SCaR: Refining Skill Chaining for Long-Horizon Robotic Manipulation via Dual Regularization

SCaR：通过双重正则化优化长时域机器人操作的技能链

Zixuan Chen [1] Ze Ji[2] Jing Huo[1*] Yang Gao [1]
陈子轩 [1] 泽 Ji[2] 景 Huo[1*] 高杨 [1]

[1] State Key Laboratory for Novel Software Technology,Nanjing University,China,
[1] 新型软件技术国家重点实验室，南京大学，中国，

[2] School of Engineering,Cardiff University,UK,
[2] 工程学院，卡迪夫大学，英国，

chenzx@nju.edu.cn, jiz1@cardiff.ac.uk, huojing@nju.edu.cn, gaoy@nju.edu.cn
chenzx@nju.edu.cn, jiz1@cardiff.ac.uk, huojing@nju.edu.cn, gaoy@nju.edu.cn

1 Abstract

2 摘要

Long-horizon robotic manipulation tasks typically involve a series of interrelated sub-tasks spanning multiple execution stages. Skill chaining offers a feasible solution for these tasks by pre-training the skills for each sub-task and linking them sequentially. However, imperfections in skill learning or disturbances during execution can lead to the accumulation of errors in skill chaining process, resulting in execution failures. In this paper, we investigate how to achieve stable and smooth skill chaining for long-horizon robotic manipulation tasks. Specifically, we propose a novel skill chaining framework called Skill Chaining via Dual Regularization (SCaR). This framework applies dual regularization to sub-task skill pre-training and fine-tuning, which not only enhances the intra-skill dependencies within each sub-task skill but also reinforces the inter-skill dependencies between sequential sub-task skills, thus ensuring smooth skill chaining and stable long-horizon execution. We evaluate the SCaR framework on two representative long-horizon robotic manipulation simulation benchmarks: IKEA furniture assembly and kitchen organization. Additionally, we conduct a simple real-world validation in tabletop robot pick-and-place tasks. The experimental results show that, with the support of SCaR, the robot achieves a higher success rate in long-horizon tasks compared to relevant baselines and demonstrates greater robustness to perturbations.
长时域机器人操作任务通常涉及跨越多个执行阶段的一系列相互关联的子任务。技能链通过预训练每个子任务的技能并顺序连接它们，为这些任务提供了可行的解决方案。然而，技能学习中的不完善或执行过程中的干扰可能导致技能链过程中误差的累积，进而导致执行失败。本文研究如何实现长时域机器人操作任务中稳定且流畅的技能链。具体而言，我们提出了一种新颖的技能链框架——基于双重正则化的技能链（Skill Chaining via Dual Regularization，SCaR）。该框架在子任务技能的预训练和微调中应用双重正则化，不仅增强了每个子任务技能内部的依赖关系，还强化了连续子任务技能之间的依赖，从而确保技能链的流畅性和长时域执行的稳定性。我们在两个具有代表性的长时域机器人操作仿真基准——宜家家具组装和厨房整理——上评估了SCaR框架。此外，我们还在桌面机器人抓取与放置任务中进行了简单的现实验证。实验结果表明，在SCaR的支持下，机器人在长时域任务中取得了比相关基线更高的成功率，并表现出更强的抗扰动能力。

3 1 Introduction

4 1 引言

Long-horizon robotic manipulation tasks are characterized by sequences of diverse and interdependent sub-tasks, which makes it crucial to maintain the stability of multi-stage sequential execution. For instance, in the robotic assembly of a stool (Fig. 1) involving two sub-tasks of leg installation, overall success is evaluated based on both the sequential installation success and factors affecting the assembly within environmental constraints. Although recent advances in deep reinforcement learning (RL) and imitation learning (IL) show promise in training robots for such complex tasks [1, 2, 3, 4, 5, 6, 7], managing long-horizon tasks with a scratch RL or IL policy remains challenging due to computational demands, extensive exploration, and intricate step dependencies [8, 9]. Skill chaining, which involves decomposing long-horizon tasks into smaller sub-tasks, pre-training skills for each, and executing them sequentially, offers a practical solution [10, 11]. However, as shown in Fig. 1(a)(b), such methods tend to fail when sub-task skills are insufficiently trained or unexpected states arise due to disturbances, especially when applied to high-degree-of-freedom robots performing contact-rich, long-horizon tasks. [12, 13, 14, 15, 16, 17].
长时域机器人操作任务的特点是包含多样且相互依赖的子任务序列，这使得维持多阶段顺序执行的稳定性至关重要。例如，在机器人组装凳子（图1）时涉及两个腿部安装的子任务，整体成功不仅取决于顺序安装的成功，还受环境约束下组装因素的影响。尽管深度强化学习（RL）和模仿学习（IL）的最新进展在训练机器人完成此类复杂任务方面展现出潜力[1, 2, 3, 4, 5, 6, 7]，但由于计算需求大、探索范围广及步骤依赖复杂，使用从零开始的RL或IL策略管理长时域任务仍然具有挑战性[8, 9]。技能链通过将长时域任务分解为更小的子任务，预训练各子任务技能并顺序执行，提供了一种实用的解决方案[10, 11]。然而，如图1(a)(b)所示，当子任务技能训练不足或因干扰产生意外状态时，尤其是在高自由度机器人执行接触丰富的长时域任务时，这类方法往往会失败[12, 13, 14, 15, 16, 17]。

In this paper, we argue that the coordination and enhancing of dependencies within and between subtask skills is necessary for stable and smooth skill chaining of long-horizon robotic manipulation [10].
本文认为，为实现长时域机器人操作中稳定且流畅的技能链，必须协调并增强子任务技能内部及其之间的依赖关系[10]。

*Corresponding author.
*通讯作者。

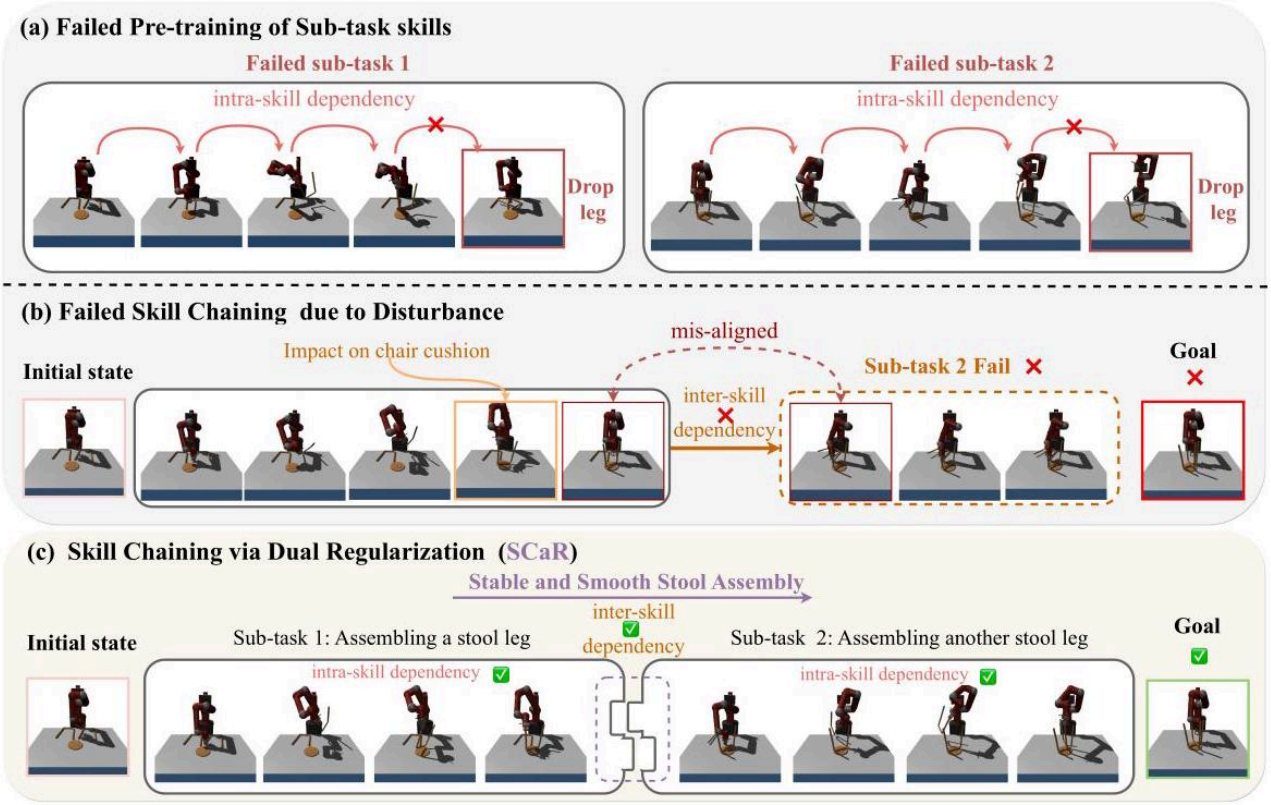


Figure 1: Illustration of the problem setting and the motivation of SCaR, using the example of a stool assembly task with two sub-tasks. Best viewed when zoomed in.
图1：以凳子组装任务中包含两个子任务为例，说明问题设置及SCaR的动机。建议放大查看。

For instance, as depicted in Fig. 1 (a)(b), the robot must consider following two points to ensure the overall task is accomplished: 1) ensuring the gripper consistently grasps and installs the stool leg stably within each sub-task skill range. and 2) ensuring the terminal state of previous skill aligns with the initial state of next skill for smooth skill chaining. We define the above two points as intra-skill dependencies between sequential actions within each sub-task skill and inter-skill dependencies between sequential sub-task skills, respectively. In this context, we propose a novel robotic skill chaining framework, Skill Chaining via Dual Regularization (SCaR). This framework enhances the aforementioned dependencies alternately through dual regularization during sub-task skill learning and chaining, aiming to provide stability for the execution of long-horizon robotic manipulation.
例如，如图1(a)(b)所示，机器人必须考虑以下两点以确保整体任务完成：1）确保夹持器在每个子任务技能范围内始终稳定地抓取并安装凳腿；2）确保前一技能的终止状态与下一技能的初始状态相匹配，以实现技能链的平滑衔接。我们将上述两点分别定义为子任务技能内连续动作之间的内部依赖（intra-skill dependencies）和连续子任务技能之间的外部依赖（inter-skill dependencies）。基于此，我们提出了一种新颖的机器人技能链框架——基于双重正则化的技能链（SCaR）。该框架通过在子任务技能学习和链式连接过程中交替应用双重正则化，增强上述依赖关系，旨在为长时域机器人操作的执行提供稳定性。

Specifically, in the pre-training phase of each sub-task skill, we propose the adaptive sub-task skill learning shceme, which employs a two-part policy learning objective that focuses on what sub-tasks the robot should perform (via RL) and how the robot should perform that task (via IL), and utilizes a novel adaptive equilibrium scheduling (AES) regularization to balance these two parts based on the robot's learning progress. This process aims to reinforce the intra-skill dependencies, ensuring a coherent sequence of actions in each sub-task skill. Subsequently, bi-directional adversarial learning is introduced in the fine-tuning phase of SCaR for better chaining sequential sub-task skills. This mechanism uses bi-

directional regularization to bring the terminal state of the current skill close to the initial state of its successor, and also to bring the initial state of the successor close to the terminal state of the current skill. This bi-directional alignment aims to reinforce robust inter-skill dependencies between sequential skills. Through the two innovative designs described, SCaR ensures coordination between the intra-skill and inter-skill dependencies, provides dual constraints for skill learning and skill chaining, as described in Fig. 1 (c), leading to a smooth skill chaining from the inside (within the sub-task skills) to the outside (between sub-task skills). Experimental results show that compared to scratch-training and skill chaining baselines, SCaR provides better task execution performance and stronger robustness to environmental perturbations in various long-horizon and contact-rich robotic manipulation simulation tasks. In addition, we conduct a simple validation in real-world tabletop robot pick-and-place tasks, and the results show that SCaR achieves a higher task success rate compared to previous skill-chaining methods.
具体来说，在每个子任务技能的预训练阶段，我们提出了自适应子任务技能学习方案，该方案采用由两部分组成的策略学习目标，分别关注机器人应执行哪些子任务（通过强化学习，RL）以及机器人应如何执行该任务（通过模仿学习，IL），并利用一种新颖的自适应平衡调度（AES）正则化方法，根据机器人的学习进度平衡这两部分。该过程旨在强化技能内部的依赖关系，确保每个子任务技能中的动作序列连贯。随后，在SCaR的微调阶段引入了双向对抗学习，以更好地串联顺序子任务技能。该机制通过双向正则化，使当前技能的终止状态接近其后继技能的初始状态，同时使后继技能的初始状态接近当前技能的终止状态。此双向对齐旨在强化顺序技能之间的稳健依赖关系。通过上述两项创新设计，SCaR确保了技能内部与技能间依赖的协调，为技能学习和技能串联提供了双重约束，如图1(c)所示，实现了从内部（子任务技能内）到外部（子任务技能间）的平滑技能串联。实验结果表明，与从零训练和技能串联基线相比，SCaR在多种长时序且接触丰富的机器人操作仿真任务中表现出更优的任务执行性能和更强的环境扰动鲁棒性。此外，我们在真实桌面机器人抓取放置任务中进行了简单验证，结果显示SCaR相比以往技能串联方法取得了更高的任务成功率。

The principal contributions of our work are delineated as follows: 1) We propose a novel robotic skill chaining framework via dual regularization, SCaR, for smoothly executing long-horizon manipulation tasks. 2) We introduce an adaptive sub-task skill learning scheme that acts as a regularization to enhance intra-skill dependencies between sequential actions within each sub-task skill. 3) We develop a bi-directional adversarial learning mechanism that serves as a regularization for reinforcing inter-skill dependencies between sequential sub-task skills. 4) In all eight simulated long-horizon robotic manipulation tasks and simple real-world pick-and-place tasks, SCaR demonstrates significantly better performance than scratch-training and skill-chaining baselines. Video demonstrations are available at: https://sites.google.com/view/scar8297
我们工作的主要贡献如下：1）我们提出了一种通过双重正则化实现平滑执行长时序操作任务的新型机器人技能串联框架SCaR。2）我们引入了一种自适应子任务技能学习方案，作为正则化手段，增强每个子任务技能内顺序动作之间的内部依赖关系。3）我们开发了一种双向对抗学习机制，作为正则化方法，强化顺序子任务技能之间的技能间依赖。4）在所有八个仿真长时序机器人操作任务及简单的真实抓取放置任务中，SCaR均显著优于从零训练和技能串联基线。视频演示见：https://sites.google.com/view/scar8297

5 2 Related Work

6 2 相关工作

6.1 2.1 Long-horizon Robotic Manipulation

6.2 2.1 长时序机器人操作

Training robots from scratch for complex, long-horizon tasks using reinforcement learning (RL) and imitation learning (IL) is challenging due to computational demands and distributional errors Solutions involve decomposing tasks into reusable sub-tasks [18]. Typically, such algorithms consist of a set of sub-policies that can be obtained through various methods, such as unsupervised exploration [19, 20, 21, 22, 23], learning from demonstrations [5, 6, 24, 25], and predefined measures [26, 27, 28, 29, 14]. Despite the merits of each of these approaches, they do not address well the challenges of long-horizon robot manipulation in environments that are object-rich, contact-rich, and characterized by multi-stage tasks [28, 29, 14]. Thus, even when pre-trained skills are provided, ensuring a smooth connection between manipulation policies remains a formidable challenge.
使用强化学习（RL）和模仿学习（IL）从零训练机器人完成复杂的长时序任务具有挑战性，原因在于计算需求高且存在分布偏差。解决方案通常是将任务分解为可复用的子任务[18]。此类算法通常由一组子策略组成，这些子策略可通过多种方法获得，如无监督探索[19, 20, 21, 22, 23]、示范学习[5, 6, 24, 25]以及预定义度量[26, 27, 28, 29, 14]。尽管这些方法各有优点，但它们未能很好地解决对象丰富、接触丰富且多阶段任务环境中长时序机器人操作的挑战[28, 29, 14]。因此，即使提供了预训练技能，确保操作策略之间的平滑衔接仍是一项艰巨任务。

6.3 2.2 Skill Chaining for Long-horizon Tasks

6.4 2.2 长时序任务的技能串联

Previous skill chaining methods for long-horizon tasks mainly focus on updating each sub-task policy to encompass the terminal state of the previous policy [11, 14, 30], implementing option chains [11, 31, 32] to forge logical skill sequences, or utilizing modulated skills to facilitate smoother transitions [33, 34, 35, 36, 14, 16]. However, these methods, while effective, often lead to a broad range of skill start and end states, a challenge in complex robotic manipulation tasks. T-STAR [15] is closely related to our work, addressing this by regularizing the learning process with a discriminator to control the expansion of the terminal state space. However, it focuses only on uni-directional dependencies between skills and ignores intra-skill dependencies within sub-task skills under long-horizon goals. Sequential Dexterity [17] centers on dexterous hand manipulation, introducing an optimization process to backpropagate long-term rewards across a policy chain. However, its scope still primarily emphasizes strengthening the dependencies between sub-task skills. GSC [37] attempts to solve skill chaining by employing diffusion models. It trains and chains primitive skills (pick, place, push, pull) through a Transformer-based skill diffusion model. However, due to the use of Transformer-based techniques, GSC requires high computational resources and cannot scale well to task environments with object-rich and contact-rich conditions. Our method instead employs simple and intuitive dual regularization constraints based on the lightweight policy network. By coordinating the dependencies within and between skills, we achieve refinement within sub-task policies and bi-directional alignment between them. This allows for stable skill chaining while also being scalable to various long-horizon manipulation tasks.

以往针对长时域任务的技能链方法主要集中在更新每个子任务策略以涵盖前一策略的终止状态[11, 14, 30]，实现选项链（option chains）[11, 31, 32]以构建逻辑技能序列，或利用调制技能促进更平滑的过渡[33, 34, 35, 36, 14, 16]。然而，这些方法虽然有效，但往往导致技能的起始和终止状态范围较广，这在复杂的机器人操作任务中是一个挑战。T-STAR[15]与我们的工作密切相关，通过使用判别器对学习过程进行正则化来控制终止状态空间的扩展，从而解决了这一问题。但其仅关注技能间的单向依赖，忽视了长时域目标下子任务技能内部的依赖关系。Sequential Dexterity[17]聚焦于灵巧手操作，引入一种优化过程，将长期奖励反向传播至策略链中，但其重点仍主要在于强化子任务技能间的依赖。GSC[37]尝试通过扩散模型解决技能链问题，利用基于Transformer的技能扩散模型训练并串联基础技能（抓取、放置、推、拉）。然而，由于采用Transformer技术，GSC计算资源需求高，难以扩展到物体丰富且接触复杂的任务环境。我们的方法则基于轻量级策略网络，采用简单直观的双重正则约束。通过协调技能内部及技能间的依赖，实现子任务策略的精细化及双向对齐，从而在保证技能链稳定性的同时，具备良好的长时域操作任务扩展性。

7 3 Preliminaries

8 3 预备知识

Among several related works on skill chaining, we consider a challenging yet practical problem setting that deals with long-horizon manipulation tasks through a combination of reinforcement learning (RL) and imitation learning (IL). In each sub-task in the long-horizon task, we consider robotic agents acting within a finite-horizon Markov Decision Process [38] $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma, d_{\mathcal{I}}, T)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P}(s' \mid s, a)$ is the transition function, $r(s, a, s')$ is the reward function, γ is the discount factor, $d_{\mathcal{I}}$ is the initial state distribution,and T is the episode horizon of sub-task. We define a policy $\pi : \mathcal{S} \to \mathcal{A}$ that maps states to actions and correspondingly moves the robotic agent to a new state according to the transition probabilities. This sub-task policy is trained to maximize the expected sum of discounted rewards $\mathbb{E}_{(s,a)\sim\pi}\left[\sum_{t=1}^{T} \gamma^t r(s_t, a_t, s_{t+1})\right]$. We assume that each sub-task policy has an initial state set $\mathcal{I} \in \mathcal{S}$ and a terminal state set $\beta \in \mathcal{S}$,where the initial set $\mathcal{I}$ contains all the initial states that lead to the successful execution of the policy and the terminal state set β contains all the final states of the successful execution. The environment

在众多关于技能链的相关工作中，我们考虑一个具有挑战性且实用的问题设定，即通过强化学习（RL）和模仿学习（IL）相结合来处理长时域操作任务。在长时域任务的每个子任务中，我们考虑机器人智能体在有限时域马尔可夫决策过程（Markov Decision Process, MDP）[38] $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma, d_{\mathcal{I}}, T)$ 中的行为，其中 $\mathcal{S}$ 是状态空间，$\mathcal{A}$ 是动作空间，$\mathcal{P}(s' \mid s, a)$ 是转移函数，$r(s, a, s')$ 是奖励函数，γ 是折扣因子，$d_{\mathcal{I}}$ 是初始状态分布，T 是子任务的回合时长。我们定义策略 $\pi : \mathcal{S} \to \mathcal{A}$，将状态映射到动作，并根据转移概率相应地将机器人智能体移动到新状态。该子任务策略通过最大化期望折扣奖励总和 $\mathbb{E}_{(s,a)\sim\pi}\left[\sum_{t=1}^{T} \gamma^t r(s_t, a_t, s_{t+1})\right]$ 进行训练。假设每个子任务策略具有初始状态集 $\mathcal{I} \in \mathcal{S}$ 和终止状态集 $\beta \in \mathcal{S}$，其中初始状态集 $\mathcal{I}$ 包含所有能成功执行该策略的初始状态，终止状态集 β 包含所有成功执行的最终状态。环境

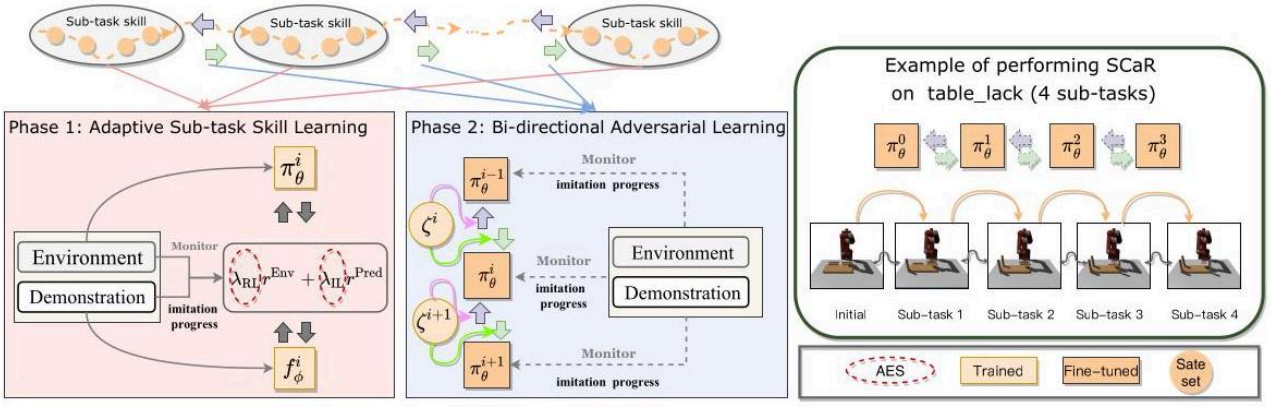


Figure 2: The Pipeline of Skill Chaining via Dual Regularization (SCaR). (Left) Phase 1: Sub-task skill pre-training ( ) merges environmental feedback and expert guidance, using adaptive equilibrium scheduling (AES) regularization to balance learning, which enhances intra-skill dependencies within skills. (Middle) Phase 2: Bi-directional discriminators (*) coupled with AES to fine-tune pre-trained sub-task skills, as regularization for reinforcing inter-skill dependencies. (Right) Evaluation: Evaluation of SCaR on long-horizon manipulation.*

图2：通过双重正则化实现技能链（*SCaR*）的流程。（左）阶段*1*：子任务技能预训练（ ）融合环境反馈与专家指导，采用自适应平衡调度（*AES*）正则化以平衡学习，增强技能内部依赖。（中）阶段*2*：双向判别器（）结合AES对预训练子任务技能进行微调，作为强化技能间依赖的正则化。（右）评估：在长时域操作任务中对SCaR进行评估。

provides the environmental feedback for each step taken by the agent and success metrics for each sub-task, derived from the terminal states of sub-task policy. For instance, as shown in Fig. 1(c), the alignment of the back and legs of the stool triggers the connect action and the realization of the sub-task goal, which indicates the successful completion of the sub-task. Additionally, we posit that during each sub-task policy learning, the agent receives a set of pre-defined expert demonstrations, $\mathbb{D}^E = \left\{\tau_1^E, \ldots, \tau_N^E\right\}$,to facilitate the IL process. Here, N represents the number of episodes,and each demonstration comprises a sequence of state-action pairs, $\tau^E = (s_1, a_1, \ldots, s_{T-1}, a_{T-1}, s_T)$.

为智能体每一步提供环境反馈及每个子任务的成功指标，这些指标来源于子任务策略的终端状态。例如，如图1(c)所示，凳子背部和腿部的对齐触发连接动作并实现子任务目标，表明子任务成功完成。此外，我们假设在每个子任务策略学习过程中，智能体接收一组预定义的专家示范，$\mathbb{D}^E = \{\tau_1^E, \ldots, \tau_N^E\}$，以促进模仿学习（IL）过程。这里，$N$表示回合数，每个示范包含一系列状态-动作对，$\tau^E = (s_1, a_1, \ldots, s_{T-1}, a_{T-1}, s_T)$。

9 4 Method

10 4 方法

In Section 4.1,we present the pipeline of the SCaR framework. Sections 4.2 and 4.3 provide further elaboration on the key design elements.
在第4.1节中，我们介绍了SCaR框架的流程。第4.2节和4.3节对关键设计元素进行了进一步阐述。

10.1 4.1 Overall Pipeline

10.2 4.1 总体流程

As illustrated in Fig. 2, the SCaR framework has two phases: (a) pre-training (adaptive sub-task skill learning) and (b) fine-tuning (bi-directional adversarial learning). In the pre-training phase, the agent co-learns sub-task skills by integrating environmental feedback and expert demonstrations In the fine-tuning phase, it refines these skills through bi-directional adversarial learning, enabling sequential integration of sub-task skills. After fine-tuning, SCaR can smoothly chain sub-task skills to complete long-horizon manipulation tasks. Specific modules and mechanisms for these phases are detailed in Sections 4.2 and 4.3
如图2所示，SCaR框架包含两个阶段：(a) 预训练（自适应子任务技能学习）和 (b) 微调（双向对抗学习）。在预训练阶段，智能体通过整合环境反馈和专家示范共同学习子任务技能。在微调阶段，通过双向对抗学习优化这些技能，实现子任务技能的顺序整合。微调后，SCaR能够顺畅地串联子任务技能以完成长时序操作任务。各阶段的具体模块和机制详见第4.2节和4.3节。

10.3 4.2 Adaptive Sub-task Skill Learning

10.4 4.2 自适应子任务技能学习

Weighted Reward Function To learn sub-task skills better, we combine goal-conditional RL and generative adversarial imitation learning (GAIL) [39], to pre-train skills that enable the agent to perform challenging sub-tasks in a desired expert behavioral style [40, 15]. More specifically, we consider the weighted reward function that is used to train each sub-task policy π_i^θ consists of two components specifying: what sub-task the agent should perform - learning from environmental feedback, and 2) how the agent should perform that task - learning from expert demonstrations:
加权奖励函数 为了更好地学习子任务技能，我们结合目标条件强化学习（goal-conditional RL）和生成对抗模仿学习（GAIL）[39]，预训练使智能体能够以期望的专家行为风格[40, 15]执行复杂子任务。更具体地，我们考虑用于训练每个子任务策略π_i^θ的加权奖励函数，该函数包含两个部分：1）智能体应执行何种子任务——从环境反馈中学习，2）智能体应如何执行该任务——从专家示范中学习：

$$r\left(s_t, a_t, s_{t+1}; \phi\right) = \lambda_{\mathrm{RL}} r_i^{\mathrm{Env}}\left(s_t, a_t, s_{t+1}, g\right) + \lambda_{\mathrm{IL}} r_i^{\mathrm{Pred}}\left(s_t, a_t; \phi\right). \tag{1}$$

As shown in Eq. 1,the first component is represented by a task-specific reward $r_i^{\mathrm{Env}}\left(s_t, a_t, s_{t+1}, g\right)$, which defines general objectives that the agent should satisfy to fulfill a given sub-task goal g for current MDP $\mathcal{M}$ (e.g. assembling a stool leg). The second component is represented through a learned task-agnostic predict-reward $r_i^{\mathrm{Pred}}\left(s_t, a_t; \phi\right)$,which specifies manipulation details of the behaviors that the agent should adopt when performing the sub-task (e.g., the expert way to grab a stool leg and attach it),and $r_i^{\mathrm{Pred}}\left(s_t, a_t; \phi\right)$ is the predicted reward by a least-square GAIL discriminator f_ϕ^i [41,40,15],which is more stable than the standard GAIL objective using the sigmoid cross-entropy loss function. Therefore, the predicted reward is:
如公式1所示，第一部分由任务特定奖励$r_i^{\mathrm{Env}}\left(s_t, a_t, s_{t+1}, g\right)$表示，定义智能体为完成当前马尔可夫决策过程（MDP）$\mathcal{M}$中的给定子任务目标g（例如组装凳子腿）应满足的一般目标。第二部分通过学习得到的任务无关预测奖励$r_i^{\mathrm{Pred}}\left(s_t, a_t; \phi\right)$表示，指定智能体执行子任务时应采用的操作细节（例如专家抓取凳子腿并连接的方式），$r_i^{\mathrm{Pred}}\left(s_t, a_t; \phi\right)$是由最小二乘GAIL判别器$f_\phi^i$[41,40,15]预测的奖励，该判别器比使用sigmoid交叉熵损失函数的标准GAIL目标更稳定。因此，预测奖励为：

$$r_i^{\mathrm{Pred}}\left(s_t, a_t; \phi\right) = \max\left[0, 1 - 0.25 \cdot \left[f_\phi^i\left(s_t, a_t\right) - 1\right]^2\right]. \tag{2}$$

We adopt the training objective of the least-squares GAIL discriminator [41] with a gradient penalty term [42, 43], This penalty term mitigates the instability of the training dynamics due to the interplay between the discriminator and the policy [40], as follows:
我们采用带梯度惩罚项[42, 43]的最小二乘GAIL判别器[41]训练目标，该惩罚项缓解了判别器与策略之间相互作用导致的训练动态不稳定性[40]，具体如下：

$$\operatorname{argmin}_{f_\phi^i} \mathbb{E}_{(s)\sim\mathbb{D}^E}\left[\left(f_\phi^i(s) - 1\right)^2\right] + \mathbb{E}_{(s)\sim\pi_\theta^i}\left[\left(f_\phi^i(s) + 1\right)^2\right] + \frac{\eta^{\mathrm{gp}}}{2}\mathbb{E}_{(s)\sim\mathbb{D}^E}\left[\left\|\nabla_s f_\phi^i(s)\right\|^2\right], \tag{3}$$

where η^{gp} is a manually-specified coefficient. The scales of r^{Env} and r^{Pred} in previous related works are set by fixed weights and linearly combined into the final reward function [40, 15]. This could lead to the agent rigidly imitating experts and curbing self-exploration, finding it difficult to adjust intra-skill dependencies and adapt to dynamic task perturbations. We propose a principle to counter this: If the agent fails to imitate the expert's demonstration well, it should shift focus to self-learning from the environment. Conversely, effective imitation should continue, focusing on

the expert to mitigate low sample efficiency in reinforcement learning. Accordingly, we extend the automatic discount scheduling (ADS) solution [9] to our problem setting, and propose adaptive equilibrium scheduling (AES) to regularize the scales of r^{Env} and r^{Pred} in sub-task skill learning for adaptive scheduling the focus of reinforcement and imitation learning, as shown in Fig. 3

其中 η^{gp} 是手动指定的系数。以往相关工作中 r^{Env} 和 r^{Pred} 的权重通过固定权重设定，并线性组合成最终的奖励函数[40, 15]。这可能导致智能体僵化地模仿专家，抑制自我探索，难以调整技能内依赖关系并适应动态任务扰动。我们提出一条原则以应对这一问题：若智能体未能良好模仿专家示范，应转而侧重于从环境中自我学习；反之，若模仿有效，则应继续聚焦专家，以缓解强化学习中的低样本效率。基于此，我们将自动折扣调度（ADS）方案[9]扩展至本问题设置，提出自适应平衡调度（AES），用于正则化子任务技能学习中 r^{Env} 和 r^{Pred} 的权重，实现强化学习与模仿学习关注点的自适应调度，如图3所示。

Adaptive Equilibrium Scheduling (AES) Regularization Specifically, AES balances the scales of r^{Env} and r^{Pred} during the learning process of each skill through adaptive scheduling of λ_{RL} and λ_{IL} , according to how well the agent imitates the expert's demonstration. To capture the agent's imitation progress,AES refers to the solution in ADS [9] and uses the imitation identifier Φ to continuously monitor whether the agent is imitating the expert demonstration well enough.

