we use an input normalization layer followed by two fully connected and an LSTM layer. The memory in the high-level allows it to perform state estimation and system identification, similar to the low-level. In each reuse task, we only have a single task-specific reward term and no longer any shaping terms that influence the locomotion style. We use (single-objective) VMPO [60] to train the new high-level controllers.

During reuse the latent AR(1) prior is used in two ways: initialization and regularization. The regularization is the same as applied to the encoder during skill module training (cf. Equation (6)) above). In addition, we found it beneficial for early exploration to match the initial temporal statistics of the latent actions to the AR(1) prior. When using the skill module generatively by sampling latent actions from the prior, we observe more “structured” behavior, where the robots tend to maintain balance and walk around randomly, in stark contrast to a Gaussian prior directly on the primitive actions. Video L shows an example on the ANYmal robot in simulation<sup>4</sup>. We anticipate that this kind of behavior provides better exploration. Therefore, we parameterize the latent action policy as

$$\begin{aligned} \pi_{\text{HL}}(z_t | z_{t-1}, o_{\leq t}, y_t) = \\ \mathcal{N}(\theta(z_{t-1}, o_{\leq t}, y_t) \cdot \mu_{\text{HL}}(z_{t-1}, o_{\leq t}, y_t) \\ + (1 - \theta(z_{t-1}, o_{\leq t}, y_t)) \cdot z_{t-1}, \\ \Sigma_{\text{HL}}(z_{t-1}, o_{\leq t}, y_t)), \end{aligned} \quad (7)$$

where  $\mu_{\text{HL}}$  and  $\Sigma_{\text{HL}}$  are functions of the previous latent command  $z_{t-1}$  (we stop gradients flowing through  $z_{t-1}$  during optimization), observation history  $o_{\leq t}$  and task-specific observations  $y_t$  and are the standard Gaussian mean and variance policy output, with  $\mu_{\text{HL}}$  clipped to the range of latents during MoCap imitation using the tanh function. The value  $\theta \in [0, 1]$  is another (state-conditional) high-level policy output that represents a filtering constant and is initialized to the

<sup>4</sup>Note that ANYmal will eventually encounter a “sitting” pose from which it cannot recover. We observe that during imitation ANYmal is unable to get back up using the current MoCap retargeting procedure and that the skill module has converged to using this sitting pose as a graceful failure mode.

value of the prior parameter  $\alpha = 0.95$ . This formulation initializes  $\pi_{\text{HL}}$  to match the prior but still allows it to invert the filtering operation.

**Controllable Walking** The aim of this RL task is to produce controllers that allow the robot to be steered with a joystick. Like previous work [13], we achieve this by conditioning the controller on commands that specify the desired yaw rate, forward and lateral velocities of the base of the robot. We sample target velocities in a range that roughly covers 99% of the velocities seen during imitation and do not expect the low-level policy to generalize beyond that. Samples follow a memoryless random process such that the high-level does not learn to anticipate command changes. This makes it feasible to use joystick commands during deployment. See *Supplementary Materials* for details.

The task-specific observations  $y_t$  consist of the three target velocities. Otherwise the observations and general simulation setup are the same as during MoCap imitation, including the dynamics variations (except we apply larger perturbations during reuse). We terminate episodes when there are self-collisions, any part of the robot touches the floor besides the feet, or the tilt of the robot base exceeds 45 degrees. Finally, for ANYmal specifically, we change the ground surface from a plane to a procedurally-generated hilly terrain based on Perlin noise [66] as shown in Figure 2C. This encourages robustness to changes in inclination and small obstacles, as shown in Video M. We find that the low-level controller reliably generalizes to moderate terrain variations despite no exposure to them during training.

The single-term task reward is defined as  $r_t = \exp(-\|\mathbf{v}_t - \hat{\mathbf{v}}_t\|_2^2 / \phi)$ , where  $\hat{\mathbf{v}}_t$  is the target velocity at step  $t$  and  $\phi$  scales the resolution. There are no shaping rewards. We find that training without reusing the low-level controller or other forms of regularization results in very erratic, unsafe behavior (see for example Video N) Moreover, we find for ANYmal specifically that we can no longer solve the task on the procedurally-generated terrain with the same hyperparameters and need to resort to a flat surface.

We train policies for ANYmal and OP3 exclusively in simulation. These policies can then be deployed on the robots zero-shot, without further adaptation. They can be controlled by a human operator via joystick relying only on on-board sensors, see Videos F and G<sup>5</sup>. For a more hands-off deployment, we implement a trajectory tracking controller on top which generates velocity commands according to  $\hat{\mathbf{v}}_t = \bar{\mathbf{v}}_t + P \cdot (\bar{\mathbf{p}}_t - \mathbf{p}_t)$ , where  $\bar{\mathbf{v}}_t$  and  $\bar{\mathbf{x}}_t$  are the velocity and position (including heading) of the trajectory to follow. The current position and orientation of the robot  $\mathbf{x}_t$  are measured via a MoCap system in our experiments, but could also come from a state estimator. The proportional feedback term is useful for compensating for drift due to accumulating errors in velocity tracking accuracy.

**Ball Dribbling** In this task, the robot needs to dribble a ball to a target position. The setup is very similar to the one for controllable walking. Besides the robot, we now also simulate a ball, which is always initialized in front of the robot. Its size and weight depend on the robot and are also randomized, though no significant effort was put into system identification otherwise. The task-specific observations  $y_t$  consist of the current 3D position of the ball and target, both in egocentric coordinates of the robot<sup>6</sup>. The reward is defined similar to the one used for velocity tracking, with  $r_t = \exp(-\|\mathbf{p}_t^{\text{ball}} - \hat{\mathbf{p}}_t\|_2^2 / \phi)$ , where  $\mathbf{p}_t^{\text{ball}}$  is the current position of the ball and  $\hat{\mathbf{p}}_t$  the target position. This reward is independent of the state of the robot, making it fairly sparse and posing an overall harder exploration problem. Similar to the velocity tracking task, the target position follows a random process. Episodes now also terminate when the distance between the robot and ball or ball and target becomes too large, to prevent unnecessarily long episodes without reward.

<sup>5</sup>Note that for ANYmal we still use a network tether but only for remote monitoring.

<sup>6</sup>We track the real ball with the MoCap system alongside the robot to provide the required egocentric observations during transfer.

