Using guide model in reinforcement learning to teach robot to walk

Wenhao Zhang
School of Engineering and Materials Science
Queen Mary University of London
London, UK
ml22505@quml.ac.uk

***Abstract*—As deep reinforcement learning (DRL) becomes increasingly mature in practical applications of image recognition, and with the rapid advancements in the field of language models in recent years, research on its application in robotics has also been on the rise. DRL provides an effective framework and toolset for designing complex robot motion behaviours that are difficult to manually program. This framework gradually completes motion design through trial-and-error interactions between the agent and the environment. However, despite the great potential of DRL in training robotic motion control, it still faces several key challenges, particularly the significantly increased training costs associated with handling high-dimensional continuous actions and states, as well as the challenge of achieving efficient learning from limited experiences.**

Taking bipedal walking in robots as an example, traditional methods of setting reward functions in reinforcement learning can generate coherent gaits, but the resulting movements differ significantly from normal human walking. This discrepancy arises because the highly simplified models in deep reinforcement learning fail to fully simulate and infer the factors that lead humans to choose bipedal walking.

This report aims to propose an auxiliary system by drawing an analogy to the experience of infants learning to walk. Using the simplest bipedal walking as an example, the paper introduces a symbolic walking-derived reward function to optimize the agent's training efficiency and outcomes, ultimately generating a periodic, minimal bipedal controller. Through computer simulation experiments, this paper will provide a detailed analysis and comparison of the performance and results of this approach.

The code repository for this report can be accessed here.

Keywords—deep reinforcement learning, bipedal walking, symbolic behaviour, infant behaviour

I. INTRODUCTION

With the success of large language models in the field of natural language processing, there is a growing interest in applying their principles to various levels of robotic control to address the limitations of traditional control methods in terms of generalization. However, this emerging field is currently facing numerous challenges. Methodologically, this endeavor intersects with multiple disciplines, including robotics kinematics, dynamics, physics, and biomechanics. During the simulation phase, the accurate creation of digital twins presents additional complexities.

At present, training based on large-scale neural networks in robotics remains in its infancy, primarily due to the scarcity of real-world data. Although some companies have begun to open-source robotic motion datasets, analogous to ImageNet in the image processing domain [1], the current scale of available data is still insufficient.

This study focuses on the field of deep reinforcement learning, where agent-environment interactions are leveraged to design behaviours [2]. Reinforcement learning primarily involves two key components: the design of the reward function and the selection of the optimization algorithm. When addressing a given problem, reinforcement learning algorithms seek to maximize rewards by capturing the essence of the problem through human-engineered representations and employing optimization algorithms. Therefore, improving the effectiveness and efficiency of learning fundamentally hinges on these two aspects.

This study focuses on the field of deep reinforcement learning, where agent-environment interactions are leveraged to design behaviors. Reinforcement learning primarily involves two key components: the design of the reward function and the selection of the optimization algorithm [3]. When addressing a given problem, reinforcement learning algorithms seek to maximize rewards by capturing the essence of the problem through human-engineered representations and employing optimization algorithms. Therefore, improving the effectiveness and efficiency of learning fundamentally hinges on these two aspects.

This research begins by establishing a well-designed reward mechanism, aiming to enhance training outcomes and efficiency through the appropriate representation and quantification of the problem. For instance, the symbolic construction for bipedal walking proposed by Harnack has been shown to significantly reduce training time and increase walking speed by generating effective reward functions [4]. Drawing inspiration from the role of parental assistance in an infant's learning to walk, this study introduces an auxiliary support system, referred to as the "Parental System," designed to accelerate the training process of the agent and guide it toward a more natural gait.

II. RELATED WORK

A. Infant behaviour

Meltzoff and Moore reported that infants as young as two to three weeks old are capable of imitating adult behaviours [5]. In the field of neuroscience, Marshal and Meltzoff 's research on the neural mirroring mechanisms and imitation in human infants suggests that they are prolific imitators [6]. Even before the onset of language, infants' brains exhibit an observation-execution connection. While the origins of this imitative behaviour remain unclear, imitation serves as a crucial pathway for infants to acquire various types of knowledge and motor skills, as indicated by Jones's study [7]. Investigating the principles of infant motor learning can

provide valuable insights for this research, making deep reinforcement learning an appropriate method to simulate the process of infant walking.