自适应平衡调度（AES）正则化具体而言，AES通过自适应调度 λ_{RL} 和 λ_{IL}，在每个技能的学习过程中平衡 r^{Env} 和 r^{Pred} 的权重，依据智能体对专家示范的模仿程度。为捕捉智能体的模仿进展，AES参考ADS[9]中的方案，使用模仿识别器 Φ 持续监测智能体是否足够良好地模仿专家示范。

At the beginning of training, the agent is assigned two initial balance factors $\lambda_{\text{RL}} = \alpha, \lambda_{\text{IL}} = 1 - \alpha$,where base exponent $\alpha \in [0, 1]$. We set $\alpha = 0.5$ in the experiments and the agent is assigned two identical balance factors $\lambda_{\text{RL}} = \lambda_{\text{IL}} = 0.5^2$,indicating that at the beginning of learning, the agent imitates the expert's behavior with the same weight as the behavior of environment exploration according to the task goal. As training progresses, the imitation progress recognizer Φ is queried periodically to monitor the progress of the agent's imitation of the expert's behavior. Φ receives the agent's collected trajectories and infers the agent's current imitation progress $p \in [0, T)$, where p in an integer and T is the step of the entire episode.

训练初期，智能体被赋予两个初始平衡因子 $\lambda_{\text{RL}} = \alpha, \lambda_{\text{IL}} = 1 - \alpha$，其中基底指数为 $\alpha \in [0, 1]$。实验中我们设定了 $\alpha = 0.5$，智能体被赋予两个相同的平衡因子 $\lambda_{\text{RL}} = \lambda_{\text{IL}} = 0.5^2$，表示学习初期，智能体以与环境探索行为相同的权重模仿专家行为，符合任务目标。随着训练推进，模仿进展识别器 Φ 会定期被查询，以监控智能体对专家行为的模仿进度。Φ 接收智能体收集的轨迹并推断当前模仿进展 $p \in [0, T)$，其中 p 为整数，T 是整个回合的步数。

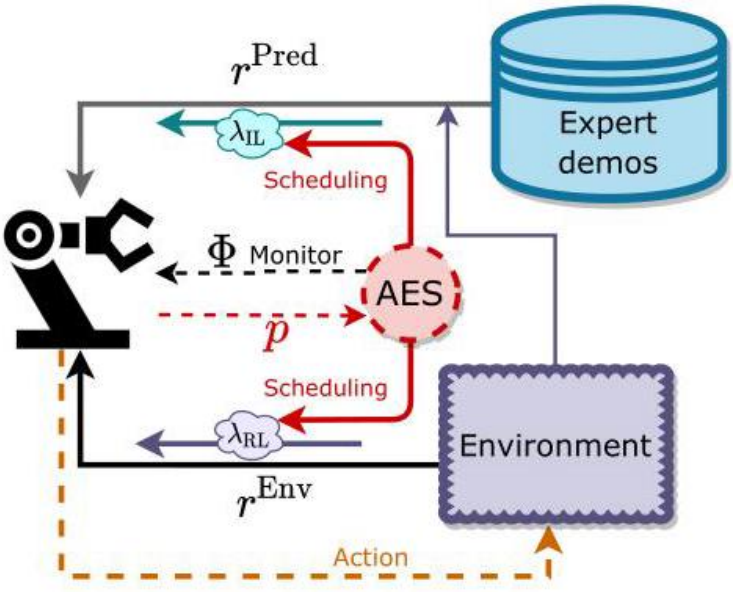


Figure 3: AES regularization for subtask skill learning.

图3：子任务技能学习的AES正则化。

The construction of Φ ,with reference to ADS,first requires the construction of a sequence $\mathbf{Q}(q_1, \ldots, q_T)$, where $q_i = \text{argmin}_j c\left(s_i, s_j^E\right)$ is the index of the nearest neighbor of s_i in τ^E, c is the cosine similarity. The progress alignment between τ and τ_j^E is measured as the length of the longest increasing subsequence (LIS) in $\mathbf{Q}$,denoted as $\text{LIS}\left(\tau, \tau^E\right)$. Specifically,the agent's imitation progress p is increased by 1 if the following inequality holds:

Φ 的构建参考ADS，首先需要构建序列 $\mathbf{Q}(q_1, \ldots, q_T)$，其中 $q_i = \text{argmin}_j c\left(s_i, s_j^E\right)$ 是 s_i 在 τ^E, c 中的最近邻索引，τ^E, c 是余弦相似度。τ 与 τ_j^E 之间的进展对齐通过 $\mathbf{Q}$ 中最长递增子序列（LIS）的长度衡量，记为 $\text{LIS}\left(\tau, \tau^E\right)$。具体而言，当满足以下不等式时，智能体的模仿进展 p 增加1：

$$\max_{\acute{\tau}^E \in \mathbb{D}^E} LIS\left(\tau_{1:p+1}, \acute{\tau}_{1:p+1}^E\right) \geq \rho \times \min_{\acute{\tau}^E, \acute{\acute{\tau}}^E \in \mathbb{D}^E} LIS\left(\acute{\tau}_{1:p+1}^E, \acute{\acute{\tau}}_{1:p+1}^E\right), \tag{4}$$

where $\hat{\tau}^E \neq \hat{\tau}^E$,the subscript $1:p+1$ denotes the first $p+1$ steps of the trajectory,and $\rho \in [0,1]$ controls the strictness of the imitation progress monitoring. This suggests that the similarity of the agent trajectory to its best matching expert trajectory at time step $p+1$ exceeds the minimal similarity criterion within the expert demonstration. See Appendix B for detailed explanation of AES.
其中 $\hat{\tau}^E \neq \hat{\tau}^E$，下标 $1:p+1$ 表示轨迹的前 $p+1$ 步，$\rho \in [0,1]$ 控制模仿进展监测的严格程度。该条件表明智能体轨迹在时间步 $p+1$ 与其最佳匹配专家轨迹的相似度超过专家示范中的最小相似度标准。详见附录B中AES的详细说明。

After obtaining the current imitation progress p of the agent,AES then adopts a mapping function $\varphi_\lambda(p)$ to schedule the two new balance discount factors λ_{RL} and λ_{IL} . Straightforward idea of setting $\varphi_\lambda(p)$ is that If p is larger and reaches a certain threshold,i.e.,the agent is able to imitate the expert behavior well, then the more the agent tends to imitate the expert's behavior in subsequent training,and vice versa. Therefore,we set the threshold as $\frac{T}{2}$. If $p \in \left[0, \frac{T}{2}\right)$,we propose $\varphi_\lambda(p) = 1 - e^{\left(-\frac{p}{k}\right)}$; if $p \in \left[\frac{T}{2}, T\right)$,we propose $\varphi_\lambda(p) = e^{\left(-\frac{p-\frac{T}{2}}{k}\right)}$,where k is used to flatten the curve of the mapping function. Then λ_{RL} and λ_{IL} are scheduled to be :
在获得代理当前的模仿进度p后，AES采用映射函数$\varphi_\lambda(p)$来调度两个新的平衡折扣因子λ_{RL}和λ_{IL}。设置$\varphi_\lambda(p)$的直接思路是：如果p较大并达到某一阈值，即代理能够较好地模仿专家行为，那么代理在后续训练中更倾向于模仿专家行为，反之亦然。因此，我们将阈值设为$\frac{T}{2}$。如果$p \in \left[0, \frac{T}{2}\right)$，我们提出$\varphi_\lambda(p) = 1 - e^{\left(-\frac{p}{k}\right)}$；如果$p \in \left[\frac{T}{2}, T\right)$，我们提出$\varphi_\lambda(p) = e^{\left(-\frac{p-\frac{T}{2}}{k}\right)}$，其中$k$用于平滑映射函数的曲线。然后，$\lambda_{\mathrm{RL}}$和$\lambda_{\mathrm{IL}}$被调度为：

$$\begin{cases} \lambda_{\mathrm{RL}} = \alpha^{\varphi_\lambda(p)}, \lambda_{\mathrm{IL}} = 1 - \alpha^{\varphi_\lambda(p)}. & \text{if } p \in \left[0, \frac{T}{2}\right) \\ \lambda_{\mathrm{IL}} = \alpha^{\varphi_\lambda(p)}, \lambda_{\mathrm{RL}} = 1 - \alpha^{\varphi_\lambda(p)}. & \text{if } p \in \left[\frac{T}{2}, T\right) \end{cases} \tag{5}$$

---

[2] We further explore what effect different α would have in the Ablation Experiments.
[2]我们在消融实验中进一步探讨不同α的影响。

Consequently, the RL and IL components of sub-task skill learning can be adaptively scheduled and regularized through AES, effectively enhancing intra-skill dependencies between sequential actions. The pseudo-code of adaptive sub-task skill learning is outlined in Algorithm 1 in Appendix A.1
因此，子任务技能学习中的强化学习（RL）和模仿学习（IL）组件可以通过AES自适应调度和正则化，有效增强序列动作间的技能内依赖。自适应子任务技能学习的伪代码见附录A.1中的算法1。

10.5 4.3 Bi-directional Adversarial Learning for Skill Chaining

10.6 4.3 技能链的双向对抗学习

Executing pre-trained sub-task skills sequentially without considering inter-skill dependencies may lead to failure. To address this, we propose bi-directional adversarial learning to further refine and better integrate sequential sub-task skills. The pseudo-code of bi-directional adversarial learning is outlined in Algorithm 2 in Appendix A.2
顺序执行预训练的子任务技能而不考虑技能间依赖可能导致失败。为此，我们提出双向对抗学习，以进一步优化和更好地整合序列子任务技能。双向对抗学习的伪代码见附录A.2中的算法2。

Bi-directional Regularization In contrast to previous uni-directional regularization schemes that only augment the initial state set $\mathcal{I}_i$ or regularize the terminal state set β_i [12,15],we impose the bi-directional constraints $(\mathcal{C}_1, \mathcal{C}_2)$ on inter-skill dependencies,facilitating smooth skill chaining,as shown in Fig 4 With the bi-directional constraint, we implement the bi-directional adversarial learning,centered on the joint training of a bi-directional discriminator,denoted by ζ_ω^i ,which is adept at distinguishing between the terminal state set of the preceding policy and the initial state set of the subsequent policy. The bi-directional constraints $\mathcal{C}_1, \mathcal{C}_2$ are defined as Eq. 10:
双向正则化 与以往仅增强初始状态集$\mathcal{I}_i$或正则化终止状态集β_i的单向正则化方案[12,15]不同，我们对技能间依赖施加双向约束$(\mathcal{C}_1, \mathcal{C}_2)$，促进技能链的平滑衔接，如图4所示。基于双向约束，我们实现了以双向判别器ζ_ω^i联合训练为核心的双向对抗学习，该判别器擅长区分前一策略的终止状态集与后一策略的初始状态集。双向约束$\mathcal{C}_1, \mathcal{C}_2$定义如公式10：

next initial → current terminal: $\mathcal{C}_1 = \mathbb{E}_{s_{\mathcal{I}} \sim \mathcal{I}_{i+1}}\left[\zeta_{\omega_1}^i(s_{\mathcal{I}}) - 1\right]^2 + \mathbb{E}_{s_T \sim \beta_i}\left[\zeta_{\omega_1}^i(s_T)\right]^2$
下一初始→ 当前终止： $\mathcal{C}_1 = \mathbb{E}_{s_{\mathcal{I}} \sim \mathcal{I}_{i+1}}\left[\zeta_{\omega_1}^i(s_{\mathcal{I}}) - 1\right]^2 + \mathbb{E}_{s_T \sim \beta_i}\left[\zeta_{\omega_1}^i(s_T)\right]^2$ (6)
previous terminal → current initial: $\mathcal{C}_2 = \mathbb{E}_{s_T \sim \beta_{i-1}}\left[\zeta_{\omega_2}^i(s_T) - 1\right]^2 + \mathbb{E}_{s_{\mathcal{I}} \sim \mathcal{I}_i}\left[\zeta_{\omega_2}^i(s_{\mathcal{I}})\right]^2$
前一终端 → 当前初始： $\mathcal{C}_2 = \mathbb{E}_{s_T \sim \beta_{i-1}}\left[\zeta_{\omega_2}^i(s_T) - 1\right]^2 + \mathbb{E}_{s_{\mathcal{I}} \sim \mathcal{I}_i}\left[\zeta_{\omega_2}^i(s_{\mathcal{I}})\right]^2$

$\zeta_{\omega_1}^i$ and $\zeta_{\omega_2}^i$ are two separate networks,each used to minimize the adversarial learning process in two different directions, and the parameters of the two networks are averaged and combined into ζ_ω^i . In summary, ζ_ω^i is trained for each policy to minimize the objective function[3] | $\mathcal{L}_i(\omega) = \frac{1}{2}\mathcal{C}_1 + \frac{1}{2}\mathcal{C}_2$. Guided by ζ_ω^i ,the bi-directional adversarial learning not only steers the terminal state set of the current policy towards the initial state set of the subsequent policy, but also ensures alignment of the initial state set of the subsequent policy with the terminal state set of current policy. This dual alignment establishes a balanced mapping between the initial and terminal states of sequential skills to reinforce inter-

skill dependencies, ensure consistency and stability in multistage tasks, and guarantee smooth transitions between sequential skills. Accordingly, the bi-directional regularization can be added to the overall objective function of policy learning in the form of the following reward term: $r_i^{\mathrm{Bi}}(s;\omega) = \mathbb{1}_{s\in\beta_i}\zeta^{i+1}(s) + \mathbb{1}_{s\in\mathcal{I}_i}\zeta^{i-1}(s)$.
$\zeta^i_{\omega_1}$ 和 $\zeta^i_{\omega_2}$ 是两个独立的网络，分别用于在两个不同方向上最小化对抗学习过程，两个网络的参数被平均并合并为 ζ^i_ω 。总之，ζ^i_ω 针对每个策略进行训练以最小化目标函数 tion[3] $\mid \mathcal{L}_i(\omega) = \frac{1}{2}\mathcal{C}_1 + \frac{1}{2}\mathcal{C}_2$ 。在 ζ^i_ω 的引导下，双向对抗学习不仅将当前策略的终端状态集引导至后续策略的初始状态集，还确保后续策略的初始状态集与当前策略的终端状态集对齐。这种双重对齐建立了顺序技能初始状态与终端状态之间的平衡映射，强化了技能间的依赖关系，确保多阶段任务中的一致性和稳定性，并保证顺序技能之间的平滑过渡。因此，双向正则化可以作为以下奖励项形式加入策略学习的整体目标函数中：$r_i^{\mathrm{Bi}}(s;\omega) = \mathbb{1}_{s\in\beta_i}\zeta^{i+1}(s) + \mathbb{1}_{s\in\mathcal{I}_i}\zeta^{i-1}(s)$.

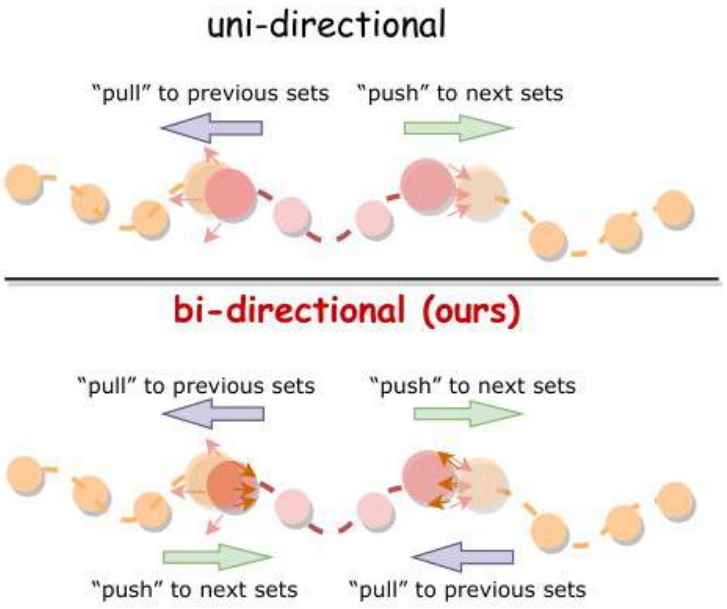


Figure 4: Bi-directional regularization for sub-task skill chaining.
图4：子任务技能链的双向正则化。

Overall Objective Function So far, the objective function via dual regularization, i.e., AES regularization and bi-directional regularization, to pre-train, fine-tune and chain sub-task skills can be rewritten as a weighted sum of the individual reward terms:
整体目标函数 迄今为止，通过双重正则化，即AES正则化和双向正则化，对子任务技能进行预训练、微调和链式连接的目标函数可以重写为各个奖励项的加权和：

$$r_i(s_t, a_t, s_{t+1}; \phi) = \underbrace{\lambda_{\mathrm{RL}} r_i^{\mathrm{Env}}(s_t, a_t, s_{t+1}, g) + \lambda_{\mathrm{LL}} r_i^{\mathrm{Pred}}(s_t, a_t; \phi)}_{\text{AES regularization}} + \underbrace{\lambda_{\mathrm{Bi}} r_i^{\mathrm{Bi}}(s_{t+1}; \omega)}_{\text{bi-directional regularization}}, \tag{7}$$

where λ_{Re} is the weighting factor of the bi-directional regularization. The objective function features AES regularization and bi-directional regularization to enhance intra- and inter-skill dependencies. It enables the agent to adaptively pre-train skills that can solve different sub-tasks well through environmental feedback and expert guidance, and further fine-tune them through the bi-directional discriminator to achieve dual alignment between sequential skills. At the same time, the fine-tuned sub-task skills help to collect terminal and initial states to refine the bi-directional discriminator. This iterative process ensures smooth long-horizon task skill chaining.
其中 λ_{Re} 是双向正则化的权重因子。该目标函数结合了AES正则化和双向正则化，以增强技能内部及技能间的依赖关系。它使智能体能够通过环境反馈和专家指导，自适应地预训练能够解决不同子任务的技能，并通过双向判别器进一步微调，实现顺序技能之间的双重对齐。同时，微调后的子任务技能有助于收集终端和初始状态，以优化双向判别器。该迭代过程确保了长时域任务技能链的平滑衔接。

[3] We explore the impact of different scales of $\mathcal{C}_1$ and $\mathcal{C}_2$ in Appendix D.3
[3] 我们在附录D.3中探讨了不同规模的 $\mathcal{C}_1$ 和 $\mathcal{C}_2$ 的影响

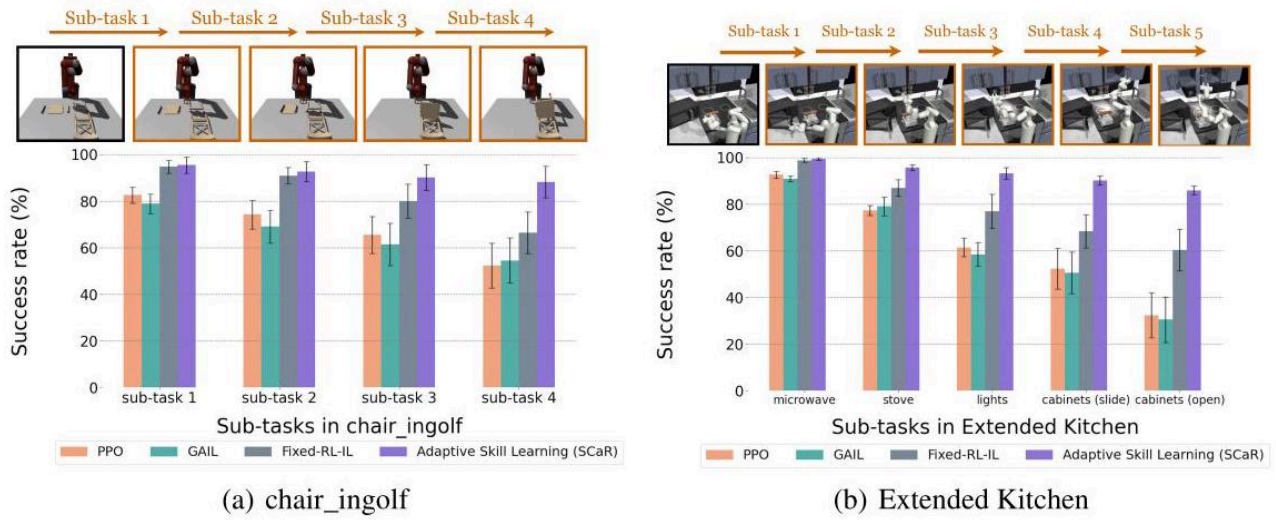


Figure 5: Evaluation Performance of Sub-task Skill Learning. Best viewed zoomed.
图5：子任务技能学习的评估表现。建议放大查看。

11 5 Experiments

12 5 实验

12.1 5.1 Experiment Setup

12.2 5.1 实验设置

We conduct simulation experiments on six IKEA furniture assembly tasks and two kitchen organization tasks, and also perform long-horizon pick-and-place experiments on the real Sagittarius K1 robot. Please refer to the Appendix for more detailed simulation experiment setup (Appendix G), network architecture (Appendix H), training details (Appendix T), more quantitative (Appendix D) and qualitative results (Appendix E) of the simulation tasks, and the real-robot experiments (Appendix F).
我们在六个宜家家具组装任务和两个厨房整理任务上进行了仿真实验，并在真实的Sagittarius K1机器人上进行了长时域的抓取与放置实验。更多详细的仿真实验设置（附录G）、网络架构（附录H）、训练细节（附录T）、仿真任务的更多定量（附录D）和定性结果（附录E），以及真实机器人实验（附录F），请参见附录。

Furniture Assembly We conduct experiments in six IKEA furniture assembly tasks in [44]: chair_agne, chair_bernhard, chair_ingolf, table_lack, toy_table, and table_dockstra.
家具组装 我们在文献[44]中的六个宜家家具组装任务上进行了实验：chair_agne、chair_bernhard、chair_ingolf、table_lack、toy_table 和 table_dockstra。

1. chair_agne: Two stool legs need to be picked up and aligned with the cross notches on the stool back. 2) chair_bernhard: The two chair supports need to be taken and aligned with the slots at the bottom of the chair surface. 3) chair_ingolf: Two chair supports and front legs need to be attached to the chair seat, which must then be secured to the chair back while avoiding collision with each other. 4) table_lack: The four table legs need to be picked up and aligned with the corners of the tabletop. 5) toy_table: The four table legs need to be picked up and aimed and inserted with the four notches on the table back. 6) table_dockstra: After supporting the two bases with table leg, the table top needs to be mounted while preventing collision. For each assembly task, we define the assembly of individual parts as sub-tasks. We collect 200 demonstrations per sub-task using a procedural assembly policy for imitation learning. Each demonstration consists of 150 steps.
2. chair_agne：需要拾取两个凳子腿，并与凳子靠背上的交叉缺口对齐。2) chair_bernhard：需要拿起两个椅子支撑件，并与椅面底部的槽口对齐。3) chair_ingolf：需要将两个椅子支撑件和前腿安装到椅座上，然后将椅座固定到椅背上，同时避免相互碰撞。4) table_lack：需要拾取四个桌腿，并与桌面四角对齐。5) toy_table：需要拾取四个桌腿，并对准桌背上的四个缺口插入。6) table_dockstra：在用桌腿支撑两个底座后，需要安装桌面，同时防止碰撞。对于每个组装任务，我们将单个部件的组装定义为子任务。我们使用程序化组装策略进行模仿学习，每个子任务收集200个示范。每个示范包含150个步骤。

Kitchen Organization We use the Franka Kitchen tasks in D4RL [45] and collect 200 demonstrations per sub-task for imitation learning. Specifically, we refer to the kitchen task in [46] and further extend the task sequence: in the Kitchen task, the 7-DoF Franka Emika Panda arm needs to perform 4 sequential sub-tasks, namely Turn on the microwave - Move the kettle - Turn on the stove - Turn on the light. In the Extended Kitchen task, the robot needs to perform 5 sequential sub-tasks: Turn on the microwave - Turn on the stove - Turn on the light - Slide the cabinet to the right - Open the cabinet, in which the sub-tasks have a lower probability of switching and is more challenging.
厨房整理 我们使用D4RL[45]中的Franka厨房任务，并为每个子任务收集200个示范用于模仿学习。具体来说，我们参考文献[46]中的厨房任务，并进一步扩展任务序列：在厨房任务中，7自由度Franka Emika Panda机械臂需要执行4个顺序子任务，分别是打开微波炉——移动水壶——打开炉灶——打开灯。在扩展厨房任务中，机器人需要执行5个顺序子任务：打开微波炉——打开炉灶——打开灯——将橱柜向右滑动——打开橱柜，其中子任务切换的概率较低，难度更大。

Baselines We compare SCaR with the following two types of baselines:
基线 我们将SCaR与以下两类基线进行比较：

Table 1: Long-horizon tasks execution performance (varies by sub-task completion progress): *tasks with 2 sub-tasks progress by 0.5 per sub-task, *tasks with 4 sub-tasks by 0.25, *tasks with 5 sub-tasks by 0.2 , and table_dockstra with 3 sub-tasks by 0.3 , where 0.9 indicates completion of all tasks. Best viewed zoomed.*
表1：长时序任务执行性能（根据子任务完成进度变化）：含2个子任务的任务每个子任务进度为0.5，含4个子任务的任务每个子任务进度为0.25，含5个子任务的任务每个子任务进度为0.2，table_dockstra含3个子任务每个子任务进度为0.3，0.9表示所有任务完成。建议放大查看。

