面向机器人操作的大型行为模型协同训练：数据模态与策略的系统性研究

Fanqi Lin[†‡]、Kushal Arora[†]、Jean Mercat[†]、Haruki Nishimura[†]、Paarth Shah[†]、Chen Xu[†]、Mengchao Zhang[†]、Mark Zolotas[†]、Maya Angeles[†]、Owen Pfannenstiehl[†]、Andrew Beaulieu[†]、Jose Barreiros[†]

[†]丰田研究院，美国马萨诸塞州剑桥和加利福尼亚州洛斯阿尔托斯

[‡]清华大学，中国北京

arXiv:2602.01067v1 [cs.RO] 1 Feb 2026

摘要—**大型行为模型（LBMs）通过将模仿学习扩展至基于大规模多任务机器人数据的训练，展现出强大的灵巧操作能力，但其泛化能力仍受限于现有机器人数据覆盖范围不足。为扩大数据覆盖范围，同时避免高昂的额外数据收集成本，近期研究越来越多地采用**协同训练**：联合利用目标机器人数据和异构数据模态进行学习。然而，不同协同训练数据模态和训练策略对策略性能的影响仍缺乏深入理解。我们开展了一项大规模实证研究，考察五种协同训练数据模态——标准视觉语言数据、机器人轨迹的密集语言标注、跨具身机器人数据、人类视频和离散机器人动作令牌——以及单阶段和多阶段训练策略。我们的研究利用 4,000 小时的机器人与人类操作数据，以及 5,000 万个视觉-语言样本，训练视觉-语言-动作（VLA）策略。我们对 89 个策略进行了评估，涵盖 58,000 次仿真运行轨迹和 2,835 次现实世界运行轨迹。结果表明，利用多种形式的视觉-语言数据和跨具身机器人数据进行协同训练，能够显著提升模型在分布偏移、未见任务和语言跟随方面的泛化能力，而离散动作令牌变体并未带来统计上显著的收益。此外，结合有效模态可带来累积增益，并通过微调实现对未见长时域灵巧任务的快速适应。我们发现，仅使用机器人数据训练会削弱视觉-语言模型骨干网络的视觉语言理解能力，而通过有效数据模态进行协同训练则可恢复这些能力；这一结论由标准视觉-语言、空间推理和多模态推理基准测试结果衡量得出。最后，在仿真基准中，明确以从协同训练数据中学习到的思维链轨迹为条件进行动作生成，并未提升性能。综上，这些结果系统地揭示了协同训练的特点，并为构建可扩展的通用机器人策略提供了实用指导。**

项目主页：https://co-training-lbm.github.io/

I. 引言

机器人学习正日益转向通用模型，使其能够感知、理解并在物理世界中采取行动。近期的研究重点是训练大型行为模型（LBMs） [1]——在大规模多任务机器人数据集上训练的具身基础模型——以生成灵巧操作策略。在这一类模型中，视觉-语言-动作模型（VLAs） [37, 100, 21, 6, 33, 67, 5] 是一种具有代表性的 LBM 子类，它们融合视觉和语言输入以生成动作。尽管取得了进展，LBM 在语义与空间理解以及开放世界泛化方面仍落后于视觉-语言模型（VLMs） [32, 16, 65, 73], 等非具身基础模型。这一局限可归因于数据规模 [25]的巨大差异：机器人数据的规模比用于训练 VLM 的互联网规模文本与图像语料小了几个数量级。

为弥合这一数据鸿沟，近期许多研究 [33, 44, 5, 39, 13, 61, 84]采用协同训练——将目标机器人（即部署具身形态）的数据与异构数据模态结合起来进行联合学习，旨在增强模型对物理世界的理解及其泛化能力。这些协同训练数据模态包括：标准视觉-语言（VL）数据 [33, 99, 98, 39], 机器人轨迹的密集语言标注 [92, 44, 93, 84, 61, 11],跨具身机器人数据 [1, 50, 48, 69, 37, 38], 人类视频 [10, 13, 88, 52, 90, 36],以及离散机器人动作令牌 [93, 22, 35, 33]。尽管人们对此日益关注，当前研究通常只使用这些模态中的一部分，并采用不一致的实验设置，因此协同训练的实证有效性仍在很大程度上未得到探索。

在这项工作中，我们通过大规模实验，系统地研究不同的数据模态和协同训练策略如何影响策略性能，以构建通用大型行为模型。我们采用由预训练VLM骨干网络和动作头组成的VLA架构。模型通过流匹配 [46, 49] 来预测连续的机器人动作，并通过下一词元目标来预测离散令牌。图 1 展示了本研究的概览。

我们评估了五种主要的协同训练数据模态（图 2）：**(1) 标准视觉语言数据**，涵盖视觉问答、物体定位和空间推理任务，为物理世界提供丰富的常识知识。**(2) 机器人轨迹的密集语言标注**，通过启发式脚本编写和基于VLM的描述生成来生成，为动作、目标和物体关系提供明确的语义标签。**(3) 跨形态机器人数据**，涵盖不同机器人形态和环境下的操作演示。**(4) 大规模第一人称人类视频**，