In reinforcement learning, the agent's process of observing the environment, receiving feedback, and then executing actions to obtain new feedback mirrors the observation-execution structure seen in infants. Zaadnoordijk's research suggests that understanding the developmental science of infant cognition may hold the key to unlocking the next generation of unsupervised machine learning methods [8]. One discussed factor influencing the infant learning process is guidance and constraint, which can significantly impact the performance of reinforcement learning.

During the early stages of an infant's learning to walk, the primary guidance and constraints come from adults, typically the parents. Initially, adults guide the infant by holding their hands and slowly moving them to help the infant experience walking and learn the process. As training progresses, the adult gradually reduces the support, allowing the infant to develop independent walking skills, ultimately leading to the infant mastering the ability to walk.

This simplified version of the process inspires this study to incorporate a physical support system into reinforcement learning—a "cyber-parent" for a "cyber-baby," or what we refer to as the "Parental System." This system is designed to guide the agent's training, with the level of assistance varying according to the training duration and effectiveness.

B. Reinforcement learning- optimization algorithm

The design of the reward function and the selection of the optimization algorithm are two critical components of reinforcement learning. In terms of optimization algorithms for reinforcement learning, there are various approaches, broadly categorized into model-based and model-free policy search methods. Among model-based approaches, guided policy search, as proposed by Levine, is an algorithm that leverages traditional controllers and supervised learning to generate smooth action sequences [9]; However, it requires examples data for learning. On the other hand, model-free methods such as stochastic policy gradient methods have been shown to successfully train deep neural networks with billions of parameters [10].

Subsequent research has introduced various advancements in optimization algorithms, including Trust Region Policy Optimization (TRPO) [11] and DDPG method, which serves as the foundation for Deep Deterministic Policy Gradient (DDPG) [12]. The DDPG algorithm has been proven to be highly effective for policy search problems in robotics [13] [14]. In deterministic policy methods, the action is deterministic, meaning that if the deterministic policy gradient exists, solving for the policy gradient does not require sampling and integrating over the action space. Therefore, compared to stochastic policy methods, deterministic policies require fewer sample data, particularly for agents with large action spaces, such as the bipedal robot in this study, where the dimensionality of the action space is substantial. Using a stochastic policy would necessitate extensive sampling within these large action spaces. Hence, this report employs DDPG as the optimization algorithm.

C. Symbolic behaviour modelling

Symbolic behaviour modelling, or traditional behaviour modelling, remains significant in the current landscape of machine learning. While machine learning excels at extracting models from complex, highly coupled parameters that traditional methods cannot easily provide, this does not imply that modelling or quantification has lost its relevance. Research by Yang and Chester demonstrates the potential of integrating symbolic systems with reinforcement learning frameworks [15] [16]. Additionally, Landajuela's work shows that symbolic optimization can address early decision-making and initialization bias in the automatic mathematical computation tasks of neural networks [17], thereby improving machine learning performance, sample efficiency, and the complexity of solutions.

In the realm of reinforcement learning, an improperly designed reward function can lead to slow initial learning and issues such as local optimization. Effective symbolic motion modelling can significantly enhance the training outcomes of models. For instance, Harnack's research indicates that guiding the reward function with a symbolic representation of bipedal walking can improve both training speed and walking velocity. This experiment will utilize Harnack's approach as the primary method for deriving the reward function, which simplifies the bipedal robot model into a compass walker by classifying various gaits by Perry's research [18], and then performs dynamic analysis on the compass walker to derive a formalized model of ideal walking. This model is represented as a hybrid automaton composed of four cyclic orthogonal sequences. The next section will provide a detailed discussion of this approach.

D. Simulated environment

In this experiment, MIMo [19] is used for simulation, a library designed for studying cognitive development and multisensory learning. MIMo retains the application programming interface (API) of the Gymnasium environment and utilizes MuJoCo for physical simulation.

