ByteDance | Seed

DemoBot：利用第三人称人类视频，高效学习灵巧手双手操作

字节跳动 Seed

完整作者名单见贡献者

arXiv:2601.01651v1 [cs.RO] 4 Jan 2026

摘要

要实现通用具身智能，机器人必须能够高效、大规模地获取多样化操作技能。然而，当前的数据驱动范式受到高质量机器人数据稀缺的根本制约，因为通过遥操作收集示范成本高昂且难以扩展。直接从海量互联网规模的人类视频中学习，为解决这一数据瓶颈提供了一个有前景的方案，但也带来了一系列新挑战：人类手部与机器人手之间显著的*具身差距*，以及*模态差距*——视觉数据缺少任务执行所需的关键物理动力学信息。因此，直接模仿人类视频对于复杂任务仍然效果不佳，也难以实现。

为应对这些挑战并弥合视觉观察与物理执行之间的差距，我们提出了**DemoBot**，这是一个新颖的框架，可从单个未标注的RGB-D视频中学习长时序双手灵巧操作技能。DemoBot的核心思想是将提取的人类轨迹用作灵活的运动先验 以提供引导，而非严格的模仿目标。首先，数据处理模块从原始视频数据中提取双手与物体的结构化运动轨迹。这些轨迹作为运动先验，供一个新颖的强化学习（RL）流程使用；该流程通过高接触交互学习对轨迹进行优化，从而无需从头开始学习。为进一步应对学习长时程操作技能的挑战，我们提出：（1）基于时间分段的强化学习，以确保当前状态与示范在时间上对齐；（2）成功门控重置策略，以平衡对已掌握技能的优化和对后续任务阶段的探索；（3）采用自适应阈值设定的事件驱动奖励课程，引导强化学习掌握高精度操作。这一新颖的数据处理与强化学习框架成功完成了长时序同步及异步双手装配任务，为直接从人类视频中获取技能提供了一种可扩展的方法。实验表明，DemoBot能够通过单个视频示范，高效掌握高接触双手技能，为从人类视觉示范中学习提供了一条可扩展的途径。

日期：2025年12月12日**联系邮箱：**
yucheng.xu@bytedance.com**项目页面：**
https://demobot-seed.github.io/

1 引言

直接通过观察人类来学习复杂操作技能的能力，是大规模实现具身通用人工智能所必需的关键能力。真正的通用型机器人应当能够学习新的能力，而不是依赖数月耗时的遥操作——这一过程从根本上缺乏可扩展性。

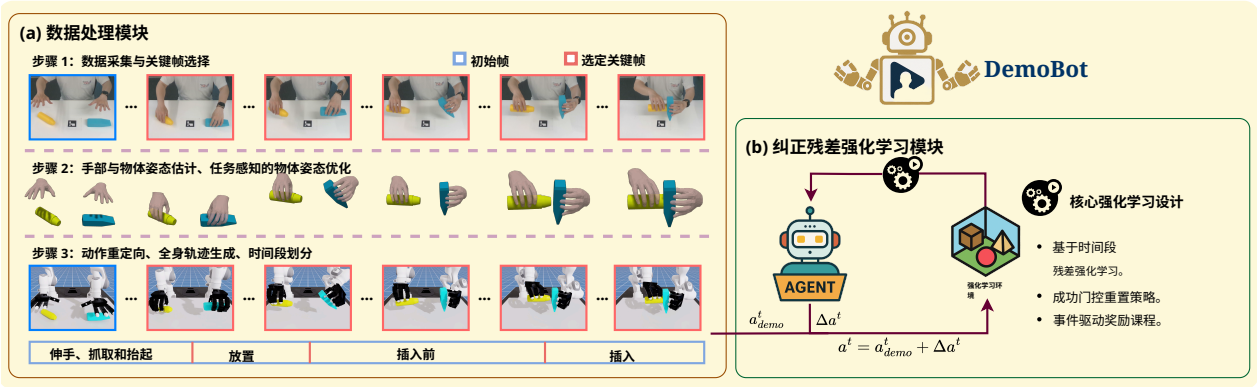

图1 DemoBot框架：从单个视觉演示中学习双手技能。(a) **数据处理模块** 分三个步骤将原始RGB-D视频转换为结构化运动先验：(1) 录制人类演示，并通过关键帧进行人工分段。(2) 手部和物体估计器生成三维手部和物体姿态，随后使用任务感知优化对其进行细化。(3) 将细化后的人类运动重定向至机器人，生成全身轨迹，并依据关键帧将其划分为有意义的时间片段。(b) 随后，**纠正残差强化学习模块** 将这些片段用作运动先验。强化学习智能体学习一种纠正策略，输出残差动作 Δa，并将其加到运动先验 $a = a_{demo} + \Delta a$ 上，使机器人能够掌握原始视觉数据中未体现的丰富接触物理动力学，并完成任务。

实现泛化的一条更有前景的路径，是让机器人从互联网上海量的人类视频中学习 [20]。然而，对于这样一种可扩展的学习系统而言，一个根本性挑战在于需要高效的数据处理和有效的技能学习。要让机器人从海量视频中学习技能，一项核心竞争优势在于，能够仅通过一次视频示范，稳健而高效地掌握一项新技能。这种单次学习能力——类似于学徒只观察熟练工匠一次——是最关键的基础模块之一 [20]。实现这一点，有望释放利用海量互联网规模人类活动数据集的潜力，并创造新一代通用机器人。然而，将视觉演示直接转化为成功的机器人技能，尤其是针对双手灵巧任务，迄今仍是研究领域尚未解决的挑战。

需要克服的首要障碍，是弥合人类示范者与机器人智能体之间的巨大差距。现有的视觉演示学习方法通常依赖专用硬件来获取运动数据，例如带有虚拟现实设备 [4, 6, 37] 的远程操作系统、可穿戴设备 [18, 19, 32, 34, 44]，或动作捕捉系统 [3, 14, 39, 45]，。直接从单个未标注的RGB-D视频中学习——这种设置可扩展性和可及性都高得多——并非易事。数据缺少动作标签，且天生带有噪声；物体和手部经常被遮挡；此外，人手与机器人末端执行器之间还存在根本性的具身不匹配。

即便从视觉演示中提取出了三维运动轨迹，如何基于不完美的数据学习稳健的策略仍是一项重大挑战。长时程操作不仅需要准确捕捉运动学运动，还需要掌握物理动力学。传统的模仿学习（IL）方法 [6, 18, 19, 32, 34, 39, 44, 45] 通常针对高质量、运动学上准确且物理上可行的示范而设计。然而，由于存在模态差距，这一假设并不适用于源自视觉演示的数据，因为这类数据缺少有关物理动力学的关键信息。强化学习（RL）方法 [2, 3] 能够通过与物理世界进行大量交互来学习接触丰富的技能，但也面临自身的一系列困难。由于多指手对应的状态-动作空间维度很高，从头开始通过强化学习学习长时程双手灵巧操作技能，众所周知既样本效率低又困难。此外，还存在与示范的时间错位 [12]，、在广阔的状态-动作空间中进行有效探索，以及在长时间尺度上分配信用等挑战 [7, 22]。

为应对这些挑战，我们提出了一种新颖的框架 **DemoBot**，能够从第三人称人类视频中学习双手

灵巧操作技能，所需数据甚至只需一段未标注的 RGB-D 视频示范。该框架由一个鲁棒的数据处理模块和一个新颖的残差强化学习流程组成。数据处理模块从 RGB-D 视频中提取物体和人手运动，进而将其映射为全身机器人动作轨迹，以及一组待实现的物体子目标。残差强化学习流程解决了两大挑战：（1）从单次含噪且不完美的示范中学习机器人技能；（2）解决长时序、复杂的双手操作任务。首先，从单次演示中学习的核心思想，是将从视觉演示中提取的不完美运动轨迹视为合理但不完美的运动先验，用作引导，而非必须模仿的目标。强化学习流程通过与物理仿真交互并完成任务，学习一种纠正性残差策略，在局部对运动先验进行修正，以补足原始视觉演示中缺失的物理动力学信息。其次，为应对长时序复杂双手操作任务带来的挑战，我们在强化学习流程中引入了三项主要创新设计：（1）**基于时间分段的强化学习**范式，将整个任务拆分为一系列分段，以缓解采集到的长时序示范与当前强化学习状态之间的时间错位。随后，训练强化学习智能体依次完成每个分段的目标。采用时间分段后，强化学习智能体可以独立学习当前阶段运动先验的纠正性残差，避免受到其他阶段误导性信息的影响；（2）**成功门控重置策略**，通过随机将失败环境重置到其上一个成功终止状态，兼顾早期技能的保持与后续技能的深入探索。这确保了采集到的学习样本能够覆盖整个任务；（3）**事件驱动奖励课程**，通过结合密集奖励和稀疏奖励，并将其进一步分解为与事件相关的项（例如伸手、抬起、重新放置等），提供平滑的学习信号。当满足特定条件时，这些奖励项便会受到控制并被激活。此外，我们还设计了一种课程学习策略，根据训练进度逐步降低稀疏奖励的阈值，从而逐步提高操作精度。所提出的强化学习流程将人类视觉示范提供的大幅度运动引导与强化学习中的丰富接触探索相结合，同时克服长时序复杂双手操作任务中的典型挑战。

