Multi-granularity Knowledge Transfer for Continual Reinforcement Learning

Chaofan Pan[1], **Lingfei Ren**[1], **Yihui Feng**[1], **Linbo Xiong**[1],
Wei Wei[2], **Yonghao Li**[1] and **Xin Yang**[1]
[1]Southwestern University of Finance and Economics
[2]Shanxi University
pan.chaofan@foxmail.com, renlf@swufe.edu.cn, yihuifeng@foxmail.com,
224081200021@smail.swufe.edu.cn, weiwei@sxu.edu.cn, liyonghao@swufe.edu.cn,
yangxin@swufe.edu.cn

Abstract

Continual reinforcement learning (CRL) empowers RL agents with the ability to learn a sequence of tasks, accumulating knowledge learned in the past and using the knowledge for problem-solving or future task learning. However, existing methods often focus on transferring fine-grained knowledge across similar tasks, which neglects the multi-granularity structure of human cognitive control, resulting in insufficient knowledge transfer across diverse tasks. To enhance coarse-grained knowledge transfer, we propose a novel framework called MT-Core (as shorthand for **M**ulti-granularity knowledge **T**ransfer for **Co**ntinual **re**inforcement learning). MT-Core has a key characteristic of multi-granularity policy learning: 1) a coarse-grained policy formulation for utilizing the powerful reasoning ability of the large language model (LLM) to set goals, and 2) a fine-grained policy learning through RL which is oriented by the goals. We also construct a new policy library (knowledge base) to store policies that can be retrieved for multi-granularity knowledge transfer. Experimental results demonstrate the superiority of the proposed MT-Core in handling diverse CRL tasks versus popular baselines.

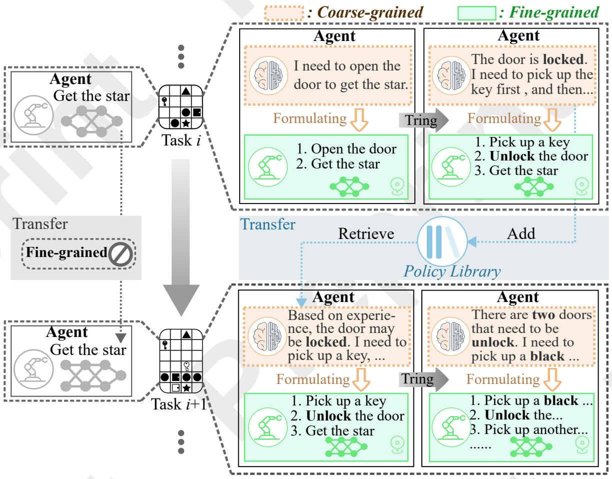


Figure 1: A simple illustration of our idea. The text has been simplified and modified for a better understanding. Traditional RL agents (**left**) primarily focus on fine-grained knowledge transfer, such as sharing the policy network, which may be ineffective across diverse tasks, particularly when the tasks have disparate state spaces. In contrast, our agent (**right**) leverages coarse-grained knowledge transfer within a multi-granularity structure to improve adaptability.

1 Introduction

Reinforcement learning (RL) is a powerful paradigm in artificial intelligence (AI) that enables agents to learn optimal behaviors through interactions with environments. The capacity to continuously adapt and learn from new tasks without starting from scratch is a defining characteristic of human learning [Kudithipudi *et al.*, 2022] and a desirable trait for RL agents deployed in the dynamic and unpredictable real world. To achieve this, the research of continual reinforcement learning (CRL, a.k.a. lifelong reinforcement learning) has emerged [Rolnick *et al.*, 2019; Kessler *et al.*, 2022]. It extends traditional RL by empowering agents with the ability to learn from a sequence of tasks, preserving knowledge from previous tasks, and using this knowledge to enhance learning efficiency and performance on future tasks. CRL aims to emulate the human capacity for lifelong learning and represents an ambitious field of research that tackles the challenges of long-term, real-world applications by addressing problems of diversity and non-stationarity [Khetarpal *et al.*, 2022].

Despite progress in CRL, the field continues to face a critical challenge: insufficient transfer of knowledge across diverse tasks. Sufficient knowledge transfer is essential for a robust CRL system, yet many existing methods struggle to achieve this, leading to suboptimal performance and a failure to fully exploit the benefits of continual learning (CL) [Wang *et al.*, 2024c; Li *et al.*, 2024]. A primary reason for this shortfall is the neglect of the sophisticated multi-granularity (hierarchical) structure of human cognitive control, which is primarily located in the prefrontal cortex and is complemented by the cerebellum for precise action control, enabling coarse-grained cognition for envisioning the future and planning [D'Mello *et al.*, 2020; Friedman and Robbins, 2022]. This structure enhances humans' ability to tackle complex tasks, and crucially, transfer skills and coarse-grained knowledge across significantly diverse tasks. In contrast, existing CRL methods often focus on transferring fine-grained knowledge, such as sharing the policy network (left side of Figure 1). This limits them to sequences of highly similar tasks with slight variations only in goals or parameters [Kessler *et al.*, 2022; Gaya *et al.*, 2023], and results in ineffective transfer.

To take advantage of abstract similarities between diverse tasks, it is essential to represent coarse-grained knowledge effectively and then employ robust models capable of extracting and transferring this knowledge. Human language, with its inherent abstraction, is a natural fit for representing coarse-grained knowledge. Furthermore, recent breakthroughs in large-scale language research have demonstrated the impressive reasoning and in-context learning abilities of the large language model (LLM) [Zhao *et al.*, 2023; Xi *et al.*, 2025; Wang *et al.*, 2024b]. Based on these insights, we propose a new framework named MT-Core (**M**ulti-granularity knowledge **T**ransfer for **Co**ntinual **re**inforcement learning), which integrates the powerful capabilities of an LLM into the CRL paradigm to enhance coarse-grained knowledge transfer and eventually improve learning performance.