通讯作者：jose.barreiros@tri.global

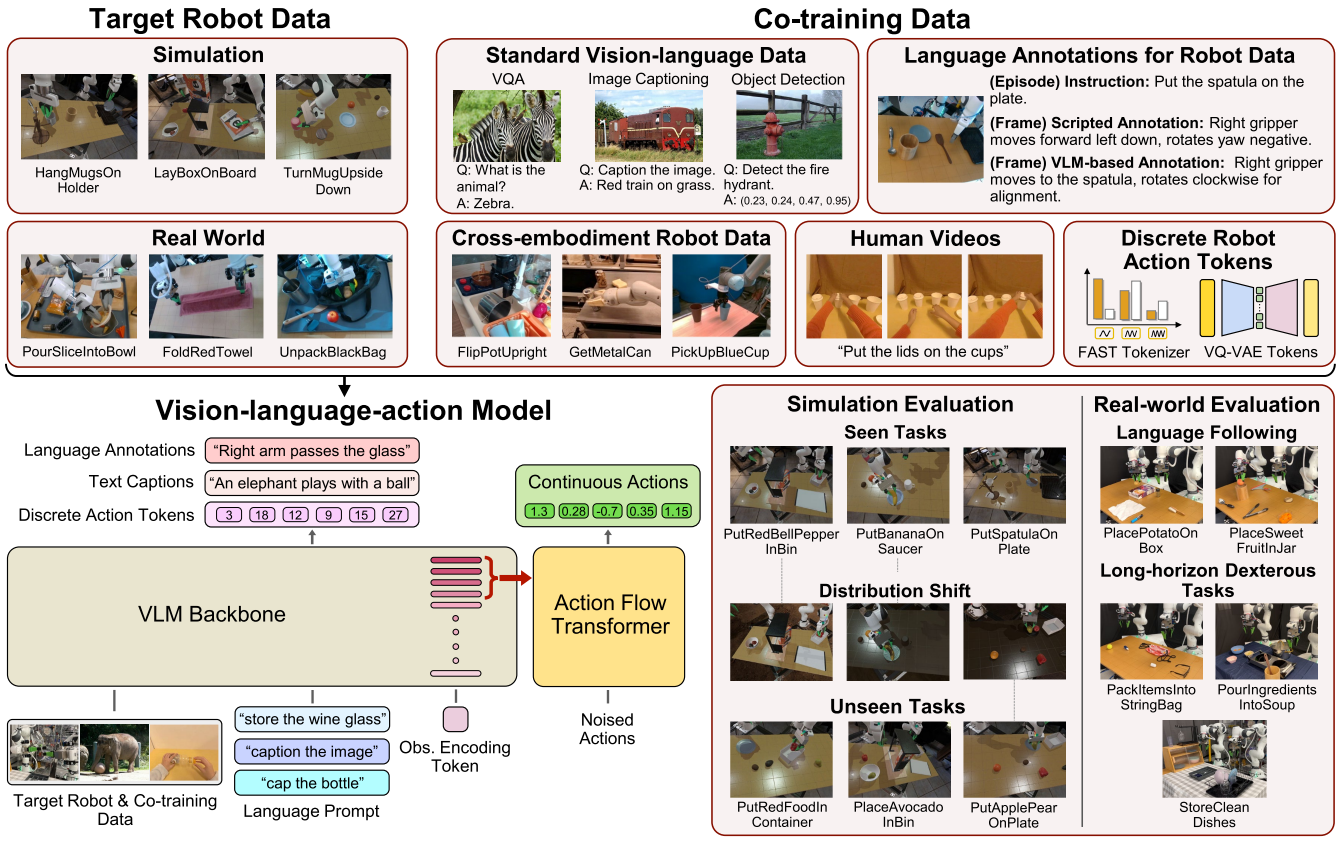


图 1. **数据、模型架构与评估设置概览。** 我们的策略基于预训练视觉语言模型骨干网络，并结合 Action Flow Transformer 构建。该策略在目标机器人数据上训练，同时使用异构协同训练模态，包括标准视觉语言数据、机器人数据的密集语言标注、跨具身机器人数据、人类视频以及离散机器人动作令牌。我们在仿真中对已见任务和未见任务、在标称条件和分布偏移下评估策略，并在真实世界中评估语言跟随和长时程灵巧操作。

使模型能够接触到机器人轨迹之外的多样化视觉场景、物体交互和运动模式。我们探索了两种利用人类视频的方法：第一，从帧序列中提取离散潜在动作标记；第二，向视觉语言模型提供视觉上下文和任务指令，生成语言标注，逐帧描述动作、目标和物体。**(5) 离散机器人动作令牌**，通过基于频率的方法（例如 FAST [58]）或基于向量量化的方法（例如 VQ-VAE [72]）将连续机器人动作压缩到低维离散空间，以探究这类抽象是否能提升泛化能力。

