Learning to Act Through Contact: A Unified View of Multi-Task Robot Learning

Shafeef Omar SHAFEEF.OMAR@TUM.DE
Majid Khadiv MAJID.KHADIV@TUM.DE
Munich Institute of Robotics and Machine Intelligence, Germany
Technical University of Munich, Germany

Editors: G. Sukhatme, L. Lindemann, S. Tu, A. Wierman, N. Atanasov

Abstract

We present a unified framework for *multi-task* locomotion and manipulation policy learning grounded in a contact-explicit representation. Instead of designing different policies for different tasks, our approach unifies the definition of a task through a sequence of contact goals–desired contact positions, timings, and active end-effectors. This enables leveraging the shared structure across diverse contact-rich tasks, leading to a single policy that can perform a wide range of tasks. In particular, we train a goal-conditioned reinforcement learning (RL) policy to realise given contact plans. We validate our framework on multiple robotic embodiments and tasks: a quadruped performing multiple gaits, a humanoid performing multiple biped and quadrupedal gaits, and a humanoid executing different bimanual object manipulation tasks. Each of these scenarios is controlled by a single policy trained to execute different tasks grounded in contacts, demonstrating versatile and robust behaviours across morphologically distinct systems. Our results show that explicit contact reasoning significantly improves generalisation to unseen scenarios, positioning contact-explicit policy learning as a promising foundation for scalable loco-manipulation. Video available at: https://youtu.be/L1vjmQqvc4M

Keywords: Goal-Conditioned RL, Task-Agnostic Policy, Contact-Explicit

1. Introduction

Advances in reinforcement learning (RL) have enabled robots to master complex motor skills, from agile quadruped locomotion Hoeller et al. (2023); Cheng et al. (2023) to dexterous object manipulation OpenAI et al. (2019); Singh et al. (2025); Yin et al. (2025). Yet, prevailing RL policies are often trained with task-specific objectives, making them difficult to transfer to unseen scenarios without retraining from scratch. For example, perceptive locomotion policies are typically tasked to train on velocity and/or position commands that work well on various rough terrain scenarios they have been trained on Rudin et al. (2022b); Miki et al. (2022). Nonetheless, they cannot be directly transferred to environments with sparser footholds and riskier terrains Zhang et al. (2024a), such as stepping stones, even though the required motions are similar to those on which it has been trained on. Similarly, in object manipulation, tasks like lifting an object from a table share similar motor skills with more complex tasks, such as stacking objects. However, the traditional approach of training policies on specific tasks makes it difficult to generalize across different types of physical interactions. Moreover, training a robot on a new task from scratch is not feasible each time it encounters one. This motivates seeking a more fundamental abstraction: one that also unifies locomotion and manipulation. In this light, we propose using *contacts* as a common representation to enable better generalization and adaptability across various physical interaction tasks and embodiments.

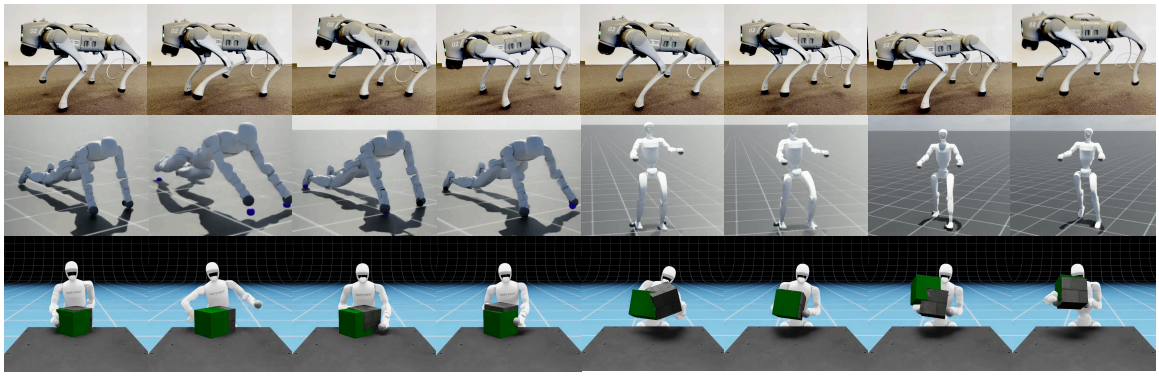

Figure 1: Snapshots of our contact-explicit framework in action. (Row 1): the quadruped demonstrates diverse gaits; (Row 2): the humanoid demonstrates locomotion using hand-assisted gaits such as quadrupedal jump, pace and bipedal gaits such as walk, jump; (Row 3): the humanoid carries out different bimanual manipulation tasks, where the green box represents the target pose.

Why Contacts? Contacts govern nearly all loco-manipulative behaviors. Locomotion requires coordinated foot placements with the ground, and manipulation relies on purposeful hand-object interactions—both inherently contact-driven. Despite their centrality, contacts are often treated as incidental in reinforcement learning (RL) frameworks, emerging implicitly as a byproduct of motion optimization. This abstraction leads to policies with limited generalization across tasks that share underlying motor principles. In contrast, humans intuitively decompose complex behaviors into contact-explicit subgoals: a climber plans handholds and footholds before ascending, and a parkour athlete sequences hand and foot placements relative to environmental features. These skills transfer seamlessly across structurally similar tasks.

