arXiv:2611.31115v1 [cs.RO] 23 Nov 2026

RUIBench：什么样的机器人使用接口才算好?

Yanzhe Chen* Zhijun Cao* Zechen Bai* Wenzheng Zeng* Kevin Yuchen Ma YiqiLin GuoqiangLiang Xiaokang Liu Ke Sun Kevin Qinghong Lin Mike Zheng Shou†

新加坡国立大学 Show Lab

摘要

基础模型能够对复杂的物理任务进行推理，但要将这种智能转化为可靠的机器人执行仍然面临挑战。我们认为，一个关键却常被忽视的因素是**机器人使用接口（RUI）**，它决定了智能体如何表达并执行物理动作。为系统地研究其影响，我们推出**RUIBench**，其中包含 30 项操作任务，分属五个套件，涵盖世界知识、上下文学习、发现、动态交互和高接触操作。我们进一步开发了**RUIKit**这一统一框架，将六种动作抽象与两种交互形式相结合，在共享感知和控制的条件下形成 13 种接口。在仿真中，仅改变 RUI，就能使前沿智能体的成功率从**45% 提升至 81%**，并使一个经微调的 4B 模型的成功率从**13% 提升至 47%**。我们的分析表明，有效的接口**能够减少不必要的决策，同时保留必要的反馈**，其优势受智能体行为和物理不确定性的影响。我们还发现了**验证差距**：智能体经常在存在可观察失败的情况下仍宣称成功，这表明除了验证单个动作之外，还需要进行任务级验证。总而言之，这些发现确立了机器人用户界面设计是影响具身智能体可靠性的关键因素。

网站 [基准测试 & 代码 & 模型]：https://showlab.github.io/RUIBench

1 引言

基础模型正从理解物理世界迈向在其中采取行动 [2, 39, 74]。视觉-语言-动作模型学习预测机器人动作 [6, 7, 43], ，而近期的智能体系统则使通用模型能够通过工具调用和可执行程序控制机器人 [15, 16, 27, 45,52, 76]。这带来了一个超越模型本身的关键设计选择：智能体应如何与机器人交互？我们将**机器人使用接口（RUI）**定义为智能体能够指令执行的动作及其组织执行的方式。从底层运动指令 [3, 4] 和图像定位点 [25, 57]，到语义技能 [1, 37] 和可执行程序 [52, 71], ，RUI 决定了智能体必须做出哪些决策、哪些任务可以委托执行，以及控制权何时返回智能体。

在复杂的物理任务中，这一区别至关重要。这些任务要求智能体收集缺失的信息、应对动态事件，并通过物理反馈纠正错误。如果每次细微移动都要求模型做出新的决策，执行过程就会成本高昂且容易出错；而如果提供的反馈不足，当世界偏离计划时，系统可能无法应对。因此，可靠执行不仅取决于智能体掌握的信息，还取决于**决策与反馈如何在**智能体和机器人之间分配。

然而，现有具身基准测试在固定界面下比较模型， [18, 23, 47, 82, 84],而接口研究通常只考察经过挑选的设计选项 [27, 64, 68]。因此，接口设计的影响仍未得到充分理解。在我们的实验中，仅改变 RUI 就使

*贡献相同。†通讯作者



图 1. 机器人使用接口（RUI）会影响具身智能体的表现。上： RUIKit 将六种动作抽象和两种交互形式组合成 13 种 RUI，并采用共享的感知与控制。 **下：** (a) 接口选择会显著影响智能体成功率。(b) 决策压缩和程序内反馈带来互补的提升。(c) 接口偏好从仿真到真实环境会发生变化。(d) 针对特定 RUI 的微调可提升小模型在两种交互形式下的表现。

Claude Code 的成功率从 45% 提升至 81%，远超两名前沿智能体之间的平均性能差距（图 1a）。这引出了我们的核心问题：**什么样的机器人使用接口才是好的?**

为回答这一问题，我们推出 RUIBench，这是一个将 30 项智能体操作任务与受控评估协议相结合的基准测试。这些任务超越了简单的抓取与放置，要求智能体进行推理、通过交互获取信息，并适应不断变化的物理条件。它们涵盖五个套件：世界知识、上下文内学习、发现、动态交互和高接触操作。该协议将智能体的输入限制为 RGB-D观测和机器人状态，在模型推理期间让环境持续演化，并独立于智能体的自述来评估任务成功率。我们还开发了 RUIKit 这一统一框架，将两项设计选择区分开来：**动作抽象**，它决定智能体可以发出哪些指令；以及**交互形式**，它决定动作如何组织和执行。RUIKit 将六种动作抽象与两种交互形式交叉组合，构建出 12 种接口，另加一种自由选择接口；所有接口均置于同一个共享测试框架中，唯一变化的是 RUI（图 1）。

在仿真中的前沿模型和小型模型，以及前沿智能体执行的 650 个真实机器人回合中，我们发现，有效的 RUI**能够减少不必要的决策，同时保留必要的反馈**（图 1b），这也让小型模型更容易学习（图 1d）。细粒度交互可能会将清晰的计划拆解为过多的模型决策，这种失败我们称为**执行碎片化**。当智能体能够有效利用多步执行和中间反馈时，将动作组合成程序可以减轻这一负担。然而，真实世界中的执行会改变哪些接口更有效，因为物理不确定性要求提供更多局部校正的机会（图 1c）。我们还发现了**验证差距**：尽管存在可观察到的失败，智能体仍频繁宣称成功；而检查单个动作也无法保证任务完成。综上，这些发现凸显了在适当层级分配决策、反馈和验证的重要性。

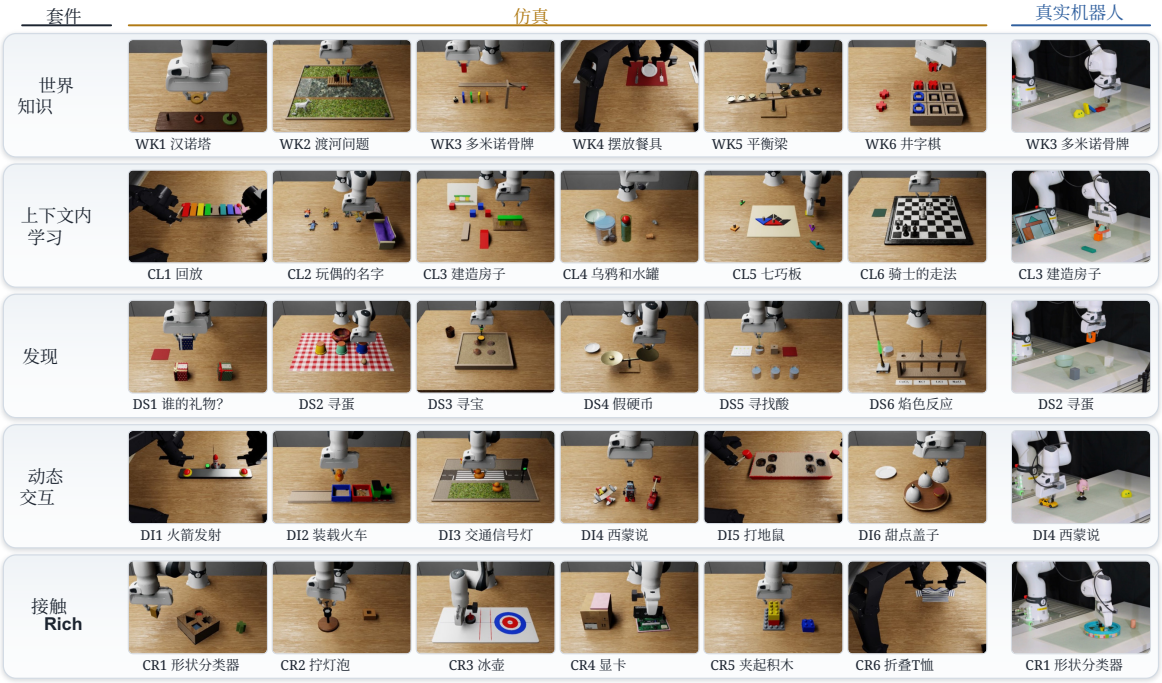


图2. RUIBench 的 30 项任务。 每行是一套由六项模拟任务组成的套件，根据成功所需的能力定义：运用先验知识进行推理、使用先前展示的视觉示例、通过物理探索、响应时间敏感事件，或在接触过程中进行精确调整。右侧列展示了每套任务中也可在真实机器人上运行的任务。

我们的主要贡献总结如下：

- **智能体基准测试。** 我们推出 RUIBench，涵盖五个套件中的 30 项操作任务，并制定受控评估协议，以系统地研究 RUI。
- **统一框架。** 我们开发了 RUIKit，涵盖 13 种界面、六种动作抽象和两种交互形式，并采用共享的感知与控制。
- **实证洞见。** 我们揭示了决策负担、物理反馈和任务级验证如何影响智能体与环境中的具身性能。

2 相关工作

面向基础模型的机器人使用接口。 基础模型通过学习型动作策略 [6, 7, 43] 或连接通用模型与物理动作的接口来控制机器人。现有系统采用语义技能和学习策略工具 [1, 13, 28, 37, 48, 51, 61, 63, 67, 83, 85, 89],可执行程序 [17, 35, 40, 52, 71, 76, 79], 空间约束和价值图 [36, 38, 60],图像定位点和路径 [25, 33, 45, 50, 57, 74, 86, 88], 末端执行器位姿 [44], 以及离散运动指令 [3, 15, 31]。然而，这些接口通常是在各个系统内部开发的，因此难以单独剥离并分析它们各自的影响。近期研究比较了选定的动作抽象 [27], 程序式执行与逐动作执行 [19, 64, 68], 接口族 [4, 53, 65, 91],测试框架组件 [12, 24, 73, 80], 以及规划器-策略接口 [34]。但这些设计选择如何共同影响智能体行为，仍未得到充分理解。RUIKit 正在解决

这一问题：将动作抽象 与交互形式解耦，从而能够对它们各自及共同作用于智能体行为的影响进行受控比较。

具身基础模型的基准测试。现有具身基准测试评估机器人操作、泛化和具身推理 [18, 47, 49, 82]。近期研究开始拓展至基础模型智能体，引入了需要多步推理、规划和物理交互的任务 [14, 23, 29, 30, 46, 66, 78, 84]。评估协议也变得更加贴近现实，纳入了非特权观测、连续物理时间和独立结果评分 [23, 65, 84], , 同时考虑时间和成本 [26, 41, 42, 81, 87]。然而，这些基准测试很少以受控方式改变界面，因此我们难以了解接口设计如何影响面临不同交互需求的智能体。此外，物理任务成功并不一定意味着智能体正确识别了完成状态，

表1. 与相关工作的比较。 **交叉设计** 评估逐动作和程序执行下匹配的动作抽象。 **错误声明**依据物理结果评估完成声明。 **我们的：RUIBench+RUIKit**。 △：部分覆盖。

Work	接口设计			评估		
	动作抽象。	指令 vs. program	交叉设计	错误声明	小型模型	Real 机器人
界面研究与系统						
CaP-X [27]	✓	×	×	×	✓	✓
ALRM [64]	×	✓	×	×	✓	×
Berman 等人 [4]	✓	✓	×	×	×	△
Show-Harness [15]	✓	×	×	×	✓	✓
RobotUse [45]	△	△	×	×	×	✓
智能体基准						
Espinola 等 [23]	×	×	×	△	×	×
RobotWorld [84]	×	×	×	△	×	×
Ours	✓	✓	✓	✓	✓	✓

。先前研究探讨了视觉成功检测和失败推理 [21, 22, 54, 90],，但即使任务尚未完成，智能体仍可能宣称成功 [23, 58, 84]。RUIBench 通过涵盖五种交互需求的 30 项智能体任务，以及将物理结果与智能体的完成声明分开评估的协议，补充了这些研究。RUIBench 与 RUIKit 共同衡量接口设计如何影响执行和完成判断（见表 1）。

3 RUIBench

RUIBench 提供任务（图 2）和评估协议（第 3.2 节），用于研究 RUI 如何影响具身智能体的表现。对于给定的智能体和任务，它固定机器人、感知、计时和评分，并通过 RUIKit（第 4 节）改变 RUI。

3.1 智能体任务设计

任务设计。 明确的任务目标并不意味着实现目标的路径也已完全明确。成功完成操作往往要求智能体超越初始指令进行推理，通过交互获取信息，并根据物理反馈调整行为。RUIBench 通过五个互补的套件中的 30 项任务来体现这些要求（图 2）。 **世界知识**要求基于先验知识进行推理，例如汉诺塔。 **上下文学习** 要求利用先前展示的视觉示例，例如根据图片搭建营垒。 **发现** 要求进行物理探索，例如用天平找出一枚假硬币。**动态交互** 要求对时间敏感事件作出响应，例如西蒙说。 **高接触操作** 要求在接触过程中进行精确调整，例如插入显卡。阴性对照证实，任务确实依赖其预期能力（附录 A）。一个脚本化参考智能体

表2。 RUIBench 的统计数据。 来源列统计改编自心理学（Psych.）或取自日常游戏、谜题、学校实验和家务（Everyday）的任务数量。能力列统计分别强调长时程（LH）、推理（RS）、记忆（ME）、精细控制（FG）、速度（SP）和双手协调（BI）的任务数量。任务统计数据是每项任务的指令长度（Instr.，词数）、场景物体、里程碑和脚本策略时间（Ref.，秒）的平均值。每项任务均提供 15 个开发任务和 20 个测试任务。

套件	来源		核心能力						任务统计			
	心理学	日常	LH	RS	ME	FG	SP	BI	指令对象	里程碑		参考（秒）
世界知识（WK）	3	3	4	4	0	1	0	1	16	13	3.8	36
上下文学习（CL）	2	4	2	2	4	2	0	1	14	16	3.7	40
发现（DS）	0	6	4	2	5	1	0	0	10	10	3.5	34
动态交互（DI）	4	2	0	0	1	1	6	2	15	9	4.0	61
高接触（CR）	2	4	1	1	0	5	0	1	14	4	4.0	22
All	11	19	11	9	10	10	6	5	14	11	3.8	38

解决全部 70 个对照任务实例的智能体，在忽略示范、首帧之后的观测或回合中变化时，只能成功完成其中 6 项。在 30 项任务中，11 项改编自发展心理学与比较心理学范式，19 项源自日常游戏、谜题、学校实验和家务（表 2 和 7）。这些熟悉的情境让目标直观易懂、结果易于解读，同时保留了智能体决策所面临的挑战。

