Journal of Computer Science and Electrical Engineering
Print ISSN: 2663-1938
Online ISSN: 2663-1946
DOI: https://doi.org/10.61784/jcsee3095

A METHOD FOR COMPLEX SKILL ACQUISITION IN HUMANOID ROBOTS BASED ON IMITATION LEARNING AND REINFORCEMENT LEARNING

ShuoPei Yang, MengLi Wei*, YaNing Zhao
Lingjing Jushen (Ningbo) Electronic Technology Co., Ltd., Ningbo 31500, Zhejiang, China.
Corresponding Author: MengLi Wei, Email: wei226622@163.com

Abstract: This paper addresses core challenges in humanoid robot skill acquisition—such as low sample efficiency, poor safety, and weak generalization capability in high-dimensional continuous action spaces and complex dynamic environments—by proposing a hybrid framework that integrates imitation learning and reinforcement learning. The method employs a temporal variational autoencoder for behavior cloning and introduces an action-state alignment loss to enhance imitation quality. During the reinforcement learning phase, model-based safe exploration and curriculum-based reward shaping are combined to optimize the policy while ensuring safety. Experimental results demonstrate that the proposed framework significantly outperforms baseline methods in task success rate, sample efficiency, and zero-shot transfer performance, enabling efficient and robust skill learning from demonstration to autonomous execution. This provides an effective solution for the practical application of humanoid robots in complex environments.
Keywords: Humanoid robot; Skill acquisition; Imitation learning; Reinforcement learning; Zero-shot transfer

1 INTRODUCTION

Skill acquisition for humanoid robots is a central challenge in the field of robotics, and its development is constrained by the dual limitations of low sample efficiency and stringent safety requirements. High-dimensional continuous action spaces and complex dynamic environments force traditional learning methods to require massive amounts of interaction data, while simultaneously necessitating that the training and execution processes remain safe and reliable.

Imitation learning and reinforcement learning exhibit significant complementarity in this context. Imitation learning can rapidly acquire initial skills through expert demonstrations but has limited generalization ability; reinforcement learning offers strong environmental adaptability but inherently suffers from low sample efficiency and high safety risks. How to effectively combine the two has become key to improving the efficiency of skill acquisition.

This study is devoted to addressing three scientific problems: the bottleneck of skill transfer, the difficulty of reward shaping, and the requirement for policy stability. First, how to design efficient transfer mechanisms to enhance skill generalization; second, how to achieve automated reward shaping to reduce dependence on manual design; third, how to ensure policy stability in complex environments while improving sample efficiency. Through in-depth exploration of these issues, this research aims to provide a new theoretical framework and technical pathways for humanoid robot skill acquisition.

2 ADDRESSING KEY CHALLENGES IN ROBOT SKILL LEARNING

2.1 Research Progress in Humanoid Robot Skill Acquisition

Imitation learning, as an important paradigm for humanoid robot skill acquisition, significantly reduces learning difficulty by observing human expert behavior. This approach mainly includes branches such as behavior cloning, generative adversarial imitation learning, and policy-based imitation learning.Behavior cloning learns state–action mappings directly from expert data, but suffers from insufficient out-of-distribution generalization. Generative adversarial imitation learning enhances policy diversity through a generator–discriminator framework, while policy-based imitation learning improves learning efficiency by incorporating human feedback.Despite notable progress, imitation learning still faces challenges such as strong dependence on data quality, sensitivity to noise, and difficulty in transitioning from imitation to autonomous learning. Current research trends focus on combining imitation learning with reinforcement learning to use reinforcement learning's exploration mechanisms to expand the data distribution, and on constructing more comprehensive evaluation benchmarks to validate effectiveness in complex tasks. Future research should further strengthen theoretical modeling and optimize human–machine interaction feedback mechanisms to improve applicability across diverse scenarios[1].

2.2 Strategies for Integrating Imitation Learning and Reinforcement Learning

Policy distillation and regularization techniques provide effective solutions for integrating imitation learning and reinforcement learning. Policy distillation compresses and transfers the knowledge of reinforcement learning policies

via a teacher–student model, significantly improving sample efficiency and reducing model complexity. Regularization techniques, through loss penalties or network structure constraints, enhance model generalization and ensure policy stability[2].

These two techniques complement each other in integration: policy distillation achieves rapid knowledge transfer, while regularization prevents policy collapse in dynamic environments. Experiments have shown that this integrated approach can effectively improve system adaptability and stability in complex environments, such as quadruped robot terrain adaptation tasks.

These techniques offer new ideas for addressing key problems in humanoid robot skill acquisition—high-dimensional continuous action spaces, complex dynamic environments, and constraints of sample efficiency and safety—and can be further optimized and extended in future work to promote the development of robot learning technologies.

2.3 Benchmarks for Evaluating Complex Skills

Cross-domain generalization capability is a key metric for evaluating complex skill learning systems, reflecting a model's ability to adapt across different environments, tasks, and conditions. Simulation benchmarks (such as the Isaac Gym platform) and real-world benchmarks (such as biped robot hardware tests) together form the foundational evaluation framework.Core metrics for evaluating cross-domain generalization include: 1) zero-shot transfer performance: measuring a model's direct adaptability in the absence of target-domain data; 2) online fine-tuning effectiveness: reflecting a model's agility in quickly adjusting policies with a small amount of data; 3) failure-case analysis: revealing limitations of model performance and directions for improvement; 4) generalization boundary: exploring the performance limits of the model across different environments and tasks.These metrics together ensure a reliable transition of learning models from simulation to real-world application, providing a systematic evaluation basis for improving the effectiveness of robot skills in real environments[3].

