arXiv:2611.32803v1 [cs.RO] 24 Nov 2026

MRBench：异构多机器人操作中的物理必要性推理基准测试

Yanzhe Chen* Zhijun Cao* Ruihe An* Chang Gong* Mike Zheng Shou†
Show Lab，新加坡国立大学
*共同贡献 †通讯作者



图 1。**物理必要性取决于任务需求和可用资源**。左：**两个物理瓶颈**：并发任务需要同步动作；可达性任务需要空间覆盖。右：**相同目标，不同团队**：三个机械臂处于精确状态；一个冗余机械臂处于盈余状态；处于短缺状态而不可行；两个机械臂与不同机器人进行交换，或使用借用的被动替代体。

摘要

近期多机器人操作领域的进展展示了机器人如何协作完成复杂任务。然而，协同成功并不能揭示究竟需要哪些机器人。同一任务可能因机器人能力发生变化、资源不可用，或物体取代机器人支持，而需要不同的团队。我们认为，可靠的多机器人操作需要物理必要性推理：理解任务需求、哪些资源能够满足这些需求，以及哪些资源是真正必要的。我们引入***MRBench***，这是一个针对异构多机器人操作中物理必要性推理的基准测试。它包含八项具有不同并发性和可达性约束的任务，每项任务都在五种配置下进行评估；这些配置会改变所需资源，同时保持任务目标不变。我们的评估表明，现有的协调智能体在使用参考团队时表现良好，但当机器人具身形态发生变化，或被动对象能够取代机器人支持时，表现就会变差。为弥补这一差距，我们提出***MRScissors***，将任务需求与测得的物理能力分离，并通过约束搜索确定最小可行团队。在两种模型骨干网络上，*MRScissors* 的团队选择准确率分别达到 *85.4%* 和 *91.2%*，较最强基线方法分别高出 *10.4* 和 *9.5* 个百分点。总的来说，这些发现揭示了协同机器人与理解其物理必要性之间的关键差距。

网站： https://showlab.github.io/MRBench

1. 引言

“若无必要，切勿设定多个实体。”——威廉·奥卡姆

基础模型智能体的最新进展，使机器人能够通过任务分

解、协作规划和共享执 [4, 6, 28, 29, 36]行来协同完成操作任务。**然而，成功的协同并不能揭示实际需要哪些机器人**。在实际工作单元中，机器人可能无法使用、被不同模型的机器人替换，或由简单夹具接管其角色。这引出了一个超越协同的问题：一项任务真正需要哪些物理能力，以及需要哪些机器人或物体来提供这些能力？

答案取决于任务的物理约束，而不仅仅是机器人的数量。有些任务要求同时执行多个独立动作，而另一些任务则涉及超出单个机械臂可达范围的位置（图1A）。关键在于，**所需的物理角色并不总是需要额外的机器人**。如图1B所示，在参考配置下，同一个番茄导流槽任务需要三条机械臂；移除一条关键机械臂后，任务变得不可行，但如果另一条机械臂的可达范围更大，或使用合适的滑槽替代缺失的支撑，则两条机械臂即可完成任务。任务目标保持不变，但其最小可行团队发生了变化。这促使我们进行**物理必要性推理**：确定任务的需求、可用资源能够提供什么，以及哪种组合是最低限度的充分条件。

经典的多机器人任务分配研究基于能力进行团队组建 [18, 30]，，而近期的具身智能体则将机器人选择与协同扩展到了异构团队 [4, 27, 29, 56]。然而，现有基准测试大多在固定团队中评估协同能力，或从可用机器人中进行选择 [36, 42]。它们并未系统地测试智能体能否推断任务需求、评估物理能力，以及识别被动对象何时能够替代机器人的角色。**成功执行只能证明团队是足够的，并不能证明其是必要的**。

为探究这一差距，我们引入**MRBench**，这是一个用于异构多机器人操作中物理必要性推理的基准测试。MRBench 包含八项任务，涵盖多臂必要性的两类来源：并发性和可达性约束（图2）。每项任务均在五种配置下进行评估——精确、盈余、短缺、交换和借用——这些配置在保持目标不变的同时改变可用的物理资源。真实最小团队根据几何约束和精确搜索计算得出，可行解决方案则在 Isaac Sim 中进行验证 [39]。这一设计检验智能体能否区分必要、冗余、缺失和可替代的机器人支持。

我们的评估揭示了一个显著的间隙：**能够成功协同参考团队的智能体，在物理资源发生变化时往往举步维艰**。现有的协同方法在参考配置下能够取得较高的准确率，但当机器人具身形态被交换，或被动对象可以替代机器人的角色时，其性能会大幅下降。为弥合这一间隙，我们提出了**MRScissors**，将任务需求与测得的物理能力分离。与团队无关的任务表示能够识别所需的

表 1。**最相关工作**。比较依据包括所确定的团队、答案所表达的内容以及其评估方式（异构性：异构）。✓：已报告； △：证据较为有限； ×：缺失。

Work	Team		答案		评估	
	异构性 arms	选择 team	不可行答案	被动替代体	匹配的变化	物理学执行
机器人团队的基础模型智能体						
RoCo [36]	✓	×	×	×	×	✓
SMART-LLM [29]	×	✓	△	×	×	✓
HMAS-2 [6]	×	×	×	×	×	✓
EMOS [4]	✓	△	×	×	×	×
CoEnv [28]	✓	×	×	×	×	✓
Tool-RoCo [56]	✓	✓	×	×	×	✓
多机器人基准						
RoboFactory [42]	×	×	×	×	×	✓
VIKI-Bench [27]	✓	✓	×	△	×	×
MRBench（我们的工作）	✓	✓	✓	✓	✓	✓

任务，物理工具决定哪些机器人或物体能够执行这些任务，而约束搜索则生成一个最小的可行团队，并给出可验证的分配方案。

在两种模型骨干网络上，MRScissors 的团队选择准确率分别达到 85.4% 和 91.2%，比最强基线方法分别高出 10.4 和 9.5 个百分点。在模型生成的能力上运行的同一求解器则分别落后 19.2 和 13.1 个百分点，凸显了物理测量的重要性。此外，当选定的团队正确时，MRScissors 的计划所实现的执行成功率是竞争方法的两倍以上。然而，被动替代仍然具有挑战性，揭示了协调机器人与理解其物理必要性之间长期存在的间隙。我们的贡献主要有以下三点：

• **实证洞见**。 我们揭示了这样一个间隙：随着机器人能力和环境支持发生变化，多机器人协调与物理必要性推理之间存在差距。

• **MRBench**。 我们提出了一个基准测试，用于评估在任务目标不变的情况下，不同异构配置之间的可行性、最小性和功能替代。

• **MRScissors**。 我们提出了一个将任务需求与测得的物理能力分离的框架，从而提升团队选择和下游执行效果。

2. 相关工作

多机器人任务基准测试。基准测试已从单个机械臂 [34] 扩展到两个 [5, 19, 25, 40, 59] 乃至多个机械臂，并围绕协同构建任务：共享一张桌子的机械臂（RoCoBench [36]）、组合约束（Robo- Factory [42]）、人形机器人团队（CoHuB [41]）以及与人类伙伴协作（PARTNR [2]）；其他套件则在固定任务上改变场景、指令、模型或测试框架 [7, 12, 14, 15, 54, 60]。这些紧密耦合的任务需要团队协作完成，因为单个机器人会因手不够或可达性不足而无法独立完成，

因此团队成功运行只能说明团队足以完成任务，而不能说明每个机械臂都是必需的。VIKI-Bench [27] 和 GroundAct [52]确实会询问任务需要哪些机器人，但不包含不可行任务；据我们所知，还没有基准测试询问紧密耦合任务所需的最少机械臂数，也不接受不存在可行团队的情况。MRBench 两者兼顾（表1）：每项任务有五种配置，在同一条指令下得出不同答案，类似对比集 [17],；其 40 个版本中有 14 个不存在可行团队，类似于 SQuAD 2 中不可回答的问题。0 [43]。

多机器人任务的规划器与智能体。大多数多机器人操作工作会为任务指定的团队制定计划 [10, 11, 13, 16, 20, 22, 31, 32, 47–49, 51, 58], , 通常针对高度耦合的任务；而选择机械臂的工作则从完整模型或明确的需求出发 [3, 18, 21,23, 30, 37, 38, 46, 53]。相反，基础模型智能体会在运行时决定每个机器人的工作：RoCo [36] 让机械臂智能体进行交流，直到就子任务和路径点达成一致；HMAS-2 [6] 由中央模型提出动作，再由每个机器人进行检查；SMART-LLM [29]根据技能组建联盟，并可能让机器人处于未分配状态或拒绝制定计划；其他方法则通过消息、共享程序，或规划器与验证器进行协调 [4, 9, 26, 28, 33, 35, 45, 50, 55–57], , 单个智能体也会学习寻求帮助或拒绝执行 [1, 8, 24, 44]。这些工作会评估成功率、步骤数、词元数或利用率，但不会评估任务所需的机械臂数量，也不会评估作出决策需要多少次模型会话。MRScissors通过拆分这一决策过程解决了这两个问题：对机械臂信息一无所知的模型编写任务所需的任务清单，代码测量每个机械臂能够执行的操作，而精确搜索则返回最少机械臂数，并提供说明每个机械臂为何不可或缺的证书，整个过程只需约四次会话。

3. MRBench

MRBench 询问：当目标保持不变且最小可行团队发生变化时，一项任务真正需要哪些机械臂，或需要哪些替代它们的被动对象。五种配置中的八项任务（图 2）产生 40 个版本和 480 个场景。

3.1. 物理必要性作为最小支持集

任务是由任务 和执行者组成的集合；当位于不同位置的任务必须同时进行（并发），或超出单个机械臂的可达性范围时，就需要多个机械臂。

任务。 每项任务 $j \in J$ 都有位置 S_j，即手必须执行操作的位姿，以及一个失败信号，当任务未完成时由模拟器报告。操作 任务需要一个能够到达所有位置的执行者；搬运任务 则由一个执行者，或在交接点会合的一串执行者，将物体从 s_j^0 搬运至 s_j^1。组 $g \in \mathcal{G} \subseteq 2^J$ 中的任务必须同时进行，例如在箱子被拉出时保持门打开。

执行者。执行者是机械臂 A，每个机械臂都是位于固定位置的一个机器人模型，以及桌子上的被动对象 P，它们可以从机械臂那里接手一项作业。机械臂 a 可以执行一项操作作业 j，写作 $c(a, j) = 1$，当它以足够的裕度到达每个位置，并且其手指张开得足够大时：

$$\rho_a(s) \geq \delta_j \quad \forall s \in S_j \qquad \text{and} \qquad w_j \leq o_a, \tag{1}$$

其中，$\rho_a(s)$ 是 s, δ_j处的逆运动学（IK）裕度；所需裕度为 2–8.5cm，w_j 表示抓取宽度，o_a 表示手指开度。当机械臂链的两端均可达 s_j^0 和 s_j^1，且相邻机械臂之间共享一个手部交接点，并且所有位置都具有裕度时，该机械臂链即可完成搬运任务。物体 p 可以执行 j, $c(p, j) =1$；当某个测量属性 $\phi_j(p)$ 满足 Θ_j （重量大于杠杆的杆）时，必须由机械臂将其放置（任务 j_p）。

最小支持集。 如果对于某个 $U \subseteq P$，存在一个分配 σ，使每个任务 $j \in J_U = J \cup \{j_p$ 都由 $p \in U\}$ 执行者 $\sigma(j) \subseteq U \cup T$，则团队 $T \subseteq A$ 是可行的：

$$\begin{aligned} &c(\sigma(j), j) = 1 \qquad \forall j \in J_U, \\ &\sigma(j) \cap \sigma(j') = \varnothing \quad \forall g \in \mathcal{G},\ j \neq j' \in g. \end{aligned} \tag{2}$$

最小支持集 （MSS）是规模最小的可行团队，

$$\mathrm{MSS}(A, P) = \underset{T \subseteq A \text{ feasible}}{\arg\min} |T|, \tag{3}$$

其中任意一个都可作为正确答案；当不存在这样的团队时，即 A 本身不可行时，该任务不可行。因此，一个答案可以同时确定最小性 （需要哪些成员）、可行性（任务是否能够完成）以及功能替代（什么可以替代机械臂）。

3.2. 任务设计

并发与可达性。 公式（2）的两个条件给出了两套各包含四个任务的套件。 并发 任务需要同时使用三只手，比双手机器人拥有的手多一只：同时抓住两个物体，由第三只手工作（Con-01、Con-02）；抓住一个物体，同时由另外两只手工作（Con-03）；或同时抓住三个物体（Con-04）。在 可达性任务中，没有两只机械臂会同时动作，但没有任何固定机械臂能够到达所有位置：物体通过放下的方式在机械臂之间传递，或者每只机械臂仅在自身可达的区域内工作（图 2）。我们排除了单只机械臂无法承受的负载：特权抓取不对抓握能力或有效载荷进行建模。

五种配置。 它们只改变任务的执行者；除参考配置外，每种配置都能击破一种跳过测量的捷径。(i) 参考配置精确地 需要三条机械臂。(ii)盈余在 N 处增加一条冗余的 Franka机械臂；它击破了按在场情况配备人员。(iii)短缺 移除一条必需的机械臂，因此没有任何团队可行；答案必须指出没有执行者的工作；它击破了假定可行性。(iv) 交换 一条机械臂与另一种具身形态的机械臂，这可能导致每个位置都有人配备但任务仍不可行，或让两条机械臂完成任务；它击破了按位置分配角色。(v) 借用 用一个能够完成其工作的道具和一个无法完成该工作的干扰物，替换一条必需的机械臂；除一个测量属性外，二者完全相似；它击破了按外观替换。

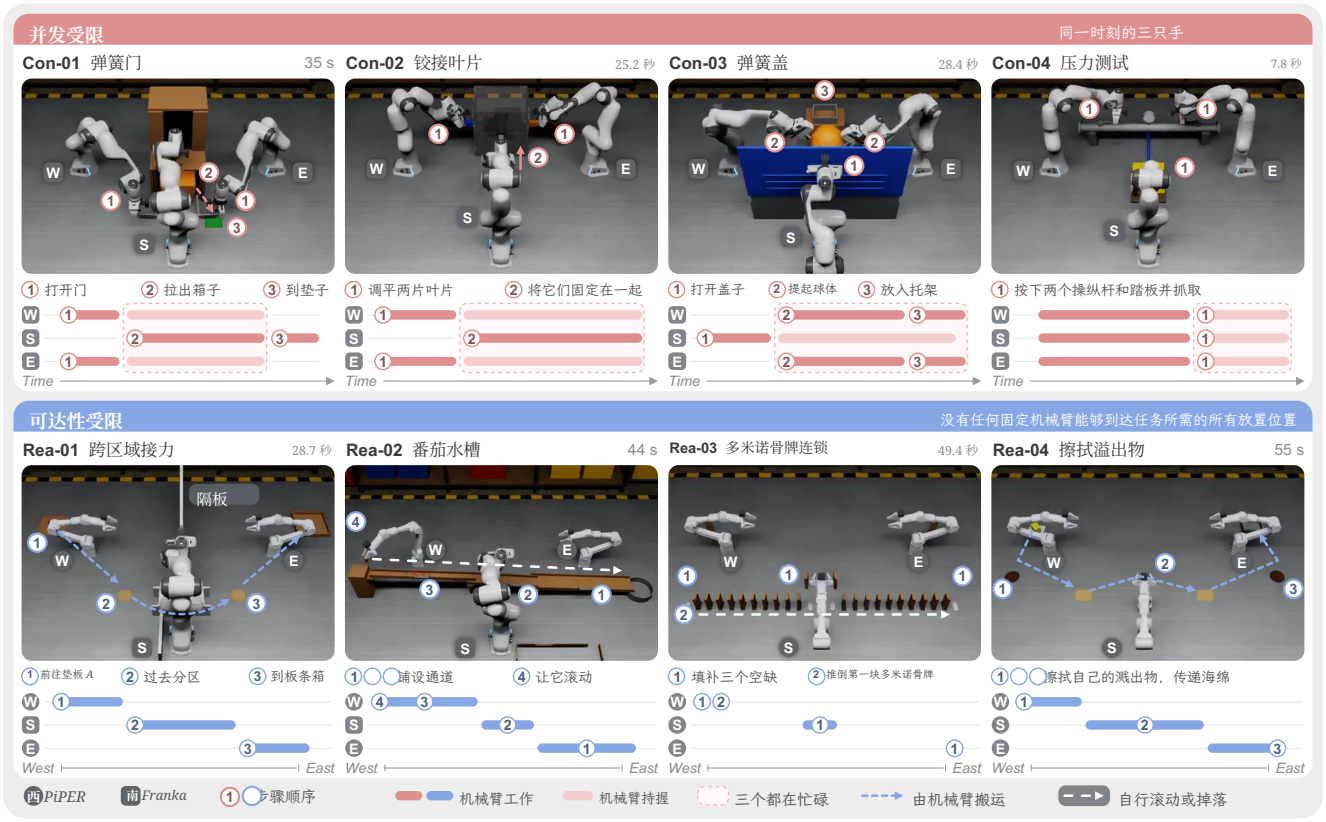


图 2。**MRBench 中八项任务的概览**。每项任务均展示其脚本化参考中的一个关键时刻（精确、简单级别），按顺序列出其步骤，以及每个机械臂何时（并发）或在哪里（可达性）工作。指令未指定任何机械臂。