Method	Furniture Assembly							Kitchen Organization		
	chair_agne	chair_bernhard	chair_ingolf	table_lack	toy_table	table_dockstra	All	Kitchen	E-Kitchen	All
PPO (Scratch RL)	${0.54} \pm {0.18}$	${0.42} \pm {0.12}$	${0.14} \pm {0.03}$	${0.09} \pm {0.01}$	${0.00} \pm {0.00}$	${0.31} \pm {0.12}$	${0.25} \pm {0.15}$	${0.13} \pm {0.05}$	${0.03} \pm {0.00}$	${0.08} \pm {0.04}$
GAIL (Scratch IL)	${0.31} \pm {0.05}$	${0.23} \pm {0.02}$	${0.00} \pm {0.00}$	${0.00} \pm {0.00}$	${0.00} \pm {0.00}$	${0.21} \pm {0.04}$	${0.12} \pm {0.09}$	${0.00} \pm {0.00}$	${0.00} \pm {0.00}$	${0.00} \pm {0.00}$
Fixed-RL-IL	${0.68} \pm {0.12}$	${0.53} \pm {0.07}$	${0.22} \pm {0.08}$	${0.21} \pm {0.11}$	${0.13} \pm {0.02}$	${0.43} \pm {0.07}$	${0.37} \pm {0.15}$	${0.33} \pm {0.06}$	${0.18} \pm {0.02}$	${0.26} \pm {0.06}$
SkiMo	${0.75} \pm {0.09}$	${0.62} \pm {0.05}$	${0.47} \pm {0.03}$	${0.58} \pm {0.14}$	${0.34} \pm {0.06}$	${0.62} \pm {0.11}$	${0.56} \pm {0.11}$	${0.57} \pm {0.08}$	${0.21} \pm {0.04}$	${0.39} \pm {0.13}$
Policy Sequencing	${0.89} \pm {0.08}$	${0.82} \pm {0.09}$	${0.77} \pm {0.12}$	${0.63} \pm {0.28}$	${0.45} \pm {0.18}$	${0.61} \pm {0.14}$	${0.70} \pm {0.16}$	${0.53} \pm {0.11}$	${0.36} \pm {0.09}$	${0.44} \pm {0.09}$
T-STAR	${0.92} \pm {0.02}$	${0.90} \pm {0.04}$	${0.89} \pm {0.04}$	${0.90} \pm {0.07}$	${0.71} \pm {0.21}$	${0.77} \pm {0.09}$	${0.85} \pm {0.09}$	${0.68} \pm {0.13}$	${0.48} \pm {0.08}$	${0.58} \pm {0.10}$
SCaR w/o Bi	${0.93} \pm {0.04}$	${0.92} \pm {0.02}$	${0.91} \pm {0.01}$	${0.93} \pm {0.02}$	${0.80} \pm {0.10}$	${0.79} \pm {0.02}$	${0.88} \pm {0.05}$	${0.75} \pm {0.08}$	${0.57} \pm {0.14}$	${0.66} \pm {0.09}$
SCaR w/o AES	${0.95} \pm {0.03}$	${0.94} \pm {0.03}$	${0.93} \pm {0.02}$	${0.95} \pm {0.04}$	${0.85} \pm {0.06}$	${0.80} \pm {0.03}$	${0.91} \pm {0.05}$	${0.77} \pm {0.07}$	${0.61} \pm {0.13}$	${0.74} \pm {0.05}$
SCaR (Ours)	${0.98} \pm {0.02}$	${0.96} \pm {0.04}$	${0.95} \pm {0.03}$	${0.97} \pm {0.03}$	${0.92} \pm {0.05}$	$\mathbf{{0.88}} \pm {0.02}$	$\mathbf{{0.94}} \pm {0.03}\left({{12}\% \uparrow }\right)$	${0.84} \pm {0.16}$	${0.73} \pm {0.17}$	$\mathbf{{0.78}} \pm {0.12}\left({{18}\% \uparrow }\right)$

方法	家具组装							厨房整理		
	chair_agne	chair_bernhard	chair_ingolf	table_lack	玩具桌	table_dockstra	全部	厨房	电子厨房	全部
PPO（策略优化，强化学习）	${0.54} \pm {0.18}$	${0.42} \pm {0.12}$	${0.14} \pm {0.03}$	${0.09} \pm {0.01}$	${0.00} \pm {0.00}$	${0.31} \pm {0.12}$	${0.25} \pm {0.15}$	${0.13} \pm {0.05}$	${0.03} \pm {0.00}$	${0.08} \pm {0.04}$
GAIL（生成对抗模仿学习）	${0.31} \pm {0.05}$	${0.23} \pm {0.02}$	${0.00} \pm {0.00}$	${0.00} \pm {0.00}$	${0.00} \pm {0.00}$	${0.21} \pm {0.04}$	${0.12} \pm {0.09}$	${0.00} \pm {0.00}$	${0.00} \pm {0.00}$	${0.00} \pm {0.00}$
固定强化学习-模仿学习	${0.68} \pm {0.12}$	${0.53} \pm {0.07}$	${0.22} \pm {0.08}$	${0.21} \pm {0.11}$	${0.13} \pm {0.02}$	${0.43} \pm {0.07}$	${0.37} \pm {0.15}$	${0.33} \pm {0.06}$	${0.18} \pm {0.02}$	${0.26} \pm {0.06}$
SkiMo	${0.75} \pm {0.09}$	${0.62} \pm {0.05}$	${0.47} \pm {0.03}$	${0.58} \pm {0.14}$	${0.34} \pm {0.06}$	${0.62} \pm {0.11}$	${0.56} \pm {0.11}$	${0.57} \pm {0.08}$	${0.21} \pm {0.04}$	${0.39} \pm {0.13}$
策略排序	${0.89} \pm {0.08}$	${0.82} \pm {0.09}$	${0.77} \pm {0.12}$	${0.63} \pm {0.28}$	${0.45} \pm {0.18}$	${0.61} \pm {0.14}$	${0.70} \pm {0.16}$	${0.53} \pm {0.11}$	${0.36} \pm {0.09}$	${0.44} \pm {0.09}$
T-STAR	${0.92} \pm {0.02}$	${0.90} \pm {0.04}$	${0.89} \pm {0.04}$	${0.90} \pm {0.07}$	${0.71} \pm {0.21}$	${0.77} \pm {0.09}$	${0.85} \pm {0.09}$	${0.68} \pm {0.13}$	${0.48} \pm {0.08}$	${0.58} \pm {0.10}$
无Bi的SCaR	${0.93} \pm {0.04}$	${0.92} \pm {0.02}$	${0.91} \pm {0.01}$	${0.93} \pm {0.02}$	${0.80} \pm {0.10}$	${0.79} \pm {0.02}$	${0.88} \pm {0.05}$	${0.75} \pm {0.08}$	${0.57} \pm {0.14}$	${0.66} \pm {0.09}$
无AES的SCaR	${0.95} \pm {0.03}$	${0.94} \pm {0.03}$	${0.93} \pm {0.02}$	${0.95} \pm {0.04}$	${0.85} \pm {0.06}$	${0.80} \pm {0.03}$	${0.91} \pm {0.05}$	${0.77} \pm {0.07}$	${0.61} \pm {0.13}$	${0.74} \pm {0.05}$
SCaR（本方法）	${0.98} \pm {0.02}$	${0.96} \pm {0.04}$	${0.95} \pm {0.03}$	${0.97} \pm {0.03}$	${0.92} \pm {0.05}$	$\mathbf{{0.88}} \pm {0.02}$	$\mathbf{{0.94}} \pm {0.03}\left({12}\% \uparrow \right)$	${0.84} \pm {0.16}$	${0.73} \pm {0.17}$	$\mathbf{{0.78}} \pm {0.12}\left({18}\% \uparrow \right)$

Scratch Training: 1) PPO is a model-free RL algorithm [47] that utilizes environmental rewards to learn tasks from scratch. 2) GAIL [39] is an adversarial imitation learning method to learn tasks from scratch, with a trained discriminator for distinguishing state-action distributions of experts and agents. 3) Fixed-RL-IL [40] uses fixed-weight environmental rewards and GAIL rewards to train policies from scratch. 4) SkiMo [46] is a model-based hierarchical RL approach that learns dynamic skill models for predicting outcomes in downstream tasks, which is used to test if modularly skill chaining method can surpass model-based scratch-training method on long-horizon tasks.
从零训练：1）PPO是一种无模型强化学习（RL）算法 [47]，通过环境奖励从零学习任务。2）GAIL [39]是一种对抗式模仿学习方法，通过训练判别器区分专家与智能体的状态-动作分布，从零学习任务。3）Fixed-RL-IL [40]结合固定权重的环境奖励和GAIL奖励，从零训练策略。4）SkiMo [46]是一种基于模型的分层强化学习方法，学习动态技能模型以预测下游任务的结果，用于测试模块化技能链方法在长时序任务上是否能超越基于模型的从零训练方法。

Skill Chaining: 1) Policy Sequencing [12] focuses on sequentially expanding the initial sets in skill chaining. 2) T-STAR [15] incorporates a discriminator to uni-directionally regularize the terminal states of sub-skills in a skill chaining. 3) SCaR w/o Bi reference to T-STAR during the fine-tuning phase, only uni-directional regularization of the terminal state set is performed to verify the validity of the proposed bi-directional regularization. 4) SCaR w/o AES fixes the scales of the two reward terms at 0.5 at all times to verify the effectiveness of the proposed AES regularization.
技能链：1）Policy Sequencing（策略序列）[12]关注于在技能链中依次扩展初始集合。2）T-STAR [15]引入判别器，对子技能的终止状态进行单向正则化。3）SCaR w/o Bi在微调阶段仅进行终止状态集合的单向正则化，以验证所提出的双向正则化的有效性。4）SCaR w/o AES将两个奖励项的权重始终固定为0.5，以验证所提出的AES正则化的有效性。

12.3 5.2 Quantitative Results

12.4 5.2 定量结果

Sub-task Skill Learning Performance First, we evaluate the proposed adaptive sub-task skill learning scheme in the sub-tasks of furniture assembly and kitchen organization. Specifically, we treat each sub-task as a separate task for policy learning and take the success rate of the trained policy tested in the reset sub-task as the criterion. All methods are trained in each sub-task with 5 random seeds, 150 million environment steps, and evaluated with the average success rate over 100 testing episodes. As shown in the Fig. 5, in chair_ingolf and Extended Kitchen tasks, even with the increase of objects in the environment and the increase of unpredictable perturbations, our proposed adaptive skill learning learns good sub-

task skills and consistently maintains a task success rate of more than 85% in all stages of the sub-task. In contrast, the PPO (only RL rewards), GAIL (only IL rewards), and Fixed-RL-IL (fixed RL and IL reward weights) baselines fail to maintain good sub-task success rates as the number of sub-task stages increases. This result well validates that our proposed adaptive weighted reward function based on AES regularization enhances intra-skill dependencies for multi-stage sub-task learning and brings effectiveness and stability.
子任务技能学习表现 首先，我们在家具组装和厨房整理的子任务中评估所提出的自适应子任务技能学习方案。具体而言，我们将每个子任务视为独立任务进行策略学习，并以训练好的策略在重置子任务中的成功率作为评判标准。所有方法均在每个子任务中使用5个随机种子、1.5亿环境步数进行训练，并在100次测试中取平均成功率进行评估。如图5所示，在chair_ingolf和Extended Kitchen任务中，即使环境中的物体数量增加且不可预测的扰动增多，我们提出的自适应技能学习依然能够学习到良好的子任务技能，并在子任务的各个阶段持续保持超过85%的任务成功率。相比之下，PPO（仅RL奖励）、GAIL（仅IL奖励）和Fixed-RL-IL（固定RL和IL奖励权重）等基线方法，随着子任务阶段数量增加，无法维持良好的子任务成功率。该结果充分验证了我们提出的基于AES正则化的自适应加权奖励函数能够增强多阶段子任务学习中的技能内部依赖性，并带来有效性与稳定性。

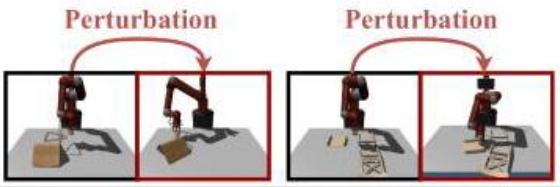
Long-horizon Execution Performance We then demonstrate the performance of SCaR in performing 8 long-horizon tasks in IKEA furniture assembly and kitchen organization. Table 1 shows the mean and standard deviation for these 8 tasks across 200 testing episodes with 5 different seeds. The PPO and GAIL baselines show minimal success on tasks with 4 and 5 sub-tasks, indicating the difficulty of learning complex multi-stage tasks solely from reward signals or expert demonstrations. The fixed RL-IL baseline, although improved compared to PPO and GAIL, mostly completed only one sub-task, which highlights the limitations of using fixed RL and IL reward weights in long-horizon tasks. While SkiMo achieves better success rates than model-free methods by building dynamic skill models, its performance remains inconsistent on long-horizon tasks due to its scratch learning nature. The performance of these scratch baselines demonstrates the importance of effective staged sub-task learning for long-horizon tasks. The results in Table 1 further highlight the superiority of the SCaR framework. By reinforcing intra- and inter-skill dependencies, task success rates are considerably higher than previous skill chaining approaches such as Policy Sequencing and T-STAR, which primarily address uni-directional inter-skill dependencies. Compared to T-STAR, SCaR increases average success rates by more than 12% on six furniture assembly tasks and 18% on two kitchen tasks [4]
长时序执行表现 接下来，我们展示SCaR在宜家家具组装和厨房整理的8个长时序任务中的表现。表1展示了这8个任务在5个不同种子下、200次测试中的均值和标准差。PPO和GAIL基线在包含4个和5个子任务的任务上几乎没有成功，表明仅依靠奖励信号或专家演示难以学习复杂的多阶段任务。固定RL-IL基线虽然较PPO和GAIL有所提升，但大多只完成了一个子任务，凸显了在长时序任务中使用固定RL和IL奖励权重的局限性。SkiMo通过构建动态技能模型，相较无模型方法取得了更高的成功率，但由于其从零学习的特性，在长时序任务上的表现仍不稳定。这些从零训练基线的表现说明了有效分阶段子任务学习对于长时序任务的重要性。表1中的结果进一步突出SCaR框架的优越性。通过强化技能内部和技能间的依赖性，任务成功率显著高于以往的技能链方法，如Policy Sequencing和T-STAR，这些方法主要处理单向技能间依赖性。与T-STAR相比，SCaR在六个家具组装任务上的平均成功率提升超过12%，在两个厨房任务上提升了18%[4]

12.5 5.3 Robustness to Perturbations

12.6 5.3 对扰动的鲁棒性

Perturbation tests are conducted to evaluate the robustness of skill chaining for two furniture assembly tasks. As shown in the top figure of Table 2, for the chair_bernhard task, the perturbation involves applying external joint torque to the robotic arm, moving the chair back before assembling the second support. For the chair_ingolf task, the perturbation is applied by exerting external torque on the robotic arms, causing them to move slightly before mounting the assembled chair seat to the chair back. The results in Table 2 highlight the detrimental impact of environmental perturbations on the success rates of baseline methods during the execution of multiple sub-task skills. Methods like Policy Sequencing and T-STAR, which focus solely on inter-skill dependencies through uni-directional regularization, struggle to complete tasks after perturbations. In contrast, SCaR, demonstrates more robust performance even under unseen perturbations. These results further support the advantages of our proposed dual regularization for stable skill chaining on long-horizon manipulation tasks.
进行了扰动测试以评估技能链在两项家具组装任务中的鲁棒性。如表2顶部图所示，对于chair_bernhard任务，扰动包括对机械臂施加外部关节力矩，在组装第二个支撑件之前移动椅背。对于chair_ingolf任务，扰动通过对机械臂施加外部力矩实现，导致机械臂在将组装好的椅座安装到椅背之前发生轻微移动。表2的结果突出显示了环境扰动对基线方法在执行多个子任务技能时成功率的负面影响。诸如Policy Sequencing和T-STAR等仅通过单向正则化关注技能间依赖的方法，在扰动后难以完成任务。相比之下，SCaR即使在未见过的扰动下也表现出更强的鲁棒性。这些结果进一步支持了我们提出的双重正则化在长时域操作任务中实现稳定技能链的优势。

Table 2: Comparison of the robustness of skill chaining in perturbed environments.
表2：扰动环境下技能链鲁棒性的比较。



Method	*chair_bernhard*		*chair_ingolf*	
	No Perturb	**Perturb**	**No Perturb**	**Perturb**
Policy Sequencing	0.82 ± 0.09	0.51 ± 0.04	0.77 ± 0.12	0.50 ± 0.10
T-STAR	0.90 ± 0.04	0.60 ± 0.08	0.89 ± 0.04	0.59 ± 0.04
SCaR w/o Bi	0.92 ± 0.02	0.65 ± 0.11	0.91 ± 0.01	0.63 ± 0.05
SCaR w/o AES	0.94 ± 0.03	0.74 ± 0.09	0.93 ± 0.02	0.71 ± 0.07
SCaR (Ours)	**0.96** ± 0.04	**0.85** ± 0.11	**0.95** ± 0.03	**0.80** ± 0.13

12.7 5.4 Ablations and Analysis

12.8 5.4 消融实验与分析

We perform ablation studies to explore the important factors that affect the performance of SCaR.
我们进行了消融研究，以探讨影响SCaR性能的重要因素。

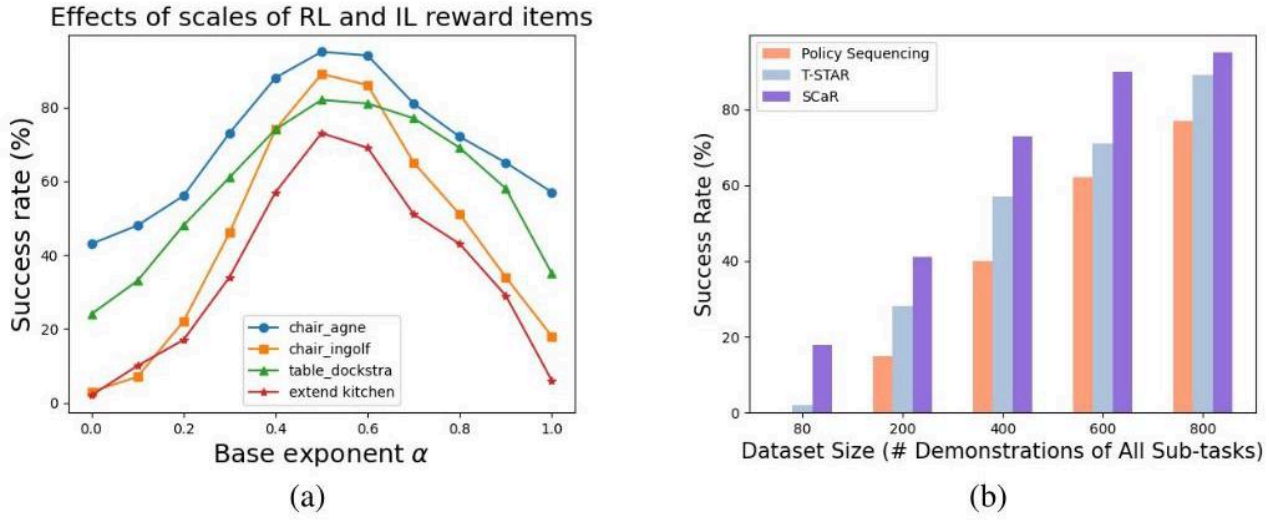


Figure 6: Ablation experiments.
图6：消融实验。

Modular Ablation We investigate how the adaptive sub-task skill learning and bi-directional adversarial learning impact skill chaining through SCaR w/o Bi and SCaR w/o AES. As shown in Table 1,without bi-directional regularization, SCaR w/o Bi experiences significant performance drops in tasks with more than two sub-tasks but still outperforms T-STAR. This is because SCaR w/o Bi maintains the adaptive scheduling of AES during sub-task skill learning,underscoring the importance of focusing on the intra-skill dependencies between successive actions. Similarly, the absence of AES regularization reduces SCaR w/o AES's performance, though it still maintains stable outcomes. This underscores the importance of reinforcing inter-skill dependencies on long-horizon tasks and reaffirms the contribution of bi-directional regularization. As shown in Table 2, SCaR w/o Bi, though slightly more robust than T-STAR due to the presence of AES, still faces challenges in adapting to perturbations and maintaining stable skill chaining because of its uni-directional fine-tuning limitations. SCaR w/o AES manages to maintain a certain level of performance stability under perturbations, thanks to bi-directional regularization, which ensures the bi-directional alignment of initial and terminal states between skills. The results show that the pre-trained skills via AES exhibit enhanced intra-skill dependencies within sub-tasks, and

bi-directional regularization ensures stable long-horizon execution, even in the presence of perturbations, by reinforcing inter-skill dependencies.
模块消融 我们通过SCaR w/o Bi和SCaR w/o AES研究自适应子任务技能学习和双向对抗学习对技能链的影响。如表1所示，缺少双向正则化的SCaR w/o Bi在包含两个以上子任务的任务中性能显著下降，但仍优于T-STAR。这是因为SCaR w/o Bi在子任务技能学习过程中保持了AES的自适应调度，强调了关注连续动作间技能内依赖的重要性。同样，缺少AES正则化的SCaR w/o AES性能有所下降，但仍保持稳定结果。这强调了在长时域任务中强化技能间依赖的重要性，并再次确认了双向正则化的贡献。如表2所示，尽管由于AES的存在，SCaR w/o Bi比T-STAR稍显鲁棒，但由于其单向微调的限制，仍面临适应扰动和维持稳定技能链的挑战。得益于双向正则化，SCaR w/o AES在扰动下能够维持一定的性能稳定性，确保技能间初始和终止状态的双向对齐。结果表明，通过AES预训练的技能在子任务内表现出增强的技能内依赖，双向正则化通过强化技能间依赖，确保了即使在扰动存在时的长时域稳定执行。

---

[4] The overall increase is somewhat modest due to averaging the success rates of the 2,3,and 4 sub-tasks and the 4 and 5 sub-tasks, respectively.
[4] 总体提升较为有限，原因是分别对2、3和4个子任务以及4和5个子任务的成功率进行了平均。

Parametric Ablation We further investigate the impact of different scales of RL and IL reward terms, as well as the size of expert demonstration datasets. The effect of varying the base exponent α on task success rates is tested across four tasks: chair_agne,chair_ingolf,table_dockstra,and extend kitchen. As depicted in Fig. 6(a), SCaR achieves the highest success rates in all four tasks when $\alpha = 0.5$,indicating a balance between RL and IL at the beginning of learning. When α becomes smaller, emphasizing IL at the start, performance decreases more steeply. Conversely, as α becomes larger,giving more weight to RL,performance also declines but at a slower rate. We also evaluate the impact of different sizes of expert datasets on three skill chaining methods: Policy Sequencing, T-STAR, and SCaR, specifically in the chair_ingolf task. We vary the overall task expert data size from 80, 120, 200, 400, 600, to 800 demos. As shown in Fig. 6(b), the results indicate significant performance improvement when increasing the dataset size from 400 to 800 demos, while the improvement is less pronounced when going from 80 to 120 demos. This demonstrates the importance of the demo dataset size in the effectiveness of data-driven approaches like skill chaining.
参数消融 我们进一步研究了不同规模的强化学习（RL）和模仿学习（IL）奖励项，以及专家示范数据集大小的影响。基底指数α对任务成功率的影响在四个任务中测试：chair_agne、chair_ingolf、table_dockstra和extend kitchen。如图6(a)所示，当$\alpha = 0.5$时，SCaR在所有四个任务中均达到最高成功率，表明在学习初期RL与IL之间取得了平衡。当α变小时，强调学习初期的IL，性能下降更为陡峭。相反，当α变大，赋予RL更大权重时，性能也下降但速度较慢。我们还评估了不同专家数据集大小对三种技能链方法（Policy Sequencing、T-STAR和SCaR）在chair_ingolf任务中的影响。我们将整体任务专家数据量从80、120、200、400、600增加到800个示范。如图6(b)所示，结果显示当数据集规模从400增加到800时性能显著提升，而从80增加到120时提升不明显。这表明示范数据集规模对数据驱动方法如技能链的有效性具有重要影响。

13 6 Discussion

14 6 讨论

Limitation and future directions The primary limitation of our work is that the sub-task division for long-horizon tasks is predefined and does not incorporate visual or semantic processing of objects. Expanding our framework to handle longer-horizon visual manipulation tasks is a direction we aim to explore in future research. For example, integrating a more scalable architecture [48] and performing large-scale pre-training on extensive datasets [49, 50] are promising directions. Another avenue worth exploring is applying our framework to real-world robotic furniture assembly tasks, rather than only staged pick-and-place tasks. Constructing a deployment environment for real-world furniture assembly and ensuring the complete insertion of each furniture module presents significant challenges We discuss additional limitations and potential solutions in further detail in Appendix J.
局限性与未来方向 我们工作的主要局限在于长时序任务的子任务划分是预定义的，且未结合对物体的视觉或语义处理。未来研究中，我们计划扩展框架以处理更长时序的视觉操作任务。例如，集成更具扩展性的架构[48]，并在大规模数据集上进行大规模预训练[49, 50]，是有前景的方向。另一个值得探索的方向是将我们的框架应用于真实机器人家具组装任务，而不仅限于分阶段的拾取与放置任务。构建真实家具组装的部署环境并确保每个家具模块的完全插入，面临重大挑战。我们在附录J中详细讨论了更多局限性及潜在解决方案。

Conclusion In this paper, we introduce SCaR, a novel skill chaining framework that ensures smooth and stable execution of long-horizon robotic manipulation tasks via dual regularization within and between sub-task skills. Extensive experiments demonstrate that the SCaR framework achieves better task success rates than the baseline methods in both simulated and real-robot manipulation tasks, while being robust against perturbations. We hope this work will inspire future research to further explore the potential of skill chaining for long-horizon robotic manipulation.
结论 本文提出了SCaR，一种新颖的技能链框架，通过子任务技能内外的双重正则化，确保长时序机器人操作任务的平滑且稳定执行。大量实验表明，SCaR框架在模拟和真实机器人操作任务中均优于基线方法，且对扰动具有较强鲁棒性。我们希望本工作能激发未来研究，进一步挖掘技能链在长时序机器人操作中的潜力。

15 Acknowledgments and Disclosure of Funding

16 致谢与资金披露

This work was supported in part by the Science and Technology Innovation 2030 New Generation Artificial Intelligence Major Project under Grant 2021ZD0113303; in part by the National Natural Science Foundation of China under Grant 62192783, Grant 62276128; in part by the Collaborative Innovation Center of Novel Software Technology and Industrialization; in part by the Fundamental Research Funds for the Central Universities under Grant 14380128; in part by the Postgraduate Research & Practice Innovation Program of Jiangsu Province under Grant KYCX24_0263. References

本工作部分由“科技创新2030新一代人工智能重大项目”资助，项目编号2021ZD0113303；部分由国家自然科学基金资助，项目编号62192783、62276128；部分由新型软件技术与产业化协同创新中心资助；部分由中央高校基本科研业务费资助，项目编号14380128；部分由江苏省研究生科研与实践创新计划资助，项目编号KYCX24_0263。参考文献

[1] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. The Journal of Machine Learning Research, 17(1):1334-1373, 2016.
[1] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. 端到端训练深度视觉运动策略。机器学习研究杂志(The Journal of Machine Learning Research), 17(1):1334-1373, 2016.

[2] Francisco Suárez-Ruiz and Quang-Cuong Pham. A framework for fine robotic assembly. In 2016 IEEE international conference on robotics and automation (ICRA), pages 421-426. IEEE, 2016.
[2] Francisco Suárez-Ruiz 和 Quang-Cuong Pham. 精细机器人组装框架。载于2016年IEEE国际机器人与自动化会议(ICRA), 页421-426。IEEE, 2016.

[3] Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. Robotics: Science and Systems XIV, 2018.
[3] Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, 和 Sergey Levine. 通过深度强化学习与示范学习复杂灵巧操作。机器人科学与系统十四(Robotics: Science and Systems XIV), 2018.

[4] Divye Jain, Andrew Li, Shivam Singhal, Aravind Rajeswaran, Vikash Kumar, and Emanuel Todorov. Learning deep visuomotor policies for dexterous hand manipulation. In 2019 international conference on robotics and automation (ICRA), pages 3636-3643. IEEE, 2019.
[4] Divye Jain, Andrew Li, Shivam Singhal, Aravind Rajeswaran, Vikash Kumar, 和 Emanuel Todorov. 学习深度视觉运动策略以实现灵巧手操作。载于2019年国际机器人与自动化会议(ICRA), 页3636-3643。IEEE, 2019.

[5] George Konidaris, Scott Kuindersma, Roderic Grupen, and Andrew Barto. Robot learning from demonstration by constructing skill trees. The International Journal of Robotics Research, 31(3):360-375, 2012.
[5] George Konidaris, Scott Kuindersma, Roderic Grupen, 和 Andrew Barto. 通过构建技能树实现机器人示范学习。国际机器人研究杂志(The International Journal of Robotics Research), 31(3):360-375, 2012.

[6] Thomas Kipf, Yujia Li, Hanjun Dai, Vinicius Zambaldi, Alvaro Sanchez-Gonzalez, Edward Grefenstette, Pushmeet Kohli, and Peter Battaglia. Compile: Compositional imitation learning and execution. In International Conference on Machine Learning, pages 3418-3428. PMLR, 2019.
[6] Thomas Kipf, Yujia Li, Hanjun Dai, Vinicius Zambaldi, Alvaro Sanchez-Gonzalez, Edward Grefenstette, Pushmeet Kohli, 和 Peter Battaglia. COMPILE：组合模仿学习与执行。载于国际机器学习会议(International Conference on Machine Learning), 页3418-3428。PMLR, 2019.

[7] Chen Wang, Linxi Fan, Jiankai Sun, Ruohan Zhang, Li Fei-Fei, Danfei Xu, Yuke Zhu, and Anima Anandkumar. Mimicplay: Long-horizon imitation learning by watching human play. arXiv preprint arXiv:2302.12422, 2023.
[7] Chen Wang, Linxi Fan, Jiankai Sun, Ruohan Zhang, Li Fei-Fei, Danfei Xu, Yuke Zhu, 和 Anima Anandkumar. MimicPlay：通过观察人类游戏实现长时序模仿学习。arXiv预印本 arXiv:2302.12422, 2023.

[8] Marcin Andrychowicz, Anton Raichuk, Piotr Stańczyk, Manu Orsini, Sertan Girgin, Raphaël Marinier, Leonard Hussenot, Matthieu Geist, Olivier Pietquin, Marcin Michalski, et al. What matters for on-policy deep actor-critic methods? a large-scale study. In International conference on learning representations, 2020.
[8] Marcin Andrychowicz, Anton Raichuk, Piotr Stańczyk, Manu Orsini, Sertan Girgin, Raphaël Marinier, Leonard Hussenot, Matthieu Geist, Olivier Pietquin, Marcin Michalski 等. 对策略内深度演员-评论家方法的重要性的大规模研究。载于学习表征国际会议(International conference on learning representations), 2020.