III. BACKGROUND

A. Policy Gradient Algorithm

The process of bipedal walking can be abstracted within the framework of a Markov Decision Process (MDP) [20]. In this framework, the system is assumed to adhere to the Markov property, which states that the next state $\boldsymbol{S_{t+1}}$ depends solely on the current state $\boldsymbol{S_t}$ and is independent of previous states. This property can be formally defined as follows:

$$P[S_{t+1}|S_t] = P[S_{t+1}|S_1, \dots, S_t] \quad (1)$$

A Markov Decision Process (MDP) can be described as a tuple $\boldsymbol{(S, A, p_0, P, R, \gamma)}$, where:

- $\boldsymbol{S}$ represents the finite state space,
- $\boldsymbol{A}$ denotes the finite action space,
- $\boldsymbol{p_0}$ is the initial state distribution of the system,
- $\boldsymbol{P}$ indicates the state transition probabilities,
- $\boldsymbol{R}$ is the reward function, and
- $\boldsymbol{\gamma}$ is the discount factor used to compute the cumulative return.

In the case of bipedal walking, the state space $\boldsymbol{S}$ consists of the three-dimensional rotational angles of each joint, while the action space $\boldsymbol{A}$ comprises the torques generated by the actuators of each joint.

In reinforcement learning, the goal is to find the optimal policy given a Markov Decision Process (MDP). The term "optimal policy" refers to the mapping from the state space to the action space that maximizes the expected cumulative return. This policy is often denoted by the symbol π\piπ, and can be formally represented as follows:

$$S \rightarrow A \tag{2}$$

Here, $\boldsymbol{\pi(s)}$ denotes the action taken by the policy $\boldsymbol{\pi}$ when the system is in state $\boldsymbol{S}$. The optimal policy $\boldsymbol{\pi^*}$ is the policy that maximizes the expected return over time, which can be expressed as:

$$\pi^* = argmax_\pi \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t | \pi\right] \tag{3}$$

where $\boldsymbol{R_t}$ represents the reward received at time step $\boldsymbol{t}$, and $\boldsymbol{\gamma^t}$ is the discount factor that determines the present value of future rewards. The latter part of Equation 3 is usually denoted as $\boldsymbol{V(s)}$, the value function of policy.

Bipedal walking is a problem that involves a continuous, high-dimensional MDP. Policy gradient algorithms have proven successful in addressing such issues. In this study, we use the Deep Deterministic Policy Gradient (DDPG) algorithm to compute the optimal policy.

DDPG is a classical algorithm for continuous control problems. The determinism in DDPG, as opposed to methods like Proximal Policy Optimization (PPO) [21], lies in its output being a direct action rather than a probability distribution. Consequently, DDPG does not involve processes like importance sampling. The algorithm is an improvement over Deep Q-Networks (DQN) [22], incorporating an actor-critic approach to handle continuous control tasks, which DQN cannot manage effectively.

The specific structure of the DDPG algorithm is as follows:

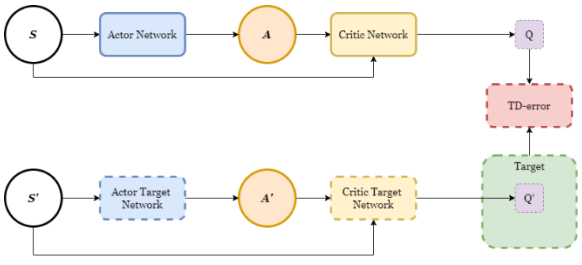

Fig 1. Structure of the DDPG algorithm

- **Actor Network:** This network is responsible for selecting actions based on the current state.
- **Critic Network:** This network evaluates the action taken by the actor by estimating the Q-value function. The critic network is trained to minimize the difference between its Q-value predictions and the target Q-values.
- **Target Networks:** Both the actor and critic networks have target networks with the same structure but with last state parameter, which are periodically updated to provide stable targets for training. To avoid the difficulties associated with changing targets during updates, DDPG employs fixed network techniques. Specifically, it involves freezing the networks used to compute the target values and then periodically updating the target networks with the parameters from the current networks.
- **Training:** The critic network is trained by minimizing the loss between the predicted Q-values and the target Q-values, while the actor network is updated using the policy gradient derived from the critic network to maximize the expected return.