任务配置。每个套件包含六项任务，每项任务至少涉及六种基础能力中的三种：长时程规划、推理、记忆、精细控制、速度和双手协调。各套件侧重这些能力的不同组合（表 2）。每项任务包含三个或四个里程碑，用于衡量阶段性进展。二十五项任务使用配备固定摄像头和腕部摄像头的 Franka Research 3 机械臂，另有五项任务使用两台配备头部摄像头和腕部摄像头的 ARX X5 机械臂。所有任务均在 Isaac Sim 中通过 Isaac Lab 实现 [56]。每项任务均提供 15 个开发任务实例和 20 个基于固定随机种子的测试任务实例。我们仅保留通过可达性、可见性和可抓取性检查，且能够由使用特权状态的脚本化参考策略解决的布局，以确保任务在仿真中的可行性。提示、任务参数和评分器均在开发任务实例上配置；所有智能体和 RUI 则均在相同的留出测试任务上进行评估。

3.2 评估协议

为在保留物理交互要求的同时隔离界面效应，RUIBench 强制采用非特权观测、连续物理时间和独立结果评估。

非特权观测。智能体会接收任务指令、提供给人工操作员的提示，以及对于上下文学习任务，先前展示过的视频或图像。所有其他信息均通过共享工具从 RGB-D 相机和机器人状态读数中获取。SAM 3 [10] 会检测并分割物体，而三维位置则根据有噪声的传感器深度数据恢复，而非来自特权模拟器状态。在 100 帧标注数据上，该处理流程检测出 93% 的物体，定位误差为 0.5 至 0.7 厘米。

连续物理时间。环境在智能体推理期间持续演变。在仿真中，每次模型响应都会产生固定的 8 秒延迟，在此期间机器人保持姿态，而其他物体继续移动。这近似于从三个前沿模型测得的 4,067 次响应的第75百分位数，可避免服务延迟差异混淆界面对比。

独立结果评估。每个回合都从全新的模拟器实例开始，并在智能体声明完成、放弃，或达到回合预算或 30 分钟时限时结束。作为参考，脚本策略平均在 38 秒内完成任务（表2），而一些动态事件仅持续 16 至 20 秒。到达目标并不会结束回合。相反，成功率根据物理轨迹独立判定，并通过执行日志离线核验，且不会告知智能体。因此，智能体必须自行判断任务何时完成，以便将其完成声明与实际物理结果进行比较。

评估指标。除了任务成功率和每回合的 API 成本外，我们还衡量接口行为的两个方面。**决策负担**通过每回合的模型轮次，以及程序每个程序执行的动作数来表征。**完成判断**通过三项比率进行评估（表6）：被判定为完成的回合占比、完成声明中错误声明的占比，以及以完成声明告终的失败回合占比。这些指标共同区分了执行性能与智能体识别任务完成状态的能力。

4 RUIKit

RUIKit 提供了一个用于构建和评估机器人使用接口的统一框架（图3）。它将接口设计与底层的感知、控制和智能体基础设施分离开来。对于给定的智能体和任务，**只有 RUI 会变化**，而测试框架、工具、评估协议和仿真器在不同 RUI 之间保持不变。

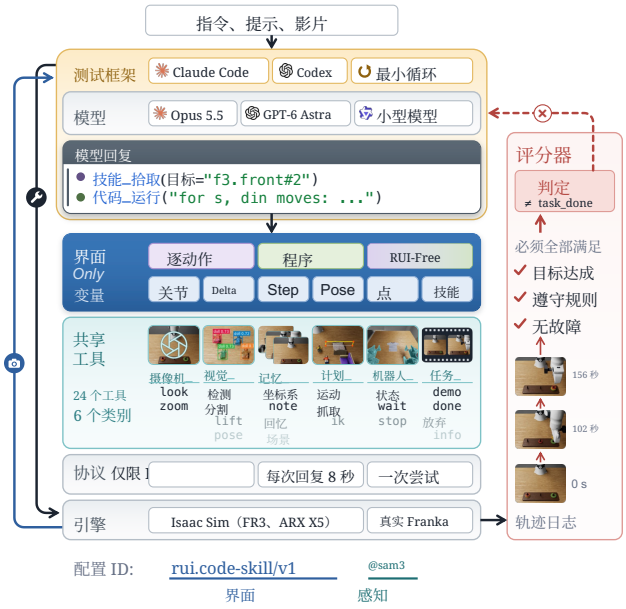


图 3。RUIKit 概览。对于固定的智能体和任务，RUI 是唯一变量，而测试框架、24 个底层工具、评估协议和机器人后端均保持共享。评分器的判定不会告知智能体。

4.1 接口设计空间

机器人使用接口在其提供的动作以及这些动作的执行组织方式上各不相同。我们从两个互补维度刻画这一设计空间：**动作抽象**和**交互形式**（图1）。

动作抽象。该维度决定智能体可以发出什么命令。我们考察六种抽象：Joint 指定关节空间目标；Delta 指定相对平移和旋转；Step 提供固定幅度运动指令；Pose 指定绝对末端执行器目标；Point 将动作定位于图像空间位置；Skill 则提供语义操作，例如抓取、放置、按压、推动和转向。它们在表示方式和控制粒度上有所不同，但并不构成严格的层级关系。

交互形式。这一维度决定了动作如何组合，以及控制权何时返回给智能体。在逐动作 形式中，每次模型回复对每条机械臂最多发出一个动作，随后控制权返回给智能体。在程序 形式中，智能体编写 Python 程序 [52]，通过序列、循环和条件分支来组合

动作。程序还可以查询观测，并在不需要模型再次回复的情况下调整后续动作。对于给定的抽象，**两种形式使用相同的动作原语**，因此它们的区别仅在于动作的组织方式。

接口配置。将六种动作抽象与两种交互形式相结合，可得到 12 种固定 RUI。我们还引入了 RUI-Free，它以两种形式提供所有抽象，并允许智能体自行选择，从而形成 13 种接口。逐动作 RUI 以其抽象命名，例如技能；基于程序的 RUI 则使用前缀 program，例如 program-skill。重要的是，Skill 和 Program 代表不同形式的委托：技能封装了预定义的机器人操作，而程序通过智能体生成的代码组合操作。因此，两者可以结合使用，而不应视为互斥选项。所有配置共享相同的基础设施，因此可以改变其中一个轴，同时固定另一个轴。

4.2 统一智能体框架

共享感知与控制。 RUIKit 提供了横跨六个类别的 24 种底层工具（图 3和表 8）。所有 RUI 共享相同的感知与控制基础设施，包括相机观测、目标检测与分割、基于深度的三维定位，以及机器人状态估计。共享工具还支持记忆、视觉标记、目标跟踪、抓取与运动规划，以及可达性检查。智能体还可通过通用工具访问任务信息，例如指令、示范和事件，并发出等待、停止、完成或放弃命令。不同 RUI 之间唯一变化的是所提供的动作命令及其交互形式。

执行与智能体集成。 RUIKit 将机器人动作和观测实现为可复用的函数。每个界面都会选择其动作原语和交互形式，并使用带版本号的标识符，例如 rui.code-skill/v1。前沿编码智能体通过模型上下文协议（MCP）调用这些函数，而较小的开源模型则使用精简的工具调用循环。程序在沙箱中执行，沙箱仅开放获准使用的函数。两类智能体使用相同的机器人侧基础设施和界面定义。

验证与部署。 RUIKit 会记录工具调用、机器人轨迹、观测结果和执行结果，以供离线评估。独立评分器根据记录的轨迹评估目标满足度、任务约束和仿真有效性。此外，每个会话都会接受审计，以检查是否存在未经授权的工具使用、文件或网络操作，以及对基准测试内部信息的访问。Franka机械臂的真实机器人后端复用相同的界面定义、工具和测试框架，因此所有 13 个 RUI 均可在无需修改的情况下运行。

5 实验

5.1 实验设置

模型。 我们评估了两个前沿编码智能体：使用 Claude Opus 5.5 的 Claude Code，以及使用 GPT-6 Astra 的 Codex；二者均采用高推理强度，并通过 MCP 连接（第 4.2 节）。我们还评估了 Qwen3.5-4B [62]，通过一个不启用思考的最小智能体循环进行测试，分别考察零样本设置和微调后的表现。对于 12 个固定 RUI 中的每一个，我们使用 300 条训练轨迹微调一个 LoRA适配器 [32]，其中 30% 来自成功的 Qwen3.8-Max 回合，70% 来自脚本策略。约一半的脚本轨迹包含注入的偏差和恢复过程。训练种子与测试种子互不重叠（附录 C.1）。

表3. 不同机器人使用接口的仿真结果。 报告五个任务套件及其平均值（平均值）的成功率。轮次：每回合模型回复的中位数；成本：每回合平均 API 成本（美元）。对于 Qwen3.5-4B，ZS/FT：零样本/微调后的成功率；训练/测试：一次性微调成本和每个测试回合的成本，按每 H200 GPU 小时 2.68 美元计算。粗体：每个智能体的最佳成功率和最低每回合成本。

RUI	Claude Code								Codex								Qwen3.5-4B			
	成功率（%）						效率		成功率（%）						效率		成功率（%）		成本（$）	
	WK	CL	DS	DI	CR	平均值	轮次	成本	WK	CL	DS	DI	CR	平均值	转向	成本	ZS	微调	训练	测试
逐动作	每次模型回复每条机械臂执行一个动作																			
关节	80	77	63	70	40	66.0	35	1.48	83	77	67	53	70	70.0	32	2.15	2.0	12.7	0.4K	2.07
Delta	93	47	50	70	53	62.7	31	1.34	80	80	67	37	67	66.0	23	1.71	3.3	21.3	0.3K	1.66
Step	67	10	63	40	47	45.3	37	2.01	87	63	77	67	67	72.0	38	2.37	0.0	20.0	0.9K	0.97
Pose	87	67	80	87	57	75.3	29	1.19	93	83	60	67	70	74.7	23	1.58	0.0	**46.7**	0.4K	0.56
点	87	50	70	50	50	61.3	28	1.49	53	67	93	33	73	64.0	20	1.42	0.0	35.3	0.3K	0.53
技能	100	63	70	63	70	73.3	19	0.90	87	90	90	57	47	74.0	16	1.09	6.7	41.3	0.2K	0.36
总体	86	52	66	63	53	64.0	30	1.40	81	77	76	52	66	70.1	25	1.72	2.0	29.6	0.4K	1.03
程序	每次回复一个 *Python* 程序，使用相同的动作原语																			
关节	100	80	67	83	50	76.0	12	0.81	63	83	87	67	67	73.3	27	1.75	2.7	22.0	0.2K	1.98
Delta	97	70	97	83	57	**80.7**	11	**0.69**	87	87	63	47	67	70.0	20	1.74	2.0	14.0	0.2K	1.96
Step	100	50	93	67	53	72.7	14	0.77	87	70	80	53	50	68.0	34	2.27	0.0	16.7	0.3K	0.55
Pose	100	70	73	67	50	72.0	11	0.78	100	83	67	83	47	76.0	21	1.33	0.0	30.0	0.2K	**0.33**
点	100	87	97	80	33	79.3	14	0.74	87	87	60	87	50	74.0	22	1.75	0.0	26.7	0.4K	0.76
技能	93	47	67	93	47	69.3	12	0.86	60	70	93	70	83	75.3	14	**0.95**	6.7	41.3	0.2K	0.52
总体	98	67	82	79	48	75.0	12	0.77	81	80	75	68	61	72.8	23	1.63	1.9	25.1	0.2K	1.02
自由选择	所有动作抽象和两种交互形式																			
RUI-Free	93	70	80	87	57	77.3	14	0.78	97	87	67	97	57	**80.7**	15	1.47	–	–	–	–
所有 RUI	92	61	75	72	51	70.1	21	1.06	82	79	75	63	63	72.2	23	1.66	1.9	27.3	0.3K	1.02

协议与指标。 每个前沿智能体在 13 个 RUI 上运行全部 30 个任务，并针对每个任务执行 5 个测试任务实例，因此每个智能体共产生 1,950 个仿真回合。小模型在 12 个固定 RUI 上进行评估，因为 RUI-Free 没有专门的训练数据。所有实验均遵循第 3.2 节中的协议。我们报告任务成功、每回合模型轮次的中位数，以及平均 API 成本。对于小模型，我们还报告一次性微调成本，以及按 GPU 租赁价格计算的每回合推理成本（表 3中的训练与测试）。

5.2 仿真中的界面效应

发现 1：具身性能取决于智能体及其界面。仅改变 RUI，就能使 Claude Code 的成功率从 45.3% 提高到 80.7%，远超前沿智能体在所有 RUI 上的平均表现之间 2.1 个百分点的差距（表 3 ）。API 成本也相差近三倍。这表明，界面不仅仅是发出动作的方式：它决定了智能体会做出多少决策，以及每项决策必须达成什么目标。重要的是，同一界面对不同智能体的帮助并不相同。对 Claude Code 而言，程序将平均成功率从 64.0% 提高到 75.0%，将中位回合数从 30 降至 12，并将平均 API 成本从 1.40 美元降至 0.77 美元。Claude Code 每个程序组合 3.3 个动作，其中 54% 的程序会在执行动作后观察世界。Codex 平均每个程序仅包含 0.9 个动作，通常只是用代码封装单个动作，因此采用程序形式仅提升了 2.7 个百分点（附录 C.2）。技能通过封装预定义操作，提供了另一种减少模型决策的方式；对于两个智能体而言，在逐动作 RUI 中，技能都实现了最少的轮次和最低的成本。RUI-Free 进一步说明了

界面选择的作用：它使 Codex 的成功率达到 80.7%，优于所有固定 RUI，同时对 Claude Code 也具有竞争力。总而言之，这些结果表明，**只有当智能体能够有效利用界面可供性时，它们才会发挥作用**。

学习使用界面。小模型实验考察这种能力能否通过微调获得，而非假定它会随模型规模自然具备。零样本情况下，Qwen3.5-4B 在 12 种固定 RUI 的回合中仅有 1.9% 的成功率；它在 Skill（6.7%）上的表现最好，因为 Skill 封装了完整操作。每种 RUI 使用 300 条轨迹进行微调后，成功率提升至 27.3%，一次性成本为 0.2K 至 0.9K 美元，但学习效果因 RUI 而异（表 3）。以目标而非运动命名的抽象（Pose、Point 和 Skill）在逐动作形式下达到 35.3% 至 46.7%，而 Joint、Delta 和 Step 则维持在 12.7% 至 21.3% 之间。它们也缩短了回合：每回合成本为 0.36 至 0.56 美元，低于所有前沿配置；而 Joint 或 Delta 的每回合成本为 1.66 至 2.07 美元。相比之下，程序使平均成功率从 29.6% 降至 25.1%，且仅对 Joint 有帮助，因为 Joint 的逐动作回合耗时最长。因此，**优秀的 RUI 也应易于学习**：抽象能减少决策、逐动作执行能保留反馈，小模型便能在这类 RUI 上学得最多。对所有三种智能体而言，Pose 和 Skill 都是表现最好的两种逐动作 RUI，尽管 Pose 的零样本成功率为 0.0%。使用逐动作 Pose 时，小模型甚至能媲美采用逐动作 Step 的 Claude Code（46.7% 对 45.3%）。