## References

- [1] C. D. Bellicoso, F. Jenelten, C. Gehring, and M. Hutter, “Dynamic locomotion through online nonlinear motion optimization for quadrupedal robots,” *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2261–2268, 2018.
- [2] M. Kalakrishnan, J. Buchli, P. Pastor, M. Mistry, and S. Schaal, “Fast, robust quadruped locomotion over challenging terrain,” in *2010 IEEE International Conference on Robotics and Automation*, pp. 2665–2670, IEEE, 2010.
- [3] M. Neunert, F. Farshidian, A. W. Winckler, and J. Buchli, “Trajectory optimization through contacts and automatic gait discovery for quadrupeds,” *IEEE Robotics and Automation Letters*, vol. 2, no. 3, pp. 1502–1509, 2017.
- [4] J. Carius, R. Ranftl, V. Koltun, and M. Hutter, “Trajectory optimization with implicit hard contacts,” *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3316–3323, 2018.
- [5] T. Apgar, P. Clary, K. Green, A. Fern, and J. W. Hurst, “Fast online trajectory optimization for the bipedal robot cassie.,” in *Robotics: Science and Systems*, vol. 101, p. 14, 2018.
- [6] N. Heess, D. TB, S. Sriram, J. Lemmon, J. Merel, G. Wayne, Y. Tassa, T. Erez, Z. Wang, S. M. A. Eslami, M. Riedmiller, and D. Silver, “Emergence of locomotion behaviours in rich environments,” *arXiv preprint arXiv:1707.02286*, 2017.
- [7] X. B. Peng, P. Abbeel, S. Levine, and M. van de Panne, “DeepMimic: Example-guided deep reinforcement learning of physics-based character skills,” *ACM Trans. Graph.*, vol. 37, July 2018.
- [8] J. Merel, S. Tunyasuvunakool, A. Ahuja, Y. Tassa, L. Hasenclever, V. Pham, T. Erez, G. Wayne, and N. Heess, “Catch & carry:

- reusable neural controllers for vision-guided whole-body tasks,” *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, pp. 39–1, 2020.
- [9] S. Liu, G. Lever, Z. Wang, J. Merel, S. Es-lami, D. Hennes, W. M. Czarnecki, Y. Tassa, S. Omidshafiei, A. Abdolmaleki, *et al.*, “From motor control to team play in simulated humanoid football,” *arXiv preprint arXiv:2105.12196*, 2021.
- [10] T. Haarnoja, S. Ha, A. Zhou, J. Tan, G. Tucker, and S. Levine, “Learning to walk via deep reinforcement learning,” *arXiv preprint arXiv:1812.11103*, 2018.
- [11] R. Hafner, T. Hertweck, P. Klöppner, M. Bloesch, M. Neunert, M. Wulfmeier, S. Tunyasuvunakool, N. Heess, and M. Riedmiller, “Towards general and autonomous learning of core skills: A case study in locomotion,” *arXiv preprint arXiv:2008.12228*, 2020.
- [12] M. Bloesch, J. Humplik, V. Patraucean, R. Hafner, T. Haarnoja, A. Byravan, N. Y. Siegel, S. Tunyasuvunakool, F. Casarini, N. Batchelor, *et al.*, “Towards real robot learning in the wild: A case study in bipedal locomotion,” in *5th Annual Conference on Robot Learning*, 2021.
- [13] J. Hwangbo, J. Lee, A. Dosovitskiy, D. Belli-coso, V. Tsounis, V. Koltun, and M. Hutter, “Learning agile and dynamic motor skills for legged robots,” *Science Robotics*, vol. 4, no. 26, 2019.
- [14] J. Tan, T. Zhang, E. Coumans, A. Iscen, Y. Bai, D. Hafner, S. Bohez, and V. Vanhoucke, “Sim-to-real: Learning agile locomotion for quadruped robots,” in *Proceedings of Robotics: Science and Systems*, 2018.
- [15] C. Yang, K. Yuan, Q. Zhu, W. Yu, and Z. Li, “Multi-expert learning of adaptive legged locomotion,” *Science Robotics*, vol. 5, no. 49, 2020.
- [16] Z. Xie, X. Da, M. van de Panne, B. Babich, and A. Garg, “Dynamics randomization re-visited: A case study for quadrupedal locomotion,” in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2021.
- [17] J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, “Learning quadrupedal locomotion over challenging terrain,” *Science Robotics*, vol. 5, no. 47, 2020.
- [18] T. Miki, J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, “Learning robust perceptive locomotion for quadrupedal robots in the wild,” *Science Robotics*, vol. 7, no. 62, p. eabk2822, 2022.
- [19] J. Siekmann, K. Green, J. Warila, A. Fern, and J. Hurst, “Blind bipedal stair traversal via sim-to-real reinforcement learning,” *arXiv preprint arXiv:2105.08328*, 2021.
- [20] Z. Xie, P. Clary, J. Dao, P. Morais, J. Hurst, and M. Panne, “Learning locomotion skills for cassie: Iterative design and sim-to-real,” in *Conference on Robot Learning*, PMLR, 2020.
- [21] Z. Li, X. Cheng, X. B. Peng, P. Abbeel, S. Levine, G. Berseth, and K. Sreenath, “Reinforcement learning for robust parameterized locomotion control of bipedal robots,” *arXiv preprint arXiv:2103.14295*, 2021.
- [22] F. Sadeghi and S. Levine, “CAD2RL: Real single-image flight without a single real image,” in *Robotics: Science and Systems*, 2017.
- [23] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. 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)*, pp. 23–30, IEEE, 2017.
- [24] X. B. Peng, M. Andrychowicz, W. Zaremba, and P. Abbeel, “Sim-to-real transfer of robotic control with dynamics randomization,” in *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 3803–3810, IEEE, 2018.