所提出框架的主要贡献如下：

- 一种新颖且鲁棒的视频处理流程，采用基于 MANO 的人手表示，并针对物体位姿进行与任务相关的改进，以优化提取出的三维手—物体运动先验，将未经标注的二维视频转换为适用于机器人学习的高质量三维运动先验——在单个图形处理器上处理一段 15 秒视频仅需 4 分钟计算时间。
- 一套专为演示增强型长时域任务设计的新颖强化学习技术：采用**基于时间分段的强化学习** 范式来缓解时间错位，采用**成功门控重置** 策略来促进深度探索，并通过**事件驱动奖励课程** 来学习高精度灵巧技能。
- 据我们所知，这是首个此类学习框架，能够从单个人类视觉演示中高效学习长时序双手灵巧操作技能。

我们通过一系列具有挑战性的长时程双手操作任务验证了 DemoBot 的有效性，这些任务要求精确的协调与接触管理。大量实验表明，所提出的残差学习范式成功弥合了视觉演示中固有的模态差距和具身差距，并缓解了长时域技能学习中的挑战。结果显示，该系统能够掌握复杂技能，而标准行为克隆和强化学习基线方法在这些技能上表现不佳；这证明，若与纠正性残差策略结合，次优视觉数据也可作为有效的运动先验。

最终，本研究无需昂贵的遥操作硬件或海量数据集，仅凭单个未标注的RGB-D视频即可实现稳健的技能获取，标志着可扩展机器人学习迈出了重要一步。DemoBot为通用机器人提供了一条切实可行的路径，使其能够借助互联网上海量、开放式的人类视频数据库来扩展技能库，从而让我们更接近具身通用人工智能这一目标。

2 相关工作

本节回顾相关研究，重点关注两个主要领域：（1）用于示范学习的数据采集；（2）示范增强强化学习。

2.1 示范学习的数据采集

相当一部分以往的示范学习（LfD）研究都是针对高保真数据设计的。遥操作装置、远程控制器 [4, 6, 37] 或可穿戴设备 [18, 19, 32, 34, 44] 能够提供干净且运动学有效的动作数据，但需要昂贵的专用硬件，并且操作人员需要付出大量努力。同样，配备多摄像头阵列的动作捕捉系统 [3, 14, 39, 45] 可以获取精准的人体运动数据，但通常受限于实验室环境，难以轻松扩展。相比之下，从被动的第三人称视频中学习具有更强的可扩展性

表 1 与近期示范学习（LfD）研究工作的比较。

	灵巧双手	LfV	单次示范	
RobotTube[41]	✗	✓	✓	✗
DexMV[25]	✓	✗	✓	✗
YOTO[45]	✗	✓	✓	✓
DexCap[39]	✓	✓	✗	✗
ManipTrans[14]	✓	✓	✗	✗
[18] 中的工作	✗	✓	✗	✓
DemoBot（我们的方法）	✓	✓	✓	✓

一种易于获取的方法，并有望利用互联网规模数据 [25–27, 41]。然而，这也带来了显著的感知和形态学挑战。仅凭单路 RGB 或 RGB-D 数据流重建三维手物交互是一个病态问题，往往会导致姿态估计存在噪声且不准确。表1总结了近期示范学习工作的主要特点。我们的工作通过提出一种稳健的数据处理流程，为这一领域作出了贡献。该流程采用现代的基于模型的手部估计器 [16, 23, 24]，并引入一种新颖的任务感知优化步骤，以改进这些不完善的估计结果。

2.2 示范增强的强化学习

为克服模仿学习的脆弱性，同时保留示范学习的样本效率，一项重要趋势是将示范与强化学习相结合 [10, 14, 25, 28, 37, 38]。这些方法利用示范来引导并加速强化学习。一种典型策略是鼓励强化学习智能体模仿示范 [10, 28, 38]。然而，这些方法本质上会受到当前强化学习状态与示范之间时间错位的影响，尤其是在长时域任务中。其他方法 [14, 25, 37]旨在从收集到的示范中进一步提取深层表征。文献 [37]采用生成对抗模仿学习，从示范中学习奖励函数 [11]。文献 [14, 25] 则从收集到的示范中学习状态-动作映射，以初始化强化学习策略。然而，这些方法需要多次示范，才能确保所学奖励函数或动作映射的可靠性；而单次演示无法满足这一要求，因此增加了规模化应用的难度。

我们的工作基于残差强化学习范式 [13, 33], ，其中示范动作作为基准轨迹，强化学习智能体则学习修正残差动作。最终动作是基准动作与学得的残差动作之和。这种结构有效地将一个困难的高维控制问题转化为更易处理的局部修正学习问题。为解决长时程任务中的时间错位问题，我们将完整的示范轨迹划分为多个时间片段，并调整残差强化学习方法，依次优化各个片段。结合这些新设计，我们能够仅凭单次演示有效地学习长时程操作技能。

3 方法论

本节将深入介绍我们提出的 DemoBot 的核心设计。首先，我们介绍用于从人类视觉示范中提取手物运动先验的数据处理模块。然后，我们介绍新颖的示范增强残差强化学习模块，该模块专为从单个人类示范中学习长时序复杂双手操作技能而设计。

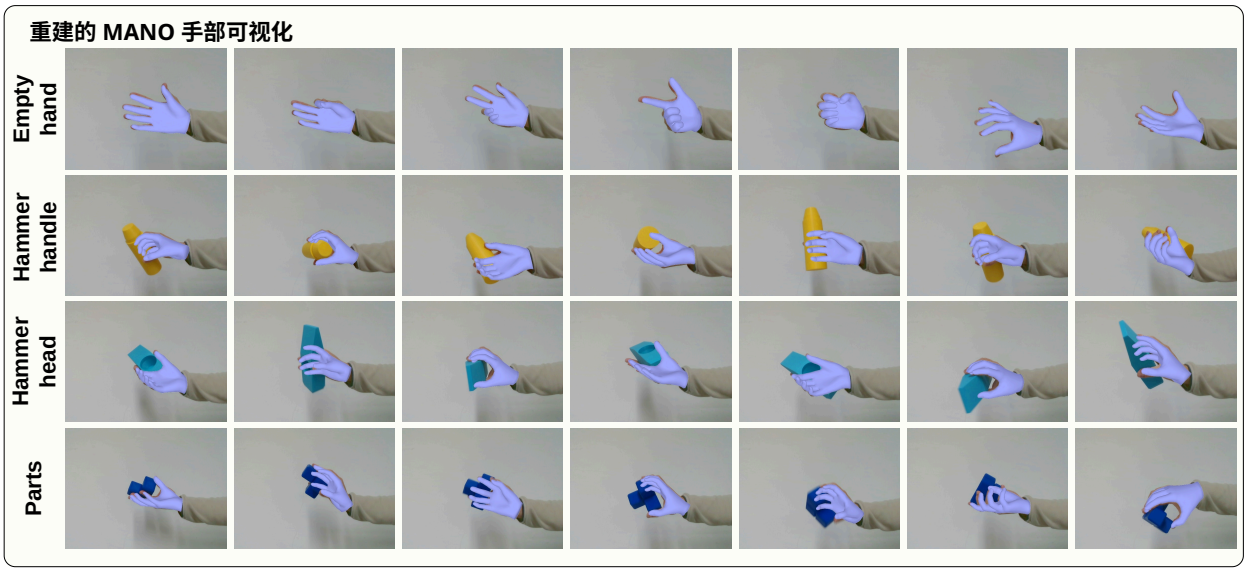

图2 在不同遮挡情况和手势下，基于 MANO 的手部姿态估计具有鲁棒性。该 MANO 手部模型由 RGB 图像重建，并通过叠加在原始 RGB 图像上进行可视化。

3.1 从视觉演示中提取手-物体运动先验

本节介绍旨在从视觉演示中提取并优化运动轨迹的数据处理流程。

3.1.1 人类演示采集

使用单个深度相机采集人类双臂演示，该相机可提供同步的 RGB-D 数据流。相机经过预先标定，以获得其内参矩阵$\mathbf{K}$ 和外参 $[\mathbf{R}|\mathbf{t}]$。数据录制过程中，操作员手动标注与任务关键阶段（如抓取、抬升、插入）对应的关键帧，用于指导后续学习阶段。

3.1.2 手部姿态估计

为表征双手的运动，我们使用基于 MANO 的手部估计器 [16, 23, 24] 从每个 RGB 帧估计三维手部姿态，得到 MANO [29] 参数 $\{\theta, \beta, \mathbf{R}^h, \mathbf{t}^h\}$，其中$\theta$ 表示局部手部关节姿态， β 表示手部形状，$\mathbf{R}^h \in \mathcal{SO}(3)$ 表示全局手部方向，$\mathbf{t}^h \in \mathbb{R}^3$ 表示全局手部平移；此外，还需估计相机内参矩阵 $\hat{}\mathbf{K}$ 并检测二维手部关节 $\mathbf{J}^{2d}$。从图 2可以看出，基于 MANO 的手部姿态估计器在不同场景下都具有较强的鲁棒性，尤其是对于人手大部分被手持物体遮挡的情况。此过程分别独立应用于左手和右手。然而，重建得到的三维 MANO 手部模型是在估计的相机