我们还研究了在不同训练阶段（即使用不同数据构成进行的训练轮次）融入这些模态的策略。此外，我们考察了结合有效的协同训练模态是否能带来累积的性能提升。我们进一步通过在未见过的长时程灵巧任务上进行微调，探究协同训练能否提升表示质量。随后，我们使用一系列标准视觉语言基准，研究有效的协同训练模态如何塑造视觉语言模型骨干网络。最后，我们研究用于动作生成的显式思维链（CoT） [77] 条件化：策略首先生成从协同训练数据中学习到的中间思维链轨迹，然后利用这些轨迹生成连续动作。

我们的策略在仿真和真实世界环境中均接受评估。总计，我们训练并比较了 89 个 VLA 策略，使用约 4,000 小时的机器人和人类操作数据，以及 5,000 万个视觉语言样本。这些策略在已见和未见任务的常规及分布偏移（DS）设置下，接受了 58,000 次仿真运行轨迹评估；此外，还通过 2,835 次现实世界运行轨迹，评估了语言跟随和长时程灵巧任务表现。

我们的结果为大型行为模型的协同训练提供了实用指导。总结如下：（1）多样化的视觉语言数据和跨具身机器人数据能够持续提升模型在分布偏移、未见任务和语言遵循方面的泛化能力，而离散动作令牌变体则没有带来任何收益。（2）结合有效模态能够累积提升效果，并通过微调更高效地适应未见的长时域灵巧任务。（3）仅使用机器人数据训练会削弱视觉语言模型骨干网络的视觉语言能力，而有效的协同训练有助于保留这种理解能力，这一点体现在标准视觉语言基准测试中性能的提升上。（4）在我们的仿真基准测试中，明确以从协同训练数据中学习到的思维链轨迹作为动作生成的条件，并不能提升性能。总之，这些发现系统地描绘了哪些协同训练信号和策略最有助于构建可扩展的通用机器人

策略。

II. 方法

本节首先在第 II-A 节介绍我们的协同训练框架，包括问题形式化、模型架构以及我们的协同训练和推理策略。随后，我们在第 II-B 节介绍如何为本研究整理目标机器人数据和多样化的协同训练数据集。

A. 协同训练框架

1）问题形式化：我们的目标是学习一种策略π_θ，使其能够利用多样化的协同训练数据模态。策略 π_θ 以一系列 n 图像 $I^{1:n}$ 和文本提示 ℓ作为输入。对于连续机器人动作，模型以流匹配（FM） [46, 49] 作为学习目标进行训练。具体而言，给定一个动作块 A_t、一个 FM 时间步 $\tau \in[0,1]$, 和采样得到的噪声 $\epsilon \sim N(0, \mathbf{I})$，我们构造加噪动作块 $A_t^\tau = \tau A_t + (1-\tau)\epsilon$。随后，通过最小化以下损失来训练模型：

$$\mathcal{L}_{\mathrm{FM}} = \left\| \pi_\theta^a\left(I_t^{1:n}, \ell, A_t^\tau, \tau\right) - (A_t - \epsilon) \right\|^2, \tag{1}$$

其中，$(A_t - \epsilon)$ 和 $\pi_\theta^a(\cdot)$ 分别是真实流向量和预测流向量。对于文本词元或离散动作词元，模型通过最小化真实词元序列 $x_{1:M}$ 与预测逻辑值 $\pi_\theta^\ell(\cdot)$之间的交叉熵（CE）损失进行优化，即：

$$\mathcal{L}_{\mathrm{CE}} = \mathcal{H}\left(x_{1:M},\ \pi_\theta^\ell\left(I_t^{1:n},\ \ell\right)\right) \tag{2}$$

在对连续和离散模态进行联合优化时，我们将各项目标组合为加权和：

$$\mathcal{L} = M_{FM}\mathcal{L}_{FM} + w * M_{CE}\mathcal{L}_{CE}, \tag{3}$$

其中，w 是应用于 CE 损失的权重，M_{FM} 是 FM 损失掩码，用于指示模型是否应针对给定样本预测连续动作；M_{CE} 是用于指定计算 CE 损失时所用词元位置的掩码。

