New Explorations in Electrical Engineering (NECE)
Department of Electrical Engineering,
Faculty of Engineering
Universiti Malaya, MALAYSIA

Vol. 2, No. 1 (2026)
pp. 31–40

Reinforcement Learning-Controlled Quadruped Robot: From Basic Gaits to Complex Motions

Lin Yashan[a], Mahidzal Bin Dahari[a,*]
[a] *Department of Electrical Engineering, Faculty of Engineering, Universiti Malaya, 50603 WPKL, Kuala Lumpur, Malaysia*
*E-mail: lithos0515@gmail.com, *mahidzal@um.edu.my*

Abstract. Quadrupedal locomotion control via deep reinforcement learning (RL) has demonstrated strong performance on symmetric gaits, yet its ability to learn non-conventional motions requiring asymmetric, temporally precise coordination remains underexplored. The objective of this study is to examine how increasing locomotion complexity affects reward design, learning dynamics, and zero-shot sim-to-sim robustness. This issue is investigated by progressively training three behaviours of increasing complexity, trotting, bounding, and a lateral alternating jumping, using Proximal Policy Optimization (PPO)-based training with curriculum learning and reference-state initialisation introduced for the more complex behaviours. Policies are evaluated via zero-shot sim-to-sim transfer from Isaac Lab to MuJoCo to assess robustness under differing physics solvers and contact models. The three policies converged after approximately 1k, 2k, and 6.5k iterations, with task-specific final mean rewards of 29, 85, and 624, confirming the need for more structured rewards as locomotion complexity increases.

Keywords: Adversarial motion priors, Locomotion complexity, Quadruped robot, Reinforcement learning, Sim-to-sim transfer.

1 Introduction

Reinforcement learning has become an effective approach for quadrupedal locomotion because it reduces reliance on manually designed foot trajectories, fixed contact schedules, and hand-crafted motion primitives. Parallel simulation, policy optimisation, and domain randomisation now enable efficient training of walking and running policies. However, most successful cases still focus on relatively regular motions, whose coordination patterns align well with common rewards such as velocity tracking, posture stabilisation, torque penalties, and action smoothness.

Non-conventional motions introduce a different learning problem. Bounding requires synchronised front-hind leg motion and often an aerial phase, while lateral alternating jumping requires strict left-right alternation, lateral centre-of-mass transfer, and stable landing. These behaviours are not merely harder versions of trotting; they require different contact structures and training signals. Rewards that naturally induce trotting may therefore lead to incorrect local optima in bounding or lateral jumping.

This paper studies this issue through a progressive-complexity sequence: trotting as a symmetric baseline, bounding as an intermediate front-back asymmetric gait, and lateral alternating jumping as a highly dynamic non-conventional motion. Rather than proposing a single controller for all behaviours, the study compares how reward design, learning dynamics, and sim-to-sim validation change as locomotion complexity increases.

NECE. All rights reserved.

The main contributions of this work are a progressive-complexity analysis of RL-based quadrupedal locomotion across trotting, bounding, and lateral alternating jumping; reward-shaping insights for different coordination demands, including contact-phase guidance, lateral centre-of-mass transfer, landing stability, and curriculum support; and a sim-to-sim evaluation perspective for analysing robustness under different physical contact models.

The overall research workflow is summarized in Fig. 1, where the three target motions are arranged from symmetric trotting to bounding and lateral alternating jumping, followed by zero-shot sim-to-sim validation from Isaac Lab to MuJoCo.