[9] Yuyang Liu, Weijun Dong, Yingdong Hu, Chuan Wen, Zhao-Heng Yin, Chongjie Zhang, and Yang Gao. Imitation learning from observation with automatic discount scheduling. arXiv preprint arXiv:2310.07433, 2023.
[9] Yuyang Liu, Weijun Dong, Yingdong Hu, Chuan Wen, Zhao-Heng Yin, Chongjie Zhang, 和 Yang Gao. 通过自动折扣调度实现的观察模仿学习。arXiv预印本 arXiv:2310.07433, 2023.

[10] Caelan Reed Garrett, Rohan Chitnis, Rachel Holladay, Beomjoon Kim, Tom Silver, Leslie Pack Kaelbling, and Tomás Lozano-Pérez. Integrated task and motion planning. Annual review of control, robotics, and autonomous systems, 4:265-293, 2021.
[10] Caelan Reed Garrett, Rohan Chitnis, Rachel Holladay, Beomjoon Kim, Tom Silver, Leslie Pack Kaelbling, 和 Tomás Lozano-Pérez. 集成任务与运动规划. 控制、机器人与自主系统年评, 4:265-293, 2021.

[11] George Konidaris and Andrew Barto. Skill discovery in continuous reinforcement learning domains using skill chaining. Advances in neural information processing systems, 22, 2009.
[11] George Konidaris 和 Andrew Barto. 在连续强化学习领域中通过技能链发现技能. 神经信息处理系统进展, 22, 2009.

[12] Alexander Clegg, Wenhao Yu, Jie Tan, C Karen Liu, and Greg Turk. Learning to dress: Synthesizing human dressing motion via deep reinforcement learning. ACM Transactions on Graphics (TOG), 37(6):1-10, 2018.
[12] Alexander Clegg, Wenhao Yu, Jie Tan, C Karen Liu, 和 Greg Turk. 学习穿衣：通过深度强化学习合成人类穿衣动作. ACM图形学汇刊 (TOG), 37(6):1-10, 2018.

[13] Youngwoon Lee, Shao-Hua Sun, Sriram Somasundaram, Edward S Hu, and Joseph J Lim. Composing complex skills by learning transition policies. In International Conference on Learning Representations, 2018.
[13] Youngwoon Lee, Shao-Hua Sun, Sriram Somasundaram, Edward S Hu, 和 Joseph J Lim. 通过学习转换策略组合复杂技能. 国际学习表征会议, 2018.

[14] Youngwoon Lee, Jingyun Yang, and Joseph J Lim. Learning to coordinate manipulation skills via skill behavior diversification. In International conference on learning representations, 2019.
[14] Youngwoon Lee, Jingyun Yang, 和 Joseph J Lim. 通过技能行为多样化学习协调操作技能. 国际学习表征会议, 2019.

[15] Youngwoon Lee, Joseph J Lim, Anima Anandkumar, and Yuke Zhu. Adversarial skill chaining for long-horizon robot manipulation via terminal state regularization. In Conference on Robot Learning (CoRL 2022), pages 406-416. PMLR, 2022.
[15] Youngwoon Lee, Joseph J Lim, Anima Anandkumar, 和 Yuke Zhu. 通过终端状态正则化实现长时域机器人操作的对抗技能链. 机器人学习会议 (CoRL 2022), 页406-416. PMLR, 2022.

[16] Jiayuan Gu, Devendra Singh Chaplot, Hao Su, and Jitendra Malik. Multi-skill mobile manipulation for object rearrangement. In The Eleventh International Conference on Learning Representations, 2022.
[16] Jiayuan Gu, Devendra Singh Chaplot, Hao Su, 和 Jitendra Malik. 多技能移动操作用于物体重新排列. 第十一届国际学习表征会议, 2022.

[17] Yuanpei Chen, Chen Wang, Li Fei-Fei, and Karen Liu. Sequential dexterity: Chaining dexterous policies for long-horizon manipulation. In Conference on Robot Learning, pages 3809-3829. PMLR, 2023.
[17] Yuanpei Chen, Chen Wang, Li Fei-Fei, 和 Karen Liu. 顺序灵巧：为长时域操作链式连接灵巧策略. 机器人学习会议, 页3809-3829. PMLR, 2023.

[18] Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. Artificial intelligence, 112(1- 2):181-211, 1999.
[18] Richard S Sutton, Doina Precup, 和 Satinder Singh. 在马尔可夫决策过程(MDPs)与半马尔可夫决策过程(semi-MDPs)之间：强化学习中时间抽象的框架. 人工智能, 112(1-2):181-211, 1999.

[19] Jürgen Schmidhuber. Towards compositional learning with dynamic neural networks. Inst. für Informatik, 1990.
[19] Jürgen Schmidhuber. 面向动态神经网络的组合学习. 信息学研究所, 1990.

[20] Pierre-Luc Bacon, Jean Harb, and Doina Precup. The option-critic architecture. In Proceedings of the AAAI conference on artificial intelligence, volume 31, 2017.
[20] Pierre-Luc Bacon, Jean Harb, 和 Doina Precup. 选项-批评者架构. 美国人工智能协会(AAAI)会议论文集, 卷31, 2017.

[21] Ofir Nachum, Shixiang Shane Gu, Honglak Lee, and Sergey Levine. Data-efficient hierarchical reinforcement learning. Advances in neural information processing systems, 31, 2018.
[21] Ofir Nachum, Shixiang Shane Gu, Honglak Lee, 和 Sergey Levine. 数据高效的分层强化学习. 神经信息处理系统进展, 31, 2018.

[22] Andrew Levy, George Konidaris, Robert Platt, and Kate Saenko. Learning multi-level hierarchies with hindsight. arXiv preprint arXiv:1712.00948, 2017.
[22] Andrew Levy, George Konidaris, Robert Platt, 和 Kate Saenko. 利用回顾学习多层次层级结构. arXiv预印本 arXiv:1712.00948, 2017.

[23] Visak CV Kumar, Sehoon Ha, and C Karen Liu. Expanding motor skills using relay networks. In Conference on Robot Learning, pages 744-756. PMLR, 2018.
[23] Visak CV Kumar, Sehoon Ha, 和 C Karen Liu. 使用中继网络扩展运动技能. 机器人学习会议, 页744-756. PMLR, 2018.

[24] Yuchen Lu, Yikang Shen, Siyuan Zhou, Aaron Courville, Joshua B Tenenbaum, and Chuang Gan. Learning task decomposition with ordered memory policy network. In International Conference on Learning Representations, 2020.
[24] Yuchen Lu, Yikang Shen, Siyuan Zhou, Aaron Courville, Joshua B Tenenbaum, 和 Chuang Gan. 通过有序记忆策略网络学习任务分解. 国际学习表征会议, 2020.

[25] Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, et al. Do as i can, not as i say: Grounding language in robotic affordances. In Conference on Robot Learning, pages 287-318. PMLR, 2023.
[25] Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, 等. 按我所能行事，而非我所言：将语言基础植入机器人可供性. 机器人学习会议, 页287-318. PMLR, 2023.

[26] Tejas D Kulkarni, Karthik Narasimhan, Ardavan Saeedi, and Josh Tenenbaum. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. Advances in neural information processing systems, 29, 2016.
[26] Tejas D Kulkarni, Karthik Narasimhan, Ardavan Saeedi, 和 Josh Tenenbaum. 分层深度强化学习：整合时间抽象和内在动机。神经信息处理系统进展，29，2016年。

[27] Junhyuk Oh, Satinder Singh, Honglak Lee, and Pushmeet Kohli. Zero-shot task generalization with multi-task deep reinforcement learning. In International Conference on Machine Learning, pages 2661-2670. PMLR, 2017.
[27] Junhyuk Oh, Satinder Singh, Honglak Lee, 和 Pushmeet Kohli. 通过多任务深度强化学习实现零样本任务泛化。国际机器学习大会论文集，第2661-2670页。PMLR，2017年。

[28] Josh Merel, Arun Ahuja, Vu Pham, Saran Tunyasuvunakool, Siqi Liu, Dhruva Tirumala, Nicolas Heess, and Greg Wayne. Hierarchical visuomotor control of humanoids. In International Conference on Learning Representations, 2018.
[28] Josh Merel, Arun Ahuja, Vu Pham, Saran Tunyasuvunakool, Siqi Liu, Dhruva Tirumala, Nicolas Heess, 和 Greg Wayne. 人形机器人分层视觉运动控制。国际表征学习会议，2018年。

[29] Chen Wang, Danfei Xu, and Li Fei-Fei. Generalizable task planning through representation pretraining. IEEE Robotics and Automation Letters, 7(3):8299-8306, 2022.
[29] Chen Wang, Danfei Xu, 和 Li Fei-Fei. 通过表示预训练实现可泛化的任务规划。IEEE机器人与自动化快报，7(3):8299-8306，2022年。

[30] Xue Bin Peng, Michael Chang, Grace Zhang, Pieter Abbeel, and Sergey Levine. Mcp: Learning composable hierarchical control with multiplicative compositional policies. Advances in Neural Information Processing Systems, 32, 2019.
[30] Xue Bin Peng, Michael Chang, Grace Zhang, Pieter Abbeel, 和 Sergey Levine. MCP：通过乘法组合策略学习可组合的分层控制。神经信息处理系统进展，32，2019年。

[31] Akhil Bagaria and George Konidaris. Option discovery using deep skill chaining. In International Conference on Learning Representations, 2019.
[31] Akhil Bagaria 和 George Konidaris. 使用深度技能链发现选项。国际表征学习会议，2019年。

[32] Zixuan Chen, Ze Ji, Shuyang Liu, Jing Huo, Yiyu Chen, and Yang Gao. Cognizing and imitating robotic skills via a dual cognition-action architecture. In Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, pages 2204-2206, 2024.
[32] Zixuan Chen, Ze Ji, Shuyang Liu, Jing Huo, Yiyu Chen, 和 Yang Gao. 通过双重认知-动作架构认知与模仿机器人技能。第23届自治代理与多智能体系统国际会议论文集，第2204-2206页，2024年。

[33] Peter Pastor, Heiko Hoffmann, Tamim Asfour, and Stefan Schaal. Learning and generalization of motor skills by learning from demonstration. In 2009 IEEE International Conference on Robotics and Automation, pages 763-768. IEEE, 2009.
[33] Peter Pastor, Heiko Hoffmann, Tamim Asfour, 和 Stefan Schaal. 通过示范学习运动技能及其泛化。2009年IEEE国际机器人与自动化会议论文集，第763-768页。IEEE，2009年。

[34] Jens Kober, Jan Peters, Jens Kober, and Jan Peters. Movement templates for learning of hitting and batting. Learning Motor Skills: From Algorithms to Robot Experiments, pages 69-82, 2014.
[34] Jens Kober, Jan Peters, Jens Kober, 和 Jan Peters. 击打和挥棒的运动模板学习。《运动技能学习：从算法到机器人实验》，第69-82页，2014年。

[35] Katharina Mülling, Jens Kober, Oliver Kroemer, and Jan Peters. Learning to select and generalize striking movements in robot table tennis. The International Journal of Robotics Research, 32(3):263-279, 2013.
[35] Katharina Mülling, Jens Kober, Oliver Kroemer, 和 Jan Peters. 学习选择和泛化机器人乒乓球击打动作。国际机器人研究杂志，32(3):263-279，2013年。

[36] Karol Hausman, Jost Tobias Springenberg, Ziyu Wang, Nicolas Heess, and Martin Riedmiller. Learning an embedding space for transferable robot skills. In International Conference on Learning Representations, 2018.
[36] Karol Hausman, Jost Tobias Springenberg, Ziyu Wang, Nicolas Heess, 和 Martin Riedmiller. 学习可迁移机器人技能的嵌入空间。国际表征学习会议，2018年。

[37] Utkarsh Aashu Mishra, Shangjie Xue, Yongxin Chen, and Danfei Xu. Generative skill chaining: Long-horizon skill planning with diffusion models. In Conference on Robot Learning, pages 2905-2925. PMLR, 2023.
[37] Utkarsh Aashu Mishra, Shangjie Xue, Yongxin Chen, 和 Danfei Xu. 生成式技能链：基于扩散模型的长时域技能规划。机器人学习会议论文集，第2905-2925页。PMLR，2023年。

[38] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. Journal of artificial intelligence research, 4:237-285, 1996.
[38] Leslie Pack Kaelbling, Michael L Littman, 和 Andrew W Moore. 强化学习综述。人工智能研究杂志，4:237-285，1996年。

[39] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. Advances in neural information processing systems, 29, 2016.
[39] Jonathan Ho 和 Stefano Ermon. 生成对抗模仿学习。神经信息处理系统进展，29，2016年。

[40] Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. ACM Transactions on Graphics (ToG), 40(4):1-20, 2021.
[40] Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, 和 Angjoo Kanazawa. AMP：用于风格化基于物理的角色控制的对抗运动先验。ACM图形学汇刊（ToG），40(4):1-20，2021年。

[41] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In Proceedings of the IEEE international conference on computer vision, pages 2794-2802, 2017.
[41] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, 和 Stephen Paul Smolley. 最小二乘生成对抗网络。IEEE国际计算机视觉会议论文集，第2794-2802页，2017年。

[42] Naveen Kodali, Jacob Abernethy, James Hays, and Zsolt Kira. How to train your dragan. arXiv preprint arXiv:1705.07215, 2(4), 2017.
[42] Naveen Kodali, Jacob Abernethy, James Hays, 和 Zsolt Kira. 如何训练你的dragan。arXiv预印本 arXiv:1705.07215, 2(4), 2017.

[43] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In International conference on machine learning, pages 3481-3490. PMLR, 2018.
[43] Lars Mescheder, Andreas Geiger, 和 Sebastian Nowozin. 哪些GANs的训练方法实际上会收敛？发表于国际机器学习大会，页码3481-3490。PMLR, 2018.

[44] Youngwoon Lee, Edward S Hu, and Joseph J Lim. Ikea furniture assembly environment for long-horizon complex manipulation tasks. In 2021 ieee international conference on robotics and automation (icra), pages 6343-6349. IEEE, 2021.
[44] Youngwoon Lee, Edward S Hu, 和 Joseph J Lim. 宜家家具组装环境用于长时域复杂操作任务。发表于2021年IEEE国际机器人与自动化会议（ICRA），页码6343-6349。IEEE, 2021.

[45] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. arXiv preprint arXiv:2004.07219, 2020.
[45] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, 和 Sergey Levine. D4RL：用于深度数据驱动强化学习的数据集。arXiv预印本 arXiv:2004.07219, 2020.

[46] Lucy Xiaoyang Shi, Joseph J Lim, and Youngwoon Lee. Skill-based model-based reinforcement learning. In Conference on Robot Learning, pages 2262-2272. PMLR, 2023.
[46] Lucy Xiaoyang Shi, Joseph J Lim, 和 Youngwoon Lee. 基于技能的模型驱动强化学习。发表于机器人学习会议，页码2262-2272。PMLR, 2023.

[47] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017.
[47] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, 和 Oleg Klimov. 近端策略优化算法。arXiv预印本 arXiv:1707.06347, 2017.

[48] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In International Conference on Machine Learning, 2021.
[48] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, 和 Joao Carreira. Perceiver：基于迭代注意力的通用感知。发表于国际机器学习大会，2021.

[49] Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. Robonet: Large-scale multi-robot learning. In Conference on Robot Learning, 2019.
[49] Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, 和 Chelsea Finn. Robonet：大规模多机器人学习。发表于机器人学习会议，2019.

[50] Abhishek Padalkar, Acorn Pooley, Ajinkya Jain, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anikait Singh, Anthony Brohan, et al. Open x-embodiment: Robotic learning datasets and rt-x models. arXiv preprint arXiv:2310.08864, 2023.
[50] Abhishek Padalkar, Acorn Pooley, Ajinkya Jain, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anikait Singh, Anthony Brohan, 等. Open x-embodiment：机器人学习数据集和RT-X模型。arXiv预印本 arXiv:2310.08864, 2023.

[51] Siddhant Haldar, Vaibhav Mathur, Denis Yarats, and Lerrel Pinto. Watch and match: Supercharging imitation with regularized optimal transport. In Conference on Robot Learning, pages 32-43. PMLR, 2023.
[51] Siddhant Haldar, Vaibhav Mathur, Denis Yarats, 和 Lerrel Pinto. 观察与匹配：通过正则化最优传输强化模仿学习。发表于机器人学习会议，页码32-43。PMLR, 2023.

[52] Abhishek Gupta, Vikash Kumar, Corey Lynch, Sergey Levine, and Karol Hausman. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning. In Conference on Robot Learning, pages 1025-1037. PMLR, 2020.
[52] Abhishek Gupta, Vikash Kumar, Corey Lynch, Sergey Levine, 和 Karol Hausman. 继电策略学习：通过模仿和强化学习解决长时域任务。发表于机器人学习会议，页码1025-1037。PMLR, 2020.

[53] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.
[53] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, 和 Adam Lerer. PyTorch中的自动微分。2017.

[54] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In International conference on machine learning, pages 9118-9147. PMLR, 2022.
[54] Wenlong Huang, Pieter Abbeel, Deepak Pathak, 和 Igor Mordatch. 语言模型作为零样本规划器：为具身智能体提取可执行知识。发表于国际机器学习大会，页码9118-9147。PMLR, 2022.

[55] Yingdong Hu, Fanqi Lin, Tong Zhang, Li Yi, and Yang Gao. Look before you leap: Unveiling the power of gpt-4v in robotic vision-language planning. arXiv preprint arXiv:2311.17842, 2023.
[55] Yingdong Hu, Fanqi Lin, Tong Zhang, Li Yi, 和 Yang Gao. 三思而后行：揭示GPT-4V在机器人视觉语言规划中的强大能力。arXiv预印本 arXiv:2311.17842, 2023.

[56] Haoxu Huang, Fanqi Lin, Yingdong Hu, Shengjie Wang, and Yang Gao. Copa: General robotic manipulation through spatial constraints of parts with foundation models. arXiv preprint arXiv:2403.08248, 2024.
[56] Haoxu Huang, Fanqi Lin, Yingdong Hu, Shengjie Wang, 和 Yang Gao. COPA：通过基础模型利用部件空间约束实现通用机器人操作。arXiv预印本 arXiv:2403.08248, 2024.

[57] Kimin Lee, Laura Smith, and Pieter Abbeel. Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. arXiv preprint arXiv:2106.05091, 2021.
[57] Kimin Lee, Laura Smith, 和 Pieter Abbeel. PEBBLE：通过重标记经验和无监督预训练实现反馈高效的交互式强化学习。arXiv预印本 arXiv:2106.05091, 2021.

[58] Younggyo Seo, Junsu Kim, Stephen James, Kimin Lee, Jinwoo Shin, and Pieter Abbeel. Multiview masked world models for visual robotic manipulation. In International Conference on Machine Learning, pages 30613-30632. PMLR, 2023.
[58] Younggyo Seo, Junsu Kim, Stephen James, Kimin Lee, Jinwoo Shin, 和 Pieter Abbeel. 用于视觉机器人操作的多视角掩码世界模型。发表于国际机器学习大会，页码30613-30632。PMLR, 2023。

[59] Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. Croco: Self-supervised pre-training for 3 d vision tasks by cross-view completion. Advances in Neural Information Processing Systems, 35:3502-3516, 2022.
[59] Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, 和 Jérôme Revaud. Croco：通过跨视图补全进行自监督预训练的3 d视觉任务。神经信息处理系统进展，35:3502-3516, 2022。

[60] Yinghao Xu, Hao Tan, Fujun Luan, Sai Bi, Peng Wang, Jiahao Li, Zifan Shi, Kalyan Sunkavalli, Gordon Wetzstein, Zexiang Xu, et al. Dmv3d: Denoising multi-view diffusion using 3d large reconstruction model. arXiv preprint arXiv:2311.09217, 2023.
[60] Yinghao Xu, Hao Tan, Fujun Luan, Sai Bi, Peng Wang, Jiahao Li, Zifan Shi, Kalyan Sunkavalli, Gordon Wetzstein, Zexiang Xu 等. Dmv3d：利用三维大规模重建模型的多视角去噪扩散。arXiv预印本 arXiv:2311.09217, 2023。

[61] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. arXiv preprint arXiv:2311.04400, 2023.
[61] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, 和 Hao Tan. Lrm：单张图像到三维的大规模重建模型。arXiv预印本 arXiv:2311.04400, 2023。

[62] Dhruv Shah, Błażej Osiński, Sergey Levine, et al. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In Conference on robot learning, pages 492-504. PMLR, 2023.
[62] Dhruv Shah, Błażej Osiński, Sergey Levine 等. Lm-nav：基于大规模预训练语言、视觉与动作模型的机器人导航。发表于机器人学习会议，页码492-504。PMLR, 2023。

[63] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. arXiv preprint arXiv:2212.06817, 2022.
[63] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu 等. Rt-1：面向大规模现实控制的机器人变换器。arXiv预印本 arXiv:2212.06817, 2022。

[64] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. arXiv preprint arXiv:2303.03378, 2023.
[64] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu 等. Palm-e：一种具身多模态语言模型。arXiv预印本 arXiv:2303.03378, 2023。

17 Technical Appendix

18 技术附录

19 A Pseudo-code

20 伪代码

Pseudo-code for adaptive sub-task skill learning and bi-directional adversarial learning are shown in Algorithm 1 and Algorithm 2 respectively. We highlight the key differences between our method and the most relevant T-STAR with a gray background.
自适应子任务技能学习和双向对抗学习的伪代码分别展示在算法1和算法2中。我们用灰色背景突出显示了本方法与最相关的T-STAR方法的关键区别。

20.1 A.1 Adaptive Sub-task Skill Learning

20.2 A.1 自适应子任务技能学习

As shown in Algorithm 1, the innovation of the sub-task skill learning scheme we propose, compared to previous methods, consists of two parts: 1) We use a more stable weighted reward function for policy learning of sub-task skills, as shown in Eq. 1 and Eq. 3 in the main paper. 2) We introduce AES regularization constraints into this weighted reward function to periodically adaptively schedule the scale of the two reward terms, as shown in line 11-14 of Algorithm 1, allowing the robot to fully explore and learn from both the environment and the expert behaviors.
如算法1所示，我们提出的子任务技能学习方案相比以往方法的创新点包括两部分：1）我们采用了更稳定的加权奖励函数用于子任务技能的策略学习，如主文中公式1和公式3所示。2）我们在该加权奖励函数中引入了AES正则化约束，以周期性自适应地调节两个奖励项的权重比例，如算法1第11-14行所示，使机器人能够充分探索并从环境和专家行为中学习。

Algorithm 1 Adaptive Sub-task Skill Learning.
算法1 自适应子任务技能学习。

Key differences to T-STAR [15] in gray.
与T-STAR [15]的关键区别以灰色标注。

```
Require: expert demonstrations ${\mathbb{D}}_{1}^{E},\ldots ,{\mathbb{D}}_{K}^{E}$ ,sub-task MDPs
${\mathcal{M}}_{1},\ldots ,{\mathcal{M}}_{K}$
要求：专家示范${\mathbb{D}}_{1}^{E},\ldots ,{\mathbb{D}}_{K}^{E}$，子任务马尔可夫决策过程（MDP）
${\mathcal{M}}_{1},\ldots ,{\mathcal{M}}_{K}$
```

2: Initialize sub-task policies $\pi_\theta^1,\ldots,\pi_\theta^K$,least-squares GAIL discriminator $f_\phi^1,\ldots,f_\phi^K$.
2: 初始化子任务策略$\pi_\theta^1,\ldots,\pi_\theta^K$，最小二乘GAIL判别器$f_\phi^1,\ldots,f_\phi^K$。

3: Initialize imitation progress recognizer Φ with $\mathbb{D}^E$,balance discount factor $\lambda_{\mathrm{RL}} \leftarrow \alpha, \lambda_{\mathrm{IL}} \leftarrow$
3: 使用$\mathbb{D}^E$和折扣因子$\lambda_{\mathrm{RL}} \leftarrow \alpha, \lambda_{\mathrm{IL}} \leftarrow$初始化模仿进度识别器$\Phi$

```
$1 - \alpha$ .
```

```
for each sub-task $i = 1,\ldots ,K$ do
对每个子任务$i = 1,\ldots ,K$执行

    for episode $= 1,2,\ldots ,N$ do
    对每个回合$= 1,2,\ldots ,N$执行

        Rollout trajectories $\tau  = \left( {{s}_{1},{a}_{1},{r}_{1}^{\mathrm{{Env}}},\ldots ,{s}_{T}}\right)$ with
${\pi }_{\theta }^{i}$
        使用${\pi }_{\theta }^{i}$展开轨迹$\tau  = \left( {{s}_{1},{a}_{1},{r}_{1}^{\mathrm{{Env}}},\ldots ,
{s}_{T}}\right)$

        // WEIGHTED REWARD FUNCTION
        // 加权奖励函数

        Compute balanced reward $\left\{  {{r}_{1},\ldots ,{r}_{T - 1}}\right\}   \leftarrow  {\lambda
}_{\mathrm{{RL}}}{r}^{\text{Env }} + {\lambda }_{\mathrm{{IL}}}{r}^{\text{Pred }}$
        计算平衡奖励 $\left\{  {{r}_{1},\ldots ,{r}_{T - 1}}\right\}   \leftarrow  {\lambda }_{\mathrm{{RL}}}
{r}^{\text{Env }} + {\lambda }_{\mathrm{{IL}}}{r}^{\text{Pred }}$

        Update ${f}_{\phi }^{i}$ with $\tau$ and ${\tau }^{E} \sim  {\mathbb{D}}_{i}^{E}$ using Eq. 3
        使用公式3用$\tau$和${\tau }^{E} \sim  {\mathbb{D}}_{i}^{E}$更新${f}_{\phi }^{i}$

        Update ${\pi }_{\theta }^{i}$ with the rewarded trajectories $\left\{  {{s}_{1},{a}_{1},{r}_{1},\ldots ,
{s}_{T}}\right\}$
        用奖励后的轨迹$\left\{  {{s}_{1},{a}_{1},{r}_{1},\ldots ,{s}_{T}}\right\}$更新${\pi }_{\theta }^{i}$

        // ADAPTIVE EQUILIBRIUM SCHEDULING REGULARIZATION
```

```
        // 自适应均衡调度正则化

        Update imitation progress recognizer $\Phi$ with $\tau$ and ${\tau }^{E} \sim  {\mathbb{D}}_{i}^{E}$
        用$\tau$和${\tau }^{E} \sim  {\mathbb{D}}_{i}^{E}$更新模仿进度识别器$\Phi$

        Query $\Phi$ about the current imitation progress $p$
        查询$\Phi$当前模仿进度$p$

        Update balance discount factor ${\lambda }_{\mathrm{{RL}}},{\lambda }_{\mathrm{{IL}}} \leftarrow  {\varphi }_{\lambda }\left( p\right)$
        更新平衡折扣因子${\lambda }_{\mathrm{{RL}}},{\lambda }_{\mathrm{{IL}}} \leftarrow  {\varphi }_{\lambda }\left( p\right)$

    end for
    结束循环

end for
结束循环
```

20.3 A.2 Bi-directional Adversarial Learning

20.4 A.2 双向对抗学习

As shown in Algorithm 2, the innovation of the bi-directional adversarial learning mechanism consists of two parts: 1) We propose a bi-directional regularization which is trained by two balanced bidirectional constraints to better chain sequential skills, as shown in line 16-17 of Algorithm 2.2) We also employ the adaptive sub-skill learning scheme during the bi-directional adversarial learning process in order to ensure inter-skill alignment while enabling the sub-task skills to be adaptively adjusted to task changes during fine-tuning as well, as shown in line 10-12 of Algorithm 2
如算法2所示，双向对抗学习机制的创新包括两部分：1）我们提出了一种双向正则化，通过两个平衡的双向约束进行训练，以更好地衔接序列技能，如算法2第16-17行所示。2）我们还在双向对抗学习过程中采用了自适应子技能学习方案，以确保技能间对齐，同时使子任务技能在微调过程中能够自适应地调整以应对任务变化，如算法2第10-12行所示。

21 B More Details on AES Regularization

22 B AES正则化的更多细节

Automatic Discount Scheduling (ADS) [9] is a mechanism for allocating more appropriate reward signals in s Imitation Learning from Observation (ILfO), based on the concept of Optimal Transport [51] and further introducing the characteristic of process dependency across tasks. Based on this, ADS focuses on adjusting the discount factor during reinforcement learning training in ILfO. Following the mechanism in ADS,our AES also employs an imitation progress recognizer Φ to monitor the extent to which the agent has assimilated the expert's behaviors. The main idea is to assess the closeness of the pair of trajectories by evaluating the agent-collected trajectory $\tau = (s_0, \ldots, s_T)$ and the expert trajectory $\tau^E = \left(s_0^E, \ldots, s_T^E\right)$ through a monotonic state-by-state alignment.
自动折扣调度（Automatic Discount Scheduling，ADS）[9]是一种在观察模仿学习（Imitation Learning from Observation，ILfO）中分配更合适奖励信号的机制，其基于最优传输（Optimal Transport）[51]的概念，并进一步引入了任务间过程依赖的特性。在此基础上，ADS关注于在ILfO的强化学习训练过程中调整折扣因子。遵循ADS的机制，我们的AES同样采用了模仿进度识别器Φ，以监控智能体对专家行为的吸收程度。其主要思想是通过对智能体采集的轨迹$\tau = (s_0, \ldots, s_T)$与专家轨迹$\tau^E = \left(s_0^E, \ldots, s_T^E\right)$进行单调的状态对齐，来评估这对轨迹的接近程度。