The right side of Figure 1 illustrates the idea of MT-Core. MT-Core is structured in two layers: the coarse-grained policy formulation and the fine-grained policy learning. At the coarse-grained, MT-Core utilizes the powerful reasoning ability of a LLM to formulate a coarse-grained policy, which is represented as a temporally extended sequence of goals. Each goal consists of a descriptive component and an intrinsic reward function that outlines the intermediate states the agent needs to reach. The coarse-grained policy is evaluated based on feedback from the fine-grained learning process, allowing the LLM to refine its policy. Following this process, the coarse-grained policy along with the corresponding information is stored in a policy library. Upon encountering a new task, MT-Core can retrieve this policy from the library as a context of the LLM for enhancing coarse-grained knowledge transfer. At the fine-grained, MT-Core employs goal-oriented RL to learn a policy that is guided by the formulated coarse-grained policy. The verified policy can be stored in the policy library to reduce the risk of catastrophic forgetting when encountering old tasks. Specifically, the coarse-grained policy is represented as *text*, while the fine-grained policy is represented as the policy *network*. The experimental results demonstrate that MT-Core significantly outperforms popular baselines. The main contributions of the work are as follows:

- We investigate the knowledge transfer problem in CRL and find that the multi-granularity structure of human cognitive control is crucial for effective knowledge transfer across diverse tasks.
- We proposed MT-Core, a novel CRL framework that leverages multi-granularity knowledge transfer, which is the first work to integrate the powerful reasoning ability of LLM into the CRL paradigm, facilitating knowledge transfer across diverse tasks.
- Extensive experiments in MiniGrid provide empirical evidence of MT-Core's effectiveness.

2 Background

2.1 Preliminaries

The continual reinforcement learning process can be formulated as a series of related Markov decision processes (MDPs) $\{< \mathcal{S}^i, \mathcal{A}^i, P^i, R^i >\}$. Each MDP represents a different task or problem instance that an agent needs to solve over its lifetime. Here, $\mathcal{S}^i$ and $\mathcal{A}^i$ denote the state and action space of task i, respectively, while $P^i : \mathcal{S}^i \times \mathcal{S}^i \times \mathcal{A}^i \to [0,1]$ is the transition probability function, and $R^i : \mathcal{S}^i \times \mathcal{A}^i \to [r^{\min}, r^{\max}]$ is the reward function. At each time step, the learning agent perceives the current state $s_t^i \in \mathcal{S}^i$ and selects an action $a_t^i \in \mathcal{A}^i$ according to its policy $\pi_\theta : \mathcal{S}^i \times \mathcal{A}^i \to [0,1]$ with parameters θ. The agent then transitions to the next state $s_{t+1}^i \sim P^i(s_t^i, a_t^i)$ and receives a reward $r_t^i = R^i(s_t^i, a_t^i, s_{t+1}^i)$. The target of an agent on the task i is to maximize the expected return $\mathbb{E}\left[\sum_{t=0}^{H} \gamma^t R^i(s_t^i, a_t^i, s_{t+1}^i)\right]$, where γ is the discount factor, and H is the horizon.

2.2 Related Works

Continual Reinforcement Learning

CRL focuses on training RL agents to learn multiple tasks sequentially without prior knowledge, garnering significant interest due to its relevance to real-world AI applications [Khetarpal *et al.*, 2022]. A central issue in CRL is catastrophic forgetting, which has led to various strategies for knowledge retention. PackNet and related pruning methods [Mallya and Lazebnik, 2018; Schwarz *et al.*, 2021] preserve model parameters but often require knowledge of task count. Experience replay techniques such as CLEAR [Rolnick *et al.*, 2019] use buffers to retain past experiences, but face memory scalability challenges. In addition, some methods prevent forgetting by maintaining multiple policies or a subspace of policies [Schöpf *et al.*, 2022; Gaya *et al.*, 2023]. Furthermore, task-agnostic CRL research indicates that rapid adaptation can also help prevent forgetting [Caccia *et al.*, 2023].

Another issue in CRL is transfer learning, which is crucial for efficient policy adaptation. Naive approaches, like fine-tuning, that train a single model on each new task provide good scalability and transferability but suffer from catastrophic forgetting. Regularization-based methods, such as EWC [Kirkpatrick *et al.*, 2017; Wang *et al.*, 2024c], have been proposed to prevent this side effect, but often reduce plasticity. Some architectural innovations have been proposed to balance the trade-off between plasticity and stability [Rusu *et al.*, 2016; Berseth *et al.*, 2022]. Furthermore, methods like OWL [Kessler *et al.*, 2022] and MAXQINIT [Abel *et al.*, 2018] leverage policy factorization and value function transfer, respectively, for improved learning.

Most existing methods perform well when applied to sequences of tasks that exhibit high environmental similarity, such as tasks where only specific parameters within the environment are altered or the objectives within the same environment are different. However, their effectiveness is greatly diminished when dealing with a sequence of diverse tasks. Our proposed framework aims to overcome this limitation by leveraging multi-granularity knowledge transfer.

Reinforcement Learning With the LLM

Recent advancements have combined the LLM with reinforcement learning to address reinforcement learning challenges such as sample efficiency and generalization. These integrations utilize LLM's vast knowledge and reasoning ability to improve agents' performance. For instance, studies have developed methods for efficient agent-LLM interactions [Hu *et al.*, 2024] and unified agent foundations with language models for better experience reuse [Palo *et al.*, 2023a].

Frameworks that merge LLMs with RL have also emerged, such as LAMP, which pretrains RL agents with Vision-Language Models [Adeniji *et al.*, 2023], and HiP, which uses foundation models for long-horizon tasks [Ajay *et al.*, 2023]. In addition, the RAFA framework offers a principled approach with guarantees to optimize reasoning and actions [Liu *et al.*, 2023]. Research into LLM-guided skill learning and reward shaping is also gaining traction. The BOSS method [Zhang *et al.*, 2023] and the work on LLM-based reward structures for robotics [Yu *et al.*, 2023] exemplify this trend, with studies like [Kwon *et al.*, 2023] focusing on LLM-driven reward design. Furthermore, VLA combines coarse-grained visual-language policies and fine-grained reinforcement learning policies, demonstrating significant zero-shot transfer performance in the real world [Yang *et al.*, 2024].