- [25] O. M. Andrychowicz, B. Baker, M. Chociej, R. Jozefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray, et al., “Learning dexterous in-hand manipulation,” *The International Journal of Robotics Research*, vol. 39, no. 1, pp. 3–20, 2020.
- [26] X. B. Peng, E. Coumans, T. Zhang, T.-W. Lee, J. Tan, and S. Levine, “Learning agile robotic locomotion skills by imitating animals,” *arXiv preprint arXiv:2004.00784*, 2020.
- [27] W. Yu, V. C. Kumar, G. Turk, and C. K. Liu, “Sim-to-real transfer for biped locomotion,” in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3503–3510, IEEE, 2019.
- [28] S. Bohez, A. Abdolmaleki, M. Neunert, J. Buchli, N. Heess, and R. Hadsell, “Value constrained model-free continuous control,” *arXiv preprint arXiv:1902.04623*, 2019.
- [29] A. Abdolmaleki, S. H. Huang, L. Hasenclever, M. Neunert, F. Song, M. Zambelli, M. F. Martins, N. Heess, R. Hadsell, and M. Reidmiller, “A distributional view on multi-objective policy optimization,” in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.
- [30] A. Iscen, K. Caluwaerts, J. Tan, T. Zhang, E. Coumans, V. Sindhwani, and V. Vanhoucke, “Policies modulating trajectory generators,” in *Conference on Robot Learning*, pp. 916–926, PMLR, 2018.
- [31] A. Safanova and J. K. Hodgins, “Construction and optimal search of interpolated motion graphs,” in *ACM SIGGRAPH 2007 papers*, pp. 106–es, 2007.
- [32] D. Holden, T. Komura, and J. Saito, “Phase-functioned neural networks for character control,” *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–13, 2017.
- [33] H. Zhang, S. Starke, T. Komura, and J. Saito, “Mode-adaptive neural networks for quadruped motion control,” *ACM Trans. Graph.*, vol. 37, July 2018.
- [34] N. Chentanez, M. Müller, M. Macklin, V. Makoviychuk, and S. Jeschke, “Physics-based motion capture imitation with deep reinforcement learning,” in *Proceedings of the 11th annual international conference on motion, interaction, and games*, pp. 1–10, 2018.
- [35] J. Merel, Y. Tassa, D. TB, S. Srinivasan, J. Lemmon, Z. Wang, G. Wayne, and N. Heess, “Learning human behaviors from motion capture by adversarial imitation,” *arXiv preprint arXiv:1707.02201*, 2017.
- [36] X. B. Peng, Z. Ma, P. Abbeel, S. Levine, and A. Kanazawa, “AMP: Adversarial motion priors for stylized physics-based character control,” *arXiv preprint arXiv:2104.02180*, 2021.
- [37] J. Merel, L. Hasenclever, A. Galashov, A. Ahuja, V. Pham, G. Wayne, Y. W. Teh, and N. Heess, “Neural probabilistic motor primitives for humanoid control,” in *International Conference on Learning Representations*, 2019.
- [38] X. B. Peng, M. Chang, G. Zhang, P. Abbeel, and S. Levine, “MCP: Learning composable hierarchical control with multiplicative compositional policies,” *arXiv preprint arXiv:1905.09808*, 2019.
- [39] L. Hasenclever, F. Pardo, R. Hadsell, N. Heess, and J. Merel, “CoMic: Complementary task learning & mimicry for reusable skills,” in *International Conference on Machine Learning*, PMLR, 2020.
- [40] N. S. Pollard, J. K. Hodgins, M. J. Riley, and C. G. Atkeson, “Adapting human motion for the control of a humanoid robot,” in *Proceedings 2002 IEEE international conference on robotics and automation (Cat. No. 02CH37292)*, vol. 2, pp. 1390–1397, IEEE, 2002.
- [41] A. Escontrela, X. B. Peng, W. Yu, T. Zhang, A. Iscen, K. Goldberg, and P. Abbeel, “Adversarial motion priors make good substitutes for complex reward functions,” *arXiv preprint arXiv:2203.15103*, 2022.

- [42] M. Hutter, C. Gehring, D. Jud, A. Lauber, C. D. Bellicoso, V. Tsounis, J. Hwangbo, K. Bodie, P. Fankhauser, M. Bloesch, *et al.*, “ANYmal - a highly mobile and dynamic quadrupedal robot,” in *2016 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 38–44, IEEE, 2016.
- [43] Robotis, “OP3.” <https://emanual.robotis.com/docs/en/platform/op3/introduction/>, 2017.
- [44] CMU Graphics Lab, “Carnegie Mellon University Graphics Lab Motion Capture Database.” <http://mocap.cs.cmu.edu/>, 2003.
- [45] D. Tirumala, A. Galashov, H. Noh, L. Hasenclever, R. Pascanu, J. Schwarz, G. Desjardins, W. M. Czarnecki, A. Ahuja, Y. W. Teh, and N. Heess, “Behavior priors for efficient reinforcement learning,” *arXiv preprint arXiv:2010.14274*, 2020.
- [46] Z. Wang, J. Merel, S. Reed, G. Wayne, N. de Freitas, and N. Heess, “Robust imitation of diverse behaviors,” *arXiv preprint arXiv:1707.02747*, 2017.
- [47] OpenAI, M. Andrychowicz, B. Baker, M. Chociej, R. Jozefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray, J. Schneider, S. Sidor, J. Tobin, P. Welinder, L. Weng, and W. Zaremba, “Learning dexterous in-hand manipulation,” *arXiv preprint arXiv:1808.00177*, 2018.
- [48] X. B. Peng, E. Coumans, T. Zhang, T.-W. E. Lee, J. Tan, and S. Levine, “Learning agile robotic locomotion skills by imitating animals,” in *Robotics: Science and Systems*, 07 2020.
- [49] J. Carius, F. Farshidian, and M. Hutter, “MPC-net: A first principles guided policy search,” *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2897–2904, 2020.
- [50] X. Da, Z. Xie, D. Hoeller, B. Boots, A. Anandkumar, Y. Zhu, B. Babich, and A. Garg, “Learning a contact-adaptive controller for robust, efficient legged locomotion,” *arXiv preprint arXiv:2009.10019*, 2020.
- [51] S. Gangapurwala, A. Mitchell, and I. Havoutis, “Guided constrained policy optimization for dynamic quadrupedal robot locomotion,” *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3642–3649, 2020.
- [52] S. Gangapurwala, M. Geisert, R. Orsolino, M. Fallon, and I. Havoutis, “Real-time trajectory adaptation for quadrupedal locomotion using deep reinforcement learning,” in *Proceedings of the International Conference on Robotics and Automation (ICRA)*, Institute of Electrical and Electronics Engineers, 2021.
- [53] P. Brakel, S. Bohez, L. Hasenclever, N. Heess, and K. Bousmalis, “Learning coordinated terrain-adaptive locomotion by imitating a centroidal dynamics planner,” *arXiv preprint arXiv:2111.00262*, 2021.
- [54] E. Todorov, T. Erez, and Y. Tassa, “MuJoCo: A physics engine for model-based control,” in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033, IEEE, 2012.
- [55] Y. Tassa, S. Tunyasuvunakool, A. Muldal, Y. Doron, S. Liu, S. Bohez, J. Merel, T. Erez, T. Lillicrap, and N. Heess, “dm\_control: Software and tasks for continuous control,” *arXiv preprint arXiv:2006.12983*, 2020.
- [56] S. Gangapurwala, M. Geisert, R. Orsolino, M. Fallon, and I. Havoutis, “RLOC: Terrain-aware legged locomotion using reinforcement learning and optimal control,” *arXiv preprint arXiv:2012.03094*, 2020.
- [57] J. Proakis, *Digital Signal Processing: Principles, Algorithms, And Applications*, 4/E. Pearson Education, 2007.
- [58] M. Gleicher, “Retargetting motion to new characters,” in *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH ’98*, (New York, NY, USA), p. 33–42, Association for Computing Machinery, 1998.
- [59] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Computation*, vol. 9, pp. 1735–1780, 11 1997.