While recent works have explored goal representations such as 3D position targets Sferrazza et al. (2024) or motion references He et al. (2024); Yin et al. (2025), they often overlook contact as a fundamental primitive. As a result, such approaches may struggle with tasks where contact timing and placement are crucial. Recent works have shown the benefits of explicitly incorporating contact—either in rewards or task representations Zhang et al. (2024b); Ciebielski and Khadiv (2024); Lin et al. (2025)—leading to better generalization and performance.

Our work builds on this contact-explicit paradigm, proposing contact goals as a unified task representation for locomotion and contact-rich manipulation, enabling a single policy to produce diverse physical behaviors. We decompose tasks into contact goals—defined by *target locations, timings*, and *active end-effectors*—together with object pose targets for manipulation. Given a high-level planner that generates sequences of these goals, a goal-conditioned RL policy learns to achieve them via joint torques across various contact modes. This bypasses task-specific reward design in the policy, instead treating contacts as fundamental physical primitives which robots interact with their environment.

We validate our framework with three demonstrations, each with a single policy: (1) on a quadruped robot to execute multiple gaits (trot, pace, bound, jump, and crawl), (2) on a humanoid robot performing multiple biped and quadruped gaits, and (3) on a humanoid robot to perform bi-

manual object manipulation such as object reorientation on a table and object pose tracking while being lifted. Our results demonstrate that using a contact-explicit representation, we can generate multiple skills with a single policy and leverage the shared information between tasks to improve generalization beyond the training distribution.

Our contributions can be summarised as follows:

1. We propose a goal-conditioned RL framework that learns to achieve given contact goals with the world. Using this framework, we train a single policy capable of performing multiple tasks.

2. We provide empirical validation on morphologically distinct robots and various tasks, showing that explicit contact reasoning enables dynamic and robust behaviors across diverse scenarios, while generalizing to unseen tasks.

2. Related Work

While locomotion and contact-rich manipulation are similarly realized through intermittent contacts with the environment, RL-based methods typically use different specialized rewards and task representations for each problem.

Locomotion. Early works represented predefined gaits (walking) with step location and robot heading as input Peng et al. (2017). More recent works avoid predefining gaits, and define the desired behavior (goal) through a desired average velocity Hwangbo et al. (2019); Rudin et al. (2022b) or reaching a desired position in the world Rudin et al. (2022a); Hoeller et al. (2023); Cheng et al. (2023). As this goal representation does not distinguish between different gaits, it is not suitable for multi-gait policy generation and its performance is highly constrained to its training distribution. To enable one policy for multiple gaits, recent approaches used either a notion of gait phase as input to the policy Margolis and Agrawal (2022); Bellegarda et al. (2024) or task-specific desired reference motion Tan et al. (2018); Li et al. (2024); Zargarbashi et al. (2024); Sleiman et al. (2024). The former representation is specific to locomotion and cyclic gaits, while the latter requires another module to generate desired trajectories for every behavior, which could become expensive. Unlike these approaches, our findings reveal that we do not need to use any parameterisation or dense reference trajectories to represent the different gaits, but rather intuitive contact goals can directly achieve them.

Manipulation/Loco-Manipulation. In manipulation, the desired behavior is usually specified through the desired object goal OpenAI et al. (2019); Lin et al. (2023) or task-specific goals such as for grasping Lum et al. (2024); Singh et al. (2025). While successful in learning single tasks, such a representation fails to work in most multi-task and the few-shot learning settings Chen et al. (2022). In loco-manipulation settings, most works simply concatenate separate locomotion and manipulation goals Pan et al. (2025); Fu et al. (2022) or define and train different tasks separately Dao et al. (2024); Liu et al. (2024); Qiu et al. (2024). However, such an approach does not enable leveraging the shared structure between locomotion and manipulation through contact. Sferrazza et al. (2024) trained whole-body controllers using 3D position targets for the robot's hands and then trained a high-level planner using these to perform loco-manipulation. They also released a suite of complex tasks as a benchmark for robot loco-manipulation. However, their low-level reaching policy ignores contacts, which are crucial for loco-manipulation and, consequently, fails in many tasks on their benchmark.

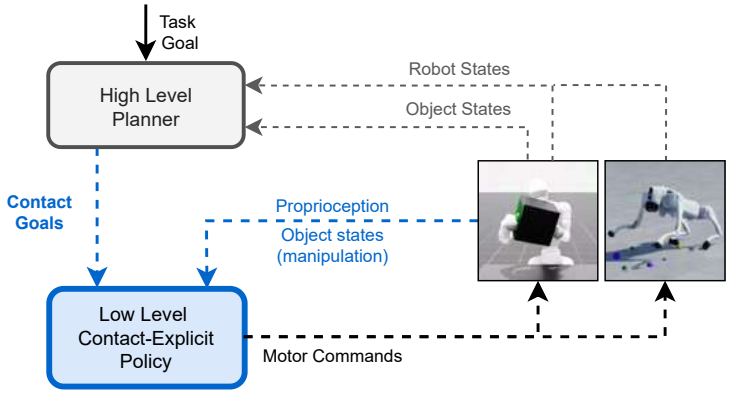

Figure 2: Overview of our contact-explicit framework. A high-level planner generates the contact goals (and object pose targets for manipulation), that is provided as immediate goals for the goal-conditioned RL policy to accomplish.