5.3 决策压缩与反馈

发现2：高效执行既需要减少决策，也需要有效反馈。细粒度的逐动作控制可能会将清晰的计划拆分成反复进行的模型决策，从而增加延迟、成本和出错的机会。我们将这种失败称为**执行碎片化**。程序可以减轻这一负担，但其优势可能同时来自动作组合和中间反馈。

表 4。 交互形式和动作级检查的影响。成功率（%）；轮次：每个回合的模型回复中位数；成本：每个回合的平均 API 成本（美元）；失败后声称完成：以完成声明结束的失败回合所占比例（%）。每个任务使用一个随机种子。

变体	成功率	效率		失败声称
		轮次	Cost	
交互形式 *Delta*、步骤、点；每种各 *90* 个回合				
逐动作	57.8	29	1.58	71.1
动作序列	65.6	16.5	0.86	77.4
程序	**75.6**	14	0.79	86.4
动作级检查 技能、程序技能；每种各 *60* 个回合				
不检查	73.3	16	0.89	87.5
报告	73.3	17.5	0.94	81.2
报告并纠正	66.7	16.5	0.91	**75.0**

为区分这些影响，我们引入一种动作序列变体：连续执行多个动作，不提供中间观察，并仅在最后返回结果。在 Delta、步骤和点界面上，Claude Code 使用动作序列时的表现从逐动作执行时的 57.8% 提升至 65.6%，同时，中位回合数从 29 降至 16.5（表 4）。允许在程序中进行观测和条件执行，可进一步将成功率提升至 75.6%，并将轮次降至 14。因此，动作组合减轻了决策负担，而程序内反馈则带来了进一步提升。

这些优势取决于任务需求（表 10）。上下文学习任务能从执行已知动作序列中显著获益，而知识、发现和动态任务则更依赖中间反馈。高接触操作并未持续从任何一种方式中受益，这表明其所需的纠正可能超出当前程序级交互所能提供的范围。因此，关键区别并不只是执行时间的长短，而在于**反馈能否影响下一步动作**。

5.4 从仿真到现实

真实机器人设置。 我们在配备固定摄像头和腕部摄像头的 Franka Research 3 机械臂上，对两种前沿智能体进行了评估，覆盖全部 13 种 RUI。评估涵盖五项任务，每个套件各一项：多米诺骨牌（WK3）、建造房屋（CL3）、寻蛋（DS2）、西蒙说（DI4）和形状分类器（CR1）。每项任务有五种布局，共进行了 650 个真实机器人回合。界面和任务预算与仿真一致，执行时间则按实际时间计算。由一名人工操作员独立评估任务成功率，其他质量评分和运行过程见附录C.3。

发现3：真实世界中的执行会改变界面偏好。 在仿真中观察到的界面优势，并不一定会延续到物理执行中（表 5）。对于 Claude Code，程序在仿真中的表现比逐动作执行高出 23.3 个百分点，但在真实机器人上不再具有这一优势（36.0% 对比 42.0%）。两者从仿真到真实场景的成功率分别下降了 54.0 和 24.7 个百分点。Codex 的变化幅度较小，在真实机器人上的逐动作执行表现略优于程序。自由选择的迁移效果最差：对于两个智能体，RUI-Free 的成功率都从仿真中的 100.0% 降至真实机器人上的 44.0%。在动作抽象中，Delta 对两个智能体的表现都最好；Pose 紧随其后，尤其适用于 Codex。

表 5。 接口设计的仿真到现实的影响。 五个匹配任务上的成功率（%）（仿真：一个种子；真实：650 个回合）。 Δ：真实减去仿真（百分点）。抽象取两种形式的平均值；粗体标记每组中真实结果的最佳值。

RUI	Claude Code			Codex		
	Sim	**Real**	Δ	Sim	**Real**	Δ
交互形式 / 选择						
逐动作	66.7	42.0	−24.7	73.3	**60.0**	−13.3
程序	90.0	36.0	−54.0	73.3	56.0	−17.3
RUI-Free	100.0	**44.0**	−56.0	100.0	44.0	−56.0
动作抽象						
关节	70.0	34.0	−36.0	60.0	54.0	−6.0
Delta	70.0	**54.0**	−16.0	60.0	**66.0**	+6.0
Step	60.0	36.0	−24.0	60.0	52.0	−8.0
Pose	80.0	42.0	−38.0	100.0	64.0	−36.0
点	100.0	32.0	−68.0	70.0	60.0	−10.0
技能	90.0	36.0	−54.0	90.0	52.0	−38.0
所有 RUI	80.0	39.4	−40.6	75.4	56.9	−18.5

这些差异凸显了在**物理不确定性下反馈和局部校正的物理不确定性**的重要性。与仿真相比，真实执行更容易受到深度噪声、校准偏移和滑脱抓取的影响；两个摄像头有时对同一对象的位置判断相差多达 1.5 厘米。成功的轨迹通常包括放置对象、检查结果，再逐步进行校正。相比之下，一段 Claude Code 程序放置了三块多米诺骨牌，却没有检查它们之间的间距，并错误地宣告任务完成。这些案例表明，动作组合的益处不仅取决于能否减少模型决策，还取决于能否**保留纠正执行错误的机会**。

当前界面的局限。 有些任务需要的反馈时间尺度或精度水平，超出了模型级交互能够可靠提供的范围。在真实机器人上，西蒙说仍然颇具挑战，因为智能体进行推理时，目标也在移动。Codex 的图像点程序提供了一个值得借鉴的例外：它利用程序内观测，在全部五种布局中跟踪并抵达目标。这表明，时间关键型反应可能需要在程序内部或低层控制器中运行。

高接触操作带来了互补性的挑战。在仿真中，插入显卡和折叠T恤在几乎所有 RUI 中都仍然很困难。这些任务的成功取决于精细调整，而现有观测难以充分分辨这些调整：T恤偏移 2 厘米，在头部摄像机中仅约占八个像素。总体而言，这些案例揭示了对反馈和控制机制的需求，即这些机制应在适当的时间和空间分辨率下运行。

5.5 验证差距

发现 4：观察到失败并不保证能够识别失败。除了执行动作之外，智能体还必须判断任务是否真正成功。然而，对于 Claude Code 和 Codex 而言，失败回合中分别有 82.1% 和 70.9% 最终以完成声明收场（表6）。我们将这种失败模式称为 **虚假完成**。

表 6。 **完成判定诊断（%）。** 在线（30 个任务，一个测试种子）：跨回合的声明率、完成声明中的错误声明，以及失败回合中的完成声明。事后（23 个任务）：由新的评审器在最终帧中检测到的失败，以及未重新检查便作出的声明。

智能体	在线判定			事后分析	
	声明 rate	错误声称	失败声称的	失败可见	No 重新检查
Claude Code	94.6	26.0	**82.1**	70.2	69.7
Codex	91.8	21.8	**70.9**	80.9	92.1

关键的是，这些失败通常可以通过视觉检测到（图 5）。在对 23 项任务进行离线分析时（附录 C.4），一个新的评审器使用仅凭智能体已经收到的最终帧，评审器识别出 Claude Code 70.2% 的错误声明和 Codex 80.9% 的错误声明。其中大多数声明是在没有再次观察的情况下作出的（分别为 69.7% 和 92.1%）。对于 Codex，离线评审器在全新上下文中使用相同的底层模型，这表明问题并非只是缺乏视觉证据。相反，**能够获取证据，并不保证能做出基于证据的完成判定**。

动作级检查不足以进行任务级验证。 我们在技能和 Program-Skill 接口中引入了针对抓取和放置的后置条件检查。这些检查仅使用可观测的机器人状态和相机测量值；与 149 个人工验证标签的一致率达到 88.6%（抓取为 97.3%，放置为 80.0%），与 524 条模拟器记录的一致率达到 94.5%。

尽管一致率较高，这些检查带来的改进仍然有限（表 4）。报告失败动作后，以完成声明告终的失败回合占比从 87.5% 降至 81.2%；自动重试则将其降至 75.0%，但并未提升任务成功率。32 次纠正中仅有 2 次挽回了失败的抓取。智能体可能成功执行了单个动作，却仍未完成总体任务。这揭示了一个根本区别： **动作级反馈无法替代任务级验证**。

6 结论

机器人智能体能够完成什么，不仅取决于其模型，还取决于它执行动作所通过的界面。我们提出了 RUIBench 和 RUIKit，以在 30 个智能体操作任务和 13 个接口中系统地研究这种关系。我们的研究结果表明，有效的 RUI**能够减少不必要的决策，同时保留必要的反馈**，其优势会受到智能体行为和物理不确定性的影响。这类界面也更易于小模型学习。我们进一步发现了一个**验证差距**：尽管存在可观察到的失败，智能体却经常宣称任务成功。总体而言，这些结果表明，机器人用户界面设计的核心问题在于，在智能体与机器人之间分配决策、执行和反馈。**关键不只是智能体能够执行哪些动作，还在于何时进行决策、何时纠正，以及如何验证成功**。

局限性与未来工作。 将真实机器人研究从单个机械臂上的五项任务扩展出去，可以检验观察到的界面变化是否具有广泛适用性。高接触和时间关键型任务对于几乎所有 RUI 仍然十分困难，因此需要在更精细的时间和空间分辨率下提供反馈和控制，例如在程序内部进行响应或使用低层控制器。若要进一步弥合验证差距，就需要任务级验证器依据观测而非单个动作来检查任务是否完成。

参考文献

[1] Michael Ahn、Anthony Brohan、Noah Brown、Yevgen Chebotar、Omar Cortes、Byron David、Chelsea Finn、Chuyuan Fu、Keerthana Gopalakrishnan、Karol Hausman 等人。《照我能做的做，而非照我说的做：将语言建立在机器人可供性之上》。载于机器人学习会议，第 287–318 页。PMLR，2023年。[2] Alisson Azzolini、Junjie Bai、Hannah Brandon、Jiaxin Cao、Prithvijit Chattopadhyay、Huayu Chen、Jinju Chu、Yin Cui、Jenna Diamond、Yifan Ding 等人。《Cosmos-Reason1：从物理常识到具身推理》。*arXiv* 预印本 *arXiv:2503.15558*，2025年。[3] Suneel Belkhale、Tianli Ding、Ted Xiao、Pierre Sermanet、Quan Vuong、Jonathan Tompson、Yevgen Chebotar、Debidatta Dwibedi 和 Dorsa Sadigh。《RT-H：使用语言的动作层级》。载于机器人学：科学与系统，2024年。[4] Shmuel Berman、Michael Ilie、Jia Deng 和 Daniel Freeman。《Claude 玩机器人学》。Anthropic 研究部，2026年。[5] Christopher David Bird 和 Nathan John Emery。《白嘴鸦使用冰壶抬高水位，以够到漂浮的蠕虫》。《当代生物学》，19(16):1410–1414，2009年。[6] Kevin Black、Noah Brown、Danny Driess、Adnan Esmail、Michael Equi、Chelsea Finn、Niccolo Fusai、Lachy Groom、Karol Hausman、Brian Ichter 等人。《 π_0：用于通用机器人控制的视觉-语言-动作流模型》。载于机器人学：科学与系统，2025年。[7] Anthony Brohan、Noah Brown、Justice Carbajal、Yevgen Chebotar、Xi Chen、Krzysztof Choromanski、Tianli Ding、Danny Driess、Avinava Dubey、Chelsea Finn 等人。《RT-2：视觉-语言-动作模型将网络知识迁移至机器人控制》。载于机器人学习会议，第 2165–2183 页。PMLR，2023年。[8] Josep Call。《猩猩（Pongo pygmaeus）、黑猩猩（Pan troglodytes）和儿童（智人，Homo sapiens）的客体永久性》。《比较心理学杂志》，115(2):159–171，2001年。[9] Susan Carey 和 Elsa Bartlett。《习得一个新词》。《儿童语言发展论文与报告》，15:17–29，1978年。[10] Nicolas Carion、Laura Gustafson、Yuan-Ting Hu、Shoubhik Debnath、Ronghang Hu、Didac Suris、Chaitanya Ryali、Kalyan Vasudev Alwala、Haitham Khedr、Andrew Huang 等人。《SAM 3：利用概念分割万物》。载于第十四届国际学习表征会议，2026年。[11] Lucy G Cheke、Elsa Loissel 和 Nicola S Clayton。《儿童如何解开伊索寓言中的谜题？》《公共科学图书馆·综合》，7 (7):e40574，2012年。[12] Kaiyuan Chen、Xinyue Zhu、Shuangyu Xie、Uksang Yoo、Thi Thu Phuong Luong、Marcus Gualtieri、Christian Juette、Liu Ren、Spencer Huang 和 Ken Goldberg。《智能体机器人学中模型、技能、反馈与课程设置的对照研究》。*arXiv* 预印本 *arXiv:2611.09030*，2026年。[13] Siyi Chen、Hugo Hadfield、Alex Zook、Mikaela Angelina Uy、Chan Hee Song、Erwin Coumans、Xuning Yang、Faisal Ladhak、Qing Qu、Stan Birchfield 等人。《VoLo：用于开放词汇长时程操作的物理协调器》。载于机器人学习会议，2026年。[14] Yanzhe Chen、Kevin Qinghong Lin 和 Mike Zheng Shou。《Code2video：教育视频生成的以代码为中心的范式》。*arXiv* 预印本 *arXiv:2510.01174*，2025年。[15] Yanzhe Chen、Zechen Bai、Zhijun Cao、Wenzheng Zeng、Kevin Qinghong Lin、Yiqi Lin、Guoqiang Liang、Kevin Yuchen Ma、Qiming Huang 和 Mike Zheng Shou。《Show-Harness：仅凭一个视觉语言模型智能体就能操控机器人》。*arXiv* 预印本 *arXiv:2609.10522*，2026年。[16] Yanzhe Chen、Ziyi Yang、Jifeng Zhu、Qiming Huang、Ruihe An、Peiyao Xu、Hesen Yang、Runda Liu、Chang Gong、Zhijun Cao 等人。《多模态具身智能体：从计算机使用到机器人使用的统一能力中心视角》。*arXiv* 预印本 *arXiv:2610.30108*，2026年。