设计规则。 指令只给出目标：不包含机械臂、角色、步骤或工具。简单和困难级别仅在一个参数上有所不同；成功与否依据状态谓词判定，而不是依据某种方法的声明；物理学参数在另外十个任务和一个原型上进行调校。

3.3. 基准测试构建

工作单元和场景。一张桌子有三个固定位置，西侧（W）和东侧（E）相距 1.4 米，中间为南侧（S），还有第四个北侧位置（N），供盈余机械臂使用；每个位置均可配置 Franka、PiPER 或 SO-101。我们使用 Isaac Sim 4.5 和 Isaac Lab 2 进行仿真。1 [39]；除非由第二只手按压物体远侧，否则对于宽度大于张开手指的物体，抓取会被拒绝。每个版本包含 12 个场景（两个层级、六个随机种子）：随机种子决定房间，为任务的物体重新着色，并添加三个日常物品。

真实值已计算。 场景的真实值（标准答案）来自带裕度的逆运动学（公式（1））以及对最小可行团队的精确搜索（公式（3））。在 Isaac Sim 中，脚本化参考方案在全部 312 个可行场景中均成功，干扰物按设计在全部 96 个 借用场景中均失败；求解器证明了不可行性，并且 Isaac 中每次双机械臂尝试都报告任务未完成。

审查。 每个关于机械臂是必要的声明，在带有裕度的逆运动学下成立，或者在逆运动学无法判定时，在 Isaac 中成立。固定规则从不进行测量，按版本来看表现良好，但按任务来看则不然：最佳规则（计算机械臂数量、选择较大的物体、将任何包含 SO-101 的团队判定为不可行）在 96 个任务场景中答对了 25 个，因此任务级准确率才是重点。

3.4. 协议与指标

该协议取消了多机器人基准测试所提供、而真实工作单元并不具备的四项便利条件。（i）不提供团队信息：方法会获得目标、每个机械臂的模型、放置位置和方向以及物体，但不会获得角色、技能列表或可达性信息。（ii）不保证可行性：40 个版本中有 14 个不可行，此时答案便是“不可行：这些任务不可能全部由执行者完成”。（iii）名称不起决定作用：道具及其干扰物看起来相似，且质量仅以 0.75–1.3 的倍数范围给出，因此必须测量适应度。（iv）失败并非证据：失败运行的 失败报告 会列出其信号被触发的任务，只有此类信号，而非抓取失败，才会对规模更小的团队不利；将任何失败都解读为需要更多机械臂，是第五种捷径，失败即证据。

工具和设置。 每种方法都具有相同的感官、测量工具（测量、重量、可达性、

谁能会合、检查抓取）和探测，并且只能通过一个程序 API 移动机械臂（检查程序、仿真试运行），提交一份表单：可行性、使用和空闲的机械臂、使用和拒绝的物体及其证据、步骤，以及缺失的内容。在一次性 设置下，它无需运行即可提交；在<style id='3'>相同预算 下，每个场景有 2 次探测和 3 次运行，并可在失败后进行修改。

指标。 团队选择准确率根据标准答案对每个场景进行评分：只有当可行性、机械臂数量和物体均匹配时才算正确，并统计所有实际移动过的机械臂，因此方法不能声明使用两条机械臂却运行三条机械臂。任务级准确率 是 96 个任务场景（某一水平和随机种子下的一个任务）中五种配置全部正确的比例。我们还报告可行场景上的成功率（SR）、*MSS-SR*（团队正确且运行成功）、成本和错误类型。

4. 基线方法

已有方法回答的是相邻问题：给定团队如何协作， [6, 36], 或者哪些机器人执行哪些子任务， [4, 29]。我们仅通过一条规则对每种方法进行改造，改变方法接收的输入和产生的输出，但从不改变其推理方式：所有方法都采用一种输出形式、被动候选物和相同的工具，按角色进行划分；它们的区别在于由谁做决策、在哪里学习机械臂能够执行什么，以及失败后修改什么。

参考方案和求解器。 *Oracle-script* 重放标准答案团队的脚本化参考，作为 SR 的上限；*Oracle-team*将标准答案团队交给执行器，因此只会在执行过程中失败。在大语言模型 +求解器中，与大语言模型+P [33] 和 FLEET [45],一样，模型将场景转换为任务、组和执行者（第 3.1节），精确搜索返回最少机械臂数或“不可行”，因此其错误属于翻译错误。

改编自以往的方法。我们针对团队组织方式的每种形式分别采用一种方法：固定流程（SMART-LLM）、机械臂之间的对话（RoCo），以及由中央提案并由机械臂进行表决（HMAS-2）。EMOS [4] 根据机器人履历进行中央分配，Tool-RoCo [56] 在中央激活机械臂智能体，或允许它们自行加入，而 CoEnv [28] 协调给定团队，因此每种方法都属于上述形式之一。*SMART-LLM* 风格 [29] 将其三个阶段——分解、联盟组建和分配——作为三个会话运行，并保留其已释放的指令，以使用最少的机器人；通过查询机械臂来组建联盟，而没有机械臂承担的子任务会使任务不可行。*RoCo* 风格 [36] 为每个机械臂配备独立的模型，该模型可查看固定摄像机和自身的腕部摄像机，仅查询自身的运动学信息，并且可以拒绝参与；机械臂轮流进行交流，最多三轮，

直到所有人都同意一项计划为止，该计划也可能宣布任务不可行。*HMAS-2* 风格 [6]拥有一个中央模型，它能够观察场景，但看不到机械臂运动学；该模型会提出团队、空闲机械臂、物体和可行性方案。每个机械臂都根据自身的可达性判断，接受或否决分配给自己的部分，最多进行三轮。_

公平比较。 在 CaP-X [15]之后，每种方法都会将其答案交给一个共享执行器：同一模型的新会话会接收完整答案（所使用和空闲的机械臂、物体、步骤，以及该方法列出的约束），为声明的机械臂编写程序，使用检查程序对其进行检查，并通过仿真试运行来运行该程序。_ Oracle-team 会提供标准答案团队和物体，但不提供步骤；而将一个可行场景判定为不可行，则计为失败。在相同预算下，每种方法在运行失败后都会收到相同的失败报告，并可以更改团队。在每个骨干网络上，所有方法共享信息、工具、输出格式、预算和调优工作量。

5. MRScissors



图3。**MRScissors 概览。** 盲证明者编写作业图，代码测量能力表，精确搜索返回证书，两名审查员检查该图。

MRScissors 将公式（2）的两个输入分开：模型编写任务所需的内容（任务 J、组 $\mathcal{G}$和可能的替代对象）作为作业图，代码测量谁能够执行这些任务（c）并将结果写入能力表，然后对该表进行精确搜索，返回最少机械臂数（图3）。它仅使用第3.4 节中的工具，并一次性给出答案。

来自盲证明者的作业图。 一个模型会话中的证明器，根据目标和 第 3.1 节 以及 ^ ^场景编写实例：任务 J 及其位置、组 $\mathcal{G}$，以及每个能够从机械臂手中接管任务的物体，并给出其工具读数的界限。证明器是 盲的：它知道手能张开多宽，却不知道桌子上有哪些机械臂、

表2。**MRBench 上的团队选择与任务成功**。按配置、按任务套件（并发：并发；可达性：可达性）报告团队选择准确率（%），涵盖全部480个回合以及任务层级（任务：96个任务场景）。SR 和 MSS-SR（%）基于312个可行回合。每回合成本：机械臂名称、数千个词元和模型会话数。粗体：每个骨干网络中的最佳结果。

方法	按配置					按套件		All	Task	任务成功		每回合成本		
	精确	盈余	短缺	交换	借用	*Con.*	*Rea.*			SR	MSS-SR	机械臂	词元	会话
Qwen3.8-Max														
大语言模型 + 求解器	82.3	87.5	60.4	50.0	51.0	62.1	70.4	66.2	10.4	15.4	15.1	2.40	327.8	1.0
SMART-LLM 风格	94.8	80.2	85.4	53.1	49.0	73.8	71.2	72.5	13.5	19.6	18.3	2.52	640.6	3.0
RoCo 风格	85.4	85.4	52.1	46.9	35.4	57.9	64.2	61.0	2.1	13.1	12.8	2.43	1155.5	6.1
HMAS-2 风格	**95.8**	81.2	86.5	64.6	46.9	75.8	74.2	75.0	17.7	18.6	17.0	2.61	764.9	7.8
MRScissors（我们的方法）	93.8	**96.9**	**93.8**	**83.3**	**59.4**	**87.5**	**83.3**	**85.4**	**43.8**	**46.8**	**46.8**	2.51	421.5	3.8
Claude Sonnet 5.5														
大语言模型 + 求解器	94.8	79.2	92.7	68.8	55.2	77.1	79.2	78.1	18.8	19.2	18.9	2.46	282.2	1.0
SMART-LLM 风格	96.9	82.3	94.8	75.0	59.4	82.5	80.8	81.7	32.3	23.4	21.8	2.52	610.0	3.0
RoCo 风格	88.5	76.0	91.7	71.9	53.1	77.9	74.6	76.2	21.9	16.7	16.7	2.46	783.8	3.6
HMAS-2 风格	95.8	77.1	93.8	71.9	55.2	76.7	80.8	78.8	32.3	23.1	20.8	2.63	765.7	6.1
MRScissors（我们的）	**99.0**	**96.9**	**97.9**	**95.8**	**66.7**	**91.7**	**90.8**	**91.2**	**59.4**	**51.6**	**51.6**	2.55	452.5	3.9

它们位于何处或能够触及什么；它唯一的工具就是感官、测量和重量。因此，它无法编写适合这些机械臂的任务，只能为外观相似者给出相同的界限。

测量能力表。代码首先将图具体化：证明器判定一只手无法抓握的物体过宽，因此需要在同一组中分两步完成；搬运任务则在被抓握的部件到达的位置进行检查。随后，它仅根据工具结果填写方程（1）的表格$\hat{c}$。谁能够——对每个地点的每条机械臂进行判断；会合——对搬运任务中的每对机械臂进行判断；以及测量——对每个物体的每项界限进行判断。可达性单元格标记为能够、不能，或者当机械臂仅差不超过 2 厘米（工具—模拟器差距）时标记为接近；只有物体的所有界限都满足时，该物体才算通过。每个单元格都保留其证据。

$\hat{}$ $\hat{}$**精确搜索与证书**。在 J、$\mathcal{G}$ 和$\hat{c}$上，代码使用大语言模型 + 求解器（第4节）的搜索方法求解方程（3）：它从一条机械臂开始，逐步尝试增加机械臂的子集，并返回所有满足以下条件的最小团队：每项任务都有一名执行者，且同一组中的各项任务由不同的执行者承担；仅当机械臂的能力不足时才使用物体。因此，两种方法的区别仅在于图和表格的来源。答案会给出一份证书：移除每条机械臂后留下哪些未完成任务；每个物体是否使用及其测量值；若不存在可行团队，则列出无法全部找到执行者的任务；以及被计为可达的邻近单元格。

替代物体。只有在搜索找不到任何团队时，MRScissors 才会寻找更多物体：仍处于盲状态的证明器会被告知哪些任务缺少执行者，并被询问桌子上的哪些物体可以接替这些任务。它可以添加物体，并为每个物体指定其边界和放置任务，但不得更改其他任务；新物体经过测量后，搜索会重新运行。

不等负担审查。图是唯一未经测量的部分，因此需要进行审查。代码会翻转每个元素（删除一项任务、将其移出所属组、移动一个位置、删除一个边界），并标记那些翻转后会改变答案的元素。两场彼此不知情的审查会话分别阅读图、证书和标记；一场寻找哪些内容被写得比任务更容易，另一场寻找哪些内容被写得更困难。每次编辑都必须引用其所属会话的一次工具调用；代码执行该调用并重新绘制证书。不产生任何变化的编辑会被删除；使任务需要更多机械臂或导致没有团队的编辑会被保留；其他编辑必须引用目标中允许该编辑的文字，并通过一个 仲裁器 会话。一个回合大约需要四场会话（凸片。2）：证明器、两名审查员，以及在需要时的替代物流程和仲裁器。

6. 实验

6.1. 实验设置

方法与骨干模型。我们将 MRScissors 与大语言模型 + 求解器、SMART-LLM 风格 [29], RoCo 风格 [36]和 HMAS-2 风格 [6]（第 4 节）进行比较，使用 Qwen3.8-Max 和 Claude Sonnet 5.5，二者均采用高推理强度。Oracle- 团队（第 4 节）在构造上必然正确，因此未列入表 2；借助 Oracle-script，它在表3 中界定了任务成功率的上界。

运行。对于每个骨干模型，每种方法都在单次设置下回答 480 个场景中的每一个（第 3.4 节）。在 312 个可行回合中，答案由共享执行器执行一次（第 4 节），最多进行三次受时间限制的 Isaac 运行；失败的运行不会反馈给该方法。

成本与统计。一个回合的成本是其中所指定团队的机械臂数量（在指定了团队的可行回合中取平均值）、其推理词元数量（平均值），以及选择该团队的模型会话数量。

差异按场景配对，并在96个任务场景上使用自助法计算95%的区间；只有当区间不包含0时，某方法才算领先。

6.2. 主要结果

在两种骨干模型上，MRScissors 最常识别出正确团队。使用 Qwen3.8-Max 和 Claude Sonnet 5.5 时，它在回合中判断正确的比例分别为 85.4% 和 91.2%；相比之下，下一种方法（HMAS-2 风格、SMART-LLM 风格）分别为 75.0% 和 81.7%。在两类任务中，它均位列第一，且两类任务的得分差异不超过 4.2 个百分点。更强的骨干模型使每种方法的表现都有所提升，提升幅度从 3.8 个百分点（HMAS-2 风格）到 15.2 个百分点（RoCo 风格）不等，但并未缩小差距：使用 Qwen3.8-Max 的 MRScissors 比使用 Claude Sonnet 5.5 的任何其他方法都更常判断正确。差距在任务层级上最大，这也是第3.4节的核心结论：MRScissors 在 43.8% 和 59.4% 的任务场景中，能够正确判断任务的全部五种配置；其他方法对应的比例分别为 2.1–17.7% 和 18.8–32.3%。因此，使用 Qwen3.8-Max 时，所有方法的表现都低于最佳固定规则（26.0%，第3.3节）；使用 Claude Sonnet 5.5 时，除 SMART-LLM 风格和 HMAS-2 风格外，其他方法也都低于该规则。按场景配对后，在两种骨干模型上，MRScissors 均领先于每种方法：在所有回合上领先 9.5 至 24.4 个百分点，在任务层级上领先 26.1 至 41.7 个百分点，且没有任何 95% 区间包含 0。

更好的翻译并不能说明机械臂能做什么。大语言模型 + 求解器在模型的单次翻译上运行 MRScissors 的搜索；在该翻译中，模型还会写明每个机械臂能做什么（第 5 节）；MRScissors 分别领先 19.2 和 13.1 个百分点。更强的骨干模型将短缺情况下的差距从 33.4 个百分点缩小到 5.2 个百分点，其中关键机械臂只是缺失了，但在交换情况下则无法缩小差距（33.3 和 27.0 个百分点），因为必须知道新机械臂能到达的位置。因此，更好的翻译并不能提供**机械臂能做什么**，而这正是 MRScissors 在代码中测量的内容。