Contacts. Recent studies have shown that including contact information in the reward design and task representation can improve multi-task learning Zhang et al. (2024b); Ciebielski and Khadiv (2024); Lin et al. (2025). In particular, Ciebielski and Khadiv (2024) showed that a contact-centric representation for multi-gait locomotion learning improves the generalization capability of the gaits when compared to other representations. However, they only showed locomotion results in a behavioral cloning setting. Zhang et al. (2024b) used the contact information in the reward design for various locomotion and loco-manipulation tasks. They proposed sparse contact-based rewards that are then combined with task-specific rewards to enable complex motions such as humanoid parkour and loco-manipulation. Compared to Zhang et al. (2024b), which learns different policies for different tasks, we show that training one multi-skill policy outperforms the generalization capabilities of the policy to unseen tasks. Furthermore, different from Zhang et al. (2024b), we present a denser reward for contact that facilitates the training procedure and qualitatively produces smoother motions. Closest to our work is a recent study that also uses contact and object pose goals to train an RL policy only to perform bimanual dextrous manipulation Lin et al. (2025). We show that our proposed contact-conditioned policy generalizes better than the (sub)task-conditioned policies in Lin et al. (2025) for object manipulation. Furthermore, we show that contact-conditioned policies are general enough to be applied to the locomotion setting as well.

3. Method

3.1. Overview

At the core, we propose a contact-explicit representation that is used to train policies for multi-gait locomotion on a quadruped/humanoid and a multi-task bimanual manipulation on a humanoid robot. In particular, we train a goal-conditioned RL policy to track contact goals, provided by a planner as shown in Fig. 2. The **contact goals** for an end-effector e, $g_e^{\text{con}} = \{p_e^{\text{con}}, S_e^{\text{con}}, \mathcal{I}_e^{\text{con}}\}$, correspond to the 3D location of contact, the contact duration, and a binary indicator to be in contact, respectively. In the case of manipulation, g_e additionally comprises the object's goal pose $\{p_{obj}, \theta_{obj}\}$. A new set of contact goals is chosen when the contact duration expires and it achieves the desired contact.

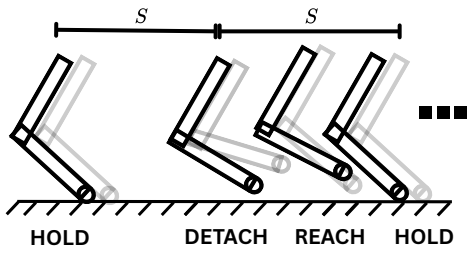

Figure 3: Simple illustration of a robot's end effector during different phases of contact, for a fixed command duration S. During the detach phase, the end effector is detached from a contact and is free to move. During the reach phase, the end effector is guided towards the desired contact location. During the hold phase, it maintains the contact.

As such, by composing several different contact goals, we can perform various long-horizon tasks. In this work, we have prespecified the contact goals required to achieve the various tasks. However, our method can be integrated with more sophisticated learned contact planners Omar et al. (2023); Dhedin et al. (2024), or even contact goals extracted from images/videos Taouil et al. (2025).

We formalize the problem of finding a multi-task policy $\pi(a_t|s_t, g_t)$ as a goal-conditioned RL problem which is formulated as a Markov Decision Process (MDP) $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma, \mathcal{G} \rangle$. This MDP is defined by states $s_t \in \mathcal{S}$, actions $a_t \in \mathcal{A}$, transtion probability $\mathcal{T}$, a reward $r_t \in \mathcal{R}$, a discount factor γ and goal $g_t \in \mathcal{G}$. The reward r_t is calculated to achieve different contact modes by following the immediate contact goals. The policy aims to maximise the expected return for achieving the contact goal g_t:

$$\max \mathbb{E}_\pi \left[\sum_t \gamma^t r_t(s_t, a_t, g_t) \right]$$

3.2. Learning to Act Through Contact

Contact Phases. We consider three phases for contact that allow us to make or break contact with the environment using any end-effector in a controlled manner, as illustrated in Fig. 3: reach (R), hold (H) and detach (D). During the reach phase of an end-effector, the robot must guide it to a desired contact location provided by the high-level planner. During the hold phase of an end-effector, the robot must maintain its contact where it was guided to during the reach phase. And during the detach phase of an end-effector, the robot is free to move it as long as it does not engage in contact. Viewing contacts from this perspective, we can develop dense rewards that allow the robot to explore several contact modes by making and breaking contacts with the world.

For an end-effector e at time t, the contact goals from the high-level planner comprise the following information: contact locations with a short horizon of two contact switches, $p^{\text{con}}_{t,e} = \left(\left[p^{\text{con}}_{t,e} \right]_1, \left[p^{\text{con}}_{t,e} \right]_2 \right)$, a binary indicator of contact for two contact switches, $(\mathcal{I}^{\text{con}}_{t,e} = \left(\left[\mathcal{I}^{\text{con}}_{t,e} \right]_1, \left[\mathcal{I}^{\text{con}}_{t,e} \right]_2 \right)$ and the command duration S of the current contact goal to be achieved. The contact phase of an end-effector is determined using the binary indicator, $\left[\mathcal{I}^{\text{con}}_{t,e} \right]_1$, and the time remaining to finish the contact command, s (where s is reset to the value of the newly sampled command duration when the previous one expires). If the remaining command duration is less than a threshold δ and the binary contact indicator is 0, we have the reach phase ($\left[\mathcal{I}^{\text{con}}_{t,e} \right]_1 = 0$ and $s < \delta$). If the binary contact