- [60] H. F. Song, A. Abdolmaleki, J. T. Springenberg, A. Clark, H. Soyer, J. W. Rae, S. Noury, A. Ahuja, S. Liu, D. Tirumala, *et al.*, “V-MPO: On-policy maximum a posteriori policy optimization for discrete and continuous control,” *arXiv preprint arXiv:1909.12238*, 2019.
- [61] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.
- [62] N. Tishby, F. C. Pereira, and W. Bialek, “The information bottleneck method,” *arXiv preprint arXiv:physics/0004057*, 2000.
- [63] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, “Deep variational information bottleneck,” *arXiv preprint arXiv:1612.00410*, 2019.
- [64] A. Galashov, S. Jayakumar, L. Hasenclever, D. Tirumala, J. Schwarz, G. Desjardins, W. M. Czarnecki, Y. W. Teh, R. Pascanu, and N. Heess, “Information asymmetry in KL-regularized RL,” in *International Conference on Learning Representations*, 2019.
- [65] A. Goyal, R. Islam, D. Strouse, Z. Ahmed, H. Larochelle, M. Botvinick, S. Levine, and Y. Bengio, “Transfer and exploration via the information bottleneck,” in *International Conference on Learning Representations*, 2019.
- [66] K. Perlin, “An image synthesizer,” in *Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '85*, (New York, NY, USA), p. 287–296, Association for Computing Machinery, 1985.

## Acknowledgments

Some of data used in this project was obtained from mocap.cs.cmu.edu. The database was created with funding from NSF EIA-0196217.

## Supplementary Materials

### Nomenclature

#### Symbols

$A$  Action space of the MDP.

$O$  Observation space of the MDP.

$P_O$  Probability of observations given states of the MDP.

$P$  State transition probability of the MDP.

$S$  State space of the MDP.

$V(\cdot)$  Value function.

$\alpha$  Time constant of the AR prior.

$\beta$  Regularization coefficient.

$\delta$  Termination metric for the imitation task.

$\eta$  Threshold on the termination metric for the imitation task.

$p$  A position.

$v$  A velocity.

$\mathcal{B}$  Set of bodies,  $\subset \mathbb{N}$ .

$\mathcal{E}$  Set of end effectors,  $\subset \mathbb{N}$ .

$\mathcal{J}$  Set of joints,  $\subset \mathbb{N}$ .

$\mathcal{N}(\mu, \Sigma)$  Normal distribution with mean  $\mu$  and covariance  $\Sigma^2$ .

$\mathcal{Z}$  Latent space.

$\phi$  Temperature coefficient to scale rewards.

$\pi(a | s)$  A policy mapping observations to action distributions.

$\theta_t$  Filtering constant at time  $t$ .

$a_t$  Low-level action at time  $t$  sent to the actuators.

$h_t$  Hidden memory state at time  $t$ .

$o_t$  Observation at time  $t$ .

$p(\cdot)$  Latent prior distribution.

$p_0$  Initial state distribution of the MDP.

$r_t$  Reward at time  $t$ .

$s_t$  State at time  $t$ .

$x_t$  Context at time  $t$  used to generate latent commands.

$y_t$  Task-specific observations at time  $t$  during downstream tasks.

$z_t$  Latent action at time  $t$  sent to the low-level controller.

### Acronyms

**AR(1)** Order 1 Autoregressive.

**DoF** Degree of Freedom.

**KL** Kullback–Leibler.

**LSTM** Long Short-Term Memory.

**MLP** Multilayer Perceptron.

**MoCap** Motion Capture.

**MPC** Model Predictive Control.

**MSE** Mean Squared Error.

**NPMP** Neural Probabilistic Motor Primitives.

**POMDP** Partially Observed MDP.

**RL** Reinforcement Learning.

**SEA** Series Elastic Actuator.

**VAE** Variational Auto-Encoder.

Table 1 | Parameters of the observation noise and delays during imitation and reuse.  $\text{Exp}(\beta)$  is the exponential distribution with scale  $\beta$ ,  $\mathcal{N}(0, \sigma^2)$  is the normal distribution with variance  $\sigma^2$ .

	Distribution	ANYmal	OP3
Delay (s)	$d_t + b$ , $d_t \sim \text{Exp}(a)$	$a = 0.0025$ , $b = 0.0025$	$a = 0.015$ , $b = 0.015$
Joint position (rad)	$d_t \sim \mathcal{N}(0, a)$	$a = 5e^{-4}$	$a = 5e^{-3}$
Angular velocity (rad/s)	$d_t \sim \mathcal{N}(0, a)$	$a = [0.1, 0.2, 0.8]$	$a = 0.002$
Linear acceleration ( $\text{m/s}^2$ )	$d_t \sim \mathcal{N}(0, a)$	$a = 0.01$	$a = 0.02$
Base orientation (rad)	$d_t \sim \mathcal{N}(0, a)$	$a = 0.01$	$a = 0.035$

## Hardware Details

**ANYmal** The quadruped considered in this work is the ANYmal B300 [42]. It stands 70cm high at rest, weighing about 33kg. It has 12 Degrees of Freedom (DoF) controlled by AnyDrive SEAs. The control stack in our work has 3 asynchronous control loops: the agent loop at 50Hz, the main control loop at 400Hz and finally the per-drive loop at 2.5KHz. Control inputs from the agent are communicated to the main control loop via ROS over TCP, and then to each individual drive via EtherCat. The control inputs in this work are position setpoints for each joint, tracked by a PD controller in each drive. We use  $P = 100\text{Nm/rad}$  and  $D = 0.25\text{Nm/rpm}$ . The main control loop is furthermore responsible for switching controllers, running state estimation, processing joystick input, etc.. In our case it also performs the delayed first-order hold [57] interpolation of the position setpoints (followed by zero-order hold at 400Hz).

**OP3** The OP3 [43] is the small humanoid robot used in this work, standing 51cm tall and weighing 3.5kg. It's 20 DoF are actuated with Dynamixel servos. Like with ANYmal, the control stack consist of 3 asynchronous loops: the agent loop at 33Hz, a separate main control loop at 200Hz and then individual control loops in the Dynamixels. Similarly communication between the agent and main control loops is via ROS, while communications to the individual servos is via the Dynamixel protocol. Control inputs are again position setpoints tracked by a P controller, with  $P = 15\text{Nm/rad}$ . We do not use any interpolation in this case.