2）模型架构：我们采用一种 VLA架构（图 1），由预训练VLM骨干网络和流变换器动作头（ActionFT）组成。VLM 使用 PaliGemma2-PT（google/paligemma2-3b-pt-224 [65]）进行初始化。它对观测到的图像和描述任务的语言提示进行编码；除提供视觉-语言表征外，还可以选择训练它生成文本或离散动作词元。为了获得用于动作生成的紧凑表示，与 [45], 类似，我们在骨干网络的词汇表中引入一个特殊的观测编码词元，并将其附加到文本提示的末尾。我们从 VLM 最后四层中提取与该词元对应的隐藏状态向量，形成单个全局条件嵌入，并将其输入 ActionFT。ActionFT 遵循 [1] 中提出的扩散变换器设计。它由 8 个流变换器层组成，每层通过自适应层归一化（adaLN）多层感知机 [56], 以观测特征和流匹配时间步为条件。ActionFT 接收全局条件嵌入、一段加噪的连续动作以及流匹配时间变量，并经过训练后，模型能够预测流向量，引导动作迭代去噪并趋近目标轨迹。与常见方法 [6], 不同，我们的方法仅使用单个词元作为视觉-语言表征，而不使用来自 VLM 所有层的注意力键和值。我们的消融研究（见附录 1A）表明，这种紧凑表示能够提升模型对未见任务和分布偏移的泛化能力。

3）协同训练与推理策略：协同训练数据可用于模型训练的不同阶段。我们探索了三种策略（表 I）：(1)单阶段协同训练：在单个阶段中，使用目标机器人数据和协同训练数据模态进行联合训练；(2) 两阶段仅第一阶段协同训练：第一阶段仅使用协同训练模态进行训练，第二阶段则使用目标机器人连续动作进行训练；(3) 两阶段完整协同训练：第一阶段与策略 (2) 相同，但第二阶段同时使用协同训练数据和目标机器人数据进行训练。

我们根据消融实验结果（附录 1B）固定训练损失权重和批次数据比例。完整的训练细节见附录 2。

除了研究何时纳入协同训练数据外，我们还探讨如何利用这些数据。虽然我们的主要方法仅将协同训练数据用作辅助监督，但近期研究 [44, 92, 93, 67, 11, 13] 表明，显式地以思维链轨迹为条件进行动作预测，可以提升策略性能。因此，我们也针对选定的协同训练模态测试了这一替代范式：机器人轨迹的语言标注，以及视频的潜在动作。

在推理时根据思维链轨迹生成动作时，视觉语言模型骨干网络首先生成一条思维链轨迹（见图 S4）。随后，将观测标记附加在其末尾，并提取由此得到的视觉语言嵌入（编码图像、任务提示和思维链轨迹），以此为 ActionFT 提供条件，进行连续动作预测。在训练期间，我们引入概率思维链条件化：以概率p，训练策略根据从协同训练数据中提取的思维链轨迹生成动作；以概率 $1-p$，策略则不使用思维链条件化，直接生成动作。重要的是，思维链条件化与其他协同训练策略的区别仅在于，预测出的协同训练标记会被直接用于构成视觉语言嵌入，而不是仅提供辅助监督。我们在第 III-F 节评估这种显式推理时思维链方法的影响。

B. 数据整理

我们整理了一个综合数据集，其中包含目标机器人专家演示数据和五种不同的协同训练模态。该数据集包含约 4,000 小时的操作数据，涵盖机器人回合和人类视频；此外还包含 5,000 万个视觉语言样本，既包括标准视觉语言数据，也包括针对机器人和人类数据的密集标注。图 2 展示了训练数据概览。