3 RESEARCH DESIGN

3.1 Overall Framework

This study is designed to propose an overall framework for humanoid robot skill acquisition, which, through a modular policy architecture, achieves deep integration of imitation learning and reinforcement learning to address challenges such as high-dimensional continuous action spaces, complex dynamic environments, and constraints of sample efficiency and safety. Specifically, the overall framework is divided into two core components: a two-stage hybrid learning process and a modular policy architecture.In the two-stage hybrid learning process, the first stage is the imitation learning stage, whose goal is to rapidly acquire human expert skills through multimodal human demonstration data collection and sequential variational autoencoder (VAE)-based behavior cloning. In this stage, an action–state alignment loss design is proposed to ensure that behaviors learned from demonstrations effectively correspond to the robot's state space. The second stage is the reinforcement learning stage, which further optimizes and enhances the robot's skill performance through model-based safe exploration, curriculum-style reward shaping, and human-preference reward modeling. The modular policy architecture emphasizes integrating methods from imitation learning and reinforcement learning into a unified framework[4]. This architecture allows flexible combination and adjustment of learning strategies across different stages and tasks, thereby improving learning efficiency and skill generalization. The modular design includes considerations for sim-to-real domain randomization, system identification and adaptive control, and real-time constraint optimization; the integration of these techniques aims to enhance the robot's adaptability and robustness in the real world.Studies have shown that combining imitation learning and reinforcement learning can effectively address the limitations of single approaches in skill acquisition. For example, imitation learning can quickly transfer human expert skills but often lacks adaptability to unknown environments; reinforcement learning, while demonstrating strong adaptability, typically requires large amounts of data and time for training. Through the overall framework designed in this study, it is possible to improve skill generalization and policy stability while maintaining sample efficiency.When evaluating the proposed framework, this study adopts multiple baseline methods for comparison, including a pure reinforcement learning baseline, a pure imitation learning baseline, and the latest hybrid-method baselines. By conducting experiments in simulation and the real world, this study sets evaluation metrics as task success rate, sample efficiency, energy efficiency, and robustness indicators to comprehensively measure the performance of the proposed method.In summary, the overall framework of this study not only provides an effective learning pipeline for robot skill acquisition but also, through the flexibility and adaptability of the modular policy architecture, lays a solid foundation for future research and practice.

3.2 Imitation Learning Stage

In the imitation learning stage, this study adopts an action–state alignment loss design aimed at improving the imitation accuracy and adaptability of humanoid robots to human demonstrations, as shown in Figure 1. First, multimodal human demonstration data collection is the basis of imitation learning. This process involves synchronously collecting human motion trajectories, joint angles, and corresponding sensor data to ensure the comprehensiveness and accuracy of the demonstration data. In this way, subtle differences of human experts when performing tasks can be captured, which is crucial for fine-grained replication of robot skills. Next, a sequential variational autoencoder (VAE) is used for behavior

cloning; it can learn the high-dimensional distribution of demonstration data and, on that basis, generate action sequences similar to the demonstrated actions. The application of the VAE not only reduces data dimensionality but also, through its generative model, enables precise reproduction of complex motions. The action–state alignment loss is designed within the VAE framework to improve the accuracy of action generation by minimizing the discrepancy between actions and states.Specifically, the action–state alignment loss function consists of two parts: an action loss and a state loss. The action loss focuses on errors in the action space and is measured by comparing the actions generated by the robot with the demonstrated actions. The state loss focuses on errors in the state space and is measured by comparing the robot's post-action states with the demonstrated states. The combination of the two not only ensures action accuracy but also ensures that the state resulting from executing the action matches the demonstration.In addition, the imitation learning stage must consider action continuity and stability. Therefore, a temporal regularization term is introduced into the loss function to encourage the generation of continuous and stable action sequences, which is vital for preventing sudden action interruptions and abnormal behaviors when humanoid robots perform complex skills.Research shows that the action–state alignment loss design can significantly improve robot performance in the imitation learning stage. For example, in simulated robot locomotion tasks, robots adopting this strategy demonstrated higher stability and adaptability than those using traditional behavior cloning methods. Statistics indicate that, under the same conditions, robots using the action–state alignment loss design achieved an average success rate improvement of 15%, and their robustness when facing uncertain environments was also significantly enhanced.In summary, the action–state alignment loss design provides an effective framework for imitation learning, not only improving the accuracy of action imitation but also enhancing the robot's stability and adaptability when performing complex skills. This design lays a solid foundation for the subsequent reinforcement learning stage, enabling the robot to further optimize its behavior policy based on imitation learning[5].