CRL diverges from the traditional RL paradigm by focusing on complex and dynamic environments where the agent constantly encounters a sequence of tasks. Unlike other research on LLM support for RL, which mainly deals with static tasks or single tasks, our work integrates LLM's powerful reasoning ability into the CRL paradigm for the first time. This distinction is crucial, as CRL poses unique challenges such as the necessity for high-level knowledge transfer across diverse tasks. Existing RL methods with LLM cannot address these challenges directly. Our framework's novelty lies in its utilization of the LLM and the policy library not only to generate high-level policies that guide the learning process but also to facilitate high-level knowledge transfer between tasks, thus improving the agent's adaptability.

3 Method

In this section, we first clarify the research problem and the overall framework. Then we elaborate on two layers of MT-Core: (1) the coarse-grained policy formulation, and (2) the fine-grained policy learning. Finally, we describe the policy library in MT-Core.

Problem Formalization. Our goal is to use the LLM's powerful reasoning ability to enhance the capability of CRL agents. We formalize the CRL problem in the goal-oriented RL framework [Chen and Luo, 2022; Colas *et al.*, 2022]. The agent perceives not only the current state of the environment but is also provided with a specific goal $g \in \mathcal{G}$, where $\mathcal{G}$ represents the space of possible goals. For task i and its corresponding goal sequence $\mathbf{g}^i = (g_1^i, \cdots, g_m^i)$, the agent's objective is to learn a policy $\pi^i(s^i)$ that effectively maps each state s^i to an appropriate action a^i. The aim of the agent is to maximize the expected cumulative reward over time, guided by the sum of the extrinsic reward $R^i(s^i, a^i, s'^i)$ and the intrinsic rewards $(R'^i(s^i, a^i, s'^i, g_1^i), \cdots, R'^i(s^i, a^i, s'^i, g_m^i))$ [Dilokthanakul *et al.*, 2019], where s'^i represents the next state in the task i. The former is determined by the environment and is usually sparse. The latter is typically defined so that the agent receives a reward if it achieves the given goal.

Overview. MT-Core is illustrated in Figure 2. For simplicity, we have omitted the task-indexed superscript in the description. The framework orchestrates the interaction between the CRL agent and the environment simulator for each task in the task sequence, aligning with standard RL protocols. For each task, the corresponding environment simulator provides the current state s_t to the agent at each timestep, which then selects an action a_t to perform, resulting in a reward r_t and the next state. Internally, the CRL agent is structured in two layers: the coarse-grained policy formulation and the fine-grained policy learning. Additionally, a policy library is maintained for the storage and retrieval of policies.

3.1 Coarse-Grained Policy Formulation

To integrate LLM into the goal-oriented RL framework, we conceptualize LLM as a function $\mathcal{F} : \mathcal{P} \times \mathcal{E} \to \mathcal{D}$, where $\mathcal{P}$ is the set of all prompts, $\mathcal{E}$ is the set of textual descriptions of the environment, and $\mathcal{D}$ is the set of textual descriptions of the sequence of goals. The description of the environment $e \in \mathcal{E}$ can be automatically obtained through a hard-coded reporter [Dasgupta *et al.*, 2022] or a vision-language model [Palo *et al.*, 2023b; Du *et al.*, 2023]. The prompt $p \in \mathcal{P}$ is a concatenation of two parts: p_{s} and p_{e}, where p_{s} is the system prompt and p_{e} is the experience retrieved from the policy library. Description of each coarse-grained policy, that is, textual output $d \in \mathcal{D}$ is mapped to the corresponding goal sequence $\mathbf{g} \in \mathcal{G}^*$ through a defined parsing function. In order to improve the stability of LLM's output, we structure the goal into a tuple: $g = (s_g, v_g, r_g)$, where $s_g \in \mathcal{E}$ is the textual description of the goal state, v_g is a verification function $v_g : \mathcal{E} \times \mathcal{S} \to \{0, 1\}$, $r_g \in \mathbb{R}$ is a intrinsic reward value of the goal. Optional validation functions are provided in p_{s}.

The intrinsic reward can not only guide the agent to learn to achieve the corresponding goal but also serve as a metric for evaluating whether the agent has achieved the goal. This metric is continuously monitored during the interactions of the CRL agent with the environment. When this average exceeds a predefined threshold, it is taken as an indication that the agent has successfully achieved the goals. Consequently, the LLM shifts its focus to the next goal in the sequence, and the process repeats. The iterative process of coarse-grained policy formulation, validation, and improvement draws inspiration from the concept of Reflexion [Shinn *et al.*, 2023]. In instances where the agent fails to meet the threshold, indicating that the goal has not been met, the LLM is prompted to adjust the coarse-grained policy. This adjustment is informed by feedback from the agent's performance. The LLM is first required to analyze the feedback to pinpoint the reasons behind the shortfall in the previous policy. With powerful reasoning ability, the LLM then formulates a new coarse-grained policy aimed at overcoming the identified challenges.

3.2 Fine-Grained Policy Learning

Fine-grained policy learning is guided by the formulated coarse-grained policy. The RL model learns the fine-grained