Algorithm 2 Bi-directional Adversarial Learning Key differences to T-STAR [15] in gray.
算法2 双向对抗学习 与T-STAR [15]的主要区别以灰色标注。

```
Require: expert demonstrations ${\mathbb{D}}_{1}^{E},\ldots ,{\mathbb{D}}_{K}^{E}$ ,sub-task MDPs ${\mathcal{M}}_{1},\ldots ,{\mathcal{M}}_{K}$ ,pre-trained sub-
需求：专家演示${\mathbb{D}}_{1}^{E},\ldots ,{\mathbb{D}}_{K}^{E}$，子任务MDP${\mathcal{M}}_{1},\ldots ,{\mathcal{M}}_{K}$，预训练子-

    task policies ${\pi }_{\theta }^{1},\ldots ,{\pi }_{\theta }^{K}$ ,pre-trained GAIL discriminator ${f}_{\phi }^{1},\ldots ,{f}_{\phi }^{K}$ .
    任务策略${\pi }_{\theta }^{1},\ldots ,{\pi }_{\theta }^{K}$，预训练的GAIL判别器（Generative Adversarial Imitation Learning）${f}_{\phi }^{1},\ldots ,{f}_{\phi }^{K}$。
```

2: Initialize bi-directional discriminator $\zeta_\omega^1,\ldots,\zeta_\omega^K$,imitation identifier Φ with $\mathbb{D}^E$,balance dis-

2：初始化双向判别器$\zeta_\omega^1,\ldots,\zeta_\omega^K$，模仿识别器$\Phi$，以及平衡折扣因子$\mathbb{D}^E$，

```
    count factor ${\lambda }_{\mathrm{{RL}}} \leftarrow  \alpha ,{\lambda }_{\mathrm{{IL}}} \leftarrow  1 - \alpha$ .
    计数因子${\lambda }_{\mathrm{{RL}}} \leftarrow  \alpha ,{\lambda }_{\mathrm{{IL}}} \leftarrow  1 - \alpha$。

           tialize initial state buffers ${\mathcal{B}}_{I}^{1},\ldots ,{\mathcal{B}}_{I}^{K}$ ,and terminal state buffers ${\mathcal{B}}_{\beta }^{1},\ldots ,{\mathcal{B}}_{\beta }^{K}$ .
           初始化初始状态缓冲区${\mathcal{B}}_{I}^{1},\ldots ,{\mathcal{B}}_{I}^{K}$和终止状态缓冲区${\mathcal{B}}_{\beta }^{1},\ldots ,{\mathcal{B}}_{\beta }^{K}$。

    for iteration $m = 0,1,\ldots ,M$ do
    对于迭代$m = 0,1,\ldots ,M$执行

        for each sub-task $i = 1,\ldots ,K$ do
        对于每个子任务$i = 1,\ldots ,K$执行

                Sample ${s}_{0}$ from environment or ${\mathcal{B}}_{\beta }^{i - 1}$
                从环境或${\mathcal{B}}_{\beta }^{i - 1}$中采样${s}_{0}$

                Rollout trajectories $\tau  = \left( {{s}_{1},{a}_{1},{r}_{1},\ldots ,{s}_{T}}\right)$ with pre-trained ${\pi }_{\theta }^{i}$
                使用预训练的${\pi }_{\theta }^{i}$展开轨迹$\tau  = \left( {{s}_{1},{a}_{1},{r}_{1},\ldots ,{s}_{T}}\right)$

                if $\tau$ is successful then
                如果$\tau$成功则

                    ${\mathcal{B}}_{I}^{i} \leftarrow  {\mathcal{B}}_{I}^{i} \cup  {s}_{1},{\mathcal{B}}_{\beta }^{i} \leftarrow  {\mathcal{B}}_{\beta }^{i} \cup  {s}_{T}$

                    // ADAPTIVE EQUILIBRIUM SCHEDULING
                    // 自适应平衡调度

                    Update imitation identifier $\Phi$ with $\tau$
                    使用$\tau$更新模仿识别器$\Phi$

                    Query $\Phi$ about the current imitation progress $p$
                    查询$\Phi$当前模仿进度$p$

                end if
                结束条件

                Update balance discount factor ${\lambda }_{\mathrm{{RL}}},{\lambda }_{\mathrm{{IL}}} \leftarrow {\varphi }_{\lambda }\left( p\right)$
                更新平衡折扣因子${\lambda }_{\mathrm{{RL}}},{\lambda }_{\mathrm{{IL}}} \leftarrow  {\varphi }_{\lambda }\left( p\right)$

                Fine-tune ${f}_{\phi }^{i}$ with $\tau$ and ${\tau }^{E} \sim  {\mathbb{D}}_{i}^{E}$
                使用$\tau$和${\tau }^{E} \sim  {\mathbb{D}}_{i}^{E}$微调${f}_{\phi }^{i}$

                // TRAIN BI-DIRECTIONAL DISCRIMINATOR
                // 训练双向判别器

                Update ${\zeta }_{\omega }^{i}$ with ${s}_{\beta } \sim  {\mathcal{B}}_{\beta }^{i - 1}$ and ${s}_{I} \sim  {\mathcal{B}}_{I}^{i}$ with ${\mathcal{L}}_{i}\left( \omega \right)  = \frac{1}{2}{\mathcal{C}}_{1} + \frac{1}{2}{\mathcal{C}}_{2}$
                用${s}_{\beta } \sim  {\mathcal{B}}_{\beta }^{i - 1}$更新${\zeta }_{\omega }^{i}$，用${\mathcal{L}}_{i}\left( \omega \right)  = \frac{1}{2}{\mathcal{C}}_{1} + \frac{1}{2}{\mathcal{C}}_{2}$更新${s}_{I} \sim  {\mathcal{B}}_{I}^{i}$

                // FINE-TUNE WITH DUAL REGULARIZATION
                // 使用双重正则化进行微调

                Update ${\pi }_{\theta }^{i}$ with ${r}_{i}\left( {{s}_{t},{a}_{t},{s}_{t + 1};\phi ,\omega }\right)$ using Eq. 7
                使用公式7用${r}_{i}\left( {{s}_{t},{a}_{t},{s}_{t + 1};\phi ,\omega }\right)$更新${\pi }_{\theta }^{i}$

        end for
        结束循环
```

```
end for
结束循环
```

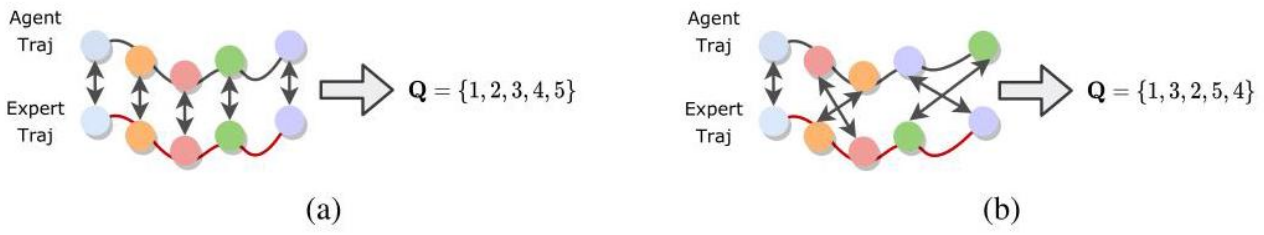


Figure 7: Visualization of the construction of the sequence $\mathbf{Q}$. To be more intuitive,we directly represent the minimum cosine similarity with double arrows.
图7：序列$\mathbf{Q}$构建的可视化。为了更直观，我们用双箭头直接表示最小余弦相似度。

To be specific, Φ receives the agent's collected trajectories τ (line 12 in Algorithm 1) and infers the agent's current imitation progress $p, p \in [0, T)$ (line 13 in Algorithm 1). The construction of Φ ,with reference to ADS,first requires the construction of a sequence $\overline{\mathbf{Q}}\left(q_1, \ldots, q_T\right)$,where $q_i = \operatorname{argmin}_j c\left(s_i, s_j^E\right)$ is the index of the nearest neighbor of s_i in τ^E, c is the cosine similarity. As shown in Fig. 7 If τ and τ^E are exactly the same,then $\mathbf{Q}$ becomes a strictly increasing sequence (Fig 7(a)). On the contrary,if τ and τ^E characterize some different behaviors,there are some unordered sequences in $\mathbf{Q}$ (Fig 7(b)).
具体来说，Φ接收智能体收集的轨迹τ（算法1第12行）并推断智能体当前的模仿进度$p, p \in [0, T)$（算法1第13行）。参考ADS，构建Φ首先需要构建序列$\overline{\mathbf{Q}}\left(q_1, \ldots, q_T\right)$，其中$q_i = \operatorname{argmin}_j c\left(s_i, s_j^E\right)$是$s_i$在$\tau^E, c$中的最近邻索引，$\tau^E, c$是余弦相似度。如图7所示，如果$\tau$和$\tau^E$完全相同，则$\mathbf{Q}$成为严格递增序列（图7(a)）。相反，如果$\tau$和$\tau^E$表现出不同的行为，则$\mathbf{Q}$中存在无序序列（图7(b)）。

After constructing $\mathbf{Q}$,the progress alignment between τ and τ^E is measured as the length of the longest increasing subsequence (LIS) in $\mathbf{Q}$,denoted as $\operatorname{LIS}\left(\tau, \tau^E\right)$. For instance,if $\mathbf{Q} = \{1, 3, 2, 5, 4\}$ as in Fig 7(b) then its LIS can be $\{1, 3, 5\}, \{1, 2, 5\}, \{1, 3, 4\}$ or $\{1, 2, 4\}$. The LIS measurement concentrates on the consistency of the macroscopic trends in these trajectories, thereby preventing overfitting to the microscopic features in the observation [9].
构建$\mathbf{Q}$后，τ与τ^E之间的进度对齐通过$\mathbf{Q}$中最长递增子序列（LIS）的长度来衡量，记为$\operatorname{LIS}\left(\tau, \tau^E\right)$。例如，如果如图7(b)所示$\mathbf{Q} = \{1, 3, 2, 5, 4\}$，则其LIS可以是$\{1, 3, 5\}, \{1, 2, 5\}, \{1, 3, 4\}$或$\{1, 2, 4\}$。LIS度量关注这些轨迹宏观趋势的一致性，从而防止对观测中的微观特征过拟合[9]。

Further, if the following inequality Eq. 8 holds, this indicates that at this time step, the agent's imitation of the expert's action is equivalent to the level of the expert's performance, then the agent's imitation progress p will increase by 1 :
此外，如果满足以下不等式公式8，表明在此时间步，智能体对专家动作的模仿达到专家表现水平，则智能体的模仿进度p将增加1：

$$\max_{\hat{\tau}^E \in \mathbb{D}^E} \operatorname{LIS}\left(\tau_{1:p+1}, \hat{\tau}_{1:p+1}^E\right) \geq \rho \times \min_{\hat{\tau}^E, \hat{\tau}^E \in \mathbb{D}^E} \operatorname{LIS}\left(\hat{\tau}_{1:p+1}^E, \hat{\tau}_{1:p+1}^E\right), \tag{8}$$

where $\hat{\tau}^E \neq \hat{\tau}^E$,the subscript $1 : p+1$ denotes the first $p+1$ steps of the extracted trajectory,and $\rho \in [0, 1]$ controls the stringency of the imitation progress monitoring.
其中$\hat{\tau}^E \neq \hat{\tau}^E$，下标$1 : p+1$表示提取轨迹的前$p+1$步，$\rho \in [0, 1]$控制模仿进度监测的严格程度。

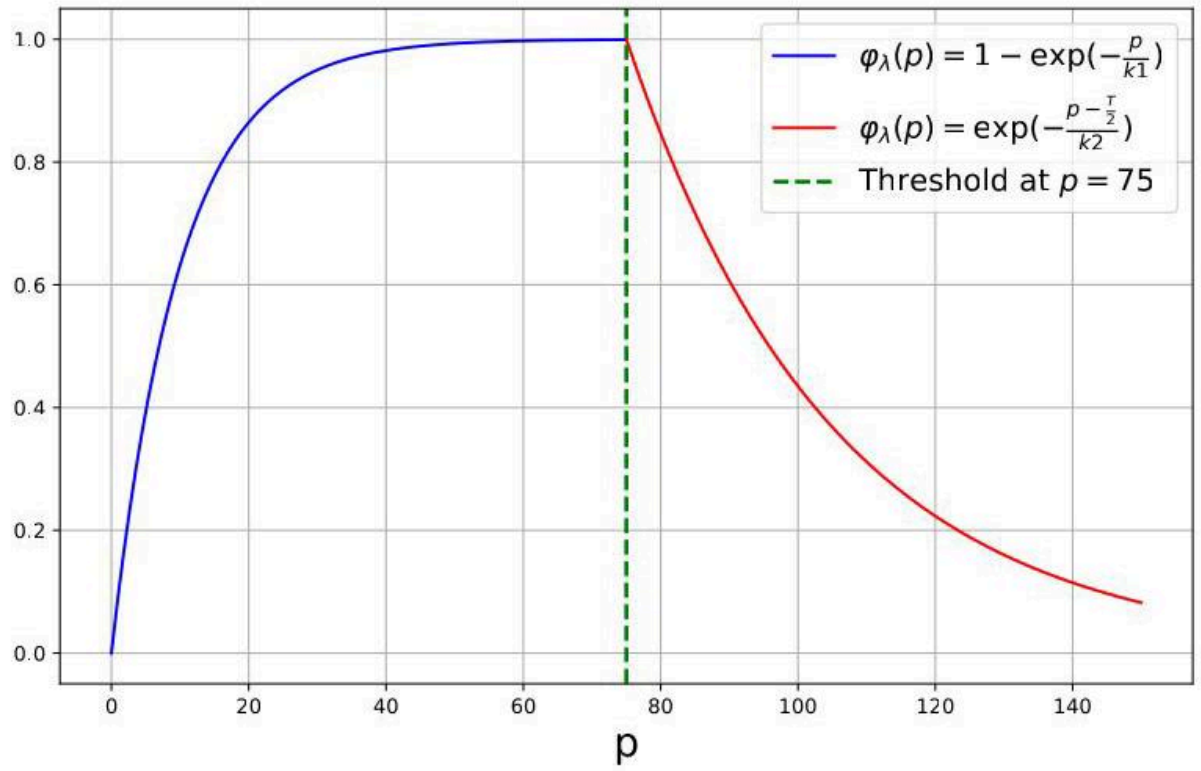


Figure 8: Visualization of the mapping function $\varphi_\lambda(p)$. In this example,we assume that $T = 150$.
图8：映射函数$\varphi_\lambda(p)$的可视化。在此示例中，我们假设$T = 150$。

After obtaining the current imitation progress p of the agent,AES then adopts a mapping function $\varphi_\lambda(p)$ to schedule the two new balance discount factors λ_{RL} and λ_{IL} . Straightforward idea of setting $\varphi_\lambda(p)$ is that If p reaches a certain threshold,i.e., the agent is able to imitate the expert's behavior well, then the more the agent tends to imitate the expert's behavior in subsequent training, and vice versa. Therefore,we set the threshold as $\frac{T}{2}$. If $p \in \left[0, \frac{T}{2}\right)$,we propose $\varphi_\lambda(p) = 1 - e^{\left(-\frac{p}{k1}\right)}$; if $p \in \left[\frac{T}{2}, T\right)$,we propose $\varphi_\lambda(p) = e^{\left(-\frac{p-\frac{T}{2}}{k2}\right)}$,where $k1$ and $k2$ are used to flatten the curve of the mapping function. The mapping function shown in Fig. 8,where $T = 150$. In our experiments,we use different flatten factors for the two stages,where $k1 = 10$ and $k2 = 30$.
获取代理当前的模仿进度p后，AES采用映射函数$\varphi_\lambda(p)$来调度两个新的平衡折扣因子λ_{RL}和λ_{IL}。设置$\varphi_\lambda(p)$的直接思路是：如果p达到某个阈值，即代理能够很好地模仿专家的行为，那么代理在后续训练中越倾向于模仿专家行为，反之亦然。因此，我们将阈值设为$\frac{T}{2}$。如果$p \in \left[0, \frac{T}{2}\right)$，我们提出$\varphi_\lambda(p) = 1 - e^{\left(-\frac{p}{k1}\right)}$；如果$p \in \left[\frac{T}{2}, T\right)$，我们提出$\varphi_\lambda(p) = e^{\left(-\frac{p-\frac{T}{2}}{k2}\right)}$，其中$k1$和$k2$用于平滑映射函数的曲线。映射函数如图8所示，其中$T = 150$。在我们的实验中，我们在两个阶段使用不同的平滑因子，分别为$k1 = 10$和$k2 = 30$。

Then λ_{RL} and λ_{IL} are scheduled to be :
然后λ_{RL}和λ_{IL}被调度为：

$$\begin{cases} \lambda_{\mathrm{RL}} = \alpha^{1-e^{\left(-\frac{p}{k1}\right)}}, \lambda_{\mathrm{IL}} = 1 - \alpha^{1-e^{\left(-\frac{p}{k1}\right)}}. & \text{if } p \in \left[0, \frac{T}{2}\right) \\ \lambda_{\mathrm{IL}} = \alpha^{e^{\left(-\frac{p-\frac{T}{2}}{k2}\right)}}, \lambda_{\mathrm{RL}} = 1 - \alpha^{e^{\left(-\frac{p-\frac{T}{2}}{k2}\right)}}. & \text{if } p \in \left[\frac{T}{2}, T\right) \end{cases} \tag{9}$$

As shown by the trend of function $\alpha^{\varphi_\lambda(p)}$ in Fig 9,when $p \in \left[0, \frac{T}{2}\right)$,the scale of $\lambda_{\mathrm{RL}} : \alpha^{1-e^{\left(-\frac{p}{k1}\right)}}$ is scheduled to be larger than $\lambda_{\mathrm{IL}} : 1 - \alpha^{1-e^{\left(-\frac{p}{k1}\right)}}$,but this gap gets smaller and smaller as p gets larger. When $p \in \left[\frac{T}{2}, T\right)$, the scale of $\lambda_{\mathrm{IL}} : \alpha^{e^{\left(-\frac{p-\frac{T}{2}}{k2}\right)}}$ is scheduled to be larger than $\lambda_{\mathrm{RL}} : 1 - \alpha^{e^{\left(-\frac{p-\frac{T}{2}}{k2}\right)}}$,while the scale of λ_{IL} increases as the agent imitates better.
如图9所示的函数$\alpha^{\varphi_\lambda(p)}$趋势，当$p \in \left[0, \frac{T}{2}\right)$时，$\lambda_{\mathrm{RL}}$:$\alpha^{1-e^{\left(-\frac{p}{k1}\right)}}$的比例被调度得大于$\lambda_{\mathrm{IL}}$:$1 - \alpha^{1-e^{\left(-\frac{p}{k1}\right)}}$，但随着$p$增大，这一差距逐渐缩小。当$p \in \left[\frac{T}{2}, T\right)$时，$\lambda_{\mathrm{IL}} : \alpha^{e^{\left(-\frac{p-\frac{T}{2}}{k2}\right)}}$的比例被调度得大于$\lambda_{\mathrm{RL}} : 1 - \alpha^{e^{\left(-\frac{p-\frac{T}{2}}{k2}\right)}}$，而随着代理模仿效果的提升，$\lambda_{\mathrm{IL}}$的比例也在增加。

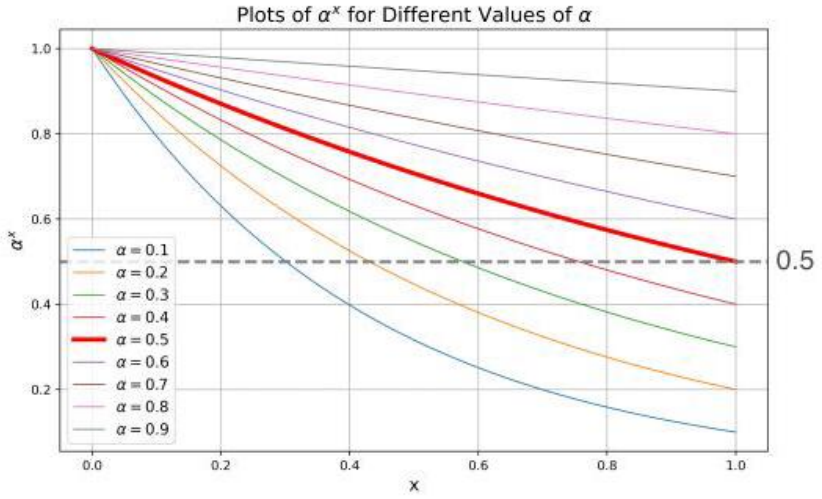


Figure 9: $\alpha^{\varphi_\lambda(p)}$ based on the variation of different α sizes in $\varphi_\lambda(p) \in [0,1]$. We use $\alpha = 0.5$ as the base in our experiments.
图9：基于不同α规模变化的$\alpha^{\varphi_\lambda(p)}$。在我们的实验中，我们使用$\alpha = 0.5$作为基准。

Thus,if p is larger and reaches a threshold step, i.e., the agent is able to imitate the expert's behavior well, then the more the agent tends to imitate the expert's behavior in subsequent training, and vice versa. The entire process is adaptively scheduled based on Φ periodic monitoring of the agent's imitation process. Consequently, the RL and IL components of sub-task skill learning can be adaptively scheduled and regularized through AES, effectively enhancing intra-skill dependencies between sequential actions.
因此，如果p较大并达到阈值步骤，即代理能够很好地模仿专家行为，那么代理在后续训练中越倾向于模仿专家行为，反之亦然。整个过程基于Φ对代理模仿过程的周期性监测自适应调度。因此，子任务技能学习中的强化学习（RL）和逆向学习（IL）组件可以通过AES自适应调度和正则化，从而有效增强连续动作之间的技能内依赖。

23 C Sub-task Skills

24 C 子任务技能

In our simulation experiments, we use sequences of sub-tasks defined internally by the environment [44, 45] as task decomposition sub-tasks. Here we list these sequential skills to emphasize the difficulty of long-horizon tasks. Each skill takes a 3D position as the input g_* .
在我们的仿真实验中，我们使用环境内部定义的子任务序列[44, 45]作为任务分解的子任务。在此我们列出这些连续技能，以强调长航时任务的难度。每个技能以一个三维位置g_*作为输入。

25 IKEA Furniture Assembly:

26 宜家家具组装：

Chair_agne (2 sub-task skills): Assemble stool leg 0 to target position $g_*^0 \rightarrow$ Assemble stool leg 1 to target position g_*^1
Chair_agne（2个子任务技能）：将凳子腿0组装到目标位置$g_*^0 \rightarrow$ 将凳子腿1组装到目标位置g_*^1

Chair_bernhard (2 sub-task skills): Assemble support leg 0 to target position $g_*^0 \rightarrow$ Assemble support leg 1 to target position g_*^1
Chair_bernhard（2个子任务技能）：将支撑腿0组装到目标位置$g_*^0 \rightarrow$ 将支撑腿1组装到目标位置g_*^1

Table_dockstra (3 sub-task skills): Assemble table leg 0 to target position $g_*^0 \rightarrow$ Assemble table leg 1 to target position $g_*^1 \rightarrow$ Assemble table top to target position g_*^3
Table_dockstra（3个子任务技能）：将桌腿0组装到目标位置$g_*^0 \rightarrow$ 将桌腿1组装到目标位置$g_*^1 \rightarrow$ 将桌面组装到目标位置g_*^3

Chair_ingolf (4 sub-task skills): Assemble chair support 0 to target position $g_*^0 \rightarrow$ Assemble chair support 1 to target position $g_*^1 \rightarrow$ Assemble front leg 0 to target position $g_*^3 \rightarrow$ Assemble front leg 1 to target position g_*^4
Chair_ingolf（4个子任务技能）：将椅子支撑0组装到目标位置$g_*^0 \rightarrow$ 将椅子支撑1组装到目标位置$g_*^1 \rightarrow$ 将前腿0组装到目标位置$g_*^3 \rightarrow$ 将前腿1组装到目标位置g_*^4

Table_lack (4 sub-task skills): Assemble table leg 0 to target position $g_*^0 \rightarrow$ Assemble table leg 1 to target position $g_*^1 \rightarrow$ Assemble table leg 2 to target position $g_*^3 \rightarrow$ Assemble table leg 3 to target position g_*^4
Table_lack（4个子任务技能）：将桌腿0组装到目标位置$g_*^0 \rightarrow$ 将桌腿1组装到目标位置$g_*^1 \rightarrow$ 将桌腿2组装到目标位置$g_*^3 \rightarrow$ 将桌腿3组装到目标位置g_*^4

Toy_table (4 sub-task skills): Assemble table leg 0 insert to target position g_*^0 → Assemble table leg 1 insert to target position g_*^1 → Assemble table leg 2 insert to target position g_*^3 → Assemble table leg 3 insert to target position g_*^4
Toy_table（4个子任务技能）：将桌腿0插入目标位置g_*^0 → 将桌腿1插入目标位置g_*^1 → 将桌腿2插入目标位置g_*^3 → 将桌腿3插入目标位置g_*^4

27 Kitchen Organization:

28 厨房整理：

Kitchen (4sub-task skills): Turn on the microwave to target position g_*^0 → Move the kettle to target position g_*^1 → Turn on the stove (rotate the stove button to target position g_*^2) → Turn on the light (rotate the light button to target position g_*^3)
厨房（4个子任务技能）：将微波炉开到目标位置g_*^0 → 将水壶移动到目标位置g_*^1 → 打开炉灶（旋转炉灶按钮至目标位置g_*^2）→ 打开灯（旋转灯按钮至目标位置g_*^3）

Extended Kitchen (5 sub-task skills): Turn on the microwave to target position g_*^0 → Turn on the stove (rotate the stove button to target position g_*^1) → Turn on the light (rotate the light button to target position g_*^2) → Slide the cabinet to the right target position g_*^3 → Open the cabinet to target position g_*^4
扩展厨房（5个子任务技能）：将微波炉开到目标位置g_*^0 → 打开炉灶（旋转炉灶按钮至目标位置g_*^1）→ 打开灯（旋转灯按钮至目标位置g_*^2）→ 将橱柜滑动到右侧目标位置g_*^3 → 打开橱柜至目标位置g_*^4

29 D More Quantitative Results

30 D 更多定量结果

We present the training curves with different skill learning methods for sub-task skills in chair_ingolf task, and we further present the evaluation performance of the pre-trained skills with different methods across sub-tasks in the other 6 long-horizon simulation tasks. Also, we test the algorithms trained from scratch in the presence of perturbations to further illustrate the importance of the execution of sub-tasks on long-horizon tasks.
我们展示了chair_ingolf任务中不同技能学习方法的子任务技能训练曲线，并进一步展示了预训练技能在其他6个长时序模拟任务中不同方法跨子任务的评估表现。此外，我们还测试了在扰动存在情况下从零开始训练的算法，以进一步说明子任务执行在长时序任务中的重要性。

Additionally,the main paper does not delve into the loss function $\mathcal{L}_i(\omega)$ concerning the different scales of the bi-directional constraints in bi-directional adversarial training. Therefore, we conduct further ablation experiments to examine the impact of different scales of the two constraints in the bi-directional discriminator.
此外，主论文未深入探讨双向对抗训练中双向约束不同尺度的损失函数$\mathcal{L}_i(\omega)$。因此，我们进行了进一步的消融实验，以检验双向判别器中两种约束不同尺度的影响。

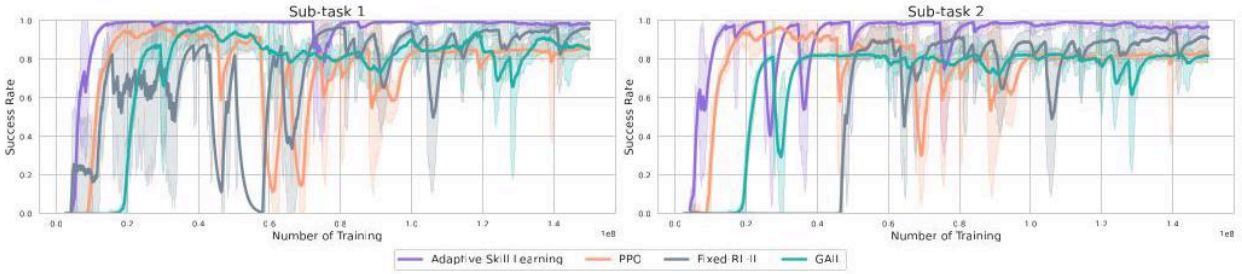
30.1 D.1 Sub-task Skill Learning Performance

30.2 D.1 子任务技能学习表现

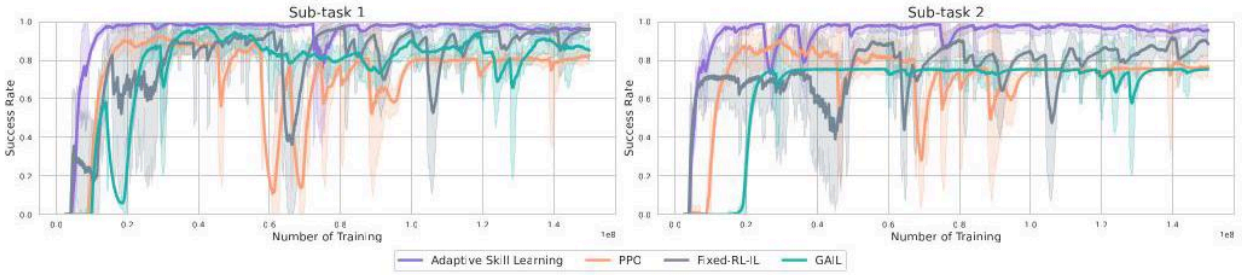
30.2.1 D.1.1 Training performance

30.2.2 D.1.1 训练表现

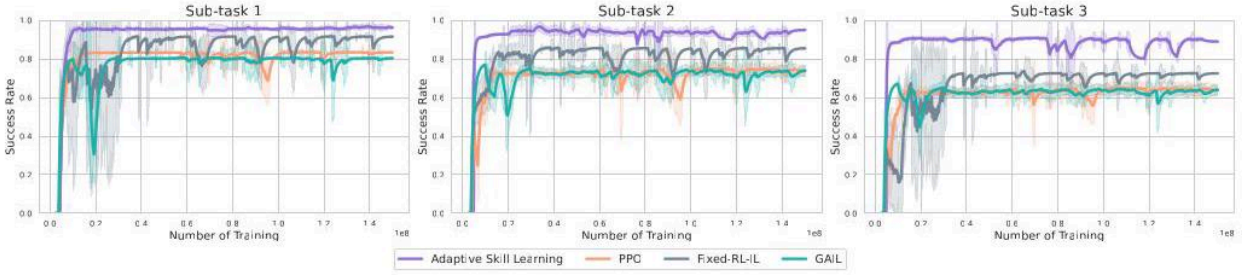
Fig. 10 shows the sub-task skill training curves in IKEA furniture assembly tasks. All methods are trained in each sub-task with 5 random seeds, 15M environment steps. As can be seen, the sub-task skill training based on PPO (learning only from environmental feedback), GAIL (learning only from expert demonstrations) and Fixed-RL-IL (learning from a fixed scale of environmental feedback and expert demonstration) cannot maintain stability and exhibits significant training performance degradation as the sub-task stage increases. In contrast, the sub-task skill training process using our proposed adaptive sub-skill learning scheme has always been relatively stable and better performing.
图10展示了宜家家具组装任务中子任务技能的训练曲线。所有方法均在每个子任务中使用5个随机种子训练，环境步数为1500万步。可以看出，仅基于PPO（仅从环境反馈学习）、GAIL（仅从专家示范学习）和Fixed-RL-IL（从固定比例的环境反馈和专家示范中学习）的子任务技能训练无法保持稳定，且随着子任务阶段增加，训练表现显著下降。相比之下，采用我们提出的自适应子技能学习方案的子任务技能训练过程始终较为稳定且表现更优。