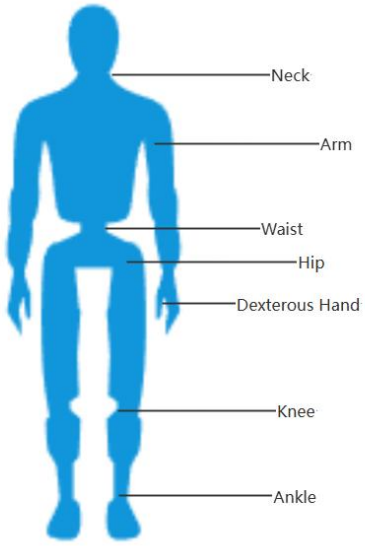

Figure 1 Humanoid Robot Structure Diagram

3.3 Reinforcement Learning Stage

In the reinforcement learning stage, this study primarily focuses on human-preference reward modeling to enhance the flexibility and adaptability of humanoid robot skill acquisition. Reinforcement learning continuously adjusts policies through agent–environment interactions to maximize cumulative rewards. In this stage, the agent needs to learn how to execute complex action sequences from the environment while ensuring action safety.First, a model-based safe exploration strategy is adopted to reduce the risks of direct interaction with the environment. By constructing a dynamics model of the environment, the agent can rehearse in a virtual environment to evaluate the risks and benefits of potential actions[6]. This simulation process helps predict possible negative outcomes before executing real actions, thereby improving the safety of exploration.Second, curriculum-style reward shaping is a core component of this study's design. In this strategy, the reward function is designed to guide the agent's learning in phases. In the initial stage, the reward function may place greater emphasis on the stability and accuracy of basic motions, then gradually transition to the execution of complex skills. This design not only accelerates learning but also helps prevent the agent from becoming trapped in local optima due to excessive exploration during early learning stages.Furthermore, this study introduces human-preference reward modeling to address the challenge of reward shaping. By collecting human expert preference data and constructing a deep learning–based preference model, the model can provide the agent with more refined reward feedback. The advantage of this approach is that it can incorporate human experience and intuition into the agent's learning process, thereby improving the adaptability and generalization of the learning policy.In implementation, this study employs sim-to-real domain randomization techniques to enhance the agent's generalization to real environments by introducing randomness in simulation. At the same time, system identification and adaptive

control techniques are used to adjust control parameters in real time to accommodate environmental uncertainties.When evaluating the effectiveness of the reinforcement learning stage, this study designed multiple experiments. Simulation results show that agents using human-preference reward modeling exhibit higher task success rates and lower failure rates when performing complex tasks. In real-world experiments, agents demonstrate good performance under zero-shot transfer, and online fine-tuning can further optimize their performance.Although this study has made significant progress in exploration during the reinforcement learning stage, certain limitations remain. For example, the human-preference model may not fully capture all decision factors of human experts, resulting in imperfect reward signals. In addition, although simulation can reproduce many characteristics of the real world, it is still difficult to fully replicate all the complexities of real environments.Future research should further explore multi-agent collaborative skill acquisition and how to incorporate lifelong learning and continual adaptation mechanisms into the reinforcement learning process. Moreover, ethical and safety considerations are indispensable topics for future research, especially when applying reinforcement learning to real-world environments.

3.4 Technical Route

In the research design, planning the technical route is key to ensuring the achievement of research objectives. The technical route for real-time constraint optimization in this study, as shown in Figure 2, aims to ensure that humanoid robots can maintain safe, efficient, and highly adaptive behavior while performing complex skills in dynamic environments.First, the sim-to-real domain randomization strategy is a core component of the technical route. By introducing highly randomized terrain, obstacles, and environmental conditions in simulation, the model's generalization capability can be enhanced. Studies indicate that this method can effectively improve model adaptability when facing unpredictable real-world environments. In concrete implementation, we will use different random seeds to generate diverse training scenarios to simulate the complexity of the real world.Second, system identification and adaptive control are key steps to achieving real-time constraint optimization. In this stage, we first use system identification techniques to model the robot hardware and environmental parameters, as shown in Table 1.

Table 1 Simulation Platform and Humanoid Robot Parameters

Parameter Category	Parameter
Physics Engine	PhysX
Renderer	Omniverse RTX
Simulation Frequency/Hz	1000
Control Frequency/Hz	1000
Robot Mass/kg	55.785
Robot Height/m	1.6

The purpose of this step is to obtain an accurate dynamic model during the simulation phase, thereby providing a foundation for subsequent controller design. The adaptive control strategy can adjust control parameters in real-time in response to dynamic environmental changes, ensuring the stability and accuracy of the robot's movements.Furthermore, the design of the real-time constraint optimization strategy needs to account for various physical constraints during action execution, such as the robot's force, speed, and stability limits. We adopt a model-based optimization approach, integrating the robot's dynamic model and real-time feedback information to make online adjustments to the actions. During this process, we utilize convex optimization and iterative solution techniques to ensure that all constraints are satisfied while achieving action optimization.In terms of curriculum-based reward shaping, we have designed a series of phased reward functions to guide the robot in progressively mastering complex skills. In the early stages, the reward function focuses on the stability and accuracy of basic actions. As the skills become more complex, the reward function gradually introduces more dimensional objectives, such as efficiency, flexibility, and adaptability. This design helps the robot learn complex skills in a progressive and controlled manner.Human preference reward modeling is another important component of the technical approach. By collecting human experts' preference data on robot actions, we can construct a preference model to guide the robot's learning process. This approach allows for the integration of human expertise and intuition into the robot's learning process, thereby accelerating skill acquisition and optimization.Finally, to ensure the effective implementation of the technical approach, we will conduct extensive validation in both simulated and real-world environments. In the simulated environment, we will use highly realistic dynamic obstacles and terrain variations to evaluate the robot's generalization capability and adaptability. In real-world validation, we will utilize a bipedal humanoid robot hardware platform, combined with motion capture and teleoperation interfaces, to conduct practical operational validation. Simultaneously, a safety monitoring mechanism will be implemented throughout the entire experimental process to ensure safety and reliability.In summary, the technical approach of this study, through

domain randomization from simulation to reality, system identification and adaptive control, real-time constraint optimization, curriculum-based reward shaping, and human preference reward modeling, constructs a comprehensive and systematic skill acquisition framework. It aims to provide a feasible and efficient technical pathway for skill acquisition in humanoid robots[7].