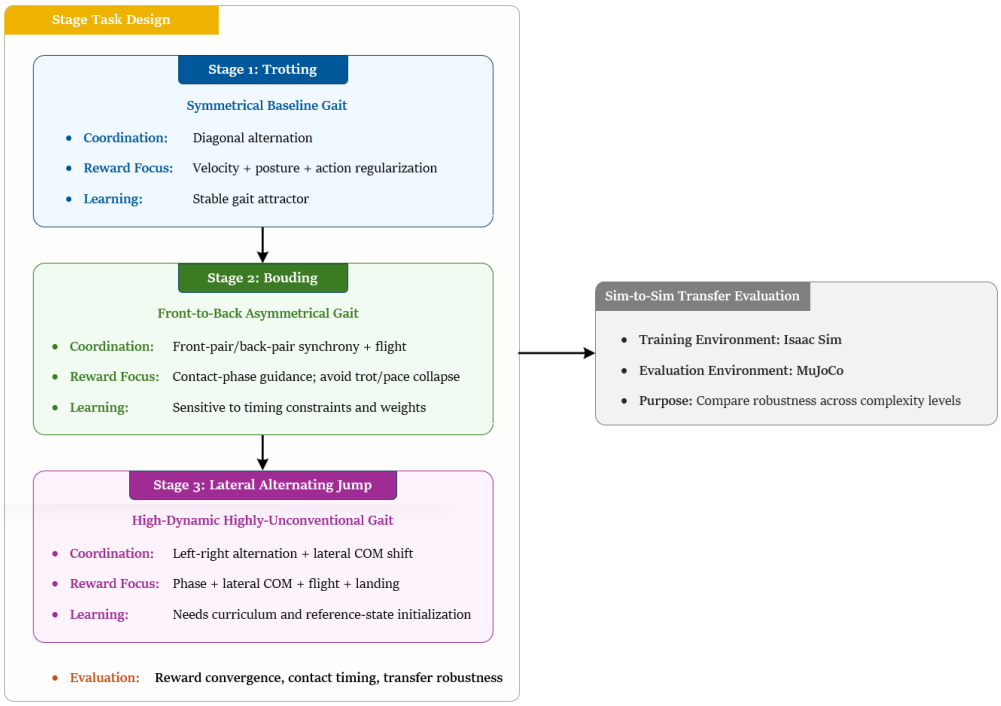


Figure 1. Overview of the progressive locomotion-complexity analysis and sim-to-sim validation workflow.

2 Background

The locomotion-control tasks are modelled as Markov decision processes. Each policy observes compact proprioceptive information, including body velocity, projected gravity, joint-position deviations, joint velocities, and previous actions, and outputs continuous joint-position offsets tracked by a low-level PD controller. Proximal Policy Optimization (PPO) is used for large-scale on-policy training. The platform is the Unitree Go2, a twelve-actuator quadruped with hip roll, hip pitch, and knee pitch on each leg. Isaac Lab and unitree_rl_lab provide the main training basis, while MuJoCo is used for sim-to-sim transfer. The bounding and lateral alternating jumping tasks extend the original codebase with custom rewards, observation terms, and gait-clock modules.

3 Related Work

Reinforcement-learning-based quadrupedal control learns policies from interaction data instead of fully hand-designed gait timing. Simulation-trained neural policies can produce agile legged locomotion when actuator behaviour and dynamics mismatch are considered. Later work extended this direction to rough terrain and perceptive locomotion, showing that proprioceptive and exteroceptive information can support stable motion in uncertain environments. Adaptation methods such as Rapid Motor Adaptation further separate policy learning from online environment-parameter estimation, enabling robots to handle unseen terrain and dynamics variations. High-speed locomotion studies show that command curricula and stability rewards can produce faster dynamic behaviours. Large-scale parallel training also shifts the practical bottleneck from data collection to task modelling, reward shaping, and curriculum design.

Dynamic and non-conventional behaviours make this bottleneck more visible. Trotting is often compatible with dense velocity and stability rewards, whereas bounding introduces front-back asymmetry and aerial phases. Prior work on high-speed running and bounding shows that action representation, contact timing, and reward formulation strongly affect the gait that emerges during training. More agile behaviours such as jumping, climbing, and parkour further narrow the effective exploration region. Hierarchical reinforcement learning and skill decomposition have been used for complex agile navigation, while curriculum-based quadrupedal jumping shows that gradually increasing task difficulty can prevent de-generate or incomplete solutions.

Adversarial Motion Priors (AMP) provide another mechanism for guiding complex locomotion. Rather than tracking a reference trajectory frame-by-frame, AMP uses a learned discriminator to encourage the policy's state transitions to resemble those in a reference dataset, enabling style transfer without rigid phase locking. This approach has been applied to character animation and quadrupedal locomotion, where it can encode stylistic preferences while leaving room for task-level optimisation.