[17] Yongchao Chen、Yilun Hao、Yang Zhang 和 Chuchu Fan。Code-as-Symbolic-Planner：通过符号代码生成实现基于基础模型的机器人规划。载于*2025*年 *IEEE/RSJ* 国际智能机器人与系统会议 *(IROS)*，第 19248–19254 页。IEEE，2025年。[18] Zhili Cheng、Yuge Tu、Ran Li、Shiqi Dai、Jinyi Hu、Shengding Hu、Jiahao Li、Yang Shi、Tianyu Yu、Weize Chen 等。EmbodiedEval：将多模态大语言模型作为具身智能体进行评估。*arXiv* 预印本 *arXiv:2501.11858*，2025年。[19] Seokju Cho、Ryo Hachiuma、Abhishek Badki、Hang Su、Byung-Kwan Lee、Chan Hee Song、Sifei Liu、Subhashree Radhakrishnan、Seungryong Kim、Yu-Chiang Frank Wang 等。SpatialClaw：重新思考智能体空间推理的动作接口。神经信息处理系统进展，39，2026年。[20] Kevin Crowley 和 Robert S Siegler。幼儿在井字棋中的灵活策略运用。认知科学，17(4)：531–561，1993年。[21] Yuqing Du、Ksenia Konyushkova、Misha Denil、Akhil Raju、Jessica Landon、Felix Hill、Nando de Freitas 和 Serkan Cabi。视觉-语言模型作为成功检测器。载于终身学习智能体会议，第 120–136 页。PMLR，2023年。[22] Jiafei Duan、Wilbert Pumacay、Nishanth Kumar、Yi Ru Wang、Shulin Tian、Wentao Yuan、Ranjay Krishna、Dieter Fox、Ajay Mandlekar 和 Yijie Guo。AHA：用于检测和推理机器人操作中失败的视觉-语言模型。载于第十三届国际学习表征会议，2025年。[23] May Espinola、Qifan Zhang 和 Yu Xiang。从推理到动作：对前沿人工智能模型在 LIBERO-10 机器人操作任务上的表现进行基准测试。*arXiv* 预印本 *arXiv:2610.30051*，2026年。[24] Liaoyuan Fan、Yuhao Zheng、Dongxing Mao、Hangyu Ran、Siyuan Hu、Weijia Mao、Rui Zhao、Weijia Wu、Guian Fang、Yanzhe Chen 等。视频生成遇上智能体运行框架：智能体视频生成、智能体世界建模与递归自我改进。*arXiv* 预印本 *arXiv:2610.40191*，2026年。[25] Kuan Fang、Fangchen Liu、Pieter Abbeel 和 Sergey Levine。MOKA：通过基于标记的视觉提示实现开放世界机器人操作。载于机器人学：科学与系统，2024年。[26] Romain Froger、Pierre Andrews、Matteo Bettini、Amar Budhiraja、Ricardo Silveira Cabral、Virginie Do、Emilien Garreau、Jean-Baptiste Gaya、Hugo Laurençon、Maxime Lecanu 等。ARE：扩展智能体环境与评估。*arXiv* 预印本 *arXiv:2509.17158*，2025年。[27] Letian Fu、Justin Yu、Karim El-Refai、Ethan Kou、Haoru Xue、Huang Huang、Wenli Xiao、Fei-Fei Li、Guanya Shi、Jiajun Wu 等。CaP-X：用于机器人操作编程智能体基准测试与改进的框架。载于国际机器学习会议，第 32155–32212 页。PMLR，2026年。[28] Liane Galanti、Dhruv Shah 和 Tri Dao。通过物理能动性弥合通用型机器人的编排鸿沟。*arXiv* 预印本 *arXiv:2607.21725*，2026年。[29] Hongcheng Gao、Jianshu Zeng、Alex Zhou、Shijia Ge、Xuan Shangguan、Yazhe Wang、Yexing Wan、Zelin Zheng、Jay Zhu、Xiping Li 等。HexaShell：用于仿真和真实世界中机器人编程智能体的通用测试框架。*arXiv* 预印本 *arXiv:2610.23479*，2026年。[30] Hongcheng Gao、Zelin Zheng、Alex Zhou、Jay Zhu、Shijia Ge、Yazhe Wang、Xiping Li、Jianshu Zeng、Di Wu、Xuan Shangguan 等。EmbodiedLab：对通过行动获取知识的具身智能体进行基准测试。*arXiv* 预印本 *arXiv:2610.35385*，2026年。[31] Meng-Hao Guo、Zhe-Han Mo、Jia-Jun Wang、Yi Zhang、Kejin Wang、Yi-Xuan Deng、Jia-Peng Zhang、Yongming Rao 和 Shi-Min Hu。将视觉语言模型的智能迁移至机器人控制。*arXiv* 预印本 *arXiv:2609.22966*，2026年。[32] Edward J Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang 和 Weizhu Chen。低秩适配：大语言模型的低秩适配。载于国际学习表征会议，2022年。

[33] Hengyuan Hu、Jensen Gao、Priya Sundaresan、Satvik Sharma、Jeannette Bohg 和 Dorsa Sadigh。VIA：用于机器人控制的视觉界面智能体。*arXiv* 预印本 *arXiv:2607.11119*，2026年。[34] Jiaheng Hu、Mohit Shridhar、Caden Lu、Dhruv Shah、Hao-Tien Lewis Chiang、Jie Tan 和 Annie Xie。机器人策略编排中的关键因素：分层视觉-语言-动作智能体的系统研究。*arXiv* 预印本 *arXiv:2606.10267*，2026年。[35] Siyuan Huang、Zhengkai Jiang、Hao Dong、Yu Qiao、Peng Gao 和 Hongsheng Li。Instruct2Act：利用大语言模型将多模态指令映射为机器人动作。*arXiv* 预印本 *arXiv:2305.11176*，2023年。[36] Wenlong Huang、Chen Wang、Ruohan Zhang、Yunzhu Li、Jiajun Wu 和 Li Fei-Fei。VoxPoser：用于借助语言模型进行机器人操作的可组合三维价值图。载于机器人学习会议，第 540–562 页。PMLR，2023年。[37] Wenlong Huang、Fei Xia、Ted Xiao、Harris Chan、Jacky Liang、Pete Florence、Andy Zeng、Jonathan Tompson、Igor Mordatch、Yevgen Chebotar 等。内心独白：通过语言模型规划实现具身推理。载于机器人学习会议，第 1769–1782 页。PMLR，2023年。[38] Wenlong Huang、Chen Wang、Yunzhu Li、Ruohan Zhang 和 Li Fei-Fei。ReKep：面向机器人操作的关系关键点约束时空推理。载于机器人学习会议，第 4573–4602 页。PMLR，2025年。[39] Yuheng Ji、Huajie Tan、Jiayu Shi、Xiaoshuai Hao、Yuan Zhang、Hengyuan Zhang、Pengwei Wang、Mengdi Zhao、Yao Mu、Pengju An 等。RoboBrain：用于机器人操作、从抽象到具体的统一大脑模型。载于计算机视觉与模式识别会议论文集，第 1724–1734 页，2025年。[40] Mengzhao Jia、Yang Lin、Xixin Zhang、Zhihan Zhang、Xiaobai Liu 和 Meng Jiang。作为机器人操作策略的智能体。*arXiv* 预印本 *arXiv:2609.12541*，2026年。[41] Sayash Kapoor、Benedikt Stroebl、Zachary S. Siegel、Nitya Nadgir 和 Arvind Narayanan。真正重要的人工智能智能体。机器学习研究汇刊，2025年。[42] Sayash Kapoor、Benedikt Stroebl、Peter Kirgis、Nitya Nadgir、Zachary S Siegel、Boyi Wei、Tianci Xue、Ziru Chen、Felix Chen、Saiteja Utpala 等。综合智能体排行榜：人工智能智能体评估所缺失的基础设施。载于第十四届国际学习表征会议，2026年。[43] Moo Jin Kim、Karl Pertsch、Siddharth Karamcheti、Ted Xiao、Ashwin Balakrishna、Suraj Nair、Rafael Rafailov、Ethan P Foster、Pannag R Sanketi、Quan Vuong 等。OpenVLA：开源视觉-语言-动作模型。载于机器人学习会议，第 2679–2713 页。PMLR，2025年。[44] Teyun Kwon、Norman Di Palo 和 Edward Johns。语言模型作为零样本轨迹生成器。*IEEE*机器人与自动化快报，9(7)：6728–6735，2024年。[45] Junhoo Lee、Injun Baek、Seungyeon Kim、Suhyun Jeon、Minkyu Kim、Baekseung Kim 和 Nojun Kwak。RobotUse：计算、上下文与决策的分配。*arXiv* 预印本 *arXiv:2610.04929*，2026年。[46] Christopher Leet、Achyutha Menon、Sravanthi Machcha、Sven Koenig、Tzu Kit Chan 和 Jay Chooi。苦涩教训降临机器人学。*arXiv* 预印本 *arXiv:2610.25287*，2026年。[47] Manling Li、Shiyu Zhao、Qineng Wang、Kangrui Wang、Yu Zhou、Sanjana Srivastava、Cem Gokmen、Tony Lee、Li E Li、Ruohan Zhang 等。具身智能体接口：对用于具身决策的大语言模型进行基准测试。神经信息处理系统进展，37：100428–100534，2024年。[48] Ruiying Li、Yunlang Zhou、Yuyao Zhu、Kylin Chen、Sukai Wang、Kongtao Hu、Minhui Yu、Bowen Jiang、Zhan Su、Jiayao Ma 等。RoboClaw：用于可扩展长时程机器人任务的智能体框架。载于欧洲计算机视觉会议，第 304–320 页。Springer，2026年。[49] Siyao Li、Jiawei Gu、Shuai Liu、Kairui Hu、Zekun Li、Linjie Li、Chengcheng Tang、Po-Chen Wu、Ivan Shugurov、Lingni Ma 等。HumanCLAW：视觉-语言模型能否通过身体采取行动？*arXiv* 预印本 *arXiv:2607.27180*，2026年。

[50] Yi Li、Yuquan Deng、Jesse Zhang、Joel Jang、Marius Memmel、Caelan Garrett、Fabio Ramos、Dieter Fox、Anqi Li、Abhishek Gupta等。HAMSTER：面向开放世界机器人操作的分层动作模型。在第十三届国际学习表征会议，2025年。[51] Zexi Li、Yehang Zhang、Wenqian Li、Haojian Huang、Chenxu Wang、Shiyuan Deng、Yangkai Wei、Tianyi Zhang、Binghui Xie、Bohan Zhou等。Robo-Harness K1：通过感知增强驾驭机器人使用智能体。*arXiv*预印本 *arXiv:2609.29389*，2026年。[52] Jacky Liang、Wenlong Huang、Fei Xia、Peng Xu、Karol Hausman、Brian Ichter、Pete Florence和Andy Zeng。代码即策略：用于具身控制的语言模型程序。在*2023*年*IEEE*国际机器人与自动化会议（*ICRA*），第9493–9500页。IEEE，2023年。[53] Haowen Liu、Xirui Li、Shaoxiong Yao、Peng Shi、Tianyi Zhou、Jia-Bin Huang、Furong Huang和Jiayuan Mao。Guava：通过机器人操作框架将前沿视觉语言模型蒸馏为紧凑型智能体。*arXiv*预印本 *arXiv:2606.18363*，2026年。[54] Zeyi Liu、Arpit Bahety和Shuran Song。REFLECT：总结机器人经验以解释和纠正失败。在机器人学习会议，第3468–3484页。PMLR，2023年。[55] Christopher G Lucas、Sophie Bridgers、Thomas L Griffiths和Alison Gopnik。儿童何时比成人更优秀（或至少更开放）地学习：学习因果关系形式的发展差异。*Cognition*，131(2)：284–299，2014年。[56] Mayank Mittal、Pascal Roth、James Tigue、Antoine Richard、Octi Zhang、Peter Du、Antonio Serrano-Muñoz、Xinjie Yao、René Zurbrügg、Nikita Rudin等。Isaac Lab：用于多模态机器人学习的GPU加速仿真框架。*arXiv*预印本 *arXiv:2511.04831*，2025年。[57] Soroush Nasiriany、Fei Xia、Wenhao Yu、Ted Xiao、Jacky Liang、Ishita Dasgupta、Annie Xie、Danny Driess、Ayzaan Wahid、Zhuo Xu等。PIVOT：迭代式视觉提示为视觉语言模型 elicitation 出可执行知识。在国际机器学习会议，第37321–37341页。PMLR，2024年。[58] Zaruhi Navasardyan、Tatul Danielyan和Hrant Davtyan。FailBench：视觉语言模型判断机器人任务成功的可靠性如何？*arXiv*预印本 *arXiv:2609.03611*，2026年。[59] Helena Örnkloo和Claes von Hofsten。将物体装入孔洞：空间认知技能的发展。*DevelopmentalPsychology*，43(2)：404–416，2007年。[60] Mingjie Pan、Jiyao Zhang、Tianshu Wu、Yinghao Zhao、Wenlong Gao和Hao Dong。OmniManip：通过作为空间约束的以物体为中心的交互基元实现通用机器人操作。在*IEEE/CVF*计算机视觉与模式识别会议论文集，第17359–17369页，2025年。[61] Jiaqi Peng、Xiqian Yu、Delin Feng、Yuqiang Yang、Wenzhe Cai、Jing Xiong、Ganlin Yang、Jinliang Zheng、Jiafei Cao、Xueyuan Wei等。Cortex：面向长时程操作的双向对齐具身智能体框架。*arXiv*预印本 *arXiv:2607.05377*，2026年。[62] Qwen团队。Qwen3.5：迈向原生多模态智能体。Qwen博客，2026年2月。[63] Krishan Rana、Jesse Haviland、Sourav Garg、Jad Abou-Chakra、Ian Reid和Niko Suenderhauf。SayPlan：利用三维场景图对大语言模型进行grounding，以实现可扩展的机器人任务规划。在机器人学习会议，第23–72页。PMLR，2023年。[64] Vitor Gaboardi dos Santos、Ibrahim Khadraoui、Ibrahim Farhat、Hamza Yous、Samy Teffahi和Hakim Hacid。ALRM：用于机器人操作的智能体式大语言模型。*arXiv*预印本 *arXiv:2601.19510*，2026年。[65] 商汤联合研究团队、Zhongang Cai、Ziang Cao、Tailai Chen、Han Chong Chua、Linjun Dai、Hanming Deng、Hokin Deng、Haiwen Diao、Chenjing Ding等。GPT-6-Astra实现具身智能了吗？*arXiv*预印本 *arXiv:2610.22471*，2026年。[66] Zeyu Shen、Haoxiang You、Yilang Liu、Zhicheng Zheng、Lihan Zha、Kashu Yamazaki、Mingtong Zhang、Suning Huang、Jiankai Sun、Qianzhong Chen等。EmbodiedSWE：面向长时程灵巧机器人技术的编程智能体。*arXiv*预印本 *arXiv:2609.27308*，2026年。