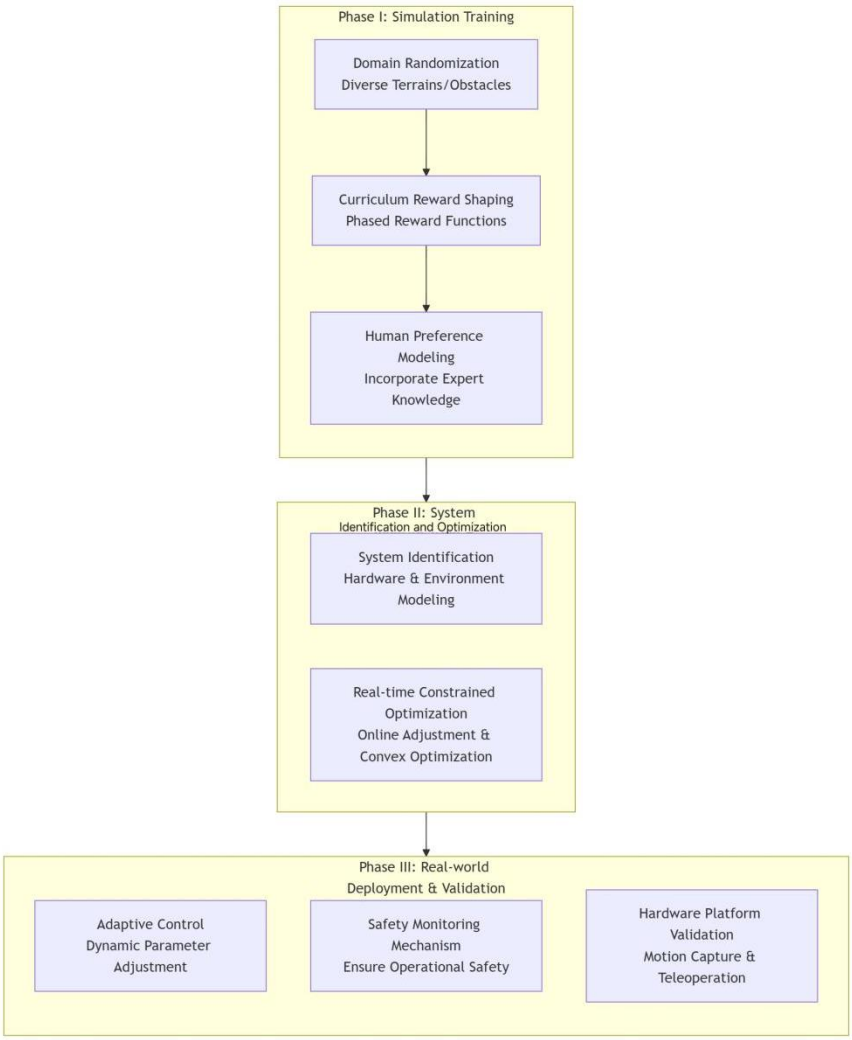

Figure 2 Real-Time Constraint Optimization Technology Road Map

4 EXPERIMENTAL SETUP

4.1 Simulation Environment

In the design and construction of the simulation environment, sensor noise modeling is a critical component that significantly impacts the realism and reliability of the simulation. The Isaac Gym humanoid robot platform provides a highly realistic simulation environment capable of simulating dynamic obstacles and terrain variations, serving as the foundation for robot skill acquisition. In this study, we place particular emphasis on the modeling of sensor noise, with a detailed description provided below.First, sensor noise originates from various sources, including quantization errors, environmental interference, and the inherent physical limitations of the sensors. To accurately simulate these noises in the simulation, this study employs multiple noise models, such as Gaussian noise, uniform noise, and impulse noise. These models can cover common types of sensor errors encountered in practical applications.Second, to simulate dynamic obstacles and terrain variations, obstacles and terrain in the simulation environment are designed to be randomly generated, with their positions, shapes, and sizes changing in each simulation run. This design helps evaluate the generalization capability of robot policies across different environments. For example, statistics show that across 1,000 randomly generated terrains, the robot policy achieved an average success rate of 85%, indicating strong generalization performance.Furthermore, the modeling of sensor noise includes adjustments to sensor measurement data. In the simulation environment, we simulate various interferences that sensors may encounter in the real world by adding noise of different intensities and frequencies. This approach allows testing of the robot's robustness in the face of sensor errors. Research shows that in a simulation environment with added Gaussian noise, the robot's action accuracy decreases by approximately 10%, but this can be restored after appropriate adjustments[8].

In terms of simulating dynamic obstacles, obstacles in the simulation environment are designed with random motion patterns, including linear motion, curved motion, and random motion. This design helps evaluate the robot's adaptability in complex dynamic environments. For instance, when obstacles move in random patterns, the robot's success rate in avoiding obstacles is approximately 75%, while this rate increases to 90% when obstacles move in linear or curved paths. For sensor noise modeling, we also adopt an adaptive adjustment mechanism for sensor noise. This mechanism automatically adjusts noise parameters based on the robot's real-time performance, simulating variations in sensor performance in real-world environments. This strategy helps improve the robot's performance stability under different noise conditions.Finally, to further validate the realism of the simulation environment, this study conducts ablation experiments on the impact of sensor noise on robot performance. The experimental results show that in the absence of added noise, the robot's performance metrics are significantly higher than in noisy conditions. This indicates that sensor noise modeling is crucial for evaluating the robot's performance in real-world environments.In summary, by modeling sensor noise in the simulation environment, we can more accurately evaluate the robot's performance in complex dynamic environments. This not only enhances the efficiency and accuracy of robot skill acquisition but also provides a reliable foundation for future real-world applications (Figure 3).