1）目标机器人数据：我们采用此前研究 [1] 中的高质量机器人数据，作为主要的分布内

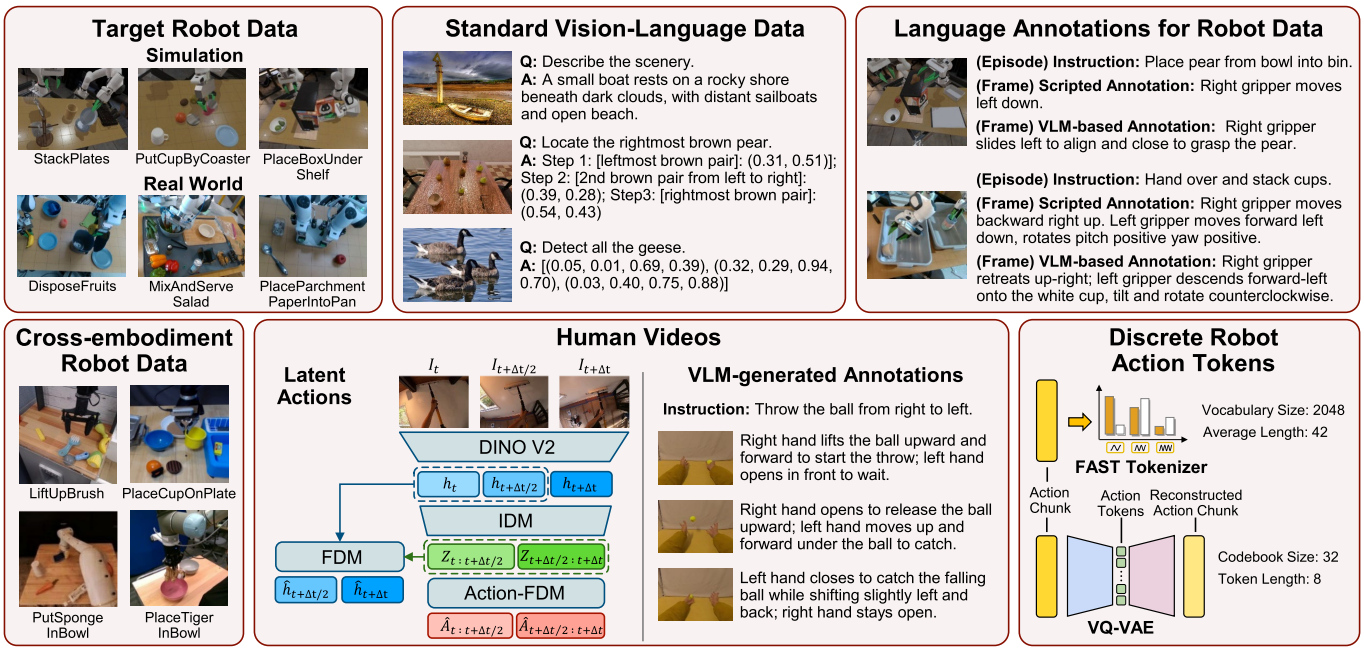


图 2. **训练数据概览**。我们的数据集包含在仿真和真实世界中采集的目标机器人数据，以及五种异构协同训练数据模态：用于常识理解、空间推理和物体定位的标准视觉语言数据；通过启发式脚本生成和基于VLM的描述生成获得的机器人数据密集语言标注；涵盖多种机器人形态和操作任务的跨具身机器人数据；人类视频，我们从中利用潜在动作模型（LAM）提取潜在动作标记，或生成视觉语言模型标注；以及离散机器人动作令牌，包括近无损 FAST 令牌和紧凑的 VQ-VAE 令牌。这些协同训练模态与目标机器人数据共同构成统一数据集，包含 ∼4,000 小时的操作数据和 50M 个视觉语言样本。

表 I 不同协同训练策略的数据构成。

协同训练策略	阶段	数据构成	
		目标机器人连续动作数据	协同训练 data
单阶段协同训练	第一阶段	✓	✓
两阶段仅第一阶段	第一阶段	–	✓
协同训练	第二阶段	✓	–
完整两阶段协同训练	第一阶段	–	✓
	第二阶段	✓	✓

训练语料库。该数据集称为 TRI-Ramen，包含 523 小时的操作数据，涵盖 403 项任务和 53,411 条示范数据。它由真实世界数据（478 小时、362 项任务、46,063 条示范数据，即 TRI-Ramen-Real）和仿真数据（45 小时、41 项任务、7,348 条示范数据，即 TRI-Ramen-Sim）组成。所有示范数据均按照 [1] 中的描述，通过遥操作双 Franka Panda 机械臂采集。

观测空间包括 i）4 张 RGB 图像（缺失图像以零填充），以及 ii）一条自然语言指令。动作空间包括 i）相对于工作站基座坐标系的末端执行器位姿，以及 ii）夹爪宽度。动作按照 [15] 中的方式表示为相对轨迹。策略仅以当前观测为条件，并预测一个预测跨度为 16 的动作块。

2) 标准视觉-语言数据： 为增强模型的多模态理解能力（例如语义、空间和规划），我们引入了专为机器人场景设计的视觉-语言数据集：RoboPoint [91] 和 RefSpatial [96]。RoboPoint 包含 130 万条数据样本和 820 万个问答对，而 RefSpatial 包含 250 万条样本和 2000 万个问答对。两个数据集都提供了针对空间指代任务的标注，涵盖定性视觉问答、关于物体及其空间属性/关系的定量查询、点坐标预测，以及多步空间推理。

3) 机器人轨迹的密集语言标注： 为使用动作关联文本描述增强 TRI-Ramen 机器人轨迹，我们采用两种互补的标注策略：**（1）脚本化标注**。遵循 [92],，我们通过在 16 步时间范围内（对应动作块的时间范围）比较未来与当前的末端执行器状态，应用启发式规则生成逐步的低层动作原语。这些标注记录了机器人的末端执行器平移运动、旋转变化以及夹爪状态转换。**（2）基于视觉语言模型的标注**。虽然脚本标注能够提供结构化的动作描述，但其语言多样性有限，并且缺乏有关物体和环境交互的上下文信息。为此，我们提示 VLM（GPT-5 [53]）生成丰富且具有上下文关联的描述。具体而言，对于每个机器人回合，我们向 VLM 提供：（i）以 2 秒为间隔下采样的帧

（约对应 1.6 秒的动作块时长），（ii）片段级任务指令，（iii）由上述启发式规则生成的动作提示，以及（iv）一张用于校准空间方向（前方/后方、左方/右方、上方/下方）的世界坐标系参考图像。随后，我们提示 VLM 生成多样化的帧级动作描述，以记录与物体及环境的交互。为获得更高的时间密度，我们以两次处理数据集，并在两次处理之间错开 1 秒，从而生成按秒频率提供的标注。更多详情请参见附录 3A。