[67] Lucy Xiaoyang Shi、Brian Ichter、Michael Robert Equi、Liyiming Ke、Karl Pertsch、Quan Vuong、James Tanner、Anna Walling、Haohuan Wang、Niccolo Fusai等。Hi Robot：使用分层视觉-语言-动作模型进行开放式指令跟随。在国际机器学习会议，第54919–54933页。PMLR，2025年。[68] Ruiyang Si、Jianxin Bi、Shunyu Yang、Rui Ni、Wenbo Huang、Qiang Wang、Shulong Jiang、Yintianrun Zhang、Feixiang Ruan、Kewei Zhang等。更少的词元，更好的动作：GPT-6 Astra机器人智能体以少65%的词元实现高14%的成功率。*arXiv*预印本 *arXiv:2610.01939*，2026年。[69] Robert S Siegler。认知发展的三个方面。《认知心理学》，8(4)：481–520，1976年。[70] Herbert A Simon。问题解决技能的功能等价性。《认知心理学》，7(2)：268–288，1975年。[71] Ishika Singh、Valts Blukis、Arsalan Mousavian、Ankit Goyal、Danfei Xu、Jonathan Tremblay、Dieter Fox、Jesse Thomason和Animesh Garg。ProgPrompt：使用大语言模型生成情境化机器人任务规划。在*2023*年*IEEE*国际机器人与自动化会议（*ICRA*），第11523–11530页。IEEE，2023年。[72] Ellen A Strommen。儿童游戏中的言语自我调节：“西蒙说”游戏中的冲动性错误。《儿童发展》，44(4)：849–853，1973年。[73] Ke Sun、Kevin Yuchen Ma、Yike Zhang、Yanzhe Chen和Mike Zheng Shou。ForceHarness：将力转化为动作，并为冻结的前沿模型引入物理学。*arXiv*预印本 *arXiv:2611.21099*，2026年。[74] Gemini Robotics团队、Abbas Abdolmaleki、Saminda Abeyruwan、Joshua Ainslie、Jean-Baptiste Alayrac、Montserrat Gonzalez Arenas、Ashwin Balakrishna、Nathan Batchelor、Alex Bewley、Jeff Bingham等。Gemini Robotics 1.5：凭借先进的具身推理、思考和运动迁移推动通用型机器人前沿发展。*arXiv*预印本 *arXiv:2510.03342*，2025年。[75] Shauna L Tominey和Megan M McClelland。红灯、紫灯：使用圆圈游戏改善学前儿童行为自我调节的随机试验结果。《早期教育与发展》，22(3)：489–519，2011年。[76] Sai H Vemprala、Rogerio Bonatti、Arthur Bucker和Ashish Kapoor。用于机器人学的ChatGPT：设计原则与模型能力。《*IEEE Access*》，12：55682–55696，2024年。[77] Brian N Verdine、Roberta M Golinkoff、Kathryn Hirsh-Pasek、Nora S Newcombe、Andrew T Filipowicz和Alicia Chang。解构积木：学前儿童的空间组装表现与早期数学技能相关。《儿童发展》，85(3)：1062–1076，2014年。[78] Kangrui Wang、Qineng Wang、Changhe Chen、Pingyue Zhang、Chi Gui、Feishi Wang、Canyu Chen、Wenli Xiao、Xiaoyu Huang、Tonghe Zhang等。具身智能首考。*arXiv*预印本 *arXiv:2611.12177*，2026年。[79] Xin Wang、Wenhao Wu、Menghao Zhang、Zhi Wang、Kun Shao、Jian Luan、Yang Li、Qing Li、Shangding Gu、Huichi Zhou等。HarnessPAI：面向物理人工智能的演化测试框架。*arXiv*预印本 *arXiv:2609.29166*，2026年。[80] Zefan Wang、Yufan Dang、Xuanwei Shi、Mingqi Gao、Cheng Yin、SikYuen Tam、Hanyu Liu、Yasheng Wang、Yaxi Lu、Songyang Zhang等。JustRUA：GPT-6机器人使用智能体的受控研究。*arXiv*预印本 *arXiv:2611.16009*，2026年。[81] Yule Wen、Yixin Ye、Yanzhe Zhang、Diyi Yang和Hao Zhu。演化环境中的实时推理智能体。在第十四届国际学习表征会议，2026年。[82] Rui Yang、Hanyang Chen、Junyu Zhang、Mark Zhao、Cheng Qian、Kangrui Wang、Qineng Wang、Teja Venkat Koripella、Marziyeh Movahedi、Manling Li等。EmbodiedBench：面向视觉驱动具身智能体的多模态大语言模型综合基准测试。在国际机器学习会议，第70576–70631页。PMLR，2025年。[83] Zhejian Yang、Yongchao Chen、Xueyang Zhou、Jiangyue Yan、Dingjie Song、Yinuo Liu、Yuting Li、Yu Zhang,

Pan Zhou、Hechang Chen 等。智能体机器人：一种面向具身智能体视觉-语言-动作模型的仿脑框架。*arXiv* 预印本 *arXiv:2505.23450*，2025年。[84] Zhiqin Yang、Chenxin Li、Xiaomeng Hu、Yibin Liu、Weidong Huang、Jiankai Sun、Haitao Li、Zijian Wu、Yuzhi Huang、Fanding Huang 等。RobotWorld：对多模态智能体在多种任务和具身形态中的机器人应用进行基准测试。*arXiv* 预印本 *arXiv:2610.10409*，2026年。[85] Haoqi Yuan、Yu Bai、Yuhui Fu、Bohan Zhou、Yicheng Feng、Xinrun Xu、Yi Zhan、Börje F Karlsson 和 Zongqing Lu。Being-0：一种结合视觉语言模型与模块化技能的人形机器人智能体。*arXiv* 预印本 *arXiv:2503.12533*，2025年。[86] Wentao Yuan、Jiafei Duan、Valts Blukis、Wilbert Pumacay、Ranjay Krishna、Adithyavairavan Murali、Arsalan Mousavian 和 Dieter Fox。RoboPoint：一种用于机器人学中空间可供性预测的视觉-语言模型。载于机器人学习会议，第4005–4020页。PMLR，2025年。[87] Alex L. Zhang、Thomas L. Griffiths、Karthik R. Narasimhan 和 Ofir Press。VideoGameBench：视觉-语言模型能否通关热门电子游戏? *arXiv* 预印本 *arXiv:2505.18134*，2025年。[88] Yehang Zhang、Haojian Huang、Yifan Chang、Jianchong Su、Bohan Zhou、Yingjie Xu、Wosong Chen、Tianhao Zhou、Chenxu Wang、Tianyi Zhang 等。World Action Agent：通过世界动作演练利用视觉语言模型进行机器人操作。*arXiv* 预印本 *arXiv:2609.29964*，2026年。[89] Yixian Zhang、Huanming Zhang、Feng Gao、Xiao Li、Zhihao Liu、Yi Nie、Chunyang Zhu、Jiaxing Qiu、Yuchen Yan、Jiyuan Liu 等。Harness VLA：通过记忆引导智能体，引导冻结的视觉-语言-动作模型实现可靠的操作原语。*arXiv* 预印本 *arXiv:2607.08448*，2026年。[90] Enshen Zhou、Qi Su、Cheng Chi、Zhizheng Zhang、Zhongyuan Wang、Tiejun Huang、Lu Sheng 和 He Wang。Code-as-Monitor：用于反应式和主动式机器人故障检测的约束感知视觉编程。载于计算机视觉与模式识别会议论文集，第6919–6929页，2025年。[91] Jian Zhou、Xunyi Zhao、Gengze Zhou、Zerui Li、Sihao Lin、Jiajun Liu 和 Qi Wu。具身智能体掌控控制权：极简界面的零样本智能体在视觉与语言导航中可媲美工业规模策略。*arXiv* 预印本 *arXiv:2607.26148*，2026年。

一个基准测试详情

表7。 RUIBench 的 30 项任务。 来源：改编自一种心理学范式（心理学），或选自日常游戏、谜题、学校实验和家务（日常）。标记为 † 的任务使用两台 ARX X5 机械臂，其余任务使用一台 Franka Research 3。限制：时间上限以模拟分钟计，为回合预算乘以 8 秒，最多 30 分钟。每项任务提供 15 个开发任务实例和 20 个测试任务实例。

ID	Task	来源	成功率要求	限制
世界知识 基于先验知识进行推理				
WK1	汉诺塔	心理学 [70]	所有圆环都在右侧柱子上，最大的在底部，且每次移动都符合规则。	16.0
WK2	渡河问题	日常	狼、山羊和水桶顺利渡河，且未留下任何禁忌组合。	16.0
WK3	多米诺骨牌	日常	摆放多米诺骨牌，使释放的摆锤将它们撞倒并敲响铃。	8.0
WK4	摆放餐具†	日常	按餐桌礼仪要求，将叉子、刀和勺子摆放在盘子周围。	12.0
WK5	平衡梁	心理学 [69]	通过将给定的砝码放在秤盘上，使横梁保持水平。	8.0
WK6	井字棋	心理学 [20]	一场进行中的游戏中，凭借三个连续的叉号获胜。	5.3
上下文学习 使用先前展示的视觉示例				
CL1	倒带播放†	日常	视频中展示的一段曲调，在木琴上重现。	5.3
CL2	玩偶的名字	心理学 [9]	视频中指定的玩偶被放到床上，没有移动其他玩偶。	4.0
CL3	建造房屋	日常	积木按照卡片上的图案摆放，整个结构保持静止。	8.0
CL4	乌鸦与水罐	心理学 [5, 11]	按照视频所示，用冰壶抬起漂浮的球，并将其移到盘子里。	14.7
CL5	七巧板	日常	七块拼板全部放置在一个轮廓上，如同完成的示例。	12.0
CL6	马步	日常	根据新规定，黑王被马步棋子吃掉。	8.0
发现 物理探索				
DS1	谁的礼物?	日常	汤姆的礼物，通过查看礼物标签找到后，放在红色垫子上。	8.0
DS2	寻蛋	日常	找到藏在三个罐子之一下面的鸡蛋，并将其放进篮子里。	16.0
DS3	寻宝游戏	日常	用探测器找到四个贝壳下面的一枚硬币，并抬起一个贝壳。	9.6
DS4	假硬币	日常	用天平找出三枚硬币中较轻的一枚，并将其放入盘子里。	10.1
DS5	寻找酸	日常	用 pH试纸从三个相同的烧杯中找出酸。	10.1
DS6	焰色反应	日常	测试浸泡在未标记的盐类中的木条时产生的绿色火焰。	11.7
动态交互 响应时间敏感事件				
DI1	火箭发射†	心理学 [55]	通过找出如何按压它的两个按钮来发射火箭。	4.0
DI2	装载火车	日常	将两个玩具装入行驶中的火车上对应的车厢。	10.7
DI3	交通信号灯	心理学 [75]	只有绿灯亮时，才把三只小鸭带过道路。	30.0
DI4	西蒙说	心理学 [72]	及时执行每一条“西蒙说”指令，不执行任何其他指令。	28.3
DI5	打地鼠†	日常	除一只外，所有地鼠都在缩回前被击中。	30.0
DI6	甜点盖子	心理学。[8]	在转向的三个盖子中，有一个下面藏着一块馅饼，只抬起一个盖子。	15.5
高接触操作 接触期间进行精确调整				
CR1	形状分类器	心理学 [59]	四块积木从对应形状的孔洞中落下。	16.0
CR2	拧灯泡	日常	将灯泡拧入灯座，直到灯亮起。	8.0
CR3	冰壶	日常	一枚冰壶被推动，而非抓取，最终停在营垒内。	8.0
CR4	显卡	日常	一张显卡被竖起并插入主板插槽中。	12.0
CR5	夹起积木	心理学 [77]	两块玩具积木完全且端正地按压在塔上。	9.3
CR6	折叠T恤†	日常	一件从下摆折叠至肩部的T恤。	8.0

任务。表7列出了30项任务及其来源、成功标准和时间限制。11项心理学任务改编自问题解决、词语学习、因果学习、工具使用、自我调节、客体永久性和空间认知等经典范式。成功标准结合了目标状态与任务规则。例如，汉诺塔还要求每一步移动都符合规则，而寻宝游戏中智能体每次只能抬起一个贝壳。

任务实例与数据划分。任务实例由固定种子生成。种子设定物体位姿和隐藏变量，例如藏有鸡蛋的锅。每项任务有 15 个开发任务（种子 0 至 14）和 20 个测试任务（种子 15 至 34）。表 3 使用每项任务的前五个测试任务。表 4 和 6 以及表 5 的仿真列使用前。5 使用前五个。

限值与评分。每项任务都有回合预算和回合时钟时间限制，时限为预算乘以 8 秒，且最多为 30 分钟。每项任务的评分器根据回合结束时记录的对象状态判定是否成功，并可根据日志重新计算。成功要求达到目标状态、不违反任务规则，且物理审计未发出警报。审计会标记数值异常、超出关节限位的关节物体、速度超过 6 米/秒的物体，以及在没有接触的情况下跳跃超过 3 厘米的物体。

阴性对照。 脚本化基准解决了全部 70 个对照任务实例，以及 14 个任务中的前 5 个测试任务。每项对照测试都会移除一种信息后运行它。在没有示范的情况下，它无法完成 CL1、CL2、CL4 和 CL5 中的全部 20 个任务实例；CL3 会在场景中的卡片上写出目标，而 CL6 会以文本陈述其规则。仅根据第一帧进行规划时，它在发现和动态任务（DS1 至 DS5、DI5 和 DI6）的 35 个任务实例中成功完成了 4 个，而且每次都是靠猜测。在忽略回合中变化的情况下，它在目标或指令会发生变化的任务（CL6、DI2 和 DI4 ）的 15 个任务实例中成功完成了 2 个。因此，**经过测试的任务依赖于定义其套件的信息**。

观测。 SAM 3 [10] 会检测并分割渲染图像中的物体，传感器深度则将其提升为三维。在仿真中，每个相机位姿都带有 3 mm 和 0.3° 的标定误差，该误差在一个任务实例中保持不变。深度还带有 1 mm 加上距离 0.3% 的逐像素误差，该误差在一个视点内保持不变，因此重复观察不会将其平均掉。因此，位置来自传感器而非仿真器；与真实机器人相比，其误差更小。对于真实机器人，两个摄像头放置同一物体时的距离最多可相差 1.5cm（第 5.4 节）。每次回复使机械臂移动后，测试框架都会返回所有摄像头的最新图像及其结果。

B 接口详情

受控接口设计。 RUI 的差异仅在于每个动作指定的内容，以及控制权何时返回给智能体（表 8）。每个 RUI 都提供相同的 24 种工具，用于感知、记忆、规划和任务，因此智能体在每种 RUI 下都拥有相同的感知和规划手段。程序可以调用其抽象中的所有逐动作工具，但 任务_放弃_除外。表 4 中的动作序列变体运行一个程序，该程序只能执行动作、等待和停止，并在程序结束时返回动作结果。因此，与逐动作形式一样，智能体仅在程序之间观察场景并做出决策。

C 实验细节

C.1 小模型

训练数据。 对于 12 个固定 RUI 中的每一个，训练集都包含来自与测试随机种子不重合的随机种子的 300 条轨迹。其中，90 条是 Qwen3.8-Max 成功回合，另外 210 条来自脚本策略。Qwen3.8-Max 在与小模型相同的最小循环和 RUI 中运行，我们只保留基准测试评分器标记为成功的回合，不包含其中的推理。脚本策略依据特权状态进行规划，但只通过 RUI 的命令执行动作，并使用与测试时相同的工具、时钟和预算，因此每条轨迹都是该 RUI 的有效回合。

表 8. 各 RUI 暴露的内容。 动作抽象仅在其动作命令上有所不同，除步骤外的每种抽象都会增加 机器人_夹爪。交互形式决定这些命令的发出方式以及控制权何时返回给智能体，而共享工具在每种 RUI 中都完全相同。