备用机械臂大多处于空闲状态，更强的模型也没有带来帮助。在盈余 回合中，每种方法至少有 76.0% 的回答正确，而 MRScissors 在两种骨干模型上均达到 96.9%。使用更强的骨干模型后，其他方法中有三种的表现下降了 4.1 至 9.4 个百分点，只有 SMART-LLM- 方法有所提升（80.2% 升至 82.3%）。在两种骨干模型上的 151 个错误盈余 回答中，有 30.5% 的回答在不需要备用机械臂时仍为其配备人员，即按在场情况配备人员；同样多的回答将机械臂数量报得过少，还有 21.2% 的回答拒绝了可行任务。

更强的骨干网络让每种方法都学会拒绝。在短缺 情况下，正确答案是不可行。使用 Qwen3.8- Max 时，MRScissors 在 93.8% 的回合中给出了这一答案，而大语言模型+ 求解器和 RoCo 风格仅在 60.4 % 和 52.1% 的回合中给出这一答案；使用 Claude Sonnet 5.5 时，每种方法至少在 91.7% 的回合中给出了这一答案，大语言模型 + 求解器和 RoCo 风格分别提升了 32.3 和 39.6 个百分点，是表 2 中最大的提升。由于“始终不可行”在每个短缺情况下都是正确答案，拒绝可行任务就是它所付出的代价：

表 3。**正确团队和错误团队的任务成功率。**312 个可行回合中的成功率（%）。团队：标注的正确团队；SR | right、SR |wrong：使用正确团队或错误团队时的成功率。

方法	Qwen3.8-Max			Claude Sonnet 5.5		
	Team	SR | 正确	SR | 错误	Team	SR | 正确	SR | 错误
Oracle-script	100	100	–	100	100	–
Oracle-team	100	58.7	–	100	60.3	–
大语言模型 + 求解器	72.4	20.8	1.2	75.3	25.1	1.3
SMART-LLM 风格	72.4	25.2	4.7	79.2	27.5	7.7
RoCo 风格	67.9	18.9	1.0	71.5	23.3	0.0
HMAS-2 风格	73.7	23.0	6.1	76.0	27.4	9.3
MRScissors（我们的方法）	**83.7**	**55.9**	0.0	**87.8**	**58.8**	0.0

MRScissors 在 312 个可行回合中拒绝了 11 个，其他方法分别拒绝了 20–31 个和 15–27 个。

交换的机械臂使各方法之间的差距最大。MRScissors 在 83.3% 和 95.8% 的交换回合中判断正确，而其他方法分别为 46.9–64.6% 和 68.8–75.0%；其领先下一方法 18.7 和 20.8 个百分点，这是其在任何配置中的最大优势。其他方法并未将 SO-101 判断为不可行：合并来看，在那些引入一台机械臂且全部使任务不可行的交换中，它们的正确率（51.0% 和 71.4%）与其他交换中的正确率（56.2% 和 72.4%）大致相当。在 Franka 替换 PiPER 后两台机械臂即可完成任务的情况下（Rea-02、Rea-04），保留原有团队、按位置分配角色是错误的；其他方法的正确率为 45.8–62.5% 和 58.3–75.0%，而 MRScissors 为 87.5% 和 91.7%。

被动对象能否代替机械臂仍是一个悬而未决的问题。借用是除使用 Qwen3.8-Max 的大语言模型 + 求解器外，所有方法在两种骨干模型上的最弱配置；在交换配置上，该求解器的表现还差 1.0 个百分点。MRScissors 在这里同样排名第一，得分分别为 59.4% 和 66.7%，而其他方法最高仅为 51.0% 和 59.4%；即使使用更强的骨干模型，它仍有三分之一的回合判断错误。按大小选择、依据外观进行替换并不能解释这些错误：在道具是较大的物体时（96 个场景中的 67 个），没有任何方法的表现能高出 5.0 个百分点以上。真正困难的是判断该物体是否具备完成任务的能力。在 450 个错误的借用答案中，28.4% 使用了干扰物，31.8% 认为任务不可行；其余 39.8% 使用了道具，但指定的机械臂数量过少，仿佛它所承担的任务超出了缺失机械臂的职责。

MRScissors 使用的词元少于所有改编方法。每回合使用 421.5 k 和 452.5k 个词元，比词元消耗最低的改编方法 SMART-LLM 风格（640.6 k 和 610.0 k）少三分之一和四分之一，且最多仅为 HMAS-2 风格和 RoCo 风格的 0.6 倍。它每回合 3.8 和 3.9 次模型会话，介于 SMART-LLM- 风格（3.0）和 HMAS-2 风格（7.8 和 6.1）之间。只有 LLM + 求解器每回合仅需一次会话，成本更低（327.8 k 和 282.2 k 个词元），但其正确率分别低 19.2 和 13.1 个百分点。

在留出种子上仍保持领先。没有任何方法是在随机种子 1–5 上开发的：在这 400 个场景中，涵盖全部 40 个版本，MRScissors 领先下一名方法 11.4 和 9.4 个百分点。它在每项任务上也都排名第一，领先幅度为 1.7 至 16.7 个百分点，并且在两个层级上均排名第一。

按错误类型分类。在 MRScissors 的 70 个和 42 个错误答案中，分别有 22 个和 14 个答案指定的机械臂太少，19 个和 4 个为不可行任务指定了团队，11 个和 12 个拒绝了可行任务，16 个和 9 个使用了干扰物而非道具，另有 2 个和 3 个指定了正确数量的机械臂，但机械臂并不正确；没有答案指定过多机械臂，而其他方法在 4 至 18 个答案中出现了这种情况。

6.3. 从正确团队到任务成功

正确团队会带来不同的解决方案。执行器会执行方法的完整答案（第 4 节），因此 SR 和 MSS-SR 既评估其解决方案，也评估其团队。MRScissors 的步骤——每个任务由哪个机械臂按什么顺序完成——可从其证书中读取，其中每个任务的执行者都是一个测量单元；对于其他方法，则由模型写出这些步骤。在全部 312 个可行回合中，MRScissors 使用正确团队取得成功（MSS-SR）的比例分别为 46.8% 和 51.6%，超过其他方法 12.8–18.3% 和 16.7–21.8% 的两倍；在它与另一种方法都指定了正确团队的回合中，它的成功率分别为 51.5–55.4% 和 56.9–58.2%，另一种方法则为 17.9–24.0% 和 24.0–27.4%。Oracle-team 获得标准答案团队后，达到 58.7% 和 60.3%（表 3）。使用正确团队时，MRScissors 自身回合的成功率为 55.9% 和 58.8%，距该上限仅差 3 个百分点。

仅任务成功会让机械臂获得过多奖励。在表2的十行中，有七行的SR超过了 MSS-SR | 错误：在这些行中，每行都有 1 至 7 个回合由并非正确团队的团队成功完成（表3），且每个团队都多使用了一台机械臂；MRScissors 不会命名这样的团队。任务成功表明某个团队足以完成任务，而不表明该团队是必需的，因此 MRBench 将任务成功评分为 MSS-SR。

6.4. 消融研究

表4逐一将 MRScissors 的组成部件添加到大语言模型 +求解器中（a–f），移除各部件（g–j），为其提供标准答案输入（k、l），并将其桌子交给两种改编方法（m、n）。

先谨慎，再恢复。前两个部分使 MRScissors 变得谨慎。在代码 (b) 中测量能力表，并在 (c) 中对证明器隐藏机械臂，这两者共同将不可行回合的准确率从 53.6% 提高到 89.3%，但使可行回合的准确率从 71.2% 降至 59.6%；其中 借用的降幅最大，因为仅凭手并不足够。替代通行方案 (d) 以及审查 (e, f) 挽回了这些回合：可行回合的准确率升至 88.5%，而不可行回合的准确率只回落了一部分增益（78.6%）。因此，总体准确率在每一步都上升，从 65.0% 升至 85.0%，尽管没有任何单列呈现这一趋势。移除盲证明者、翻转标记、审查员或仲裁器后，80 个回合中的准确率会降低 2 或 3 个百分点

表4。**MRScissors 各部件的消融研究**。在简单级别、随机种子 0 和 4 下的 80 个 Qwen3.8-Max 回合上的准确率（%）；每回合会话数。测量：实测表；审查：审查员；仲裁：仲裁器。

	测量值。	盲测。	站立。	翻转。	修订。	仲裁。	附近。	交换	借用	可行	不可行	全部会话
(a)								62.5	37.5	71.2	53.6	65.0 1.0
(b)	✓						✓	75.0	25.0	61.5	78.6	67.5 1.0
(c)	✓	✓					✓	81.2	6.2	59.6	**89.3**	70.0 1.0
(d)	✓	✓	✓				✓	81.2	43.8	71.2	**89.3**	77.5 1.6
(e)	✓	✓	✓	✓	2		✓	81.2	56.2	82.7	82.1	82.5 3.8
(f)	✓	✓	✓	✓	2	✓	✓	75.0	**68.8**	**88.5**	78.6	**85.0** 3.9
(g)	✓		✓	✓	2	✓	✓	75.0	62.5	84.6	78.6	82.5 3.8
(h)	✓	✓	✓		2	✓	✓	81.2	50.0	76.9	**89.3**	81.2 3.6
(i)	✓	✓	✓	✓	1	✓	✓	75.0	56.2	78.8	85.7	81.2 2.8
(j)	✓	✓	✓	✓	2	✓		**87.5**	56.2	82.7	**89.3**	**85.0** 3.8
(k)	金标准图 + 实测表							93.8	93.8	98.1	96.4	97.5 –
(l)	自身图 + 标准答案表							87.5	62.5	86.5	89.3	87.5 3.8
(m)	SMART-LLM 风格 +桌子							56.2	43.8	71.2	78.6	73.8 4.8
(n)	HMAS-2 风格 + 桌子							68.8	43.8	73.1	78.6	75.0 9.6

（g–i、e）；将邻近单元格视为不可达则不会降低准确率（j）。因此，各部分在**谨慎与覆盖率之间进行权衡**：后面的部分弥补了前面的部分所付出的代价。

其余大多数错误都出在作业图上。给定带有实测表（k）的金标准作业图时，MRScissors 在97.5%的回合中做出了正确判断；给定其自身图和金标准能力（l）时，这一比例为87.5%。在其12个错误中，金标准图因此消除了10个，而金标准表消除了2个。大多数错误来自**模型所写的内容**，即任务、任务所属的组以及替身的边界，而不是工具测量出的内容。

起决定作用的是流程，而不只是桌子本身。每种方法使用相同的工具，但只有 MRScissors 会在每个地点为每条机械臂调用这些工具。向 MRScissors 提供实测表后，在每回合多出1.75次会话的情况下，SMART-LLM 风格和 HMAS-2 风格的可达性分别达到73.8%和75.0%（m、n），而 MRScissors 为85.0%。因此，提升并不在于测量结果——任何方法都可以采用这些结果——而在于**将测量结果转化为答案的流程**。

7. 结论

在机械臂能够协作之前，智能体必须先决定一项任务需要多少机械臂，而这一决策必须建立在测量之上。MRBench 将其转化为一项测试：针对异构机械臂设置八项任务，每项任务包含五种配置；这些配置会改变在同一目标下的最小可行团队，并配有经审计的标准答案。MRScissors 基于实测证据给出答案：在两种骨干模型上，由求解器根据实测表进行选择的方法，分别在 85.4% 和 91.2% 的回合中选出正确团队，而其他方法最高分别仅达到 75.0% 和 81.7%。然而，MRScissors 在任务场景中正确处理全部五种配置的比例最高仅为 59.4%，并且至少有三分之一的<几率>回合未能完成：什么可以替代机械臂，仍是一个开放问题。

参考方案

[1] Michael Ahn、Anthony Brohan、Noah Brown、Yevgen Cheb- otar、Omar Cortes、Byron David、Chelsea Finn、Chuyuan Fu、Keerthana Gopalakrishnan、Karol Hausman 等。《Do as I can, not as I say：将语言落地于机器人可供性》。arXiv 预印本 arXiv:2204.01691，2022。3[2] Matthew Chang、Gunjan Chhablani、Alexander Clegg、Mikael Dallaire Cote、Ruta Desai、Michal Hlavac、Vladimir Karashchuk、Jacob Krantz、Roozbeh Mottaghi、Priyam Parashar 等。《PARTNR：具身多智能体任务的规划与推理基准测试》。载于国际学习表征会议（ICLR），2025。2[3] Hao Chen、Weiwei Wan、Keisuke Koyama 和 Kensuke Harada。《使用双臂机器人规划搭建索玛方块》。载于 IEEE 学习与发展国际会议（ICDL），2021。3[4] Junting Chen、Checheng Yu、Xunzhe Zhou、Tianqi Xu、Yao Mu、Mengkang Hu、Wenqi Shao、Yikai Wang、Guohao Li 和 Lin Shao。《EMOS：具身感知异构多机器人操作系统与大语言模型智能体》。载于国际学习表征会议（ICLR），2025。2、3、5[5] Tianxing Chen、Zanxin Chen、Baijun Chen、Zijian Cai、Yibin Liu 等。《RoboTwin 2.0：面向稳健双臂机器人操作的可扩展数据生成器与基准测试，具有强领域随机化》。载于国际机器学习会议（ICML），2026。2[6] Yongchao Chen、Jacob Arkin、Yang Zhang、Nicholas Roy 和 Chuchu Fan。《使用大语言模型实现可扩展的多机器人协作：集中式系统还是去中心化系统？》。载于 IEEE 机器人与自动化国际会议（ICRA），2024。2、3、5、6[7]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即可操控机器人》。arXiv 预印本 arXiv:2609.10522，2026。2[8] Liang Cheng、Mingsheng Cai、Jiuming Jiang 和 Luo Mai。《智能体知道自己做不到什么吗？工具使用智能体的可行性感知评估》。arXiv 预印本 arXiv:2605.28532，2026。3[9] Aidan Curtis、Nishanth Kumar、Jing Cao、Tom´as Lozano- P´erez 和 Leslie Pack Kaelbling。《相信 PRoC3S：使用大语言模型和约束满足解决长时域机器人问题》。载于机器人学习会议（CoRL），2024。3[10] Dayi Dong、Maulik Bhatt、Seoyeon Choi 和 Negar Mehr。《MIMIC-D：使用去中心化扩散策略进行多智能体协作的多模态模仿》。载于 IEEE 机器人与自动化国际会议（ICRA），2026。3[11] Ria Doshi、Tian Gao、Annie Chen、Chelsea Finn 和 Jean- nette Bohg。《CHORUS：使用单一视觉语言动作策略实现去中心化多具身协作》。arXiv 预印本 arXiv:2606.12352，2026。3[12] May Espinola、Qifan Zhang 和 Yu Xiang。《从推理到行动：在 LIBERO-10 上对前沿人工智能模型进行基准测试》机器人操作任务。arXiv预印本 arXiv:2610 30051，2026 2 . .[13] Hamed Farivarnejad 和 Spring Berman。集体运输的多机器人控制策略。《控制、机器人与自主系统年度评论》，5(1):205–219，2022。3[14] Senyu Fei、Siyin Wang、Junhao Shi、Zihao Dai、Jikun Cai 等。LIBERO-Plus：视觉-语言-动作模型的渐进式鲁棒性基准。在 IEEE/CVF 计算机视觉与模式识别会议（ CVPR ），2026。2[15] Letian Fu、Justin Yu、Karim El-Refai、Ethan Kou、Haoru Xue 等。CaP-X：用于基准测试和改进机器人操作编程智能体的框架。在机器学习国际会议（ICML），2026。2, 5[16] Jiawei Gao、Ziqin Wang、Zeqi Xiao、Jingbo Wang、Tai Wang、Jinkun Cao、Xiaolin Hu、Si Liu、Jifeng Dai 和 Jiangmiao Pang。CooHOI：结合被操作物体动力学学习协作式人-物交互。在神经信息处理系统进展（NeurIPS），2024。3[17] Matt Gardner、Yoav Artzi、Victoria Basmov、Jonathan Be-rant、Ben Bogin、Sihao Chen、Pradeep Dasigi、Dheeru Dua、Yanai Elazar、Ananth Gottumukkala、Nitish Gupta、Han-naneh Hajishirzi、Gabriel Ilharco、Daniel Khashabi、Kevin Lin、Jiangming Liu、Nelson F. Liu、Phoebe Mulcaire、Qiang Ning、Sameer Singh、Noah A. Smith、Sanjay Subramanian、Reut Tsarfaty、Eric Wallace、Ally Zhang 和 Ben Zhou。通过对比集评估模型的局部决策边界。在计算语言学协会论文集：EMNLP，第1307–1323页，2020。3[18] Brian P. Gerkey 和 Maja J. Matari´c。多机器人系统中任务分配的形式化分析与分类。《国际机器人研究杂志》，23 954，2004 9 939 2, 3 ( ): – 。[19] Markus Grotz、Mohit Shridhar、Yu-Wei Chao、Tamim As-four 和 Dieter Fox。PerAct2：机器人双臂操作任务的基准测试与学习。在机器人学习会议（CoRL）工作间“全身控制与双臂操作：在人形机器人及其他领域的应用”，2024。2[20] Huy Ha、Jingxi Xu 和 Shuran Song。学习去中心化多机械臂运动规划器。在机器人学习会议（CoRL），2020。3[21] Valentin N. Hartmann、Huang Su、Yijiang Huang 和 Stelian Coros。WRAP：无夹具、感知扳手的多机器人装配规划。arXiv预印本 arXiv:2609 29407，2026 3 . .[22] Chengyang He、Gadiel Sznaier Camps、Xu Liu、Mac Schwa-ger 和 Guillaume Sartoretti。潜在心智理论：用于协作操作的去中心化扩散架构。在机器人学习会议（CoRL），2025。3[23] Rachel Holladay、Tom´as Lozano-P´erez 和 Alberto Ro-driguez。多阶段施力操作规划。在 IEEE机器人与自动化国际会议（ICRA），2021。3[24] Wen-Han Hsieh、Elvis Hsieh、Dantong Niu、Trevor Darrell、Roei Herzig 和 David M. Chan。做什么? 教导

