arXiv:2304.00776v2 [cs.LG] 7 Jul 2024

思维链预测控制

Zhiwei Jia [1] **Vineet Thumuluri** [1] **刘方辰** [2] **陈令浩** [3] **黄志豪** [1] **苏昊** [1]

摘要

我们研究面向复杂低层控制（例如，高接触物体操作）的可泛化示范策略学习。我们提出了一种利用次优演示数据的新型分层模仿学习方法。首先，我们提出一种与观测空间无关的方法，以无监督的方式高效发现演示数据的多步子技能分解。通过将时间上接近且功能相似的动作分组成子技能级示范片段，片段边界处的观测构成了任务的一系列规划步骤，我们将其称为思维链（CoT）。接下来，我们提出一种基于 Transformer 的设计，有效学习预测思维链，将其作为子技能级指导。我们通过可学习的提示词元和混合掩码策略，将动作预测与子技能预测相结合，从而能够在测试时动态更新指导，并改善轨迹的特征表示，以实现可泛化的策略学习。我们的方法——思维链预测控制（**CoTPC**）在使用次优演示数据的挑战性操作任务中，持续超越现有强基线。更多详情请参阅我们的项目页面。

1. 引言

分层强化学习（HRL）（Hutsebaut-Buysse 等人，2022年）作为一种有望实现样本高效且具备泛化能力的策略学习方向，在人工智能领域受到广泛关注。分层强化学习通过时间抽象将复杂的序贯决策问题分解为更简单、更小的子问题（即所谓的思维链（Wei 等人，2022年））。此外，许多方法采用两阶段策略，并具备针对高层动作（即子目标或选项）进行规划的能力，以实现泛化能力。另一方面，模仿学习（IL）仍然是训练自主智能体最有效的方法之一。在没有密集标注奖励或策略在线交互的情况下，模仿学习通常将策略学习视为（自）监督学习，并有望利用大规模预收集示范；受近期大语言模型（LLMs）取得成功的启发，这类学习通常采用 Transformer。构建基础决策模型的一大障碍（Yang 等人，2023年）仍在于如何更好地利用次优演示。本文研究如何利用次优示范，在低层控制任务中开展分层模仿学习。

尽管近期取得了进展（Chen 等人，2021年；Florence 等人，2022年；Shafiullah 等人，2022年；Liu 等人，2022年；Ajay 等人，2022年；Chi 等人，2023年），通过模仿学习以可扩展的方式解决高接触物体操作等低层控制任务仍极具挑战性。由于底层接触动力学（Pfrommer 等人，2021年）以及演示数据的生成方式，示范通常本质上是次优的。这些不理想的特性，例如非马尔可夫性、噪声、不连续和随机性，给模仿者的优化与泛化都带来了巨大挑战（详见附录 A）。我们发现，将层级原则（即时间抽象和高层规划）融入基于 Transformer 的设计后，能够更好地利用次优演示数据来解决具有挑战性的任务。为此，我们首先提出一种无监督子技能发现策略，从示范中生成子技能级别的监督信号。随后，我们训练模型动态生成子技能引导，以改进低层动作预测。

具体而言，我们将任务分解为一系列规划步骤的多步子技能分解视为其思维链（Wei 等人，2022年；Yang 等人，2022年）。作为我们贡献的第一部分，我们提出一种与观测空间无关的方法，能够以无监督方式高效发现子技能；子技能定义为演示数据中的关键观测序列。我们提出将时间上接近且功能相似的动作归为一个片段。随后，变点（即片段边界）自然构成子技能序列，也就是代表高层任务完成过程的思维链。作为第二部分，我们提出一种新颖的基于 Transformer 的设计，能够有效地联合学习思维链与低层

[1]加州大学圣迭戈分校 [2]加州大学伯克利分校 [3]浙江大学。通讯作者：贾志伟 <zjia@ucsd.edu>。

《 41^{st} 国际机器学习会议，维也纳，奥地利。PMLR 235，2024年。版权所有：作者，2024年。

动作预测。我们通过在上下文历史的开头添加额外的提示词元，并采用混合掩码策略，实现这种耦合预测机制。因此，子技能引导会在每一步动态更新，并能够学习到更优的轨迹特征表示，最终提升低层动作预测过程的泛化能力。

我们的方法——思维链预测控制（CoTPC）——从优化角度来看，能够利用通常更稳健、方差更小的子技能，从次优演示数据中更快地学习。从泛化角度来看，该方法使用 Transformer（Brown 等人，2020年），通过子技能规划提升泛化能力，而子技能规划则以无监督发现的思维链作为监督信号进行学习。我们在若干具有挑战性的低层控制任务（移动迷宫、Franka-Kitch 和 ManiSkill2）上评估了 CoTPC，并通过消融研究验证了其设计。结果表明，CoTPC 的表现始终优于多个强基线。

2. 相关工作

我们在附录D中补充了相关工作。

从演示中学习（LfD）利用预先收集的演示数据学习交互式智能体，因其有效且易于扩展而广受欢迎。粗略地说，这类方法分为三种：离线强化学习、结合辅助演示的在线强化学习，以及行为克隆（BC）。离线强化学习方法（Kumar 等人，2019年；Fu 等人，2020年；Levine 等人，2020年；Kumar 等人，2020年；Kostrikov 等人，2021年；Chen 等人，2021年；Wang 等人，2022年）通常需要带有密集标注奖励的示范；而利用演示数据增强在线强化学习的方法（Hester 等人，2018年；Kang 等人，2018年；Ross 等人，2011年；Nair 等人，2020年；Rajeswaran 等人，2017年；Ho 和 Ermon，2016年；Pertsch 等人，2021年；Singh 等人，2020年）依赖同策略交互。行为克隆（Pomerleau，1988年）则构建完全监督或自监督学习问题，实用性更强，因此得到广泛应用，尤其是在机器人学领域（Zeng 等人，2021年；Florence 等人，2022年；Qin 等人，2022年；Zhang 等人，2018年；Brohan 等人，2022年；Rahmatizadeh 等人，2018年；Florence 等人，2019年；Zeng 等人，2020年）。然而，行为克隆的一个众所周知的缺点是误差累积（Ross 等人，2011年），通常由演示数据与测试时轨迹之间的分布偏移造成。为解决这一问题，人们提出了多种方法（Ross 和 Bagnell，2010年；Ross 等人，2011年；Sun 等人，2017年；Laskey 等人，2017年；Tennenholtz 等人，2021年；Brantley 等人，2019年；Chang 等人，2021年）。其他问题还包括演示数据的非马尔可夫性（Mandlekar 等人，2021年）、不连续性（Florence 等人，2022年）、随机性和噪声性（Sasaki 和 Yamashina，2020年；Wu 等人，2019年），这些问题会导致神经策略在推理过程中产生严重的累积误差（详细讨论见附录A）。

将从演示中学习视为序列建模。近期离线策略学习的一种趋势是放宽策略的马尔可夫假设，这一做法受到序列模型成功的启发（Graves 和 Graves，2012年；Chung 等人，2014年；Vaswani 等人，2017年）：在这类模型中，相较于算法层面的精巧设计，人们更看重模型的表达能力和容量。在相关研究中，Dasari 和 Gupta（2021年）以及 Mandi 等人（2021年）研究了单次模仿学习；Lynch 等人（2020年）和 Singh 等人（2020年）探索了从演示数据中提取行为先验；Chen 等人（2021年）、Liu 等人（2022年）、Janner 等人（2021年）、Shafiullah 等人（2022年）、Ajay 等人（2022年）以及 Janner 等人（2022年）则考察了策略学习的不同建模策略。特别是，基于 Transformer 的方法（Vaswani 等人，2017年；Brown 等人，2020年）因其简单有效而极受欢迎。

序列建模与强化学习中的分层方法 思维链（Wei 等人，2022年）指的是一种通用策略：将多步问题分解为一系列中间步骤来解决。近年来，这种方法已广泛应用于各种问题，例如数学推理（Ling 等人，2017年；Cobbe 等人，2021年）、程序执行（Reed 和 De Freitas，2015年；Nye 等人，2021年）、常识推理或一般推理（Rajani 等人，2019年；Clark 等人，2020年；Liang 等人，2021年；Wei 等人，2022年）以及机器人学（Xu 等人，2018年；Zhang 和 Chai，2021年；Jia 等人，2022b；Gu 等人，2022年；Yang 等，2022；Shridhar 等人，2023年；James 和 Davison，2022年）。在分层强化学习的背景下，类似思想可追溯到封建强化学习（Dayan 和 Hinton，1992年）和选项框架（Sutton 等人，1999年）。受这些方法启发，我们关注低层控制任务中的模仿学习设置。请注意，尽管过程克隆（Yang 等，2022）与本文名称相似，但它存在一些局限，因此适用范围小得多（见附录D）。

机器人任务演示 在实践中，对于机器人任务而言，演示数据的最优性假设通常并不成立。涉及低层动作的演示数据主要有三种形式：通过遥操作采集的人类演示（Kumar 和 Todorov，2015年；Vuong 等人，2021年）、由强化学习智能体生成的专家演示（Mu 等人，2020年；2021年；Chen 等人，2022年；Jia 等人，2022a），以及由规划器（例如启发式方法、采样、搜索）找到的演示数据（Gu 等，2023；Qureshi 等人，2019年；Fishman 等人，2022年）。这些演示数据通常并非最优，原因可能是人为偏差、强化学习智能体不够完善，或规划器本身的特性。本文主要使用 ManiSkill2（Gu 等，2023）中由规划器生成的演示数据。对于模仿学习而言，该基准测试目前尚未达到性能饱和，同时也具有足够的挑战性（详情见第5.3节）。

3. 预备知识

马尔可夫决策过程（MDP）形式化 通过以下方式对序列决策问题进行形式化是最常见的方法之一：

马尔可夫决策过程，或称 MDP（Howard，1960年），定义为一个六元组 $\langle S$、A、$\mathcal{T}$、$\mathcal{R}$、ρ_0、$\gamma\rangle$，其中包括状态空间S、动作空间 A、马尔可夫转移概率$\mathcal{T}$：$S \times A \to \Delta(S)$、奖励函数 $\mathcal{R}$： $S \times A \to \mathbb{R}$、初始状态分布 ρ_0以及折扣因子$\gamma \in [0,1]$。智能体根据策略 π：$S \to \Delta(A)$ 与由 $\mathcal{T}$ 和 $\mathcal{R}$ 表征的环境进行交互。我们将轨迹表示为 τ_π，即通过遵循策略 π采取动作而形成的 $(s(0)\ a(0)\ s(1)\ a(1)\ ...,\ s(t)\ a(t))$ 序列。在每个时间步，智能体都会收到奖励信号 $r_t \sim \mathcal{R}(s(t)\ a(t))$。目标是找到能够最大化期望回报$\mathbb{E}_{\tau\sim\pi}[\sum_t \gamma^t r_t]$的最优策略 π^*。值得注意的是，在机器人任务和许多现实世界的应用中，奖励充其量只是稀疏地给出（例如，二元成功信号），或者仅在轨迹结束后给出（非马尔可夫性）。

行为克隆 模仿学习（IL）中最直接的方法是行为克隆（BC）。该方法假设可以访问由专家策略生成的预先收集的演示数据 $D = \{(s(t)\ a(t))\}_{t=1}^N$，并通过最小化行为克隆损失$\mathbb{E}_{(s,a)\sim D}[-\log\ \pi(a|s)]$，在直接监督信号下学习最优策略。由于测试时$\tau_\pi$ 会超出演示数据的分布，因此学习到的策略需要能够泛化到演示数据中未见的状态。这一挑战被称为分布偏移（Ross 和 Bagnell，2010年）。近年来，一些方法，尤其是基于 Transformer 的方法，通过考虑大小为 T 的上下文历史，放宽了马尔可夫假设，即将策略表示为 $\pi(a(t)|s(t)\ ...,\ s(t-T+1))$ 或$\pi(a(t)|s(t)\ a(t)\ ...,\ s(t-T+1)\ a(t-T+1))$。实证结果表明，这一改变具有优势。

行为Transformer 行为Transformer（BeT）（Shafiullah 等人，2022年）旨在解决行为克隆中的含噪（多模态）示范问题。它采用了与决策Transformer（DT）（Chen 等人，2021年）相似的架构，同时借鉴目标检测领域的思路，使用中心加偏移表示来表征动作信号。具体而言，首先对演示数据中的所有动作进行（k 均值）聚类，将动作空间划分为 K 个区域。模型采用一个分类器和 K 个独立的回归器（每个区域对应一个），以预测各区域中的动作。这种方法大幅提高了行为克隆的优化效率。在推理时，BeT 从预测出的区域及其对应的偏移回归结果中采样。为简化起见，我们始终选择预测得分最高的区域（性能相当）。决策Transformer根据交织的动作和状态历史信息来生成动作预测，而BeT则仅依据状态。 我们在*BeT*变体中采用了决策*Transformer*的方法（既用于*BeT* 基线，也用于*CoTPC*），并发现其表现更好，因为进一步以动作为条件能更有效地区分含噪演示数据。

4. 子技能序列即思维链

我们提出了两项内容：（1）一种与观测空间无关的策略，能够以无监督方式从示范中高效发现子技能序列，并将其作为思维链监督信号；（2）一种新颖的基于 Transformer 的设计，能够在模仿学习中有效采用层级原则，即子技能规划。我们将这一方法命名为思维链预测控制（**CoTPC**）。