组件	工具	单条指令指定或返回的内容
动作抽象 因不同 *RUI* 而异		
关节	机器人_移动_关节，机器人_求解_逆运动学	关节角度；逆运动学指令返回某个姿态对应的角度
Delta	机器人_移动_Delta	相对于夹爪的平移和旋转
Step	机器人_步骤	以 0.25/1/4 厘米为一步，或转向 2/5/15°， *n* 次，或进行抓取
Pose	机器人_移动_至	夹爪的绝对位置和方向
点	机器人_移动_至_像素	通过深度将像素抬升至三维，并赋予方向
技能	技能_{拾取、放置、按压、push，move，转向}	对对象、像素或点进行的一整套操作
交互形式因 *RUI* 而异		
逐动作	该抽象的命令	每条回复中，每条机械臂最多执行一个动作
程序	代码_运行	使用逐动作工具的 Python 程序
RUI-Free	两种形式中每种抽象的工具	任意形式的任意抽象
动作序列	代码_仅对动作命令运行	按顺序排列的动作，结果置于末尾（表 4 仅）
共享工具在每个 *RUI* 中都相同		
摄像机	摄像机_{look, zoom}	固定摄像头、头部摄像头或腕部摄像头的图像
视觉	视觉_{detect, segment, 抬起, pose，点 ，标记 ，跟踪}	检测、掩码、三维姿态、指向、标记、跟踪
机器人	机器人_{状态、等待、停止}	关节和夹爪读数、等待、停止
记忆	记忆_{笔记，回忆， 坐标系，场景}	笔记、先前图像和场景记录
Plan	计划_{抓取， 运动， ik}	抓取、无碰撞路径、姿态可达性
Task	任务_{信息， 演示、事件，完成，放弃_	指令、示范、事件、回合结束

大约一半的脚本化回合会在 30% 到 60% 的移动机械臂的回复中，为手的目标位姿加入误差；无论抽象方式如何，位置上的标准差最高为 8 毫米，偏航角上的标准差最高为 4°。随后，策略会根据由此产生的状态重新规划。执行时带有此类误差的回复会保留在后续步骤的历史记录中，但不会作为训练目标，因此模型学会的是如何修正，而不是如何制造错误。

微调与推理。每个 RUI 都有自己的 LoRA适配器 [32]，秩为 64（$\alpha = 128$），应用于语言模型的所有线性层，而视觉编码器保持冻结。全部 12 个适配器共享相同的数据方案、轨迹数量和训练计划。训练共进行 10 个 epoch，学习率为 10^{-4}，采用余弦衰减，每个批次包含 32 个请求。每个训练样例都是一个请求，与 RUIKit 发送的请求完全一致，包含系统提示词、工具、文本历史记录和最新的六张图像；损失仅针对回复计算。测试时，模型在最小循环中运行，不进行思考，使用与前沿智能体相同的工具、限制和 8 秒时钟，对每个 RUI 的 150 个测试回合进行测试。

回合长度和成本。 能够命名目标的抽象可缩短小模型的回合。在逐动作形式中，Point 和 Skill 回合的回复数中位数分别为 29 和 23，而 Joint 和 Delta 分别为 53 和 36。程序将最长的 Joint 回合缩短至 37.5 条回复，这与其仅在 Joint 上取得的提升一致（第 5.2 节）。训练成本按每 H200 GPU 小时 2.68 美元、每 B200 GPU 小时 4.06 美元计费。测试成本则按照每个测试回合的实际耗时，计费一块专用推理GPU以保持模型加载状态，就像已部署机器人一样，其中包括执行动作所花费的时间。

表9. 两个智能体编写的程序。 首次测试任务中的六种程序 RUI（每个智能体180个回合）。动作：每个程序的平均动作数，仅进行观察的程序计为零； ≥5 动作、循环和观察：至少包含五个动作、包含循环，以及在执行动作后进行感知调用的程序所占比例（%）；行数：每个程序的行数中位数。配对胜利：仅有一种形式成功的任务和抽象配对中，程序形式和逐动作形式各自获胜的配对数。

智能体	动作	≥5 动作	循环	观察	行数	配对胜利
Claude Code	3.3	21	58	54	7	32 : 8
Codex	0.9	0	10	8	1	20 : 14

表10. 表4交互形式消融 按套件和抽象划分。 Claude Code 在首个测试任务实例上的成功率（%），每个套件包含18个回合，每种抽象包含30个回合。套件：世界知识（WK）、上下文学习（CL）、发现（DS）、动态交互（DI）和高接触操作（CR）。

Form	套件					抽象			All
	WK	CL	DS	DI	CR	Delta	Step	点	
逐动作	83.3	33.3	66.7	50.0	55.6	66.7	46.7	60.0	57.8
动作序列	83.3	66.7	72.2	55.6	50.0	70.0	60.0	66.7	65.6
程序	100.0	66.7	88.9	72.2	50.0	80.0	70.0	76.7	75.6

C.2 程序与反馈

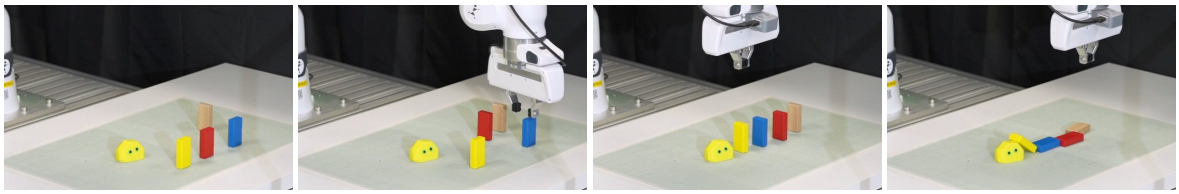
当智能体组合动作时，程序会发挥作用。 程序使 Claude Code 的成功率提高了 11.0 个百分点，但仅使 Codex 提高了 2.7 个百分点（表 3）。表9 解释了其中的差异。Claude Code 的程序平均包含 3.3 个动作，其中 54% 会在执行动作后观察世界。Codex 的程序平均包含 0.9 个动作，其程序中位数只有一行。在相同任务和抽象上，在只有一种形式成功的 40 组配对中，Claude Code 的程序赢下了 32 组，而 Codex 赢下了 34 组中的 20 组。因此，**当智能体利用程序形式结合带有反馈的动作时，程序形式才能发挥效用。**

按套件划分的组合与反馈。 表 10按套件和抽象细分了表 4 中的交互形式消融。对于三种抽象中的每一种，成功率从逐动作执行到动作序列，再到程序逐步提升；各套件也遵循了第 5.3 节所述的模式。

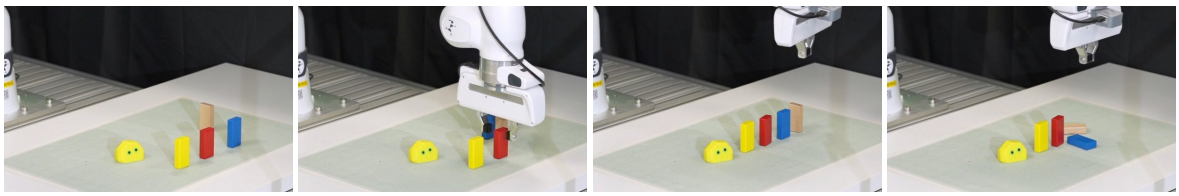
C.3 真实机器人

设置。 真实实验台使用第 5.4 节中的 Franka Research 3 机械臂，配有 Franka Hand、固定在桌子对面的 RealSense D435，以及安装在腕部的 D405。测试框架、RUI、工具和感知代码在真实机器人后端上运行，无需任何改动。后端以 50 Hz 向阻抗控制器发送设定点，控制器在速度限制内沿最小跃度路径运动。每次运动前，后端都会在运动学模型上预览路径，并检查工作空间、手部倾斜度和关节限位。无法完成的运动会在最后一个可到达的点停止，并返回 *blocked*，与仿真中的情况相同。运动过程中，如果跟踪误差、关节位置或关节速度超过限值，机械臂就会停止。对于真实传感器，感知不会添加合成噪声。桌上的 ChArUco 标定板用于校准两台摄像机，时间按挂钟时间计算，包括智能体的推理时间。另有一台智能体无法看到的独立摄像机，会录制每个回合，供图表和视频使用。操作员会监督每个回合以确保安全，并可暂停机械臂或终止回合；被终止的回合将计为失败。

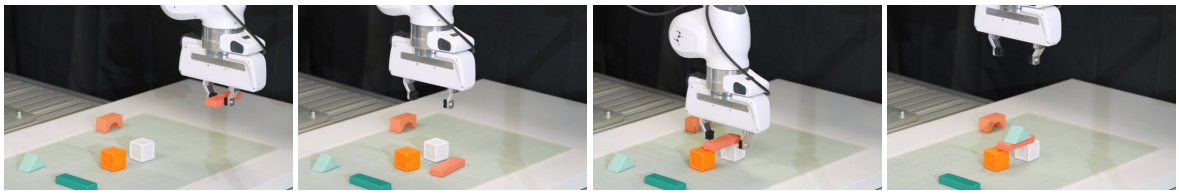
(a) 多米诺骨牌（WK3），Claude Code 使用 Delta：27 条回复后成功



(b) 多米诺骨牌（WK3），Claude Code 使用 Program-Delta：5 条回复后失败



(c) 建造房屋（CL3），Codex 使用技能：31 条回复后成功



(d) 形状分类器（CR1），Claude Code 使用技能：64 条回复后成功

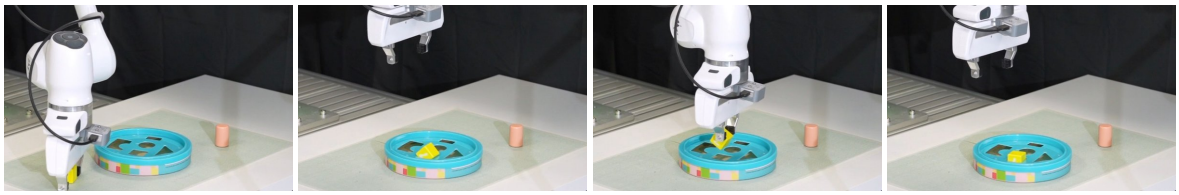
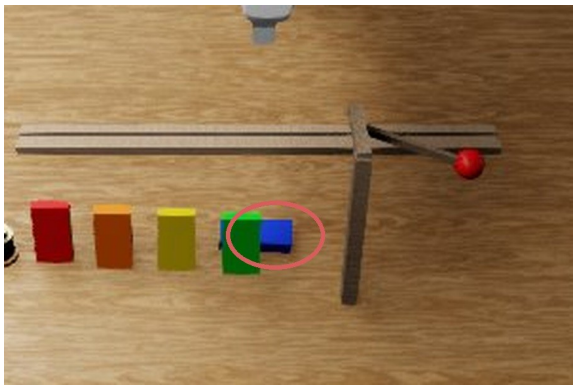


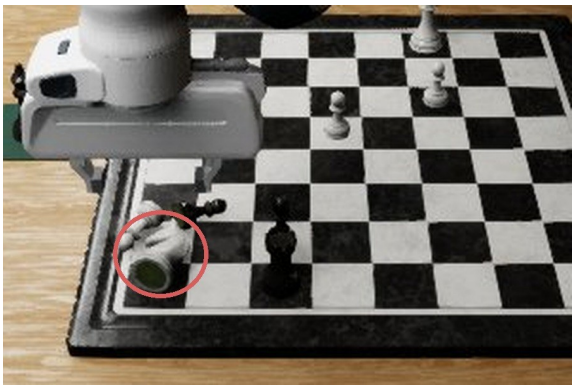
图4。 首个布局上的真实机器人运行轨迹。 每行展示第三人称摄像机拍摄的同一回合中的四帧。在相同的任务和布局下，Claude Code 使用逐动作 Delta 时会推倒整条多米诺骨牌（a），但执行其程序后，只有蓝色多米诺骨牌会侧向倒下（b）。逐动作回合还能从横梁落在支撑物旁边（c）以及立方体倾斜着停在盖子上（d）的情况中恢复。

任务与评分。 每项真实任务都保留其仿真对应任务的目标，并使用实体物体（图 2）；例如，多米诺骨牌必须撞倒玩具，而不是敲响铃。每项任务在五种固定布局上运行，所有 RUI 和智能体使用相同布局，轮次预算和时间限制也与仿真相同。由于物体各不相同，发现3在每个世界内比较接口。每个回合结束后，操作员依据真实任务的书面标准评判成功与否，并以 0.5 分为步进，在 1 至 5 分之间为每次成功评分。失败计为 0 分，评分再按比例换算为 100 分制。这些评分保留了表 5中的排序：对于 Claude Code（30.1 对 27.7）和 Codex（44.1 对 41.5），逐动作执行的表现都优于程序。

运行过程。 图 4 展示了首个布局中的四个真实机器人回合。在多米诺骨牌挑战中，使用逐动作 Delta 的 Claude Code 搭建了一条骨牌链，最终将石制玩具撞倒（a）。使用 Program-Delta 时，它放置好多米诺骨牌后便结束了回合，仅用了 5 条回复，而非 27 条，但骨牌链在第一块骨牌处断开了（b）。成功的回合通常会先放置一个对象、检查结果，再进行修正，(c) 中的横梁和 (d) 中的立方体便展示了这一过程。附录 D 中还展示了另外 17 个真实机器人回合和 29 个仿真回合。D。



(a) 多米诺骨牌（WK3），使用 Joint 的 Codex
智能体： “将四块直立的多米诺骨牌排成一列，朝向铃，并收回机械臂；操作过程中一块蓝色多米诺骨牌倒下了。”
评审器： “最靠近摆锤的蓝色多米诺骨牌平躺着，导致摆锤前方没有一条连续竖立的骨牌链可供其推倒至铃处。”



(b) 马步（CL6），使用 RUI-Free 的 Codex
智能体： “移除了黑王，并将白马放在它原来的位置；恢复后，白马侧躺着，一次失败的手腕旋转扰动了附近的两个兵。”
评审器： “……白马位于棋盘边缘附近，而不是占据王的位置。”

图 5. 仿真中的虚假完成。 每一列都截取了智能体发出 任务_完成 指令前收到的最后一张相机图像，红圈标出了评审器指出的失败。Codex 注意到了倒下的棋子，却报告称目标已达成；而同一模型在全新上下文中仅凭这些图像进行判断时，则识别出了失败：观察到失败并不意味着一定能识别失败。