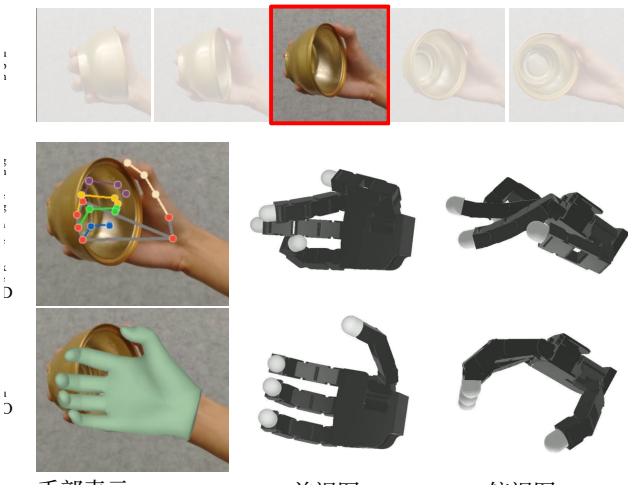

图 3 我们提出的算法能够保留完整手部姿态的优势：将我们开发的基于 MANO 的动作重定向算法与最先进的基于二维关键点的重定向方法进行比较。

ˆ内参 $\mathbf{K}$下定义的，这些内参可能与我们实际设置中标定相机的内参 $\mathbf{K}$ 不同。为解决这一差异，我们进一步通过将三维 MANO 手部关节反向投影到当前 RGB 视频帧上，并最小化投影后的 MANO 手部关节与检测到的二维手部关节之间的 l-2 距离，使三维 MANO 手部模型与观测到的人手对齐$\mathbf{J}^{2d}$:

$$\begin{aligned} \theta, \beta, \mathbf{R}^h, \mathbf{t}^h &= \underset{\theta,\beta,\mathbf{R}^h,\mathbf{t}^h}{\arg\min} ||\hat{\mathbf{J}}^{2d} - \mathbf{J}^{2d}||_2 \\ \text{where } \hat{\mathbf{J}}^{2d} &= \mathbf{K}\hat{\mathbf{J}}^{3d}, \\ \hat{\mathbf{J}}^{3d} &= \text{MANOLayer}(\theta, \beta, \mathbf{R}^h, \mathbf{t}^h) \end{aligned} \tag{1}$$

其中 MANOLayer(∗)是可微 MANO 解释器 [9] ，接收 MANO 参数并输出二维和三维手部关节。上述手部姿态估计和对齐过程会在 MANO 参数空间中生成一系列三维手部姿态， $\mathcal{T}^{hand} = \{\tau_t^{hand}\}_{t=0}^T$，其中 $\tau_t^{hand} = (\theta_t, \beta_t, \mathbf{R}_t^h, \mathbf{t}_t^h)$。

3.1.3 物体位姿估计

首先将二维图像分割器 [5]应用于 RGB 序列，生成物体分割掩码。随后，将这些掩码与 RGB-D 帧及三维物体模型一同输入现成的三维物体姿态估计器 [40] ，以得到物体的旋转 $\mathbf{R}^o \in \mathcal{SO}(3)$ 和平移 $\mathbf{t}^o \in \mathbb{R}^3$。与手部姿态估计类似，这些初始姿态估计可能存在误差，从而不利于装配等高精度任务。我们使用任务感知精化模块来优化位姿，使其符合任务专用目标函数，从而减少这些误差， f_{task} （见第 4.2节）：

$$\mathbf{R}^o, \mathbf{t}^o = \underset{\mathbf{R}^o,\mathbf{t}^o}{\arg\min}\, f_{task}(\mathbf{R}^o, \mathbf{t}^o).$$

此模块用于计算一系列物体位姿， $\mathcal{T}^{obj} = \{\tau_t^{obj}\}_{t=0}^T$，其中 $\tau_t^{obj} = (\mathbf{R}_t^o, \mathbf{t}_t^o)$。

3.1.4 基于 MANO 的手到机器人动作重定向

如图所示。3，传统的动作重定向方法 [8, 15, 27]依赖二维关键点，但当物体遮挡住手时，这些方法就会失效。基于 MANO 的表示对这类遮挡更具鲁棒性。我们将上一步得到的精细三维手部轨迹 $\mathcal{T}^{hand}$ 重定向到浮动基座机器人手。目标是估计机器人手的关节位置 q_h 以及基座位姿 p_{base} ，使其尽可能模仿 MANO 手的三维关节位置。我们通过优化求解:

$$\begin{aligned} q^h, p^{base} &= \underset{q,p}{\arg\min} ||\hat{\mathbf{J}}^{3d} - \mathbf{J}_{fk}^{3d}||_2, \\ \text{where } \mathbf{J}_{fk}^{3d} &= \text{FK}(q), \\ \hat{\mathbf{J}}^{3d} &= \text{MANOLayer}(\theta, \beta, \mathbf{R}^h, \mathbf{t}^h) \end{aligned} \tag{2}$$

FK(∗) 是正向运动学（FK）， q_h 是机器人关节位置， p_{base} 是机器人手底座的三维位姿。

算法1 示范回放与分割

$\mathcal{T}^{hand} = \{ q_t^h \ p_t^{base} \}_{t=0}^T$: **输入:** 处理后的手部轨迹（，）: **输入:** 处理后的物体轨迹 $\mathcal{T}^{obj} = \{(\mathbf{R}_t^o, \mathbf{t}_t^o)\}_{t=0}^T$
输入: 关键帧索引集合 $\mathcal{K} = \{k_1, k_2, \ldots, k_N\}$
输出: 时间片段集合 $\mathcal{S} = \{S_1, S_2, \ldots, S_N\}$
初始化回放缓冲区 $\mathcal{B} \leftarrow \emptyset$
初始化上一个关键帧索引 $t_{prev} \leftarrow 0$
对于 $t = 0$ 到 T，**执行**
    获取当前手掌基座位姿 p_t^{base}
    求解手臂关节: $q_t^{arm} \leftarrow$逆运动学(p_t^{base})
    构建全身构型: $q_t \leftarrow (q_t^{arm}, q_t^h)$
    将当前机器人状态添加到缓冲区: $\mathcal{B} \leftarrow \mathcal{B} \cup \{q_t\}$
    如果 $t \in \mathcal{K}$ **则**
        为阶段定义子目标: $g \leftarrow$(R $_t^o$, $\mathbf{t}_t^o$)
        创建阶段片段: $S_{new} \leftarrow (\mathcal{B}, g)$
        将片段添加到集合: $\mathcal{S} \leftarrow \mathcal{S} \cup \{S_{new}\}$
        清空缓冲区，以便进入下一阶段: $\mathcal{B} \leftarrow \emptyset$
        更新上一个关键帧索引 $t_{prev} \leftarrow t$
    结束 if
结束 for

3.1.5 真实到仿真全身轨迹生成

动作重定向过程仅为机器人手部生成动作，而不会为手臂生成动作。为了为完整的双臂机器人系统生成一条完整且运动学有效的轨迹，我们在 IsaacLab 模拟器中进行基于逆运动学的轨迹生成 [21]。在示范的每个时间步 t ，我们拥有

目标机器人手部的关节位置 q_t^h 以及目标手掌基座位姿 p_t^{base}（来自动作重定向模块），以及经过优化的物体位姿 （$\mathbf{R}_t^o$、$\mathbf{t}_t^o$）。为了确定所需的手臂运动，我们将手掌基座位姿 p_t^{base}作为机器人手臂的末端执行器目标，并使用逆运动学（IK）求解手臂关节位置 q_t^{arm}。于是，该时间步的全身机器人构型为 $q_t = \{q_t^{arm}, q_t^h\}$。随后，我们利用数据采集期间手动标注的关键帧（例如抓取、抬升、插入），将连续轨迹划分为 **时间片段**。这种分段演示结构详见算法 1。重放完整个示范后，我们得到一系列 N 时间片段，$\mathcal{S} = \{S_1, S_2, \ldots, S_N\}$每个片段代表操作任务中一个具有明确定义目标的独立阶段。

3.2 通过残差强化学习习得双手灵巧操作技能

以下各节介绍的数据处理流程会为双臂系统生成一条完整且运动学有效的轨迹 $\{q^t\}_{t=0}^T$。这些轨迹是强有力的运动先验，能够捕捉示范任务的整体策略和准动力学可行运动，但其本身并不完美，无法完全遵循物理规律，因此不能直接执行。

仍有两个主要挑战有待解决：

- **误差累积**。该轨迹是分层流程（位姿估计 $\rightarrow$对齐 $\rightarrow$ 动作重定向 $\rightarrow$ 逆运动学）的输出。各阶段产生的小误差可能不断累积，导致最终轨迹存在运动学误差。

- **模态差距**。通过 RGB-D 数据流采集的视觉演示缺少关于手部与物体之间实际物理接触和动力学的信息，而这些信息对于涉及大量接触的物体操作任务至关重要。

因此，生成的轨迹无法保证机器人手能够与物体建立并保持正确且稳定的接触，而这对成功操作至关重要。我们没有舍弃这一宝贵先验，而是设计学习模块来显式利用它。核心思路是学习一种纠正性残差策略，在局部修正不完美的示范，而不是从头学习控制策略。这样便将一个高维、探索需求繁重的强化学习问题，转化为更易处理、侧重局部调整的问题。为此，我们将任务构建为**残差强化学习**。

定义一个由元组 $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$ 指定的标准马尔可夫决策过程（MDP）。时间步 t, $s_t = (q_t, \dot{q}_t, s_t^{obj}, s_t^{task}) \in \mathcal{S}$ 时的状态由机器人的关节位置 q_t 和速度 $\dot{q}_t$、被操作物体的状态 s_t^{obj} 以及其他与任务相关的状态 s_t^{task}组成。模拟器执行的动作 $a_t \in \mathcal{A}$ 是机器人低层控制器的目标关节位置。在残差强化学习框架中，该动作是来自示范的底座动作与学习得到的残差动作之和： $a_t = a_t^{demo} + \Delta a_t$。这里， a_t^{demo}是当前时间步经处理轨迹中的目标关节位置，即 $a_t^{demo} = q_t$。残差动作 Δa_t 是经过训练的神经网络策略 π_ϕ的输出，该策略学习根据当前状态进行修正： $\Delta a_t \sim \pi_\phi(\cdot|s_t)$。为明确限制强化学习智能体的探索，使其保持接近底座动作轨迹，预测的残差动作 Δa_t 被裁剪至 $[-0.25, 0.25]$ 弧度。强化学习智能体的目标是学习策略参数 ϕ，以最大化期望累积折扣奖励。详细的奖励项见第 3.2 节。