This combination of actor-critic architecture with experience replays and target networks allows DDPG to effectively handle the continuous action spaces typical of problems like bipedal walking.

B. Symbolic Bipedal Walking

In Section 2, we briefly discussed modelling approaches for bipedal walking. In this study, bipedal walking is modelled as a compass walker, focusing exclusively on the hip joints to reduce complexity.

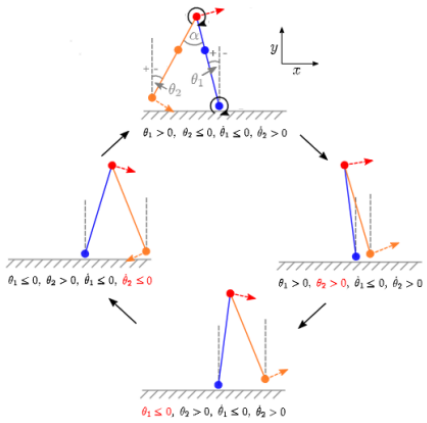

Fig 2. The four states of a compass walker in a cycle

Using this model, we can construct a hybrid automaton that integrates both continuous and discrete actions.

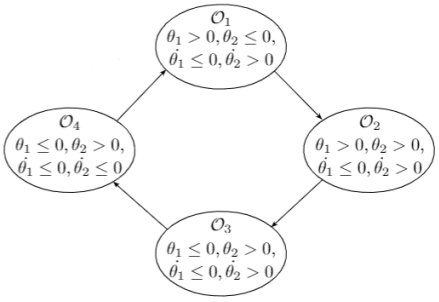

Fig 3. Hybrid automaton modelling the orthant sequences

This automaton can be represented as a tuple:

$$H = (V, Q, E, \mu_1, \mu_2, \mu_3) \quad (4)$$

Where:

- $\boldsymbol{V}$ is the set of variables related to joint angles and velocities.
- $\boldsymbol{Q}$ is the set of discrete walking states.
- $\boldsymbol{E}$ defines the transition relations between different walking states.
- $\boldsymbol{\mu_1}$ defines the continuous behavior of data variables in state **q**, which is represented by a set of differential equations.
- $\boldsymbol{\mu_2}$ specifies the invariant conditions for data variables **σ** in state **q**.
- $\boldsymbol{\mu_3}$ describes the transition relations of data variables when moving from state **σ** to next state **τ**.

From Harnack's derivation, we learn the following equation:

$$E = \{(\mathcal{O}_1, \mathcal{O}_2), (\mathcal{O}_2, \mathcal{O}_3), (\mathcal{O}_3, \mathcal{O}_4), (\mathcal{O}_4, \mathcal{O}_1)\} \quad (5)$$

$$\mu_2(\mathcal{O}_1) \Leftrightarrow \theta_1 > 0 \wedge \theta_2 \leq 0 \wedge \dot{\theta}_1 \leq 0 \wedge \dot{\theta}_2 > 0 \quad (6)$$

$$\mu_2(\mathcal{O}_2) \Leftrightarrow \theta_1 > 0 \wedge \theta_2 > 0 \wedge \dot{\theta}_1 \leq 0 \wedge \dot{\theta}_2 > 0 \quad (7)$$

$$\mu_2(\mathcal{O}_3) \Leftrightarrow \theta_1 \leq 0 \wedge \theta_2 > 0 \wedge \dot{\theta}_1 \leq 0 \wedge \dot{\theta}_2 > 0 \quad (8)$$

$$\mu_2(\mathcal{O}_4) \Leftrightarrow \theta_1 \leq 0 \wedge \theta_2 > 0 \wedge \dot{\theta}_1 \leq 0 \wedge \dot{\theta}_2 \leq 0 \quad (9)$$

$$\mu_3(p, q) = \begin{cases} \{(\theta_1, \theta_2), (\theta_2, \theta_1), (\dot{\theta}_1, \dot{\theta}_2^+), (\dot{\theta}_2, \dot{\theta}_1^+)\} \\ \quad \text{if } p = \mathcal{O}_4, q = \mathcal{O}_1 \\ \{(\theta_1, \theta_1), (\theta_2, \theta_2), (\dot{\theta}_1, \dot{\theta}_1), (\dot{\theta}_2, \dot{\theta}_2)\} \\ \quad \text{otherwise} \end{cases} \quad (10)$$

How to apply this hybrid automaton to reward functions is discussed in the next section.