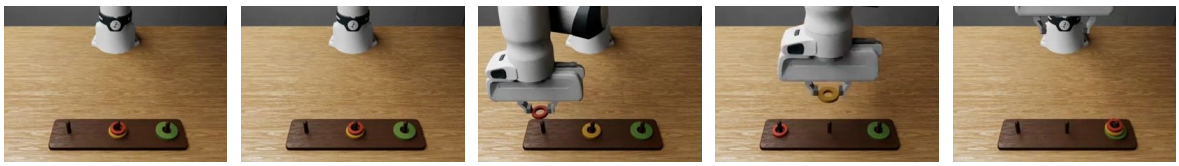
C.4 完成判断

事后评审器。 评审器是在全新上下文中运行的 GPT-6 Astra。它接收指令、提示，以及测试框架在智能体调用 任务_完成前返回给智能体的每个摄像机的最后图像。它仅根据这些图像判断成功或失败，并说明理由。表 6中的事后数据集保留了评审器能够从这些图像中识别成功的 23 个任务。在其中六个任务中，最终帧难以显示结果，例如演奏过的曲调（CL1）或已识别的酸（DS5），因此评审器会拒绝许多实际已完成的任务。我们排除了这些任务（CL1、CR2、DI4、DI5、DS5 和 WK2），以免评审器错误被计为可见失败；同时还排除了 CR6，因为两个智能体在第一次测试任务中都未成功。在这 23 个任务中，评审器确认了 Claude Code 实际完成任务的 95.4%，以及 Codex 实际完成任务的 98.7%。因此，它对错误声明的检测反映的是最终帧所显示的失败，例如图 5中展示的两个 Codex 回合。

D 回合展示集

图 6 至 10 展示了在首次测试任务中解决的 29 个任务实例各自的一次成功仿真回合，由两种智能体交替完成，并涵盖全部 13 个 RUI。图 11至13展示了首次布局中的另外 17 次真实机器人回合，并根据智能体的行为进行分组。每一行列出任务、智能体、RUI 和回复数量，其坐标系按时间从左到右排列。

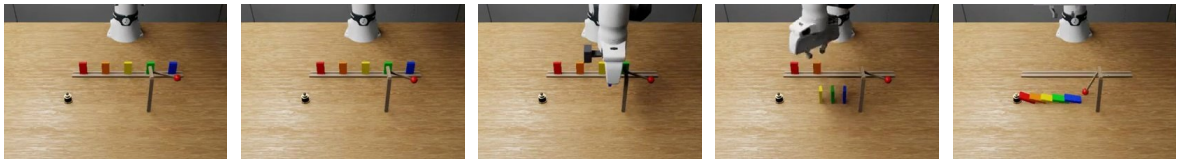
WK1，汉诺塔：Claude Code 使用程序技能，经过 8 条回复后成功



WK2，渡河问题：Codex 使用程序技能，经过 23 条回复后成功



WK3，多米诺骨牌：Claude Code 使用 Program-Pose，经过 11 条回复后成功



WK4，摆餐具：Codex 使用程序关节，经过 25 条回复后成功



WK5，平衡木：使用 RUI-Free 的 Claude Code，在回复 9 次后成功



WK6，井字棋：使用 Program-Point 的 Codex，在回复 11 次后成功

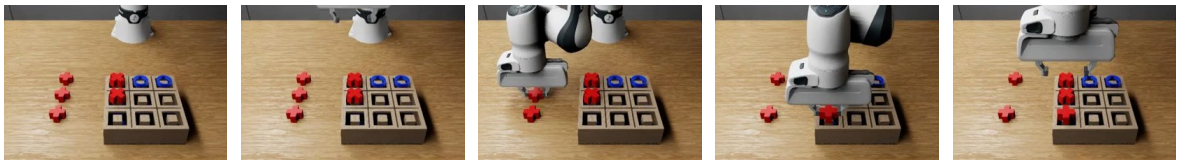
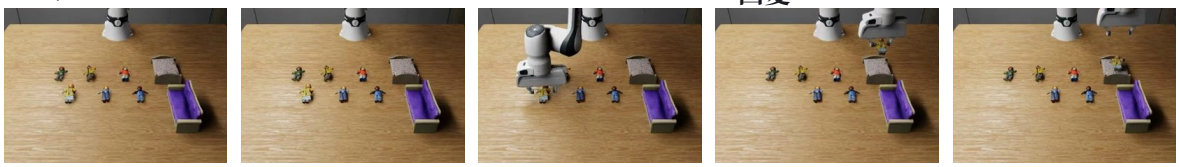


图6。模拟展开：世界知识。每项任务在第一个测试任务实例中成功完成的一个回合，画面来自智能体自身观察到的摄像机视角（固定摄像机；对于双臂任务，则为头部摄像机）。

CL1，倒带播放：Claude Code 搭配 Program-Point，经过 10 条回复后成功



CL2，玩偶的名字：Codex 搭配 Program-Delta，经过 12 条回复后成功



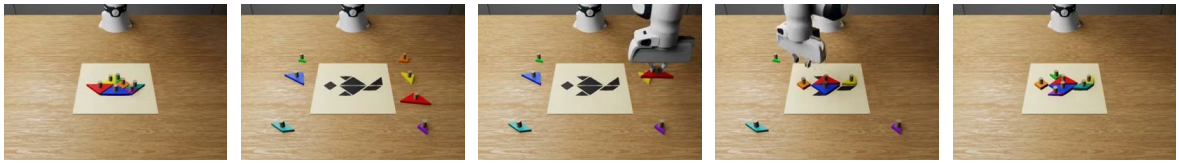
CL3，建造房屋：Claude Code 搭配姿态，在 26 条回复后成功



CL4，乌鸦与水罐：Codex 搭配姿态，在 36 条回复后成功



CL5，七巧板：Claude Code 搭配程序技能，在 24 条回复后成功



CL6，马步：Codex 搭配 Program-Pose，在 44 条回复后成功

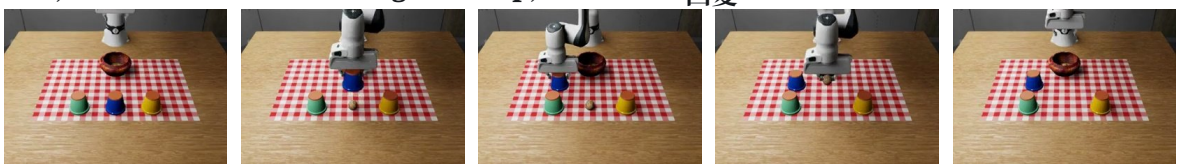


图7。 模拟展开：上下文学习。 每个任务在首个测试任务实例中的一次成功回合，视角来自智能体自身观察的摄像机（固定摄像机，或双臂任务中的头部摄像机）。

DS1，谁的礼物？：Claude Code 搭配 Delta，经过 13 条回复后成功



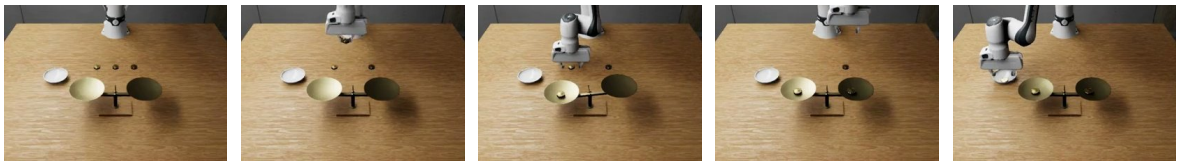
DS2，寻蛋： Codex 搭配 Program-Step，经过 28 回复



DS3，寻宝游戏：Claude Code 搭配 Program-Delta，在回复 16 次后成功



DS4，假硬币：Codex 搭配步骤，在回复 39 次后成功



DS5，寻找酸：Claude Code 搭配 Program-Point，在回复 32 次后成功



DS6，焰色反应：Codex 搭配 RUI-Free，在回复 24 次后成功

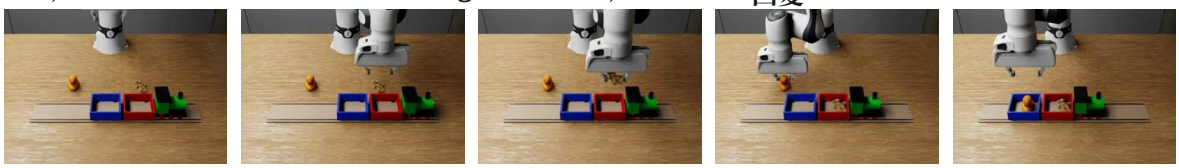


图8。 模拟展开：发现。 每项任务在第一个测试任务实例中的一次成功回合，视角为智能体自身观察到的摄像机画面（固定摄像机；双臂任务则为头部摄像机）。

DI1，火箭发射：Claude Code 搭配技能，5 条回复后成功



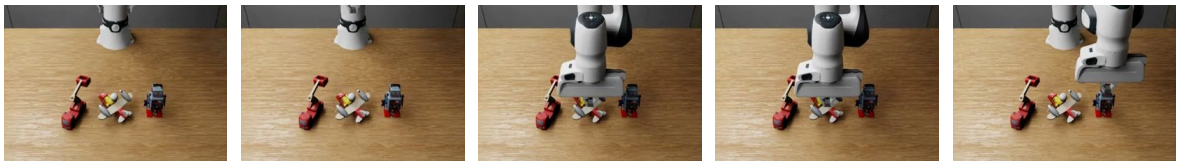
DI2，装载火车 ：Codex 搭配 Program-Point，15 条回复



DI3，交通信号灯：Claude Code 使用 Program-Step，经过 12 条回复后成功



DI4，西蒙说：Codex 使用技能，经过 10 条回复后成功



DI5，打地鼠：Claude Code 使用姿态，经过 23 条回复后成功



DI6，甜点盖子：Codex 使用程序关节，经过 27 条回复后成功

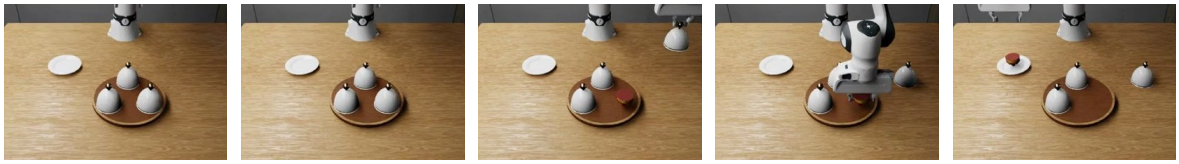
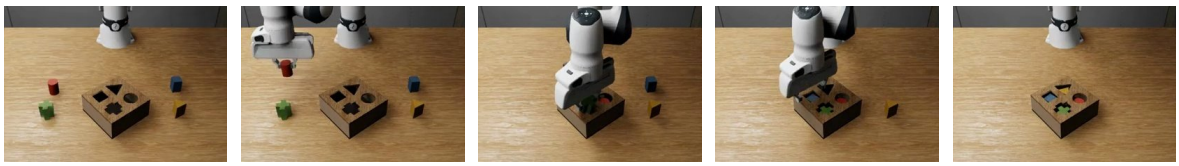
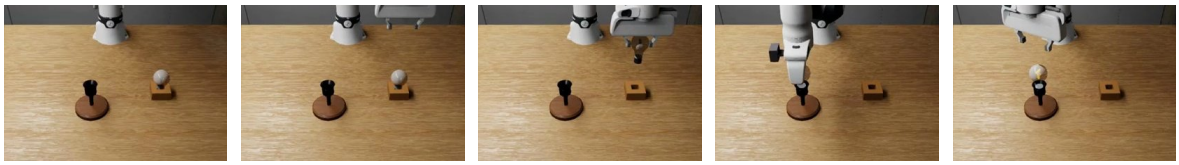


图 9。 **模拟展开：动态交互。** 每项任务在首次测试任务实例中的一次成功回合，视角来自智能体自身观察到的摄像机画面（固定摄像机，或双臂任务中的头部摄像机）。

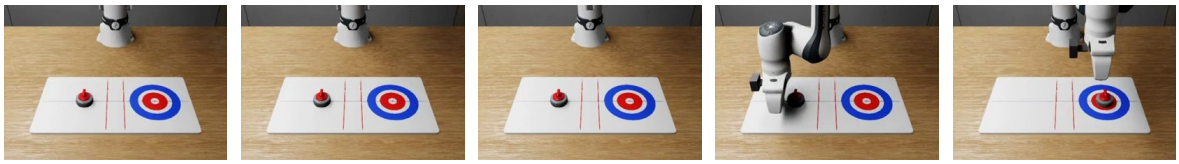
CR1，形状分类器：使用关节的 Claude Code，经过 53 条回复后成功



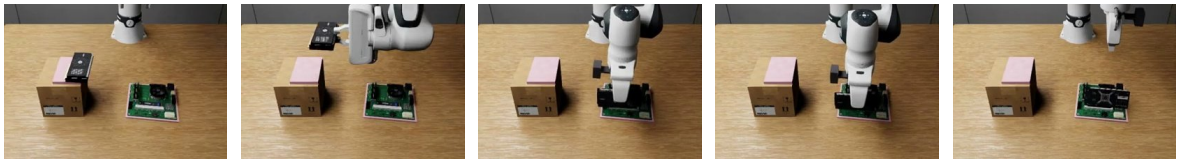
CR2，拧灯泡：Codex 搭配关节，21 条回复后成功



CR3，冰壶：Claude Code 搭配点，11 条回复后成功



CR4，显卡：Codex 搭配程序关节，37 条回复后成功



CR5，夹起积木：Claude Code 搭配步骤，43 条回复后成功

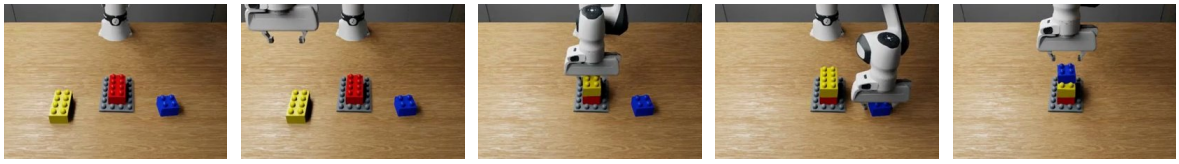
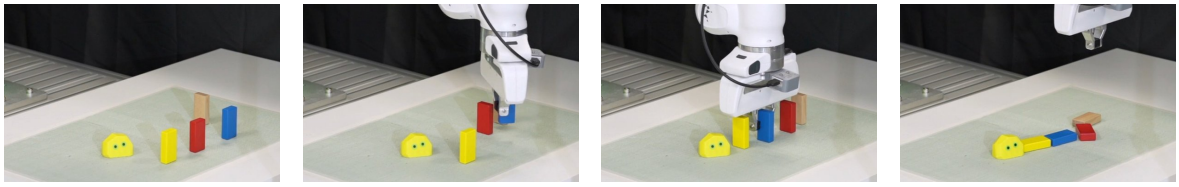
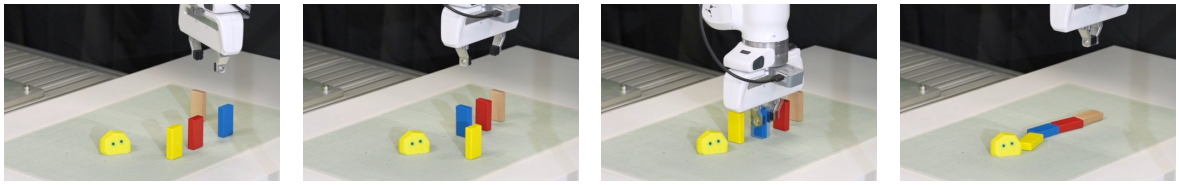