Simulation transfer is another central issue because policies may degrade under mismatched dynamics, actuation, sensing, and contact modelling. Domain randomisation improves generalisation, but its effectiveness depends on task structure and randomisation range. For highly dynamic motions, Isaac Lab to MuJoCo transfer provides an intermediate robustness test before hardware deployment.

4 Proposed Method

4.1 *Method Overview*

This section describes how the locomotion tasks are organized and compared. Rather than building a multi-task controller that switches among all behaviors, the three motions are arranged as a staged complexity sequence. This organization allows the effect of increasing coordination difficulty to be examined through reward design, contact timing, exploration difficulty, and sim-to-sim robustness.

The three stages are implemented as follows. Stage 1 is flat-ground trotting, which serves as a symmetric baseline gait. Stage 2 is bounding, a periodic synchronous jumping motion that introduces explicit support-flight timing. Stage 3 is lateral alternating jumping, a highly dynamic non-conventional motion that requires ipsilateral left-right alternation, lateral center-of-mass transfer, and stable landing recovery. The methodological question is how training signals change as the target behavior moves from continuous-contact locomotion through front-back symmetric jumping to fully asymmetric lateral alternation.

The expected foot-contact schedules of the three staged tasks are shown in Fig. 2, while their main design differences are summarized in Table 1. Together, these comparisons clarify how the tasks progress from an emergent diagonal gait to explicit support-flight timing and finally to ipsilateral left-right alternation.

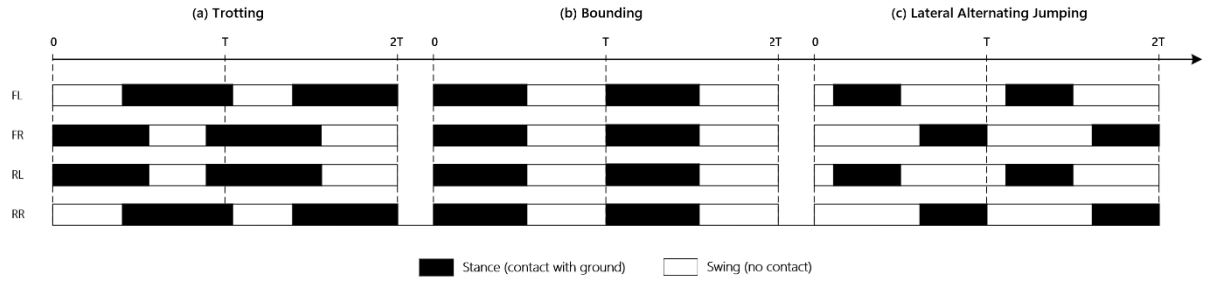


Figure 2. Expected foot-contact schedules for the staged locomotion tasks. Black and white bars denote stance and swing phases, respectively, for (a) trotting, (b) bounding, and (c) lateral alternating jumping over two gait cycles.

Table 1. Design comparison of the staged locomotion tasks

Design Aspect	Trotting	Bounding	Lateral Alternating Jumping
Target behavior	Stable diagonal gait	Synchronous support-flight jump	Strict left-right alternating jump with lateral CoM shift
Policy observation	Commands, joint states, action history	Trot observations plus contact state and gait clock	Trot observations plus contact state, gait clock; base vertical velocity excluded
Structural prior	No explicit gait phase	Explicit phase and contact targets	Explicit phase alternation, RSI, and phase joint tracking
Reward focus	Velocity tracking, stability, smooth swing	Timing, clearance, stable landing	Alternating contact, lateral lean, swing clearance, landing stability
Termination condition	Pitch/roll > 0.5 rad, height < 0.15 m	Pitch/roll > 0.8 rad, height < 0.1 m	Pitch/roll > 1.0 rad, height < 0.1 m

4.2 *Trotting*

The trotting task is a flat-ground locomotion problem driven by forward-velocity, lateral-velocity, and yaw-rate commands. The command curriculum samples forward velocity within -1.0 to 1.0 m/s, lateral velocity within -0.4 to 0.4 m/s, and yaw velocity within -1.0 to 1.0 rad/s. The policy receives command and joint observations but no explicit gait phase, so diagonal coordination emerges from dense rewards for velocity tracking, posture stability, smooth motion, foot air time, and slip reduction. Trotting therefore serves as the baseline for evaluating how far stable gait structure can emerge with limited structural priors.