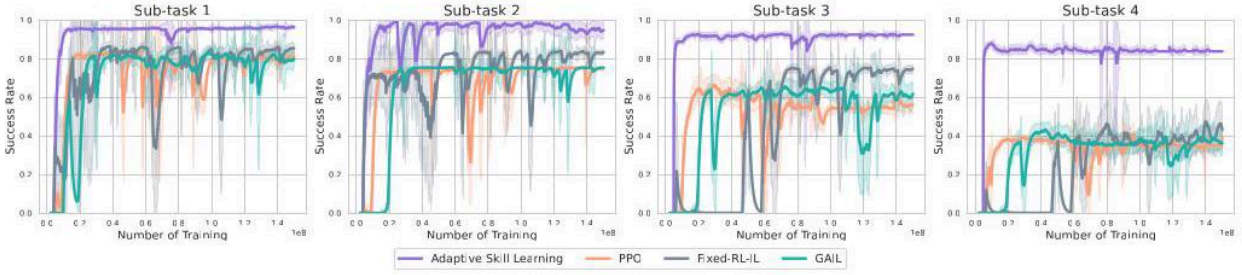
(a) chair_agne



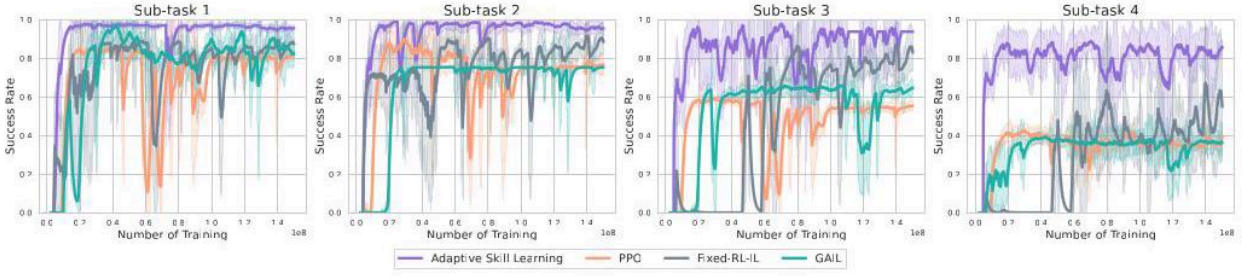
(b) chair_bernhard



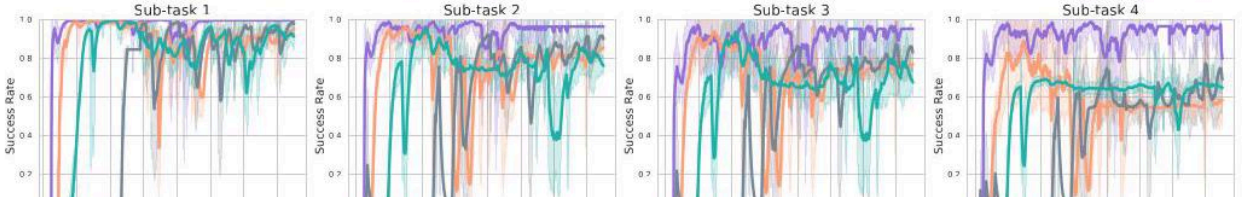
(c) table_dockstra



(d) table_lack

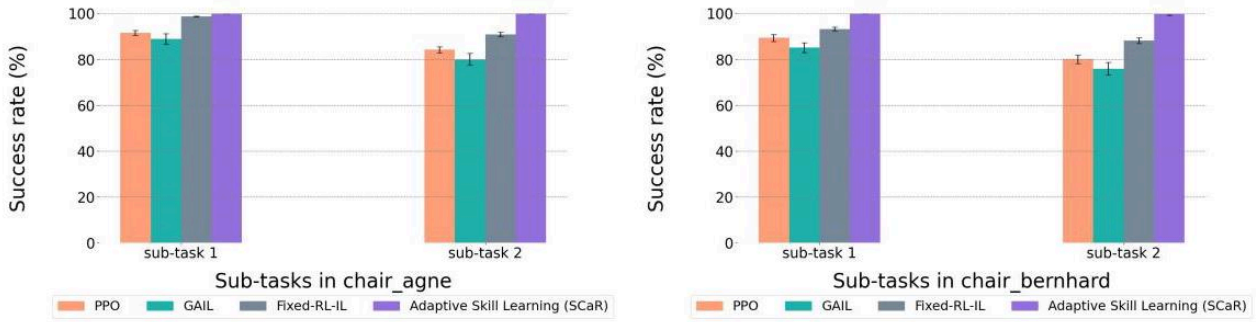


(e) toy_table



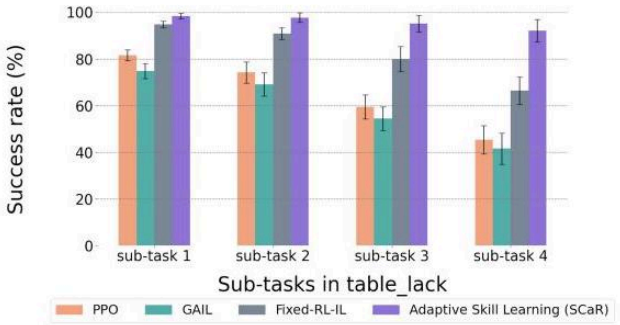
(f) chair_ingolf

Figure 10: Training curves for sub-task skills in IKEA furniture assembly tasks. The y-axis represents the success rate of the sub-task.
图10：宜家家具组装任务中子任务技能的训练曲线。纵轴表示子任务的成功率。

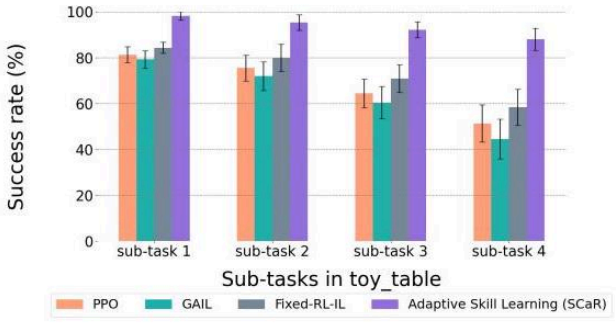


(a) chair_agne

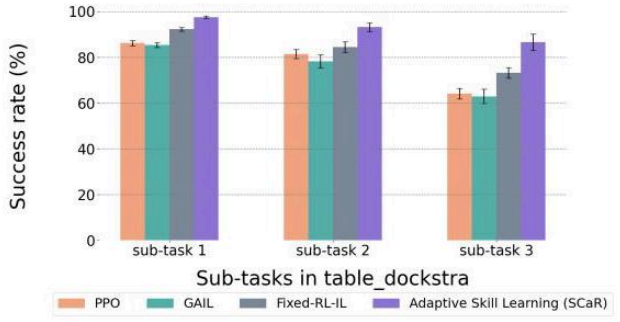
(b) chair_bernhard



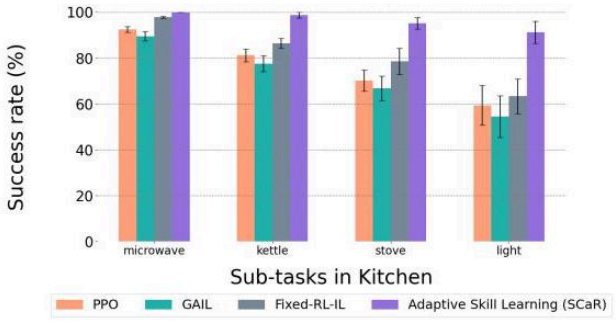
(c) table_lack



(d) toy_table



(e) table_dockstra



(f) Kitchen

Figure 11: Evaluation Performance Comparison of Sub-task Skill Learning.
图11：子任务技能学习的评估性能比较。

30.2.3 D.1.2 More evaluation performance

30.2.4 D.1.2 更多评估性能

As shown in Fig. 11, in chair_agne, chair_bernhard, table_lacktoy_table, table_dockstra, and Kitchen tasks, even with the increase of objects in the environment - and the increase of unpredictable perturbations - our proposed adaptive skill learning learns better sub-task skills. In contrast, the PPO, GAIL, and Fixed-RL-IL baselines fail to maintain well-learning sub-task skills.
如图11所示，在chair_agne、chair_bernhard、table_lacktoy_table、table_dockstra和Kitchen任务中，即使环境中物体数量增加且不可预测的扰动增多，我们提出的自适应技能学习仍能学习到更优的子任务技能。相比之下，PPO、GAIL和Fixed-RL-IL基线方法未能保持良好的子任务技能学习效果。

These results further corroborate that our proposed AES regularization can reinforce inter-step dependencies to the sequential actions within each sub-task skill, and thus pre-train better sub-task skills for long-horizon tasks.
这些结果进一步证明，我们提出的AES正则化能够加强子任务技能中各步骤间的依赖关系，从而为长时域任务预训练出更优的子任务技能。

30.3 D.2 Robustness to Perturbations

30.4 D.2 对扰动的鲁棒性

We test the algorithms trained from scratch in the presence of perturbations. As shown in Table 3 algorithms trained from scratch fail to successfully complete the task when environment perturbations occur during execution. This further illustrates the importance of dividing sub-tasks for multi-stage execution on long-horizon manipulation tasks that are contact-rich and subject to unanticipated perturbations. It also supports the significance of our work on long-horizon robotic manipulation tasks.
我们测试了在扰动存在下从零开始训练的算法。如表3所示，当执行过程中环境发生扰动时，从零开始训练的算法无法成功完成任务。这进一步说明了在接触丰富且易受意外扰动影响的长时域操作任务中，将任务划分为多阶段子任务执行的重要性，也支持了我们在长时域机器人操作任务上的工作意义。

Table 3: Success rates of completing the two sub-tasks chair_bernhard and four sub-tasks chair_ingolf in stationary and perturbed environments.
表3：在静止和扰动环境下完成两个子任务chair_bernhard和四个子任务chair_ingolf的成功率。

Method	chair_bernhard		chair_ingolf	
	No Perturb	$\mathbf{{Perturb}}$	No Perturb	$\mathbf{{Perturb}}$
PPO (Scratch RL)	${0.42} \pm {0.12}$	${0.01} \pm {0.00}$	${0.14} \pm {0.03}$	${0.00} \pm {0.00}$
GAIL (Scratch IL)	${0.23} \pm {0.02}$	${0.00} \pm {0.00}$	${0.00} \pm {0.00}$	${0.00} \pm {0.00}$
Fixed-RL-IL	${0.53} \pm {0.07}$	${0.05} \pm {0.00}$	${0.22} \pm {0.00}$	${0.00} \pm {0.00}$
SkiMo	${0.62} \pm {0.05}$	${0.10} \pm {0.00}$	${0.47} \pm {0.03}$	${0.00} \pm {0.00}$
Policy Sequencing	${0.82} \pm {0.09}$	${0.51} \pm {0.04}$	${0.77} \pm {0.12}$	${0.50} \pm {0.10}$
T-STAR	${0.90} \pm {0.04}$	${0.60} \pm {0.08}$	${0.89} \pm {0.04}$	0.59 ± 0.04
SCaR w/o Bi	${0.92} \pm {0.02}$	${0.65} \pm {0.11}$	${0.91} \pm {0.01}$	${0.63} \pm {0.05}$
SCaR w/o AES	${0.94} \pm {0.03}$	${0.74} \pm {0.09}$	${0.93} \pm {0.02}$	${0.71} \pm {0.07}$
SCaR (Ours)	$\mathbf{{0.96}} \pm {0.04}$	$\mathbf{{0.85}} \pm {0.11}$	$\mathbf{{0.95}} \pm {0.03}$	${0.80} \pm {0.13}$

方法	chair_bernhard		chair_ingolf	
	无扰动	$\mathbf{{Perturb}}$	无扰动	$\mathbf{{Perturb}}$
PPO（从零开始强化学习）	${0.42} \pm {0.12}$	${0.01} \pm {0.00}$	${0.14} \pm {0.03}$	${0.00} \pm {0.00}$
GAIL（从零开始模仿学习）	${0.23} \pm {0.02}$	${0.00} \pm {0.00}$	${0.00} \pm {0.00}$	${0.00} \pm {0.00}$
固定强化学习-模仿学习	${0.53} \pm {0.07}$	${0.05} \pm {0.00}$	${0.22} \pm {0.00}$	${0.00} \pm {0.00}$
SkiMo	${0.62} \pm {0.05}$	${0.10} \pm {0.00}$	${0.47} \pm {0.03}$	${0.00} \pm {0.00}$
策略序列	${0.82} \pm {0.09}$	${0.51} \pm {0.04}$	${0.77} \pm {0.12}$	${0.50} \pm {0.10}$
T-STAR	${0.90} \pm {0.04}$	${0.60} \pm {0.08}$	${0.89} \pm {0.04}$	0.59 ± 0.04
无双向SCaR	${0.92} \pm {0.02}$	${0.65} \pm {0.11}$	${0.91} \pm {0.01}$	${0.63} \pm {0.05}$
无AES SCaR	${0.94} \pm {0.03}$	${0.74} \pm {0.09}$	${0.93} \pm {0.02}$	${0.71} \pm {0.07}$
SCaR（本方法）	$\mathbf{{0.96}} \pm {0.04}$	$\mathbf{{0.85}} \pm {0.11}$	$\mathbf{{0.95}} \pm {0.03}$	${0.80} \pm {0.13}$

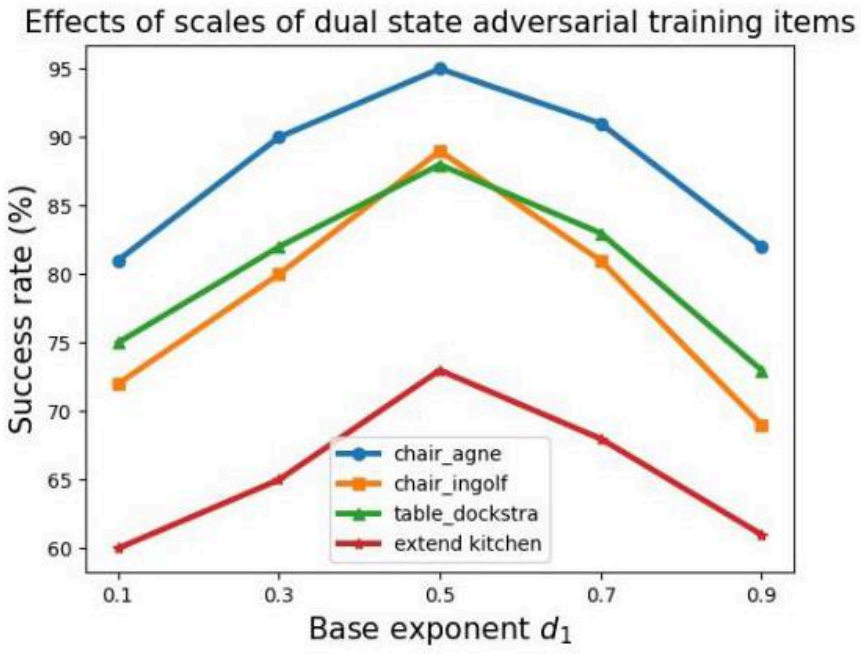


Figure 12: Impact on skill chaining performance of different scales of bi-directional constraints in SCaR.
图12：不同规模的双向约束在SCaR中对技能链性能的影响。

30.5 D.3 Further Ablation

30.6 D.3 进一步消融实验

We set the loss function for the bi-directional discriminator in the main paper as $\mathcal{L}_i(\omega) = \frac{1}{2}\mathcal{C}_1 + \frac{1}{2}\mathcal{C}_2$., where the bi-directional constraints $\mathcal{C}_1, \mathcal{C}_2$ are defined as:
我们在正文中将双向判别器的损失函数设定为$\mathcal{L}_i(\omega) = \frac{1}{2}\mathcal{C}_1 + \frac{1}{2}\mathcal{C}_2$，其中双向约束$\mathcal{C}_1, \mathcal{C}_2$定义为：

$$\text{next initial} \rightarrow \text{previous terminal:}\ \mathcal{C}_1 = \mathbb{E}_{s_{\mathcal{I}\sim\mathcal{I}_i}}\left[\zeta_\omega^i(s_\mathcal{I}) - 1\right]^2 + \mathbb{E}_{s_\mathcal{T}\sim\beta_{i-1}}\left[\zeta_\omega^i(s_\mathcal{T})\right]^2$$
$$\text{下一个初始状态}\rightarrow\text{前一个终止状态：}\ \mathcal{C}_1 = \mathbb{E}_{s_{\mathcal{I}\sim\mathcal{I}_i}}\left[\zeta_\omega^i(s_\mathcal{I}) - 1\right]^2 + \mathbb{E}_{s_\mathcal{T}\sim\beta_{i-1}}\left[\zeta_\omega^i(s_\mathcal{T})\right]^2 \quad (10)$$
$$\text{previous terminal} \rightarrow \text{next initial:}\ \mathcal{C}_2 = \mathbb{E}_{s_\mathcal{T}\sim\beta_i}\left[\zeta_\omega^i(s_\mathcal{T}) - 1\right]^2 + \mathbb{E}_{s_\mathcal{I}\sim\mathcal{I}_{i+1}}\left[\zeta_\omega^i(s_\mathcal{I})\right]^2$$
$$\text{前一个终止状态}\rightarrow\text{下一个初始状态：}\ \mathcal{C}_2 = \mathbb{E}_{s_\mathcal{T}\sim\beta_i}\left[\zeta_\omega^i(s_\mathcal{T}) - 1\right]^2 + \mathbb{E}_{s_\mathcal{I}\sim\mathcal{I}_{i+1}}\left[\zeta_\omega^i(s_\mathcal{I})\right]^2$$

The first constraint $\mathcal{C}_1$ trains the policy to have the initial states approach the terminal states of the previous policy,while the second constraint $\mathcal{C}_2$ trains the policy to have the terminal states close to the initial states of the next policy. In the experiments, these two constraints have the same scale in the training process of the bi-directional discriminator.
第一个约束$\mathcal{C}_1$训练策略使初始状态接近前一个策略的终止状态，而第二个约束$\mathcal{C}_2$训练策略使终止状态接近下一个策略的初始状态。在实验中，这两个约束在双向判别器的训练过程中具有相同的权重。

We wonder whether different scales of these two terms would lead to different performances, and for this reason, we conduct further parametric ablation experiments to explore this. Specifically, we define the scale parameter of the first term $\mathcal{C}_1$ as d_1 ,and the second term $\mathcal{C}_2$ as $d_2 = 1 - d_1$, and set0.1,0.3,0.5,0.7,0.9for d_1 respectively for comparison experiments. We test the effect of different scales of bi-directional adversarial training items d_1 and d_2 on the success rate of SCaR in each of the four tasks: chair_agne, chair_ingolf, table_dockstra, and extend kitchen. As shown in Fig. 12 the experimental result is also in line with our intuition that when the ratio of the two terms initial → previous terminal and terminal → next initial is the same,the performance is the best among the four tasks, whereas when the more imbalanced the scale of the two terms is, the worse the performance is.
我们想知道这两个项的不同权重是否会导致性能差异，因此进行了进一步的参数消融实验进行探索。具体地，我们将第一项$\mathcal{C}_1$的权重参数定义为d_1，第二项$\mathcal{C}_2$的权重定义为$d_2 = 1 - d_1$，并分别设置d_1为0.1、0.3、0.5、0.7、0.9进行对比实验。我们测试了双向对抗训练项d_1和d_2不同权重对SCaR在四个任务（chair_agne、chair_ingolf、table_dockstra和extend kitchen）成功率的影响。如图12所示，实验结果符合我们的直觉：当两项“初始状态→前一个终止状态”和“终止状态→下一个初始状态”的比例相同时，四个任务中的性能最佳；而当两项权重越不平衡时，性能越差。

This ablation result further demonstrate our statement in Sec. 4.3 in the main paper: The purpose of the bi-directional discriminator is to establish a balanced mapping relationship between the initial states and terminal states to ensure the coherence and stability of the policy. If the constraint in one direction (e.g., from initial states to terminal states) is stronger than the constraint in the other direction (e.g., from terminal states to initial states), the information transmission becomes asymmetric. This asymmetry results in better training in one direction and insufficient training in the other, thereby affecting overall performance.
该消融结果进一步验证了正文第4.3节中的观点：双向判别器的目的是建立初始状态与终止状态之间的平衡映射关系，以保证策略的连贯性和稳定

性。如果一个方向的约束（例如从初始状态到终止状态）强于另一个方向的约束（例如从终止状态到初始状态），信息传递将变得不对称。这种不对称导致一个方向训练较好，而另一个方向训练不足，从而影响整体性能。

30.7 D.4 Impact of Different Sub-task Divisions

30.8 D.4 不同子任务划分的影响

To explore the impact of different sub-task divisions, we conduct more experimental validation using the chair_ingolf task. The original sub-tasks in this task are divided as follows: "Assemble chair support 0 to target position" → "Assemble chair support 1 to target position" → "Assemble front leg 0 to target position" → "Assemble front leg 1 to target position". We have re-divided the sub-tasks into two alternative settings:
为了探讨不同子任务划分的影响，我们使用chair_ingolf任务进行了更多实验验证。该任务中的原始子任务划分如下：“将椅子支撑0组装到目标位置”→“将椅子支撑1组装到目标位置”→“将前腿0组装到目标位置”→“将前腿1组装到目标位置”。我们重新将子任务划分为两种备选方案：

1. "Assemble chair support 0 and chair support 1 to target positions" → "Assemble front leg 0 to target position" → "Assemble front leg 1 to target position".
   1.“将椅子支撑0和椅子支撑1组装到目标位置”→“将前腿0组装到目标位置”→“将前腿1组装到目标位置”。

2. "Assemble chair support 0 to target position" → "Assemble chair support 1 to target position" → "Assemble front leg 0 and leg 1 to target positions".
   2.“将椅子支撑0组装到目标位置”→“将椅子支撑1组装到目标位置”→“将前腿0和前腿1组装到目标位置”。

Table 4: The impact of different sub-task divisions on SCaR performance.
表4：不同子任务划分对SCaR性能的影响。

	chair_ingolf (setup 1)	chair_ingolf (setup 2)
SCaR	0.68	0.74
	chair_ingolf（设置1）	chair_ingolf（设置2）
SCaR	0.68	0.74

It is worth noting that, since the re-division of the sub-tasks results in only three sub-tasks, we set 90% as the success metric for all three sub-tasks being successfully executed. As seen in Table 4 compared to SCaR's success rate of about 95% with the original four sub-task divisions,the success rate for completing the first sub-task and then executing the remaining two sub-tasks is significantly reduced. This decrease is due to the increased difficulty of the first sub-task in setup 1 (which requires assembling both chair support) and the last sub-task in setup 2. These changes result in a lower overall success rate for the task. This result suggests that a reasonable division of sub-tasks in long-horizon tasks is crucial for the success rate of overall task completion.
值得注意的是，由于子任务的重新划分仅产生了三个子任务，我们将90%设定为所有三个子任务成功执行的成功指标。如表4所示，与SCaR在原有四个子任务划分下约95%的成功率相比，先完成第一个子任务再执行剩余两个子任务的成功率显著降低。这一下降是由于设置1中第一个子任务（需要组装两个椅子支架）和设置2中最后一个子任务的难度增加所致。这些变化导致整体任务的成功率降低。该结果表明，在长时域任务中，合理划分子任务对于整体任务完成的成功率至关重要。

30.9 D.5 Impact of the Number of Sub-tasks

30.10 D.5 子任务数量的影响

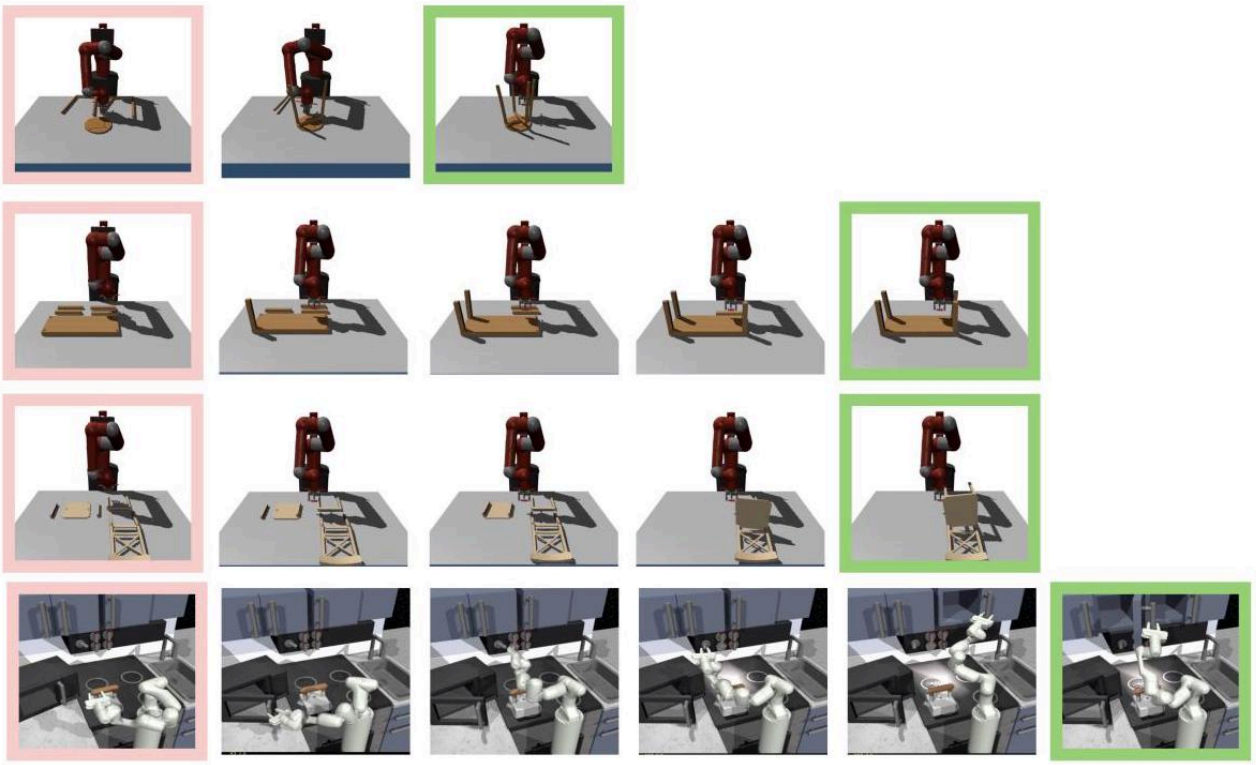
To explore the performance of skill-chaining methods as the number of sub-tasks increases, we add a sub-task to the Extended Kitchen task to evaluate SCaR's performance in manipulation tasks with longer horizons, involving 6 sub-tasks. The modified task, Longer Extended Kitchen includes: 1. Turn on the microwave; 2. Turn on the stove; 3. Turn on the light; 4. Slide the cabinet to the right target position; 5. Open the cabinet to the target position; 6. Move the kettle to the target position.
为了探究技能链方法在子任务数量增加时的表现，我们在扩展厨房任务中增加了一个子任务，以评估SCaR在包含6个子任务的长时域操作任务中的性能。修改后的任务“更长的扩展厨房”包括：1. 打开微波炉；2. 打开炉灶；3. 打开灯；4. 将橱柜滑动到右侧目标位置；5. 将橱柜打开到目标位置；6. 将水壶移动到目标位置。

Table 5: Performance Comparison on Longer Extended Kitchen Task.
表5：更长扩展厨房任务的性能比较。

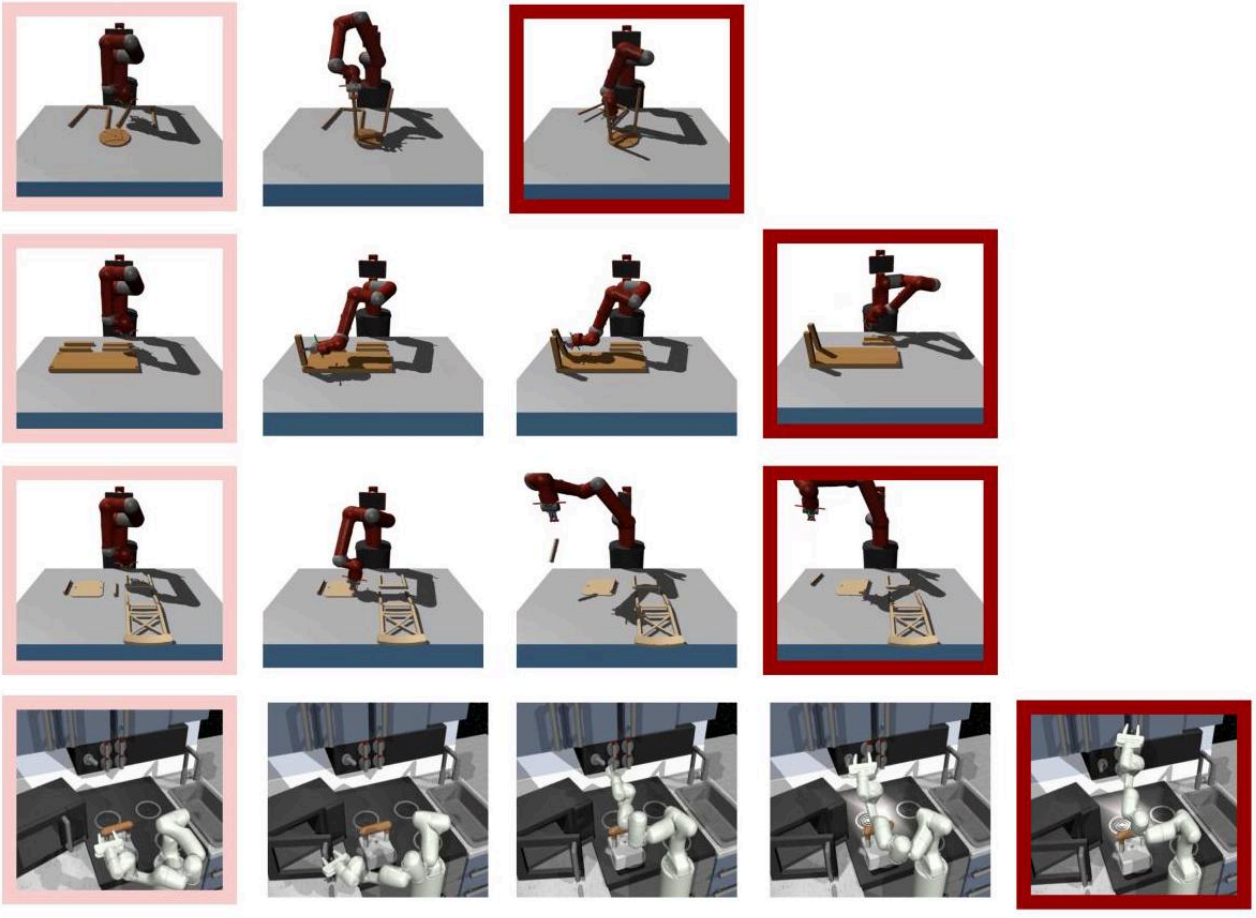
$\mathbf{{Method}}$	Longer Extended Kitchen Task
T-STAR	0.33
SCaR	0.61
$\mathbf{{Method}}$	更长时间的厨房任务
T-STAR	0.33
SCaR	0.61

As shown in Table 5, the addition of an extra sub-task increases the complexity and difficulty of skill chaining in long-horizon tasks. Nonetheless, SCaR achieves a significantly higher overall task execution success rate, surpassing T-STAR by 28%. Although there is still ample room for improvement, we believe our approach establishes a strong baseline for future research on skill-chaining methods for long-horizon manipulation tasks.