IV. GUIDE SYSTEM

A. Auxiliary systems - parent systems

This study draws inspiration from the process of infants learning to walk, highlighting the role of parents not only as models to be imitated but also as guides in this learning process. Based on this insight, we introduce an auxiliary system, referred to as the "parent system," whose purpose is to guide the bipedal agent forward and help it maintain balance.

Specifically, during the learning process, the parent system provides a set of orthogonal forces at the agent's hips to assist with standing and balancing, as well as a forward-directed force to guide the agent in walking. These two requirements are implemented using proportional-derivative (PD) controllers. The controllers generate the necessary forces to achieve the desired outcomes. In detail, the PD controller regulates the position of the agent's hips and its forward velocity. The target balance position is set to the normal standing posture, and the forward speed is aligned with the agent's desired walking speed.

The general expression of the PD controller can be represented as:

$$u(t) = K_p * e(t) + K_d * \frac{de(t)}{dt} \quad (11)$$

where $\boldsymbol{K_p}$ is the stiffness coefficient, $\boldsymbol{K_d}$ is the damping coefficient, and $\boldsymbol{e(t)}$ represents the error between the current and target states. In machine learning, during each iteration, the values of $\boldsymbol{K_p}$ and $\boldsymbol{K_d}$ can be adjusted to influence the strength of the balancing and propulsion forces.

In this study, we implement a curriculum schedule based on the learner-centre curriculum model proposed in Yu's Study [23]. This model has been shown to be more efficient and data-effective. Specifically, our approach to the PD controller involves systematically and gradually reducing the auxiliary forces, ultimately enabling the agent to perform walking motions without any assistance. Similar to adjusting the learning rate in reinforcement learning, the challenge lies in designing a scientifically sound method to decrease the auxiliary forces, balancing learning efficiency with learning effectiveness. The learner-centre curriculum scheduling algorithm follows these steps:

- Begin by applying a sufficiently simple initial auxiliary force $\boldsymbol{x_0}$ to ensure that a standard policy learning algorithm can produce a successful policy $\boldsymbol{\pi}$.
- In each iteration of the curriculum, first update the policy $\boldsymbol{\pi}$ by running the standard policy learning algorithm. If the updated policy achieves an average return that is at least 80% of the initial return R^-., proceed to update the auxiliary force $\boldsymbol{x}$ using Update Algorithm†.
- Conduct a one-dimensional line search along five directions to find the maximum step size $\boldsymbol{\alpha^*}$ that allows the current policy $\boldsymbol{\pi}$ to still achieve 60% of the initial return R^-. Update the auxiliary force $\boldsymbol{x}$ to $\boldsymbol{x+\alpha^* d^*}$, where $\boldsymbol{d^*}$ is the optimal direction identified through the search.
- Continue this process until the norm of the auxiliary force $\boldsymbol{x}$ falls below a threshold $\boldsymbol{\epsilon}$, At this point, perform a final policy learning phase without any auxiliary force.

B. Reward Function

In this section, we define the reward function based on the formalism of the hybrid automaton discussed in Section 3. By symbolizing the walking gait, we can encourage the agent to perform healthier, more symmetrical, and periodic walking behaviours.

Reward Function Defined as the following function:

† The update algorithm can be found in Yu's report (Algorithm II) and will not be repeated here.

$$r_{\mathrm{or}}(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}) = \begin{cases} +1 & \text{if} \quad \mathcal{O}(\boldsymbol{x}_{t-1}) \in Q \wedge \\ & \quad \mathcal{O}(\boldsymbol{x}_t) \in Q \wedge \\ & \quad (\mathcal{O}(\boldsymbol{x}_{t-1}), \mathcal{O}(\boldsymbol{x}_t)) \in E \\ +1 & \text{if} \quad \mathcal{O}(\boldsymbol{x}_{t-1}) \notin Q \wedge \\ & \quad \mathcal{O}(\boldsymbol{x}_t) \in Q \\ -1 & \text{else} \end{cases} \tag{12}$$

The figure below visualizes the rewards when the four state paces make all possible transfers.