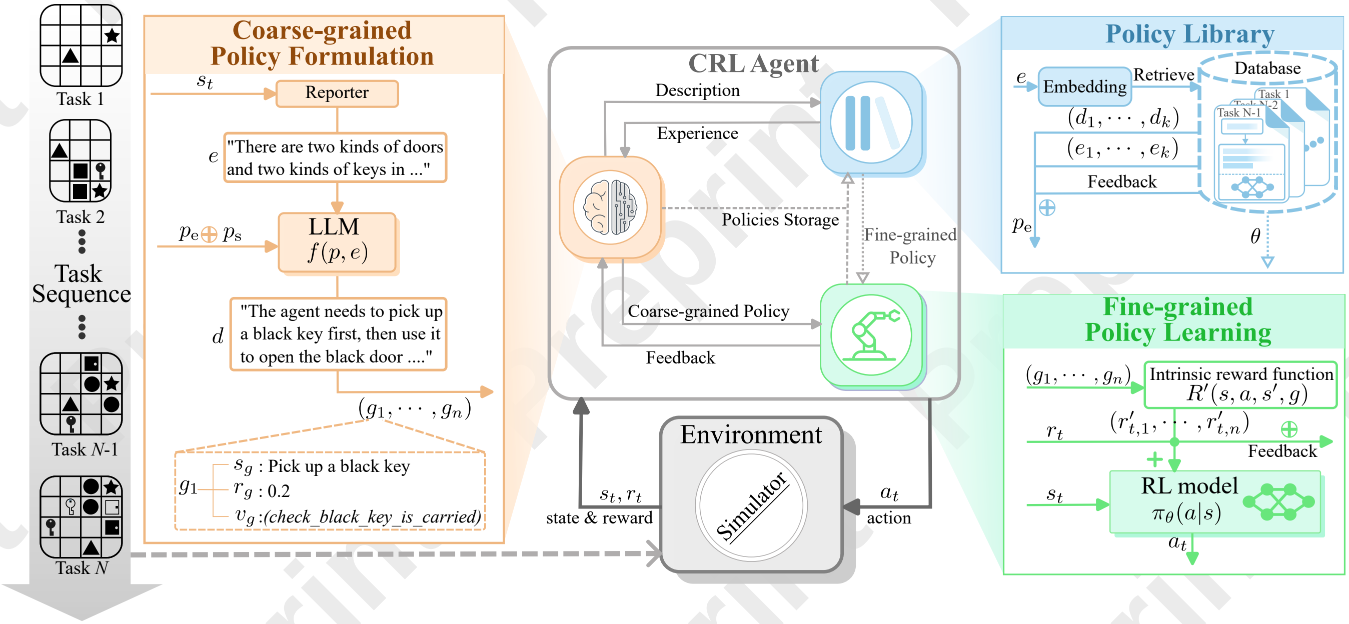


Figure 2: The illustration of the proposed framework. The middle section depicts the internal interactions (**light gray line**) and external interactions (**dark gray line**) in MT-Core. Internally, the agent is structured in two layers: the coarse-grained policy formulation (**orange**) and the fine-grained policy learning (**green**). Furthermore, the policy library (**blue**) is constructed to store and retrieve policies. The formulation of the coarse-grained policy is initiated when the agent interacts first with the environment. If the policy library is not empty, related experiences (including coarse-grained policies, corresponding environment descriptions and feedback) will be retrieved through the description of the environment to assist in the formulation. Once the coarse-grained policy is established, it guides the learning of the fine-grained policy. Through feedback from the RL learning process, the LLM gradually refines its coarse-grained policy to improve performance. Upon completion of a task, the successful policies are stored in the policy library.

policy using raw state input s_t and total intrinsic reward r'_t based on the goal sequence $\mathbf{g} = (g_1, \cdots, g_m)$. The agent's intrinsic reward function is defined as:

$$R'(s,a,s',g) = \begin{cases} r_g, & \text{if } v_g(s_g,s') = 1 \\ 0, & \text{otherwise} \end{cases}. \tag{1}$$

This reward structure provides a more granular view into the effectiveness of goal-related actions, encouraging the agent to learn a policy that is aligned with the goal more efficiently. Furthermore, the state-dependent design of this function improves the generalizability of coarse-grained policies. Then, the internal rewards for each goal $\{r'_{t,l} = R'(s_i, a_i, s_{i+1}, g_l)|l = 1, \cdots, m\}$ and the external reward r_t from the environment will be textual and will be concatenated as the feedback. The total intrinsic reward is $r'_t = \sum_{l=1}^{m} r'_{t,l}$. Therefore, the aim of RL agent with goal sequence g can be expressed as

$$max_\theta J(\theta, g) = \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{H} \gamma^t (r_t + r'_t)\right], \tag{2}$$

where θ is the parameter of fine-grained policy.

3.3 Policy Library

Recent studies [Wang *et al.*, 2024a; Zhang *et al.*, 2023] have highlighted the potential of LLMs' reasoning and planning ability in facilitating continual learning (CL) for agents, which involves continuous acquisition and update of skills [Wang *et al.*, 2024c]. Two core challenges in continual learning are catastrophic forgetting and knowledge transfer [Hadsell *et al.*, 2020; McCloskey and Cohen, 1989]. One method to address the these challenges is to design a knowledge base [Li *et al.*, 2023] or a skill library to store knowledge or skills[Wang *et al.*, 2024a]. By using validated successful experiences, agents can not only reduce catastrophic forgetting but also rapidly enhance their capabilities.

Based on the above findings, MT-Core maintains a policy library to store the policies that solve tasks successfully. The description of each coarse-grained policy d and the corresponding feedback is indexed by the embedding of the environment description e of its corresponding task, which can be retrieved in similar tasks in the future. The policy library can be continuously extended throughout the life of the CRL agent. Given a new environment description e, MT-Core can use embedding to retrieve the first k similar environment descriptions and concatenate them with goal descriptions and corresponding feedback as experience p_e. Specific policies for new tasks can be learned under the guidance of coarse-grained goals. Specifically, We adopt a vector similarity search approach, with cosine similarity as the metric. When the number of tasks is small, simple storage strategies can also be used for the policy library. In addition, for each environment description e, MT-Core also stores the parameter of the fine-grained policy θ that has been learned. This can accelerate learning on similar tasks and further reduce catastrophic forgetting if storage space is sufficient. Even if the storage space is strict, only storing coarse-grained policies (text) can perform knowledge transfer with small memory usage. By storing both coarse-grained and fine-grained learning policies in the policy library, the agent's capabilities for diverse tasks can be enhanced over time.

4 Experiments

In this section, we evaluate our framework in several continual reinforcement learning tasks. More details about the experimental implementation are provided in the supplementary material.

4.1 Environment

For our experiments, we utilized a suite of MiniGrid environments [Chevalier-Boisvert *et al.*, 2023] to evaluate the efficacy of MT-Core in addressing CRL tasks. These environments feature image-based state observations, a discrete set of possible actions, and various objects characterized by their color and type. We intentionally simplified the state and action spaces to streamline the policy learning process, enabling more rapid experimental iterations. This approach allowed us to focus on the core objective of this study: leveraging hierarchical knowledge transfer for CRL. We crafted a sequence of four distinct tasks within the MiniGrid framework. This sequence is arranged in ascending order of difficulty and aligns with the human learning process, facilitating better tracking of knowledge transfer throughout the learning process. A comprehensive account of these modified environments, including their specific configurations and visual representations, is available in supplementary material.