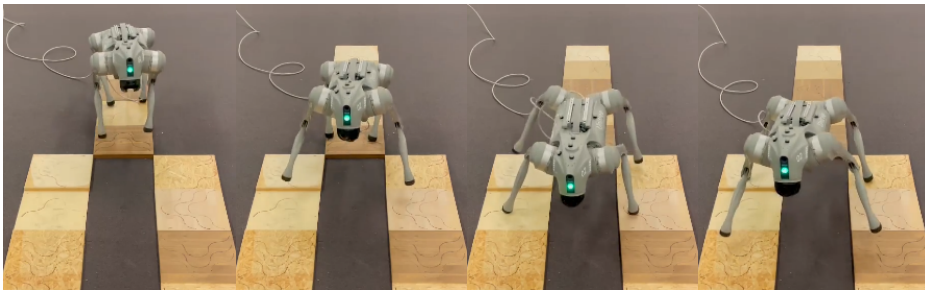

Figure 4: Quadruped robot crossing a gap with a bound gait by accurately adjusting the contact locations to remain on the wooden terrain.

indicator is 1, we have the hold phase ($\left[\mathcal{I}_{t,e}^{\text{con}}\right]_1 = 1$). We have the detach phase if the binary contact indicator is 0 and the remaining command duration is greater than the threshold δ ($\left[\mathcal{I}_{t,e}^{\text{con}}\right]_1 = 0$ and $s > \delta$). The binary contact indicators of all the end-effectors, $\mathcal{I}_{t,e}^{\text{con}}$, are stacked to form the contact sequence, referred to in the paper.

Policy Observations and Rewards. Apart from the proprioceptive inputs for the policy's observations, we provide explicit contact goals as described above, $p_{t,e}^{\text{con}}$, $\mathcal{I}_{t,e}^{\text{con}}$, and s, to achieve multiple tasks using the same shared structure. The following rewards are used to achieve the various contact phases described previously to perform multiple tasks:

$$r_{t,e}^{\text{reach}} = \exp\left(-\frac{d\left(\left[p_{t,e}^{\text{con}}\right]_1, p_{t,e}^{\text{act}}\right)}{\sigma^2}\right) \cdot \mathbb{I}\left[\left[\mathcal{I}_{t,e}^{\text{con}}\right]_1 = 0 \wedge s \leq \delta\right] \tag{1}$$

$$r_{t,e}^{\text{hold}} = \left(1 + \alpha_{\text{hold}} \cdot \exp\left(-\frac{d\left(\left[p_{t,e}^{\text{con}}\right]_1, p_{t,e}^{\text{act}}\right)}{\sigma^2}\right)\right) \cdot \mathbb{I}\left[\left[\mathcal{I}_{t,e}^{\text{con}}\right]_1 = \mathcal{I}_{t,e}^{\text{act}} = 1\right] \tag{2}$$

$$r_{t,e}^{\text{detach}} = \mathbb{I}\left[\left[\mathcal{I}_{t,e}^{\text{con}}\right]_1 = \mathcal{I}_{t,e}^{\text{act}} = 0 \wedge s > \delta\right] \tag{3}$$

where $d(\mathbf{a}, \mathbf{b})$ is the $L2$-norm between $\mathbf{a}$ and $\mathbf{b}$, and $\mathbb{I}(\cdot)$ is an indicator function. The hold reward incentivises the robot to maintain contact, and it gets a higher reward for making contact at the desired location for $\alpha_{hold} > 0$. The detach reward is a scalar reward that incentivises the agent to not make any contact during this phase.

For the case of manipulation, we additionally have a reward for tracking the object pose:

$$r_{t,obj}^{pose} = \frac{c_{pos}}{\epsilon_{pos} + \Delta p_{t,obj}} + \frac{c_{rot}}{\epsilon_{rot} + \Delta \theta_{t,obj}}$$

The total contact reward r_t^{con} is given as:

$$r_t^{\text{con}} = r_{t,obj}^{\text{pose}} + \sum_e \left(r_{t,e}^{\text{reach}} + r_{t,e}^{\text{hold}} + r_{t,e}^{\text{detach}}\right)$$

4. Experiments and Results

We evaluate our contact-explicit framework for performing multiple tasks on various robotic embodiments, such as a quadruped and a humanoid, and further conduct extensive experiments on them

for locomotion and bimanual manipulation. Our evaluations are based mainly on the multi-tasking and representation capabilities of our contact-explicit approach.

Locomotion. The quadruped is trained to perform multiple gaits, such as *trot, pace, bound, jump and crawl*, as depicted in row 1 of Fig 1. The contact locations are sampled to move in all directions. First, stride lengths and stance widths for each pair of front and hind legs are sampled for each environment upon initialization from a uniform distribution $\mathcal{U}(0.0, 0.3)m$ and $\mathcal{U}(0.1, 0.3)m$, with a sampled heading direction $\mathcal{U}[-\pi, \pi]rad$ or a with a sampled yaw rate to follow curved paths $\mathcal{U}[-\pi, \pi]rad/s$. From these locations, we further sample additional offsets to allow more versatility for each leg, $\mathcal{U}(-0.15, 0.15)m$, both in lateral and longitudinal directions. We find that sampling these values once during initialisation leads to a better policy compared to sampling upon every reset. The command durations were sampled from a narrow uniform distribution of $[0.34, 0.36]$ seconds. We use extensive domain randomisation Tan et al. (2018) to deploy our policy in the real world without additional fine-tuning. Fig 4 shows our contact-explicit policy deployed to navigate a challenging gap crossing scenario while remaining on the wooden platform, highlighting accurate tracking. A similar strategy as described above was also used to train the humanoid robot for various biped (walk and jump) and quadruped (crawl, pace, jump) locomotion modes.