视觉-语言-动作模型拒绝不可能的任务。在计算语言学协会论文集：EMNLP，2025。3[25] Tobias J¨ulg、Seongjin Bien、Simon Hilber、Yannik Blei、Pierre Krack、Maximilian Li、Sven Parusel、Rudolf Li-outikov、Florian Walter 和 Wolfram Burgard。DuoBench：仿真与真实世界中双臂操作的可复现基准。arXiv预印本 arXiv:2606.11901，2026。2[26] Subbarao Kambhampati、Karthik Valmeekam、Lin Guan、Mudit Verma、Kaya Stechly、Siddhant Bhambri、Lucas Saldyt 和 Anil Murthy。大语言模型无法规划，但可以在 LLM-modulo框架中辅助规划。arXiv预印本 arXiv:2402.01817，2024。3[27] Li Kang、Xiufeng Song、Heng Zhou、Yiran Qin、Jie Yang、Xiaohong Liu、Philip Torr、Lei Bai 和 Zhenfei Yin。VIKI-R：通过强化学习协调具身多智能体协作。在神经信息处理系统进展（NeurIPS），2025。2, 3[28] Li Kang、Yutao Fan、Rui Li、Heng Zhou、Yiran Qin、Zhe-meng Zhang、Songtao Huang、Xiufeng Song、Zaibin Zhang、Bruno N. Y. Chen、Zhenfei Yin、Dongzhan Zhou、Wangmeng Zuo 和 Lei Bai。CoEnv：通过组合式环境推动具身多智能体协作。arXiv预印本 arXiv:2604.05484，2026。2, 3, 5[29] Shyam Sundar Kannan、Vishnunandan L. N. Venkatesh 和 Byung-Cheol Min。SMART-LLM：使用大语言模型进行智能多智能体机器人任务规划。在 IEEE/RSJ智能机器人与系统国际会议（IROS），2024。2, 3, 5, 6[30] G. Ayorkor Korsah、Anthony Stentz 和 M. Bernardine Dias。多机器人任务分配的综合分类。《国际机器人研究杂志》，32(12):1495–1512，2013。2, 3[31] Matthew Lai、Keegan Go、Zhibin Li、Torsten Kr¨oger、Ste-fan Schaal、Kelsey Allen 和 Jonathan Scholz。RoboBallet：利用图神经网络和强化学习进行多机器人到达规划。《科学·机器人学》，10 (106):eads1204，2025。3[32] Stefan Lionar 和 Gim Hee Lee。TeamHOI：学习适用于任意团队规模的协作式人-物交互统一策略。在 IEEE/CVF 计算机视觉与模式识别会议（CVPR），2026。3[33] Bo Liu、Yuqian Jiang、Xiaohan Zhang、Qiang Liu、Shiqi Zhang、Joydeep Biswas 和 Peter Stone。LLM+P：赋予大语言模型最优规划能力。arXiv预印本 arXiv:2304.11477，2023。3, 5[34] Bo Liu、Yifeng Zhu、Chongkai Gao、Yihao Feng、Qiang Liu、Yuke Zhu 和 Peter Stone。LIBERO：终身机器人学习中的知识迁移基准测试。在神经信息处理系统进展（NeurIPS），2023。2[35] Kehui Liu、Zixin Tang、Dong Wang、Zhigang Wang、Xue-long Li 和 Bin Zhao。COHERENT：利用大语言模型实现异构多机器人系统协作。在 IEEE机器人与自动化国际会议（ICRA），2025。3[36] Zhao Mandi、Shreeya Jain 和 Shuran Song。RoCo：利用大语言模型进行辩证式多机器人协作。在 IEEE机器人与自动化国际会议（ICRA），2024。2、3、5、6[37] David R Martin、Brooks A Butler、Scott Nivison、Magnus Egerstedt、Mohammad Abdullah Al Faruque 和 Pramod Khargonekar。通过迭代聚类实现异构多机器人系统的协作任务分配。IEEE机器人与自动化快报，11(1):33–40，2025。3[38] Jeremy A. Marvel、Roger Bostelman 和 Joe Falco。多机器人装配策略与指标。ACM计算调查，51(1):1–32，2018。3[39] Mayank Mittal、Pascal Roth、James Tigue、Antoine Richard、Octi Zhang 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。2、4[40] Yao Mu、Tianxing Chen、Zanxin Chen、Shijia Peng、Zhiqian Lan 等。RoboTwin：带生成式数字孪生的双臂机器人基准。在 IEEE/CVF计算机视觉与模式识别会议（CVPR）上，2025。2[41] Hyunjin Park、Jebeom Chae、Minwoo Park、Sunghyun Park、Hanjun Yoo 等。CoHuB：多 humanoid 协作仿真基准。arXiv 预印本 arXiv:2609.34782，2026。2[42] Yiran Qin、Li Kang、Xiufeng Song、Zhenfei Yin、Xiaohong Liu、Xihui Liu、Ruimao Zhang 和 Lei Bai。RoboFactory：探索具有组合约束的具身智能体协作。在 IEEE/CVF国际计算机视觉会议（ICCV）上，2025。2[43] Pranav Rajpurkar、Robin Jia 和 Percy Liang。了解自己的未知：SQuAD 中的不可回答问题。在计算语言学协会年会（ACL）上，2018。3[44] Allen Z. Ren、Anushri Dixit、Alexandra Bodrova、Sumeet Singh、Stephen Tu、Noah Brown、Peng Xu、Leila Takayama、Fei Xia、Jake Varley、Zhenjia Xu、Dorsa Sadigh、Andy Zeng 和 Anirudha Majumdar。会寻求帮助的机器人：大语言模型规划器的不确定性对齐。在机器人学习会议（CoRL）上，2023。3[45] Corban Rivera、Grayson Byrd、Meghan Booker、Bethany Kemp、Allison Gaines 等。FLEET：面向异构机器人团队的形式语言引导调度。arXiv 预印本 arXiv:2510.07417，2025。3、5[46] Max Rudolph、Sonia Chernova 和 Harish Ravichandar。非常时期需要非常手段：迈向风险自适应的任务分配。在 2021 IEEE/RSJ智能机器人与系统国际会议（IROS）上，第2592–2597页。IEEE，2021。3[47] Christian Smith、Yiannis Karayiannidis、Lazaros Nalpantidis、Xavi Gratal、Peng Qi、Dimos V. Dimarogonas 和 Danica Kragic。双臂操作——综述。机器人与自主系统，60(10):1340–1353，2012。3[48] Thomas G. Sugar 和 Vijay Kumar。协作移动机械臂的控制。IEEE机器人与自动化学报，18(1):94–103，2002。[49] Albert Tung、Josiah Wong、Ajay Mandlekar、Roberto Mart´ín-Mart´ín、Yuke Zhu、Li Fei-Fei 和 Silvio Savarese。通过协作遥操作学习多臂操作。在 IEEE机器人与自动化国际会议（ICRA）上，2021。3

[50] Shu Wang、Muzhi Han、Ziyuan Jiao、Zeyu Zhang、Ying Nian Wu、Song-Chun Zhu 和 Hangxin Liu。LLM^3：基于大语言模型的任务与运动规划及运动失败推理。在 IEEE/RSJ 智能机器人与系统国际会议（IROS）上，2024。3[51] Ziye Wang、Li Kang、Yiran Qin、Jiahua Ma、Zhanglin Peng、Lei Bai 和 Ruimao Zhang。GauDP：通过扩散策略中的高斯图像协同重新发明多智能体协作。在神经信息处理系统进展（NeurIPS）上，2025。3[52] Zixuan Wang、Dingming Li、Hongxing Li、Yanrui Miao、Shuo Chen、Yuchen Yan、Wenqi Zhang、Yongliang Shen、Weiming Lu、Jun Xiao 和 Yueting Zhuang。GroundAct：大语言模型智能体能否将动作落地到环境状态中？arXiv 预印本 arXiv:2508.05614，2025。首个版本标题为 OmniEAR。3[53] Mengdi Xu、Peide Huang、Wenhao Yu、Shiqi Liu、Xilun Zhang、Yaru Niu、Tingnan Zhang、Fei Xia、Jie Tan 和 Ding Zhao。利用大语言模型进行创造性机器人工具使用。arXiv 预印本 arXiv:2310.13065，2023。3[54] Zhiqin Yang、Chenxin Li、Xiaomeng Hu、Yibin Liu、Weidong Huang、Jiankai Sun、Haitao Li、Zijian Wu、Yuzhi Huang、Fanding Huang 等。RobotWorld：在多样化任务与具身形态中对用于机器人操作的多模态智能体进行基准测试。arXiv 预印本 arXiv:2610.10409，2026。2[55] Hongxin Zhang、Weihua Du、Jiaming Shan、Qinhong Zhou、Yilun Du、Joshua B. Tenenbaum、Tianmin Shu 和 Chuang Gan。利用大语言模型模块化构建协作式具身智能体。在国际学习表征会议（ICLR）上，2024。3[56] Ke Zhang、Xiaoning Zhao、Ce Zheng、Jiahong Ning、Dandan Zhu 等。Tool-RoCo：多机器人协作中的智能体作为工具的自组织大语言模型基准。arXiv 预印本 arXiv:2511.21510，2025。2、5[57] Xiaopan Zhang、Hao Qin、Fuquan Wang、Yue Dong 和 Jiachen Li。LaMMA-P：借助语言模型驱动的PDDL规划器实现可泛化的多智能体长时域任务分配与规划。在 IEEE机器人与自动化国际会议（ICRA）上，2025。3[58] Zaibin Zhang、Junlan Xiao、Zhongbo Zhang、Yifan Wang、Li Kang、Yiran Qin、Changxing Xia、Heng Zhou、Talas Fu、Enshen Zhou、Ruimao Zhang、Zhenfei Yin、Huchuan Lu 和 Lijun Wang。MA-VLA：用于协作与组合泛化的多臂视觉-语言-动作模型。在欧洲计算机视觉会议（ECCV）上，2026。3[59] Tony Z. Zhao、Vikash Kumar、Sergey Levine 和 Chelsea Finn。利用低成本硬件学习细粒度双手操作。在机器人学：科学与系统（RSS）上，2023。2[60] Xueyang Zhou、Yangming Xu、Guiyao Tie、Yongchao Chen、Guowen Zhang 等。LIBERO-PRO：迈向超越记忆的视觉-语言-动作模型稳健且公平的评估。arXiv 预印本 arXiv:2510.03827，2025。2

A. 附录

A.1. 任务卡

每张任务卡都会给出指令、机械臂所在的W/S/E、用于设定两个层级的参数、需要三条机械臂的原因、脚本化参考及其在 Isaac Sim 中简单/困难层级下的持续时间（种子 0；如果两个层级耗时相同，则只给出一幅图）、在物体状态上检查的成功谓词，以及必要性审计所检查的内容。指令按照默认场景给出：在变化场景中，指令会根据随机种子为任务物体赋予的颜色来称呼它们；Rea-01 和 Rea-04 的指令还会说明相应层级的水果数量和擦拭时间。指令会陈述目标，其中三条还会限定过程：Rea-02 的番茄必须滚动而不是被搬运，并且成功条件要求它在离开架子后的 5 秒内到达锅中；Rea-03 的一排在被触发时必须保持完整；Rea-04 的海绵必须在每处洒落物上停留达到擦拭时间。标准答案是在这些约束下所需机械臂数量的最小值。图4 和 5展示了每个参考在变化场景中的情况。

Con-01，弹簧门。“将橙色箱子从柜子中取出，放在绿色垫子上，并保持两扇门关闭。”三台 Franka；箱子宽度为 23 / 26 厘米。一个柜子横跨桌子中部，柜门朝向南方。柜子前方有两扇与箱子等高的弹簧门，在固定天花板下向内打开；每扇门都由门外的低位杠杆操作，杠杆是一根位于侧壁旁、带旋钮的杆。箱子比单扇门的通道更宽，也比天花板上方的槽口更深、更高，因此无法通过转动将其取出：必须保持两扇门打开，同时由第三只手抓住箱子的 T 形把手将其拉出；把手的杆柱在天花板上的狭窄槽口中移动。 参考（35.0 秒）：西方和东方各自转动并保持住自己的杠杆，使对应的门保持在 88°；南方将箱子从天花板下拉出，再从上方重新抓取箱子，在两扇门关闭时将其托住，然后把它放在垫子上。 成功。箱子已离开柜子，并在两扇门关闭的状态下停留在垫子上 1 秒，直至运行结束。 审计。 天花板和 T 形把手使取物器的所有连杆避开门（逆运动学：在 81–88° 时至少保持 1.4 厘米间距）。仅由西方+南方操作时，东门保持关闭并挡住箱子。抓取动作无法抓住箱子的 23–26 厘米宽侧面。

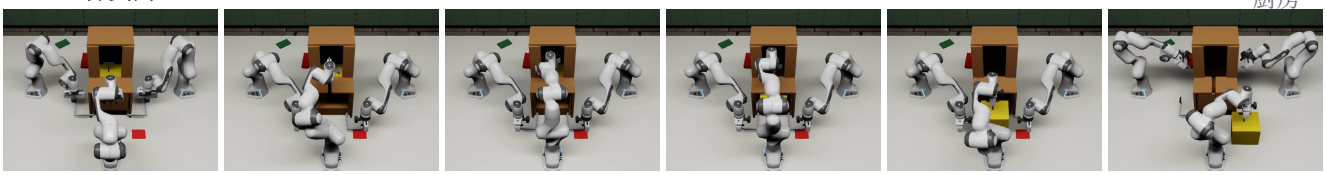
Con-02，铰接叶片。 “用灰色销钉将蓝色叶片和橙色叶片通过所有槽口连接在一起，使它们在立柱之间彼此保持水平。”三台 Franka ； 槽口间隙为 6 / 3 毫米。两片分别铰接在西侧和东侧立柱上的叶片，在空气中会合，形成玻璃护罩内的叉耳；只有在两片都保持水平时，它们的槽口才会对齐，因为一旦松手，西侧叶片就会下垂，而东侧叶片则会因弹簧作用向上弹起。每片叶片都通过其侧壁上的槽口进入护罩，只有其末端的机械臂能够触及其杆；销钉通过南壁上的一个小开口插入。参考 (25.2 秒)：西和东将它们的叶片转至水平并保持住；南将销钉从开口穿过三个槽口；插入销钉后，叶片彼此保持水平。成功。 销钉贯穿三个槽口，两个叶片在 1 秒内保持在 3° 以内的水平状态。 审计。铰链和护罩排除了双机械臂路径，例如将部件固定在桌子上并将其抬起（逆运动学扫描：没有其他机械臂能够触及叶片；一只机械臂抓取两者时，距离最近也不小于 17 厘米）。几何结构排除了将平键销钉放入一个叶片的槽口并将其用作作用于另一个叶片的杠杆的可能性，因此无需运行即可排除该路径。