如表5所示，增加一个额外的子任务提升了长时序任务中技能链的复杂性和难度。尽管如此，SCaR实现了显著更高的整体任务执行成功率，超过T-STAR 28%。虽然仍有较大提升空间，但我们认为该方法为未来长时序操作任务技能链方法的研究奠定了坚实基线。



(a) SCaR - Successful



(b) T-STAR - Failed

Figure 13: Qualitative results of successful skill chaining performance with SCaR and failed skill chaining performance with T-STAR. More qualitative results can be found on our project website https://sites.google.com/view/scar8297
图13：SCaR成功技能链表现的定性结果与T-STAR失败技能链表现的对比。更多定性结果可见于我们的项目网站 https://sites.google.com/view/scar8297

E More Qualitative Results

E 更多定性结果

Fig 13 shows the qualitative comparison of skill chaining methods. Their animated versions can be found on our project website.
图13展示了技能链方法的定性比较。其动画版本可在我们的项目网站查看。

F Real-world Validation via Sim-to-Real Transfer

F 通过仿真到现实转移的真实环境验证

Table 6: Skill chaining performance of real-world long-horizon robotic manipulation tasks.
表6：真实环境中长时序机器人操作任务的技能链表现。

Method	Success rate
T-STAR	70% (2 sub-tasks) / 50% (3 sub-tasks)
SCaR	90% (2 sub-tasks) / 70% (3 sub-tasks)
方法	成功率
T-STAR	70%（2个子任务）/ 50%（3个子任务）
SCaR	90%（2个子任务）/ 70%（3个子任务）

Real-robot Experiment Setup We also evaluate the skill chaining performance of real-robot for solving simple yet intuitive real-world long-horizon manipulation. We set up two types of desktop-level long-horizon manipulation tasks. The robotic arm needs to pick-and-place 2 and 3 blue squares in sequence.
真实机器人实验设置 我们还评估了真实机器人在解决简单但直观的现实世界长时序操作中的技能链表现。我们设置了两种桌面级长时序操作任务。机械臂需要依次抓取并放置2个和3个蓝色方块。

We built the corresponding task environment using the gazebo simulation that accompanies the K1robot[5] ,and collect 50 demonstrations of grasping skills for each square for training. With camera calibration, we deploy agents trained under simulation in a real robot desktop task to solve 2-square as well as 3-square pick-and-place tasks without the need for adaptation processes. We conduct experiments with the Sagittarius K1 and use MoveIt2 library based on ROS 2 framework for controlling the arm. We use RGB observations from RealSense D435i camera on the wrist of the robotic arm.
我们使用随K1robot[5]附带的gazebo仿真构建了相应的任务环境，并为每个方块收集了50次抓取技能示范用于训练。通过相机标定，我们将仿真中训练的智能体部署到真实机器人桌面任务中，能够无需适应过程完成2个和3个方块的抓取放置任务。我们使用Sagittarius K1机器人进行实验，并基于ROS 2框架的MoveIt2库控制机械臂。机械臂腕部配备RealSense D435i相机，采集RGB观测。

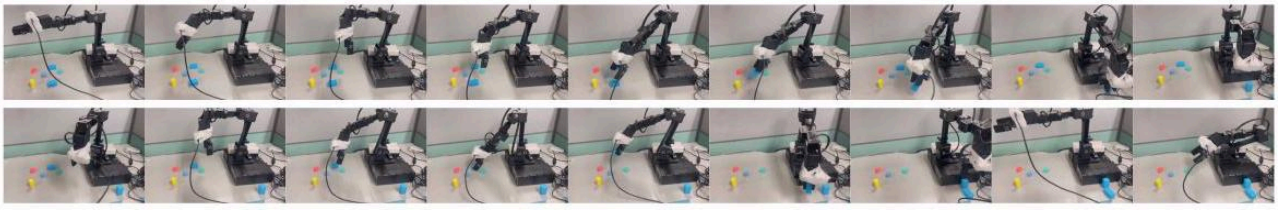


Figure 14: Visualization of the successful skill chaining in the 2-blue-square pick-and-place tasks using SCaR.
图14：使用SCaR在2个蓝色方块抓取放置任务中成功技能链的可视化。

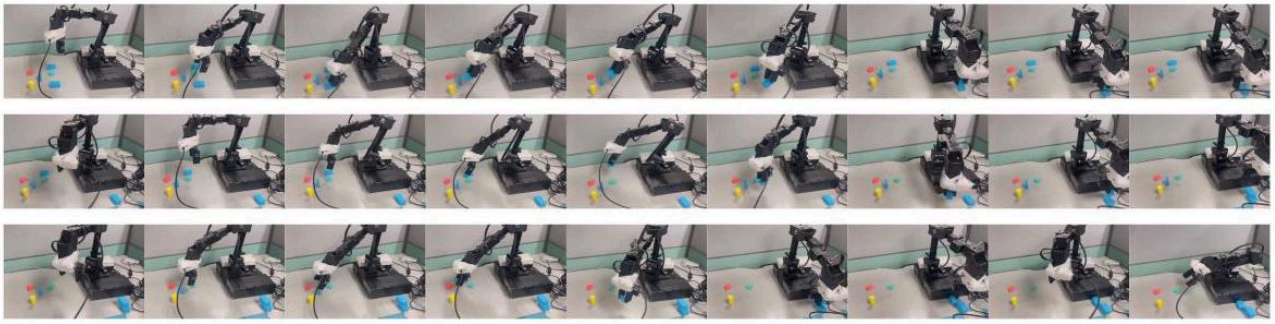


Figure 15: Visualization of the successful skill chaining in the 3-blue-square pick-and-place tasks using SCaR.
图15：使用SCaR在3个蓝色方块抓取放置任务中成功技能链的可视化。

Results For evaluation, we measure the success rate across 10 randomized square positions for each task. As shown in Table. 6, SCaR can solve the two long-horizon tasks and outperforms T-STAR baseline. Fig. 14 and Fig. 15 show the qualitative results of successful skill chaining in the 2 and 3-blue-square pick-and-place tasks using SCaR.
结果 评估时，我们测量了每个任务中10个随机方块位置的成功率。如表6所示，SCaR能够解决这两个长时序任务，且表现优于T-STAR基线。图14和图15展示了使用SCaR在2个和3个蓝色方块抓取放置任务中成功技能链的定性结果。

---

https://github.com/NXROBO/sagittarius_ws

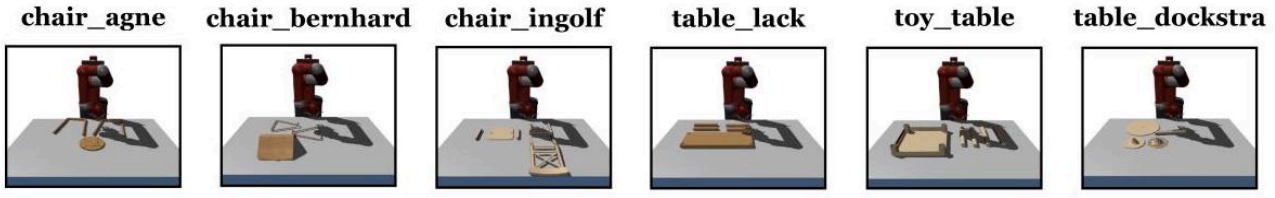


Figure 16: IKEA Furniture Assembly Environment for Long-Horizon Complex Manipulation **Tasks.**
图16：IKEA家具组装环境用于长时序复杂操作 **Tasks.**

35 G Environment Details

36 G 环境细节

36.1 G.1 IKEA Furniture Assembly

36.2 G.1 IKEA家具组装

We choose six tasks, chair_agne, chair_bernhard, chair_ingolf, table_lack, toy_table, and table_dockstra from the IKEA furniture assembly environment [6] [44] as the focal points of our experiments, as shown in Fig. 17. Our chosen robotic platform is the 7-DoF Rethink Sawyer robot, and we control it using joint velocity commands.
我们从IKEA家具组装环境[6][44]中选择了六个任务，分别是chair_agne、chair_bernhard、chair_ingolf、table_lack、toy_table和table_dockstra，作为实验重点，如图17所示。我们选用的机器人平台是7自由度的Rethink Sawyer机器人，控制方式为关节速度命令。

Observation Space The observation space comprises three key components: robot observations (29 dimensions), object observations (35 dimensions), and task phase information (8 dimensions). Robot observations encompass robot joint angles (7 dimensions), joint velocities (7 dimensions), gripper state (2 dimensions), gripper position (3 dimensions), gripper quaternion (4 dimensions), gripper velocity (3 dimensions), and gripper angular velocity (3 dimensions). Object observations include the positions (3 dimensions) and quaternions (4 dimensions) of all five furniture pieces in the scene. Task information, an 8-dimensional one-hot encoding, represents the current phase, including actions like reaching, grasping, lifting, moving, and aligning.
观测空间 观测空间包含三个关键组成部分：机器人观测（29维）、物体观测（35维）和任务阶段信息（8维）。机器人观测包括机器人关节角度（7维）、关节速度（7维）、夹爪状态（2维）、夹爪位置（3维）、夹爪四元数（4维）、夹爪速度（3维）和夹爪角速度（3维）。物体观测包括场景中五个家具部件的位置（3维）和四元数（4维）。任务信息为8维独热编码，表示当前阶段，包括到达、抓取、抬起、移动和对齐等动作。

Action space The action space includes arm movement, gripper control, and the connect action, which can vary based on different control modes: 6D end-effector space control using inverse kinematics, joint velocity control, and joint torque control.
动作空间 动作空间包括机械臂运动、夹爪控制和连接动作，具体取决于不同的控制模式：6D使用逆运动学的末端执行器空间控制、关节速度控制和关节力矩控制。

In the context of reinforcement learning (RL), we utilize a heavily shaped multi-phase dense reward obtained from the IKEA Furniture Assembly Environment [44].
在强化学习（RL）背景下，我们利用来自IKEA家具组装环境[44]的多阶段密集奖励进行强力的奖励塑形。

Environmental Reward Function The IKEA furniture assembly environmental reward function is a multi-phase reward defined with respect to a pair of furniture parts to attach (e.g., a table leg and a table top) and the corresponding manually annotated way-points,such as a target gripping point g for each part. The reward function for a pair of furniture parts consists of eight different phases as follows:
环境奖励函数 IKEA家具组装环境的奖励函数是针对一对家具部件（例如桌腿和桌面）及其对应的人工标注路径点（如每个部件的目标抓取点g）定义的多阶段奖励。该对家具部件的奖励函数包含以下八个不同阶段：

- Initial phase: The robot has to reconfigure its arm pose to an appropriate pose $\mathbf{p}_{\text{init}}$ for grasping a new furniture part. The reward is proportional to the negative distance between the end-effector $\mathbf{p}_{\text{eff}}$ and $\mathbf{p}_{\text{init}}$.
- 初始阶段：机器人需将机械臂重新配置到适合抓取新家具部件的姿态$\mathbf{p}_{\text{init}}$。奖励与末端执行器$\mathbf{p}_{\text{eff}}$与$\mathbf{p}_{\text{init}}$之间距离的负值成正比。

- Reach phase: The robot reaches above a target furniture part. The reward is proportional to the negative distance between the end-effector $\mathbf{p}_{\text{eff}}$ and a point $\mathbf{p}_{\text{reach}}$ 5 cm above the gripping point g .
- 到达阶段：机器人到达目标家具部件上方。奖励与末端执行器$\mathbf{p}_{\text{eff}}$与抓取点g上方某点$\mathbf{p}_{\text{reach}}$ 5 cm之间距离的负值成正比。

- Lower phase: The gripper is lowered onto the target part. The phase reward is proportional to the negative distance between $\mathbf{p}_{\text{eff}}$ and the target gripping points.
- 下降阶段：夹爪下降至目标部件。该阶段奖励与$\mathbf{p}_{\text{eff}}$与目标抓取点之间距离的负值成正比。

- Grasp phase: The robot learns to grasp the target part. The reward is given if the gripper contacts the part, and is proportional to the force exerted by the grippers.
- 抓取阶段：机器人学习抓取目标部件。当夹爪接触到部件时给予奖励，奖励与夹爪施加的力成正比。

- Lift phase: The robot lifts the gripped part up to $\mathbf{p}_{\text{lift}}$. The reward is proportional to the negative distance between the gripped part $\mathbf{p}_{\text{part}}$ and the target point $\mathbf{p}_{\text{lift}}$.
- 提升阶段：机器人将抓取的部件提升至$\mathbf{p}_{\text{lift}}$。奖励与抓取部件$\mathbf{p}_{\text{part}}$和目标点$\mathbf{p}_{\text{lift}}$之间的负距离成正比。

- Align phase: The robot roughly rotates the gripped part before moving it. The reward is proportional to the cosine similarity between up vectors $\mathbf{u}_A, \mathbf{u}_B$ and forward vectors $\mathbf{f}_A, \mathbf{f}_B$ of the two connectors.
- 对齐阶段：机器人在移动前粗略旋转抓取的部件。奖励与两个连接器的上向量$\mathbf{u}_A, \mathbf{u}_B$和前向量$\mathbf{f}_A, \mathbf{f}_B$之间的余弦相似度成正比。

---

https://github.com/clvrai/furniture

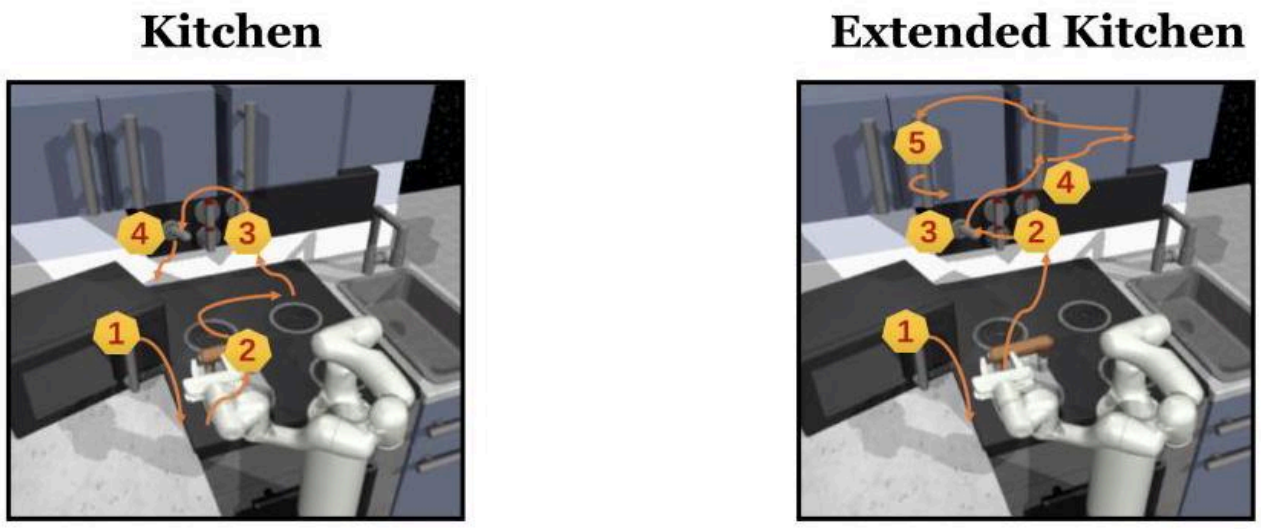


Figure 17: Kitchen Organization Environment for Long-Horizon Complex Manipulation Tasks.
图17：用于长时复杂操作任务的厨房组织环境。

- Move phase: The robot moves and aligns the gripped part to another part. The reward is proportional to the negative distance between the connector of the gripped part and a point $\mathbf{p}_{\text{move}}$ to 5 cm above the connector of another part,and the cosine similarity between two connector up vectors, $\mathbf{u}_A$ and $\mathbf{u}_B$,and forward vectors $\mathbf{f}_A$ and $\mathbf{f}_B$. Note that all connectors are labeled with aligned up vectors and forward vectors.
- 移动阶段：机器人移动并对齐抓取的部件与另一部件。奖励与抓取部件连接器和另一部件连接器上方从点$\mathbf{p}_{\text{move}}$到5 cm的负距离，以及两个连接器的上向量$\mathbf{u}_A$和$\mathbf{u}_B$、前向量$\mathbf{f}_A$和$\mathbf{f}_B$之间的余弦相似度成正比。注意，所有连接器均标注有对齐的上向量和前向量。

- Fine-grained move phase: The robot must finely align two connectors until attached. The same reward is used as the move phase with a higher coefficient, making the reward more sensitive to small changes. In addition, when the part is connectable, a reward is provided based on the activation of the connect action a [connect].
- 精细移动阶段：机器人必须精细对齐两个连接器直至连接。使用与移动阶段相同的奖励，但系数更高，使奖励对微小变化更敏感。此外，当部件可连接时，根据连接动作a[connect]的激活情况给予奖励。

Upon completion of each phase, completion rewards are given to encourage the agent to move on to the next phase. In addition to stage-based rewards, control penalties, stabilizing wrist pose rewards, and grasping rewards (i.e., opening the grasping hand only during the initial, arrival, and lower stages) are provided throughout the process. If the robot releases the grasped object, the phase ends early and a negative reward is provided. Phase completion depends on the robot and part configurations satisfying distance and angle constraints with respect to the goal configuration. After all stages are completed, the stage resets to the initial stage. This process repeats until all parts are connected.
每个阶段完成后，给予完成奖励以鼓励智能体进入下一阶段。除阶段奖励外，整个过程中还提供控制惩罚、稳定手腕姿态奖励和抓取奖励（即仅在初始、到达和放下阶段打开抓取手）。若机器人释放抓取物体，阶段提前结束并给予负奖励。阶段完成取决于机器人和部件配置是否满足与目标配置的距离和角度约束。所有阶段完成后，阶段重置至初始阶段。该过程重复，直至所有部件连接完成。

Demonstration Collection For imitation learning (IL), we gathered 200 demonstrations for each furniture part assembly using a programmatic assembly policy. Each demonstration for single-part assembly typically spans 150 steps, reflecting the overall task's inherently long-horizon nature.
示范收集 为模仿学习（IL），我们使用程序化装配策略为每个家具部件装配收集了200个示范。单部件装配的每个示范通常包含150步，反映了整体任务固有的长时性质。

Sub-tasks In our experiments, we define a sub-task as the process of assembling one part to another. Thus, the chair_agne and chair_bernhard tasks have two distinct sub-tasks, table_dockstra has three distinct sub-tasks, and chair_ingolf, table_lack, and toy_table have four distinct sub-tasks. These sub-tasks are trained independently, with their initial state sampled from the environment and random noise introduced in the $[-2\text{ cm}, 2\text{ cm}]$ and $[-3°, 3°]$ ranges of the(x,y)plane. Importantly,the decomposition of the sub-tasks is pre-determined, which means that the environment is initialized for each sub-task, and the agent receives a notification when a sub-task is successfully completed. Once the two components are firmly connected, the corresponding sub-task is considered completed and the robotic arm is guided back to its initial pose, i.e., at the center of the workspace.
子任务 在我们的实验中，子任务定义为将一个部件装配到另一个部件的过程。因此，chair_agne和chair_bernhard任务有两个不同的子任务，table_dockstra有三个不同的子任务，chair_ingolf、table_lack和toy_table有四个不同的子任务。这些子任务独立训练，其初始状态从环境中采样，并在(x,y)平面的$[-2\text{ cm}, 2\text{ cm}]$和$[-3°, 3°]$范围内引入随机噪声。重要的是，子任务的分解是预先确定的，意味着环境为每个子任务初始化，且智能体在子任务成功完成时收到通知。一旦两个组件牢固连接，对应子任务即视为完成，机械臂被引导回初始姿态，即工作空间中心。

Assembly Difficulty The difficulty of modeling furniture depends largely on the shape of the furniture. For example, the toy_table task with cylindrical legs is more difficult to grasp, whereas the table_lack task with rectangular legs is easier to grasp. Chairs are generally more difficult to assemble because of their irregular shape (e.g., seat and back). This is the reason why the success rates of the toy_table andchair_ingolf tasks are lower than the success rates of table_lack.
装配难度 家具建模的难度主要取决于家具的形状。例如，toy_table任务中带有圆柱形腿的家具更难抓取，而table_lack任务中带有矩形腿的家具则较易抓取。椅子因其不规则形状（如座椅和靠背）通常更难装配。这也是toy_table和chair_ingolf任务的成功率低于table_lack的原因。

36.3 G.2 Kitchen Organization

36.4 G.2 厨房组织

We use the Franka Kitchen tasks in D4RL [45] and refer to the experimental setup in SkiMo [46] for the sub-task extensions. Including the following two tasks: Kitchen task and Extended Kitchen task, as shown in Fig. 17.
我们使用D4RL [45]中的Franka Kitchen任务，并参考SkiMo [46]中的实验设置进行子任务扩展。包括以下两个任务：厨房任务和扩展厨房任务，如图17所示。

Kitchen The 7-DoF Franka Emika Panda robot arm is tasked with performing four sequential sub-tasks: Turn on the microwave - Move the kettle - Turn on the stove - Turn on the lights.
厨房 7自由度的Franka Emika Panda机器人手臂被要求依次完成四个子任务：打开微波炉——移动水壶——打开炉灶——打开灯。

Extended Kitchen The environment and task-agnostic data used in this experiment are consistent with those employed in the Kitchen scenario. However, we introduce a different set of sub-tasks for this experiment, namely: Turn on the microwave - Turn on the stove - Turn on the lights - Slide the cabinets to the right - Open the cabinets, as depicted in Fig. 17 (right). It's worth noting that this sequence of tasks is not aligned with the sub-task transition probabilities observed in the task-agnostic data, posing a challenge for exploration based on prior data.
扩展厨房 本实验中使用的环境和任务无关数据与厨房场景中所用数据保持一致。然而，我们为本实验引入了一组不同的子任务，即：打开微波炉——打开炉灶——打开灯——将橱柜向右滑动——打开橱柜，如图17（右）所示。值得注意的是，这一任务序列与任务无关数据中观察到的子任务转移概率并不一致，这对基于先前数据的探索提出了挑战。

Observation Space The agent operates within a 30-dimensional observation space, which includes an 11-dimensional robot proprioceptive state and 19-dimensional object states. This modified observation space removes a constant 30-dimensional goal state found in the original environment.
观测空间 智能体在一个30维的观测空间中运行，包括11维的机器人本体感知状态和19维的物体状态。该修改后的观测空间移除了原环境中恒定的30维目标状态。

Action Space The agent's action space consists of 9 dimensions, encompassing 7-dimensional joint velocity control and 2-dimensional gripper velocity control.
动作空间 智能体的动作空间为9维，包括7维的关节速度控制和2维的夹爪速度控制。

Environmental Reward Function In terms of the environmental rewards, the agent receives a reward of +1 for each completed sub-task. The total episode length is set to 280 steps, and an episode concludes once all sub-tasks are successfully accomplished. The initial state is initialized with slight noise introduced in each state dimension.
环境奖励函数 在环境奖励方面，智能体每完成一个子任务可获得+1的奖励。每个回合的总步数设为280步，当所有子任务均成功完成时，回合结束。初始状态在每个状态维度中引入了微小噪声进行初始化。

Demonstration Collection For imitation learning, we collect 200 demonstrations per sub-task with reference to the dataset in [52] that obtained through teleoperation. This dataset covers interactions with all seven manipulatable objects within the environment.
示范收集 为了模仿学习，我们参考[52]中的数据集，通过远程操作为每个子任务收集了200个示范。该数据集涵盖了与环境中所有七个可操控物体的交互。

37 H Network Architecture

38 H 网络结构

For a fair comparison, our method and the benchmark methods use the same network structure. The policy network and the critic network consist of two layers of 128 and 256 hidden units fully connected with ReLU nonlinear properties, respectively. The output layer of the actor network outputs an action distribution, which consists of the mean and standard deviation of a Gaussian distribution. The critic network outputs only one critic value. The discriminator of GAIL [39] and the bi-directional discriminator of our proposed approach use a two-layer fully connected network with 256 hidden units. The outputs of these discriminators are clipped between $[0, 1]$,following the least-square GAIL proposed by [40].
为了公平对比，我们的方法与基线方法采用相同的网络结构。策略网络和评论家网络均由两层全连接的隐藏单元组成，分别为128和256个，激活函数为ReLU。演员网络的输出层输出一个动作分布，包括高斯分布的均值和标准差。评论家网络仅输出一个评论值。GAIL [39]的判别器以及我们提出方法的双向判别器均采用两层全连接网络，隐藏单元为256个。这些判别器的输出被限制在$[0, 1]$之间，遵循[40]提出的最小二乘GAIL方法。

39 I Training Details

40 I 训练细节

40.1 I.1 Computing Resources

40.2 I.1 计算资源

Our method and all baselines were implemented using PyTorch [53]. All experiments were conducted on workstations equipped with Intel(R) Xeon(R) Gold 5218 CPUs and dual NVIDIA GeForce RTX 3080 GPUs. In the SCaR framework, the pre-training phase for each sub-task skill policy, spanning 150M time steps,took approximately 10 hours with dual regularization,compared to about 8 hours without it (Fixed-RL-IL), leading to an added computational overhead of roughly 2 hours. The full testing and evaluation process of skill chaining for a complete long-horizon task required an additional 10 to 15 hours, depending on the task's complexity. For comparison, training the skill dynamics model in SkiMo [46] took approximately 24 hours (100M steps). Training baselines such as PPO [47], GAIL [39], and Fixed-RL-IL took longer, requiring about 48 hours each, as these methods train the entire long-horizon task from scratch,with 450M time steps for each complete task. In our evaluation, we used 5 seeds, each tested over 200 episodes, resulting in an average real-time execution time of about 36-54 seconds per single long-horizon task.

我们的方法及所有基线均使用PyTorch [53]实现。所有实验均在配备Intel(R) Xeon(R) Gold 5218 CPU和双NVIDIA GeForce RTX 3080 GPU的工作站上进行。在SCaR框架中，每个子任务技能策略的预训练阶段，历时150M步，采用双重正则化约需10小时，而不采用时（Fixed-RL-IL）约需8小时，增加了约2小时的计算开销。完整长时序任务的技能链测试与评估过程还需额外10至15小时，具体取决于任务复杂度。相比之下，SkiMo [46]中技能动力学模型的训练约需24小时（1亿步）。PPO [47]、GAIL [39]和Fixed-RL-IL等基线方法的训练时间更长，因其需从零开始训练整个长时序任务，每个完整任务约需48小时，步数为450M。在评估中，我们使用5个随机种子，每个种子测试200个回合，单个长时序任务的平均实时执行时间约为36-54秒。

40.3 I.2 Algorithm Implementation Details

40.4 I.2 算法实现细节

We report the hyperparameters used in our experiments in Table 7.