为应对长时程双手操作任务，我们提出了三种用于整合到机器人学习流程中的新设计：（1）基于时间片段的强化学习；（2）成功门控重置策略；（3）事件驱动奖励课程。

基于时间片段的强化学习。在基于单次演示的演示增强强化学习中，一个关键挑战是**时间错位**问题。要求策略以锁步、按时间索引的方式跟随完整轨迹的标准方法往往很脆弱 [28]。如果智能体的执行速度偏离演示（原因可能是动力学、扰动或初始状态不同），就可能失去同步。这会导致智能体收到与任务不同阶段相对应的错误基础动作，进而引发错误不断累积，最终导致失败。

为解决这一问题， 我们用基于时间分段的强化学习范式取代对完整轨迹的修正。A s

表2 不同环境中的奖励项和物理量。

Type	符号	描述	方程	尺度	阶段		
					伸手	抓取并抬起	*goal*
稠密奖励	r_{reach}	手的末端执行器与 物体质心之间的距离， d_{h2o}	$1 - tanh(d_{h2o}/0.3)$	1.0	✔	✔	✔
	r_{grasp}	手部指尖与 物体最近表面点之间的距离， d_{f2o}	$1 - tanh(d_{f2o}/0.05)$	2.0		✔	✔
	r_{goal}	物体关键点与 当前目标关键点之间的距离， d_{goal}	$1 - tanh(d_{goal}/0.1)$	15.0			✔
稀疏奖励	B_{reach}	物体触及奖励	$1\ if\ d_{h2o} < \delta_{reach},\ else\ 0$	50	✔		
	B_{lift}	物体抬升奖励x	$1\ if\ height_{obj} > \delta_{lift},\ else\ 0$	100		✔	
	B_{goal}	目标触及奖励x	$1\ if\ d_{goal} < \delta_{goal},\ else\ 0$	1000			✔
任务特定奖励	B_{sync}	同步目标到达奖励	$1\ if\ \lvert Time(B_{goal}^{right} - B_{goal}^{left})\rvert < w, \quad else\ 0$	2000			✔
	B_{switch}	切换阶段奖励	$1\ if\ current\ stage == switch,\ else\ 0$	2000			✔

详见第 3.1.5 节和算法 1，连续演示被划分为一系列有意义的片段 $\mathcal{S} = \{S_1, S_2, \ldots, S_N\}$，其中每个片段 $S_i = (\mathcal{B}_i,\ g_i)$ 由一个轨迹片段$\mathcal{B}_i$ 和相应的基于状态的子目标 g_i组成。片段编号还会被加入到每个时间步 s_t 的状态 t 中，以便利用当前片段的信息显式地对强化学习策略进行条件化。

策略网络最终输出层的权重和偏置 π_ϕ 均初始化为零。这些设计天然支持每个阶段采用**先学习后优化**的策略，这是缓解时间错位的关键。在训练开始时，预测的残差动作 Δa_t 接近于零。因此，智能体最初会逐步忠实地跟踪示范状态。在这一阶段，时间对齐是通过设计得以维持的。随着训练的推进，智能体的能力和信心不断增强，面向任务的奖励信号会促使其超越示范中特定的时序安排，为当前阶段发展出更通用、更高效的技能。

成功门控重置策略。一种新颖的**成功门控重置**策略使强化学习探索更加高效。这种方法通过管理初始状态的分布，在保留早期技能与深入探索后续技能之间取得平衡。当智能体因当前片段 S_i中的失败而结束一个回合时，我们不会将其确定性地重置到全局初始状态 s_0，而是采用概率重置。以 p_{init}的概率，环境会重置到全局初始状态 s_0 （作为防止遗忘的正则化项）。以$1 - p_{init}$的概率，环境会重置到前一阶段的成功终止状态 s^*_{i-1}，并以此作为新的起点。这种概率重置策略可视为一种隐式课程学习：

- 重置到 s^*_{i-1} 可提供**针对性训练**，使智能体能够反复练习其认为具有挑战性的特定阶段，而无需“付出”重新执行所有前置阶段的代价。这将显著提升掌握任务后续部分的样本效率。
- 重置到 s_0 可作为**防止遗忘的正则化项**，确保策略持续从一开始复习基础技能，从而维持稳健且全面的解决方案。

超参数 p_{init} 控制着探索智能体当前能力边界与利用已掌握技能之间的平衡。

事件驱动奖励课程。将稠密距离奖励与稀疏事件驱动奖励相结合的奖励函数为策略提供引导。这种混合结构既提供细粒度反馈，也标示关键任务里程碑的达成情况。奖励按顺序激活，形成隐式课程，引导智能体从伸手到抓取，最终学会操作。任意时间步的总奖励由稠密奖励和稀疏奖励两部分组成，如表中所列。2。绿色勾号表示在整个阶段均可获得的奖励项，红色勾号表示仅在完成特定阶段时才能获得的奖励项。除了通用奖励组成部分外，还有两项任务特定奖励：B_{sync} 用于鼓励智能体在同步双手任务中学习双臂之间的协调技能，B_{switch} 用于鼓励智能体探索并通过关键的“切换阶段”。此外，还有两种课程策略：(1) 对目标到达阈值进行阈值退火（δ_{goal}），以弥合初始探索与高精度（毫米级）接触密集型操作之间的差距；(2) 采用预抓取课程，在到达阶段保持手部张开并冻结其状态。这将初始阶段简化为仅需控制机械臂，随后再在需要时引入更复杂的手指控制问题。

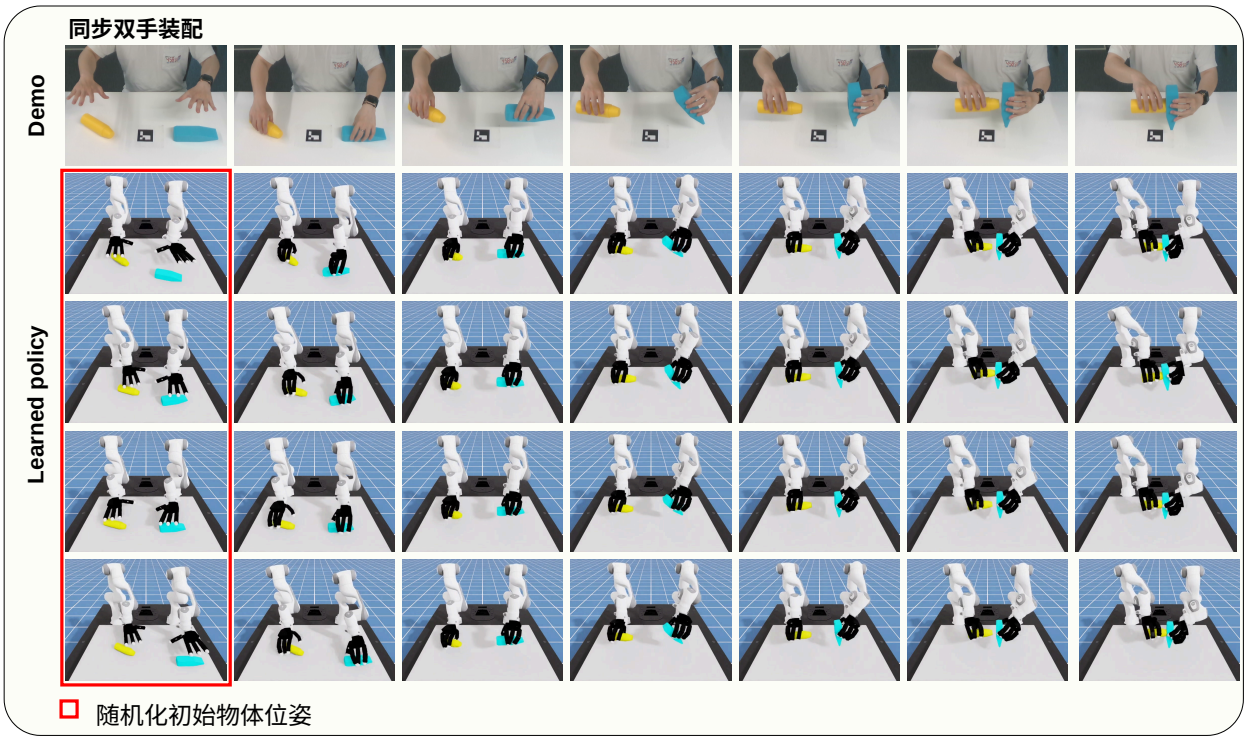

图 4 基于物理的仿真展示了从人类视频中学习到的、用于同步装配任务的双手装配技能。第一行展示采集的演示，其余各行展示使用训练好的策略得到的操作结果。值得注意的是，尽管我们的方法基于单次演示，但对物体的初始位姿进行随机化，仍能使策略泛化到随机物体初始位姿。

4 实验验证

本节首先分别介绍仿真实验和真实世界硬件实验所采用的实验设置，然后介绍核心模块的实现细节。接着，我们展示仿真结果和消融研究，以证明框架设计的有效性。最后，作为概念验证，我们报告真实环境设置下的实验结果，以证明我们方法的仿真到现实迁移能力。