图 10。 模拟展开：高接触操作。 每项任务在第一个测试任务实例中的一次成功回合，视角取自智能体自身观察所用的摄像机（固定摄像机；对于双臂任务，则为头部摄像机）。折叠T恤（CR6）在该任务实例中没有成功回合。

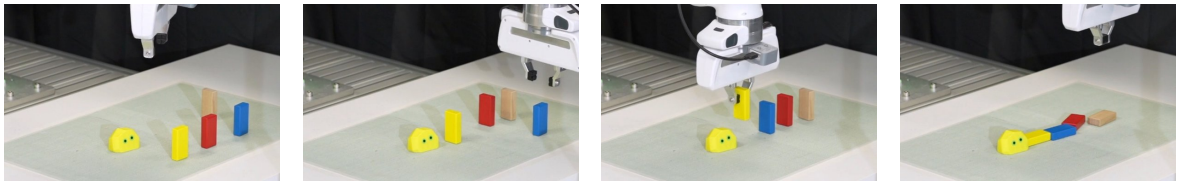
(a) 多米诺骨牌（WK3），使用姿态的 Claude Code，经过 44 条回复后成功：竖起多米诺骨牌，然后逐个调整



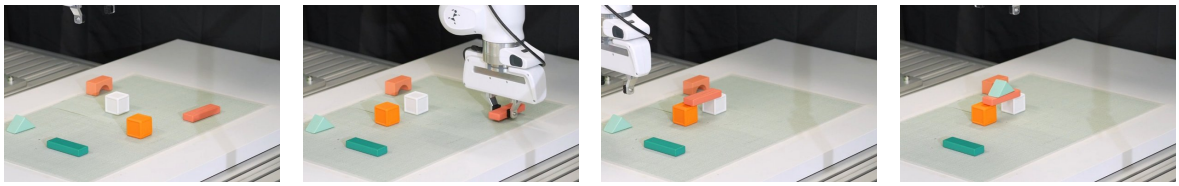
(b) 多米诺骨牌（WK3），Codex 使用姿态，53 条回复后成功率：逐个扶正这一排多米诺骨牌



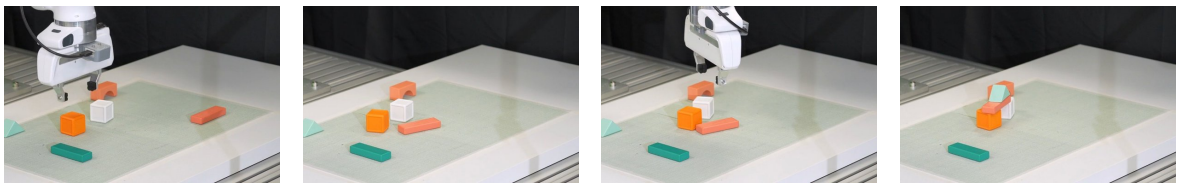
(c) 多米诺骨牌（WK3），Codex 使用 Delta，41 条回复后成功率：在首次排列后重新摆放一张多米诺骨牌



(d) 建造房屋（CL3），Codex 使用姿态，32 条回复后成功率：调整横梁在支撑物上的位置



(e) 建造房屋（CL3），Claude Code 使用 RUI-Free，13 条回复后成功率s:横梁滑落了，于是它将立方体移近一些，再次放置横梁



(f) 建造房屋（CL3），使用 Program-Pose 的 Claude Code，经过 10 条回复后成功：营垒倒塌了，程序便再次将其建起来

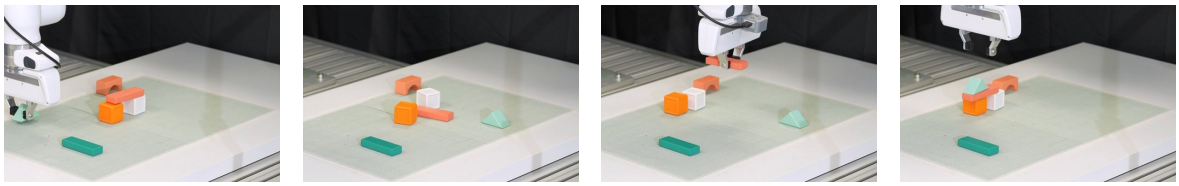
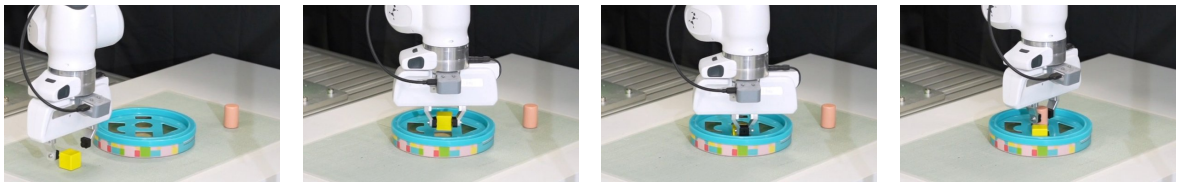
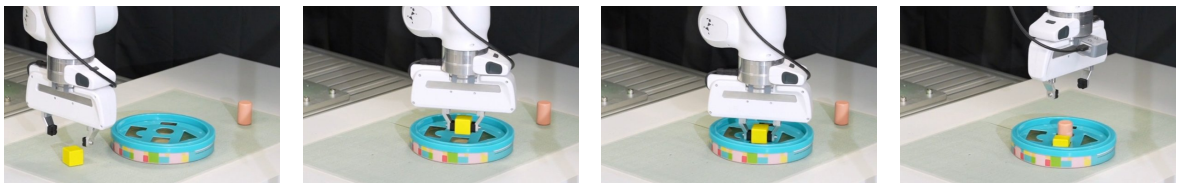


图 11. 真实机器人运行轨迹：纠正放置。 每行展示了第一种布局下一个回合的四个坐标系画面，来自第三人称摄像机，并附有任务、智能体、RUI 和回复数量。智能体会检查已放置的物体并进行调整，直到排列可行。

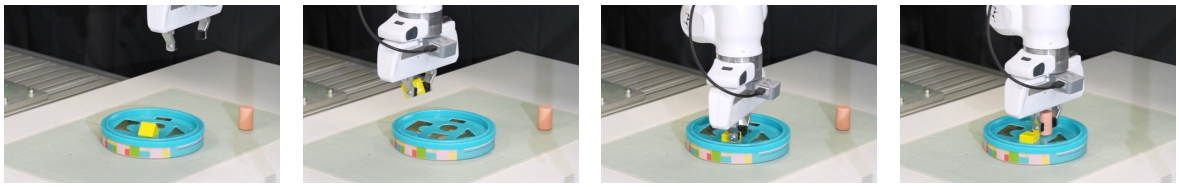
（a）形状分类器（CR1），使用 Codex 和 Delta，经过 46 条回复后成功率：
立方体卡在孔中后被推动进去



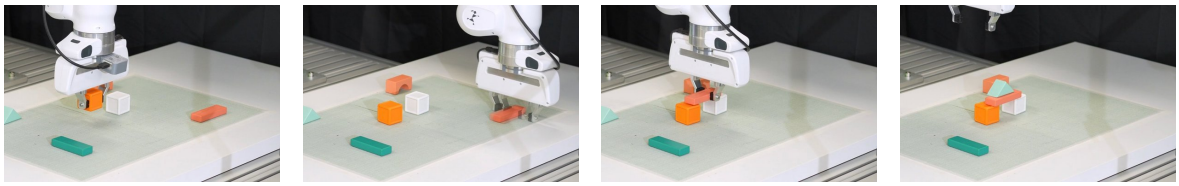
（b）形状分类器（CR1），使用 Claude Code 和 Program-Point，经过45条回复后成功率：将倾斜的立方体按压进孔中



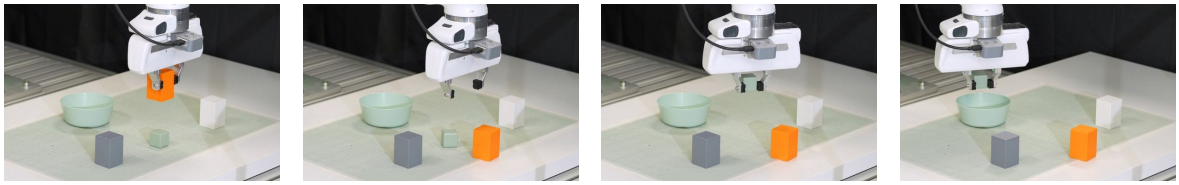
（c）形状分类器（CR1），使用 Claude Code 和 Pose，经过 89 条回复后成功率：按压、重新抓取并轻推立方体进入孔中



(d) 建造房屋（CL3），Claude Code 搭配 Delta，在 35 条回复后成功：将手转动 90° 以抓住横梁



(e) 寻蛋（DS2），Claude Code 搭配步骤，在 42 条回复后成功：将手转向小立方体的各个面



(f) 寻蛋（DS2），Claude Code 搭配点，在 47 条回复后成功：倾斜夹爪以伸入碗中

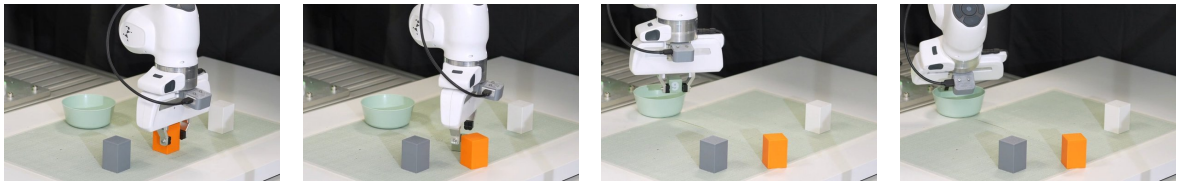
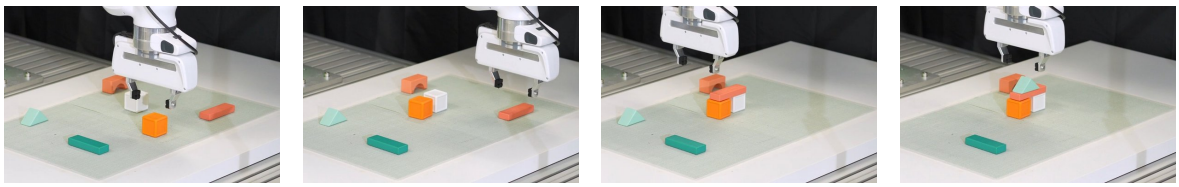
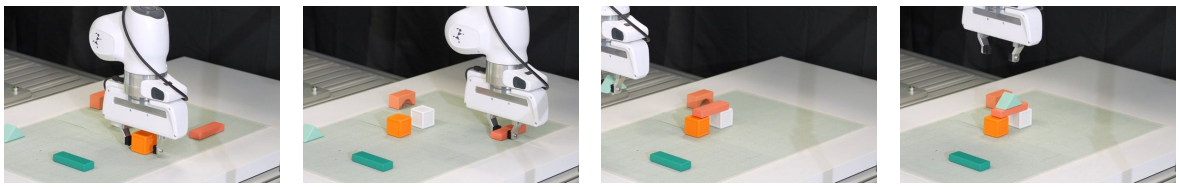


图12. 真实机器人运行轨迹：接触与手部姿态。每行展示第一种布局下一个回合的四个坐标系画面，画面来自第三人称摄像机，并附有相应的任务、智能体、RUI 和回复数量。智能体会将卡住或倾斜的部件推入孔中，并转动或倾斜手部，使其适应对象和目标。

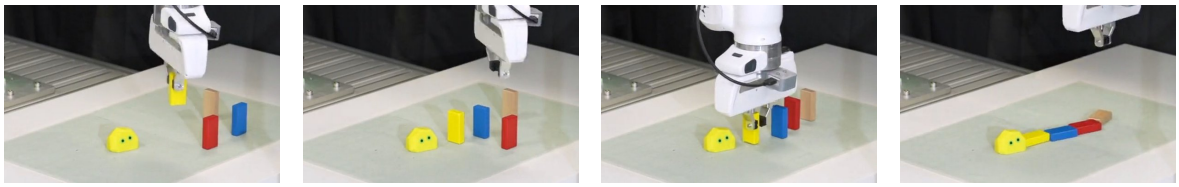
(a) 建造房屋（CL3），Claude Code 使用 Program-Step，8 条回复后成功率：通过少量程序建造营垒



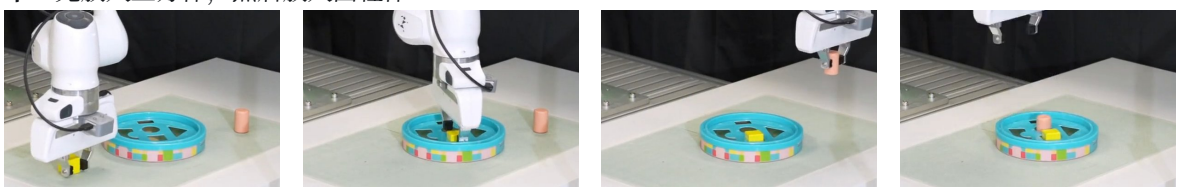
(b) 建造房屋（CL3），Claude Code 使用 Program-Delta，8 条回复后成功率：放置底座、横梁和屋顶



(c) 多米诺骨牌（WK3），Codex 使用 Program-Pose，45 条回复后成功率：在首次排列后调整多米诺骨牌



(d) 形状分类器（CR1），Codex 使用 Program-Point，37 条回复后成功率：先放入立方体，然后放入圆柱体



(e) 寻蛋（DS2），使用姿态的 Codex，回复 22 次后成功：移动橙色方块，找到它下面的小方块，并将其放入碗中

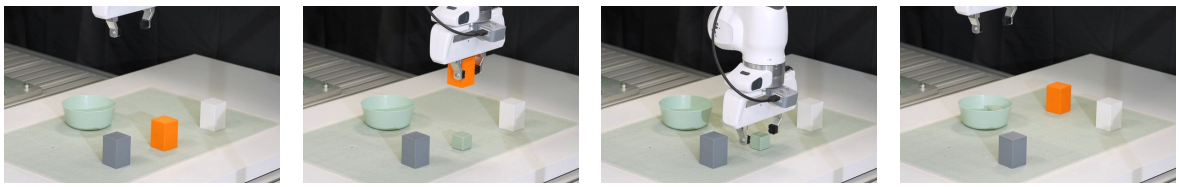


图 13。 真实机器人运行轨迹：程序与搜索。 每行展示第一个布局中一个回合的四个坐标系，画面来自第三人称摄像机，并附有任务、智能体、RUI 和回复次数。Claude Code 的程序可在每次回复中组合多个动作（a、b），而程序回合仍能修正自身操作（c）。