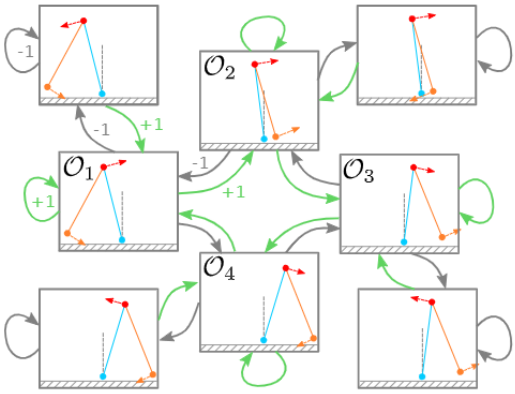

Fig 4. Visualization of the reward function

For comparison, we will also establish a traditional motion task control reward as a benchmark, specifically the forward direction reward $\boldsymbol{r_{for}}$:

This reward component assigns positive or negative points based on the agent's horizontal displacement. The agent receives a positive reward when moving forward and a negative reward when moving backward.

V. Result & Discussion

In this study, we used MIMo for simulations. To simplify the model, the degrees of freedom for the torso, head, and arm joints of the baby robot were frozen, leaving only the degrees of freedom for the leg, foot, and hip joints. The experimental test model is the one trained under the parent system combined with the symbolically guided reward function $\boldsymbol{r_{or}}$, referred to as $\boldsymbol{Model_{p+or}}$. As a control, three additional models were tested:

$\boldsymbol{Model_{or}}$: Without the parent system, trained with the reward function $\boldsymbol{r_{or}}$.

$\boldsymbol{Model_{p+for}}$: With the parent system, trained with the reward function $\boldsymbol{r_{for}}$.

$\boldsymbol{Model_{for}}$: Without the parent system, trained with the reward function $\boldsymbol{r_{for}}$.

The following includes the average learning curves for the four models and visualizations of the gait patterns trained using the guided model:

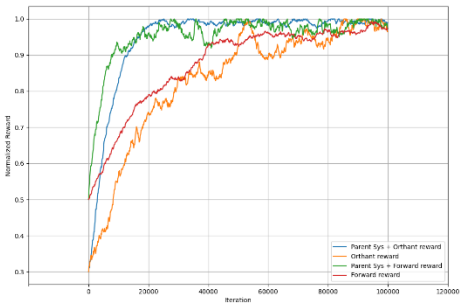

Fig 5. Average learning curves for the four models

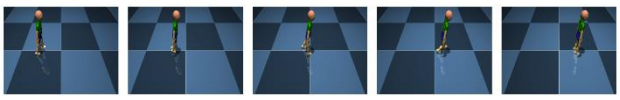

Fig 6. Results of bipedal walking under guide model training

Model Performance Overview*

$\mathbf{Model_{p+or}}$ (Parent System + Symbolically Guided Reward):

Learning Curve: It shows the fastest convergence with the smallest variance among the four models. This indicates that the combination of the parent system and the symbolically guided reward function facilitates more stable and rapid learning.

Gait Effectiveness: The gait pattern achieved is notably smooth and natural. This is attributed to the effectiveness of the parent system in providing balance and guidance, combined with the well-designed reward function that promotes symmetrical and periodic walking.

$\mathbf{Model_{or}}$ (Without Parent System + Symbolically Guided Reward):

Learning Curve: This model converges more slowly compared to **$\mathbf{Model_{p+or}}$**, suggesting that the absence of the parent system hinders the learning process. However, the learning stability remains relatively good.

Gait Effectiveness: The gait produced is less effective compared to **$\mathbf{Model_{p+or}}$** . While the reward function $\boldsymbol{r_{or}}$ encourages forward motion, the lack of a guiding system results in less refined gait patterns.

$\mathbf{Model_{p+for}}$ (Parent System + Reward $\boldsymbol{r_{for}}$):

Learning Curve: It converges at a rate similar to $\boldsymbol{Model_{p+or}}$ indicating that the parent system contributes positively to the learning speed.