4.1 实验设置

我们进行了大量仿真实验，随后开展了真实世界的概念验证演示。下面介绍仿真和真实世界实验的设置。补充视频有助于更直观地了解任务设置。

仿真环境。我们使用 IsaacLab 模拟器 [21],评估方法的关键组件，该模拟器建模了一个由两台 Franka Panda机械臂组成的双臂设置，每台机械臂均配备一个 Allegro Hand。我们评估同步双手装配和异步双手装配，以测试复杂操作的不同方面。同步 设置（5 个子目标）要求两条手臂协同抓取一个底座和一个销钉，并在空中将二者装配起来。其主要挑战是在整个轨迹中维持两条手臂之间紧密的时间与空间协调。异步 任务（11 个子目标）要求一条手臂先放置底座，然后由第二条手臂将把手插入已放置的底座中。其主要挑战在于应对显著的分布偏移，即“切换阶段”。这两种设置中的每个训练回合都会在物体的初始位姿上加入随机的平移（ $[-10, 10]$ 厘米）和旋转（ $[-0.1, 0.1]$ 弧度）变化。

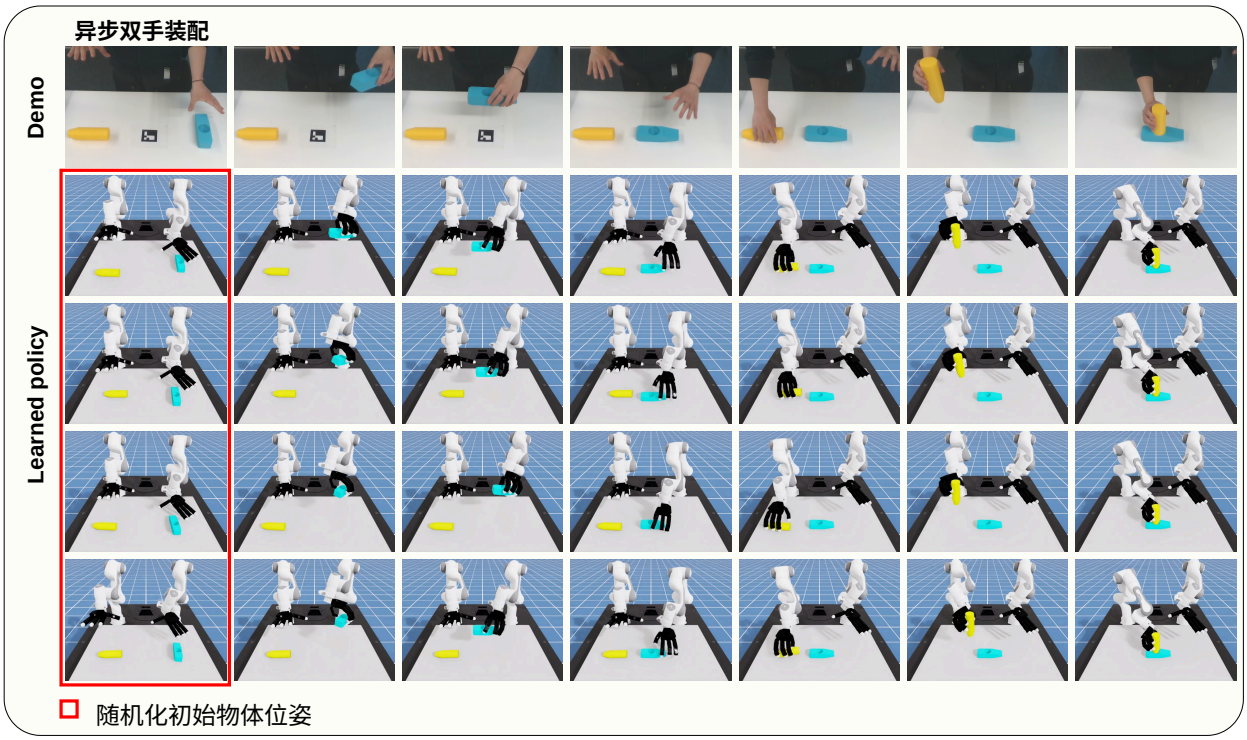

图 5 基于物理的仿真：从人类视频中学习到的双手装配技能在异步装配任务中的表现。顶行展示采集到的演示，其余各行展示使用训练好的策略得到的操作结果。值得注意的是，尽管我们的方法基于单次演示，但对物体的初始位姿进行随机化，仍能使策略泛化到随机物体初始位姿。

真实世界验证。 为展示仿真到现实的迁移能力和跨具身能力，我们将所提出的方法部署在由一台 UR-3e 机械臂和一只 XHand 灵巧手组成的实体系统上。由于硬件限制，真实世界验证仅限于单臂设置，因此我们将任务略微改写为两个变体：(1)单步装配：基准物体位于桌子上的固定位置；机器人需要抓取并插入插销物体，以完成装配；(2)多步放置与装配：机器人需要先拾取并放置基准物体，然后通过抓取并插入插销物体来完成装配任务。这些实验作为概念验证，确认了所提出的方法能够成功迁移到真实硬件。补充视频展示了更多定性结果。

4.2 实现细节

数据采集。 录制示范时，我们使用一台 RealSense D455 相机记录人类操作的同步 RGB-D 数据流。我们使用 3D 打印物体进行数据采集，随后将其 3D 模型用于物体姿态估计、优化以及物理仿真。

数据处理模块。 我们使用 WiLoR [24] 进行手部姿态估计，使用 Segmentation-and-Track Anything [5]进行物体掩码分割，并使用 FoundationPose [40] 在数据处理模块和真实机器人实验中进行三维物体姿态跟踪与估计。对于所提出的装配任务（机器人将插销状物体插入孔中），任务专用目标函数 f_{task} 是以下两项的加权和：(1) 插销-孔轴线共线损失；(2) 插销-孔端点相对位置损失。

残差强化学习模块。 物体关键点用于表示物体状态。对于圆柱体等轴对称物体，使用其主轴上的三个点作为关键点；否则，使用旋转边界框

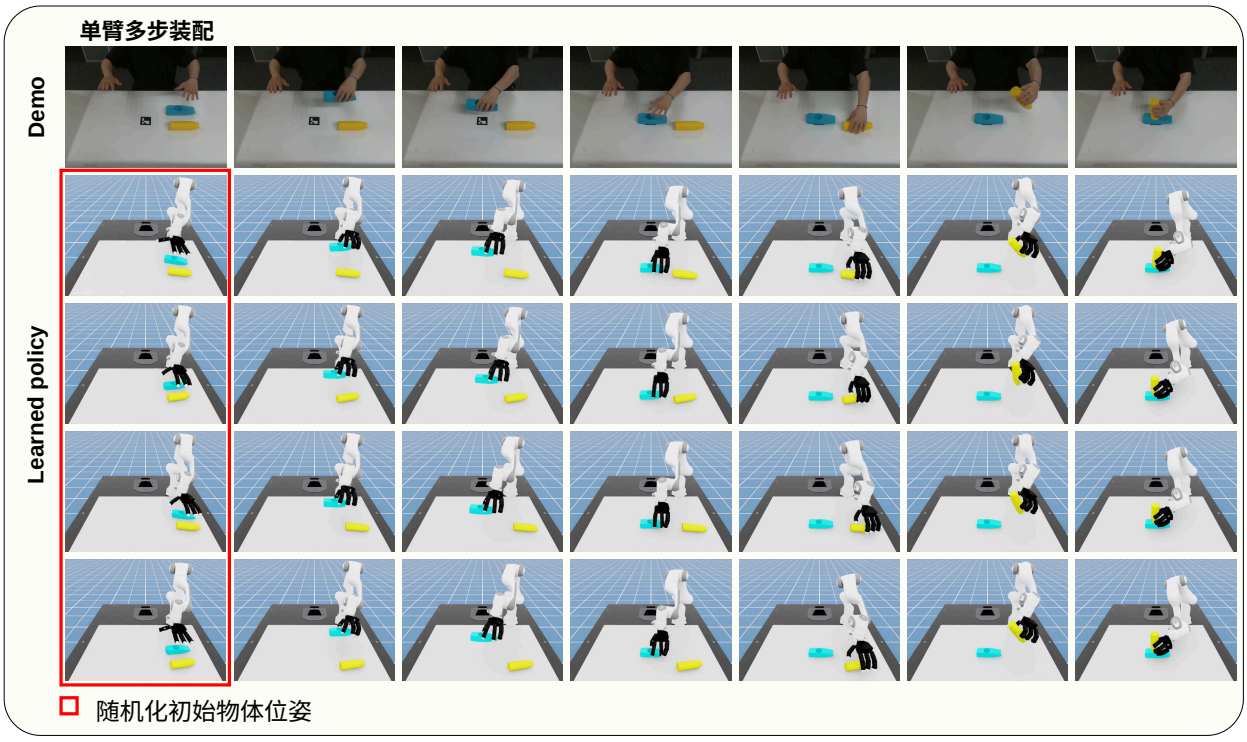

图6 从人类视频中学习到的双手装配技能在多步单臂装配任务中的基于物理的仿真。顶行展示采集到的演示，其余各行展示使用训练好的策略得到的操作结果。请注意，尽管我们的方法基于单次演示，但对物体初始位姿进行随机化仍可实现对随机物体初始位姿的泛化。

的角点作为关键点。我们通过实验发现， $p_{init} \in [0.90, 0.95]$ 是成功门控重置策略的一个理想取值。我们使用 RSL [31] 强化学习库中的 PPO_算法 [30] 来训练强化学习策略。我们通过使用 Optuna [1]进行参数优化来确定 PPO 的超参数。实验中使用的 PPO 核心超参数列于3。