4.1. 从演示数据中无监督发现子技能

我们发现，许多低层控制任务天然由一系列子技能构成。在一条成功轨迹中，存在一些关键观测，每个关键观测都标志着一个子技能的完成。例如，在图 3 所示的移动迷宫中，两座桥梁自然地将任务划分为三个子技能。我们将任务的这种多步子技能分解（即将任务分解为一连串规划步骤）称为其思维链（CoT），因为这条链被视为任务的“思考过程”。思维链提供连贯而简洁的行为指导，这是分层策略学习的典型优势。形式上，对于每条轨迹$\tau \in D$，我们将其思维链定义为其观测 $F_{cot}(\tau) \triangleq \{s^{\mathbf{cot}}(i)\}_{i=1}^m$构成的子序列，其中每个观测都标志着轨迹中相邻两个子技能片段的边界。这里，i 表示 m 子技能的索引。

我们如何定义子技能片段，又如何找到它们？我们的基本假设是，同一子技能中的动作在时间上相近，且在功能上相似。因此，我们提出采用一种无需训练的过程来将连续的动作分组成片，通过该过程找出被建模为时间序列的动作序列中的变点。具体而言，对于每条轨迹 τ，我们将其在时间 t 的p维动作视为向量 $a(t) \in \mathcal{R}^p$。子技能发现过程被表述为一个优化问题，首先找到动作序列在 τ 中的变点索引 $\{\tau_i\}_{i=1}^m$，并满足

$$\tau_1,\tau_2,...,\tau_m = \arg\min_{c_1...c_m} \mathcal{J}(\tau,\{c_i\})$$

$$\mathcal{J}(\tau,\{c_i\}) = \sum_{i=1}^{m}[\mathcal{C}(a(c_{i-1}+1),\ldots,a(c_i)) + \beta]$$

$$\mathcal{C}(v_1,\ldots,v_n) = \frac{1}{n}\sum_{i=1}^{n} d_{cos}(v_i,\bar{v}),\ \ \bar{v} = \frac{1}{n}\sum_{i=1}^{n} v_i$$

为方便起见，我们定义 $c_0 = -1$ 为 τ起始于 $t = 0$，并将最后一个变点定义为轨迹（以及最后一个子技能）结束时的最后一个时间索引 $|\tau| - 1$。我们使用 d_{cos}余弦距离（相关讨论见附录 I），而 β是一个用于惩罚子技能数量过多的超参数 m。例如，当 $\beta = 0$时，每个时间步都会成为一个变点，从而导致演示轨迹被过度分段。这个优化问题通常通过逐条轨迹进行动态规划来求解

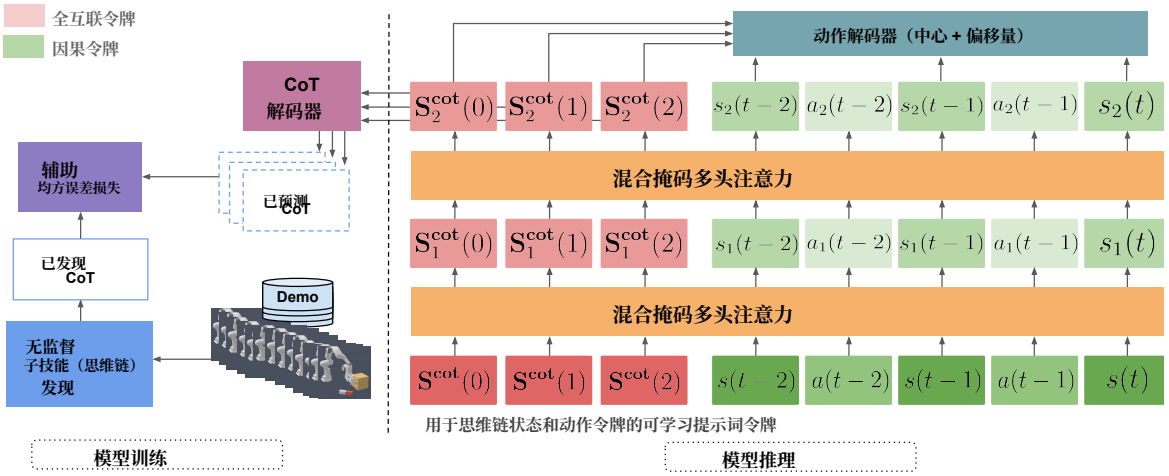


图 *1.* 训练期间，CoTPC 学习联合预测（1）从每个思维链令牌预测下一个子技能和最后一个子技能，以及（2）从每个思维链令牌（用于动作偏移量）和状态令牌（用于动作中心）预测低层动作。详见第 4.2.2 节。推理期间，不使用思维链解码器时，低层动作将在动态更新的思维链特征的引导下进行预测。思维链令牌之间采用全连接注意力（可关注任意令牌）。状态令牌和动作令牌采用因果注意力（只能关注先前的令牌和思维链令牌）。为便于展示，此处仅显示 2 个注意力层和 3 个时间步。

。这里我们采用 PELT（Killick 等人，2012年）这一改进方法，其中包含剪枝等额外技术。我们在图 2 中展示同一条轨迹中各动作之间相似度图的可视化，以说明我们用于子技能发现的直观动作分组策略。最后，我们使用观测值 $\{s^{\text{cot}}(i) \triangleq s(\tau_i)\}_{i=1}^{m}$ 作为发现出的思维链，而非动作。

任务中的子技能数量并非 先验已知，因为即使是同一任务的示范，也可能因起始配置和执行难度而有所不同。例如，在推椅子任务中，由于初始条件各异且任务本身的性质，演示数据中可能缺少某些子技能。为实现这种灵活性，我们的方法仅将 m 设为推理期间要预测的子技能数量上限，并假设同一任务中的所有演示轨迹共享相同的子技能“超序列”。在实验中，我们发现 β 相对稳健（详见附录 J ）。

我们发现，基于对所发现变点处关键观测的定性评估，我们的方法能够为具有不同动作空间的一系列多样化任务发现有意义的子技能。例如，插销插入任务的执行包括伸手、抓取、使插销与孔对齐、微调以及平稳地插入插销。附录中展示了一些子技能发现结果：G。除了无需监督之外，我们方法的另一项主要优势是与观测空间无关，发现结果不受具体传感器配置的影响（例如摄像机角度，这些角度通常会针对具体任务进行临时调整）。

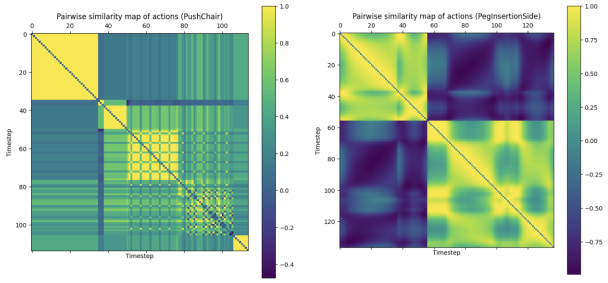


图 *2.* 推椅子（**左侧**）和插销插入（**右侧**）两条轨迹中不同时间步的动作成对相似度。动作空间为关节速度增量和关节位姿增量。对角线上的可视化识别区块被分组，其中的动作在短时间内相近且功能相似。这与人类对各子技能的直觉高度吻合（参见附录 G）。

4.2. 子技能引导的动作建模

本节介绍我们基于 Transformer 的设计，该设计利用子技能发现过程获得的监督信息，学习对思维链和动作进行建模。

4.2.1. 带混合掩码的可学习思维链提示词元

我们的架构以近期提出的行为Transformer（BeT）（Shafiullah 等人，2022年）的改进版本为基础。该方法通过离散中心加连续偏移预测策略，使 GPT（Brown 等人，2020年）能够建模多样且带噪声的动作序列。BeT 架构简介请参阅第 3 节。为预测思维链，我们提出添加一组可学习提示词令牌（Zhou 等人，2022年），称为思维链令牌，并将其置于状态和动作上下文历史的开头。

我们训练这些令牌从序列中提取特征，并通过思维链解码器预测思维链（类似于检测 Transformer（Carion 等人，2020年）中的目标查询令牌）。我们设计了一种混合掩码机制：在推理期间，思维链令牌彼此之间可以相互关注，并可查看上下文历史中的所有动作和状态令牌；状态令牌或动作令牌则关注过去的令牌（标准因果掩码），其中也包括思维链令牌。这样，动作解码便能由与思维链令牌对应的已提取思维链特征引导。

具体而言，给定上下文长度 T，我们选择使用 T 个 CoT 标记$\mathbf{S^{cot}}(i)$，其中每个标记都是不同的可学习提示词标记 $\in \mathcal{R}^{128}$；Transformer 的嵌入维度为 128。因此，截至时间 t 的演示片段由 T 个思维链令牌、 T 个状态令牌和 $T-1$ 个动作令牌组成：

$$\tau_T(t) = \mathbf{S^{cot}}(0),\ ...,\ \mathbf{S^{cot}}(T-1),\ s(t-(T-1)),\ a(t-(T-1)),\ ...,\ s(t-1),\ a(t-1),\ s(t) \quad (1)$$

通过应用混合掩码多头注意力 $\mathcal{M}[\cdot]$，总共 J 个注意力层的特征表示为

$$h_j(\tau_T(t)) = \mathcal{M}[F_{enc}(\tau_T(t))],\ j=1$$
$$h_j(\tau_T(t)) = \mathcal{M}[h_{j-1}(\tau_T(t))],\ j>1$$

其中， F_{enc} 分别通过编码器 $f_a(\cdot)$ 和 $f_s(\cdot)$ 对每个动作令牌和状态令牌进行编码（CoT 令牌不使用编码器）。此处省略位置嵌入，以及标准 Transformer 中注意力层之间的额外操作。因此，最后一个注意力层的特征可写为

$$h_J(\tau_T(t)) = \mathbf{S}_J^{\mathbf{cot}}(0),\ ...,\ \mathbf{S}_J^{\mathbf{cot}}(T-1),\ s_J(t-(T-1)),\ a_J(t-(T-1)),\ ...,\ s_J(t-1),\ a_J(t-1),\ s_J(t)$$

随后，将 CoT 令牌对应的思维链特征输入思维链解码器 g_{cot}，以预测 $\{g_{cot}(\mathbf{S}_J^{\mathbf{cot}}(t))\}_{t=0}^{T-1}$（稍后将说明为何使用 T 个 CoT 令牌和 T 个输出）。在我们的实验中，我们使用一个具有 ReLU 激活函数的两层多层感知机来实现 g_{cot}。

尽管 Transformer 中的可学习提示词令牌已广泛用于语言、视觉与语言等相关任务，但将其应用于低层连续控制问题时仍缺乏充分探索；正如我们的消融研究所示，这极具挑战性。

4.2.2. 使用共享思维链令牌的耦合动作与子技能预测

受 BeT 启发，我们采用一对中心和偏移量解码器来预测动作，分别记为 $g_a^{\mathbf{ctr}}(\cdot)$ 和 $g_a^{\mathbf{off}}(\cdot)$。对于动作 $a(t')$，我们使用 $g_a^{\mathbf{ctr}}(\cdot)$ 来预测其中心，表示为

$$g_a^{\mathbf{ctr}}(s_J(t')),\ t' \in \{t-(T-1), ..., t-1, t\}$$

其中 t遵循 $h_J(\tau_T(t))$ 中的前述记法。与原始版 BeT 不同，我们使用以下方式预测其偏移量

$$g_a^{\mathbf{off}}(\mathbf{S}_J^{\mathbf{cot}}(t'')),\ t'' = t' - [t-(T-1)]$$

我们实质上使用 i^{th} CoT 标记的特征来预测 i^{th} 动作的偏移分量。因此，我们需要 T个 CoT 标记，以便利用每个 CoT 标记的特征来预测其中一个 T不同时间步的动作偏移量。整体架构见图 1 。思维链解码器$g_{cot}(\cdot)$使用与动作偏移解码器相同的输入，从而将动作预测与思维链预测耦合起来，以增强子技能引导。另一种设计是仅使用一个 CoT 标记来生成思维链预测，如 $g_{cot}(\mathbf{S^{cot}})$所示。动作偏移量预测则基于状态令牌（类似于原始版 BeT）。然而，我们发现，这种解耦策略生成的策略泛化能力要差得多。一种耦合程度更高的预测方法是让中心解码器$g_a^{\mathbf{ctr}}(\cdot)$ 也与 $g_{cot}(\cdot)$共享输入；但由于 CoT 标记直接负责所有预测，这会导致优化不稳定性。有关实证证据，请参阅我们的消融研究。