## Simulation Details

**Observation Noise Models** To increase the realism of the simulations, we emulate (additive) noisy and delayed observations for our controllers. Noise magnitudes are upper bounds to what we typically measure on hardware to provide additional robustness. Delays incorporate any effects from asynchronous control loops, inference time and communication delay, and are for simplicity shared between observations (i.e. we sample a single delay per control step). We use the same noise and delay distribution in the imitation and reuse phases of our approach. Note that we don't use delays for any privileged observations (e.g. ground truth state) as observed by e.g. the critic. Table 1 lists the different distributions used for ANYmal and OP3.

**Domain Randomization** During both imitation and reuse stages of our approach we apply a set of domain randomizations to improve robustness and invariance of our controllers. Table 2 list the model randomizations we use for both ANYmal and OP3. Besides these model randomization, we also apply perturbations on the robots in the form of random lateral forces on the robot's base. The magnitude of the forces is sampled from  $\text{Exp}(a)$  and are held with a duration sampled from  $\text{Exp}(b)$ , with  $\text{Exp}(\beta)$  being the exponential distribution with scale  $\beta$ . The time between impulses is sampled from  $\text{Exp}(c)$ . Table 3 lists the parameters used. Note that we use smaller and shorter but more frequent perturbations during imitation, to balance out robustness with early terminations as described in *Task Definition*. For the dribbling task we also slightly randomize the ball size and weight, see

Table 2 | Robot model randomization settings during imitation and reuse. We randomize various attributes of different types of model elements. Variations can either be directly setting shared global values, scales that are applied multiplicatively to separate elements, or offsets that are applied additively.  $\mathcal{B} \subset \mathbb{N}$  represents the set of body indices,  $\mathcal{J} \subset \mathbb{N}$  those of the joints. Subscript  $g$  indicates a single, global value applied across all elements.

Element	Attribute	Type	Distribution	ANYmal	OP3
Body	Mass (kg)	Scale	$(1 + s_g) \cdot (1 + s_i), \forall i \in \mathcal{B}, s_g \sim \mathcal{U}(-a, a), s_i \sim \mathcal{U}(-b, b)$	$a = 0.3,$ $b = 0.1$	$a = 0.3,$ $b = 0.1$
	Centre of mass (m)	Offset	$d_i \sim \mathcal{U}(-a, a), \forall i \in \mathcal{B}$	$a = 0.02$	$a = 0.02$
Joint	Position (m)	Offset	$d_i \sim \mathcal{U}(-a, a), \forall i \in \mathcal{J}$	$a = 0.02$	$a = 0.005$
	Reference (rad)	Offset	$d_i \sim \mathcal{U}(-a, a), \forall i \in \mathcal{J}$	$a = 0.1$	$a = 0.1$
	Damping (Nm/rad/s)	Global	$d_g + d_i + c, \forall i \in \mathcal{J}, d_g \sim \mathcal{U}(0, a), d_i \sim \mathcal{U}(0, b)$	$a = 0.1,$ $b = 0.02$ $c = 0.$	$a = 0.1,$ $b = 0.02$ $c = 1.084$
	Friction loss (Nm)	Global	$(1 + s_g) \cdot (1 + s_i) \cdot c, \forall i \in \mathcal{J}, s_g \sim \mathcal{U}(-a, a), s_i \sim \mathcal{U}(-b, b)$	$a = 0.5,$ $b = 0.1,$ $c = 0.1$	$a = 0.5,$ $b = 0.1,$ $c = 0.03$
Geom	Friction (—)	Global	$d_g + b, d_g \sim \mathcal{U}(-a, a)$	$a = 0.2,$ $b = 0.6$	$a = 0.2,$ $b = 0.6$
Actuator	P gain (Nm/rad)	Global	$d_g + b, d_g \sim \mathcal{U}(-a, a)$	—	$a = 2.,$ $b = 15.$
	Torque limit (Nm)	Global	$d_g + b, d_g \sim \mathcal{U}(-a, a)$	—	$a = 0.1,$ $b = 4.1$

Table 4. Finally, for the controllable walking task with ANYmal specifically, we also sample procedural terrains for a fixed list of 128 pre-calculated terrains with a maximum height difference of 0.3m.

**ANYmal Actuator Model** To decrease the simulation-to-reality gap on the ANYmal platform specifically, we use learned actuator networks to model the [SEAs](#) in simulation, following previous work [13, 56]. We however use a different, more flexible architecture, as shown in Figure 9. The model is split in two parts. First is a classic PID controller, generating reference torque  $\hat{\tau}$  based on open-loop torque  $\bar{\tau}$ , joint velocity  $\dot{\theta}$  and joint position error  $\epsilon$ . The learned part of the actuator model then models the torque trans-

fer function of the actuator, translating (a history of) reference torque  $\hat{\tau}$ , joint velocity  $\dot{\theta}$ , temperature  $T$  and battery voltage  $V$  to actual instantaneous joint torque  $\tau$  and current  $I$ . This split allows us to change the control mode and gains on the fly without having to recollect data or retrain the actuator model. The learned part of the model is similar to an auto-regressive Wavenet-style [72] stack of 4 layers of 1D strided & dilated convolutions, creating a total receptive field of 8 timesteps. Each convolution is followed by a tanh non-linearity. The output of the model is the residual with respect to the previous timestep’s output. We collect a dataset of half an hour of robot data in total, containing various different controllers and control modes. We use a 10:1:1 split of the actuators for train, validation and test

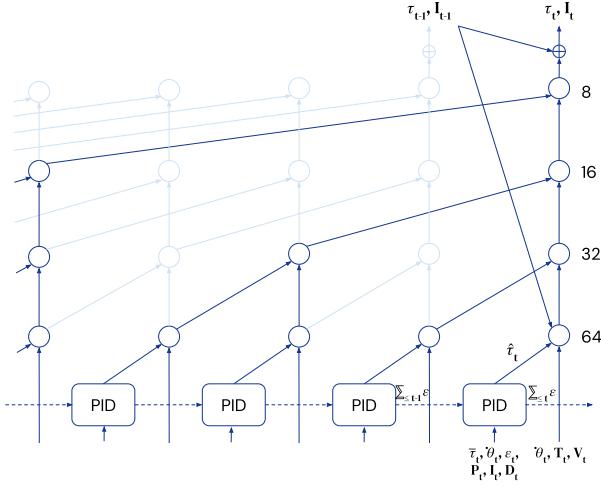


Figure 9 | Architecture of the actuator model used to simulate the SEA of ANYmal, consisting of an exact PID controller followed by a learned stack of 1D convolutions predicting both instantaneous torque and current autoregressively. Numbers on the right indicate the size of the feature vector.

sets, giving us a total of 5 hours of training data at 400Hz. We train the model by minimizing the **Mean Squared Error (MSE)** over a mini-batch of 16 and an unroll length of 1600 steps (4 seconds) for backpropagation-through-time. ADAM with a learning rate of  $1e^{-3}$  is used as the optimizer. The resulting model has a train/test root **MSE** of 0.50/0.54Nm and 1.48/1.64A.