用于仿真到现实迁移的基于物理的驱动动力学。为了弥合真实机器人上的非线性电动机与仿真中理想执行器之间的差距，我们设计了一种随机化的基于物理的驱动模型，将基本的电机动力学和系统性的参数变化直接嵌入仿真中的低层关节控制回路。这样的随机化使强化学习策略能够直接部署到真实机器人上。物理仿真中的默认PD控制器使用理想扭矩源来模拟关节级位置控制，这也是仿真中的传统PD控制。在这种情况下，扭矩源是完美的，模拟执行器的行为无法体现真实电机及其齿轮传动系统的详细电气特性和动力学。

表3 PPO核心超参数

参数	值
num_步骤_per_env	24
经验 _归一化	True
值 _归一化	True
值 _loss_coef	0.5
use_裁剪 _值 _loss	True
clip_参数	0.2
熵 _coef	5.0e-5
num_学习 _训练周期	4
num_mini_批次	32
学习 _rate	5.0e-4
计划	自适应
伽马	0.99
lam	0.95
期望的_kl	0.008
max_grad_norm	1.0

不同于传统的理想PD控制仿真方式，我们的建模明确考虑了真实执行器的三个关键方面：（1）增益变化和校准误差，（2）与速度相关的扭矩饱和，以及（3）传感器测量偏差。执行器动力学通过物理信息方程表示，这些方程捕捉了基本的

电机约束。

期望扭矩的计算方式如下：

$$\tau_{\text{des}} = K_p' \cdot (q_{\text{des}} - (\hat{q} + b)) + K_d' \cdot (\dot{q}_{\text{des}} - \dot{q}),$$

其中，$K_p' = \alpha_p K_p$ 和 $K_d' = \alpha_d K_d$ 表示随机化的刚度和阻尼增益，α_p，$\alpha_d \in \mathcal{U}(0。9，1。1)$。项 $b \in \mathcal{U}(-0.1，0。1)$ 弧度表示恒定的位置测量偏差，而 $\hat{q}$ 表示偏置关节位置。

为模拟真实的电机扭矩限制，我们采用依赖速度的饱和模型，以反映直流电机的扭矩-速度特性：

$$\tau_{\max}(\omega) = \frac{\tau_{\text{stall}}}{1-\nu}\left(1 - \frac{|\omega|}{\omega_{\max}}\right)，\quad \tau_{\min}(\omega) = \frac{\tau_{\text{stall}}}{1-\nu}\left(-1 - \frac{|\omega|}{\omega_{\max}}\right)$$

其中，τ_{stall} 是堵转扭矩，ω 是关节角速度，$\omega_{\max}$ 是空载角速度。参数 $\nu \in \mathcal{U}(13，23)$ 表示扭矩开始饱和时的归一化速度，用于反映不同执行器在电机常数和电气特性方面的差异。最终施加的扭矩计算为 $\tau_{\text{applied}} = \gamma \cdot \text{clip}(\tau_{\text{des}}，\tau_{\min}(\omega)\ \tau_{\max}(\omega))$，其中 $\gamma \in \mathcal{U}(0。9，1。1)$ 引入了整体电机强度的变化，以反映电机常数和放大器增益的差异。

训练期间，每个回合开始时，都会为每个关节独立重新采样所有随机化参数（α_p、α_d、b、ν、γ）。

4.3 仿真实验与消融研究

本节介绍仿真实验结果，并评估关键设计选择的有效性，包括：（1）运动先验：比较基于演示引导的强化学习与从零开始学习；（2）残差动作裁剪：考察将策略限制在参考轨迹附近对训练稳定性的重要性；（3）预抓取课程：评估我们专门设计的伸手

表4 仅使用提取的运动先验、从头开始训练的强化学习，以及我们提出的强化学习流程（运动先验 +强化学习）所达到的子目标数量。

	同步任务	异步任务
仅运动先验	0/5	0/11
仅强化学习	0/5	1/11
运动先验+强化学习（**我们的方法**）	**5/5**	**11/11**

阶段是否有助于实现更稳健、稳定的抓取；（4）基于时间分段的强化学习：比较将任务分段与基于完整轨迹学习对长时域性能的影响；（5）成功门控重置策略：分析所提出的重置机制对探索效率和训练稳定性的影响。图中展示了仿真中的同步与异步装配任务以及多步骤单臂装配任务的快照。4、5、6 分别展示了这些内容。

我们对同步和异步装配任务都进行了消融研究。所有实验均使用不同的随机种子重复五次。表4 中的结果表明，无论是重放提取的运动先验，还是从头开始训练强化学习智能体，都无法完成任务。相应的学习曲线见图7。实线表示所有运行结果的均值，阴影区域表示标准差。完整方法（红色）在两项任务中均实现了最高性能和样本效率，并成功完成所有子目标。下面将详细分析各个有助于实现这一最佳总体性能的组成部分。

演示引导和基于时间分段的强化学习的有效性。无论是从头开始学习（曲线 无运动先验），还是从单条未分段的轨迹中学习（曲线 无时间分段），都会导致两个任务都彻底失败。这归因于高维状态-动作空间，在其中进行随机探索难以实现。此外，从一整条长轨迹中学习会造成严重的时间错位；智能体会接收到来自不同任务阶段、毫无意义的基础动作，从而污染学习信号并阻碍进展。这些结果证实，将分段演示用作运动先验是我们核心方法不可或缺的一环。

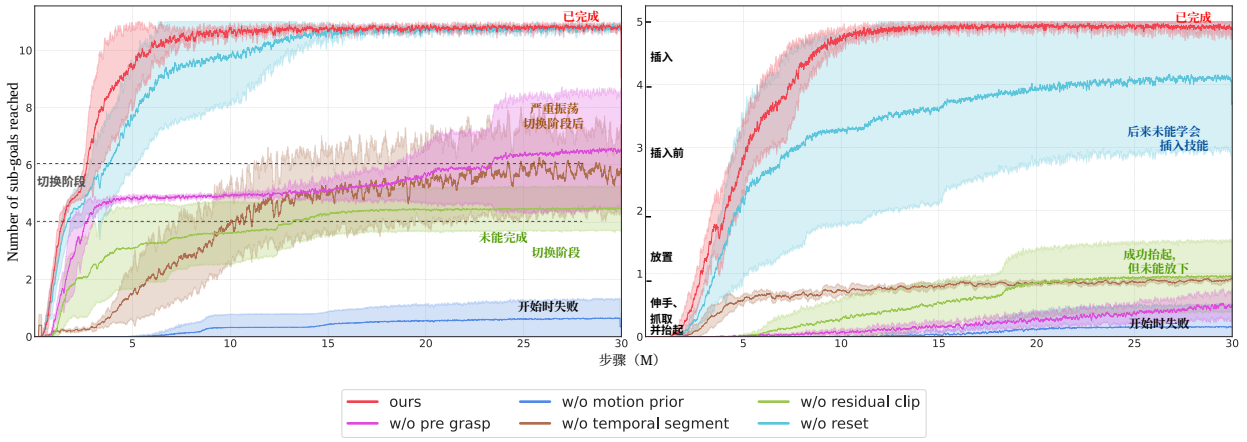

图7 异步（左）和同步（右）双手装配任务上的定量比较与消融结果。图中展示了不同消融版本的所提算法所达到的子目标平均数量（纵轴）与训练步数之间的关系。

残差动作裁剪的有效性。禁用残差裁剪（曲线 无残差裁剪）会使智能体偏离稳定运动先验过远，从而降低两个任务的表现。这种影响在同步任务中更为明显，其策略的表现停滞在低得多的水平。完成该任务所需的精细协调几乎没有容错空间，而无约束探索产生的大幅、无规律动作很容易破坏这种协调。

预抓取课程的有效性。移除预抓取（曲线 无预抓取）会导致同步任务彻底失败。双手稳定地同时抓取是同步任务成功的先决条件；从一开始就同时学习手臂与手指控制，会导致伸手行为不稳定并立即失败。在异步任务中，尽管学习初期受到显著阻碍，智能体最终仍能恢复。这表明，预抓取课程始终有益，但在协调约束严格的任务中最为关键。

成功门控重置的有效性。移除重置策略（曲线 无重置）揭示了长时程学习中的关键瓶颈。尽管异步任务中的智能体最终能够成功，但其切换后阶段的学习率显著降低。更具挑战性的同步任务则彻底失败：智能体很早便陷入平台期，始终无法掌握最后的协调阶段。同步任务更严格的成功条件放大了数据不平衡问题；如果没有重置策略提供针对性练习，智能体无法从少数几次随机进入这些困难状态的机会中学习。因此，成功门控重置对于高效探索和掌握复杂的序列任务至关重要。

4.4 真实世界实验

为验证所提出框架的仿真到现实的迁移能力和跨具身能力，我们在第4.1节描述的任务上开展了真实世界实验。我们使用与仿真实验完全相同的流程收集并处理单次演示，但将收集到的演示重新定向后，在真实世界机器人配置（UR-3e +XHand）上重放。在策略训练期间，为弥合真实电机与仿真中理想电机之间的差距，我们采用了额外的基于物理的驱动随机化，详见第4.2节。

我们对单步装配任务进行了 20 次试验，并改变初始物体状态，以评估训练好的智能体的成功率。这些实验的快照见图8。智能体在 20 次运行中成功完成了 18 次，成功率为 90%。而在多步装配任务中，智能体成功完成了 12 次，成功率为 60%。在真实世界实验中，失败主要是由于物体被机器人手遮挡时，物体位姿估计发生漂移。这种估计误差对多步装配任务的影响更为显著，因为基准物体放置错误会严重影响