4) 跨具身机器人数据：我们采用了 [1]，中使用的同一跨具身数据集，即 OXE-Ramen；该数据集是 Open X-Embodiment 数据集 [55] 的精选子集。该数据集涵盖多种机器人形态和操作场景，共计 1,150 小时的数据，涉及 12 种机器人配置、924 项任务和 466,415 条示范数据。观测空间和动作空间与目标机器人数据保持一致。

5) 人类视频：为了从人类视频中提取丰富的运动和动作信息（例如，向前移动和抓取），我们探索了两种不同的方法：**(1) 潜在动作**。我们使用公开可用的第一人称人类视频数据集（例如，Ego4D [28]、EgoDex [31]、Something-Something V2 [27]、Epic Kitchen [18]、HoloAssist [75]），经过筛选后总计 2,271 小时（详细数据构成见表 S3）。我们在人类视频、TRI-Ramen 和 OXE-Ramen 上联合训练潜在动作模型（LAM），以学习统一的离散动作表示。给定连续帧 I_t、$I_{t+\frac{\Delta t}{2}}$、$I_{t+\Delta t}$，我们使用预训练的 DINOv2 [54] 视觉编码器对其进行编码，以获得视觉特征 h_t、$h_{t+\frac{\Delta t}{2}}$、$h_{t+\Delta t}$。遵循 [13, 10]，LAM 使用三个模块学习潜在动作的量化码本（码本大小为 C）：逆动力学模型（IDM）、视觉前向动力学模型（FDM）和动作前向动力学模型（ActionFDM）。

IDM 预测两个潜在动作片段：

$$Z_{t:t+\frac{\Delta t}{2}},\ Z_{t+\frac{\Delta t}{2}:t+\Delta t} = \mathrm{IDM}\Big(h_t,\ h_{t+\frac{\Delta t}{2}},\ h_{t+\Delta t}\Big), \quad (4)$$

而 FDM 则重建未来的视觉特征：

$$\hat{h}_{t+\frac{\Delta t}{2}} = \mathrm{FDM}\Big(h_t,\ Z_{t:t+\frac{\Delta t}{2}}\Big), \quad (5)$$

$$\hat{h}_{t+\Delta t} = \mathrm{FDM}\Big(h_{t+\frac{\Delta t}{2}},\ Z_{t+\frac{\Delta t}{2}:t+\Delta t}\Big). \quad (6)$$

为了鼓励潜在令牌在捕捉视觉变化的同时也捕捉物理动力学，我们还通过重建真实动作块，在机器人数据（TRI-Ramen、OXE-Ramen）上进行监督：

$$\hat{A}_{t:t+\frac{\Delta t}{2}} = \mathrm{ActionFDM}\Big(Z_{t:t+\frac{\Delta t}{2}}\Big), \quad (7)$$

$$\hat{A}_{t+\frac{\Delta t}{2}:t+\Delta t} = \mathrm{ActionFDM}\Big(Z_{t+\frac{\Delta t}{2}:t+\Delta t}\Big). \quad (8)$$

对于人类视频，由于缺少真实动作，我们不进行动作重建。训练完成后，我们运行推理以获得每个时间步的 $Z_{t:t+\frac{\Delta t}{2}}$ 和 $Z_{t+\frac{\Delta t}{2}:\,t+\Delta t}$，— —

将它们拼接起来形成 Z_t，并将每个 Z_t 量化为从大小为 $C = 32$的码本中取出的 8 个离散令牌。对于机器人视频，我们使用 $\Delta t = 1.6$ 秒（对应 16 步的动作块）；而对于人类视频，我们使用 $\Delta t = 1.0$ 秒，以适应更快的运动。实现细节见附录 3B.1。**（2）视觉语言模型生成的标注**。作为潜在动作的替代方案，语言也可以作为不同本体间的统一表示。它能够捕捉有关动作、目标和物体的丰富语义信息，同时天然兼容视觉-语言-动作模型。具体而言，我们向 GPT-5 提供：（i）以 1 秒间隔下采样的人类视频帧；（ii）片段级指令；以及（iii）一张描绘世界坐标系三轴的参考图像。我们提示视觉语言模型生成对双手动作的简洁描述，其中包含有关其与物体及环境交互的丰富信息。我们使用 Ego4D、EgoDex 和 Something-Something V2 数据集，获得了 900 万个标注数据样本。训练期间，我们将这些样本视为另一种形式的视觉-语言数据。更多详情见附录 3B.2。