Gait Effectiveness: Despite the effective learning curve, the gait patterns produced are less natural compared to $\boldsymbol{Model_{p+or}}$. The reward function $\boldsymbol{r_{for}}$ primarily promotes forward motion without emphasizing gait symmetry and periodicity, leading to less natural and more erratic steps.

$\mathbf{Model_{for}}$ (Without Parent System + Reward $\boldsymbol{r_{for}}$):

Learning Curve: It shows the slowest convergence and highest variance, highlighting the challenges in learning without both the parent system and an effective reward function.

Gait Effectiveness: The gait produced is the least effective, characterized by unnatural and inconsistent steps. The absence of the parent system, combined with a reward function that focuses solely on forward motion, results in suboptimal walking patterns.

The presence of the parent system significantly enhances both the learning speed and the quality of the gait patterns. The parent system's role in providing balance and guidance helps the agent achieve a more stable and effective learning process. This effect is particularly evident in the comparison between $\mathbf{Model_{p+or}}$ and $\boldsymbol{Model_{for}}$.

The reward function $\boldsymbol{r_{or}}$ is more effective in guiding the agent towards a natural gait compared to $\boldsymbol{r_{for}}$. Reward function $\boldsymbol{r_{or}}$, which emphasizes symmetry and periodicity, results in smoother and more natural walking patterns. In contrast, $\boldsymbol{r_{for}}$ focuses mainly on forward motion, which may lead to less coherent gait patterns even with the parent system's assistance.

The combination of the parent system and a well-designed reward function provides the best results in terms of learning efficiency and gait quality. $\mathbf{Model_{p+or}}$ outperforms the other models in both convergence speed and gait effectiveness, demonstrating the importance of integrating both effective guidance and reward structures.

The experiment demonstrates that incorporating a parent system alongside a symbolically guided reward function significantly enhances both the learning speed and the quality of bipedal walking, as evidenced by the superior performance of $\mathbf{Model_{p+or}}$. However, this study also highlights several critical limitations and areas for improvement. The use of a simplified bipedal model and symbolic gait modeling may not fully capture the complexity of real-world walking dynamics, potentially limiting the generalizability of the findings. The reward function $\boldsymbol{r_{or}}$, while effective in promoting more natural gait patterns, may still be insufficient for addressing all aspects of gait refinement, particularly in more complex or dynamic environments. Additionally, the absence of the parent system in other models clearly hinders learning, but the specific mechanisms by which the parent system facilitates better gait patterns remain underexplored. Future work should address these limitations by incorporating more detailed and realistic models of bipedal locomotion to better simulate real-world scenarios. Researchers should also investigate advanced guidance mechanisms and reward structures that can adapt to more nuanced gait characteristics. Furthermore, optimizing the reward functions to balance gait symmetry, smoothness, and efficiency while evaluating their performance across a broader range of environments could yield significant improvements. By addressing these areas, future studies can enhance the robustness and applicability of the findings, ultimately leading to more effective and generalizable solutions for bipedal locomotion training.

VI. Conclusion & Outlook

This study draws inspiration from the process of infants learning to walk and introduces a novel parental assistance system. This system, combined with a symbolic gait modeling approach, enables the efficient generation of bipedal walking patterns that are appropriately paced, alternately symmetrical, and periodic through reinforcement learning algorithms. The first key aspect of this method is the design of a PD controller-based assistance system that helps the agent maintain balance and swiftly reach the training target position. The second crucial element is the use of symbolic gait modeling to guide the formulation of the reward function. Comparative analysis with traditional training methods demonstrates that the approach used in this study achieves higher efficiency and better performance.

While the results of this study show promising training outcomes, it is important to note that the bipedal model employed is still a simplified representation, and its performance cannot yet fully match real-world data. Future research could expand this bipedal model to more complex human-like models and explore the use of more advanced optimization algorithms to further enhance training efficiency.

Acknowledgment

I sincerely thank my advisor, Lorenzo Jamone, for their invaluable guidance. My deepest gratitude goes to my family and my girlfriend, Bai, for their constant support and encouragement.

* The code repository for this experiment can be accessed here.