对于子技能预测，我们通过训练 $g_{cot}(\cdot)$ 从每个 CoT标记的特征中同时解码下一个子技能和最后一个子技能，构建了一种自回归预测策略。对于轨迹 τ 中时间 t 的情况，下一个子技能记为 $s_{next}^{\mathbf{cot}}(t) = s(\tau_i)$，即满足 $\tau_i > t$ 的首个 τ_i 所对应的子技能。最后一个子技能始终是演示轨迹中任务完成时的终点，记为 $s_{last}^{\mathbf{cot}}(t) = s(|\tau|-1)$。形式化地，假设 q 维状态或观测空间为 $\mathcal{R}^q$，对于每个满足 $t = \{0, 1, ..., T-1\}$ 的 CoT标记 $\mathbf{S^{cot}}(t)$，思维链解码器生成 $g_{cot}(\mathbf{S}_J^{\mathbf{cot}}(t)) \in \mathcal{R}^{2q}$。 TCoT预测使每个 CoT标记提取有助于子技能规划的有用特征，从而改善所有时间步上的动作偏移预测。我们通过实验证实，同时进行即时规划和长期规划的这一策略优于另外两种替代方案。

由于思维链预测具有自回归特性，我们的方法允许同一任务的不同演示轨迹包含灵活数量的子技能。此外，由于思维链特征会在每个时间步动态更新，CoTPC 能够处理涉及动态控制的任务（例如，移动迷宫和推椅子）。我们以思维链为方法命名，因为我们的子技能预测与大语言模型中的思维链相呼应：文本输出与推理链共同生成。请注意，在推理阶段不使用思维链解码器，仅使用思维链特征（用于预测动作偏移量）。有关这一主题的进一步讨论，请参见第6 节。另请参见附录5 中的图C ，其中比较了不同版本中动作解码器（偏移量和中心）及思维链令牌的数据流。

4.2.3. 模型训练

整体流程如图 1 所示。模型通过行为克隆损失以及基于均方误差（MSE）的辅助CoT预测损失进行训练，$\mathcal{L}_{cot}$ 该损失由系数 $\lambda = 0.1$ 加权，从而得到总体目标：

$$\mathcal{L}_{total} = \mathop{\mathbb{E}}_{(s_t,a_t)\in D} \mathcal{L}_{bc}(a_t, \hat{a}_t) + \mathop{\mathbb{E}}_{\tau_T(t)\sim D} \mathop{\mathbb{E}}_{t'} \mathcal{L}_{cot}([s^{\textbf{cot}}_{next}(t'), s^{\textbf{cot}}_{last}(t')], g_{cot}(\mathbf{S}^{\textbf{cot}}_J(t'')))$$

其中 $\hat{a}_t$ 是通过 $g^{\textbf{ctr}}_a(\cdot)$ 预测的动作， $g^{\textbf{off}}_a(\cdot)$。遵循公式 1 和上一节中的记号，$\tau_T(t)$ 是从演示集 D中随机采样的、上下文长度为 T 的片段， $t'' = t' - [t - (T-1)]$ 是CoT标记的移位索引。由于共有 T 个CoT标记，因此有 T 个损失项。

在训练过程中，我们对动作令牌和状态令牌应用随机注意力掩码，使思维链令牌关注长度各异的上下文。对于小批次中的每个数据点，我们均匀采样 $l \in \{0, 1, ..., T-1\}$，并手动将所有思维链令牌指向与以下内容对应的状态令牌和动作令牌的注意力掩码设为 $s(t')$ 和 $a(t')$ 零（即禁用），条件是 $t' > t - l$。这种正则化技术的直觉是：对于同一段历史信息的不同长度片段，我们应该预测出相似的未来子技能。此外，由于思维链令牌之间是全连接的，如果不采用这种技术，它们在训练时可能会关注未来令牌，从而在一定程度上使子技能预测变得轻而易举。相关消融研究见第 5.4 节。

5. 实验

本节介绍主要实验结果及消融研究。现有基准测试大多已在模仿学习方面趋于饱和（DMControl（Tassa 等人，2018年）、D4RL（Fu 等人，2020年）等），或缺少演示数据（例如 Mine- Dojo（Fan 等人，2022年）），因此我们选择了一系列介于两者之间的多样化任务。我们首先使用二维连续空间移动迷宫和 Franka-Kitchen 的一个变体（Gupta 等人，2019年）检验我们的方法。随后，我们使用 ManiSkill2（Gu 等人，2023年）中的 5 项物体操作任务（难度从较易到极具挑战性）进行广泛比较。这些任务涵盖多个类别（导航、静态／移动操作、软体操作等）、多种动作空间（关节位姿增量、速度增量等），演示数据来源各异（人类、启发式方法等），并各具独特挑战（详情见第5.3 节和附录 B ）。

5.1. 移动迷宫

我们展示了一个二维迷宫，其动作空间为位移 $(\delta x, \delta y)$。如图 3 （左）所示，在这个 S形迷宫中，智能体从右上方绿色方块区域内随机初始化的位置出发。目标是到达左下方的绿色区域。每次环境重置时，这两个区域和两座绿色矩形桥的位置都会随机变化。除顶部方块外，其余方块均以随机恒定速度（彼此独立地）来回移动。一旦智能体落在移动方块上，该方块就会静止。智能体不能越过迷宫边界（但越界不会导致其死亡）。为简单起见，我们采用10维状态观测，其中包含智能体当前的位置以及四个绿色区域的位置。此任务需要动态控制/规划。

演示 为使策略能够从示范中学习，我们采用多种启发式方法以及一种借助测试时不可用的事后信息的 RRT 风格规划器（遵循近期关于利用机器生成演示数据的研究（Gu 等人，2023年）。据此，我们整理了演示轨迹（每条轨迹对应不同的随机环境配置）。

训练与评估 对于此任务，我们将 CoTPC 与原始行为克隆、行为Transformer（BeT）（Shafi- ullah 等人，2022年）、决策Transformer（DT）（Chen 等人，2021年）进行比较。DT 最初是为使用稠密奖励示范数据的离线强化学习提出的。我们将其调整用于行为克隆设置，方法是忽略奖励令牌。我们为 BeT 添加动作令牌（与 DT 类似），并基于 BeT 构建 CoTPC。为确保比较公平，我们使用共享的 Transformer 配置实现 CoTPC、DT 和 BeT。我们使用 400 条来自不同环境配置的演示轨迹训练所有方法，并在 100 个未见过的环境配置上进行评估（结果见表 1）。

5.2. Franka Kitchen

此前已有研究探讨了 Franka Kitchen环境的不同变体和任务设置（Gupta 等人，2019年；Sasaki 和 Yamashina，2020年；Fu 等人，2020年）。我们提出一种设置：要求智能体按照目标指定的顺序完成 4 项物体操作（共 7 个不同选项）。我们采用严格的判定标准，即只有完成全部 4 个子任务，任务才算成功。必须按要求的顺序完成子任务才能成功。若执行了指定的 4 个子任务之外的其他子任务，环境将终止。动作空间基于机器人的关节速度（8 维）。我们使用原始状态观测，并附加经过修改的目标嵌入。

示范 我们重放了（Gupta 等人，2019年）中最初提出的人类演示的一部分。具体而言，我们使用了 50 条长度为150 至300步的演示轨迹，并为每条轨迹标注所执行的子任务及其顺序。因此，许多有序子任务组合最多只对应一条演示轨迹。更多详情请参见附录。

训练与评估 我们使用与移动迷宫相同的一组基线。我们使用 90 个未见过的环境配置进行评估，这些配置在初始场景配置上有所不同（所有有序



图 *3.* 移动迷宫（**左侧**）、Franka 厨房（**中间**）以及 ManiSkill2 中采样的若干任务（**右侧**），即转动水龙头、插销插入和推椅子。详细说明分别见第 5.1、5.2 和 5.3节。



图 *4.* 推椅子、转动水龙头和插销插入任务的几何变化采样。不同环境配置中的插销和盒子尺寸以及孔的相对位置各不相同。

表 *1.* 移动迷宫和（Franka Kitchen 的一个变体）上的测试性能。成功率（%）表示任务成功率（对于 Franka Kitchen，表示完成全部 4 个子任务）。# s-tasks 表示每次轨迹展开中完成的子任务平均数量。最佳结果以**粗体**标出。

	原始行为克隆	DT	BeT	CoTPC（本文方法）
迷宫（成功率）	9.0	23.0	33.0	**44.0**
厨房（任务数/成功率）	1.7/6.7	1.6/6.7	1.8/14.4	**2.1/25.6**

子任务组合都曾在演示数据中出现过）。由于人类演示数据有限，且有序子任务序列种类繁多，这项任务需要具备泛化能力的模仿学习。另请参见表 1。

5.3. ManiSkill2

ManiSkill2 提供了多种低层物体操作任务，任务所在环境采用逼真的物理仿真（例如，完全动态的抓取动作）。我们选择了 5 项任务（见图 3）：堆叠方块，即拿起一个方块，将其放在另一个方块上，并让夹爪离开堆叠；转动水龙头，即打开不同的水龙头；插销插入，即将一个长方体插销 横向 插入不同几何形状和尺寸的盒子上的孔中；推椅子，即通过移动机器人将不同的椅子模型推至指定目标位置；以及倾倒，即将液体从瓶子倒入目标烧杯，使液面达到指定高度。推椅子采用关节速度增量控制（19 维，双臂加移动底座）；倾倒采用末端执行器位姿增量控制（8 维）；其余任务采用关节位姿增量控制（8 维）。我们使用状态观测和点云观测开展实验。

任务复杂度 这些任务的挑战来自多个方面。首先，所有任务中的物体位姿在环境重置时都会完全随机化（位移约为 0.3 米，且 360° 旋转）。这与 Franka Kitchen 等环境不同。其次，转动水龙头、插销插入和推椅子任务中的目标物体，其几何形状和尺寸变化很大（见图 4）。此外，操作过程中主要是推动而非抓取水龙头（欠驱动控制）；孔的间隙仅为 3 毫米（需要高精度控制），且至少需要将插销的一半横向推入孔中（比其他基准测试中的类似任务更难（Xu 等人，2019年））；椅子模型完全关节化，包含大量关节；倾倒任务则要求动作平稳且不洒出液体。此外，ManiSkill2 采用阻抗控制器，可实现比基于位置的控制器更平滑的运动轨迹（这对倾倒等任务很重要），但低层动作建模更困难（例如，执行器可能存在明显滞后）。

示范任务的复杂性还体现在演示数据并非最优（例如，原始行为克隆在全部 5 项任务上都表现不佳）。这些演示数据由多阶段运动规划和基于启发式规则的策略生成（借助模拟器中的特权信息）。我们为堆叠方块和转动水龙头任务使用了 500 条演示轨迹（分布在 10 个水龙头上），为插销插入和推椅子任务使用了 1000 条演示数据（分布在 5 把椅子上），并为倾倒任务使用了 150 条演示数据。

训练与评估 除了原始行为克隆、决策变换器和 BeT，我们还加入了 Decision Diffuser (DD) (Ajay 等人，2022年) 和扩散策略 (DP) (Chi 等人，2023年)，作为两种基线方法，它们探索将扩散模型 (Ho 等人，2020年) 用于策略学习：一种方法先生成仅包含状态的轨迹，再利用已训练的逆动力学模型预测动作；另一种方法则直接执行扩散步骤来预测动作。值得注意的是，基于扩散的策略依赖反复执行去噪步骤，通常比 BeT 或 CoTPC 慢得多，因此难以实现动态控制。*ManiSkill2* 中的任务具有多样的物体级变化，能够帮助深入了解两方面的问题：

表 2. 使用状态观测时，ManiSkill2 任务的测试性能（未见任务和零样本成功率）。最佳结果以**粗体**显示。基于扩散的方法通常速度较慢，数据效率也可能较低（参见此处的讨论），这使动态控制更具挑战性。附录 L中列出了 3 次运行结果的均值和标准差。

	堆叠 CUBE		PEG 插入		TURN 水龙头		PUSH 椅子	
	未见	零样本	未见	零样本	未见	零样本		
原始行为克隆	1.0	0.0	0.0	0.0	0.0	0.0		
决策Transformer	19.0	17.5	40.0	27.0	25.6	17.0		
Decision Diffuser	26.0	12.6	17.0	5.0	**56.0**	20.0		
扩散策略	84.5	57.6	**52.0**	35.0	54.0	38.0		
行为Transformer	73.0	42.5	49.6	32.5	44.0	33.4		
CoTPC（本文方法）	**86.0**	**59.3**	50.0	**39.3**	51.2	**41.0**		

表 3. 在未见设置和零样本设置下，针对点云观测的 ManiSkill2 任务的测试性能（成功率）。倾倒任务仅支持视觉观测。最佳结果以**粗体**标出。此处仅展示密切相关的 BeT 基线。

	CUBE	PEG	POUR	水龙头		椅子	
	未见	零样本	未见	未见	零样本	未见	零样本
行为 Transformer	70.0	35.0	24.0	50.0	20.0	26.0	13.4
CoTPC（本文方法）	**81.0**	**44.0**	**32.0**	**58.0**	**27.5**	**32.0**	**16.7**

有效的模仿者能够学习潜在行为及其泛化能力。我们使用 5 个任务进行评估，分别采用未见设置（物体见过，但场景配置未见）和零样本设置（物体几何形状未见）。具体而言，除插销插入外，所有任务均采用未见设置；转动水龙头、推椅子和插销插入采用零样本设置。我们使用任务成功率（成功率）作为度量。结果见表 2 （状态观测）和表 3（点云观测），结果表明 CoTPC 具有明显优势。为避免针对每个基线调整点云编码器（我们使用从头训练的轻量级 PointNet（Qi 等人，2017年）），我们在点云观测实验中仅使用 BeT 基线。对于 ManiSkill2 上所有方法的主要结果，我们报告 3 次训练运行中最后 20 个检查点的最佳表现；（Chi 等人，2023年）采用了这一评估方案。附录L中提供了 3 次运行的额外均值和标准差结果。