6）离散机器人动作令牌：多项研究表明，对连续和离散机器人动作表征进行协同训练，能够提高样本效率和泛化能力。受此启发，我们探索了两种形式的离散机器人动作令牌：**（1）FAST 令牌**。我们采用 FAST [58] 将连续动作块转换为压缩且近乎无损的离散令牌序列。我们使用现成的分词器，不进行微调，因为我们发现，在我们的数据上微调后，平均令牌长度或重建误差均未显著改善（见附录 3C）。将 FAST 应用于 TRI-Ramen 数据集后，生成的序列平均长度为 42.1，词汇表包含 2,048 个令牌。**（2）VQ-VAE 离散动作令牌**。我们在 TRI-Ramen 和 OXE-Ramen 上使用 VQ-VAE [72]，将动作块压缩为 8 个离散令牌，码本大小为 32。值得注意的是，这些维度与从视频中学习到的潜在动作维度完全相同。与 FAST 令牌相比，这种表示更加紧凑，维度也更低。

III. 实验

为系统地研究协同训练数据模态和策略的有效性，我们开展了大规模实验，以回答以下研究问题：

1) 在不同训练阶段引入不同的协同训练数据模态，会如何影响策略在各个方面的表现（分布内表现、对分布偏移的泛化、未见任务以及语言跟随）？ 2) 结合有效的协同训练模态能否带来累积的性能提升？ 3) 协同训练能否提升所学表征的质量，从而通过微调实现对未见的长时程灵巧任务的快速适应？ 4) 有效的协同训练模态会如何塑造视觉语言模型骨干网络?

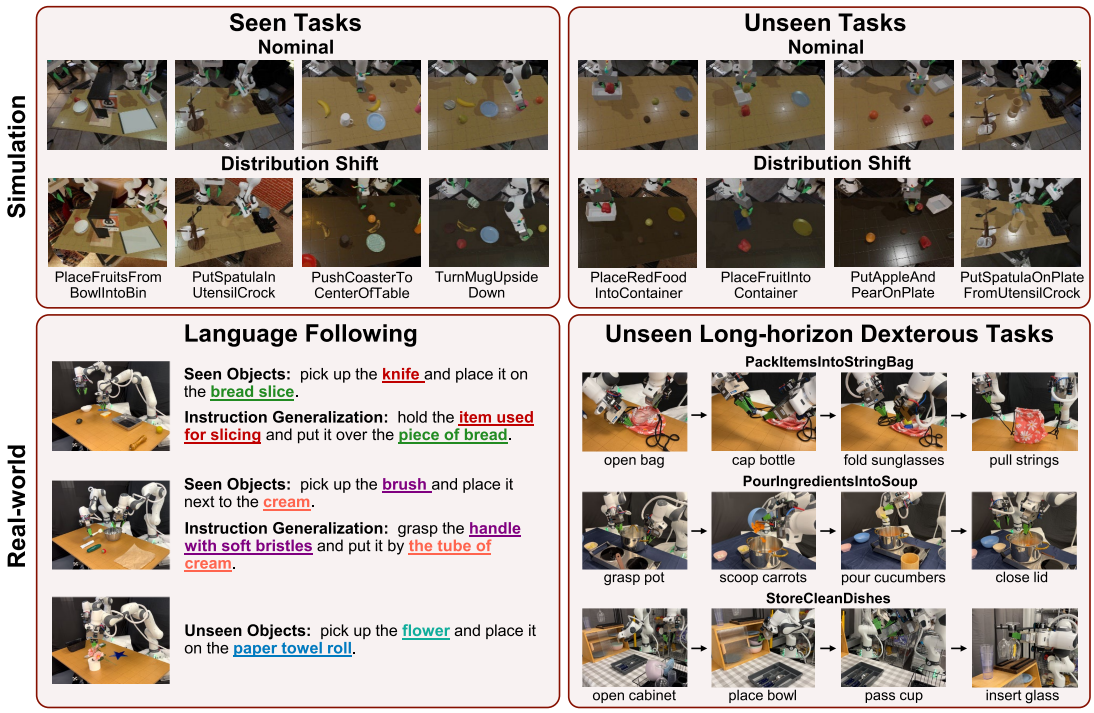


图 3. **仿真与真实世界评估。** 我们在名义条件和分布偏移（DS）条件下，评估策略在 13 个已见任务和 8 个未见任务上的表现；DS 会引入外观变化（例如光照、纹理、干扰物和相机参数）。真实世界评估包括使用已见物体进行的语言跟随实验、通过释义改写测试指令泛化，以及使用未见物体进行的实验；此外，还通过微调评估策略对未见长时域灵巧任务的适应能力。

5) 显式地以从协同训练数据中学到的思维链为条件进行动作生成，会如何影响策略表现?