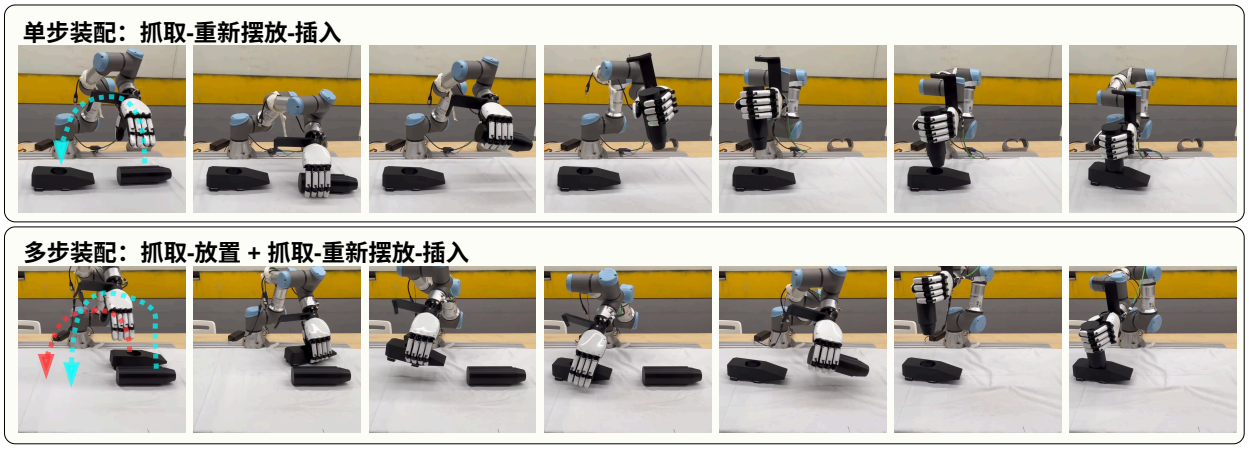

图8 真实机器人实验：（上）通过依次执行抓取-重新摆放-插入进行单步装配；（下）通过依次抓取-放置基准物体，再抓取-重新摆放-插入另一物体进行多步装配。

后续插入阶段的结果。估计误差会导致输入强化学习智能体的状态不准确，使其无法将物体重新定位到正确的插入位姿。总体而言，真实世界实验成功展示了所提出的 DemoBot 框架的仿真到现实迁移能力和跨具身能力，彰显了其对机器人学界的贡献。

5 讨论

本技术报告介绍了 DemoBot，这是一种新颖的框架，可利用单个 RGB-D视频高效学习复杂的灵巧双手操作技能。主要贡献包括一个全面的数据处理模块，可将非结构化视频转换为用于机器人学习的结构化运动先验；以及一种新颖的纠正式残差强化学习流程，其中包含一系列创新设计，以应对学习长时程灵巧双手操作技能所面临的挑战。尽管该方法具有显著优势，但其局限性包括：任务分解依赖人工选定的关键帧，物体位姿估计依赖深度信息和3D 打印物体，以及物体位姿精化依赖人工指定的任务目标。不过，我们认为，该方法为从非结构化人类视频中学习复杂操作技能提供了一条很有前景的途径。未来工作将着力克服这些局限，并探索如何将该方法扩展到更多样、甚至不完美的示范，从而逐步实现基于野外环境视频的可扩展操作技能学习。

6 未来工作

尽管所提出的 DemoBot 在从原始人类视频中学习灵巧操作技能方面迈出了重要一步，但为简化这一具有挑战性的研究问题，我们作出了一些核心松弛处理：（1）在示范过程中使用 RGB-D 相机采集 RGB 和深度信息；（2）使用现有的三维物体模型进行物体位姿估计，并在仿真中进行策略学习；（3）需要人工标注，从原始视频中选取关键帧；（4）需要进行任务特定优化，以优化原始估计的物体位姿。此外，目前使用 6D 位姿表示物体状态，但这种表示仅适用于刚性物体。

未来工作需要社区共同努力来消除这些松弛项。结合深度估计模型 [42, 43] 和 3D资产生成模型 [35, 36]，有望将当前框架从 RGB-D视频扩展到任意纯RGB视频，并摆脱对现有 3D物体模型的依赖，从而消除松弛项 (1) 和 (2)。视觉语言模型（VLMs）可用于通过视觉理解序列任务，并选择与各任务阶段相对应的关键帧。此外，近期研究 [17] 表明，大语言模型（LLMs）有望自动生成

用于强化学习训练的奖励函数。因此，将 VLMs 纳入当前框架即可消除松弛项 (3) 和 (4)。

此外，未来的工作可以着眼于将当前框架扩展到更多类型的被操作物体，例如关节物体和可变形物体，并采用更通用的物体状态表示方法。例如，可以结合6D位姿与关键点：6D位姿可用于粗略表示整个物体当前的姿态，而关键点则表示物体的形变或关节结构。

7 结论

总而言之，本研究率先提出了一种有效框架，可从任意互联网视频中学习双手灵巧操作技能，并通过仿真和真实环境设置中的大量实验加以验证。该框架展示了其在系统级工程设计与集成方面的有效性，运用了从基于视觉的手部重建、物体姿态估计到新颖的演示增强残差强化学习等多种技术。

此外，这项工作为实现可扩展的机器人学习迈出了重要一步。DemoBot 从单个人类视频中提取运动先验，即使该视频并不完美，也能绕过遥操作带来的数据瓶颈。我们的结果表明，将次优运动先验与残差强化学习相结合，能够使机器人掌握接触密集型长时域任务——由于探索复杂性，这类问题难以通过从零开始的强化学习解决；而由于数据需求高昂，也难以通过模仿学习解决。这一范式为利用互联网规模的视频数据打开了大门，使我们离通用机器人更近一步：它们只需通过“观看”人类，就能习得新的灵巧技能。

8 贡献与致谢

以下贡献按技术投入和研究投入分类，并在每个类别中列出作者姓名。

- 研究构想与概念化：Yucheng Xu、Robert B. Fisher、Zhibin Li。
- 数据收集与处理模块：Yucheng Xu。
- 残差强化学习模块：Yucheng Xu、Xiaofeng Mao、Elle Miller。
- 仿真实验：Yucheng Xu、Xiaofeng Mao、Elle Miller。
- 真实机器人实验：Xinyu Yi、Yang Li、Zhibin Li。
- 硬件与软件基础设施：Yucheng Xu、Xinyu Yi、Yang Li。
- 论文撰写：全体作者。
- 研究方向与团队负责人：Robert B. Fisher、Zhibin Li。

我们衷心感谢 Jianran Liu、Yufei Liu 和 Lisuo Li 在硬件搭建方面给予的大力支持。我们诚挚感谢 Wenjia Zhu 鼓励我们专注于高影响力研究。我们也由衷感谢 Yonghui Wu 凭借其战略愿景帮助我们确定研究重点。此外，我们谨此感谢系主任 Hang Li 在整个项目期间给予的宝贵组织支持。