5.4. 消融研究

我们针对 ManiSkill2 中的两个任务（插销插入和推椅子）开展了基于状态观测的消融研究。结果汇总于表 4 ，本节其余部分将介绍具体细节。

对子技能和动作进行解耦预测 在这种称为解耦的变体中，我们训练另一个 Transformer（称为 CoT Transformer），以相同的状态和动作序列作为输入来预测思维链（下一个和最后一个子技能），而不是依靠一个 Transformer 同时预测动作和子技能。在推理期间，CoT Transformer 预测的子技能会输入原始 Transformer，以预测动作。这种解耦的子技能和动作预测策略不如耦合策略，因为后者不仅能学会利用预测的子技能作为指导，还能通过共享这些预测任务的特征，促进对轨迹更好的特征表示。在早期研究中，我们发现，在不同的动作空间、模型规模和上下文长度下，耦合策略的效果都更好。

从思维链令牌中预测哪些子技能？ 在名为仅最后一步的变体中，我们要求思维链解码器 $g_{cot}(\cdot)$仅预测最后一个子技能。在名为仅下一步的变体中，我们要求它仅预测下一个子技能。我们发现，同时预测两者的效果最好，因为这能为模型提供即时和长期规划。在名为随机的变体中，我们要求它将未来随机选取的一个观测预测为随机子技能（并非来自我们的子技能发现过程）。这种方法的结果最差，因为动作引导帮助不大，甚至可能产生误导。

用于子技能和动作预测的共享思维链令牌一种联合预测子技能和动作的原始设计，是只使用一个 CoT 标记，将中心解码器和偏移解码器置于状态令牌之上，并将思维链解码器置于 CoT 标记之上，称为原始版。我们进一步强制

表 4. 消融研究结果（推椅子的未见场景成功率和插销插入的零样本成功率）。

	解耦	仅最后一步	仅下一步	随机	原始版	O-SHARED	交换版	BET+PROMPT	-随机掩码	COTPC
PEG	47.0	52.0	49.0	41.0	45.0	39.0	46.0	43.0	48.5	59.3
椅子	36.0	36.0	37.0	31.0	35.0	29.0	32.0	32.7	34.2	41.0

中心解码器和偏移解码器都以思维链令牌为输入，称为 `o-shared`。此变体让所有解码器过度共享思维链令牌，因此会遭遇优化不稳定性。在最后一种变体中，我们称之为 交换版，让偏移解码器以动作令牌为输入，中心解码器则以思维链令牌为输入。我们发现，这种替代设置的表现较差。我们在附录C中展示了这些变体的数据流。

性能提升是否来自额外的网络容量？ 在标记为 BeT+prompt 的变体中，我们向 BeT 基线（没有辅助损失，也没有思维链解码器）添加了额外的提示标记。该变体与 CoTPC 具有相同的容量，但其表现与 BeT 基线非常相似，这表明我们方法所实现的性能提升并不完全来自额外的网络容量。

随机注意力掩码有多大帮助？ 我们提出在训练期间随机屏蔽思维链令牌与动作/状态令牌之间的注意力。该技术非常有帮助，因为移除它后性能会显著下降（参见 -随机掩码 变体）。该技术无法应用于 BeT 基线（其没有 CoT 标记，而且 BeT 中的所有其他令牌在涉及任何未来令牌时，其注意力掩码均已手动设为 0）。请注意，BeT+prompt 变体采用了该技术，但训练时没有使用辅助思维链损失。其较差的结果表明，仅凭这种正则化技术也无法为 BeT 带来有意义的性能提升。

5.5. 附加实验

我们进行了两项实验，探索基于状态观测的 CoTPC 在仿真到现实迁移方面的可能性。我们在附录 K 中给出了初步的积极结果。

6. 讨论与局限性

与观测空间无关的子技能发现 我们提出的子技能发现方法只依赖动作（并且适用于不同的动作空间），因此能够提供良好的功能抽象。相反，利用额外的观测信息有明显的缺点。首先，视觉观测的维度比动作更高，也更具挑战性。其次，使用视觉信息取决于具体的传感器配置（摄像机角度等），因此结果的鲁棒性较差。另一种方法是使用从某些预训练编码器中提取的视觉特征，但这仍是一个尚未解决的问题（Hansen 等人，2022年）。另一方面，即便是在模拟器中的任务，人类设计的状态观测通常也无法获取（例如，Pour 仅支持基于软体的视觉观测）。

局限性 CoTPC 的一个局限在于，它假设相似且相邻的动作应归入同一子技能。在包含大量微调的复杂操作中（例如站在绳子上保持平衡），这一假设可能不成立。在这种情况下，可能需要额外的力反馈，才能区分那些仅有细微差异的子技能。另一个局限是，我们的方法完全离线，无法避免次优演示带来的影响。我们认为，额外的在线微调可以增强其鲁棒性。

低层任务与高层任务 许多研究“长时程”机器人任务的现有工作（SayCan（Ahn 等人，2022年）、ALFRED（Shridhar 等人，2020年）等）都假设低层控制问题已经解决，或任务层级已给定。与此相反，本文研究如何利用无监督发现的层级信息作为监督信号，更好地学习解决低层控制任务。我们认为，CoTPC 可以扩展至多任务学习设置；在拥有足够演示数据的情况下，还可用作低层控制的基础模型。

CoTPC 是否真正具有层级性？ 虽然 CoTPC 并非严格意义上的基于模型的规划方法，而且由于推理过程中不使用思维链解码器，因此也不能算是“100%”具有层级性，但它隐式地利用了会动态更新的层级信息（思维链特征）。思维链解码器是一个轻量级网络，而思维链特征与预测的子技能并无显著差异。我们发现，使用表达能力更强的思维链解码器会降低思维链特征对预测子技能的能力，从而损害动作解码。

7. 结论

在这项工作中，我们提出了 CoTPC，这是一种层级模仿学习算法，旨在从规模可扩展但次优的演示数据中学习具有泛化能力的策略。我们提出了一种基于新型 Transformer 架构的设计，在离线策略学习中以思维链（即子技能）指导的形式，将层级原则形式化，并提供了一种从示范中以无监督方式获取子技能监督信号的有效方法。我们证明，CoTPC 能够解决广泛且具有挑战性的低层控制任务，并始终优于许多现有方法。

影响声明

本文旨在推动基于学习的控制策略领域的发展。我们的工作可能会带来许多潜在的社会影响，但我们认为没有哪一项需要在此特别指出。

参考文献

Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., 等。《照我所能做，而非照我所说：以机器人可供性为基础的语言》。*arXiv* 预印本 *arXiv:2204.01691*，2022年。

Ajay, A., Du, Y., Gupta, A., Tenenbaum, J., Jaakkola, T. 和 Agrawal, P.《条件生成模型是否就是决策所需的一切？》*arXiv* 预印本 *arXiv:2211.15657*，2022年。

Brantley, K., Sun, W. 和 Henaff, M.《带分歧正则化的模仿学习》。发表于国际学习表征会议，2019年。

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., 等。Rt-1：用于大规模现实世界控制的机器人变换器。*arXiv* 预印本 *arXiv:2212.06817*，2022年。

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shya m, P., Sastry, G., Askell, A., 等。语言模型是少样本学习者。神经信息处理系统进展，33：1877–1901，2020年。

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A. 和 Zagoruyko, S. 使用 Transformer 进行端到端目标检测。载于欧洲计算机视觉会议，第 213–229 页。施普林格，2020年。

Chang, J., Uehara, M., Sreenivas, D., Kidambi, R., 和 Sun, W. 通过具有部分覆盖的离线数据缓解模仿学习中的协变量偏移。神经信息处理系统进展，34:965–979，2021年。

Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., 和 Mordatch, I. 决策 Transformer：通过序列建模实现强化学习。神经信息处理系统进展，34:15084–15097，2021年。

Chen, T., Xu, J., 和 Agrawal, P. 一种用于手内物体重新定向的通用系统。发表于机器人学习会议，第 297–307 页。PMLR，2022年。

Chen, X. and He, K. 探索简单的孪生表示学习。载于 *IEEE/CVF*计算机视觉与模式识别会议论文集，第 15750–15758 页，2021年。

Chi, C., Feng, S., Du, Y., Xu, Z., Cousineau, E., Burch- fiel, B., and Song, S. 扩散策略：通过动作扩散进行视觉运动策略学习。*arXiv*预印本 *arXiv:2303.04137*，2023年。Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. 门控循环神经网络在序列建模上的实证评估。*arXiv*预印本 *arXiv:1412.3555*，2014年。Clark, P., Tafjord, O., and Richardson, K. 将 Transformer 作为语言的软推理器。*arXiv*预印本 *arXiv:2002.05867*，2020年。

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., 等。训练验证器以解答数学文字题。*arXiv* 预印本 *arXiv:2110.14168*，2021年。

Dasari, S. 和 Gupta, A. 用于单次视觉模仿的 Transformer。载于机器人学习会议，第 2071–2084 页。PMLR，2021年。

Dayan, P. 和 Hinton, G. E. 封建式强化学习。神经信息处理系统进展，第 5 卷，1992年。

Eysenbach, B.、Gupta, A.、Ibarz, J. 和 Levine, S. 多样性就是你所需的一切：无需奖励函数即可学习技能。*arXiv* 预印本 *arXiv:1802.06070*，2018年。

Fan, L.、Wang, G.、Jiang, Y.、Mandlekar, A.、Yang, Y.、Zhu, H.、Tang, A.、Huang, D.-A.、Zhu, Y. 和 Anand-kumar, A. MineDojo：利用互联网规模知识构建开放式具身智能体。*arXiv* 预印本 *arXiv:2206.08853*，2022年。

Fishman, A., Murali, A., Eppner, C., Peele, B., Boots, B., 和 Fox, D. 运动策略网络。*arXiv* 预印本 *arXiv:2210.12209*，2022年。

Fishman, A., Murali, A., Eppner, C., Peele, B., Boots, B., 和 Fox, D. 运动策略网络。载于机器人学习会议，第 967–977 页。PMLR，2023年。

Florence, P., Manuelli, L. 和 Tedrake, R. 视觉运动策略学习中的自监督对应关系。*IEEERobotics and Automation Letters*，5(2):492–499，2019年。

Florence, P., Lynch, C., Zeng, A., Ramirez, O. A., Wahid, A., Downs, L., Wong, A., Lee, J., Mordatch, I. 和 Tompson, J. 隐式行为克隆。载于机器人学习会议，第 158–168 页。PMLR，2022年。

Fu, J., Kumar, A., Nachum, O., Tucker, G. 和 Levine, S. D4RL：深度数据驱动强化学习数据集。*arXiv* 预印本 *arXiv:2004.07219*，2020年。

Graves, A. 和 Graves, A. 长短期记忆。使用循环神经网络进行监督序列标注，第 37–45 页，2012年。

Gu, J.、Chaplot, D. S.、Su, H. 和 Malik, J. 多技能移动操作与物体重排。*arXiv* 预印本 *arXiv:2209.02778*，2019年。

Gu, J.、Xiang, F.、Li, X.、Ling, Z.、Liu, X.、Mu, T.、Tang, Y.、Tao, S.、Wei, X.、Yao, Y.、Yuan, X.、Xie, P.、Huang, Z.、Chen, R. 和 Su, H. ManiSkill2：可泛化操作技能的统一基准测试。载于国际学习表征会议，2023年。

Gupta, A.、Kumar, V.、Lynch, C.、Levine, S. 和 Hausman, K. 接力策略学习：通过模仿学习和强化学习解决长时程任务。*arXiv* 预印本 *arXiv:1910.11956*，2019年。

Hansen, N.、Yuan, Z.、Ze, Y.、Mu, T.、Rajeswaran, A.、Su, H.、Xu, H. 和 Wang, X. 关于视觉运动控制的预训练：重新审视从零开始学习的基线。*arXiv*预印本*arXiv:2212.05749*，2022年。

Hester, T.、Vecerik, M.、Pietquin, O.、Lanctot, M.、Schaul, T.、Piot, B.、Horgan, D.、Quan, J.、Sendonaris, A.、Osband, I. 等。基于示范的深度Q学习。载于*AAAI*人工智能会议论文集，第32卷，2018年。

Ho, J. 和 Ermon, S. 生成式对抗模仿学习。神经信息处理系统进展，29，2016年。

Ho, J.、Jain, A. 和 Abbeel, P. 去噪扩散概率模型。神经信息处理系统进展，33:6840–6851，2020年。

Howard, R. A. 动态规划与马尔可夫过程。1960年。

Hutsebaut-Buysse, M., Mets, K., and Latr´e, S. 分层强化学习：综述与开放研究挑战。机器学习与知识提取，4 (1):172–221，2022年。

James, S. and Davison, A. J. Q注意力：实现基于视觉的机器人操作的高效学习。*IEEERobotics and Automation Letters*，7(2):1612–1619，2022年。

Janner, M., Li, Q., and Levine, S. 将离线强化学习视为一个大型序列建模问题。神经信息处理系统进展，34: 1273–1286，2021年。