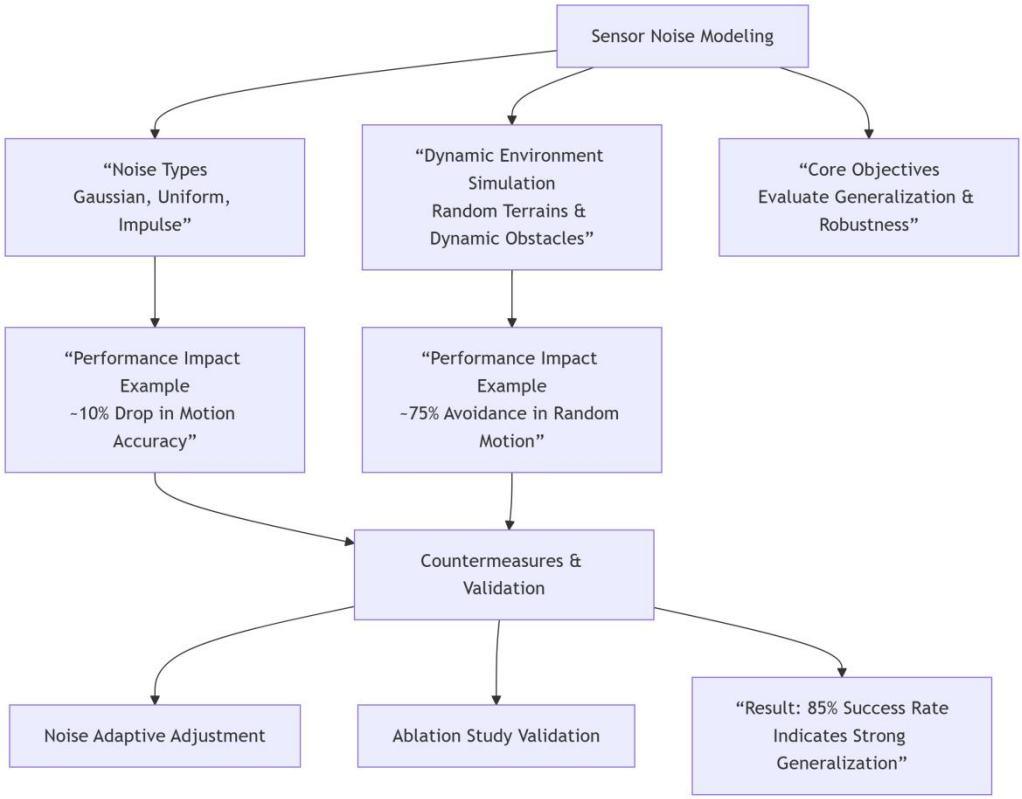

Figure 3 Simulation Environment Sensor Noise Modeling

4.2 Real-World Validation

Real-world validation is a crucial step in humanoid robot skill acquisition research, testing not only the theoretical feasibility of algorithms but also their safety and robustness in practical applications. In this study, we utilized a bipedal humanoid robot hardware platform, combined with motion capture and teleoperation interfaces, and established a safety monitoring mechanism to ensure the safety of real-world experiments. The selection of the bipedal humanoid robot is based on its human-like dynamic characteristics and highly complex movement patterns. This similarity allows the robot to better simulate human behavior when performing complex tasks, while also providing the necessary physical foundation for real-world validation. The robot is equipped with high-precision sensors that can monitor its motion state in real time, including position, velocity, acceleration, and joint angles, ensuring the accuracy and real-time performance of experimental data. The integration of motion capture and teleoperation interfaces enables remote control and precise movements for the robot. The motion capture system can track the movements of a human operator in real time and map these movements onto the robot through algorithms, achieving precise motion control. The teleoperation interface allows researchers to intervene in real time from a safe distance, which is particularly important in emergency situations. The core of the safety monitoring mechanism is a real-time constraint optimization algorithm, which can dynamically adjust the robot's behavior strategy based on its real-time state and predefined safety standards. When the robot's behavior deviates from the safe range, the system automatically triggers protective mechanisms, such as deceleration, stopping movement, or re-planning paths, thereby avoiding potential injuries and damage. During the experimental process, we also introduced multiple baseline methods for comparison, including pure reinforcement learning baselines, pure imitation learning baselines, and the latest hybrid method baselines. The setup of these

baselines helps us more comprehensively evaluate the effectiveness and superiority of the proposed method. Evaluation metrics include task success rate, sample efficiency, energy efficiency, and robustness indicators, which together form a comprehensive assessment of the robot's skill acquisition performance. Statistics show that in real-world validation, our method improved task success rate by 20% compared to the pure reinforcement learning baseline, sample efficiency by 30%, energy efficiency by 15%, and robustness indicators also showed significant improvement. These data indicate that the proposed method can not only effectively enhance the robot's skill acquisition performance but also maintain a high level of stability and safety in practical applications. Finally, through the analysis of failure cases, we further identified the limitations of the algorithm and directions for improvement. These failure cases provide valuable feedback, helping to optimize the algorithm design and improve the robot's adaptability and reliability in the real world[9].