参考文献

[1] Takuya Akiba、Shotaro Sano、Toshihiko Yanase、Takeru Ohta 和 Masanori Koyama。《Optuna：新一代超参数优化框架》。载于第25届ACM SIGKDD知识发现与数据挖掘国际会议，第2623–2631页，2019年。[2] Yuanpei Chen、Yiran Geng、Fangwei Zhong、Jiaming Ji、Jiechuang Jiang、Zongqing Lu、Hao Dong 和 Yaodong Yang。《Bi-dexhands：迈向人类水平的双手灵巧操作》。IEEE模式分析与机器智能汇刊，46(5)：2804–2818，2023年。[3] Yuanpei Chen、Chen Wang、Yaodong Yang 和 C Karen Liu。《基于人体运动数据的以物体为中心的灵巧操作》。arXiv:2411.04005，2024年。[4] Xuxin Cheng、Jialong Li、Shiqi Yang、Ge Yang 和 Xiaolong Wang。《Open-television：采用沉浸式主动视觉反馈的遥操作》。arXiv:2407.01512，2024年。[5] Yangming Cheng、Liulei Li、Yuanyou Xu、Xiaodi Li、Zongxin Yang、Wenguan Wang 和 Yi Yang。《分割并跟踪任意物体》。arXiv:2305.06558，2023年。[6] Runyu Ding、Yuzhe Qin、Jiyue Zhu、Chengzhe Jia、Shiqi Yang、Ruihan Yang、Xiaojuan Qi 和 Xiaolong Wang。《Bunny-visionpro：用于模仿学习的实时双手灵巧遥操作》。arXiv:2407.03162，2024年。[7] Abhishek Gupta、Vikash Kumar、Corey Lynch、Sergey Levine 和 Karol Hausman。《Relay策略学习：通过模仿学习和强化学习解决长时域任务》。arXiv预印本 arXiv:1910.11956，2019年。[8] Ankur Handa、Karl Van Wyk、Wei Yang、Jacky Liang、Yu-Wei Chao、Qian Wan、Stan Birchfield、Nathan Ratliff 和 Dieter Fox。《Dexpilot：基于视觉的灵巧机械手臂系统遥操作》。载于2020年IEEE机器人与自动化国际会议（ICRA），第9164–9170页，2020年。[9] Yana Hasson、Gül Varol、Dimitris Tzionas、Igor Kalevatykh、Michael J. Black、Ivan Laptev 和 Cordelia Schmid。《学习手部与被操作物体的联合重建》。载于计算机视觉与模式识别会议，2019年。[10] Todd Hester、Matej Vecerik、Olivier Pietquin、Marc Lanctot、Tom Schaul、Bilal Piot、Dan Horgan、John Quan、Andrew Sendonaris、Ian Osband 等。《基于示范的深度Q学习》。载于AAAI人工智能大会，第32卷，2018年。[11] Jonathan Ho 和 Stefano Ermon。《生成对抗模仿学习》。载于《神经信息处理系统进展》，第29卷，2016年。[12] William Huey、Huaxiaoyue Wang、Anne Wu、Yoav Artzi 和 Sanjiban Choudhury。《从单个时间错位视频中进行模仿学习》。arXiv预印本 arXiv:2502.05397，2025年。[13] Tobias Johannink、Shikhar Bahl、Ashvin Nair、Jianlan Luo、Avinash Kumar、Matthias Loskyll、Juan Aparicio Ojea、Eugen Solowjow 和 Sergey Levine。《用于机器人控制的残差强化学习》。载于2019年国际机器人与自动化会议（ICRA），第6023–6029页。IEEE，2019年。[14] Kailin Li、Puhao Li、Tengyu Liu、Yuyang Li 和 Siyuan Huang。《ManipTrans：通过残差学习实现高效的灵巧双手操作迁移》。载于计算机视觉与模式识别会议论文集，第6991–7003页，2025年。[15] Shuang Li、Xiaojian Ma、Hongzhuo Liang、Michael Görner、Philipp Ruppel、Bin Fang、Fuchun Sun 和 Jianwei Zhang。《使用端到端深度神经网络对Shadow灵巧手进行基于视觉的遥操作》。载于2019年国际机器人与自动化会议（ICRA），第416–422页。IEEE，2019年。[16] Kevin Lin、Lijuan Wang 和 Zicheng Liu。《使用Transformer进行端到端人体姿态与网格重建》。载于计算机视觉与模式识别会议，2021年。[17] Yecheng Jason Ma、William Liang、Guanzhi Wang、De-An Huang、Osbert Bastani、Dinesh Jayaraman、Yuke Zhu、Linxi Fan 和 Anima Anandkumar。《Eureka：通过编写大语言模型代码实现人类水平的奖励设计》。arXiv预印本 arXiv:2310.12931，2023年。[18] Xiaofeng Mao、Yucheng Xu、Ruoshi Wen、Mohammadreza Kasaei、Wanming Yu、Efi Psomopoulou、Nathan F Lepora 和 Zhibin Li。《利用真实示范数据中的触觉感知学习精细的捏取技能》。CoRR，2023年。

[19] Xiaofeng Mao、Gabriele Giudici、Claudio Coppola、Kaspar Althoefer、Ildar Farkhatdinov、Zhibin Li 和 Lorenzo Jamone。Dexskills：利用触觉数据进行技能分段，以学习自主长时程机器人操作任务。发表于 IEEE/RSJ智能机器人与系统国际会议，2024年。[20] Robert McCarthy、Daniel CH Tan、Dominik Schmidt、Fernando Acero、Nathan Herr、Yilun Du、Thomas G Thuruthel 和 Zhibin Li。迈向从互联网视频中进行通用机器人学习：一项综述。《人工智能研究杂志》，83，2025年。[21] Mayank Mittal、Calvin Yu、Qinxi Yu、Jingzhou Liu、Nikita Rudin、David Hoeller、Jia Lin Yuan、Ritvik Singh、Yunrong Guo、Hammad Mazhar、Ajay Mandlekar、Buck Babich、Gavriel State、Marco Hutter 和 Animesh Garg。Orbit：面向交互式机器人学习环境的统一仿真框架。IEEE机器人与自动化快报，8(6):3740–3747，2023年。doi：10.1109/LRA.2023.3270034。[22] Tianwei Ni、Michel Ma、Benjamin Eysenbach 和 Pierre-Luc Bacon。Transformer何时在强化学习中表现出色？将记忆与信用分配解耦。神经信息处理系统进展，36:50429–50452，2023年。[23] Georgios Pavlakos、Dandan Shan、Ilija Radosavovic、Angjoo Kanazawa、David Fouhey 和 Jitendra Malik。利用Transformer重建三维手部。发表于计算机视觉与模式识别会议，2024年。[24] Rolandos Alexandros Potamias、Jinglei Zhang、Jiankang Deng 和 Stefanos Zafeiriou。WiLoR：野外环境下端到端的三维手部定位与重建。发表于计算机视觉与模式识别会议论文集，第12242–12254页，2025年。[25] Yuzhe Qin、Yueh-Hua Wu、Shaowei Liu、Hanwen Jiang、Ruihan Yang、Yang Fu 和 Xiaolong Wang。DexMV：从人类视频中进行灵巧操作的模仿学习，2021年。[26] Yuzhe Qin、Hao Su 和 Xiaolong Wang。从一只手到多只手：基于单摄像头遥操作的灵巧操作模仿学习。IEEE机器人与自动化快报，7(4):10873–10881，2022年。[27] Yuzhe Qin、Wei Yang、Binghao Huang、Karl Van Wyk、Hao Su、Xiaolong Wang、Yu-Wei Chao 和 Dieter Fox。Anyteleop：一种通用的基于视觉的灵巧机器人臂手遥操作系统。发表于机器人学：科学与系统，2023年。[28] Aravind Rajeswaran、Vikash Kumar、Abhishek Gupta、Giulia Vezzani、John Schulman、Emanuel Todorov 和 Sergey Levine。利用深度强化学习和示范学习复杂的灵巧操作。arXiv:1709.10087，2017年。[29] Javier Romero、Dimitrios Tzionas 和 Michael J. Black。具身化的手：手与身体的联合建模与捕捉。ACM图形学汇刊（亚洲计算机图形学国际会议论文集），36(6)，2017年11月。[30] Nikita Rudin、David Hoeller、Philipp Reist 和 Marco Hutter。利用大规模并行深度强化学习，在数分钟内学会行走。发表于第五届机器人学习会议，机器学习研究会议录第164卷，第91–100页。PMLR，2022年。网址：https://proceedings.mlr.press/ v164/rudin22a.html。[31] 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年。[32] Kenneth Shaw、Yulong Li、Jiahui Yang、Mohan Kumar Srirama、Ray Liu、Haoyu Xiong、Russell Mendonca 和 Deepak Pathak。面向复杂任务的双手灵巧操作。arXiv:2411.13677，2024年。[33] Tom Silver、Kelsey Allen、Josh Tenenbaum 和 Leslie Kaelbling。残差策略学习。arXiv:1812 06298，2018年。[34] Tony Tao、Mohan Kumar Srirama、Jason Jingzhou Liu、Kenneth Shaw 和 Deepak Pathak。Dexwild：用于野外环境机器人策略的人类灵巧交互。arXiv:2505.07813，2025年。[35] 腾讯混元3D团队。混元3D 2.0：扩展扩散模型以生成高分辨率纹理三维资产，2025年。[36] 腾讯混元3D团队。混元3D 2.5：迈向细节极致的高保真三维资产生成，2025年。网址：https://arxiv.org/abs/2506.16504。[37] Eleftherios Triantafyllidis、Fernando Acero、Zhaocheng Liu 和 Zhibin Li。利用机器人操作网络 ROMAN 的混合分层学习解决复杂序列任务。《自然·机器智能》，5(9):991–1005，2023年。

[38] Mel Vecerik、Todd Hester、Jonathan Scholz、Fumin Wang、Olivier Pietquin、Bilal Piot、Nicolas Heess、Thomas Rothörl、Thomas Lampe 和 Martin Riedmiller。利用示范，在稀疏奖励的机器人问题上进行深度强化学习。arXiv:1707.08817，2017年。[39] Chen Wang、Haochen Shi、Weizhuo Wang、Ruohan Zhang、李飞飞和 C Karen Liu。DexCap：可扩展、便携的灵巧操作动作捕捉数据采集系统。arXiv:2403.07788，2024年。[40] Bowen Wen、Wei Yang、Jan Kautz 和 Stan Birchfield。FoundationPose：新物体的统一6D位姿估计与跟踪。发表于IEEE/CVF计算机视觉与模式识别会议论文集，第17868–17879页，2024年。[41] Haoyu Xiong、Haoyuan Fu、Jieyi Zhang、Chen Bao、Qiang Zhang、Yongxi Huang、Wenqiang Xu、Animesh Garg 和 Cewu Lu。Robotube：利用仿真孪生环境从人类视频中学习家庭环境操作。发表于第六届机器人学习年会，2022年。[42] Lihe Yang、Bingyi Kang、Zilong Huang、Xiaogang Xu、Jiashi Feng 和 Hengshuang Zhao。Depth Anything：释放大规模无标注数据的力量。发表于计算机视觉与模式识别会议，2024年。[43] Lihe Yang、Bingyi Kang、Zilong Huang、Zhen Zhao、Xiaogang Xu、Jiashi Feng 和 Hengshuang Zhao。Depth Anything V2。arXiv:2406.09414，2024年。[44] Han Zhang、Songbo Hu、Zhecheng Yuan 和 Huazhe Xu。Doglove：利用低成本开源触觉力反馈手套实现灵巧操作。arXiv:2502.07730，2025年。[45] Huayi Zhou、Ruixiang Wang、Yunxin Tai、Yueci Deng、Guiliang Liu 和 Kui Jia。只教一次：从视频示范中学习单次示范双臂机器人操作。arXiv:2501.14208，2025年。