Con-03、弹簧盖和篮球。“将橙色球从带弹簧盖的箱子中取出，并放入箱子北侧的托架中。”三台 Franka 机械臂；盖子弹簧刚度为 0.3 / 0.6 牛顿·米/弧度。一个带有光滑凸缘内壁的长箱子装着一个直径 22 厘米的球，其宽度超过任何一只手的最大张开宽度（8 厘米）；箱盖沿南侧边缘通过弹簧铰接，没有任何角度能使其保持打开。一只手将盖子保持打开，另外两只手从球的两侧夹住球，将其抬起并搬运至北侧 54 厘米处的托架中。参考 （28.4 / 30.9 秒）： S 通过盖子上的凸片将盖子向上打开至竖直状态；W 和 E 将闭合的指尖压在球的两侧，抬起并将其搬运至托架；S 放下盖子。成功。球在托架中停留 1 秒，并保持到运行结束。审计。一条机械臂将球沿墙向上推：28 次设置中有 26 次没有逆运动学解决方案，另外两次在 Isaac 中失败，因为球沿墙面滑动。单手无法抓取球是由抓握宽度规则决定的：除非另一只手按住物体的远侧，否则该规则拒绝宽度超过张开手指间距的物体。

Con-04，压力测试。“让钢管在测试压力下保持一秒：两个泄压阀关闭，同时脚踏阀让空气进入。”三台 Franka；阀门弹簧刚度为 2.0 / 3.0 牛顿·米/弧度。一根钢制测试管东西向放置在两个支架上，顶部装有压力表，并通过软管从 S 前方的脚踏阀吸入空气。只有在两个泄压阀关闭且脚踏阀让空气进入时，它才能保持测试压力。每个泄压阀的关闭状态由按下钢管顶部的杠杆来实现；松手时，弹簧会再次将杠杆抬起；脚踏阀的踏板同样会在松手时升起并关闭。参考（7.8 / 7.9 秒）： W 和 E 分别将各自杠杆的抓握处按下，S 将踏板的脚尖杆按下，三者在测试期间持续保持。成功。两个杠杆和踏板同时保持按下状态 1 秒。审计。每个杠杆的抓握处只能由桌子两端的机械臂触及（S 距离两者均为 1.04 米，远端机械臂距离为 1.02 米）；当两台末端机械臂操作杠杆时，就没有手可用于踏板；在 Isaac 中，双机械臂尝试在全部 12 个场景中均失败。

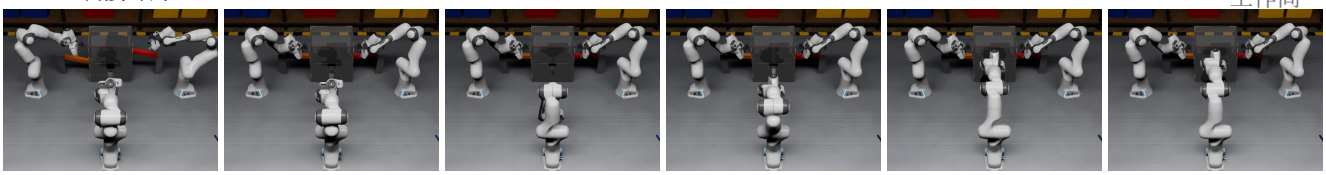
Rea-01，跨区域接力。“将苹果从桌子最西端的起始区域移动到最东端的板条箱中。”西 PiPER，南 Franka，东 PiPER；水果共 3 个部件中的第 1 个。水果起始于西侧后方，必须最终放入东侧后方、相距 2.1 米的板条箱中。中间横贯着一道 50 厘米高的隔板，末端机械臂可在此处伸至高处，使它们无法共享任何位置；隔板南侧有一道延伸至桌子边缘的路缘，唯一的通道是一条低矮通道，其上方覆盖着顶棚，通道宽 30 厘米

Con-01 弹簧门 厨房



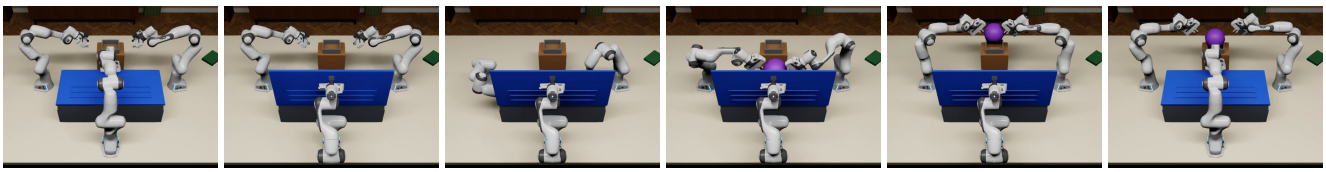
(1) 西、东抓取旋钮 (2) 并将门保持打开 (3) 南抓取箱子 (4) 将其拉出 （5）并出去；门关闭 （6）垫子上的箱子

Con-02铰接叶片 工作间



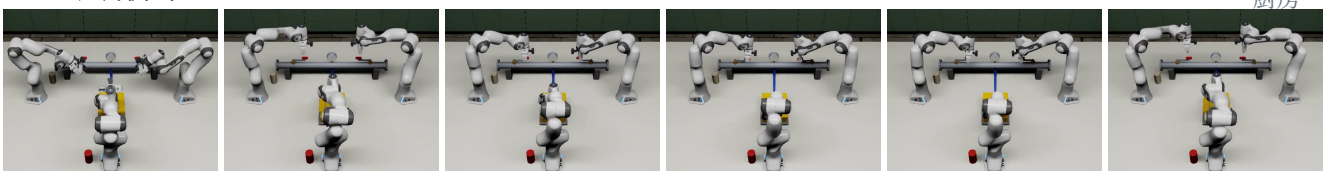
（1）西、东方向可达叶片 （2）并将其转至水平 （3）南拿起销钉 （4）并将其对齐 （5）将其滑入槽口 （6）销钉固定且保持水平

Con-03 弹簧盖 home



（1）南拿住盖子的凸片 （2）并抓住盖子使其保持打开 （3）西、东按压球 （4）并将其抬出 （5）将其搬运至托架 （6）放入；南让盖子落下

Con-04压力测试 厨房



（1）开始 （2）手过来 （3）西、东、南拿走它们 （4）已按下：保持压力 （5）它们松手 （6）弹簧将它们抬起

图4。**Isaac 中的并发参考方案**。每个任务在一个变化的场景中处于简单级别的脚本化参考，如其任务卡所示；房间、任务物体的颜色以及桌子上的日常物体均遵循随机种子。

的中部，任何一端的机械臂都无法到达。参考（28.7 / 48.9 秒）：西 → 垫板 A → 南 → 垫板 B → 东 → 板条箱。成功。每个水果部件都在板条箱中停留 1 秒，并持续到运行结束。审计。末端 PiPER 机械臂的可达距离为 0.75 米，超出其标称的 0.626 米；隔板以及通道中部的顶棚使它们无法会合。

Rea-02，番茄水槽。“将番茄从桌子西端架子顶部滚入东端的锅中，沿着铺设在支架上的轨道滚动。不要搬运它。” W PiPER，S Franka，E PiPER；适配房间 6 / 3mm。三条木制槽道（0.80、0.55 和 0.80 米，每条都比前一条更宽）首尾相接地铺设在两个支架上，每条槽道的下端都位于下一条槽道的上端内。每条槽道的起点都只有一台机械臂能够触及，中间那条较短的槽道位于 S 的后方，只有 W 能够触及架子边缘的停止装置；铺设完成后，槽道会保持在原位，因此番茄滚动时没有机械臂抓取任何物体。参考 （44.0 秒）：E 铺设最低通道，S 铺设中间通道，W 铺设顶部通道；W 抬起停止装置。成功。番茄在离开架子后的 5 秒内到达锅中，从未躺在桌子上，并在那里停留 1 秒。审计。较短的中间通道位于 S 的后方，超出末端机械臂的可达范围；托架排除了在没有中间机械臂参与的情况下沿途耙动番茄的可能性。

Rea-03，多米诺骨牌连锁。 “从多米诺骨牌连锁的西端开始，使其倒下并一路倒至东端。”三台 PiPER 机械臂；间距为 75 / 79 mm。一排 25 个多米诺骨牌沿桌子向下排列 1.8 / 1.9 m，位于 S 前方 20 cm 处，中间有三个间隙：第一个位置、最后一个位置，以及中间的第 12–14 个位置；每个间隙均仅在一台机械臂的可达范围内，且只有 W 能够到达第一个位置。放置在间隙中的多米诺骨牌会保持竖立，因此在另一台机械臂工作时，没有机械臂持握任何物体。参考（49.4 / 50.2 s）： 末端机械臂填补两端的位置，S 填补中间的三个位置；W 将第一个多米诺骨牌向东推倒。成功。 当第一个多米诺骨牌倒下，且其上的每个多米诺骨牌都已倒下并倾斜超过 15°、持续 1s 时，这一排才算完整



图 5. **Isaac 中的 可达性 参考方案**。 与图 4相同；每个机械臂只在桌子上属于自己的区域内工作，工作任务在机械臂之间传递。每个画面都展示了该机械臂的工作区域。

。审计。 所有裕度均至少保持 3cm；由于这一排必须完整，将其两半分别推倒不算成功。

Rea-04，擦拭溢出物。 “用海绵擦拭台面上的三个咖啡洒 *spill*，每个洒落物上都按住海绵 *1* 秒，然后将海绵放入东端的水槽。” 三个 PiPER 机械臂；擦拭时间为 1 / 2 秒。三个溢出物沿台面分布，每个机械臂负责的区域各有一个；一个海绵放在最西端的盘子上；盘子仅属于西，水槽仅属于东。海绵通过两个相邻机械臂都能到达的垫块从一个机械臂传到另一个机械臂，并放下后保持不动。中间的溢出物位于一个高 6 厘米的升降台上，在南方前方 19 厘米处，距离末端机械臂 0.87 米。参考 （55.0 / 58.0 秒）： 西擦拭其溢出物，并将海绵放在垫板 A 上；南擦拭中间的溢出物，并将其放在垫板 B 上；东擦拭其溢出物，并将海绵放入水槽。成功。 海绵在每个溢出物上都保持放下状态，其中心与溢出物中心的距离在 2.5 厘米以内，对于 the 擦拭时间，随后在水槽中停留 1 秒，直至运行结束。审计。 中间的洒落物位于南面前方，超出末端机械臂的可达性（裕度保持在 8.5 厘米，手握住海绵末端）。

A.2. 版本、证据与道具

各版本的答案。表格5给出了每个版本的标准答案和凸片。7 其证据，以及图。6 展示了每个任务的每个版本在不同场景开始时的情况。在盈余中，一台 Franka 位于北。其答案是精确包含北，在 Con-01、Con-03、Rea-03 和 Rea-04 中处于空闲状态；在 Con-02、Con-04 和 Rea-02 中则是若干可选机械臂组合之一（在 Rea-02 中，北沿最低通道向南移动，而东保持空闲）；以及西+北在 Rea-01 中，其中西将每个部件递给北高隔板以西的机械臂，而南和东保持空闲。在交换中，替换后的机械臂可能是：原本 Franka 所在位置换成 PiPER 或 SO-101，原本 PiPER 所在位置换成 SO-101，或原本

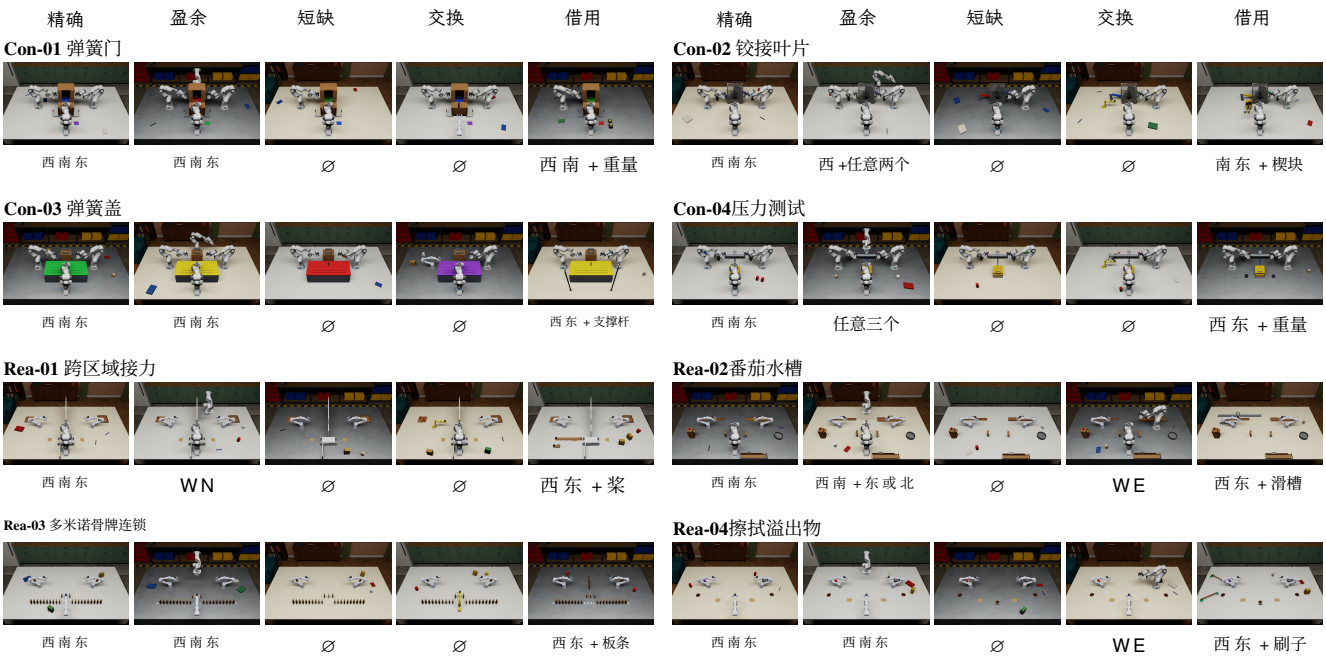


图6。 **每个任务的五种配置**：每个场景的开始状态，其下方为标准答案、最小支持集中的机械臂（如有需要还包括道具），或 ∅，不可行（凸片5）。随机种子和水平会发生变化，房间和物体也随之变化；目标保持不变。

表 5。 **任务及标准答案**。 位于 W/S/E 的机械臂 (F：Franka，Pi：PiPER) 以及各版本的标准答案：最小支持集中的机械臂（杆分隔备选项），或 ∅，不可行，此时任务没有执行者。 盈余 会在 N处增加一台 Franka； 短缺 和 借用会移除同一台机械臂， E 对应 Con-01， W 对应 Con-02， S 对应其他六个任务； 交换 表示被替换的机械臂； 借用 表示道具，并在括号中列出干扰物。两种水平的旋钮见任务卡（第 A.1节）。