Janner, M., Du, Y., Tenenbaum, J. B., 和 Levine, S. 利用扩散进行灵活行为合成的规划。*arXiv* 预印本 *arXiv:2205.09991*，2022年。

Jia, Z., Li, X., Ling, Z., Liu, S., Wu, Y., 和 Su, H. 通过通才-专才学习改进策略优化。载于国际机器学习会议，第 10104–10119 页。PMLR，2022a。Jia, Z., Lin, K., Zhao, Y., Gao, Q., Thattai, G., 和 Sukhatme, G. 学习以具备可供性意识的多模态神经 SLAM 采取行动。*arXiv* 预印本 *arXiv:2201.09862*，2022b。Jiang, Y., Liu, E., Eysenbach, B., Kolter, J. Z., 和 Finn, C. 通过压缩学习选项。神经信息处理系统进展，35:21184–21199，2022年。

Kang, B., Jie, Z., and Feng, J. 基于示范的策略优化。发表于国际机器学习会议，第 2469–2478 页。PMLR，2018年。

Killick, R., Fearnhead, P. 和 Eckley, I. A. 以线性计算成本最优检测变点。美国统计学会期刊，107(500):1590–1598，2012年10月。doi: 10.1080/01621459.2012.737745。URLhttps://doi.org/10.1080%2F01621459.2012.737745。Kipf, T., Li, Y., Dai, H., Zambaldi, V., Sanchez-Gonzalez, A., Grefenstette, E., Kohli, P. 和 Battaglia, P. Com- pile：组合模仿学习与执行。载于国际机器学习会议，第3418–3428页。PMLR，2019年。

Konidaris, G., Kuindersma, S., Grupen, R. 和 Barto, A. 通过构建技能树实现从示范中进行机器人学习。国际机器人研究期刊，31 (3):360–375，2012年。

Kostrikov, I., Fergus, R., Tompson, J. 和 Nachum, O. 使用 Fisher散度评论家正则化的离线强化学习。载于国际机器学习会议，第 5774–5783 页。PMLR，2021年。Kujanp¨a¨a, K., Pajarinen, J. 和 Ilin, A. 使用向量量化模型的分层模仿学习。*arXiv* 预印本 *arXiv:2301.12962*，2023年。Kumar, A., Fu, J., Soh, M., Tucker, G. 和 Levine, S. 通过自举误差降低稳定离策略Q学习。神经信息处理系统进展，32，2019年。

Kumar, A., Zhou, A., Tucker, G. 和 Levine, S. 面向离线强化学习的保守Q学习。神经信息处理系统进展，33：1179–1191，2020年。

Kumar, V. 和 Todorov, E. MuJoCo Haptix：一种用于手部操控的虚拟现实系统。载于 *2015*年 *IEEE-RAS* 第*15*届国际人形机器人会议（*Humanoids*），第 657–663 页。IEEE，2015年。

Laskey, M.、Lee, J.、Fox, R.、Dragan, A. 和 Goldberg, K. Dart：用于稳健模仿学习的噪声注入。载于 机器人学习会议，第 143–156 页。PMLR，2017年。

Levine, S., Kumar, A., Tucker, G., and Fu, J. 离线强化学习：教程、综述及开放问题展望。*arXiv*预印本 *arXiv:2005.01643*，2020年。

Liang, Z., Bethard, S., and Surdeanu, M. 通过内部独白实现可解释的多跳言语推理。载于计算语言学协会北美分会*2021*年会议论文集：人类语言技术，2021年。

Ling, W., Yogatama, D., Dyer, C., and Blunsom, P. 通过推理生成进行程序归纳：学习解决并解释代数文字题。*arXiv*预印本 *arXiv:1705.04146*，2017年。

Liu, F., Liu, H., Grover, A., and Abbeel, P. 用于可扩展且可泛化决策的掩码自编码。*arXiv*预印本 *arXiv:2211.12740*，2022年。

Lynch, C., Khansari, M., Xiao, T., Kumar, V., Tompson, J., Levine, S. 和 Sermanet, P. 从玩耍中学习潜在计划。载于机器人学习会议，第 1113–1132 页。PMLR，2020年。

Mandi, Z., Liu, F., Lee, K. 和 Abbeel, P. 朝着更具泛化能力的单次视觉模仿学习迈进。*arXiv* 预印本 *arXiv:2110.13423*，2021年。

Mandlekar, A., Xu, D., Wong, J., Nasiriany, S., Wang, C., Kulkarni, R., Fei-Fei, L., Savarese, S., Zhu, Y. 和 Mart´ın-Mart´ın, R. 从离线人类示范中学习机器人操作时哪些因素至关重要。*arXiv* 预印本 *arXiv:2108.03298*，2021年。

Mu, T., Gu, J., Jia, Z., Tang, H. 和 Su, H. 利用自监督物体提议实现组合泛化能力的策略重构。神经信息处理系统进展，33:8883–8894，2020年。

Mu, T., Ling, Z., Xiang, F., Yang, D., Li, X., Tao, S., Huang, Z., Jia, Z. 和 Su, H. ManiSkill：具有大规模示范的通用操作技能基准测试。*arXiv* 预印本 *arXiv:2107.14483*，2021年。

Nair, A., Dalal, M., Gupta, A. 和 Levine, S. 利用离线数据集加速在线强化学习。*arXiv* 预印本 *arXiv:2006.09359*，2020年。

Nye, M., Andreassen, A. J., Gur-Ari, G., Michalewski, H., Austin, J., Bieber, D., Dohan, D., Lewkowycz, A., Bosma, M., Luan, D. 等。《展示你的工作：语言模型中间计算的草稿区》。*arXiv* 预印本 *arXiv:2112.00114*，2021年。

Pan, Y., Li, Y., Zhang, Y., Cai, Q., Long, F., Qiu, Z., Yao, T. 和 Mei, T.《Silver-bullet-3d 在 ManiSkill 2021 中：用于物体操作的从演示中学习方法和启发式规则方法》。*arXiv* 预印本 *arXiv:2206.06289*，2022年。

Paster, K., McIlraith, S. 和 Ba, J. 不能指望运气：决策Transformer为何在随机环境中失效。*arXiv*预印本 *arXiv:2205.15967*，2022年。

Peng, S., Liu, Y., Huang, Q., Zhou, X., 和 Bao, H. PVNet：用于六自由度位姿估计的逐像素投票网络。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 4561–4570 页，2019年。

Pertsch, K., Lee, Y., 和 Lim, J. 利用学习得到的技能先验加速强化学习。载于机器人学习会议，第 188–204 页。PMLR，2021年。

Pfrommer, S., Halm, M., 和 Posa, M. ContactNets：使用平滑的隐式表示学习不连续接触动力学。载于机器人学习会议，第 2279–2291 页。PMLR，2021年。

Pickett, M. 和 Barto, A. G. PolicyBlocks：一种用于在强化学习中创建实用宏动作的算法。载于国际机器学习大会，第 19 卷，第 506–513 页，2002 年。

Pomerleau, D. A. Alvinn：神经网络中的一辆自主陆地车辆。神经信息处理系统进展，1，1988年。

Qi, C. R., Su, H., Mo, K. 和 Guibas, L. J. Pointnet：用于三维分类和分割点集的深度学习。载于*IEEE*计算机视觉与模式识别会议论文集，第652–660页，2017年。

Qin, Y., Su, H. 和 Wang, X. 从一只手到多只手：通过单摄像头遥操作进行灵巧操作的模仿学习。*arXiv*预印本 *arXiv:2204.12490*，2022年。

Qureshi, A. H., Simeonov, A., Bency, M. J. 和 Yip, M. C. 运动规划网络。载于*2019*年国际机器人与自动化会议（*ICRA*），第2118–2124页。IEEE，2019年。

Rahmatizadeh, R., Abolghasemi, P., B¨ol¨oni, L., 和 Levine, S. 基于视觉的多任务操作：通过端到端示范学习实现低成本机器人的操作。载于*2018*年 *IEEE* 国际机器人与自动化会议 *(ICRA)*，第 3758–3765 页。IEEE，2018年。

Rajani, N. F., McCann, B., Xiong, C. 和 Socher, R. 请解释自己！利用语言模型进行常识推理。*arXiv* 预印本 *arXiv:1906.02361*，2019年。

Rajeswaran, A., Kumar, V., Gupta, A., Vezzani, G., Schulman, J., Todorov, E. 和 Levine, S. 利用深度强化学习和示范学习复杂的灵巧操作。*arXiv* 预印本 *arXiv:1709.10087*，2017年。

Reed, S. 和 De Freitas, N. 神经程序员解释器。*arXiv* 预印本 *arXiv:1511.06279*，2015年。

Ross, S. 和 Bagnell, D. 面向模仿学习的高效归约。在第十三届人工智能与统计国际会议论文集，第 661–668 页。JMLR研讨会与会议论文集，2010年。

Ross, S.、Gordon, G. 和 Bagnell, D. 将模仿学习和结构化预测归约为无悔在线学习。在第十四届人工智能与统计国际会议论文集，第 627–635 页。JMLR研讨会与会议论文集，2011年。

Saatc¸i, Y.、Turner, R. D. 和 Rasmussen, C. E. 高斯过程变点模型。在第 *27* 届国际机器学习会议（*ICML-10*）论文集，第 927–934 页，2010年。

Salter, S., Wulfmeier, M., Tirumala, D., Heess, N., Riedmiller, M., Hadsell, R. 和 Rao, D. MO2：基于模型的离线选项。载于终身学习智能体会议，第 902–919 页。PMLR，2022年。

Sasaki, F. 和 Yamashina, R. 基于含噪演示数据的行为克隆。载于国际学习表征会议，2020年。

Shafiullah, N. M. M., Cui, Z. J., Altanzaya, A. 和 Pinto, L. 行为变换器：用一块石头克隆 k 模式。*arXiv* 预印本 *arXiv:2206.11251*，2022年。

Shridhar, M., Thomason, J., Gordon, D., Bisk, Y., Han, W., Mottaghi, R., Zettlemoyer, L. 和 Fox, D. ALFRED：用于解读日常任务中基于环境的指令的基准测试。载于*IEEE/CVF* 计算机视觉与模式识别会议论文集，第 10740–10749 页，2020年。

Shridhar, M., Manuelli, L. 和 Fox, D. Perceiver-actor：一种用于机器人操作的多任务变换器。载于机器人学习会议，第 785–799 页。PMLR，2023年。

Singh, A., Liu, H., Zhou, G., Yu, A., Rhinehart, N., 和 Levine, S. Parrot：用于强化学习的数据驱动行为先验。*arXiv* 预印本 *arXiv:2011.10024*，2020年。

Sun, W., Venkatraman, A., Gordon, G. J., Boots, B., 和 Bagnell, J. A. Deeply aggrevated：用于序列预测的可微模仿学习。载于国际机器学习会议，第 3309–3318 页。PMLR，2017年。

Sutton, R. S., Precup, D. 和 Singh, S. 从 MDP 到半马尔可夫决策过程：强化学习中时间抽象的框架。人工智能，112(1-2)：181–211，1999年。

Tassa, Y., Doron, Y., Muldal, A., Erez, T., Li, Y., Casas, D. d. L., Budden, D., Abdolmaleki, A., Merel, J., Lefrancq, A. 等。DeepMind 控制套件。*arXiv* 预印本 *arXiv:1801.00690*，2018年。

Tennenholtz, G., Hallak, A., Dalal, G., Mannor, S., Chechik, G. 和 Shalit, U. 模仿学习和强化学习中潜在混杂因素的协变量偏移。*arXiv* 预印本 *arXiv:2110.06539*，2021年。

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J.,Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. 你所需要的只是注意力。神经信息处理系统进展，30，2017年。

Vuong, Q., Qin, Y., Guo, R., Wang, X., Su, H., and Christensen, H. 用于通用机器人操作的单 RGB-D 相机远程操作。*arXiv* 预印本 *arXiv:2106.14396*，2021年。

Wang, Z., Hunt, J. J., and Zhou, M. 扩散策略作为一种适用于离线强化学习的表达能力强的策略类别。*arXiv* 预印本 *arXiv:2208.06193*，2022年。

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., and Zhou, D. 思维链提示能够激发大语言模型中的推理能力。*arXiv* 预印本 *arXiv:2201.11903*，2022年。

Wu, Y.-H., Charoenphakdee, N., Bao, H., Tangkaratt, V., and Sugiyama, M. 从不完美示范中进行模仿学习. 载于国际机器学习会议，第 6818–6827 页. PMLR, 2019年。

Xu, D., Nair, S., Zhu, Y., Gao, J., Garg, A., Fei-Fei, L., and Savarese, S. 神经任务编程：学习在分层任务间泛化. 载于 *2018*年 *IEEE* 国际机器人与自动化会议 *(ICRA)*，第 3795–3802 页. IEEE, 2018年。

Xu, J., Hou, Z., Liu, Z., and Qiao, H. 比较基于接触模型的控制与无接触模型学习：机器人插销入孔装配策略综述. *arXiv* 预印本 *arXiv:1904.05240*，2019年。

Yang, M.、Schuurmans, D.、Abbeel, P. 和 Nachum, O. 思维链模仿与程序克隆。*arXiv* 预印本 *arXiv:2205.10816*，2022年。