### Motion Capture Retargeting Details

**ANYmal** We point-cloud retargeting procedure similar to [48, 58] to retarget the dog **MoCap** dataset from [33] to the ANYmal robot. Specifi-

Table 3 | Parameters of the perturbation applied to the robot’s base during imitation and reuse.

	ANYmal	OP3
Imitation	$a = 5,$ $b = 0.5,$ $c = 2$	$a = 1,$ $b = 0.5,$ $c = 2$
	$a = 40,$ $b = 1,$ $c = 5$	$a = 4,$ $b = 1,$ $c = 5$
Reuse		

cally, we optimize the following objective:

$$\theta^*, \mathbf{q}_t^{*0..T} = \arg \min_{\theta} \sum_{t=0}^T \arg \min_{\mathbf{q}_t} \|f(\theta, \mathbf{q}_t) - \mathbf{p}_t^{ref}\|_2^2, \quad (8)$$

where  $\mathbf{q}_t$  is the per-timestep vector of joint positions,  $\theta$  are the coordinates of the markers, in the frame of the body they are attached to, that correspond to (fixed) markers on the reference dog,  $\mathbf{p}_t^{ref}$  are the per-timestep global positions of the markers on the dog and  $f(\cdot)$  is the forward kinematics function. We place corresponding markers on the feet, hips, shoulders and the center of the base. We alternate optimizing joint positions per frame in the dataset with optimizing the position of the markers on the robot jointly over all frames. Optimizing the joint positions is a standard least-squares problem with known Jacobian so converges fast. We furthermore initialize the inner optimization problem at  $t$  with the solution at  $t - 1$ .

Since the ANYmal robot is proportionally wider than the reference dog, the robot’s legs tend to fold inwards, leading to less stable poses due to the smaller support polygon. This is resolved by adding a small regularization penalty to Equation (8) towards a stable standing pose for ANYmal,  $\beta \cdot \|\mathbf{q}_t - \mathbf{q}_t^{ref}\|_2^2$  with  $\beta = 0.01$ , which causes the markers on the feet to move inwards with respect to the feet themselves. Furthermore, we enforce left-right and front-back symmetry of the markers, which allows us to trivially mirror the reference motions as a form of data augmentation and allows e.g. the ANYmal to walk backwards (as no backwards walking is otherwise present in the dataset).

After retargeting, we filter out parts of reference trajectories that have joint positions or velocities which exceed the actuator specifications, where the height of all the feet deviates too much from the ground plane or where the dog is not significantly moving for an extended period. We further chunk the clips into segments of a length of maximum 10 seconds. In total this gives us 2.5 hours of reference trajectories. Finally we interpolate the trajectories using cubic and SQUAD interpolation. This allows us to train imitation policies at control rates different from the **MoCap**

Table 4 | Parameters of the model randomizations applied to the ball.

	Distribution	ANYmal	OP3
Radius	$d_g + b, d_g \sim \mathcal{U}(-a, a)$	$a = 0.0025, b = 0.1075$	$a = 0.005, b = 0.065$
Mass	$d_g + b, d_g \sim \mathcal{U}(-a, a)$	$a = 0.02, b = 0.43$	$a = 0.02, b = 0.182$

frame rate.

**OP3** For the OP3, we start from roughly 1.5 hours of walking and running trajectories from [37] that have already been retargeted to the CMUHumanoid model from the `dm_control` [55] **RL** environment package. These trajectories originally came from the CMU motion capture dataset [44]. This allows us to transfer most joint positions from the CMUHumanoid model directly to the OP3 model. However, since the OP3 has no degrees of freedom in the torso, we have to combine all upper body movements into the hip joints. Specifically, we treat the entire upper body as a single rigid body and rotate the OP3 model so that its torso orientation agrees with the uppermost spine frame of the CMU humanoid, and use the three hip joints to match the orientation of the upper leg relative to the torso. We also scale translations along each trajectory by the ratio between leg lengths of the two walker models, which works well for trajectories where there are regular contacts between a foot and the ground. The advantages of this approach compared to full point-cloud retargeting are that we do not need to determine marker placements, and that we leverage known-good trajectories that were successfully used in a number of other projects in the past.

### Task Details

**Motion Capture Imitation** The reward for **MoCap** imitation consists of several terms, following [39]:

$$r = \frac{1}{2} \cdot r_{trunc} + \frac{1}{2} \cdot (a \cdot r_{com} + r_{vel} + b \cdot r_{app} + c \cdot r_{quat}), \quad (9)$$

where  $r_{trunc} = 1 - \delta/d$  with  $\delta$  as defined in Equation (2). In this work  $d$  is 0.3. The remaining terms penalize the difference between the current and reference center of mass, joint velocities,

end effector positions and body quaternions respectively:

$$r_{com} = \exp \left( -d \cdot \left\| p_{com} - p_{com}^{ref} \right\|_2^2 \right) \quad (10)$$

$$r_{vel} = \exp \left( -e \cdot \sum_{i \in \mathcal{J}} \left\| q_{vel,i} - q_{vel,i}^{ref} \right\|_2^2 \right) \quad (11)$$

$$r_{app} = \exp \left( -f \cdot \sum_{i \in \mathcal{E}} \left\| p_{app,i} - p_{app,i}^{ref} \right\|_2^2 \right) \quad (12)$$

$$r_{quat} = \exp \left( -g \cdot \sum_{i \in \mathcal{B}} \left\| q_{quat,i} \ominus q_{quat,i}^{ref} \right\|_2^2 \right), \quad (13)$$

where  $\mathcal{B} \subset \mathbb{N}$  is the set of bodies,  $\mathcal{E} \subset \mathbb{N}$  is the set of end effectors,  $\mathcal{J} \subset \mathbb{N}$  is the set of joints and  $\ominus$  is the quaternion difference. Values for ANYmal and OP3 are listed in Table 5.