4.2 Experiment Setup

Baselines. We consider the following baselines:

- **Single-Task (SG)**: This baseline represents a traditional RL setup where a distinct agent is trained exclusively on each task. There is no sharing or transfer of knowledge between tasks in this method. This serves as a foundational comparison point to underscore the advantages of CL, as it lacks any mechanism for knowledge retention or transfer.
- **Fine-Tuning (FT)**: Building upon the standard RL algorithm, this baseline differs from SG by using a single agent that is sequentially fine-tuned across different tasks. The agent trained on the preceding task is adapted for the subsequent task, effectively using the prior model as a starting point. This baseline provides a basic measure of an agent's capacity to maintain knowledge of earlier tasks while encountering new tasks.
- **Fine-Tuning with L2 Regularization (FT-L2)**: Based on FT, this baseline incorporates L2 regularization during the fine-tuning process. The addition of L2 regularization aims to mitigate catastrophic forgetting by penalizing significant changes to the weights that are important for previous tasks. This helps in preserving performance on earlier tasks while learning new ones. FT-L2 can be viewed as the simplest implementation of regularization-based CRL methods [Kirkpatrick *et al.*, 2017; Wang *et al.*, 2024c].
- **PackNet** [Mallya and Lazebnik, 2018]: This baseline is a representative CRL method based on parameter isolation [Wang *et al.*, 2024c]. It utilizes the network pruning technique to efficiently allocate the neural network's capacity throughout the continual learning of tasks. After training on a task, PackNet identifies the crucial weights to retain, prunes the less essential ones, and thus creates room for the network to learn additional tasks. PackNet is a representative CRL method based on parameter isolation [Wang *et al.*, 2024c].

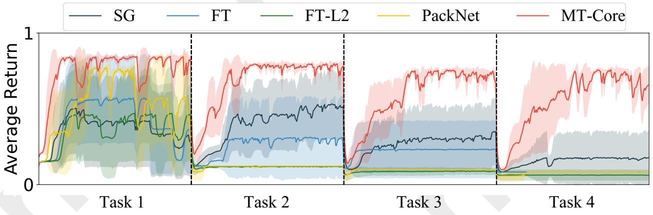


Figure 3: Performance during training across four tasks (smoothed with exponential moving average). The x-axis represents the training progress for each task. The y-axis represents the normalized average return of each task. The shaded area depicts the means ± the standard deviation over five random seeds.

Note that our method is not directly comparable with certain methods, such as CLEAR [Rolnick *et al.*, 2019] and ClonEx-SAC [Wolczyk *et al.*, 2022], as it requires access to the data from previous tasks for training. Our framework can continually learn without the need to revisit past task data, which is a common constraint in real-world applications.

Metrics. To evaluate the effectiveness of MT-Core, we use the normalized average return to measure the performance of the trained agents. Following standard practice in supervised continual learning literature [Wolczyk *et al.*, 2021; Li *et al.*, 2024], we define a suite of metrics based on the agent's performance throughout different phases of its training process. Based on the agent's normalized average return, we evaluated the continual learning performance of our framework and baselines using the following metrics: *average performance*, *forward transfer* and *forgetting*, as detailed in the supplemental material.

4.3 Competitive Results

To evaluate our framework, we compare MT-Core with other baselines across four challenging tasks that have the same state and action space. Each experiment is trained in 5M steps and replicated with five random seeds of environments to ensure reliability.

Overall Performance. As illustrated in Figure 3, the baselines perform well on the first Task. However, their performance gets worse progressively as the difficulty of the tasks increases. A common trend of performance degradation in the transition between tasks is evident across all methods.

Metric	SG	FT	FT-L2	PackNet	MT-Core
$A_N(\uparrow)$	0.30	0.25	0.16	0.20	**0.65**
$FW(\uparrow)$	0.00	-0.05	-0.14	-0.09	**0.17**
$FG(\downarrow)$	0.12	0.09	**0.05**	0.06	0.12

Table 1: CL performance of MT-Core and four baselines.

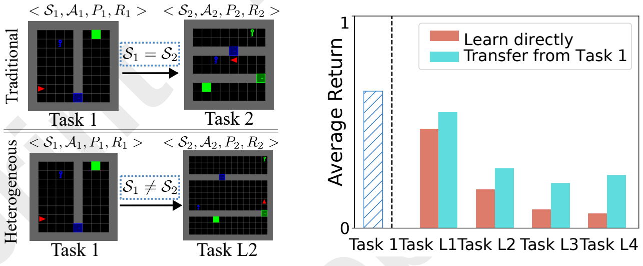
Specifically, the SG baseline, while effective on the first task, is unable to maintain its performance due to its inherent limitation. Other baselines, such as FT, FT-L2, and PackNet, also start strong on the first task but encounter difficulties in the subsequent tasks. This phenomenon suggests that a singular shared policy may not be robust enough to handle a sequence of highly varied tasks. The balance between policy stability and plasticity is critical, with higher stability often leading to reduced adaptability, as observed in the performance degradation of these baselines. In contrast, MT-Core consistently surpasses the other baselines, showcasing remarkable performance across all tasks. Although it is not immune to performance degradation when transitioning between tasks with significant differences, it demonstrates a rapid recovery in performance. This can be attributed to its multi-granularity structure, which leverages the powerful capabilities of LLM and promotes the transfer of coarse-grained policy knowledge, thereby enhancing the agent's ability to adapt to new and diverse tasks.

Continual Learning Performance. Table 1 shows the evaluation results in terms of CL metrics. The average performance metric clearly demonstrates MT-Core's superiority in task sequence with diverse tasks, significantly outperforming other baseline methods. The forward transfer metric is particularly important, as it measures the capacity of an agent to utilize knowledge from previous tasks. All baselines, with the exception of MT-Core, present forward transfer metrics of less than or equal to zero. In addition, the average performace of MT-Core is more than twice that of the best baseline. This suggests that traditional approaches struggle to transfer previously acquired knowledge to new tasks effectively. Baselines that rely on regularization strategies, such as FT-L2, or parameter isolation techniques, like PackNet, may even hinder the learning of new tasks.