		各版本的标准答案			
Task	W/S/E	精确　盈余	短缺	交换	借用
并发 同一时刻的三只手					
Con-01 弹簧门	F/F/F	西 南 东 西 东南	∅东门或箱子	∅ (南→Pi)	西 南 + 9厘米重量（4 厘米）
Con-02 铰接叶片	F/F/F	西 东南 西 东南 | 西 北 东 | 西 南北	∅西侧叶片	∅ （西→SO-101）	南 东 + 车轮挡块（低型或宽型）
Con-03 弹簧盖	前/前/前	西 南 东 西 西南	∅盖子或球	∅ (西→Pi)	西 东 + 0.72 米支撑杆（±12厘米）
Con-04 压力测试	F/F/F	西 东南 西 东南 | 牛顿 南 东 | 西 北 东北 | 西 南 北	∅三个把手，两个机械臂	∅ (西→SO-101)	西 东 + 4.8 厘米重量（2.4 厘米）
可达性 没有一台固定机械臂能够到达任务所经过的每个放置位置					
Rea-01 跨区域接力	Pi/F/Pi	西 东南 西 北	∅搬运任务经过隔板	∅ (西→SO-101)	W E + 67厘米桨（16 厘米/高）
Rea-02 番茄滑道	Pi/F/Pi	西 东南 西 东南 | 西 南 北	∅中间通道	W E（E→F）	西 东 + 1.30 米滑槽（0.45 米）
Rea-03 多米诺骨牌连锁	Pi/Pi/Pi	西 东南 西 东南	∅放置 12–14	∅ (南→SO-101)	西 东 + 30厘米板条（24 / 36 厘米）
Rea-04 擦拭溢出物	Pi/Pi/Pi	西 东南　西 南 东	∅中间洒落物	西 东（东→F）	西 东 + 海绵刷（夹子窄 / 宽）
该配置击败的快捷方式–		按在场情况配备人员	假定可行性	按位置分配角色	按外观替换

原本 PiPER 所在位置换成 Franka；六次交换会使任务不可行，而在 Rea-02 和 Rea-04 中，东侧的 Franka 可以完成中间机械臂的工作，因此 W+E 即可满足要求。在 Rea-04 中，东侧的 Franka 夹住海绵的末端，这是唯一能够够到外伸 0.87 米升降台的抓握方式。

标准答案的依据。 每个标准答案都经过检验（表 7）：标准团队在全部 312 个可行场景中均成功，而干扰物在全部 96 个借用场景中均失败。在 Isaac 中成功的参考对象表明其团队足够胜任；求解器在带有裕度的逆运动学条件下，从最小团队开始逐一尝试，表明没有更小的团队能够胜任并且在不可行版本中，所有任务无法各自拥有执行者。在主要运行中，针对可行场景里答案所指机械臂数量少于标准答案的306个答案均进行了执行，但没有一个成功。在每次双机械臂尝试中，短缺团队的两只机械臂重放脚本化参考动作，并移除缺失机械臂的动作；其中一只机械臂先接管该机械臂的抓取。在全部96个任务场景中，每次尝试均失败，并显示缺失任务的信号。每张任务卡的审计（第A.1节）都会说明为什么两只机械臂不足。在 Con-04短缺中，每个抓取都有一只机械臂负责，但三个动作必须同时发生，而剩下的只有两只机械臂。

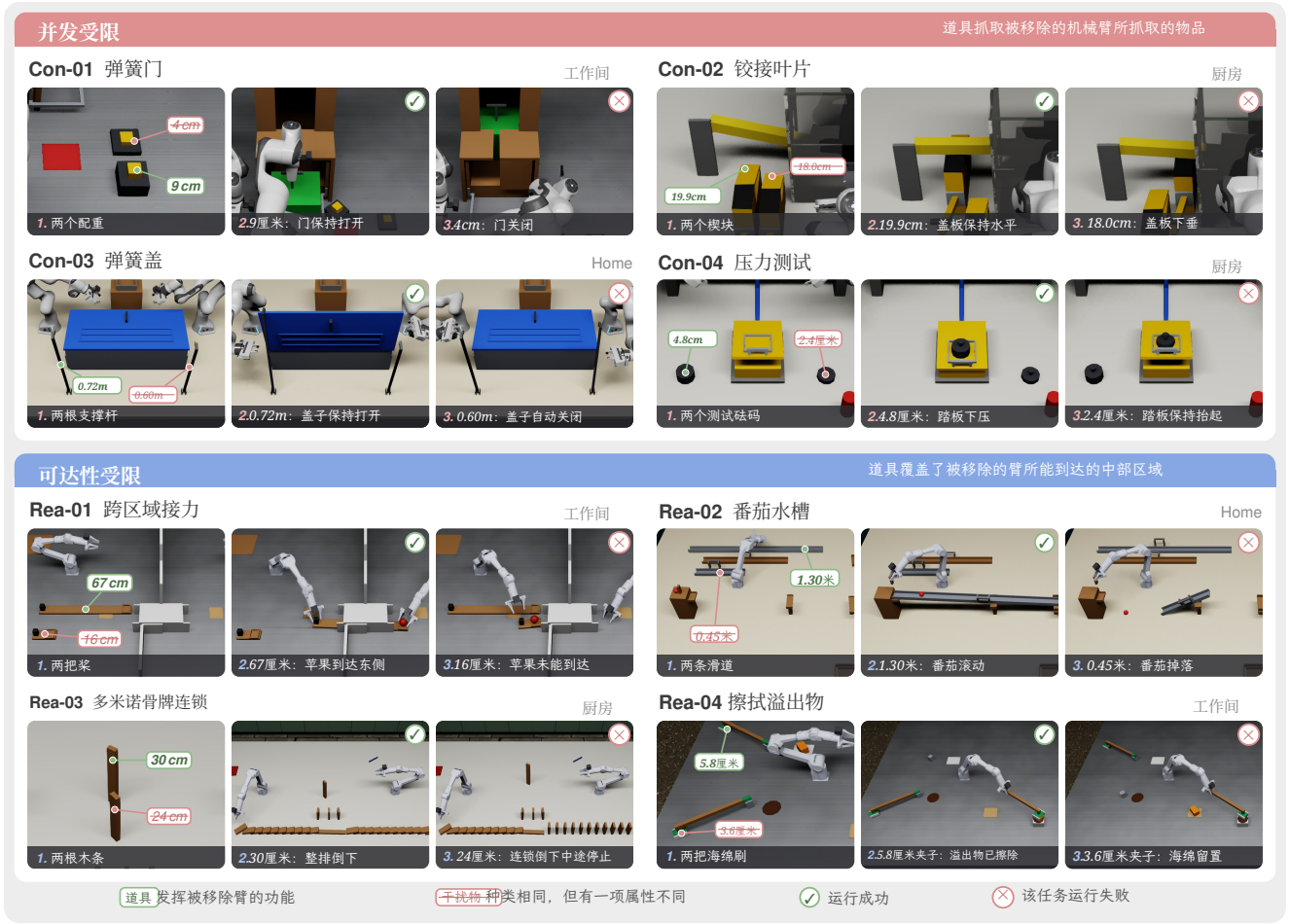


图 7。 **Isaac 中的道具和干扰物**。 每个 借用 简单级别的参考均位于一个多样化场景中：(1) 两个物体在使用前的状态，并标注它们存在差异的属性（在 Con-02 中为楔块的顶部高度）；(2) 使用道具的运行；(3) 在道具位置放置干扰物的运行，并在其任务失败时停止。每个场景包含其任务的一个干扰物（表 6）。

道具与干扰物。表6 列出了每一对物体。道具及其干扰物属于同一种物体，并在一个可测量几何属性上有所不同，绝不会仅在质量或隐藏状态上不同；Con-04 的干扰物高度为一半，质量也为一半。当任务有两个干扰物时，每个干扰物都在一个属性上与道具不同。在存在较大的干扰物但其会失败的情况下，随机种子决定道具或干扰物哪个更大；在其余场景中，道具更大，固定规则利用了表9 。图7 分别借用参考动作，并使用其道具和干扰物：道具执行被移除臂的功能，而干扰物则因该功能对应的信号而失败。

随机种子的影响。在同一版本和层级内，随机种子决定房间、颜色和三个日常物体，将任务中的物体移动不超过 1.6 厘米，并在借用中抽取道具和干扰物所处的位置以及哪一个更大。因此，这 480 个场景包含 80 个不同的物理问题（40 个版本 × 2 层级），每个问题对应六个随机种子

这些随机种子会改变外观，并将物体移动不超过 1.6 厘米：随机种子用于测试对外观和放置变化的鲁棒性。

A.3. 工具、程序API与隔离

表 8 列出了共享工具。三项边界确保其公平性。 重量 是一种测量，就像 测量：它不计入探测次数，并且必须指明一条能够触及物体的机械臂。每个会话获得其方法设计所赋予的工具（第 A.4节）。粗略质量是每一行都会获得的一项观察值：真实质量乘以 0.75–1.3 之间的一个因子，并且同一任务的每个版本所用因子相同。在按机械臂作出决定的行中，每条机械臂都能看到固定摄像机，但只能看到自己的腕部摄像机，读取机械臂 并且 能够触及。

工具知道什么。 没有任何工具会读取任务、角色、成功谓词或脚本化参考；在卡片中，任务简报给出了目标以及未显示的信息，例如弹簧，而 Con-01 的垫子这类绘制地点是

表6。**道具和干扰物**。每个借用版本都会移除一条机械臂；其道具执行该机械臂的功能，而其干扰物是同类物体，仅在一个可测量属性上存在差异（Con-04 的干扰物高度减半，质量也减半）。质量仅给出大致数值，因此必须根据物体所承担的功能对其进行测量。

Task	被移除臂的功能	Prop	干扰物	差异在于
并发道具抓取被移除臂所抓取的物体				
Con-01 弹簧门	东：抓取东门 open	东侧杠杆路径上有一个9厘米配重后退：杆位于上方5厘米处，无法通过它	高4厘米：杠杆从它和门的上方通过门关闭	高度
Con-02 铰接叶片	W：抓取蓝色叶片水平	一个推入悬挂杆下方的车轮挡块；其顶部低于水平杆0.8毫米	低1.9厘米（杆低4°或更多），或宽32厘米（间隙为25.5厘米）	高度、宽度
Con-03 弹簧盖	南：抓取盖子使其保持打开	一根安装在枢转支脚上的杆，其钩子伸出 0.72 米，摆过盖子边缘	长 12 厘米（钩子越过自由边缘）或短 12 厘米（它始终无法到达）	长度
Con-04 压力测试	南：抓取脚踏阀 down	一个高 4.8 厘米、重 2.4 千克的测试砝码，放在踏板凹槽中：它将踏板压下	高度和重量均减半，即 2.4 厘米和 1.2 千克：踏板不移动	高度、质量
可达性 道具覆盖了被移除的机械臂所能到达的中部区域				
Rea-01 跨区域接力	S：跨区域搬运水果中间	一块由西侧推动的67厘米桨板，穿过通道，直到水果处于东侧的可达范围内	16厘米（水果处于双方的可达范围之外），或一根12厘米的边缘（通道宽11厘米）	长度、高度
Rea-02 番茄水槽	南：铺设中间通道	一条 1.30 米长的滑槽，由西侧从架子旁铺设，经过第一个支架，进入最低通道	0.45 米：它远远短于最低通道	长度
Rea-03 多米诺骨牌连锁	南：填充放置 12–14	一根 30 厘米的板条竖立在放置 11 处；它倒下时横跨将位置 12–14 放置到位置 15 的多米诺骨牌上	24厘米（它会掉入间隙）或 36厘米（它会将位置 15 的多米诺骨牌压下）	长度
Rea-04 擦拭溢出物	南：擦拭中间位置洒落物	夹子抓取海绵的海绵刷；西将其抬到升降台上并擦拭	夹子太窄（被抓取规则拒绝），或太宽（它只能推动海绵）	夹子宽度

表7。**40个版本的状态**。I：标准答案在该版本的全部12个场景中均能在 Isaac 中成功（两个层级、随机种子0–5；每组中有多个正确答案时亦然），并且在借用干扰物在全部12个场景中均失败；S：求解器在作业模型（带裕度的逆运动学）中显示不可行性。

Task	精确	盈余	短缺	交换	借用
Con-01	I	I	S	S	I
Con-02	I	I（每组）	S	S	I
Con-03	I	I	S	S	I
Con-04	I	I（每组）	S	S	I
Rea-01	I	I	S	S	I
Rea-02	I	I（每组）	S	I	I
Rea-03	I	I	S	S	I
Rea-04	I	I	S	I	I

locate 返回的场景物体。方法询问的每个点都是独立的，取自 locate 返回的包围盒。可达性以及谁能够执行操作，是根据机械臂的运动学确定的，并与桌子和固定物体进行检查，但不会沿路径检查，也不会考虑其他机械臂或有效载荷；meet 会在两个机械臂基座之间搜索这样的地点：两个机械臂都能从各自一侧接触某个物体，且不与固定物体相交；check hold 通过静力学和粗略质量计算，判断方法列出的把手是否能使物体在其重量及任何弹簧作用下保持静止。没有任何工具会说明物体的用途或其服务的任务。失败报告会列出信号触发的任务，但它仅来自仿真试运行；而在单次设置中，团队会在任何运行开始前提交。

隔离。 每个会话开始时都无法访问基准测试的代码：要么通过关闭其所有自有工具的命令行智能体，要么对于主表格，通过模型 API 配合我们自己的循环。在这两种情况下，唯一可用的工具都是环境提供的工具，上下文中仅包含方法的提示、工具及其回复。

一个回合，逐步进行。 在 Rea-01 中，智能体只会收到目标和任务简报。它观察桌子，找到苹果、板条箱、隔板和垫块；它写下任务（开始时抓取苹果，将其搬运至隔板另一侧，再把它放入板条箱）；它询问谁能到达各个位置，发现只有 W 能到达开始位置，只有 E 能到达板条箱，而且隔板使得没有任何位置能同时被两侧的末端机械臂到达；随后提交 W+S+E，并为每条机械臂说明理由。共享执行器随后为声明的手臂分阶段编写程序，检查程序会检查其路径，仿真试运行则运行该程序；运行失败后，执行器可以编写新程序，最多运行三次，但任何内容都不会返回给方法。计分团队由声明的手臂以及任何发生过移动的手臂组成。

A.4. 基线方法详述

主论文最接近工作表格的范围。 该表报告的是形式化方案和评估的得分，而不是某种方法能够完成的事情。异构机械臂指不同的机械臂模型。选择团队包括动态激活，不一定会给出最少的机器人；被排除的机器人会在整个任务中都不参与，而不是暂时处于 WAIT 状态。不可行答案是对整个任务的判定，而不是一次失败的运行。被动替代体接管机械臂的支撑作用。匹配答案变更在保留指令的同时改变标准答案团队或可行性答案；我们的变更由数值裕度（相对于物理界限的余量）定义，而不是由物体尺寸、错误消息或布尔逆运动学结果定义，并且所列工作中没有任何一项报告过此类变更。

表 8。**共享工具**。每种方法都通过一个环境服务器访问它们；机械臂只能通过 programAPI 移动。每个回合中，仿真试运行最多三次，主要运行则按照相同预算执行，其中探测运行最多两次。

Tool	返回内容
感知与测量	
任务简报	目标；机械臂模型和位姿；粗略质量；固定物体
look	来自固定相机或腕部摄像机的一张图像
读取机械臂	关节角及其限位、工具位姿、手指开度、力矩
定位	物体或部件的箱子：模拟器的，噪声为 4 毫米和 1.5°
测量	物体的部件和尺寸：长度、高度、开口
heft	由处于可达范围内的机械臂将质量控制在百分之几以内（已计算）
能够到达	在某个位姿下的自身逆运动学：裕度或距离不足（1–2 厘米）
能够做到的对象	每个机械臂在某一点的裕度
meet	两条机械臂可以进行交接的地方，也可以位于障碍物上方
场景图	场景中的物体及其关系
检查抓取	手、停止或重量是否能够使物体保持静止
note	一份记录已确认事实及其证据的笔记本
探测	一段最长 8 秒、在 Isaac 中播放并回读的运动
控制	
程序 API	工具位姿、上方、移动、后退、其中、移动，抓握、推、释放、张开手指、阶段；一个程序设置一个计划，并且只导入数学
检查程序	沙盒运行中的冲突：所有路径均使用逆运动学；间隙为 3 厘米机械臂之间为 1 厘米，与场景之间为 1 厘米，抓取物体的间隙为 1.5 厘米
仿真试运行	物理运行（关节目标、其中的特权附着宽度检查后 2 厘米处、布料夹取）：四个关键帧、事件日志、成功、失败报告
输出	
提交团队计划 -	是否可行；使用的机械臂、空闲机械臂及其原因；使用的物体以及它们的任务，附证据拒绝；缺少什么
提交程序	最终程序
回复	其所在行中的会话的结构化回复 '的程序：子任务、联盟、对话轮次、检查
大语言模型 +求解器专用	
求解团队	在作业图上进行精确搜索：每个最小团队以及谁负责什么，或是什么阻碍了它