Yang, S.、Nachum, O.、Du, Y.、Wei, J.、Abbeel, P. 和 Schuurmans, D. 用于决策的基础模型：问题、方法与机遇。*arXiv* 预印本 *arXiv:2303.04129*，2023年。

Zeng, A., Song, S., Lee, J., Rodriguez, A., and Funkhouser, T. Tossingbot：利用残差物理学习抛掷任意物体。*IEEE*机器人学汇刊, 36(4): 1307–1319, 2020年。Zeng, A., Florence, P., Tompson, J., Welker, S., Chien, J., Attarian, M., Armstrong, T., Krasin, I., Duong, D., Sindhwani, V., et al. Transporter网络：重排用于机器人操作的视觉世界。载于机器人学习会议, pp. 726–747。PMLR, 2021年。Zhang, T., McCarthy, Z., Jow, O., Lee, D., Chen, X., Goldberg, K., and Abbeel, P. 通过虚拟现实遥操作，对复杂操作任务进行深度模仿学习。载于*2018*年 *IEEE* 国际机器人与自动化会议 *(ICRA)*, pp. 5628–5635。IEEE, 2018年。Zhang, Y. and Chai, J. 利用统一Transformer和自我监控，从语言指令中进行分层任务学习。*arXiv* 预印本 *arXiv:2106.03427*, 2021年。Zhou, K., Yang, J., Loy, C. C., and Liu, Z. 学习为视觉语言模型设计提示。国际计算机视觉期刊, 130(9):2337–2348, 2022年。

A. 从次优示范中进行模仿学习，以完成低层控制任务所面临的一般挑战

非马尔可夫性 虽然演示数据中的每条轨迹都可以用马尔可夫策略表示，但通过完美模仿合并后的演示数据而将这些策略线性组合起来，若不同示范之间存在冲突，就可能产生负协同效应。这是因为这些演示数据可能由不同的智能体，或同一算法的不同运行过程生成。当示范本身由非马尔可夫智能体（例如人类或基于规划的算法）生成时，情况会更加严重。相比之下，非马尔可夫策略更具通用性，并能通过将历史信息作为额外上下文纳入其中来区分不同的演示数据，从而解决冲突。

噪声性 有时，演示轨迹本身带有噪声：对于相同的状态，会产生不同的动作。例如，基于搜索的规划器可能会根据相同的动作和状态历史，返回多个可用于完成任务的动作。有时，演示动作甚至呈均匀分布（例如，示范者使用运动规划算法时）。这会增加克隆策略的不确定性和方差，从而加剧误差累积。请注意，多模态性是一个相关但正交的问题（Shafiullah 等人，2022年），即对（连续）动作分布进行单峰估计会导致回报显著降低。

不连续性 对于低层控制任务，演示策略通常包含明显的值变化或拓扑变化（例如，由接触变化引起）。底层状态到动作的映射存在此类不连续性，使得学习稳健且准确的模型变得困难，进而损害泛化能力。近期的一种方法（Florence 等人，2022年）通过采用基于能量的隐式模型来替代显式模型，以应对这一问题。尽管该方法在理论上合理，但研究表明（Shafiullah 等人，2022年）它对于非马尔可夫隐式模型而言不太实用，而且后续出现的若干非马尔可夫显式模型表现优于它。

随机性 次优演示中通常存在实际或表面上的不可预测性，这可能是因为示范者的中间计算过程未在演示数据中呈现（例如，广度优先搜索生成的最短路径并未揭示中间搜索过程），也可能是因为示范者本身具有非确定性（例如依赖拒绝采样）。这一特性会降低模仿学习的鲁棒性，因为演示数据中的决策模式可能不明确、难以学习，且无法泛化（Paster 等人，2022年）。例如，在连续动作空间的迷宫中，通过随机搜索找到的解多少就像一张中奖的彩票，其模式可能很难泛化。

B. 我们对行为克隆进行评估时所用任务面临的具体挑战

我们评估所用的任务存在多项挑战。首先，这些任务涉及关节物体的大幅几何变化，这非常复杂，且在文献中研究不足（更多详情请参见 https://maniskill2.github.io/）。其次，动力学方面也存在挑战，包括：（1）高精度操作（一些 ManiSkill2 任务，例如插销插入，要求达到毫米级精度）；（2）欠驱动控制（例如推椅子和转动水龙头）；以及（3）由于底层阻抗控制器的存在，低层动作建模更加困难。第三，由于底层接触动力学或演示的生成方式（含噪、不连续等），演示本身并非最优。最后，演示数量可能有限（例如我们对 Franka Kitchen 的设置）。

C. 网络数据流消融研究示意图

为更好地说明第 5.4 节第三项消融研究中各个变体的差异（即“子技能与动作预测共享思维链令牌”），我们在图 5 中展示了这三种变体以及原始设计的网络数据流。

D. 关于相关工作的补充讨论

过程克隆（PC） PC（Yang 等人，2022年）最近被提出，它将示范者的中间计算输出作为额外监督信号，以提升行为克隆策略的泛化能力。然而，该方法假设能够全面了解示范者，包括通常隐藏的计算过程，而这些过程可能包含大量中间结果。例如，在原论文使用的图搜索示例中，PC 要求了解广度优先搜索算法的遍历路径，例如每个节点的状态——无论是最终回报结果中包含的节点，还是被排除（因而省略）的节点。相比之下，CoTPC 无需此类信息，因为思维链是结果的一部分，而不是

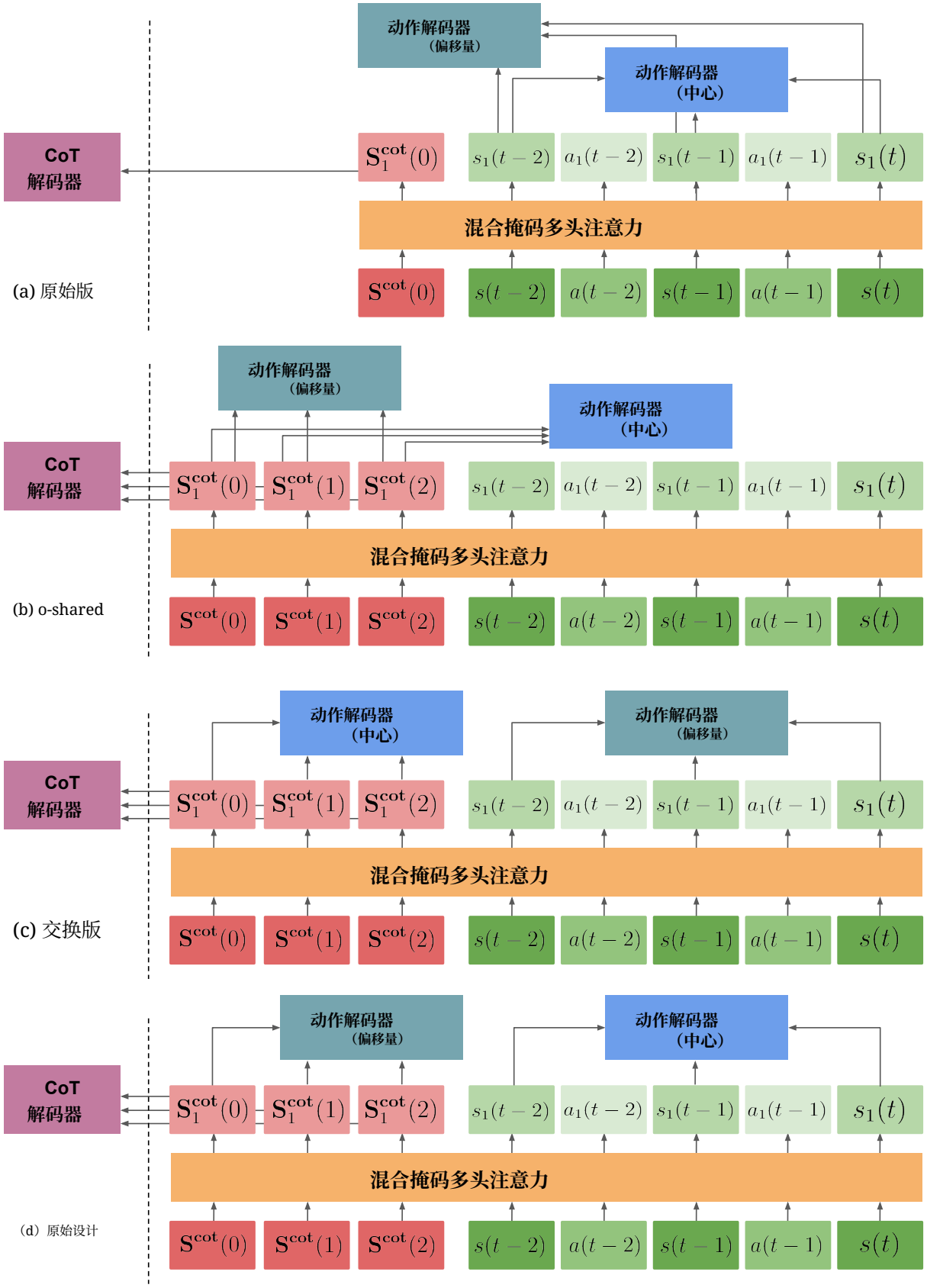


图5. 我们第三项消融研究中名为原始版、o-shared和交换版 的变体以及原始设计的网络数据流示意图。16

隐藏的中间计算，而思维链监督本身可以通过无监督发现方法获得。此外，机器生成的示范可以通过众包方式收集，而示范者通常被视为黑箱，这构成了 PC 的一项局限。

结合运动规划的策略学习 一些现有工作（Shridhar 等人，2023年；James 和 Davison，2022年）采用了与 CoTPC 类似的策略，将关键状态（航路点）预测为高层策略。然而，它们使用运动规划器作为低层策略，而 CoTPC 则直接学习预测低层控制。我们方法的主要优势在于能够处理需要动态（或反应式）控制的环境（如移动迷宫和推椅子）；在这些环境中，关键状态/观测必须在每一步更新，因此基于运动规划器的策略难以奏效。此外，PerAct（Shridhar 等人，2023年）由于采用离散化动作而受到一定限制，尤其是在需要高精度操作的任务中（例如插销插入）。此外，一些现有工作（Qureshi 等人，2019年；Fishman 等人，2023年）仅使用运动规划器作为示范者来获取神经策略，因此在一定程度上只能用于涉及准静态控制的任务。

机器人变换器-1 (RT-1) RT-1（Brohan 等人，2022年）是一项同期开展的工作，同样使用 Transformer 直接对低层控制动作建模。它得益于历时 17 个月预先收集的大规模真实世界机器人示范数据，以及对视觉输入（RGB图像）和低层动作进行的标记化。尽管 RT-1 在开发用于机器人学的决策基础模型方面展现出巨大潜力，但它采用了传统的自回归变换器，没有明确利用低层控制任务中体现的结构知识。此外，它的计算开销极大，通常每秒只能处理不到 5 个控制信号。我们的工作 CoTPC 是对此方向的早期探索，我们相信它将启发未来适用于各类机器人任务的通用模型设计。另一个区别是，由于 RT-1 对动作空间进行了离散化，因此在需要高精度的任务（例如插销插入）中可能会出现性能下降。

从离线数据中发现子技能 仅从离线示范集中发现技能或子技能极具挑战性，因为几乎没有任何有用的监督信号。在相关文献中，基于选项的方法是一种常用策略。例如，MO2（Salter 等人，2022年）是一种采用瓶颈状态原理的离线选项学习框架，研究表明，它在连续控制问题中学习选项时效果良好。然而，这类方法的缺点是依赖良好的状态空间表示。目前尚不清楚 MO2 是否能有效处理高维视觉观测（例如，由于软体操作的特点，倾倒任务仅支持视觉观测）。我们基于动作的方法与观测无关，因此避免了这个问题。此外，基于选项的方法还需要为选项学习良好的初始条件和终止条件，而这本身就是一个难题（Gu 等人，2022年）。通常，还需要进一步进行在线学习，以实现技能串联，从而弥补选项次优的问题（在通过离线数据集学习并固定选项后，MO2 仍然需要在线学习）。我们提出的另一种方案是，利用变点检测领域用于时间序列建模的方法来进行动作分段。变点检测方法大致可分为基于预测模型的方法（例如，（Saatc¸i 等人，2010年））和基于优化的方法（例如，（Killick 等人，2012年））。前者需要底层预测模型（UPM），其中（Saatc¸i 等人，2010年）选择了高斯过程。然而，目前尚不清楚它能否对高维且复杂的控制信号建模，因为动作序列本身就可能难以建模（（Saatc¸i 等人，2010年）指出，这可能需要大量敏感的超参数）。后者不直接对时间序列建模，而是通过优化代价函数来寻找变点。在这种情况下，选择合适的度量至关重要。