In contrast, MT-Core exhibits positive transfer performance, underscoring the benefits of transferring coarse-grained policies. When examining the forgetting metrics, FT, FT-L2, and PackNet show relatively low scores. This can be attributed, in part, to their emphasis on model stability. However, a more critical factor is their inherently lower performance on subsequent tasks, such as Task 2, Task 3, and Task 4. Although MT-Core does not achieve the lower score in the forgetting metric, its exceptional forward transfer capability and strong average performance accentuate its proficiency in handling diverse task sequences. Note that MT-Core does not resort to fine-tuning strategy in experiments to focus on coarse-grained knowledge transfer, although it could be employed in situations where minimizing forgetting is a priority.

4.4 Heterogeneous Tasks Experiments

Most existing works have focused on the transfer ability of CRL agents in environments that have the same state space.



(a)

(b)

Figure 4: Heterogeneous tasks experiments. **(a)** A simple illustration. The setting of traditional CRL requires the same state spaces of the tasks (**above**). In our heterogeneous task experiments, the state space of each task may be different (**below**). **(b)** Experiment results. The left side of the dashed line represents the performance on the original Task 1. The right side of the dashed line represents the performances on tasks with larger state space.

In contrast, our study explores the application of CRL agents to heterogeneous tasks, where environments have different state spaces or action spaces. For the simplification of the problem, we concentrate on tasks with differing state spaces.

As depicted in Figure 4a, heterogeneous tasks present a unique challenge. While it is possible to accommodate differences in the size of the state space during knowledge transfer by employing a masking technique, this approach necessitates prior knowledge of the maximum state space size, which is impractical for RL agents expected to learn continually in dynamic environments. On the contrary, the coarse-grained policy transfer of MT-Core can be applied to heterogeneous tasks without the need for masking.

In our experiments, we evaluate the effectiveness of MT-Core across two heterogeneous tasks. The first task has a relatively small state space, as shown in the left below of Figure 4a. In contrast, the second task has a larger state space, such as the right below of the same figure. Figure 4b reports the results of these experiments. The first task is Task 1, while the second task is Task L1, Task L2, Task L3, or Task L4, each representing a larger state space variant of Task 1 through Task 4. Direct learning agents who are exposed to tasks with large state spaces from the outset may struggle to learn effective policies within the set number of steps, resulting in poor performance. This phenomenon is more obvious on more difficult tasks, such as Task L3 and Task L4. However, agents that are learned on Task 1 can transfer coarse-grained knowledge to subsequent tasks. Despite the increasing difficulty of these tasks, the agents still demonstrate comparable performance. This evidence suggests that MT-Core retains transfer ability even when applied to heterogeneous tasks.

4.5 Ablation Study

In this part, we conduct an ablation study to investigate what affects MT-Core's performance in CRL. We consider several variants of MT-Core to understand the effects:

- **MT-Core-FT**: MT-Core fine-tuning the policy from the previous task when encountering a new task.

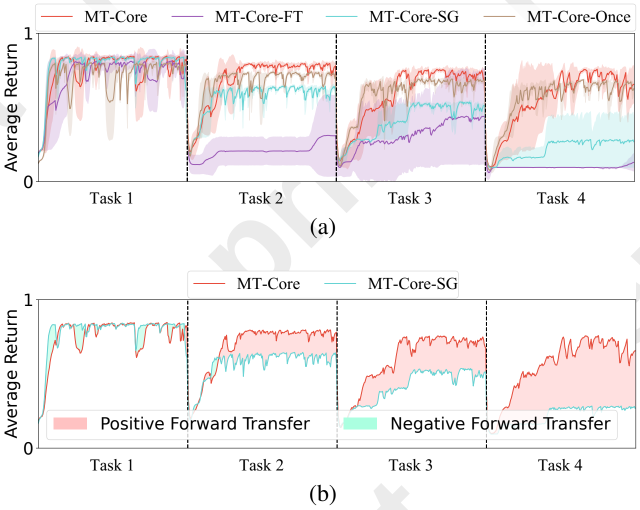


Figure 5: Ablation study results. **(a)** Performance of MT-Core and its variants during training. **(b)** Forward transfer performance of MT-Core during training. The red areas between the curves represent positive forward transfer and green represent negative forward transfer.

	$A_N(\uparrow)$	$FW(\uparrow)$	$FG(\downarrow)$
SG	0.30	0.00	0.10
MT-Core	**0.65**	**0.17**	0.12
MT-Core-FT	0.33	−0.14	**0.09**
MT-Core-SG	0.47	0.00	0.19
MT-Core-Once	0.62	0.15	0.13

Table 2: CL performance of the variants of MT-Core.

- **MT-Core-SG**: MT-Core without policy library. This variant treats each task as an isolated learning problem, akin to training single agents for each task.
- **MT-Core-Once**: MT-Core without feedback mechanisms. In this variant, the coarse-grained policy formulation of LLM is performed only once for each task.

The results of our study are presented in Table 2 and Figure 5. For better comparison, we also include the performance of the SG baseline in the Table. Our findings are as follows: Firstly, MT-Core-SG shows superior performance compared to the SG baseline, which confirms that the LLM's powerful capabilities can help RL. An RL agent can take advantage of reasoning ability in the LLM to significantly reduce the difficulties in solving these tasks. Secondly, while MT-Core-SG outperforms the baseline, it falls short of the average performance and forward transfer metrics achieved by the MT-Core. The gap between MT-Core with MT-Core-SG emphasizes the value of the policy library in facilitating knowledge transfer, ultimately boosting the agent's overall performance (see the figure in 5b). Furthermore, MT-Core-Once exhibits slightly lower performance on Task 2, Task 3 and Task 4 than MT-Core. This phenomenon suggests that feedback can improve the performance of LLM in dealing with these challenging tasks. Lastly, the MT-Core-FT variant underperforms on all tasks except Task 1. This observation aligns with the previous finding in competitive results that methods designed to enhance policy stability may decrease the agent's performance on subsequent tasks when those tasks differ significantly from one another.