For both the embodiments, the policy observes proprioceptive states such as joint positions, joint velocities, and task observations such as the current and next contact sequence of all feet, the current and next contact locations of all feet in base frame, the command duration, relative distance of the robot's feet to its desired contact locations. Notably, we do not use any contact sensing on the robot's feet since it did not make any difference in simulation performance. Apart from the rewards mentioned in 3.2, we additionally add penalties typically used in locomotion settings such as the base angular velocities, joint velocities, accelerations, torques, joint deviations, and action rate for smoother behaviours. We also update the goals if the robot's base when projected to the ground remains within a threshold of the desired contact locations and provide a bonus reward for discovering more goals.

Bimanual Manipulation. Using our method, the humanoid is trained to perform the following bi-manual manipulation tasks as shown in row 3 of Fig. 1: 1) *Repose*: The robot must continuously maintain contact on two surfaces of a box (cuboid) and must track several positions and orientations for the object, sampled from a uniform distribution. These object poses are in the air, hence the humanoid must learn to lift the object and not let it slip from its hand while tracking the various poses. 2) *Reorient*: The robot must make and break contact with the box repeatedly to keep rotating it 45° on the table each time it makes contact. The contact locations were predefined so that the correct surfaces could be chosen to rotate the object continuously. The command durations are sampled from a uniform distribution of $\mathcal{U}[1.0, 1.5]s$.

The policy observes proprioceptive robot states, object states, and task observations such as the current and next contact sequence of the robot's hands, the current and next contact location on the object's surface, the command duration, and the goal pose relative to the object's pose. Apart from the rewards mentioned in 3.2, we additionally provide penalties to penalise joint velocities, accelerations, torques, action rate and high impact forces. We only update the goals if the desired object pose is within a threshold of the actual object pose and provide a bonus reward for it.

Training. For all the tasks, we use the Proximal Policy Optimization (PPO) Schulman et al. (2017) algorithm with a recurrent architecture (GRU) and entropy decay to train the policy in IsaacLab Mittal et al. (2023) with 8192 parallel environments. We use PPO as our algorithm of choice since it is effective in learning low-level motion primitives, as also observed in Yin et al. (2025).

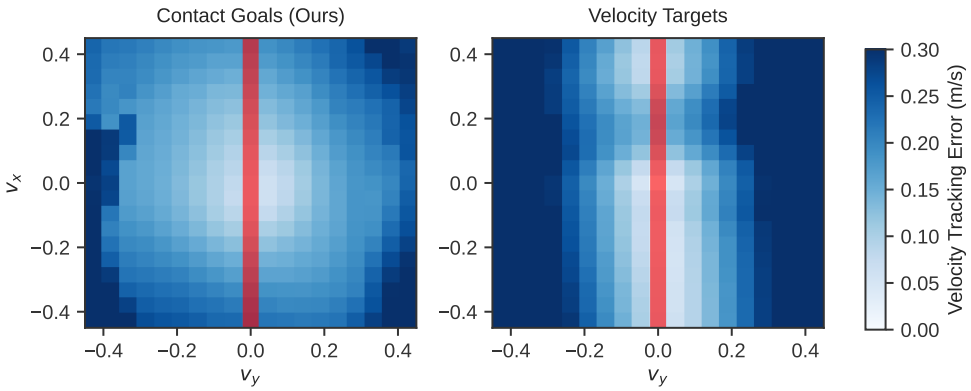

Figure 5: Comparison of velocity tracking error in all x-y directions. Each cell in the grid is a combination of x and y velocity. The red line denotes the velocity combinations seen during training. The results were averaged over 500 episodes (each lasting 15 seconds of simulation time).

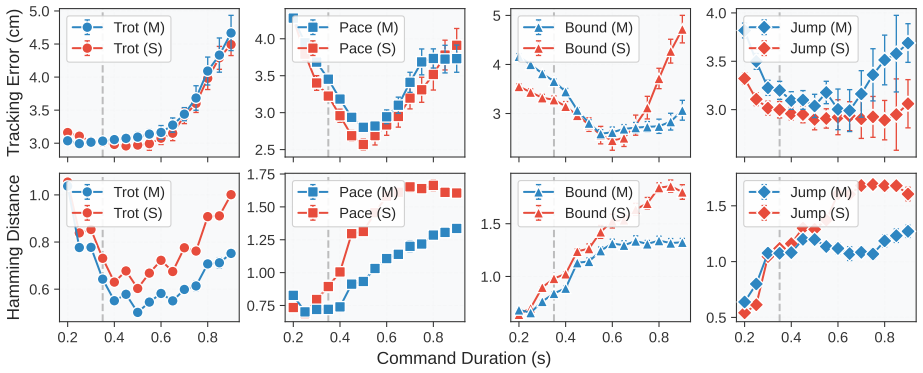

Figure 6: Comparison of contact location tracking ($L2$-norm) and contact plan deviation (Hamming distance) for the multi-gait policy (denoted by M) against single-gait policies (denoted by S) trained exclusively on one gait, evaluated over a broader range of command durations. Results are averaged over 1000 episodes, each lasting 15 seconds of simulation time. The dotted vertical grey line indicates the command duration seen during training.

In the future, we'd also like to explore off-policy RL algorithms to make use of goal relabelling Andrychowicz et al. (2018).