分层强化学习 +关于子技能分解的示范 尽管利用层级原则进行策略学习这一高层理念与分层强化学习 + 示范文献中的方法相同（例如，Jiang 等人，2022；Eysenbach 等人，2018；Konidaris 等人，2012；Pickett 与 Barto，2002），但这些方法都依赖于与环境进行在线交互：要么在迁移学习设置中使用在线强化学习，学习如何利用已学到的（子）技能（例如，Jiang 等人，2022），要么通过在线探索首先获取这些（子）技能（例如，Eysenbach 等人，2018）。与此相反，我们的方法是一种完全离线的模仿学习方法。能够用于下游模仿学习的纯离线子技能发现极具挑战性，而近期工作大多依赖强化学习来提供额外的监督信号（Kujanp¨a¨a 等人，2023）。在这种设置下，与我们的方法相近的工作包括 CompILE（Kipf 等人，2019）和 PC（Yang 等人，2022）。CompILE 使用变分自编码器对演示轨迹进行软分段。然而，其评估仅在一组有限且相对简单的控制任务上进行。另一方面，PC 借助额外的程序级监督信号进行分层策略学习，这限制了它的适用性，也使其难以直接比较。请注意，由于这一方向（子技能发现 + 分层

离线策略学习）的先前研究相对较少，CompILE 和 PC 都仅采用非分层模仿学习方法作为基线。类似地，在我们的比较中，我们也纳入了几种强大的非分层基线。

E. 环境详情

移动迷宫 移动迷宫是一个二维迷宫，具有位移连续动作空间（$\delta x,\ \delta y$），其中两个分量 $\in [1.5, 4]$。这是一个S形迷宫，高度为80、宽度为60，智能体从顶部右侧方形区域（绿色）内随机初始化的位置出发，目标是到达左下方的方形区域（同样为绿色）。每次环境重置时，两个区域（起始方形区域和目标方形区域）以及两座矩形桥（绿色）的初始位置都会随机化。具体而言，两个方形区域分别随机出现在顶部蓝色岛屿的右侧和底部蓝色岛屿的左侧。它们的初始位置在垂直方向上的变化范围为20。两座桥的初始位置在水平方向上的变化范围同样为20。在游戏过程中，除顶部方块外，它们中的每一个都会以随机恒定速度 $\in [1, 2]$（独立地）来回移动。当智能体降落在移动方块上时，该方块会变为静止状态。智能体无法穿过迷宫边界（但不会因此死亡）。为简单起见，我们采用10维状态观测，包括智能体的当前位置和四个绿色区域的位置。该任务需要动态控制/规划。

Franka Kitchen 我们提出了一种设定（因此也是原始设计 Franka Kitchen 任务的一个变体），要求智能体按照目标指定的顺序完成 4 项物体操作（共 7 种不同的选项）。这 7 项任务是：打开/关闭底部炉灶、打开/关闭顶部炉灶、打开/关闭灯、打开/关闭推拉式橱柜、打开/关闭铰链式橱柜、打开/关闭微波炉，以及将水壶推至/移动到目标位置。与其他变体相比，我们采用严格的判定标准：只有完成全部 4 个子任务，任务才算成功；每个子任务都必须按指定顺序完成，才计为已完成。如果执行了指定的 4 个子任务以外的子任务，环境就会终止。动作空间基于机器人的关节速度（8 维）。我们采用原始的 30 维状态观测，其中包含所有相关物体的位姿和本体感知信号，以及额外的 14 维目标嵌入。该嵌入为 7 个潜在子任务中的每一个分配一个 2 维向量。每个向量的取值为 $[0, 0], [0, 1], [1, 0], [1, 1]$（表示相应子任务的完成顺序）或 $[-1, -1]$（表示不应完成该子任务）。我们不在状态观测中包含物体的目标位姿（也就是说，我们要求智能体从示范中学习目标位姿）。

ManiSkill2 ManiSkill2（Gu 等人，2023年）是近期提出的一个全面的低层物体操作任务基准测试。我们选择了 5 个任务作为测试平台。(1) 堆叠方块：抓起一个方块，将其放置在另一个方块上方，然后使夹爪离开堆叠物。(2) 转动水龙头：打开不同的水龙头。(3) 插销插入：将长方体插销从侧面插入不同几何形状和尺寸的盒子中的孔内。(4) 推椅子：通过移动机器人，将不同的高度多关节椅子模型推到指定的目标位置。(5) 倾倒：将液体从瓶子倒入目标烧杯中，使其达到指定液位。所有任务中的物体位姿在环境重置时都会完全随机化（位移约为 0.3 米，且 360°旋转随机）（这不同于 Franka Kitchen 等环境）。请注意，插销插入任务中的孔仅有 3 毫米的间隙，因此需要高度精确的操作；此外，插销至少有一半需要从侧面推入孔中（不同于类似但更简单的任务（Xu 等人，2019年））。我们选择的任务同时涉及静态操作和移动操作，并覆盖 3 种动作空间（推椅子使用关节速度增量控制，倾倒使用末端执行器位姿增量控制，其余任务使用关节位姿增量控制）。对于状态观测，我们在这些任务中使用维度各异的状态（详见 ManiSkill2 论文）。对于转动水龙头任务，我们对默认状态观测进行了轻微修改，附加了一个额外的 3 维向量（水龙头连杆的位姿），以便智能体更容易区分不同的水龙头模型。相应的示范也进行了修改。对于点云观测，我们使用 ManiSkill2 提供的默认预处理策略，在每个时间步获得固定长度为 1200 个点的点云（RGB 和 XYZ）。

F. 示范详情与评估协议

移动迷宫 我们采用多种启发式方法和 RRT 风格规划器，并利用测试时无法获得的事后信息来筛选示范。对于每种随机环境配置，我们从找到的路径中随机选择一条从起始方格通往目标方格的路径作为示范轨迹。我们将迷宫划分为 6 个区域：三个岛屿（每座桥都归属其下方的岛屿；起始方格和目标方格分别归属右上方和左下方的区域），每个岛屿又进一步划分为两个区域（沿中线垂直切分）。我们使用 RRT 风格采样器，依次寻找连接相邻区域的路径（从初始位置到

第二个区域）。我们将每条穿过两个相邻区域的路径的步数限制为 ≤ 13，使示范轨迹的最大总长度为 $\leq 13 \times 5 = 65$ 步。为了在动态环境中实现这种规划，我们实际上先在静态版本的迷宫中生成示范，然后再协调地为移动元素添加动画效果。由于这需要事后信息，因此在推理阶段无法做到这一点。我们使用 400 条示范轨迹进行训练，并在这 400 种配置以及一个包含 100 种未见环境配置的留出集上评估所有智能体。我们将任务成功率作为主要度量。推理阶段，我们将最大步数设为 60。

Franka Kitchen 我们在模拟器中重放了（Gupta 等人，2019年）中最初提出的一部分人类演示。具体而言，我们随机选取了 50 条长度介于150 至300 之间的演示轨迹，这些轨迹成功完成了 7 个选项中的 4 种不同子任务。我们利用特权信息为这些轨迹重新标注，以构建前文所述的目标嵌入。请注意，对于每条轨迹，该嵌入向量在不同时间步保持固定。50 份示范中共包含 20 种不同的有序子任务组合，其中大多数组合仅有 ≤ 3 条轨迹。在这种情况下，子任务的组合泛化难度过高（共有$35 \times 4! = 840$ 种组合）；因此，我们重点评估模型对初始机器人/物体位姿的泛化能力。我们使用原始人类演示中呈现的 90 种未见环境配置（每种都需要完成 4 个子任务）进行评估（仅纳入已见过的子任务组合）。我们报告任务成功率（要求一条轨迹完成全部 4 个子任务）以及每条轨迹成功完成的子任务平均数量作为度量。推理期间，我们将最大步数设为 280。

ManiSkill2 除推椅子以外的所有任务，我们都使用 ManiSkill2 提供的原始示范数据，这些数据由任务与运动规划求解器和启发式方法混合生成。详情请参阅原论文（不过，实际用于生成这些示范数据的代码并未公开）。对于堆叠方块和转动水龙头任务，我们随机抽取了 500 条演示轨迹作为训练数据。对于转动水龙头任务，我们使用来自 10 种不同水龙头模型的轨迹作为演示数据，并在 4 个未见过的水龙头上评估零样本泛化能力（每个水龙头对应 100 种不同的场景配置）。需要注意的是，转动水龙头任务的示范中，大多数水龙头是通过推动而非抓取来操作的，即采用欠驱动控制。对于插销插入任务，每种不同的环境配置都对应不同形状和尺寸的盒子（带孔）及插销。我们从 ManiSkill2 的原始示范数据中随机抽取了 1000 条轨迹作为训练数据，并抽取了 400 种未见过的环境配置用于零样本泛化评估。对于推椅子任务，我们采用了文献（Pan 等人，2022年）中一种复杂的、基于启发式方法的方法，并将其改编用于 ManiSkill2（该方法使用额外的特权信息，任务成功率约为 50%），作为示范者。我们在 6 种椅子模型上收集了 1000 条成功轨迹作为训练数据，并在 3 种未见过的椅子模型所对应的 300 种环境配置上评估零样本泛化性能。对于倾倒任务，我们使用 ManiSkill2 提供的 150 条成功重放的示范轨迹来生成点云观测序列（该任务涉及软体，因此不支持状态观测）。我们采用 ManiSkill2 原论文中的成功条件来报告任务成功率。推理时，堆叠方块、推椅子、插销插入、转动水龙头和倾倒任务允许的最大步数分别设为 200、200、200、250 和 300。

G. 发现的子技能序列示意图

在图 6中，我们展示了使用第 4.1 节所述的自动子技能发现过程，从两条轨迹中提取的子阶段（以及相应的子技能）。

H. 网络架构与训练的实现细节

原始行为克隆：我们使用一个马尔可夫策略，该策略由一个三层多层感知机实现，隐藏层大小为 256，采用 ReLU 非线性函数。我们使用 Adam 优化器，以 $1e-3$ 的固定学习率和 32 的批量大小进行训练，训练迭代次数分别为 150K（移动迷宫）、300K（Franka Kitchen）和 500K（ManiSkill2）。我们发现，即使采用 L2 正则化，延长训练时间也会导致过拟合。

Decision Diffuser 我们采用 DD 作者提供的参考实现，并对扩散模型作出以下调整：100 个扩散步数、20 的上下文长度和 4 的时域长度（在我们的实验中，发现更长的时域长度效果更差）。扩散模型和逆动力学模型总共有 ∼1.6M 个参数。由于 DD 适用于固定序列长度，我们在训练期间对初始状态和结束状态进行填充，在推理期间则仅对初始状态进行填充。