4.3 *Bounding*

The bounding task is a periodic synchronous jumping problem that requires coordinating support, lift-off, flight, and landing according to a target temporal structure. A 0.6 s gait clock defines support in the first half of the cycle and flight in the second half. The observation space adds foot-contact state and periodic phase to the trotting formulation, enabling the policy to align actions with the intended contact state. The reward retains stability and command-response terms while adding phase-related incentives for support contact, flight-phase lift-off, clearance, body height, and suppression of walking-like solutions.

4.4 *Lateral Alternating Jumping*

Lateral alternating jumping is the most complex motion in the sequence and is driven by forward-velocity, lateral-velocity, and yaw-rate commands. A 1.0 s gait clock alternates between left-pair (FL+RL) and right-pair (FR+RR) support phases with short aerial transitions between them.

Single-stage training proved unreliable, as the policy consistently collapsed to diagonal trotting or shuffling, so a two-step procedure is adopted:

Lateral base policy (Step 1): Trains ipsilateral left-right alternation using an observation space extended with contact state, periodic phase, and base height. RSI initialises each episode at a random trajectory point to ensure full phase coverage. Rewards reinforce ipsilateral contact, penalise diagonal shortcuts, encourage stance-side lateral lean, and promote swing-side clearance. The result is reliable left-right alternation, though aerial phases may remain insufficient.

Lateral-AMP jump policy (Step 2): Refines stepping into genuine jumping via AMP. A discriminator trained on the base policy's best rollouts provides a style reward matching the reference's kinematic and temporal profile without frame-by-frame tracking. Task rewards additionally target airborne transitions, landing stability, and lateral CoM transfer. Vertical base velocity is excluded from observations to prevent aerial suppression.

4.5 *Training and Evaluation Protocol*

All three stages follow the same PPO-based training and evaluation pipeline. The Unitree Go2 model, joint-position action space, Isaac Lab training environment, policy network family, and low-level PD control interface are kept consistent so that differences in learning behaviour can be attributed mainly to stage-specific reward and observation design. Command curricula and environment randomisation are applied across the stages to improve robustness.

Evaluation is organised around both common and stage-specific criteria. Mean reward convergence is used as a common training diagnostic, while contact-pattern analysis is adapted to each target motion: diagonal alternation for trotting, support-flight synchronisation for bounding, and left-right ipsilateral alternation with lateral body shift for lateral alternating jumping. After training, each policy is transferred from Isaac Lab to MuJoCo without retraining. This zero-shot sim-to-sim validation provides a common robustness check under a different physics solver and contact model.

The shared PPO settings and task-specific training configurations used for the three stages are summarized in Table 2.

Table 2. Training configuration for all three stages

Parameter	Trotting	Bounding	Lateral Alternating Jumping
Parallel environments	2048	2048	2048
Policy network	MLP [512, 256, 128]	MLP [512, 256, 128]	MLP [512, 256, 128]
Learning rate	1e-3	1e-3	1e-3
Discount factor γ	0.99	0.99	0.99
GAE λ	0.95	0.95	0.95
PPO clip ratio ε	0.2	0.2	0.2
Training iterations	5000	5000	4200 (Step 1) + 2800 (Step 2)
PD gains (Kp, Kd)	(20, 0.5)	(20, 0.5)	(20, 0.5)
Gait period (s)	N/A (emergent)	0.6	1.0

5 Experimental Validation

The validation focuses on whether each stage produces its intended contact structure and whether the learned behaviour transfers from Isaac Lab to MuJoCo without retraining. Mean reward and loss curves are used as training diagnostics, while foot-contact patterns and MuJoCo rollout decomposition are used as the primary behavioural evidence.

5.1 *Trotting Results and Validation*

As shown in Fig. 3(a), the trotting policy learns stably, with the mean reward converging after approximately 1k iterations and reaching a final task-specific mean reward of 29, as summarized in Table 3. The value-function loss remains low after the early training period, as shown in Fig. 3(b). The contact diagram in Fig. 3(c) shows alternating FL+RR and FR+RL support, confirming a diagonal trot rather than walking or dragging. The rollout snapshots in Fig. 3(d) further show that the Isaac Lab and MuJoCo rollouts preserve the same lift-contact-lift sequence.