Locomotion generalization to unseen velocities/directions. To demonstrate that contact-explicit representations truly generalize better to out-of-distribution scenarios and has better representation capabilities, we compare two policies with different goal representations, contact-explicit (ours) trained to only perform trotting gait against velocity targets (typical in learning quadrupedal locomotion as in Hwangbo et al. (2019); Rudin et al. (2022b) and known to converge to a trotting gait), each trained only to move forward/backwards up to a maximum speed of 0.65 m/s (i.e.,

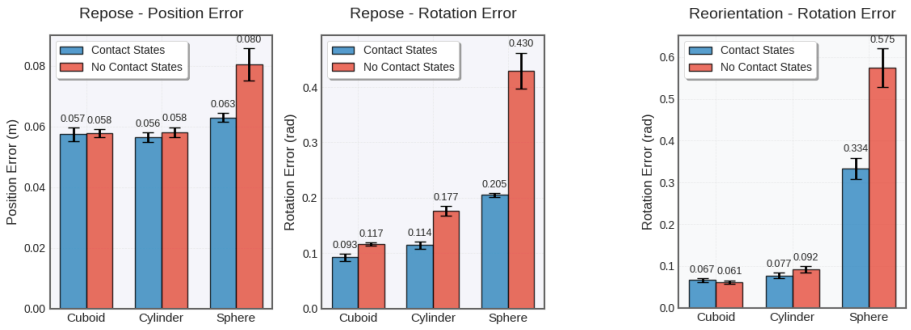

Figure 7: Tracking error comparison on unseen object shapes across various tasks. We compare our contact-explicit policy (uses contact states) with a baseline policy that uses one-hot task encoding (no contact states) to represent the task. Results are averaged over 1000 episodes, each lasting 20 seconds of simulated time.

$v_x \in [-0.65, 0.65]$ m/s, $v_y = 0$). We evaluate the velocity tracking error of these two policies when commanded to move along all directions as shown in Fig. 5.

Although the policy trained with velocity targets has slightly less tracking error while extrapolating velocity targets along the trained direction, it becomes apparent that the contact-explicit policy can cover a much broader range of velocities, compared to the one trained with velocity targets. Especially in the case of lateral/sideways walking ($v_x = 0$), the velocity-conditioned policy hesitates to move. In contrast, the contact-explicit policy moves sideways, even though it hadn't received any reward for the lateral movements during training, but was solely trained to track the contact goals.

Multi-task versus single-task. Contact-explicit task representation enables learning multiple gaits in a single policy, which helps leverage the shared structure between different gaits to interpolate between them (even though it has not seen those states during training). To test our claims, we compare our multi-gait quadrupedal locomotion policy against separate policies trained on single gaits, as shown in Fig. 6. The policies were trained with command durations sampled from a narrow uniform distribution of range $[0.34s, 0.36s]$ and evaluated over a broader range of $[0.2s, 0.9s]$. We use two metrics: contact location tracking error measured using $L2$-norm between the actual end-effector locations and the planned contact locations while making contact, and the contact plan deviation measured using the Hamming distance between the desired contact plan and the actual contact status of the end-effectors.

From Fig. 6, we observe that the multi-gait (M) policy has the lowest contact plan deviation across all the evaluated command durations and for all gaits. We hypothesize that this is mainly due to our contact-explicit representation that enables learning multiple gaits in a single policy. We also observe that generally, the tracking error of both the single-gait and multi-gait policies are similar, i.e. within 2 cm of difference.

Manipulation generalization to unseen object shapes. We compare our bimanual manipulation policy against a policy that instead uses one-hot task encoding to distinguish the tasks. Here, we evaluate the performance of the two policies, one with contact goals in the state (ours, blue) and another that uses a one-hot task encoding (baseline, red) instead to distinguish the tasks. For the

Task / Repose	**Contact-Explicit**	**One-hot task**
Position Error (m)	**0.115** $\pm$ 0.003	0.129 $\pm$ 0.003
Rotation Error (rad)	**0.390** $\pm$ 0.006	0.455 $\pm$ 0.007
Task / Reorientation	**Contact-Explicit**	**One-hot task**
Rotation Error (rad)	**0.109** $\pm$ 0.002	0.191 $\pm$ 0.004

Table 1: Quantitative evaluation of our contact-explicit policy against a one-hot task encoding baseline on out-of-distribution object poses. Object poses were sampled from $\mathcal{U}_{\text{train}}$ / $\mathcal{U}_{\text{eval}}$, where the first range corresponds to training and the second to evaluation: $X \sim \mathcal{U}_{\text{train}}(0.0, 0.1)$ / $\mathcal{U}_{\text{eval}}(0.0, 0.2)$, $Y \sim \mathcal{U}_{\text{train}}(-0.15, 0.15)$ / $\mathcal{U}_{\text{eval}}(-0.3, 0.3)$, $Z \sim \mathcal{U}_{\text{train}}(0.05, 0.25)$ / $\mathcal{U}_{\text{eval}}(0.05, 0.4)$, and roll, pitch, yaw $\sim \mathcal{U}_{\text{train}}(-0.6, 0.6)$ / $\mathcal{U}_{\text{eval}}(-1.2, 1.2)$.

baseline, we additionally provide the object dimensions as an observation to the policy. The policy was only trained on cuboidal shapes and evaluated here with cylindrical and spherical shapes. Our results are summarised in Fig. 7.

As with the contact-explicit locomotion policy being able to generalize to unseen contact locations (velocity directions), we witness with the contact-explicit bimanual manipulation policy that we can better generalize to unseen object shapes. Using the contact-explicit approach, the policy learns to track the contact locations on the different shapes much better than other implicit goal representations. Especially in the case of spherical shapes for object reorientation on the table, we observe that our policy comes up with emergent retrying behavior to track the object poses as the object starts rolling on the table.