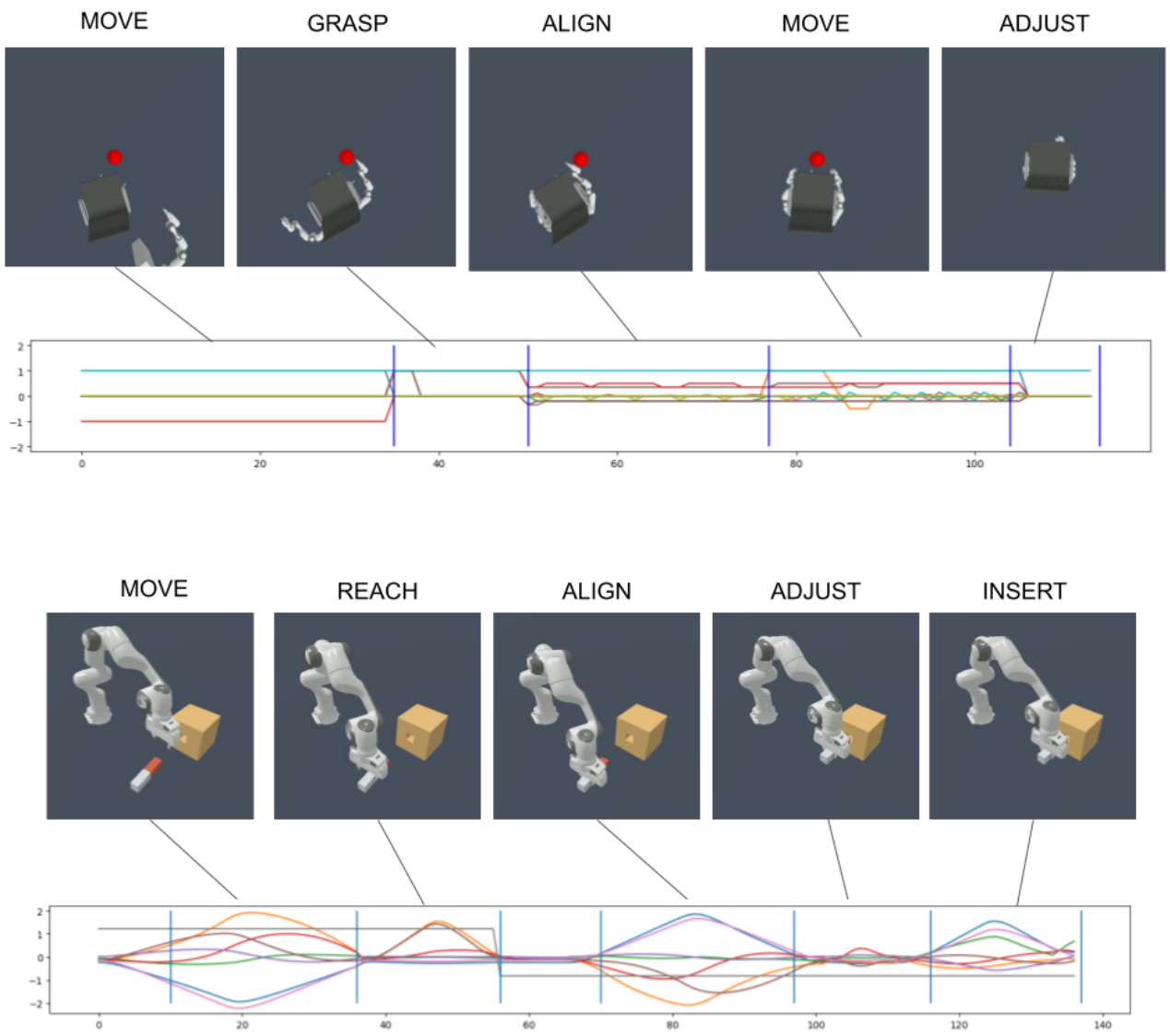


图*6*。 展示了两个任务中不同阶段对应的动作及相关观测：推椅子（上）和插销插入（下）。这些阶段由我们的无监督子技能发现方法通过将动作分组成子技能而发现。

扩散策略 我们采用 DP 作者提供的参考实现，同时调整模型规模，使其具有~1.2M 个参数（与其他基线及我们的 CoTPC 相当）。

决策Transformer 我们采用 minGPT 实现，并对所有任务使用相同的超参数配置（特征嵌入维度为 128，包含 4 个掩码多头注意力层，每层有 8 个注意力头），可学习参数总数略高于 1M。动作解码器是一个 3 层多层感知机，包含两个隐藏层，每层大小为 256，并使用 ReLU 非线性激活函数（但对于 Franka Kitchen，我们使用一个包含 2 层、隐藏层大小为 1024 的多层感知机）。我们使用学习率 $5e-4$，通过较短的预热期和余弦衰减计划将学习率降至 $5e-5$，训练所有任务（Franka Kitchen 除外，其终止学习率为 $5e-6$），并使用 Adam 优化器，批量大小为 256。Moving Maze 和 Franka Kitchen 训练 200K 次迭代，ManiSkill2 中的所有任务训练 500K 次迭代。所有任务的权重衰减设为0.0001，但 Franka Kitchen 使用0.1，Push Chair 使用 0.001。ManiSkill2 任务的上下文长度设为 10，Moving Maze 设为 20，Franka Kitchen 设为 10。我们参照 DT 论文，为状态和动作令牌使用可学习的位置嵌入。

行为Transformer 本文使用 BeT 的一种修改版本，通过结合过去的动作和过去的状态作为上下文来预测动作。我们从原论文中用于 Franka Kitchen 任务的配置开始。我们将 K 均值的分箱数量改为 1024（我们发现，对于我们的任务，较少的分箱效果较差），并将上下文长度改为 10（与其他基于 Transformer 的模型一致）。Transformer 主干网络的参数数量（约 ~1M）与 CoTPC 和决策变换器大致相同。我们训练模型约 5 万次迭代（我们发现，延长训练容易导致 BeT 过拟合，这可能是因为其离散化策略以及用于行为克隆的演示数据有限）。对于 ManiSkill2 任务，我们使用与决策变换器相同的架构，但使用中心加偏移解码器来解码动作。我们使用包含 128 个簇的 K 均值对动作空间进行划分。转动水龙头、推椅子和倾倒任务训练 10 万次迭代。插销插入和堆叠方块任务训练 20 万次迭代。训练动作偏移回归器时，我们使用 100 的系数（即，动作偏移回归器损失相较于动作中心分类损失放大 100 倍）。除非另有说明，其他超参数均遵循决策变换器部分的描述。

CoTPC 除非此处另有说明，其他配置（包括模型训练和网络架构）均与 BeT 相同。我们不为思维链令牌添加位置嵌入，因为它们本身就是可学习提示。思维链解码器是一个具有 ReLU 非线性的两层多层感知机，隐藏层大小为 256。所有任务的辅助均方误差损失系数均设为 $\lambda = 0.1$。训练期间，我们对动作令牌和状态令牌应用随机掩码，使思维链令牌能够关注不同长度的历史信息（从第一步到随机选取的第 t 步）。更多详情请参阅正文。对于所有任务，我们使用与 BeT 基线相同数量的簇来划分动作空间。另一项实现细节是，我们发现，使用最后一个注意力层之外的思维链特征 $\mathbf{S}_j^{\mathbf{cot}}(\cdot)$ $j \leq J$ 作为思维链解码器的输入，可以进一步提升最终性能。我们将其视为每个任务的超参数。通常，我们发现使用 $\mathbf{S}_J^{\mathbf{cot}}(\cdot)$是一个不错的起点。除非另有说明，其他超参数均遵循 BeT 中的设定。

I. 无监督子技能发现的度量选择

我们通过分解动作序列来发现子技能，这一方法源于这样的直觉：在时间上相近且功能相似的动作应归入同一子技能。如何选择度量来衡量这种功能相似性，仍是一个开放问题。我们尝试了余弦相似度、L2距离和更复杂的豪斯多夫距离。在这些方法中，我们发现余弦相似度兼具简洁性和泛化能力。在实验中，我们发现它适用于不同的动作空间（例如，移动迷宫中的位置增量、插销插入中的关节位姿增量，以及推椅子中的关节速度增量）；这些动作空间的动作统计特征因生成方式不同而异（例如，RRT 等基于采样的方法、启发式方法等），且该方法能够良好地扩展到更高维的动作空间（例如，移动迷宫使用二维动作，推椅子则使用双臂移动机器人的19维动作）。实际上，在机器人任务中，维度高得多的动作空间相对少见。

J. 子技能发现粒度的超参数调优

我们在子技能发现过程中调整 β，以控制发现的子技能序列的粒度。针对不同的动作空间（例如，双臂与单臂），我们将少量（5–10 条）演示轨迹作为验证集进行可视化，以调整 β。我们发现，对于所有单臂控制任务，β 对参数取值相对不敏感，设为 0.05；对于所有双臂控制任务，则设为 0.02

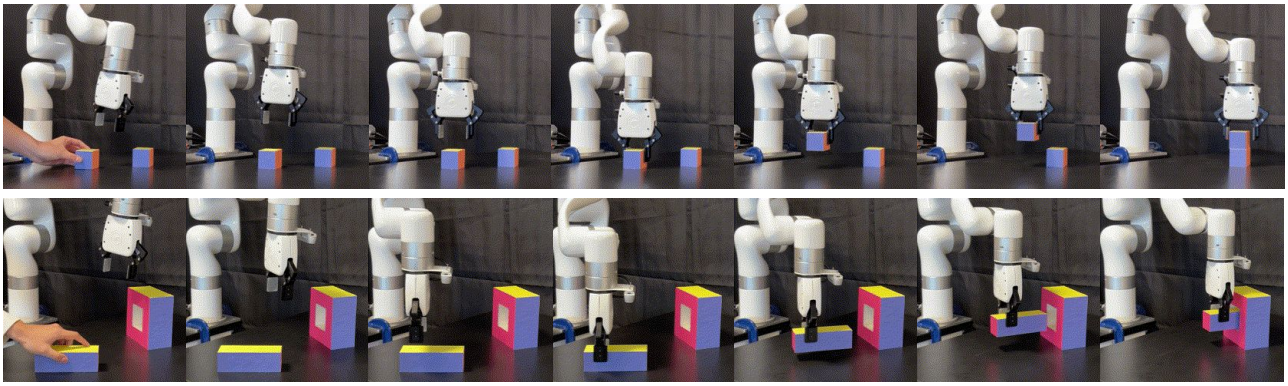


图7。 在真实机器人环境中，分别针对堆叠方块和插销插入任务采样得到的两条成功轨迹。所用的是仅基于模拟器中的演示数据训练的状态式 CoTPC 策略。作为初步检验，我们将插销插入的间隙从 3 毫米（模拟环境）增大到 10 毫米（真实环境），并且只使用几何形状固定的插销和带孔方块模型。

表5。额外测试性　　能（成功率）：在 ManiSkill2 任务中，使用状态观测时未见设置和零样本设置下的表现。

	堆叠方块	插销插入	转动水龙头	推动椅子
	未见	零样本	零样本	零样本
行为Transformer	73.0 (68.7±4.0)	42.5 (39.5±2.7)	32.5 (31.7±0.8)	33.4 (32.3±1.0)
CoTPC（本文方法）	86.0 (84.0±2.6)	59.3 (51.5±7.4)	39.3 (35.4±3.6)	41.0（36.4±4.0）

任务。此外，我们提出的模型对于 β 具有较强的鲁棒性。在一项消融研究中，我们手动调整模型，使发现的子技能平均长度有所变化。我们发现，在插销插入任务（采用状态观测策略设置）中，平均而言，子技能长度缩短约一个会导致性能略有下降（以零样本成功率计，59.3 对比 54.0）；平均而言，子技能长度增加约一个也会产生类似影响（以零样本成功率计，59.3 对比 55.3）。然而，当发现的子技能长度显著增加（约为 7）时，性能会大幅下降，因为此时过度细分的子技能无法为策略学习提供多少有价值的结构信息。

K. 仿真到现实迁移的初步结果

我们在堆叠方块和插销插入这两个任务中，以零样本设置考察基于状态的 CoTPC 进行仿真到现实迁移的可行性。图中展示了真实世界实验的设置以及两条采样得到的成功轨迹。7。借助 PVNet（Peng 等人，2019年）等现成的姿态估计框架，我们仅使用仿真数据训练的基于状态的 CoTPC 策略即可取得合理的性能。项目页面上提供了视频。作为初步考察，我们仅进行了定性评估。

L. 附加 ManiSkill2 上的额外结果（3 次运行的均值和标准差）

除了主论文中的结果（我们报告原始设计论文三次训练运行中的最佳结果，与 (Chi 等人，2023年) 中的评估流程类似）外，我们还在表 5 中报告均值和标准差。格式为最佳值（均值 ± 标准差）。

M. 基于点云的 CoTPC 详情

为处理点云观测，我们采用轻量级的 PointNet（Qi 等人，2017年）（~27k 个参数），并与 Transformer 一起从头开始以端到端方式训练。我们将额外的本体感知信号与点云特征拼接后输入状态令牌。在 CoTPC 中，思维链解码器经过训练以预测点云思维链，我们让解码器改为预测思维链的 PointNet 特征。我们发现，辅助点云思维链损失会导致

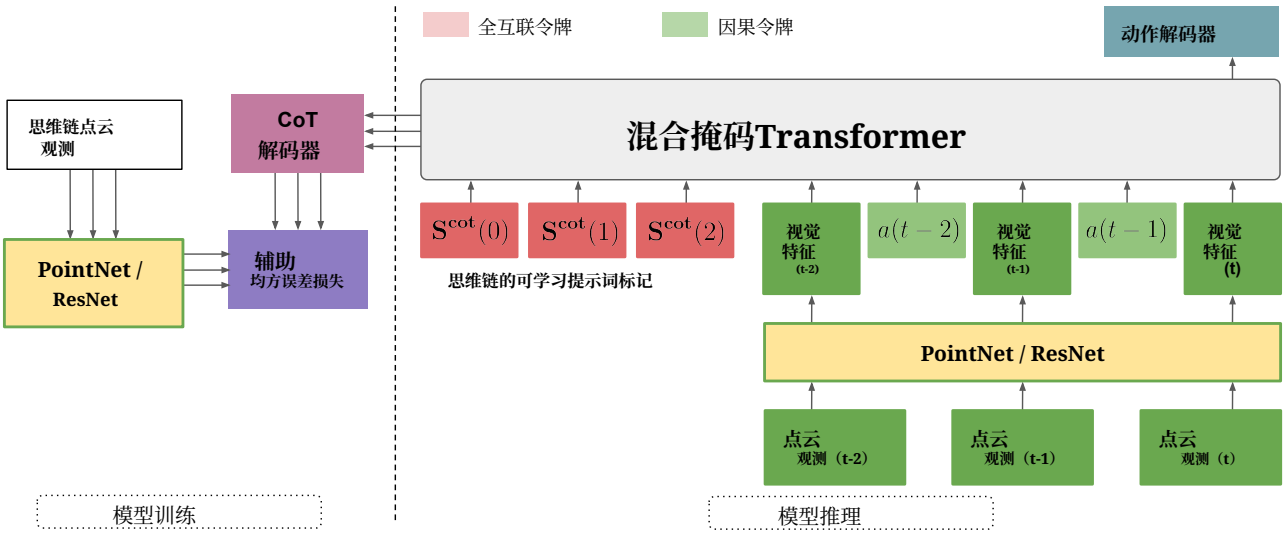


图 *8.* 基于点云的 CoTPC 示意图。与基于状态的 CoTPC 相比，我们添加了一个 PointNet，用于处理点云观测和点云思维链。我们省略了输入本体感知信号到模型的数据路径，也省略了动作解码器的详细数据流。

PointNet 编码后的表示发生坍塌。受（Chen 和 He，2021年）的启发，我们在点云思维链编码路径中采用停止梯度操作以防止这种情况。我们在图 8中展示了网络架构。训练策略（我们使用相同的一组超参数）和评估协议与基于状态的实验相似。

N. 补充视频

我们在项目页面上提供了基于状态的 CoTPC 在 ManiSkill2 任务上的推理结果视频（.gif 动画）。