我们在表7中报告了实验中使用的超参数。

Table 7: Hyperparameters used in our experiments.

表7：我们实验中使用的超参数。

Hyperparameter	Value
Rollout Size	1024
Learning Rate	0.0003
Learning Rate Decay	Linear decay
Mini-batch Size	128
Discount Factor	0.99
Entropy Coefficient	0.003
Reward Scale	0.05
State Normalization	True
Discriminator learning rate	$1{e}^{-4}$
Sub-task training steps	150000000
#Workers	20
#Epochs per Update	10
Base exponent for balancing α	0.5
${k}_{1}$ (used to flatten the mapping function during $p \in \left\lbrack {0,\frac{T}{2}}\right)$)	10
${k}_{2}$ (used to flatten the mapping function during $p \in \left\lbrack {\frac{T}{2},T}\right)$)	30
Weighting factor ${\lambda }_{\mathrm{{Bi}}}$	10000
ρ (for imitation progress recognizer Φ)	0.9
Penalty coefficient ${\eta }^{\mathrm{{gp}}}$	10

超参数	数值
展开大小	1024
学习率	0.0003
学习率衰减	线性衰减
小批量大小	128
折扣因子	0.99
熵系数	0.003
奖励缩放	0.05
状态归一化	是
判别器学习率	$1{e}^{-4}$
子任务训练步数	150000000
工作线程数	20
每次更新的训练轮数	10
平衡的基数指数 α	0.5
${k}_{1}$（用于在 $p \in \left\lbrack {0,\frac{T}{2}}\right)$ 期间平滑映射函数）	10
${k}_{2}$（用于在 $p \in \left\lbrack {\frac{T}{2},T}\right)$ 期间平滑映射函数）	30
加权因子 ${\lambda }_{\mathrm{{Bi}}}$	10000
ρ（用于模仿进度识别器 Φ）	0.9
惩罚系数 ${\eta }^{\mathrm{{gp}}}$	10

Table 8: Comparison to prior work and ablated methods.
表8：与先前工作及消融方法的比较。

Method	Model-based	Skill-based	Scratch training	Joint training
PPO [47] and GAIL [39]	✗	✗	✓	✗
Fixed-RL-IL [40]	✗	✗	✓	✓
SkiMo [46]	✓	✓	✓	✓
Policy Sequencing [12]	✓	✓	✗	✓
T-STAR [15]	✗	✓	✗	✓
SCaR (Ours) and SCaR w/o Bi and SCaR w/o AES	✓	✓	✗	✓

方法	基于模型	基于技能	从零训练	联合训练
PPO [47] 和 GAIL [39]	✗	✗	✓	✗
Fixed-RL-IL [40]	✗	✗	✓	✓
SkiMo [46]	✓	✓	✓	✓
策略序列 [12]	✓	✓	✗	✓
T-STAR [15]	✗	✓	✗	✓
SCaR（本方法）及无 Bi 版本和无 AES 版本	✓	✓	✗	✓

For the baseline implementations, we use the official code for PPO [47], GAIL [39], Fixed-RL-IL [40], SkiMo [46], Policy Sequencing [12] and T-STAR [15]. The table below (Table 8) compares key components of **SCaR** with model-based,model-free and skill-based baselines and ablated methods, where joint training indicates whether or not reinforcement learning combined with imitation learning is used for training.
对于基线实现，我们使用了PPO [47]、GAIL [39]、Fixed-RL-IL [40]、SkiMo [46]、Policy Sequencing [12]和T-STAR [15]的官方代码。下表（表8）比较了**SCaR**与基于模型、无模型和基于技能的基线及消融方法的关键组成部分，其中联合训练表示是否使用强化学习与模仿学习相结合进行训练。

PPO [47] Any reinforcement learning algorithm can be used for policy optimization, in this paper we choose to use Proximal Policy Optimization (PPO) and use the default hyperparameters of PPO [47].
PPO [47] 任何强化学习算法都可用于策略优化，本文选择使用近端策略优化（Proximal Policy Optimization，PPO）并采用PPO [47]的默认超参数。

GAIL [39] In this paper we choose to use Generative Adversarial Imitation Learning (GAIL) [39] as the learning algorithm for imitation learning and use the default hyperparameters of GAIL [39]. We specifically use an agent states s to discriminate agent and expert trajectories,instead of state-action pairs(s,a).
GAIL [39] 本文选择使用生成对抗模仿学习（Generative Adversarial Imitation Learning，GAIL）[39]作为模仿学习的算法，并采用GAIL [39]的默认超参数。我们特别使用代理状态s来区分代理和专家轨迹，而非状态-动作对(s,a)。

Fixed-RL-IL [12] We adopt the AMP [40] solution combining environmental rewards and least square GAIL with $\lambda_{\mathrm{RL}} = \lambda_{\mathrm{IL}} = 0.5$. For implementation details of least square GAIL training and GAIL rewards, see original paper [40].
Fixed-RL-IL [12] 我们采用结合环境奖励和最小二乘GAIL的AMP [40]方案与$\lambda_{\mathrm{RL}} = \lambda_{\mathrm{IL}} = 0.5$。关于最小二乘GAIL训练和GAIL奖励的实现细节，详见原文[40]。

SkiMo [46] We use the official implementation of the original paper and use the hyperparameters suggested in the official implementation.
SkiMo [46] 我们使用原文的官方实现，并采用官方实现中建议的超参数。

Policy Sequencing [12] We employ the official implementation and the hyperparameters provided by [15].
Policy Sequencing [12] 我们采用官方实现及[15]提供的超参数。

T-STAR [15] We use the official implementation of the original paper and use the hyperparameters suggested in the official implementation [15].
T-STAR [15] 我们使用原文的官方实现，并采用官方实现[15]中建议的超参数。

SCaR (ours) We refer to T-STAR and use $\lambda_{\text{Re}} = 10000$ for bi-directional regularization. We take 50% of the initial state samples from the start environment of each policy, 50% of the terminal state samples at the end,and 50% of the initial state buffer and 50% of the terminal state buffer from the previous skill, respectively.
SCaR（本方法）我们参考T-STAR并使用$\lambda_{\text{Re}} = 10000$进行双向正则化。我们分别取每个策略起始环境的初始状态样本的50%，终止状态样本的50%，以及前一技能的初始状态缓冲区的50%和终止状态缓冲区的50%。

41 J Limitations and Potential Solutions

42 J 限制与潜在解决方案

While SCaR demonstrates strong capabilities, it does have certain limitations. In the following sections, we outline these limitations and propose potential solutions.
尽管SCaR展现了强大的能力，但仍存在一定的局限性。以下章节中，我们将概述这些限制并提出潜在的解决方案。

Limited Task Generalization SCaR primarily focuses on predefined, structured environments to validate its mechanism for chaining pre-trained sub-task skills within long-horizon tasks. Consequently, our current study does not address SCaR's adaptability across varied or novel task environments. While we demonstrate SCaR's robustness to unknown perturbations within a single environment (e.g., unexpected forces applied to the robot arm joints in specific sub-tasks), the system's ability to generalize to entirely new or unfamiliar environments remains unexplored. A potential solution is to leverage foundational models to expand SCaR's applicability. In long-horizon tasks where sub-task definitions are unclear or missing, foundational models can use their powerful task-planning capabilities to divide tasks into logical sub-tasks [54, 55, 56].When unexpected subtask demands or changes in overall task goals arise, these models can re-plan sub-tasks accordingly. Another approach involves human-in-the-loop learning, incorporating human guidance in the training pipeline, such as using human priors for sub-task division [32] or employing methods to manage the subjectivity of human-labeled rewards through preference learning [57].
任务泛化能力有限 SCaR主要聚焦于预定义的结构化环境，以验证其在长时域任务中串联预训练子任务技能的机制。因此，当前研究未涉及SCaR在多样或新颖任务环境中的适应性。虽然我们展示了SCaR对单一环境中未知扰动（如特定子任务中施加于机器人臂关节的意外力）的鲁棒性，但系统对全新或陌生环境的泛化能力尚未探讨。潜在的解决方案是利用基础模型扩展SCaR的适用性。在子任务定义不明确或缺失的长时域任务中，基础模型可利用其强大的任务规划能力将任务划分为逻辑子任务[54, 55, 56]。当出现意外的子任务需求或整体任务目标变化时，这些模型能够相应地重新规划子任务。另一种方法是引入人机交互学习，在训练流程中融入人类指导，如利用人类先验进行子任务划分[32]，或通过偏好学习[57]管理人类标注奖励的主观性。

Extensive Retraining for New Tasks SCaR faces limitations in adapting to diverse long-horizon manipulation tasks or different robotic setups, as it requires extensive retraining of sub-task skills for each new task. This reliance on full retraining restricts SCaR's efficiency and scalability when addressing evolving or variable task requirements, particularly in complex, real-world environments. A potential solution is to integrate online reinforcement learning into SCaR's skill-learning process, allowing it to adapt to task variations, such as different table designs, without full retraining, enabling more efficient adaptation.
新任务需大量再训练 SCaR在适应多样的长时域操作任务或不同机器人配置时存在局限，因为每个新任务都需对子任务技能进行大量再训练。这种对全面再训练的依赖限制了SCaR在应对不断变化或多样化任务需求时的效率和可扩展性，尤其是在复杂的现实环境中。潜在的解决方案是将在线强化学习整合进SCaR的技能学习过程中，使其能够适应任务变化（如不同的桌面设计），无需全面再训练，从而实现更高效的适应。

Lack of visual input handling SCaR currently lacks the capability to process encoded image and semantic state inputs, limiting its applicability in tasks that rely on visual or semantic information for effective performance. The current setup is largely dependent on environments with direct access to state information, as both simulated environments are based on state representations. While effective in controlled setups, enabling SCaR to learn from image encodings as states would enhance its robustness and applicability in tasks requiring nuanced visual and semantic processing. Given recent advancements in learning multi-view representations [58, 59] and generating 3D models [60, 61], incorporating these improvements and 3D priors into the encoder design could enhance the sample efficiency of SCaR.
缺乏视觉输入处理能力 SCaR目前无法处理编码后的图像和语义状态输入，限制了其在依赖视觉或语义信息以实现有效表现的任务中的适用性。当前设置主要依赖于能够直接访问状态信息的环境，因为两个模拟环境均基于状态表示。虽然在受控环境中效果良好，但使SCaR能够从图像编码中学习状态，将提升其在需要细致视觉和语义处理任务中的鲁棒性和适用性。鉴于近期在多视角表示学习[58, 59]和三维模型生成[60, 61]方面的进展，将这些改进及三维先验融入编码器设计，有望提升SCaR的样本效率。

43 K Potential negative impacts

44 K 潜在负面影响

Since our method is currently limited to applications in simulated environments and simple desktop-level robot manipulation, it is not expected to have a significant negative impact on society. However, privacy concerns may arise if our method is applied to real-world long time-series tasks with mobility, as imitation learning agents used in applications such as autonomous driving [62] or real-time control [63, 64] require large amounts of data that often contain controversial information. In addition, the imitation learning policy is a challenge because it imitates a specified demonstration that may include bad behavior. If the expert demonstration includes some nefarious behaviors (e.g., training data for a mobile manipulation task includes behaviors that may be violent towards pedestrians), then the policy may have a significant negative impact on the user. To address this issue, future directions should focus on developing agents with safety adaptations in addition to improving performance.
由于我们的方法目前仅限于模拟环境和简单桌面级机器人操作的应用，预计不会对社会产生重大负面影响。然而，如果将我们的方法应用于具有移动性的真实世界长时间序列任务，可能会引发隐私问题，因为用于自动驾驶[62]或实时控制[63, 64]等应用的模仿学习代理需要大量数据，这些数据往往包含有争议的信息。此外，模仿学习策略存在挑战，因为它模仿的是指定的示范，而示范中可能包含不良行为。如果专家示范中包含一些恶意行为（例如，移动操作任务的训练数据中包含可能对行人具有暴力倾向的行为），那么该策略可能对用户产生显著负面影响。为解决此问题，未来的研究方向应侧重于在提升性能的同时，开发具备安全适应性的代理。

45 NeurIPS Paper Checklist

46 NeurIPS 论文清单

47 1. Claims

48 1. 论断

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
问题：摘要和引言中提出的主要论断是否准确反映了论文的贡献和范围？

Answer: [Yes]
回答：[是]

Justification: We propose a new skill chaining framework for long time-series robotic manipulation tasks that improves overall task completion performance by providing dual regularization for intra- and inter-skill dependencies. We hope this work will inspire future research to further explore the potential of skill chaining for long-horizon robotic manipulation.
理由：我们提出了一种用于长时间序列机器人操作任务的新技能链框架，通过对技能内和技能间依赖关系进行双重正则化，提升了整体任务完成性能。我们希望这项工作能激发未来研究，进一步探索技能链在长远机器人操作中的潜力。

Guidelines:
指导原则：

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- 答案为NA表示摘要和引言中未包含论文中的论断。
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- 摘要和/或引言应明确陈述所提出的论断，包括论文的贡献及重要假设和限制。对此问题回答“No”或“NA”将不被评审认可。
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- 所提出的论断应与理论和实验结果相符，并反映结果在其他环境中的推广程度。
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.
- 作为动机包含理想目标是可以的，但需明确这些目标并未在论文中实现。

49 2. Limitations

50 2. 限制

Question: Does the paper discuss the limitations of the work performed by the authors? Answer: [Yes]
问题：论文是否讨论了作者所做工作的限制？回答：[是]

Justification: We discuss limitations in the last section of the main paper: limitations mainly exist in that 1) the sub-tasks in our framework are predefined, 2) we did not test our method on a more challenging real robot furniture assembly task due to limited hardware. Guidelines:
理由：我们在论文最后一节讨论了限制：主要存在于1）框架中的子任务是预定义的，2）由于硬件限制，我们未在更具挑战性的真实机器人家具组装任务上测试方法。指导原则：

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- 答案为NA表示论文无任何限制，答案为No表示论文存在限制但未讨论。

- The authors are encouraged to create a separate "Limitations" section in their paper.
- 鼓励作者在论文中单独设立“限制”章节。

- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- 论文应指出任何强假设及其对这些假设违背的结果稳健性（例如，独立性假设、无噪声环境、模型良好设定、渐近近似仅在局部成立）。作者应反思这些假设在实际中可能被违背的情况及其影响。

- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- 作者应反思所提出结论的适用范围，例如该方法是否仅在少数数据集或少量运行中测试过。一般而言，实证结果往往依赖隐含假设，这些假设应予以明确说明。

- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- 作者应反思影响该方法性能的因素。例如，面部识别算法在图像分辨率低或光线昏暗时可能表现不佳；语音转文字系统可能无法可靠地为在线讲座提供闭幕字幕，因为它无法处理专业术语。

- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- 作者应讨论所提算法的计算效率及其随数据集规模的扩展性。

- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- 如适用，作者应讨论其方法在解决隐私和公平性问题上的可能局限性。

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.
- 虽然作者可能担心对局限性的完全坦诚会被评审作为拒稿理由，但更糟的情况是评审发现论文未承认的局限。作者应以最佳判断行事，认识到个人对透明度的坚持在形成维护社区诚信的规范中起重要作用。评审将被特别指示不因坦诚局限而予以惩罚。

51 3. Theory Assumptions and Proofs

52 3. 理论假设与证明

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?
问题：对于每个理论结果，论文是否提供了完整的假设集合及完整（且正确）的证明?

Answer: [Yes]
回答：[是]

Justification: We provide further explanation in the Appendix to explain the assumptions presented in the main paper.
理由：我们在附录中进一步解释了主文中提出的假设。

Guidelines:
指导原则：

- The answer NA means that the paper does not include theoretical results.
- 答案为NA表示论文不包含理论结果。

- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- 论文中的所有定理、公式和证明应编号并相互引用。

- All assumptions should be clearly stated or referenced in the statement of any theorems.

- 所有假设应在任何定理陈述中明确说明或引用。
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- 证明可以出现在主文或补充材料中，但若在补充材料中，鼓励作者提供简短的证明概要以便直观理解。
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- 反之，论文主体中提供的任何非正式证明应辅以附录或补充材料中的正式证明。
- Theorems and Lemmas that the proof relies upon should be properly referenced.
- 证明所依赖的定理和引理应当被适当引用。

4. Experimental Result Reproducibility

4. 实验结果的可复现性

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?
问题：论文是否充分披露了复现论文主要实验结果所需的全部信息，且这些信息影响论文的主要论断和/或结论（无论是否提供代码和数据）？

Answer: [Yes]

答案：[是]

Justification: We further describe the network architecture, training details, dataset, and the open source codebase on which the method is based in the Appendix.
理由：我们在附录中进一步描述了网络架构、训练细节、数据集以及该方法所基于的开源代码库。

Guidelines:

指南：

- The answer NA means that the paper does not include experiments.
- 答案为NA表示论文不包含实验内容。
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- 如果论文包含实验，回答“否”将不被评审认可：使论文可复现非常重要，无论是否提供代码和数据。
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- 如果贡献是数据集和/或模型，作者应描述为使结果可复现或可验证所采取的步骤。
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- 根据贡献类型，可复现性可通过多种方式实现。例如，若贡献是新颖架构，完整描述架构可能足够；若贡献是特定模型及其实证评估，则可能需要使他人能够用相同数据集复现模型，或提供模型访问权限。通常，发布代码和数据是实现这一目标的有效方式，但也可通过详细的复现说明、托管模型访问（如大型语言模型）、发布模型检查点或其他适合所做研究的方式实现可复现性。
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
- 虽然NeurIPS不强制要求发布代码，但会议要求所有投稿提供合理的可复现途径，具体取决于贡献性质。例如

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
(a) 若贡献主要是新算法，论文应明确说明如何复现该算法。

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(b) 若贡献主要是新模型架构，论文应清晰完整地描述该架构。

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(c) 若贡献是新模型（如大型语言模型），应提供访问该模型以复现结果的方式，或提供复现模型的方法（如开源数据集或构建数据集的说明）。

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.
(d) 我们理解某些情况下复现较为困难，作者可描述其提供复现的具体方式。对于闭源模型，访问可能有限制（如仅限注册用户），但应确保其他研究者有途径复现或验证结果。

5. Open access to data and code

5. 数据和代码的开放获取

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?
问题：论文是否提供了数据和代码的开放访问，并附有足够的说明以忠实复现主要实验结果，如补充材料中所述？

Answer: [NA]
回答：[不适用]

Justification: All simulation environments, datasets, and open-source code libraries used to implement our method are described in the Appendix.
理由：所有用于实现我们方法的仿真环境、数据集和开源代码库均在附录中进行了描述。

Guidelines:
指导原则：

- The answer NA means that paper does not include experiments requiring code.
- 答案为不适用（NA）表示论文不包含需要代码的实验。
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/ public/guides/CodeSubmissionPolicy) for more details.
- 详情请参见NeurIPS代码和数据提交指南（https://nips.cc/public/guides/CodeSubmissionPolicy）。
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- 虽然我们鼓励发布代码和数据，但理解这可能不可行，因此“不”也是可接受的答案。除非代码是贡献的核心（例如新的开源基准），否则不能仅因未包含代码而拒稿。
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https: //nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- 说明应包含运行复现结果所需的准确命令和环境。详情请参见NeurIPS代码和数据提交指南（https://nips.cc/public/guides/CodeSubmissionPolicy）。
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- 作者应提供数据访问和准备的说明，包括如何访问原始数据、预处理数据、中间数据和生成数据等。
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- 作者应提供脚本以复现新提出方法和基线的所有实验结果。如仅部分实验可复现，应说明哪些实验未包含在脚本中及原因。
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- 提交时，为保持匿名，作者应发布匿名版本（如适用）。
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.
- 建议在补充材料（附于论文后）中提供尽可能多的信息，但允许包含数据和代码的URL链接。

6. Experimental Setting/Details

6. 实验设置/细节

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?
问题：论文是否明确了所有训练和测试细节（例如数据划分、超参数及其选择方式、优化器类型等），以便理解结果？

Answer: [Yes]
回答：[是]

Justification: We describe partial details in the main paper and provide further details in the Appendix.
理由：我们在正文中描述了部分细节，并在附录中提供了更多细节。

Guidelines:
指导原则：

- The answer NA means that the paper does not include experiments.
- 答案为NA表示论文中不包含实验内容。
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- 实验设置应在论文主体中详细呈现，达到能够理解和评估结果所必需的程度。
- The full details can be provided either with the code, in appendix, or as supplemental material. 7. Experiment Statistical Significance
- 详细信息可以随代码提供，或附录中，或作为补充材料。7. 实验统计显著性

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?
问题：论文是否适当且正确定义了误差条，或提供了其他关于实验统计显著性的相关信息？

Answer: [Yes]
回答：[是]

Justification: Our experiments perform means and standard deviations for the five seed results.
理由：我们的实验对五个随机种子结果计算了均值和标准差。

Guidelines:
指导原则：

- The answer NA means that the paper does not include experiments.
- 答案为NA表示论文中不包含实验内容。
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- 如果结果附带误差条、置信区间或统计显著性检验，至少针对支持论文主要论点的实验，作者应回答“是”。
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- 应明确说明误差条所反映的变异因素（例如，训练/测试划分、初始化、某些参数的随机抽取，或在给定实验条件下的整体运行）。
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- 应解释误差条的计算方法（闭式公式、调用库函数、自助法等）。
- The assumptions made should be given (e.g., Normally distributed errors).
- 应说明所做的假设（例如，误差服从正态分布）。
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- 应明确误差条是标准差还是均值的标准误。
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96%CI ,if the hypothesis of Normality of errors is not verified.
- 报告1σ误差条是可以接受的，但应明确说明。若误差正态性假设未验证，作者最好报告2σ误差条，而非仅声明存在96%CI。
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- 对于非对称分布，作者应注意避免在表格或图中显示对称误差条，导致结果超出合理范围（例如负误差率）。
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.
- 如果在表格或图表中报告了误差线，作者应在正文中说明其计算方法，并在正文中引用相应的图表。

63 8. Experiments Compute Resources

64 8. 实验计算资源

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?
问题：对于每个实验，论文是否提供了足够的信息说明复现实验所需的计算资源（计算工作节点类型、内存、执行时间）？

Answer: [Yes]
回答：[是]

Justification: We illustrate the computational resources and the time required for the experiments in the Appendix.
理由：我们在附录中说明了实验所需的计算资源和时间。

Guidelines:
指导原则：

- The answer NA means that the paper does not include experiments.
- 答案为NA表示论文中不包含实验内容。
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- 论文应指明计算工作节点类型，如CPU或GPU，内部集群或云服务提供商，并包括相关的内存和存储信息。
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- 论文应提供每次单独实验运行所需的计算量，并估算总计算量。
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).
- 论文应披露整个研究项目是否比论文中报告的实验需要更多计算资源（例如，未纳入论文的初步或失败实验）。

9. Code Of Ethics

9. 伦理守则

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?
问题：论文中进行的研究是否在各方面均符合NeurIPS伦理守则 https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]
回答：[是]

Justification: Yes, the research conducted in the paper conforms in every respect with the NeurIPS Code of Ethics.
理由：是的，论文中进行的研究在各方面均符合NeurIPS伦理守则。

Guidelines:
指导原则：

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- 答案为NA表示作者未审阅NeurIPS伦理守则。
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- 如果作者回答否，应说明需要偏离伦理守则的特殊情况。
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).
- 作者应确保匿名性（例如，若因其司法管辖区的法律或法规有特殊考虑）。

10. Broader Impacts

10. 更广泛的影响

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?
问题：论文是否讨论了所开展工作的潜在积极社会影响和消极社会影响?

Answer: [Yes]
回答：[是]

Justification: We elaborate on these in the final section of the Appendix.
理由：我们在附录的最后部分对此进行了详细阐述。

Guidelines:
指导原则：

- The answer NA means that there is no societal impact of the work performed.
- 答案为NA表示该工作没有社会影响。
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- 如果作者回答NA或否，应解释其工作为何没有社会影响或为何论文未涉及社会影响。

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- 消极社会影响的例子包括潜在的恶意或非预期用途（例如，虚假信息、生成虚假身份、监控）、公平性考虑（例如，部署可能对特定群体产生不公平影响的技术决策）、隐私考虑和安全考虑。

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- 会议预期许多论文属于基础研究，且不针对特定应用，更不用说部署。然而，如果存在通向任何负面应用的直接路径，作者应予以指出。例如，指出生成模型质量的提升可能被用于生成用于虚假信息的深度伪造（deepfakes）是合理的。另一方面，无需指出通用的神经网络优化算法可能使人们更快训练生成深度伪造的模型。

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- 作者应考虑技术在按预期使用且正常运行时可能产生的危害，技术按预期使用但产生错误结果时的危害，以及技术被（有意或无意）滥用时的危害。

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).
- 若存在负面社会影响，作者还可讨论可能的缓解策略（例如，模型的分级发布、提供防御措施以配合攻击、滥用监控机制、监控系统随时间从反馈中学习的机制、提升机器学习的效率和可及性）。

69 11. Safeguards

70 11. 保障措施

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?
问题：论文是否描述了为负责任地发布存在高滥用风险的数据或模型（例如，预训练语言模型、图像生成器或爬取的数据集）而采取的保障措施？

Answer: [NA]
回答：[NA]

Justification: Our paper does not release data or models with high risks.
理由：我们的论文未发布具有高风险的数据或模型。

Guidelines:
指导原则：

- The answer NA means that the paper poses no such risks.
- 答案NA表示论文不存在此类风险。

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- 对于存在高误用风险或双重用途风险的已发布模型，应附带必要的安全措施以实现模型的受控使用，例如要求用户遵守使用指南或访问限制，或实施安全过滤器。

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- 从互联网抓取的数据集可能存在安全风险。作者应说明如何避免发布不安全的图像。

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.
- 我们认识到提供有效的安全措施具有挑战性，且许多论文不要求此项，但我们鼓励作者考虑这一点并尽最大努力。

71 12. Licenses for existing assets

72 12. 现有资源的许可

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?
问题：论文中使用的资源（如代码、数据、模型）的创作者或原始所有者是否得到了适当的致谢，且许可和使用条款是否明确提及并得到遵守?

Answer: [Yes]
回答：[是]

Justification: We have cited or provided URLs to all the code, data, and models used in the paper.
理由：我们已引用或提供了论文中使用的所有代码、数据和模型的链接。

Guidelines:
指导原则：

- The answer NA means that the paper does not use existing assets.
- 答案NA表示论文未使用现有资源。
- The authors should cite the original paper that produced the code package or dataset.
- 作者应引用产生代码包或数据集的原始论文。
- The authors should state which version of the asset is used and, if possible, include a URL.
- 作者应说明所用资源的版本，并在可能的情况下附上链接。
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- 每个资源应包含许可名称（例如CC-BY 4.0）。
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- 对于从特定来源（如网站）抓取的数据，应提供该来源的版权和服务条款。
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- 若发布资源，应提供包内的许可、版权信息和使用条款。对于流行数据集，paperswithcode.com/datasets整理了一些数据集的许可，其许可指南可帮助确定数据集的许可类型。
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- 对于重新打包的现有数据集，应同时提供原始许可和派生资源的许可（若有变更）。
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.
- 如果这些信息无法在线获取，建议作者联系资源的创建者。

72.1 13.New Assets

72.2 13.新资源

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?
问题：论文中引入的新资源是否有充分的文档说明，且文档是否随资源一同提供?

Answer: [NA]
回答：[不适用]

Justification: Our paper does not release new assets.
理由：我们的论文未发布新资源。

Guidelines:
指导原则：

- The answer NA means that the paper does not release new assets.
- 回答“不适用”表示论文未发布新资源。

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- 研究人员应通过结构化模板在提交时提供数据集/代码/模型的详细信息，包括训练细节、许可、限制等。
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- 论文应讨论是否以及如何获得使用资源相关人员的同意。
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.
- 提交时请记得对资源进行匿名处理（如适用）。可以创建匿名链接或包含匿名压缩文件。

14. Crowdsourcing and Research with Human Subjects

14. 众包与涉及人类受试者的研究

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?
问题：对于众包实验和涉及人类受试者的研究，论文是否包含提供给参与者的完整指令文本和截图（如适用），以及补偿细节（如有）？

Answer: [NA]
回答：[不适用]

Justification: Our paper does not do crowdsourcing experiments and research on human subjects.
理由：我们的论文未进行众包实验或涉及人类受试者的研究。

Guidelines:
指导原则：

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- 回答“不适用”表示论文不涉及众包或人类受试者研究。
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- 将此信息包含在补充材料中是可以的，但如果论文的主要贡献涉及人体受试者，则应尽可能多地在正文中详细说明。
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.
- 根据NeurIPS伦理守则，参与数据收集、整理或其他劳动的工作人员应至少获得数据收集者所在国家的最低工资。

15. Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects

15. 涉及人体受试者研究的机构审查委员会（IRB）批准或同等审批

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?
问题：论文是否描述了研究参与者可能面临的风险，这些风险是否已告知受试者，以及是否获得了机构审查委员会（IRB）批准（或根据您所在国家或机构要求的同等审批/审查）？

Answer: [NA]
回答：[不适用]

Justification: Our paper does not involve crowdsourcing nor research with human subjects. Guidelines:
理由：我们的论文不涉及众包或人体受试者研究。指导原则：

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- 回答“不适用”表示论文不涉及众包或人体受试者研究。
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- 根据研究所在国家的不同，人体受试者研究可能需要IRB批准（或同等审批）。如果您获得了IRB批准，应在论文中明确说明。
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- 我们认识到此类程序在不同机构和地区可能存在显著差异，期望作者遵守NeurIPS伦理守则及其所在机构的相关指南。

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.
- 对于初稿提交，请勿包含任何可能破坏匿名性的内容（如进行审查的机构名称）。