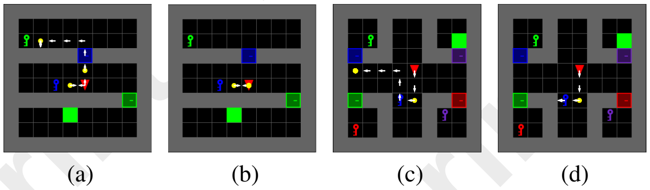


Figure 6: Early-stage policy visualization for agents on Task 2 (0.5M steps) and Task 4 (1M steps), following learning on Task 1 (5M steps). **(a) and (c)** depict the policies learned by MT-Core for Task 2 and Task 4 respectively, while **(b) and (d)** represent the policies by MT-Core-SG for the same tasks. The forward actions of agents are denoted by white arrows, while the yellow dots denote interactive actions such as pickup and toggle. For the clearness, steering actions are omitted from these visual representations.

Policy Visualization. Figure 6 further illustrates the difference between MT-Core and MT-Core-SG on Task 2 and Task 4. As shown in Figure 6a, the agent of MT-Core demonstrates the capability to swiftly learn to pick up the blue key and unlock the corresponding door. In contrast, the agent of MT-Core-SG struggles to progress beyond the initial room (Figure 6b). This performance gap can be attributed to the previous learning of the MT-Core agent on Task 1, which shares common elements with Task 2, such as the blue key and door. By drawing on previous experience, MT-Core formulate relevant initial goals, accelerating learning on Task 2. Similarly, when comparing Figure 6c and Figure 6d for Task 4, we observe a consistent pattern. Although Task 4 presents a higher level of difficulty, the agent with pre-acquired coarse-grained knowledge exhibits more effective behaviors than the agent learning from scratch.

5 Conclusion

In this work, we took a step in the direction of improving the generalization ability of CRL agents. We investigated the limitations of existing CRL methods in transferring fine-grained knowledge across diverse tasks and proposed MT-Core, a novel framework that aligns with the multi-granularity structure of human cognitive control, enhancing the coarse-grained knowledge transfer of CRL. MT-Core leverages a LLM for coarse-grained policy formulation, employs goal-oriented RL for fine-grained policy learning, and constructs the policy library to store and retrieve policies. It addresses the challenge of learning across a sequence of diverse tasks by transferring coarse-grained knowledge, which has been demonstrated in our experiments. Further results indicate that MT-Core can be applied to heterogeneous tasks and achieves comparable transfer performance.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Nos. 62476228), the Sichuan Science and Technology Program (Nos. 2024ZYD0180), and the Graduate Representative Achievement Cultivation Project of Southwest University of Finance and Economics (Nos. JGS2024068).

References

[Abel *et al.*, 2018] David Abel, Yuu Jinnai, Sophie Yue Guo, George Dimitri Konidaris, and Michael L. Littman. Policy and value transfer in lifelong reinforcement learning. In *ICML*, pages 20–29, 2018.

[Adeniji *et al.*, 2023] Ademi Adeniji, Amber Xie, Carmelo Sferrazza, Younggyo Seo, Stephen James, and Pieter Abbeel. Language reward modulation for pretraining reinforcement learning. *arXiv preprint arXiv:2308.12270*, 2023.

[Ajay *et al.*, 2023] Anurag Ajay, Seungwook Han, Yilun Du, Shuang Li, Abhi Gupta, Tommi S. Jaakkola, Joshua B. Tenenbaum, Leslie Pack Kaelbling, Akash Srivastava, and Pulkit Agrawal. Compositional foundation models for hierarchical planning. In *NeurIPS*, 2023.

[Berseth *et al.*, 2022] Glen Berseth, Zhiwei Zhang, Grace Zhang, Chelsea Finn, and Sergey Levine. CoMPS: Continual meta policy search. In *ICLR*, 2022.

[Caccia *et al.*, 2023] Massimo Caccia, Jonas Mueller, Taesup Kim, Laurent Charlin, and Rasool Fakoor. Task-Agnostic Continual Reinforcement Learning: Gaining Insights and Overcoming Challenges. In *CoLLAs*, 2023.

[Chen and Luo, 2022] Liyu Chen and Haipeng Luo. Near-optimal goal-oriented reinforcement learning in non-stationary environments. *Advances in Neural Information Processing Systems*, 35:33973–33984, 2022.

[Chevalier-Boisvert *et al.*, 2023] Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo De Lazcano Perez-Vicente, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and J K Terry. Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. In *NeurIPS Datasets and Benchmarks Track*, 2023.

[Colas *et al.*, 2022] Cédric Colas, Tristan Karch, Olivier Sigaud, and Pierre-Yves Oudeyer. Autotelic agents with intrinsically motivated goal-conditioned reinforcement learning: A short survey. *Journal of Artificial Intelligence Research*, 74:1159–1199, 2022.

[Dasgupta *et al.*, 2022] Ishita Dasgupta, Christine Kaeser-Chen, Kenneth Marino, Arun Ahuja, Sheila Babayan, Felix Hill, and Rob Fergus. Collaborating with language models for embodied reasoning. In *LaReL*, 2022.

[Dilokthanakul *et al.*, 2019] Nat Dilokthanakul, Christos Kaplanis, Nick Pawlowski, and Murray Shanahan. Feature control as intrinsic motivation for hierarchical reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 30(11):3409–3418, 2019.

[Du *et al.*, 2023] Yuqing Du, Olivia Watkins, Zihan Wang, Cédric Colas, Trevor Darrell, Pieter Abbeel, Abhishek Gupta, and Jacob Andreas. Guiding pretraining in reinforcement learning with large language models. In *ICML*, volume 202, pages 8657–8677, 2023.

[D'Mello *et al.*, 2020] Anila M. D'Mello, John D.E. Gabrieli, and Derek Evan Nee. Evidence for hierarchical cognitive control in the human cerebellum. *Current Biology*, 30(10):1881–1892.e3, 2020.

[Friedman and Robbins, 2022] Naomi P Friedman and Trevor W Robbins. The role of prefrontal cortex in cognitive control and executive function. *Neuropsychopharmacology*, 47(1):72–89, 2022.

[Gaya *et al.*, 2023] Jean-Baptiste Gaya, Thang Doan, Lucas Caccia, Laure Soulier, Ludovic Denoyer, and Roberta Raileanu. Building a Subspace of Policies for Scalable Continual Learning. In *ICLR*, 2023.