REFERENCES

[1] W. D. R. S. L.-J. L. K. L. a. L. F.-F. Jia Deng, "ImageNet: A Large-Scale Hierarchical Image Database," *conference on Computer Vision and Pattern Recognition,* 2009.

[2] D. Bertsekas, "Reinforcement learning and optimal control," *Athena Scientific,* 2019.

[3] R. S. S. a. A. G. Barto, Reinforcement learning: An introduction., MIT press, 2018.

[4] C. L. Daniel Harnack, "Deriving Rewards for Reinforcement Learning from Symbolic Behaviour Descriptions of Bipedal Walking," *https://arxiv.org/pdf/2312.10328,* 2023.

[5] M. M. Meltzoff AN, "Imitation of facial and manual gestures by human neonates.," *Science.*, no. 198, pp. 75-78, 1977.

[6] A. N. M. Peter J. Marshall, "Neural mirroring mechanisms and imitation in human infants," *Philos Trans R Soc Lond B Biol Sci. ,* 2014.

[7] S. S. Jones, "The development of imitation in infancy," *Philos Trans R Soc Lond B Biol Sci.,* pp. 2325-2335, 2009.

[8] T. R. B. &. R. C. Lorijn Zaadnoordijk, "Lessons from infant learning for unsupervised machine learning," *Nature Machine Intelligence volume,* no. 4, pp. 510-520, 2022.

[9] V. K. Sergey Levine, "Learning Complex Neural Network Policies with Trajectory Optimization," *Proceedings of the 31st International Conference on Machine Learning, PMLR,* vol. 2, no. 32, pp. 829-837, 2014.

[10] Y. B. &. G. H. Yann LeCun, "Deep learning," *Nature,* no. 521, pp. 436-444, 2015.

[11] S. L. P. M. M. I. J. P. A. John Schulman, "Trust Region Policy Optimization," 2015.

[12] J. J. H. A. P. N. H. T. E. Y. T. D. S. D. W. Timothy P. Lillicrap, "Continuous control with deep reinforcement learning," *https://arxiv.org/abs/1509.02971,* 2015.

[13] T. H. J. S. F. W. Mel Vecerik, "Leveraging Demonstrations for Deep Reinforcement Learning on Robotics Problems with Sparse Rewards," *https://arxiv.org/pdf/1707.08817,* 2018.

[14] B. M. Ashvin Nair, "Overcoming Exploration in Reinforcement Learning," *https://arxiv.org/pdf/1709.10089,* 2018.

[15] D. L. Fangkai Yang, "PEORL: Integrating Symbolic Planning and Hierarchical Reinforcement Learning for Robust Decision-Making," *https://arxiv.org/pdf/1804.07779,* 2018.

[16] M. D. F. Z. a. J. T. Andrew Chester, "SAGE: Generating Symbolic Goals for Myopic Models in Deep Reinforcement Learning," *https://arxiv.org/pdf/2203.05079,* 2022.

[17] B. K. P. S. K. K. Mikel Landajuela, "IMPROVING EXPLORATION IN POLICY GRADIENT SEARCH: APPLICATION TO SYMBOLIC OPTIMIZATION," *https://arxiv.org/pdf/2107.09158,* 2021.

[18] J. Perry, "Gait Analysis — Normal and Pathological Function," *NJ: SLACK Incorporated,* 1992.

[19] "MIMo's documentation," 2023. [Online]. Available: https://mimo.readthedocs.io/en/latest/. [Accessed 2024].

[20] B. Everitt, "The Cambridge Dictionary of Statistics.," *2002.*

[21] F. W. P. D. A. R. O. K. John Schulman, "Proximal Policy Optimization Algorithms," *https://arxiv.org/abs/1707.06347,* 2017.

[22] K. K. D. S. A. G. I. A. D. W. M. R. Volodymyr Mnih, "Playing Atari with Deep Reinforcement Learning," *https://arxiv.org/abs/1312.5602,* 2013.

[23] G. T. L. WENHAO YU, "Learning Symmetric and Low-Energy Locomotion," *https://arxiv.org/pdf/1801.08093,* 2018.

[24] V. K. Sergey Levine, "Learning Complex Neural Network Policies with Trajectory Optimization," *PMLR,* vol. 2, no. 32, pp. 829-837, 2014.