For ANYmal specifically we add an additional term to the overall imitation reward that penalizes the current draw of the actuators:

$$r_{amp} = -5e^{-4} \cdot \sum_{i \in \mathcal{J}} I_i^2. \quad (14)$$

**Controllable Walking** For the controllable walking task we generate target velocities according to a memoryless random process. Specifically, we change velocities every  $\tau \sim \text{Exp}(5)$  seconds, with every component being updated according to

Table 5 | List of coefficients used in the MoCap imitation task.

Coefficient	ANYmal	OP3
$a$	0.1	0.1
$b$	0.15	0.15
$c$	0.65	0.65
$d$	20.	40.
$e$	0.1	0.1
$f$	80.	160.
$g$	2.	2.

Table 6 | Parameters of the random process generating target velocities for the controllable walking task, consisting of a platform-specific range  $a$  as well as a probability  $b$  that each component is  $\neq 0$ .

	ANYmal	OP3	Probability $\neq 0$
Forward	1.5	0.4	0.9
Lateral	0.4	0.2	0.25
Yaw rate	1.2	1	0.5

$$x_{k+1} = x_k - w_k \cdot (x_k - y_k \cdot z_k), \quad (15)$$

where  $x_k$  is the previous target velocity,  $y_k \sim \mathcal{U}(-a, a)$  a uniformly sampled velocity,  $w_k \sim \text{Bern}(b)$  the probability the velocity is  $\neq 0$  and  $w_k \sim \text{Bern}(0.5)$  a binary random variable indicating whether we retain the previous target or not.  $z_k$  makes sure that we put enough emphasis on the target velocity = 0, which we find to be required for the agent to learn to stand still.  $w_k$  on the other hand ensures not every component of the target vector switches every time, to decorrelate changes in one dimension from the others. Note that  $a$  and  $b$  are different for each of the target velocity components, see Table 6 for the values used. We use the same ranges for  $a$  to scale joystick inputs in  $[-1, 1]$  to target velocities when giving control of the real robot to a user.

The reward is based on the error between the current velocity  $v_t$  and target velocity  $\hat{v}_t$ :

$$r_t = \exp\left(-\|v_t - \hat{v}_t\|_2^2 / \eta\right), \quad (16)$$

where we use a value for  $\eta$  of 0.5 and 0.05 for ANYmal and OP3 respectively, to account for the difference in expected velocities. For OP3 we also found it useful to replace the instantaneous velocity  $v_t$  with an exponentially filtered velocity  $\tilde{v}_t$ , where we use a filter constant of 0.95. This averages out the velocity over approximately a gait cycle and allows the controller to “sway” the robot’s base more without affecting the reward.

**Ball Dribbling** Similar to the controllable walking task, we generate ball target positions according to a random process. The target position gets

updated every  $\tau \sim \text{Exp}(10)$  seconds, by translating the previous target position in a random direction by a distance  $\delta \sim \mathcal{U}(a, b)$ . We use values  $(a, b)$  of  $(0.5, 2.)$  and  $(0.3, 1.5)$  for ANYmal and OP3 respectively.

The reward is based on the error between the current velocity  $v_t$  and target velocity  $\hat{v}_t$ :

$$r_t = \exp\left(-\|\mathbf{x}_t^{\text{ball}} - \hat{\mathbf{x}}_t\|_2^2 / \eta\right) \quad (17)$$

where we use a value for  $\eta$  of 1. and 0.5 for ANYmal and OP3 respectively.

Besides the terminations listed in *Downstream Tasks* in the main text, we terminate the episode when the distance between the walker and the ball or the ball and the target exceeds 5 meters, to avoid long stretches of episode without any interaction with the ball.

## Training Details

We train our imitation controllers using MO-VMPO [29] and reuse controllers using standard VMPO [60]. We use a training setup similar to IMPALA [82] where we have a large number of parallel, asynchronous actors collecting environments which are added to a shared queue. A single learner process running on a 1x1 TPUv2 [83] then samples batches from this queue to optimize the controllers’ parameters. Each sample in the batch is a fixed-length sequence of environment steps, used for backpropagation-through-time and n-steps returns to train the value function. Table 7 lists the hyperparameters used for (MO-)VMPO used during both phases of our approach. We use identical parameters for both ANYmal and OP3, and for all reuse tasks. Experiments terminate after a maximum number of environment steps and take 1.5 to 2 days to complete.

**Network Architecture** We refer back to Figure 8 for an overview of the controller architectures. Furthermore Table 8 lists the various inputs for the separate neural network components in our approach and their dimensionality. Unless otherwise mentioned we use tanh activations. The imitation controller takes as input the MoCap reference frames  $x_t$  as well as the previous latent

Table 7 | Hyperparameter settings for (MO-)VMPO. Parameters are identical for ANYmal & OP3, and for all reuse tasks.  $k$  is the current number of environment steps processed.

Hyperparameter	Imitation	Reuse
Number of parallel actors	8192	8192
Number of environment steps	$3e^{10}$	$3e^{10}$
Optimizer	ADAM	ADAM
Learning rate	$1e^{-4}$	$1e^{-4}$
KL regularization coefficient	$0.3 \cdot \left(1 - \left(1 - \min\left(1, \frac{k}{1.5e^{10}}\right)\right)^{0.2}\right)$	0.01
Batch size	512	512
Unroll length	20	20
Target network update period	100	100
E-step advantages	Top 50%	Top 50%
E-step $\epsilon$	$1e^{-2}$ for all terms except $1e^{-4}$ for current term	$1e^{-1}$
M-step $\epsilon_\mu$	$1e^{-1}$	$1e^{-1}$
M-step $\epsilon_\Sigma$	$1e^{-5}$	$1e^{-5}$
Discount $\gamma$	0.98	0.99

command  $z_t$  and encodes it using a two-layer MLP to output  $\pi_{HL}(z_t | z_{t-1}, x_t)$  per Equation (3). Each layer has 1024 units and we also LayerNorm [84]. In the reuse phase we replace the encoder with a new high-level policy  $\pi_{HL}(z_t | z_{t-1}, o_{\leq t}, y_t)$ . We concatenate  $z_{t-1}$ ,  $o_t$  and  $y_t$  and first pass them through an input normalisation layer. This is a single linear layer with 256 outputs followed by LayerNorm and activation, the output of which is passed through a two-layer MLP with 256 units per layer and finally an LSTM [59] layer with 256 cells. We parameterize the output according to Equation (7), with  $\Sigma_{HL}$  initialized to 0.5.