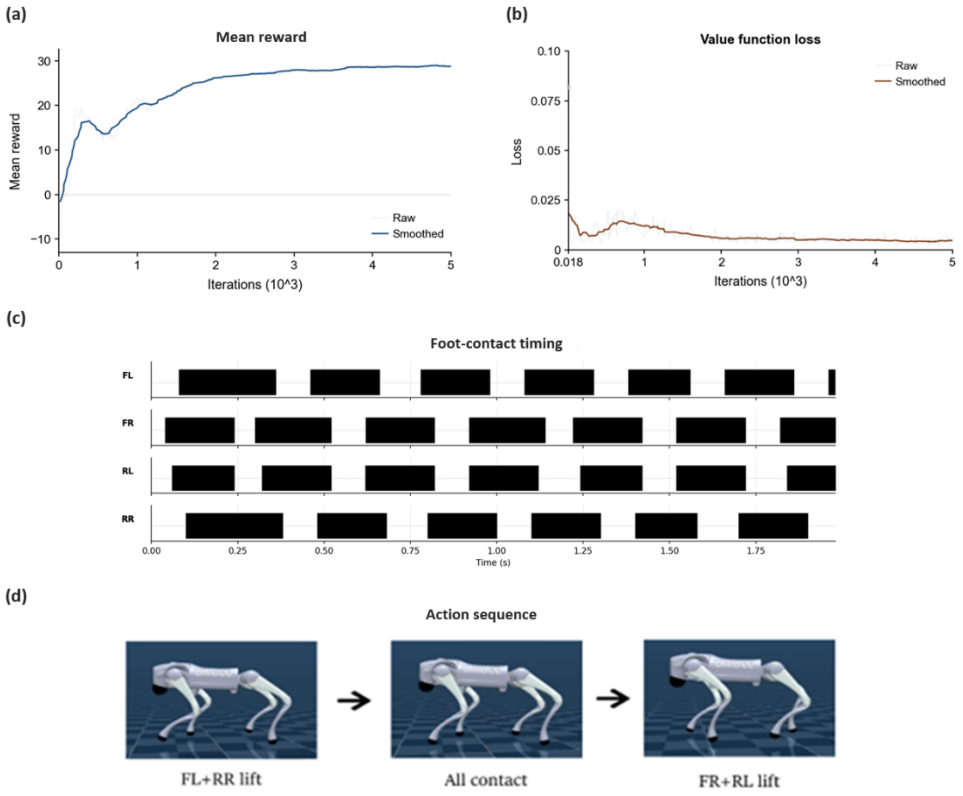


Figure 3. Trotting results and validation. (a) Mean reward convergence during policy training. (b) Value-function loss during training. (c) Foot-contact timing diagram showing alternating FL+RR and FR+RL diagonal support. (d) Rollout snapshots showing the learned trotting action sequence: FL+RR lift, all-foot contact, and FR+RL lift.

5.2 *Bounding Results and Validation*

As shown in Fig. 4, the bounding policy validates the need for explicit phase-contact guidance. The mean reward in Fig. 4(a) converges after approximately 2k iterations and reaches a final task-specific

mean reward of 85, while the value-function loss in Fig. 4(b) stabilizes as the policy learns the 0.6 s support-flight clock. The contact diagram in Fig. 4(c) shows synchronized four-foot contact and clear no-contact intervals, and the rollout snapshots in Fig. 4(d) illustrate push-off, landing, retraction, and the next push-off. These results confirm that explicit phase-contact guidance is required for stable bounding.