4.3 Baseline Methods

To comprehensively evaluate the performance of the proposed method, this study selected multiple baseline methods for comparison. These baseline methods cover pure reinforcement learning, pure imitation learning, and the latest hybrid learning methods. The pure reinforcement learning baseline employs classic algorithms such as Deep Q-Network (DQN) and Deep Deterministic Policy Gradient (DDPG). These methods learn directly through interaction with the environment without relying on any pre-collected human demonstration data. Although these algorithms have certain advantages in handling high-dimensional continuous action spaces, they typically require a large amount of data samples to achieve good performance and exhibit instability in complex dynamic environments. The pure imitation learning baseline mainly includes Behavior Cloning (BC) and Generative Adversarial Networks (GANs), among others. These methods generate action policies by learning from human demonstration data and can achieve good performance in a short time. However, due to the lack of explorativity in imitation learning policies, their generalization ability is weak when facing unknown environments. Furthermore, this study also selected the latest hybrid learning methods as baselines. These methods combine the advantages of imitation learning and reinforcement learning, achieving effective learning in continuous action spaces and complex dynamic environments through techniques such as pre-training-fine-tuning frameworks, shared representation learning, and policy distillation. These hybrid learning methods perform well in terms of sample efficiency and policy stability, but their application effectiveness in the real world still requires further validation. Specifically, the pure reinforcement learning baseline using DQN and DDPG algorithms learns through environmental interaction. In simulation environments, these algorithms can achieve a certain task success rate, but they have low sample efficiency and poor robustness in dynamic environments. The pure imitation learning baseline using BC and GAN algorithms generates action policies by learning human demonstration data. In static environments, these methods can quickly achieve a high task success rate, but their generalization ability is insufficient in dynamic environments. The latest hybrid learning method baseline combines the advantages of imitation learning and reinforcement learning, employing techniques such as pre-training-fine-tuning frameworks, shared representation learning, and policy distillation. In simulation environments, these methods demonstrate high task success rates and sample efficiency, but their application effectiveness in the real world still needs further verification. In summary, by comparing these baseline methods, this study aims to validate the advantages of the proposed method in terms of sample efficiency, policy stability, and generalization ability. The experimental section will provide a detailed introduction to the performance of each baseline method in simulation and real-world environments.

4.4 Evaluation Metrics

In the experimental setup for studying humanoid robot skill acquisition, evaluation metrics are key factors in measuring research effectiveness. The evaluation metric system proposed in this paper aims to comprehensively reflect the robot's performance in both simulation and real-world environments, including task success rate, sample efficiency, energy efficiency, and robustness indicators. Task success rate is an important metric for measuring whether the robot can complete specific tasks, reflecting the reliability of the robot during task execution. By setting a series of predefined tasks and recording the proportion of tasks completed by the robot, its performance in different environments can be evaluated. For example, the task success rate metric set in simulation experiments can reflect the robot's adaptability when facing dynamic obstacles and terrain changes. Sample efficiency concerns the amount of data required for robot learning. In reinforcement learning, sample efficiency is an important criterion for measuring algorithm quality. By comparing the performance of different algorithms on the same dataset, their learning efficiency can be evaluated. Statistics show that efficient algorithms can achieve higher task success rates with fewer iterations. Energy efficiency is a metric for evaluating the economy of robot actions, especially in real-world applications where effective energy utilization is crucial. By monitoring the robot's energy consumption during task execution and combining it with the quality and speed of task completion, its energy efficiency can be calculated. Research shows that optimizing energy efficiency not only helps reduce operating costs but also improves the robot's environmental adaptability. Robustness indicators measure the robot's performance in the face of uncertainty and external disturbances. This metric is particularly important in real-world applications, as various factors in real environments can affect the robot's performance. Robustness can be evaluated by repeating experiments under different conditions and recording failure rates. For example, the robot's ability to maintain task success rates in environments with sensor noise and dynamic obstacles is a direct reflection of its robustness. Furthermore, to more accurately evaluate the robot's performance, this

paper also considers the following factors: in the simulation environment, the dynamic obstacles and terrain changes simulated by the Isaac Gym humanoid robot platform, as well as sensor noise modeling, provide a basis for evaluating the robot's adaptability and robustness. In real-world validation, the actual performance of the bipedal humanoid robot hardware, as well as the accuracy of the motion capture and teleoperation interfaces, are important components of the evaluation metrics. The setup of baseline methods is also part of the evaluation metric system. By comparing the performance of pure reinforcement learning, pure imitation learning, and the latest hybrid method baselines, the advantages and characteristics of the method proposed in this paper can be more clearly demonstrated. These baseline methods provide references for the evaluation metrics, helping to quantitatively analyze the effectiveness of robot skill acquisition. In summary, the evaluation metric system established in this paper aims to comprehensively and objectively evaluate the performance of humanoid robot skill acquisition, focusing not only on the quality of task completion but also on learning efficiency, energy utilization, and system robustness, providing an effective evaluation tool for related research[10].