物理执行使用动力学或硬件。对于 Tool-RoCo，机械臂模型和执行单元格也参考了其公开代码。其机器人在智能体激活之前保持非活动状态，因此某个机器人可以在整个运行过程中都不参与；但报告的四种设置中，没有任何机器人被激活的情况全部失败，并且未被选中的机器人不计分。SMART-LLM 公开的分配提示要求选择所需的最少机器人，但该论文并未对最小团队进行评分。EMOS 以几何方式执行计划，关闭物理学和碰撞检测，并将其示例中未分配的机器人称为低效，而不是展示更小的团队能够成功。
VIKI-Bench 的异构机械臂单元，即 VIKI-R 的基准测试，基于一项在机械臂模型之间进行选择的硬件案例研究。

为何进行适配。 SMART-LLM 在技能缺失的文字任务中可能会遗漏机器人，并在文字描述中拒绝任务，但它不会对不可行性进行评分。其发布的提示词已经要求使用最少的机器人。
RoCo 的 WAIT 仅会让一个机械臂暂停一个步骤，而其夹爪损坏示例涉及任务的其余部分。EMOS 以定量方式测试可达性，也可能不为机器人分配任务，但不会报告可行性裕度。HMAS-2 会检查每个已分配的动作，但没有针对整个团队的目标函数。VIKI-R 会对选定的机器人集合和计划进行评分，而不是对物理任务成功进行评分；其箱子支撑示例也没有表明可以省略某个机械臂。RoCo 中有两台 UR5e 机械臂共同执行一项任务，HMAS-2 运行三台或六台同类型机械臂，而 VIKI-R 的一个示例启用了两台 Panda机械臂；EMOS 则让四种不同类型的机器人组成团队。RoCo、EMOS 和 VIKI-R 仅依据执行前进行的检查来修改计划：验证反馈、每个机器人根据其简历进行的检查，以及计划检查器。HMAS-2 还会在已执行的动作失败时重新计划，例如箱子从夹爪中滑落的情况。我们的适配为它们提供了 MRBench 的通用输出和被动候选物。

失败运行后的修订。 在相同预算下，每一行都会收到相同的失败报告，并且可以更换其团队。SMART-LLM 风格将上一次计划和报告放入其第一阶段（分解阶段）的提示中，并重新运行全部三个阶段。RoCo 风格将报告作为环境消息发送给所有机械臂，并开启新一轮对话；与第一轮一样，所有机械臂达成一致或经过三轮对话后，提交最新提案。HMAS-2 风格将报告发送给中央模型和每个机械臂，并开启新一轮提案与否决；经过三轮后仍被否决的计划将按原样提交。大语言模型 + 求解器的搜索涉及机械臂、被动对象以及机械臂之间的转移。运行失败后，新会话会接收上一次计划和报告，重新编写作业图，然后求解器再次进行求解。在每一行中，仅由执行失败导致的问题都可以只通过更改程序来修复。

固定规则。 凸片规则。 9 根据所有 480 个场景的标准答案进行评分，并且仅使用当前存在的机械臂、其模型和物体的尺寸：“机械臂数量、较大的物体”答案 W+S+E 当存在三条或更多机械臂时；当存在两条机械臂和一对外观相似的物体时，选择抓取两个相似物体中较大的物体的机械臂；否则判定为不可行；每个缩进行都会增加一条条款：当存在 SO-101 时判定为不可行，随后当两台 Franka 与一台 PiPER 并列时也判定为不可行。

比较规则。 表格中的每一行都使用相同的骨干网络（Qwen3.8-Max 或 Claude Sonnet 5.5，并采用高推理强度）、相同的信息、根据设计在该行各会话之间分配的相同工具、相同的输出形式、相同的探测和运行次数，以及相同的调优工作量。每一行中的抓取均通过附着完成，并且都在执行相同的宽度检查之后进行。

A.5. 评分

每种配置。 对于精确 答案，必须是标准答案团队；对于盈余，四个机械臂中任意一个包含最少机械臂数量的集合；对于

表 9. **固定规则**。 与计算出的标准答案完全一致的场景数，每种配置为 96 个，全部为 480 个；任务：与标准答案一致的任务场景数，为 96 个。缩进规则会为其上方的规则添加一个条款。

Rule	精确	盈余	短缺	交换	借用	全部	任务
始终不可行	0	0	96	72	0	168	0
始终 西+南+东	96	84	0	0	0	180	0
手臂数量，较大的物体	96	84	96	0	67	343	0
+SO-101：不可行	96	84	96	48	67	391	25

短缺，不可行，并报告缺失任务但不计分；对于 交换，采用求解器的答案（Rea- 02 和 Rea-04 中为两个机械臂，另外六种配置中为不可行）；对于 借用，答案为两个机械臂和该道具，且不使用干扰物。仅在借用情况下对物体计分：答案必须指明该道具，且不得指明干扰物；它指明的任何其他物体均不计分。标准答案列出每个最小团队（表 5），其中任意一个都算正确；某个团队中的机械臂不必出现在所有团队中，正如 Con-04 的盈余情况所示，其四个最小团队不含任何共同机械臂。

团队使用情况。 团队得分所依据的是声明的手臂数量，加上在一次成功运行中实际移动的每一条机械臂；由于检查程序会拒绝移动答案要求保持空闲的机械臂的程序，因此在主要运行中，这就是声明的团队。在单次设置中，答案在任何运行之前给出，因此得分团队就是声明的团队。使用少于标准答案数量的团队若成功完成任务，则判定为正确，并会重新检查任务的标准答案。

任务层级与固定规则。 当一个任务场景的全部五个配置均正确时，该任务场景才算正确。表 9给出了固定规则。每个配置最多可得 100%；正文中的规则在 480 个场景中的 391 个场景（81.5%）中正确；在任务层级上，除非模型读懂机械臂模型，否则不得分：SO-101 仅在其能完成任务时才会被引入，因此该规则在 25 个任务场景中正确。MRScissors 在两种骨干模型上、跨所有场景（85.4% 和 91.2%）以及任务层级（43.8% 和 59.4%）均优于每条规则。对于答案是当前机械臂集合真子集的三个版本，没有任何规则能够正确处理，即 Rea-01盈余 以及 Rea-02 和 Rea-04 交换：在这 36 个场景中，MRScissors 使用任一骨干模型时均有 33 个场景正确；其他方法使用 Qwen3.8-Max 时正确 20 至 26 个，使用 Claude Sonnet 5.5 时正确 22 至 29 个。

错误与统计。 每个错误答案仅归入以下一类：多余机械臂、缺失机械臂、使用最少机械臂数的另一团队、可行性判断颠倒（将团队分配给不可行任务，或拒绝可行任务）、使用干扰物，或遗漏道具（团队需要该道具却将其遗漏）；在没有任务失败信号的情况下将失败运行作为证据，即协议的最后一类，在单次设置中不会出现。在主要运行中，在单次设置下，每种方法针对每个骨干网络的 480 个场景各回答一次；我们报告

表 10。**MRScissors 减去各方法**。以点值表示，按场景配对，涵盖所有回合及任务层级；采用95%自助法置信区间，对96个任务场景进行重采样，每个场景包含五种配置（10000次重采样）；点值为两个准确率之差，准确率按正文中的方式列出。

MRScissors 减法	Qwen3.8-Max		Claude Sonnet 5.5	
	All	Task	All	Task
大语言模型 + 求解器	19.2 [13.8, 24.6]	33.4 [20.8, 44.8]	13.1 [9.2, 17.1]	40.6 [27.1, 54.2]
SMART-LLM 风格	12.9 [8.3, 17.7]	30.3 [16.7, 42.7]	9.5 [5.4, 13.8]	27.1 [13.5, 40.6]
RoCo 风格	24.4 [19.2, 29.6]	41.7 [31.2, 52.1]	15.0 [11.2, 19.0]	37.5 [25.0, 50.0]
HMAS-2 风格	10.4 [5.8, 14.8]	26.1 [14.6, 37.5]	12.4 [8.3, 16.7]	27.1 [13.5, 40.6]

在 96 个任务场景上进行重采样得到的 95%区间均值，按场景对方法进行配对比较，以及留出种子上的准确率。会话统计提供答案的模型会话，包括格式检查，但不包括实际执行任务的执行器。

A.6. 主要运行详情

本节的表格和图示扩展了论文的主要结果，除非图注另有说明，否则使用相同的回合。

区间。 按场景配对时，在所有回合中，MRScissors 减去各方法的结果为 9.5 至 24.4 个百分点；在任务层级上则为 26.1 至 41.7 个百分点，并且每个区间都不包含 0（表 10）：在两种骨干模型上，MRScissors 均领先于各方法。最小裕度出现在使用 Claude Sonnet 5.5 的 SMART-LLM 风格方法上，为 9.5 个百分点 [5.4，13。8]。将 16 个任务—层级配对而非 96 个任务—场景进行重采样时，每个区间同样都不包含 0；最低界限在所有回合中为 5.0 个百分点，在任务层级上为 11.5 个百分点；对 8 个任务进行重采样也得到相同结果。针对每种方法，在两种骨干模型上，MRScissors 均在全部 8 个任务中领先（图 8）：对任务进行单侧符号检验得到$p = 1/256 \approx 0.004$。

按部件。 表 13 给出了在每个配置、套件和水平下，MRScissors 相较于其他最佳方法的领先幅度，并列出了任务—层级配对上的区间。

相同的正确团队。 在 MRScissors 和某方法都指定了正确团队的可行回合中，MRScissors 使用 Qwen3.8-Max 时的成功率为 51.5% 至 55.4%，使用 ClaudeSonnet 5.5 时为 56.9% 至 58.2%；其他方法分别为 17.9% 至 24.0% 和 24.0% 至 27.4%（凸片14）。

执行器的上限。仅提供标准答案团队时，共享执行器在 312 个可行场景中分别有 58.7% 和 60.3% 的成功率（Oracle-team）；使用正确团队时，MRScissors 的回合成功率分别为 55.9% 和 58.8%，其他方法分别为 18.9% 至 25.2% 和 23.3% 至 27.5%。除 精确之外，

表11。**主要运行的鲁棒性**。按水平和房间（厨房、工作间、主页：随机种子0和3、1和4、2和5）统计的团队正确率（%），按带入的机械臂进行交换，并根据道具或干扰物是否为较大的物体进行借用，其中“拿较大的物体”会失败。

方法	水平		Room			交换		借用	
	easy	hard	厨房	工作间	home	SO-101 在内	其他	道具较大	较小的道具
Qwen3.8-Max									
大语言模型 + 求解器	64.2	68.3	63.1	66.9	68.8	45.8	54.2	52.2	48.3
SMART-LLM 风格	74.6	70.4	70.6	75.6	71.2	60.4	45.8	49.3	48.3
RoCo 风格	62.9	59.2	63.8	60.6	58.8	37.5	56.2	34.3	37.9
HMAS-2 风格	74.2	75.8	76.2	75.0	73.8	60.4	68.8	47.8	44.8
MRScissors（我们的方法）	**89.2**	**81.7**	**84.4**	**83.8**	**88.1**	**77.1**	**89.6**	**56.7**	**65.5**
Claude Sonnet 5.5									
大语言模型 + 求解器	75.4	80.8	79.4	77.5	77.5	68.8	68.8	56.7	51.7
SMART-LLM 风格	79.2	84.2	83.1	81.2	80.6	75.0	75.0	59.7	58.6
RoCo 风格	75.4	77.1	73.8	77.5	77.5	77.1	66.7	44.8	**72.4**
HMAS-2 风格	80.8	76.7	76.9	79.4	80.0	64.6	79.2	47.8	**72.4**
MRScissors（我们的方法）	**91.2**	**91.2**	**91.9**	**93.8**	**88.1**	**100.0**	**91.7**	**67.2**	65.5

表12。 **主要运行中按类型统计的错误答案**。 每行包含480个回合，每个错误答案对应一种类型（第 A.5节）； 分配团队和 拒绝 分别表示可行性的两个相反方向，即为不可行任务分配团队，以及拒绝可行任务； 其他团队：另一个使用最少机械臂数的团队。主要运行中没有缺失答案。

方法	错误	多余机械臂	缺失机械臂	分配团队	拒绝	使用干扰项	遗漏道具	其他团队
Qwen3.8-Max								
大语言模型 + 求解器	162	6	55	76	20	3	0	2
SMART-LLM 风格	132	9	30	46	31	14	0	2
RoCo 风格	187	4	46	87	30	18	0	2
HMAS-2 风格	120	15	13	38	20	28	0	6
MRScissors（我们的）	70	0	22	19	11	16	0	2
Claude Sonnet 5.5								
大语言模型 + 求解器	105	8	44	28	16	5	0	4
SMART-LLM 风格	88	10	31	23	15	8	0	1
RoCo 风格	114	6	38	25	27	10	0	8
HMAS-2 风格	102	18	13	27	21	17	0	6
MRScissors（我们的方法）	42	0	14	4	12	9	0	3

Oracle-script 为标准答案团队执行预设金标准路线。根据配置，MRScissors 的 MSS-SR 使用 Qwen3.8-Max 时为 42.7% 至 50.0%，使用 Claude Sonnet 5.5 时为 47.9% 至 55.2%；其他方法分别为 8.3% 至 21.9% 和 12.5% 至 25.0%。

留出种子。 表格 15 给出了在 400 个场景的随机种子 1–5 上各行的结果，其中 MRScissors 领先其他最佳方法 11.4 和 9.4 个百分点；以及在随机种子 0 的 80 个场景上的结果。

鲁棒性。 在不同层级之间，没有任何方法的变化超过 7.5 个百分点（MRScissors 搭配 Qwen3.8-Max 时，简单水平为 89.2%，困难水平为 81.7%），并且在两种骨干模型的两个层级上，MRScissors 均排名第一（表 11）；在不同房间之间，没有任何方法的变化超过 5.7 个百分点（搭配 Qwen3.8-Max 的大语言模型 +求解器，以及搭配 Claude Sonnet 5.5 的 MRScissors），并且 MRScissors 是

每个房间中首先如此。在交换情况下，当引入 SO-101（这总会使任务不可行）时，其他方法答对的频率与引入另一机械臂时大致相同（Qwen3.8-Max 分别为 51.0% 和 56.2%，Claude Sonnet 5.5 分别为 71.4% 和 72.4%）：它们没有将 SO-101 识别为不可行。在借用情况下，当道具较大时，拿取较大的物体才是正确的；而当干扰物较大时则是错误的。对于 MRScissors，差距为 −8.8 和 1.7 个百分点；对于其他方法，差距为 −27.6 至 5.0 个百分点，因此没有任何方法会根据大小进行选择。有些答案在同一任务场景中彼此矛盾，无论标准答案是什么：盈余向精确情况增加了一条机械臂，因此在精确可行的情况下，不可能不可行或需要更多机械臂；短缺减少了一条机械臂，因此在精确不可行的情况下，不可能可行，也不可能在机械臂更少时可行；在 Rea-02 和 Rea-04 中，交换