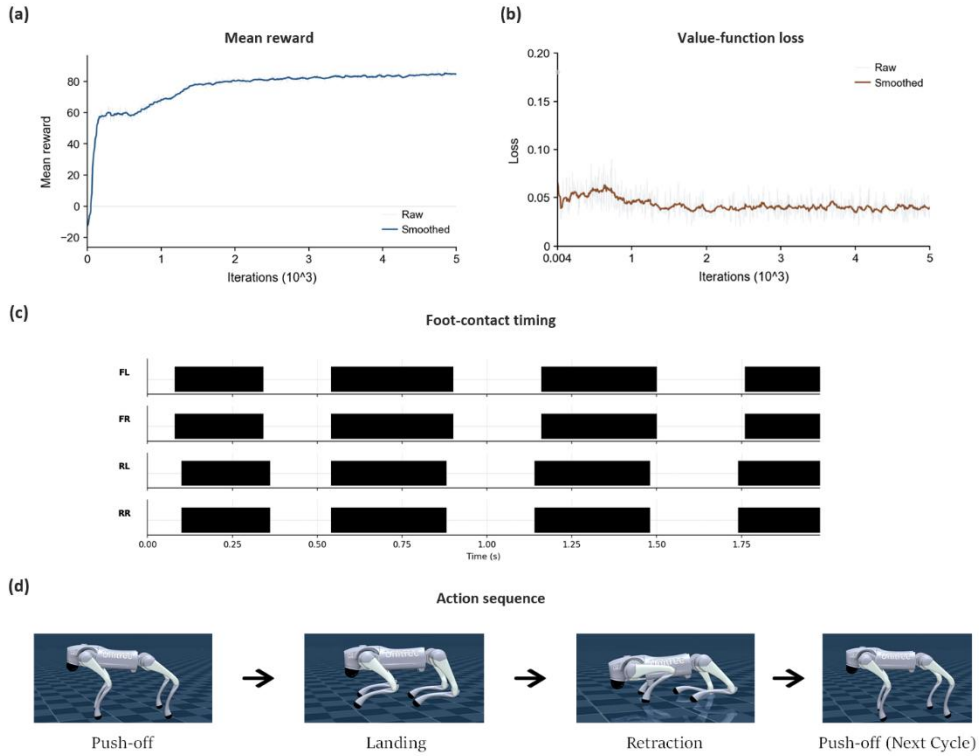


Figure 4. Bounding results and validation. (a) Mean reward convergence and (b) value-function loss during training. (c) Foot-contact timing showing synchronised four-foot support and no-contact intervals, confirming the learned support-flight structure. (d) Rollout snapshots illustrating the bounding cycle from push-off to landing, retraction, and the next push-off.

5.3 *Lateral Alternating Jumping Results and Validation*

As shown in Fig. 5, the lateral alternating jumping policy produces the intended ipsilateral alternation through two training steps. The mean reward in Fig. 5(a) converges after approximately 2.5k iterations for the lateral base policy and 4k iterations for the lateral-AMP jump policy, reaching a final task-specific mean reward of 624. The contact diagram in Fig. 5(c) confirms ipsilateral FL+RL / FR+RR alternation rather than diagonal trotting or shuffling, while the rollout snapshots in Fig. 5(d) show the corresponding FR+RR lift, all-swing transition, all-contact recovery, and FL+RL lift.