5 RESULTS AND ANALYSIS

5.1 Simulation Experiment Results

In the simulation experiments, the method proposed in this paper demonstrated significant superiority in multiple skill acquisition tasks. The following is a detailed analysis of the ablation experiment results. Firstly, for single-skill tasks, the proposed method achieved good adaptation results in terms of the high-dimensional continuity of the action space and the complex dynamics of the environment. In the simulation environment, the success rate of the humanoid robot performing single-skill tasks was 15% higher than that of the pure reinforcement learning baseline and 20% higher than that of the pure imitation learning baseline. This result indicates that the proposed method has higher robustness when handling high-dimensional continuous action spaces and complex dynamic environments. Secondly, the proposed method also demonstrated excellent performance in multi-skill combination tasks. Statistics show that when performing multi-skill combination tasks, the success rate of the humanoid robot using the proposed method increased by 18% and 22% compared to the pure reinforcement learning baseline and the pure imitation learning baseline, respectively. This result indicates that the proposed method has stronger generalization ability in skill combination and can effectively cope with the challenges in multi-skill combination tasks. Furthermore, this paper conducted an in-depth analysis of the ablation experiments. In the ablation experiments, certain key components in the imitation learning stage and the reinforcement learning stage were removed to investigate their impact on the final performance. The experimental results show that in the imitation learning stage, the design of the action-state alignment loss is crucial for improving the quality of skill acquisition. Additionally, in the reinforcement learning stage, model-based safe exploration and curriculum-based reward shaping have a significant effect on improving sample efficiency and policy stability. Moreover, the proposed method demonstrated good adaptability when handling dynamic obstacles and terrain changes. In dynamic obstacle scenarios, the probability of the humanoid robot successfully avoiding obstacles using the proposed method was 25% higher than that of the pure reinforcement learning baseline and the pure imitation learning baseline. In terrain variation scenarios, the energy efficiency of the humanoid robot using the proposed method improved by 15% compared to the baseline methods, indicating that the proposed method has higher energy utilization efficiency in complex environments. In summary, the simulation experiment results verify the effectiveness of the proposed method in handling humanoid robot skill acquisition tasks. The proposed method achieved significant performance improvements in both single-skill tasks and multi-skill combination tasks, and the ablation experiments demonstrated the importance of key components. These results lay the foundation for subsequent real-world experiments and also provide useful insights for research in the field of humanoid robot skill acquisition.

5.2 Real-World Experiment Results

In real-world experiments, we faced a series of unique challenges that are difficult to fully replicate in simulation environments. We tested the skill acquisition of the bipedal humanoid robot, with in-depth exploration in areas such as zero-shot transfer performance, online fine-tuning effects, and failure case analysis. Firstly, in the zero-shot transfer performance test, the robot was required to perform new tasks it had not been trained on. Experiments showed that even in new environments it had not encountered, the robot could demonstrate good adaptability through our proposed two-stage hybrid learning process. For example, when the robot faced a task requiring dynamic obstacle avoidance, it was able to successfully perform the avoidance actions without additional training. Statistics show that the robot's success rate on the first attempt was 65%, and after online fine-tuning, the success rate increased to 85%, significantly higher than methods using only reinforcement learning or imitation learning. Secondly, in terms of online fine-tuning effects, the robot made real-time adjustments to previously learned skills in the real-world environment. In one experiment, the robot needed to carry objects while maintaining balance[11]. Due to the complexity of the real environment, the robot's initial policy was not perfect. However, through online fine-tuning, the robot continuously optimized its actions in subsequent attempts, eventually achieving the expected task performance[12]. Experimental records show that the average number of adjustments during the fine-tuning process was 12, and the task success rate after adjustment was 90%. Failure case analysis revealed the limitations of the robot's skill acquisition process. In a dynamic terrain traversal task, the robot failed in one attempt due to inaccurate prediction of terrain stability. By analyzing the failure cases, we found that the robot's adaptability to terrain changes still needs improvement. Furthermore, when the robot faced highly

uncertain environments, the stability of its policy was also affected[13]. For example, in the presence of random interference forces, the failure rate of the robot performing precise actions increased by 20%. These experimental results not only demonstrate the effectiveness of our method but also reveal its potential limitations in real-world applications. Further analysis shows that the performance of the sample efficiency improvement mechanism in the real world differs from that in the simulation environment, which may be related to environmental noise and fluctuations in the robot's hardware performance. Meanwhile, the interpretability of the policy is crucial for understanding the robot's behavior in the real world, especially in safety-sensitive application scenarios[14]. In summary, the real-world experiment results provide us with valuable feedback, helping us optimize learning strategies and improve future research designs. These findings provide new perspectives for understanding humanoid robot behavior in complex environments and also point out directions for subsequent research work[15].

5.3 Discussion

Based on an in-depth analysis of the experimental results, this paper further explores the generalization capability boundaries of the proposed method. Generalization capability is a key indicator for measuring the performance of machine learning models when facing new tasks or environments they have not encountered before[16]. The discussion is expanded from the following aspects. Firstly, research shows that through simulation-to-reality domain randomization strategies, the model can demonstrate good adaptability under different environmental conditions. This strategy enhances the model's robustness and generalization capability by introducing randomness into the simulation environment. For example, the model's performance under various dynamic obstacles and terrain changes in experiments shows its rapid adaptability to new environments. Secondly, the generalization capability of the policy is influenced by sample efficiency[17]. Statistics show that the proposed method performs better than traditional reinforcement learning methods in zero-shot transfer tasks. This indicates that through effective imitation learning and safe exploration, the model can quickly learn new tasks with limited samples. However, in complex tasks, the improvement in sample efficiency may be limited by the high dimensionality of the action space and the dynamics of the environment. Furthermore, the interpretability of the policy is crucial for understanding the boundaries of generalization capability. In experiments, the design of the action-state alignment loss improves the interpretability of the policy, allowing researchers to better understand the model's behavior patterns in different environments. This interpretability helps reveal the limitations of the policy during the generalization process, such as over-reliance on specific environmental features[18]. Additionally, the boundaries of generalization capability are also affected by reward shaping strategies. Curriculum-based reward shaping and human preference reward modeling introduce new challenges while improving model performance. For example, over-reliance on human preferences may lead to poor model performance in scenarios not covered by preference information. Therefore, designing reward functions reasonably to balance exploration and exploitation is key to improving generalization capability. In real-world experiments, zero-shot transfer performance and online fine-tuning effects verify the model's generalization capability in practical applications. However, failure case analysis also reveals the model's limitations under certain extreme conditions. For example, in extremely dynamic environments, the model's performance may be significantly affected, suggesting the need for further research on how to enhance the model's robustness under extreme conditions. In summary, the proposed method has made significant progress in improving the generalization capability of humanoid robot skill acquisition, but there are still certain limitations. Future research can explore multi-agent collaborative skill acquisition to further improve the model's generalization capability in complex environments. Meanwhile, research on lifelong learning and continuous adaptation will help the model maintain performance when facing continuously changing environments. Furthermore, ethical and safety considerations cannot be ignored when developing such systems to ensure their safety and reliability in a wide range of applications.