Manipulation generalization to unseen object poses. We compare the two bimanual manipulation policies mentioned in the previous experiment, contact-explicit (ours) and one-hot task encoding (baseline), to track object poses that were outside the training distribution. This was done by extrapolating the range of object poses from those seen in the training distribution. The results, summarized in Table 1, demonstrate that contact-explicit policy consistently outperforms the one-hot task policy, with lower variability indicating more stable task accomplishment. This underscores the value of contact-aware policies for robust bimanual manipulation in uncertain scenarios.

5. Conclusion

We presented a unified contact-explicit task representation for learning a wide range of locomotion and manipulation skills through reinforcement learning. By treating contact as a central physical primitive, rather than a byproduct of motion optimization, our approach enables a single policy to generalize across morphologically diverse platforms and tasks. Empirical results demonstrate that contact-explicit policies offer stronger generalization to out-of-distribution goal configurations.

Moving forward, we aim to extend this framework to hierarchical reinforcement learning by coupling our low-level contact-conditioned policy with a learned high-level planner. This would allow autonomous long-horizon loco-manipulation in complex environments. We also plan to explore prehensile interactions and improve sim-to-real robustness for real-world deployment.

Acknowledgments

This work was partially supported by the Huawei-TUM joint laboratory.

References

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay, 2018. URL https://arxiv.org/abs/1707.01495.

Guillaume Bellegarda, Milad Shafiee, and Auke Ijspeert. Allgaits: Learning all quadruped gaits and transitions. *arXiv preprint arXiv:2411.04787*, 2024.

Yuanpei Chen, Tianhao Wu, Shengjie Wang, Xidong Feng, Jiechuan Jiang, Zongqing Lu, Stephen McAleer, Hao Dong, Song-Chun Zhu, and Yaodong Yang. Towards human-level bimanual dexterous manipulation with reinforcement learning. *Advances in Neural Information Processing Systems*, 35:5150–5163, 2022.

Xuxin Cheng, Kexin Shi, Ananye Agarwal, and Deepak Pathak. Extreme parkour with legged robots. *arXiv preprint arXiv:2309.14341*, 2023.

Michal Ciebielski and Majid Khadiv. Contact-conditioned learning of locomotion policies. *arXiv preprint arXiv:2408.00776*, 2024.

Jeremy Dao, Helei Duan, and Alan Fern. Sim-to-real learning for humanoid box loco-manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 16930–16936. IEEE, 2024.

Victor Dhedin, Adithya Kumar Chinnakkonda Ravi, Armand Jordana, Huaijiang Zhu, Avadesh Meduri, Ludovic Righetti, Bernhard Schölkopf, Majid Khadiv, et al. Diffusion-based learning of contact plans for agile locomotion. In *2024 IEEE-RAS 23rd International Conference on Humanoid Robots (Humanoids)*, pages 637–644. IEEE, 2024.

Zipeng Fu, Xuxin Cheng, and Deepak Pathak. Deep whole-body control: Learning a unified policy for manipulation and locomotion, 2022. URL https://arxiv.org/abs/2210.10044.

Tairan He, Wenli Xiao, Toru Lin, Zhengyi Luo, Zhenjia Xu, Zhenyu Jiang, Jan Kautz, Changliu Liu, Guanya Shi, Xiaolong Wang, et al. Hover: Versatile neural whole-body controller for humanoid robots. *arXiv preprint arXiv:2410.21229*, 2024.

David Hoeller, Nikita Rudin, Dhionis Sako, and Marco Hutter. Anymal parkour: Learning agile navigation for quadrupedal robots, 2023. URL https://arxiv.org/abs/2306.14874.

Jemin Hwangbo, Joonho Lee, Alexey Dosovitskiy, Dario Bellicoso, Vassilios Tsounis, Vladlen Koltun, and Marco Hutter. Learning agile and dynamic motor skills for legged robots. *Science Robotics*, 4(26):eaau5872, 2019.

Zhongyu Li, Xue Bin Peng, Pieter Abbeel, Sergey Levine, Glen Berseth, and Koushil Sreenath. Reinforcement learning for versatile, dynamic, and robust bipedal locomotion control. *The International Journal of Robotics Research*, page 02783649241285161, 2024.

Toru Lin, Kartik Sachdev, Linxi Fan, Jitendra Malik, and Yuke Zhu. Sim-to-real reinforcement learning for vision-based dexterous manipulation on humanoids. *arXiv:2502.20396*, 2025.

Yijiong Lin, Alex Church, Max Yang, Haoran Li, John Lloyd, Dandan Zhang, and Nathan F Lepora. Bi-touch: Bimanual tactile manipulation with sim-to-real deep reinforcement learning. *IEEE Robotics and Automation Letters*, 8(9):5472–5479, 2023.

Minghuan Liu, Zixuan Chen, Xuxin Cheng, Yandong Ji, Ri-Zhao Qiu, Ruihan Yang, and Xiaolong Wang. Visual whole-body control for legged loco-manipulation. *arXiv preprint arXiv:2403.16967*, 2024.

Tyler Ga Wei Lum, Martin Matak, Viktor Makoviychuk, Ankur Handa, Arthur Allshire, Tucker Hermans, Nathan D. Ratliff, and Karl Van Wyk. Dextrah-g: Pixels-to-action dexterous arm-hand grasping with geometric fabrics, 2024. URL https://arxiv.org/abs/2407.02274.