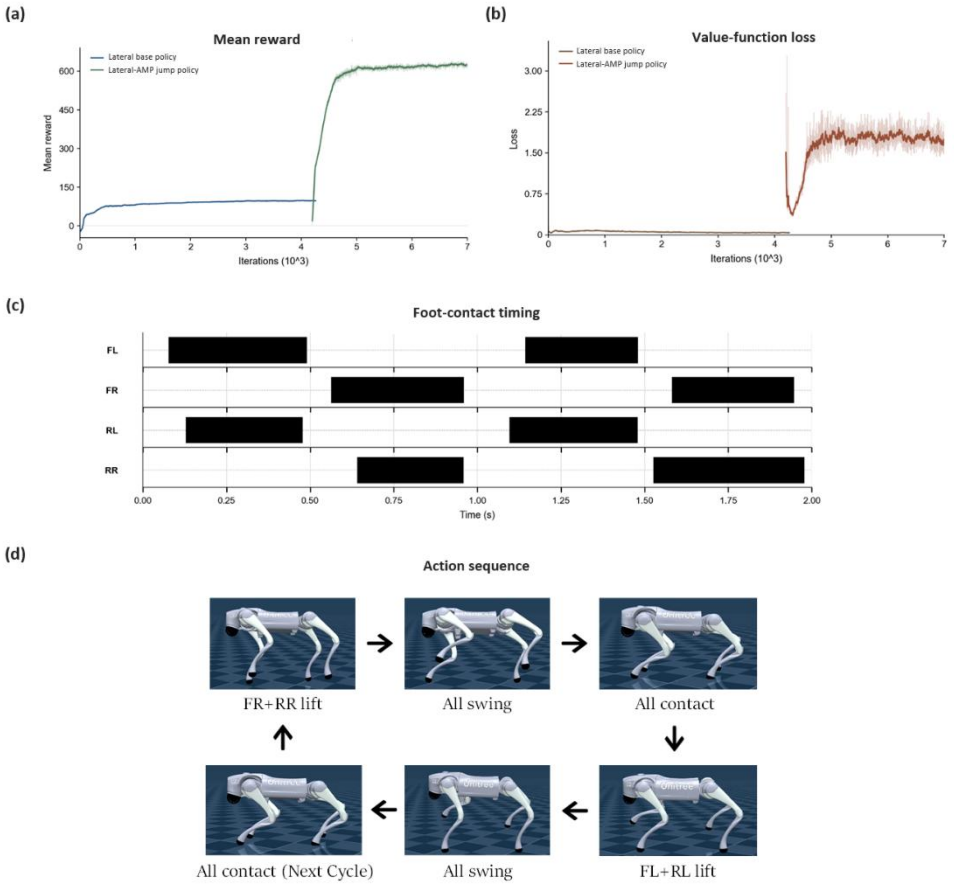


Figure 5. Lateral alternating jumping results. (a) Mean reward convergence across two training steps (Step 1: lateral base policy; Step 2: lateral-AMP jump policy). (b) Value-function loss over the same training process. (c) Foot-contact timing diagram showing ipsilateral FL+RL / FR+RR alternation. (d) Rollout snapshots of one full jump cycle: FR+RR lift, all-swing, all-contact recovery, and FL+RL lift.

Reward values are not directly comparable across tasks because each task uses a different reward formulation.

Table 3. Quantitative comparison of the three locomotion policies

Metric	Trotting	Bounding	Lateral Alternating Jumping
Final mean reward	29	85	624
Convergence iteration ($\times 10^3$)	1	2	2.5 (Step 1) + 4 (Step 2)
Mean episode length (steps)	1184	1146	1162 (Step 1) + 1161 (Step 2)

6 Discussion

Increasing locomotion complexity changes the nature of the learning problem, not merely its difficulty. Trotting emerges naturally from standard velocity-tracking and regularisation rewards; bounding requires explicit phase-contact guidance, otherwise the policy degrades to simpler behaviours; lateral alternating

jumping is most constrained, requiring simultaneous avoidance of diagonal shortcuts, ipsilateral coordination, and aerial transitions.

Reward ceilings primarily reflect differences in reward formulation, not task success. The more meaningful signal is initial reward behaviour: trotting starts from a comparatively high baseline, bounding begins slightly negative (support-flight clock must be satisfied), and lateral alternating jumping starts deepest (multiple constraints active simultaneously). This supports the central argument: reward specificity and anti-shortcut design must scale with locomotion complexity.

The lateral alternating jump further shows why a single-stage reward is insufficient: the policy requires both a style prior for left-right alternation and explicit incentives for aerial motion.

Qualitative gait patterns transfer under a different contact solver, suggesting the learned behaviours are not simulator-specific contact artefacts. However, this does not imply hardware readiness, as real deployment still requires actuator-delay handling, impact safety, friction calibration, and sim-to-real validation, particularly for the lateral alternating jump where asymmetric landing and timing errors are most critical.

7 Conclusion

This work presented a progressive-complexity analysis of RL-based quadrupedal locomotion on the Unitree Go2, covering trotting, bounding, and lateral alternating jumping. The results show that increasing locomotion complexity requires progressively more structured reward design, contact guidance, and staged training, while qualitative gait patterns can still transfer from Isaac Lab to MuJoCo without retraining. Future work will extend the framework to hardware deployment and investigate a unified mode-conditioned policy for smooth switching among behaviours.