[Hadsell *et al.*, 2020] Raia Hadsell, Dushyant Rao, Andrei A Rusu, and Razvan Pascanu. Embracing change: Continual learning in deep neural networks. *Trends in cognitive sciences*, 24(12):1028–1040, 2020.

[Hu *et al.*, 2024] Bin Hu, Chenyang Zhao, Pu Zhang, Zihao Zhou, Yuanhang Yang, Zenglin Xu, and Bin Liu. Enabling intelligent interactions between an agent and an llm: A reinforcement learning approach. In *RLC*, 2024.

[Kessler *et al.*, 2022] Samuel Kessler, Jack Parker-Holder, Philip J. Ball, Stefan Zohren, and Stephen J. Roberts. Same state, different task: Continual reinforcement learning without interference. In *AAAI*, pages 7143–7151, 2022.

[Khetarpal *et al.*, 2022] Khimya Khetarpal, Matthew Riemer, Irina Rish, and Doina Precup. Towards continual reinforcement learning: A review and perspectives. *Journal of Artificial Intelligence Research*, 75:1401–1476, 2022.

[Kirkpatrick *et al.*, 2017] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

[Kudithipudi *et al.*, 2022] Dhireesha Kudithipudi, Mario Aguilar-Simon, Jonathan Babb, Maxim Bazhenov, Douglas Blackiston, Josh Bongard, Andrew P Brna, Suraj Chakravarthi Raja, Nick Cheney, Jeff Clune, et al. Biological underpinnings for lifelong learning machines. *Nature Machine Intelligence*, 4(3):196–210, 2022.

[Kwon *et al.*, 2023] Minae Kwon, Sang Michael Xie, Kalesha Bullard, and Dorsa Sadigh. Reward Design with Language Models. In *ICLR*, 2023.

[Li *et al.*, 2023] Miaomiao Li, Jiaqi Zhu, Xin Yang, Yi Yang, Qiang Gao, and Hongan Wang. Cl-wstc: Continual learning for weakly supervised text classification on the internet. In *WWW*, pages 1489–1499, 2023.

[Li *et al.*, 2024] Yujie Li, Xin Yang, Hao Wang, Xiangkun Wang, and Tianrui Li. Learning to prompt knowledge transfer for open-world continual learning. *AAAI*, 38(12):13700–13708, Mar. 2024.

[Liu *et al.*, 2023] Zhihan Liu, Hao Hu, Shenao Zhang, Hongyi Guo, Shuqi Ke, Boyi Liu, and Zhaoran Wang. Reason for Future, Act for Now: A Principled Framework for Autonomous LLM Agents with Provable Sample Efficiency. *arXiv preprint arXiv:2309.17382*, 2023.

[Mallya and Lazebnik, 2018] Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *CVPR*, pages 7765–7773, 2018.

[McCloskey and Cohen, 1989] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pages 109–165. 1989.

[Palo *et al.*, 2023a] Norman Di Palo, Arunkumar Byravan, Leonard Hasenclever, Markus Wulfmeier, Nicolas Heess, and Martin Riedmiller. Towards A Unified Agent with Foundation Models. In *ICLR*, 2023.

[Palo *et al.*, 2023b] Norman Di Palo, Arunkumar Byravan, Leonard Hasenclever, Markus Wulfmeier, Nicolas Heess, and Martin Riedmiller. Towards a unified agent with foundation models. In *ICLR RRL Workshop*, 2023.

[Rolnick *et al.*, 2019] David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy P. Lillicrap, and Gregory Wayne. Experience replay for continual learning. In *NeurIPS*, pages 348–358, 2019.

[Rusu *et al.*, 2016] Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive Neural Networks. *arXiv preprint arXiv:1606.04671*, 2016.

[Schöpf *et al.*, 2022] Philemon Schöpf, Sayantan Auddy, Jakob Hollenstein, and Antonio Rodriguez-sanchez. Hypernetwork-PPO for Continual Reinforcement Learning. In *NeurIPS DeepRL Workshop*, 2022.

[Schwarz *et al.*, 2021] Jonathan Schwarz, Siddhant M. Jayakumar, Razvan Pascanu, Peter E. Latham, and Yee Whye Teh. Powerpropagation: A sparsity inducing weight reparameterisation. In *NeurIPS*, pages 28889–28903, 2021.

[Shinn *et al.*, 2023] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *NeurIPS*, 2023.

[Wang *et al.*, 2024a] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*, 2024.

[Wang *et al.*, 2024b] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024.

[Wang *et al.*, 2024c] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[Wolczyk *et al.*, 2021] Maciej Wolczyk, Michal Zajac, Razvan Pascanu, Lukasz Kucinski, and Piotr Milos. Continual world: A robotic benchmark for continual reinforcement learning. In *NeurIPS*, pages 28496–28510, 2021.

[Wolczyk *et al.*, 2022] Maciej Wolczyk, MichałZając, Razvan Pascanu, Ł ukasz Kuciński, and Piotr Mił oś. Disentangling transfer in continual reinforcement learning. In *NeurIPS*, volume 35, pages 6304–6317, 2022.

[Xi *et al.*, 2025] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101, 2025.

[Yang *et al.*, 2024] Sherry Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *ICLR*, 2024.

[Yu *et al.*, 2023] Wenhao Yu, Nimrod Gileadi, Chuyuan Fu, Sean Kirmani, Kuang-Huei Lee, Montserrat Gonzalez Arenas, Hao-Tien Lewis Chiang, Tom Erez, Leonard Hasenclever, Jan Humplik, Brian Ichter, Ted Xiao, Peng Xu, Andy Zeng, Tingnan Zhang, Nicolas Heess, Dorsa Sadigh, Jie Tan, Yuval Tassa, and Fei Xia. Language to Rewards for Robotic Skill Synthesis. In *CoRL*, 2023.

[Zhang *et al.*, 2023] Jesse Zhang, Jiahui Zhang, Karl Pertsch, Ziyi Liu, Xiang Ren, Minsuk Chang, Shao-Hua Sun, and Joseph J. Lim. Bootstrap Your Own Skills: Learning to Solve New Tasks with Large Language Model Guidance. In *CoRL*, 2023.

[Zhao *et al.*, 2023] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.