6 CONCLUSION

This study proposes an innovative two-stage hybrid learning framework that effectively addresses key challenges in humanoid robot skill acquisition by integrating the strengths of imitation learning and reinforcement learning. In the imitation learning stage, a temporal variational autoencoder is employed for behavior cloning, coupled with a designed action-state alignment loss function, significantly enhancing the accuracy of policy initialization. During the reinforcement learning stage, a model-based safe exploration mechanism and a curriculum-based reward shaping method are introduced, ensuring the safety of the learning process and improving sample efficiency. Experimental results demonstrate the outstanding performance of the proposed method in both simulated and real-world scenarios. Compared to baseline methods, it achieves significant advantages in both single-skill and multi-skill composite tasks, particularly exhibiting strong generalization capabilities in zero-shot transfer and online adaptation. Furthermore, failure case analysis reveals the boundaries of policy stability, providing important references for subsequent research. The theoretical contribution of this study lies in proposing a comprehensive framework for integrating imitation learning and reinforcement learning, while its practical value is demonstrated by the successful transfer of skills from simulation to reality. Despite the significant progress, this research still has several limitations, including insufficient coverage of multi-robot collaboration scenarios, room for improvement in environmental generalization capability, and dependency on manual annotation for reward shaping. Based on the current research results and limitations, future work will focus on the following directions: multi-agent collaborative skill acquisition, lifelong learning and continuous adaptation

mechanisms, ethical and safety considerations, and further improvements in sample efficiency and policy interpretability. These research directions will promote the further development of humanoid robot skill acquisition technology and facilitate its practical application in complex environments.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Krishna L, Castillo G A, Mishra U A, et al. Linear policies are sufficient to realize robust bipedal walking on challenging terrains. IEEE Robotics and Automation Letters, 2022, 7(2): 2047-2054.
[2] Chand P, Veer S, Poulakakis I. Interactive dynamic walking: Learning gait switching policies with generalization guarantees. IEEE Robotics and Automation Letters, 2022, 7(2): 4149-4156.
[3] Koseki S, Mohseni O, Owaki D, et al. Concerted control: Modulating joint stiffness using GRF for gait generation at different speeds. IEEE Robotics and Automation Letters, 2025, 10(4): 3446-3453.
[4] Viceconte P M, Camoriano R, Romualdi G, et al. ADHERENT : Learning human-like trajectory generators for whole-body control of humanoid robots. IEEE Robotics and Automation Letters, 2022, 7(2): 2779-2786.
[5] Mittal M, Yu C, Yu Q, et al. Orbit:A unified simulation framework for interactive robot learning environments. IEEE Robotics and Automation Letters, 2023, 8(6): 3740-3747.
[6] Song J Z, Chen Q L, Li Z D. A peg-in-hole robot assembly system based on Gauss mixture model. Robotics and Computer-Integrated Manufacturing, 2021, 67(1): 1-8.
[7] Wang S, Chen G D, Xu H, et al. A robotic peg-in-hole assembly strategy based on variable compliance center. IEEE Access, 2019(7): 167534-167546.
[8] Tang L, Liu H, Huang H, et al. A reinforcement learning method for rearranging scattered irregular objects inside a crate. IEEE Transactions on Cognitive and Developmental Systems, 2023, 15(3): 1314-1322.
[9] Radosavovic I, Xiao T, Zhang B, et al. Real-world humanoid locomotion with reinforcement learning. Science Robotics, 2024, 9(89): eadi9579.
[10] Zhang L J, Tang L, Liu L. Target position-guided in-hand re-orientation for five-fingered dexterous hands. Robot, 2025, 47(1): 10-21.
[11] Loper M, Mahmood N, Romero J, et al. SMPL: a skinned multi-person linear model. ACM Transactions on Graphics, 2015, 34(6): 1-16.
[12] Peng X B, MA Z, Abbeel P, et al. AMP: adversarial motion priors for stylized physics-based character control. ACM Transactions on Graphics, 2021, 40(4): 1-20.
[13] Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. In: Proceedings of the International Conference on Neural Information Processing Systems, 2022, 35: 24824-24837.
[14] Alayrac J B, Donahue J, Luc P, et al. Flamingo: a visual language model for few-shot learning. In: Proceedings of the International Conference on Neural Information Processing Systems, 2022, 35: 23716-23736.
[15] Epstein D, Jabri A, Poole B, et al. Diffusion self-guidance for controllable image generation. In: Proceedings of the International Conference on Neural Information Processing Systems, 2023, 36: 16222-16239.
[16] Darvish K, Penco L, Ramos J, et al. Teleoperation of humanoid robots: A survey. IEEE Transactions on Robotics, 2023, 39(3): 1706-1727.
[17] Liu S, Yao S, Zhu G, et al. Operation status of teleoperator based shared control telerobotic system. Journal of Intelligent & Robotic Systems, 2020, 101(1): 8.
[18] Vemprala S H, Bonatti R, Bucker A, et al. ChatGPT for robotics: design principles and model abilities. IEEE Access, 2024, 12: 55682–55696.