The low-level controller is slightly more involved. After an initial input normalisation layer applied to the (noisy and delayed) proprioception  $o_t$ , it splits off in two branches. The first branch is similar to  $\pi_{HL}$  with a two-layer MLP with 256 units per layer and LSTM with 256 cells. The output of this branch is concatenated with the output of the input normalisation layer and the latent command  $z_t$  and passed to the second branch. This is another two-layer MLP with 256 units per layer. Finally the output of the two branches is concatenated and followed by a single linear layer to output  $\pi_{LL}$  as in Equation (4), with  $\Sigma_{LL}$  initialized to 0.2.

During the imitation phase, the value function

$V$  gets as input the ground truth state of the simulation  $s_t$ , the MoCap reference  $x_t$  as well as embeddings of the specific MoCap clip being imitated and the randomized physics parameters. This then gets passed through a three-layer MLP with 1024 units per layer and LayerNorm, with a final linear layer on top to predict the value(s). The clip embedding is a lookup table that maps the identified of the clip to a fixed-length real-valued embedded that is learned together with the value function. The physics parameters are embedded by first passing them through an input normalisation layer of 512 units and then a two-layer MLP with 128 and 32 units. In the reuse phase we train a new value function, omit the clip embedding and replace  $x_t$  by the task-specific observations  $y_t$ .

## Supplementary References

- [67] M. Hutter, C. Gehring, D. Jud, A. Lauber, C. D. Bellicoso, V. Tsounis, J. Hwangbo, K. Bodie, P. Fankhauser, M. Bloesch, *et al.*, “ANYmal - a highly mobile and dynamic quadrupedal robot,” in *2016 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 38–44, IEEE, 2016.
- [68] J. Proakis, *Digital Signal Processing: Prin-*

Table 8 | The inputs for the neural networks in our approach and their dimensionality. Notation  $[a, b]$  indicates a dimensionality of  $a$  for ANYmal and  $b$  for OP3. See Table 2 for the list of parameters included in the model randomizations. Note that  $o_t$  are noisy and delayed, whereas  $s_t$  are ground truth values.

Input	Component	Dimensionality
MoCap reference $x_t$	Relative body positions	$[13, 20] \times 3 \times 5$
	Relative body orientations	$[13, 20] \times 4 \times 5$
Latent command $z_t$	—	$[12, 20]$
Proprioception $o_t$	Joint positions	$[12, 20]$
	Position setpoints	$[12, 20]$
	Angular velocity	3
	Linear acceleration	3
	Gravity vector	3
Walking obs. $y_t$	Target velocity	3
Dribbling obs. $y_t$	Ball position	3
	Target position	3
Ground truth state $s_t$	Joint positions	$[12, 20]$
	Position setpoints	$[12, 20]$
	Joint velocities	$[12, 20]$
	Linear velocity	3
	Angular velocity	3
	End-effector position	$4 \times 3$
	Gravity vector	3
MoCap clip ID embedding	—	30
Model randomizations	—	[238, 807]

- ciples, Algorithms, And Applications, 4/E. Pearson Education, 2007.
- [69] Robotis, “OP3.” <https://emanual.robotis.com/docs/en/platform/op3/introduction/>, 2017.
- [70] J. Hwangbo, J. Lee, A. Dosovitskiy, D. Belluscio, V. Tsounis, V. Koltun, and M. Hutter, “Learning agile and dynamic motor skills for legged robots,” *Science Robotics*, vol. 4, no. 26, 2019.
- [71] S. Gangapurwala, M. Geisert, R. Orsolino, M. Fallon, and I. Havoutis, “RLOC: Terrain-aware legged locomotion using reinforcement learning and optimal control,” *arXiv preprint arXiv:2012.03094*, 2020.
- [72] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “Wavenet: A generative model for raw audio,” *arXiv preprint arXiv:1609.03499*, 2016.
- [73] X. B. Peng, E. Coumans, T. Zhang, T.-W. E. Lee, J. Tan, and S. Levine, “Learning agile robotic locomotion skills by imitating animals,” in *Robotics: Science and Systems*, 07 2020.
- [74] M. Gleicher, “Retargetting motion to new characters,” in *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH ’98*, (New York, NY, USA), p. 33–42, Association for Computing Machinery, 1998.
- [75] H. Zhang, S. Starke, T. Komura, and J. Saito, “Mode-adaptive neural networks for quadruped motion control,” *ACM Trans. Graph.*, vol. 37, July 2018.
- [76] J. Merel, L. Hasenclever, A. Galashov, A. Ahuja, V. Pham, G. Wayne, Y. W. Teh, and N. Heess, “Neural probabilistic motor primitives for humanoid control,” in *International Conference on Learning Representations*, 2019.
- [77] Y. Tassa, S. Tunyasuvunakool, A. Muldal, Y. Doron, S. Liu, S. Bohez, J. Merel, T. Erez, T. Lillicrap, and N. Heess, “dm\_control: Software and tasks for continuous control,” *arXiv preprint arXiv:2006.12983*, 2020.
- [78] CMU Graphics Lab, “Carnegie Mellon University Graphics Lab Motion Capture Database.” <http://mocap.cs.cmu.edu/>, 2003.
- [79] L. Hasenclever, F. Pardo, R. Hadsell, N. Heess, and J. Merel, “CoMic: Complementary task learning & mimicry for reusable skills,” in *International Conference on Machine Learning*, PMLR, 2020.
- [80] A. Abdolmaleki, S. H. Huang, L. Hasenclever, M. Neunert, F. Song, M. Zambelli, M. F. Martins, N. Heess, R. Hadsell, and M. Reidmiller, “A distributional view on multi-objective policy optimization,” in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.
- [81] H. F. Song, A. Abdolmaleki, J. T. Springenberg, A. Clark, H. Soyer, J. W. Rae, S. Noury, A. Ahuja, S. Liu, D. Tirumala, *et al.*, “V-MPO: On-policy maximum a posteriori policy optimization for discrete and continuous control,” *arXiv preprint arXiv:1909.12238*, 2019.
- [82] L. Espeholt, H. Soyer, R. Munos, K. Simonyan, V. Mnih, T. Ward, Y. Doron, V. Firoiu, T. Harley, I. Dunning, S. Legg, and K. Kavukcuoglu, “IMPALA: Scalable distributed deep-rl with importance weighted actor-learner architectures,” *arXiv preprint arXiv:1802.01561*, 2018.
- [83] Google, “Cloud tpu.” <https://cloud.google.com/tpu/>, 2018.
- [84] J. L. Ba, J. R. Kiros, and G. E. Hinton, “Layer normalization,” *arXiv preprint arXiv:1607.06450*, 2016.
- [85] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Computation*, vol. 9, pp. 1735–1780, 11 1997.