Acknowledgements

I am grateful to Prof. Mahidzal Bin Dahari for the research guidance and to the reviewers for their helpful suggestions.

References

Atanassov, V., Ding, J., Kober, J., Havoutis, I., Della Santina, C.: Curriculum-based reinforcement learning for quadrupedal jumping: A reference-free design. IEEE Robotics & Automation Magazine 32(2), 35-48 (2025). https://doi.org/10.1109/MRA.2024.3487325

Bellegarda, G., Nguyen, C., Nguyen, Q.: Robust quadruped jumping via deep reinforcement learning. Robotics and Autonomous Systems 182, 104799 (2024). https://doi.org/10.1016/j.robot.2024.104799

Hoeller, D., Rudin, N., Sako, D.V., Hutter, M.: ANYmal parkour: Learning agile navigation for quadrupedal robots. Science Robotics 9(88), eadi7566 (2024). https://doi.org/10.1126/scirobotics.adi7566

Hwangbo, J., Lee, J., Dosovitskiy, A., Bellicoso, C.D., Tsounis, V., Koltun, V., Hutter, M.: Learning agile and dynamic motor skills for legged robots. Science Robotics 4(26), eaau5872 (2019). https://doi.org/10.1126/scirobotics.aau5872

Kumar, A., Fu, Z., Pathak, D., Malik, J.: RMA: Rapid Motor Adaptation for Legged Robots. In: Proceedings of Robotics: Science and Systems XVII (2021). https://doi.org/10.15607/RSS.2021.XVII.011

Lee, J., Hwangbo, J., Wellhausen, L., Koltun, V., Hutter, M.: Learning quadrupedal locomotion over challenging terrain. Science Robotics 5(47), eabc5986 (2020). https://doi.org/10.1126/scirobotics.abc5986

Margolis, G.B., Yang, G., Paigwar, K., Chen, T., Agrawal, P.: Rapid locomotion via reinforcement learning. The International Journal of Robotics Research 43(4), 572-587 (2024). https://doi.org/10.1177/02783649231224053

Miki, T., Lee, J., Hwangbo, J., Wellhausen, L., Koltun, V., Hutter, M.: Learning robust perceptive locomotion for quadrupedal robots in the wild. Science Robotics 7(62), eabk2822 (2022). https://doi.org/10.1126/scirobotics.abk2822

Peng, X.B., Andrychowicz, M., Zaremba, W., Abbeel, P.: Sim-to-real transfer of robotic control with dynamics randomization. In: 2018 IEEE International Conference on Robotics and Automation (ICRA) (2018). https://doi.org/10.1109/ICRA.2018.8460528

Peng, X.B., Ma, Z., Abbeel, P., Levine, S., Kanazawa, A.: AMP: Adversarial Motion Priors for Stylized Physics-Based Character Animation. ACM Transactions on Graphics 40(4), Article 144 (2021). https://doi.org/10.1145/3450626.3459670

Rudin, N., Hoeller, D., Reist, P., Hutter, M.: Learning to walk in minutes using massively parallel deep reinforcement learning. In: Proceedings of the 5th Conference on Robot Learning, PMLR 164, 91-100 (2022)

Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. 2nd edn. MIT Press, Cambridge (2018)

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv:1707.06347 (2017). https://doi.org/10.48550/arXiv.1707.06347

Tan, J., Zhang, T., Coumans, E., Iscen, A., Bai, Y., Hafner, D., Bohez, S., Vanhoucke, V.: Sim-to-real: Learning agile locomotion for quadruped robots. In: Proceedings of Robotics: Science and Systems XIV (2018). https://doi.org/10.15607/RSS.2018.XIV.010

Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., Abbeel, P.: 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 (2017). https://doi.org/10.1109/IROS.2017.8202133

Unitree Robotics: unitree_rl_lab. GitHub repository, https://github.com/unitreerobotics/unitree_rl_lab, last accessed 2026/05/10

Xie, Z., Da, X., van de Panne, M., Babich, B., Garg, A.: Dynamics randomization revisited: A case study for quadrupedal locomotion. In: 2021 IEEE International Conference on Robotics and Automation (ICRA) (2021). https://doi.org/10.1109/ICRA48506.2021.9560837