Gabriel B Margolis and Pulkit Agrawal. Walk these ways: Tuning robot control for generalization with multiplicity of behavior. *Conference on Robot Learning*, 2022.

Takahiro Miki, Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. Learning robust perceptive locomotion for quadrupedal robots in the wild. *Science Robotics*, 7(62):eabk2822, 2022. doi: 10.1126/scirobotics.abk2822. URL https://www.science.org/doi/abs/10.1126/scirobotics.abk2822.

Mayank Mittal, Calvin Yu, Qinxi Yu, Jingzhou Liu, Nikita Rudin, David Hoeller, Jia Lin Yuan, Ritvik Singh, Yunrong Guo, Hammad Mazhar, Ajay Mandlekar, Buck Babich, Gavriel State, Marco Hutter, and Animesh Garg. Orbit: A unified simulation framework for interactive robot learning environments. *IEEE Robotics and Automation Letters*, 8(6):3740–3747, 2023. doi: 10.1109/LRA.2023.3270034.

Shafeef Omar, Lorenzo Amatucci, Giulio Turrisi, Victor. Barasuol, and Claudio Semini. Safesteps: Learning safer footstep planning policies for legged robots via model-based priors. In *IEEE-RAS International Conference on Humanoid Robots*, 2023.

OpenAI, Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, Jonas Schneider, Nikolas Tezak, Jerry Tworek, Peter Welinder, Lilian Weng, Qiming Yuan, Wojciech Zaremba, and Lei Zhang. Solving rubik's cube with a robot hand, 2019. URL https://arxiv.org/abs/1910.07113.

Guoping Pan, Qingwei Ben, Zhecheng Yuan, Guangqi Jiang, Yandong Ji, Shoujie Li, Jiangmiao Pang, Houde Liu, and Huazhe Xu. Roboduet: Learning a cooperative policy for whole-body legged loco-manipulation. *IEEE Robotics and Automation Letters*, 2025.

Xue Bin Peng, Glen Berseth, Kangkang Yin, and Michiel Van De Panne. Deeploco: Dynamic locomotion skills using hierarchical deep reinforcement learning. *ACM Trans. Graph.*, 36(4): 41:1–41:13, July 2017. ISSN 0730-0301. doi: 10.1145/3072959.3073602. URL http://doi.acm.org/10.1145/3072959.3073602.

Ri-Zhao Qiu, Yuchen Song, Xuanbin Peng, Sai Aneesh Suryadevara, Ge Yang, Minghuan Liu, Mazeyu Ji, Chengzhe Jia, Ruihan Yang, Xueyan Zou, et al. Wildlma: Long horizon loco-manipulation in the wild. *arXiv preprint arXiv:2411.15131*, 2024.

Nikita Rudin, David Hoeller, Marko Bjelonic, and Marco Hutter. Advanced skills by learning locomotion and local navigation end-to-end, 2022a. URL https://arxiv.org/abs/2209.12827.

Nikita Rudin, David Hoeller, Philipp Reist, and Marco Hutter. Learning to walk in minutes using massively parallel deep reinforcement learning, 2022b. URL https://arxiv.org/abs/2109.11978.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Carmelo Sferrazza, Dun-Ming Huang, Xingyu Lin, Youngwoon Lee, and Pieter Abbeel. Humanoidbench: Simulated humanoid benchmark for whole-body locomotion and manipulation, 2024.

Ritvik Singh, Arthur Allshire, Ankur Handa, Nathan Ratliff, and Karl Van Wyk. Dextrah-rgb: Visuomotor policies to grasp anything with dexterous hands, 2025. URL https://arxiv.org/abs/2412.01791.

Jean-Pierre Sleiman, Mayank Mittal, and Marco Hutter. Guided reinforcement learning for robust multi-contact loco-manipulation. In *8th Annual Conference on Robot Learning (CoRL 2024)*, 2024.

Jie Tan, Tingnan Zhang, Erwin Coumans, Atil Iscen, Yunfei Bai, Danijar Hafner, Steven Bohez, and Vincent Vanhoucke. Sim-to-real: Learning agile locomotion for quadruped robots. *arXiv preprint arXiv:1804.10332*, 2018.

Ilyass Taouil, Haizhou Zhao, Angela Dai, and Majid Khadiv. Physically consistent humanoid loco-manipulation using latent diffusion models, 2025. URL https://arxiv.org/abs/2504.16843.

Zhao-Heng Yin, Changhao Wang, Luis Pineda, Francois Hogan, Krishna Bodduluri, Akash Sharma, Patrick Lancaster, Ishita Prasad, Mrinal Kalakrishnan, Jitendra Malik, Mike Lambeta, Tingfan Wu, Pieter Abbeel, and Mustafa Mukadam. Dexteritygen: Foundation controller for unprecedented dexterity, 2025. URL https://arxiv.org/abs/2502.04307.

Fatemeh Zargarbashi, Jin Cheng, Dongho Kang, Robert Sumner, and Stelian Coros. Robotkeyframing: Learning locomotion with high-level objectives via mixture of dense and sparse rewards, 2024. URL https://arxiv.org/abs/2407.11562.

Chong Zhang, Nikita Rudin, David Hoeller, and Marco Hutter. Learning agile locomotion on risky terrains, 2024a. URL https://arxiv.org/abs/2311.10484.

Chong Zhang, Wenli Xiao, Tairan He, and Guanya Shi. Wococo: Learning whole-body humanoid control with sequential contacts, 2024b.