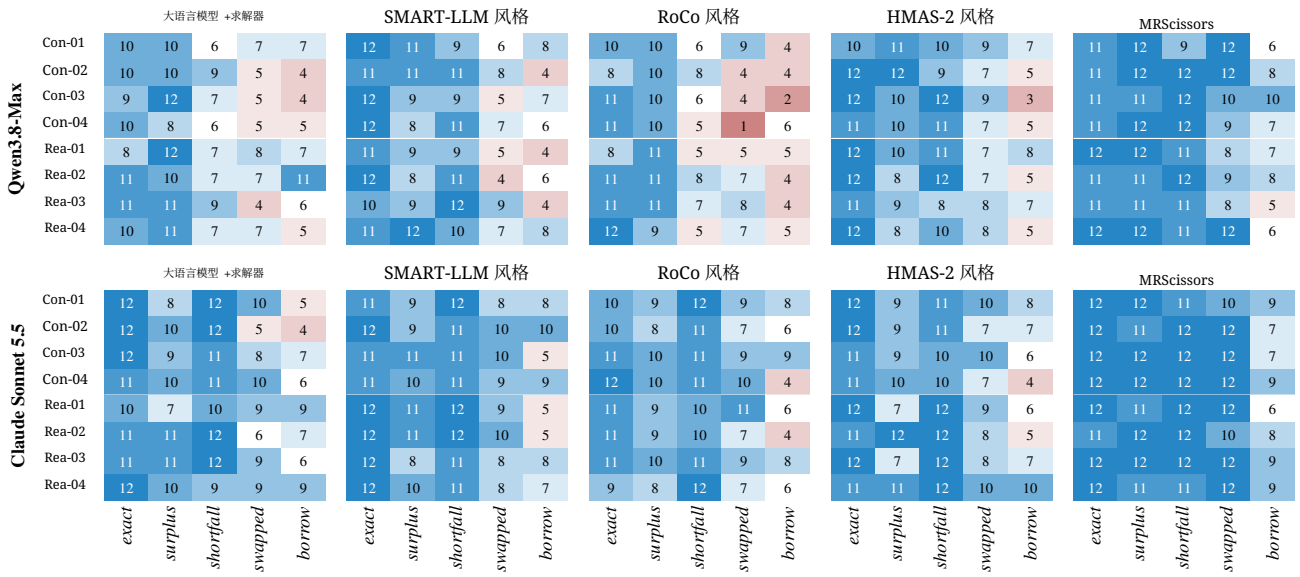


图 8。**按任务和配置划分的正确答案**。每个版本包含 12 个回合（两个层级、六个随机种子），按方法和骨干网络划分，颜色从红色（0）经过白色（6）过渡到蓝色（12）。

表13。**MRScissors 与其他最佳方法的差值，按部件划分**。以点数表示，按场景与每个部件中准确率最高的方法配对，并采用对 16 个任务—层级配对进行重采样的 95%自助法置信区间（10000 次重采样）；每个点表示两种准确率之差，准确率数值如正文和表11中所示。

Part	Qwen3.8-Max	Claude Sonnet 5.5
精确	−2.0 [−8.3, 4.2]	2.1 [−2.1, 6.2]
盈余	9.4 [2.1, 16.7]	14.6 [6.2, 24.0]
短缺	7.3 [1.0, 15.6]	3.1 [−1.0, 7.3]
交换	18.7 [8.3, 29.2]	20.8 [13.5, 28.1]
借用	8.4 [−10.4, 26.0]	7.3 [−6.2, 19.8]
并发	11.7 [2.9, 18.3]	9.2 [2.9, 15.0]
可达性	9.1 [1.7, 15.8]	10.0 [5.8, 14.2]
easy	14.6 [7.5, 20.8]	10.4 [3.3, 16.7]
hard	5.9 [−2.5, 13.8]	7.0 [1.2, 12.5]

表14。**相同正确团队下的成功率**。在两者都指定正确团队的 n 个可行回合中，MRScissors（我们的方法）和各方法（方法）的成功率（%）。

方法	Qwen3.8-Max			Claude Sonnet 5.5		
	n	Ours	方法	n	Ours	方法
大语言模型 +求解器	193	55.4	20.2	208	57.2	25.5
SMART-LLM 风格	196	51.5	24.0	223	57.0	27.4
RoCo 风格	184	54.3	17.9	204	56.9	24.0
HMAS-2 风格	199	54.8	22.1	220	58.2	26.8

将一台 Franka 放在 PiPER 的位置上，因此在精确可行的情况下，不可能不可行或需要更多机械臂。在每种方法的这 216 对样本中，MRScissors 与 Qwen3.8-Max 的答案有 12 对自相矛盾，与 Claude Sonnet 5.5 的答案有 3 对自相矛盾，而其他方法

表15。**留出种子**。在随机种子 1–5 的 400 个场景和随机种子 0 的 80 个场景上的主要运行（%）；可行、不可行：标准答案为可行或不可行。

方法	随机种子 1–5（400）			随机种子 0（80）		
	可行	Inf.	All	可行	Inf.	All
Qwen3.8-Max						
大语言模型 +求解器	72.3	56.4	66.8	73.1	46.4	63.8
SMART-LLM 风格	73.1	72.9	73.0	69.2	71.4	70.0
RoCo 风格	67.3	46.4	60.0	71.2	57.1	66.2
HMAS-2 风格	74.2	75.7	74.8	71.2	**85.7**	76.2
MRScissors（我们的）	**83.8**	**90.7**	**86.2**	**82.7**	78.6	**81.2**
Claude Sonnet 5.5						
大语言模型 +求解器	75.0	84.3	78.2	76.9	78.6	77.5
SMART-LLM 风格	79.2	86.4	81.8	78.8	85.7	81.2
RoCo 风格	73.5	86.4	78.0	61.5	78.6	67.5
HMAS-2 风格	76.2	82.1	78.2	75.0	92.9	81.2
MRScissors（我们的方法）	**87.7**	**97.9**	**91.2**	**88.5**	**96.4**	**91.2**

第 29 至 60 题和第 16 至 25 题中的方法，而 MRScissors 从未在可行之后回答盈余不可行精确。

逐任务分析。 图 8展示了各方法出错的位置：一个借用 单元格对两个骨干模型下的每种方法而言都是最困难的（单独或并列），但使用 Qwen3.8- Max 的 RoCo 风格方法除外，其最困难的是 Con-04 交换 （8.3%）。MRScis- sors 在两个骨干模型下的每个任务中均排名第一：使用 Qwen3.8-Max 时领先 3.3 至 16.7 个百分点，使用 Claude Sonnet 5.5 时领先 1.7 至 13.3 个百分点；其在使用 Qwen3.8-Max 时于 Con-02 和 Con-03 上的领先幅度最大（16.7 和 13.3 ），在使用 Claude Sonnet 5.5 时于 Rea-03 和 Con-04 上的领先幅度最大（13.3 和 11.7）。

按错误类型统计。 MRScissors 从未指定过多机械臂，而其他方法在使用 Qwen3.8-Max 时出现此类错误 4 至 15 次，在使用 Claude Sonnet 5.5 时出现 6 至 18 次（表 12）；它将团队分配给不可行任务的次数分别为 19 次和 4 次，而其他方法分别为 38 至 87 次和 23 至 28 次。其余错误包括缺失机械臂（22 次和 14 次）、使用干扰项（16 次和 9 次）、拒绝可行任务（11 次和 12 次），以及为最少机械臂数的任务分配了另一个团队（2 次和 3 次）。

MRScissors 按配置划分的错误。 借用 占据了 MRScissors 大多数错误答案：在 Qwen3.8-Max 中为 70 个答案中的 39 个，在 Claude Sonnet 5.5 中为 42 个答案中的 32 个。其中，它会拿取干扰物（16 个和 9 个）、在需要两条机械臂时只为道具指定一条机械臂（15 个和 11 个），或宣称任务不可行（8 个和 12 个）。对于 Qwen3.8-Max，另有 16 个错误出现在交换配置中，其中 13 个是分配团队发生交换后导致某项工作没有执行者。

A.7. 提示词

每个模型会话都会以一条作为系统消息的提示词开始：其中一行用于说明会话的角色，共享部分由提示词包含 {COMMON}，以及该方法自身的指令。我们逐字给出共享部分和每种方法自身指令中的关键段落；[...] 表示省略的文本，方括号标签用于标明会话，不会发送给模型，完整提示词将随代码一同发布。带大括号的字段会在调用前填充：{ARM}和 {ARMS}填入机械臂名称和团队名称，{SIDE}填入审查员的两个方向之一。会话只能通过观察接收图像，而工具则使用 Tab 的工具。8。提示词将被动对象称为物体。

共享部分。它说明了问题和回答的形式，每种方法都相同。

共享部分 ({COMMON})

问题

一组机器人手臂围绕一张桌子安装，桌上摆放着物体和一个目标。问题是完成该目标需要多大的团队：

– 需要哪些机械臂所需的最少机械臂数量，以及哪些机械臂可以保持空闲；- 桌上的物品能否代替机械臂 [...]；- 或者，现有条件无法完成任务，还缺少什么。

答案：`提交_团队_计划`

- `可行`；使用的 `机械臂_` （完成任务所需的最少机械臂数；如果无法完成则为空）； `空闲` （其他每个机械臂，并说明原因）；- 使用的 `物体_` （每个物体说明其执行的任务，以及如果它接替了某个机械臂的任务，则说明接替的是哪个）； [...] - `缺失项` （如果无法完成：说明哪个任务没有执行者，以及什么类型的机械臂或物体可以执行该任务）；- `步骤` （谁执行什么任务，以及同时发生什么）； `关键_约束条件` （哪些因素决定机械臂数量，并列出具体数字）。[...]

基线方法。 SMART-LLM 风格将其三个阶段作为三个会话运行，每个会话都会接收前一阶段的输出。在 RoCo 风格中，每个机械臂都在独立的会话中发言，按照当前机械臂的 W、S、E、N 顺序进行，最多进行三轮，并接收截至目前的对话内容和最新提案。在 HMAS-2 风格中，中央规划器接收计划板及其先前的提案和机械臂的回答；每个机械臂则在独立的会话中检查新提案，最多进行三轮。LLM +求解器使用一个会话。

基线方法：自定义指令

[SMART-LLM 风格 / 分解]观察场景（`任务简报`、 `观察`、 `定位`、 ` 测量`），并将目标分解为子任务。 [...] 暂时不要分配机械臂。

[SMART-LLM 风格 / 联盟形成]对于每个子任务，找出能够完成它的对象：- **机械臂**：询问每个机械臂是否具备完成该子任务所需的技能—— `能否_在其位置` 处达到目标，指出它所需的方向（读取裕度）； `读取_机械臂` 以确定手指的张开程度； `确定`在质量决定结果的情况下所能承受的重量；- 桌子上可以代替机械臂完成任务的**物体** [...]： `测量` 它们，并说明每个物体为何能够或不能完成任务。没有任何机械臂或物体能够完成的子任务，其联盟为空。

[SMART-LLM 风格 / 任务分配]将每个子任务分配给其联盟中的一名成员，同时使用数量最少的机器人，确保必须同时进行的子任务由不同的执行者完成，并保持每个子任务的顺序

[.....] 如果某个子任务的联盟为空，则该任务无法完成：请说明是哪个子任务，以及需要哪种机械臂或物体来完成它。

[RoCo 风格 / 每个机械臂]你是 {ARM} 机械臂，是安装在同一张桌子周围的机器人手臂（{ARMS}）之一，每个机械臂都由一个独立的你运行 [.....]。你与其他机械臂交流，以共同制定团队计划；在你们全部同意某个计划之前，任何物体都不会移动。[.....]你只能查询自己的关节（`读取_机械臂`）和自己的运动学（`可_达`）；请在消息中询问其他机械臂有关其自身情况的信息。[.....] - `数据.参与_部件`：表示你是否参与（你可以选择不参与）；[.....]当每个机械臂都同意某个提案时，对话结束。

[HMAS-2 风格 / 中央规划器]你提出完整计划；每个机械臂使用自身的运动学检查分配给自己的部件，并接受或否决该计划。你不会自行查询机械臂的运动学[.....] 如果计划板上显示有否决意见，请修改计划以回应这些意见。

[HMAS-2 风格 / 每个机械臂]
使用你自己的工具检查你自己的部件—— `是否_可达` 要求你工作的各个位置（读取裕度）， `读取_机械臂`， [.....] ——或者，如果它让你处于空闲状态，检查这样是否正确。然后 `回应`：接受或否决（`data.accept`）、理由以及你的证据（工具结果）。

[大语言模型 + 求解器]将任务转换为作业图；求解器计算所需的机械臂数量 [.....]。[.....]

2. **编写任务。** 明确手（或物体）必须处于什么位置、必须做什么，以及所需的力度和持续时间。用“同时”连接必须同时发生的任务。 [ ] .....。**寻找执行者，并提供证据。** [.....] 只有有证据表明某个物体确实执行某项任务时，它才是该任务的执行者 [.....] 并且必须先由某个机械臂将其放置到位：添加该任务。4. **求解。** `使用任务和转移来求解_团队` 。如果遇到阻塞，查看原因；只有依据证据修改图，然后再次求解。如果仍无济于事，则任务无法完成：说明缺少什么。

MRScissors。 证明器会接收任务简报和不含机械臂的场景图；对于手，只会接收最宽的开口。它没有可达性工具，而它的相机图像会显示包括机械臂在内的整张桌子。替代物流程会接收不含执行者的作业图和任务；它添加的物体会附带放置该物体的任务，而已经写入的任务保持不变。两名审查员共享一个提示词，其中 {SIDE} 由两方之一填充，并接收目标、作业图、证书和经过翻转测试的元素；仲裁器接收目标、作业图以及存在争议的编辑内容。

MRScissors：自身指令

[证明器]
你要将桌面上的任务建模为一个**作业图**：目标要求完成什么、在哪里完成，以及哪些事情必须同时发生。机器人手臂将执行这些操作，但不会告知你有哪些机械臂、它们位于何处或能够到达什么位置——这些信息会在之后由代码针对你指定的每个位置和每个机械臂进行测量。 [...] - **任务**——必须完成什么，以及一只手（或代替手的物体）必须位于何处才能完成。[...] - **组**——具有相同组名的任务会同时完成，每项任务分别由一只手执行。 [...] - **物体**——桌面上的物体，不是目标所指明的物体；计划会将其放置到位，使任务能够完成。 [...] 列出 `物体`中的每个候选物及其决定性尺寸，代码会测量每个候选物，并保留符合要求的那个。[...]写出目标所需的每项任务；写出一项目标不需要的任务同样是错误。

[替代物流程]
你现在只需问一个问题：**桌子上是否有一个物体，放置到位后可以完成其中的任何一项任务？** [...] 任务本身保持你所写的样子：在此处不能更改、删除或添加任务。 [...]

[审查员]
每个单元格（谁到达何处、谁与谁会合）都是工具结果。 [.....] 可能错误的是模型所写的内容：遗漏或不需要的任务、把放置写在并不存在该物体的位置、哪些事项是或不是同时完成的、谁可以执行某项任务，以及某个物体的规格。 [.....]{SIDE}[.....]每次编辑都需要 `证据`：作为其依据的工具结果，**必须来自你在本次会话中的自行调用**；证据未指明你调用过任何工具的编辑将被拒绝。使任务更容易的编辑还需要 `引文`：作为其依据的目标原文，必须逐字复制。[.....]

代码会应用每次编辑，测量其中指明的内容，并重新绘制答案。不产生任何变化的编辑将被丢弃。使机械臂的任务更困难的编辑将根据其证据予以保留。使任务更容易的编辑不会仅凭自身被保留：它会连同引文和证据一起提交给仲裁器，只有在目标和场景能够证实该编辑时，仲裁器才会保留它。

[{SIDE}：太简单]观察图中遗漏的内容，或比任务实际情况描述得更简单的内容：- 目标所需但缺失的一项任务 [...]；- 必须同时完成但未归入同一组的任务。 [...]

[{SIDE}：太困难]
观察图中要求而目标不要求的内容：- 目标不需要的一项任务；- 同一组中不必同时完成的任务（其中一项可以先完成并放手，然后再完成另一项）；[...] - 任何一只手都无法完成、但桌上的物品放置到位后可以完成的一项任务 [...]

[仲裁器]
对于每个编辑单元，只决定一件事：按文字所写的目标，以及工具所呈现的场景，是否支持这项编辑？只有在支持时才保留。如果无法确定，就拒绝该编辑：模型按原文保持不变。

共享执行器与格式检查。 执行器使用 Tab 的程序 API，为某个方法的团队计划编写程序。8。RoCo 风格或 HMAS-2 风格提交的计划若未通过格式检查，则会被发送一次到一个仅修复其格式的会话中。

共享执行器与格式检查

[执行器]你要编写一个程序，由安装在同一张桌子周围的机器人手臂团队运行，以完成一项任务。哪些机械臂参与、哪些保持空闲，以及桌上使用哪些物体，已经在第一条消息中的团队计划里确定。 [ ] .... 遵循团队计划：[ ]
不要移动其中要求保持空闲的机械臂。

[格式检查]你要将已达成一致的团队计划填入提交表单。 [... ...] 不要更改该决定。