A. 评估

我们从多个重要维度评估策略性能：分布内性能、对分布偏移（DS）的鲁棒性、对未见任务的泛化能力，以及语言跟随能力。此外，我们还评估策略通过微调适应未见长时域灵巧操作任务的能力。为此，我们开展了大规模仿真和真实世界实验（图 3）。

1）仿真基准：我们采用 [1] 中提出的仿真基准，该基准构建于 Drake [70] 之上。我们的基准包含 13 个已见任务和 8 个未见任务（其中 3 个未见任务定义于 [1]，另有 5 个新引入的任务，用于进一步探究泛化能力；参见附录 6）。每个策略在名义条件和 DS 条件下，对每项任务分别进行 50 次滚动评估，并以成功率衡量性能。

13个已见任务属于训练分布，而8个未见任务则旨在探测模型在该分布之外的泛化能力。这些未见任务涵盖了几项挑战：(i) 语义理解（例如，在*PlaceRedFood-IntoContainer*中识别“红色食物”，或在*PlaceFruitIntoContainer*中区分水果和蔬菜）；(ii) 多步操作（例如，在*PutAp-pleAndPearOnPlate*中依次将苹果和梨放在盘子上）；以及 (iii) 组合泛化（例如，根据将物体 A 放在 C 上、将物体 B 放在 D 上的训练示范，泛化到将物体 A 放在 D 上）。这里，“未见任务”指训练数据中未出现的技能，尽管其中涉及的物体和环境确实出现在训练数据中。

为评估模型对外观变化的鲁棒性，我们使用仿真基准的分布偏移条件，这些条件会在光照、环境背景、相机参数、物体和桌面纹理以及颜色方面引入相对于训练分布的变化。标称条件在这些因素上与训练数据保持一致。

2) 真实世界评估：我们在双机械臂Franka机器人平台上，于以下三种设置中评估策略。策略部署的细节见附录4。

语言跟随。 为评估模型遵循自然语言指令的能力，我们设计了一组语言引导的拾取与放置实验，用于评估三种不同场景：**(i) 已见物体：** 此设置中的物体均出现在训练数据中。指令遵循一个简单的模板：“拿起 [物体 A] 并将其放在 [物体 B]里面/上面/旁边”，其中会明确以名称指代物体。**(ii) 指令泛化：** 此设置通过改述指令，测试模型理解自然语言背后含义的能力。具体而言，策略必须：(1) 理解语义物体类别（例如，“书写工具”指笔）；(2) 根据物体属性识别物体（例如，“带有软毛的手柄”指刷子）；以及 (3) 展现对释义改写的鲁棒性（例如，改变句法和动词选择）。**(iii) 未见物体：** 此设置采用目标机器人的训练

数据中未出现的物体。指令遵循与已见物体设置相同的模板。

在全部三种设置中，我们评估了 15 种空间布局，每种布局包含 6–8 个桌面物体。对于每种布局，我们使用三条针对不同物体的语言指令，因此每种设置共进行 45 次滚动评估。在指令泛化设置中，我们采用与已见物体设置相同的空间布局和目标操作结果，但改变指令的措辞。完整的评估套件共涵盖 49 个已见物体和 52 个未见物体。我们报告平均任务完成百分比作为评估指标（评分标准、实验流程和评分标准质量保证（质量保证）见附录7）。

长时程灵巧操作。为探究协同训练能否促进预训练模型快速适应预训练期间未见过的新型挑战性任务，我们设计了三项长时程灵巧操作任务：*PackItemsIntoStringBag*、*PourIngredientsIntoSoup*和*StoreCleanDishes*。每项任务平均包含 13 个步骤，执行时间为 93 秒。这些任务需要超越简单抓取与放置操作的精细操作技能，例如给瓶子盖上盖子、用锅铲将碗中的食物舀入锅中，或将酒杯倒置插入碗碟架。对于每项任务，我们收集 200 条用于微调的示范数据。每个检查点针对每项任务进行 30 次滚动评估，并报告平均任务完成百分比作为评估指标（评分标准、实验流程和评分标准质量保证见附录7）。

3）统计分析框架：我们采用与近期研究类似的严格统计分析方法 [1], , 通过成对假设检验 [78, 64]和紧凑字母显示法（CLD） [59]进行比较。未共享任何 CLD 字母的协同训练策略，其平均表现存在显著差异，族错误率（FWER）为 5%。我们报告各策略的贝叶斯不确定性估计，并通过叠加在条形图上的小提琴图来可视化后验不确定性。圆点和水平线分别表示经验均值和后验均值。任务完成结果的原始经验分布见附录 5，其中还提供了统计框架的更多细节。

B. 在不同训练阶段纳入不同协同训练模态，会如何影响策略表现?

我们采用第 II-A3 节所述的三种策略，评估每种协同训练数据模态：单阶段协同训练、两阶段仅第一阶段协同训练和完整两阶段协同训练。我们将采用这些策略得到的策略，与使用流匹配目标（M_{CE}=0）仅基于来自 TRI-Ramen 的连续机器人动作训练的基线策略（即无协同训练基线）进行比较。首先在仿真中评估所有策略，结果汇总于图 4。随后，在真实世界语言跟随实验中评估表现有效的模态（图 5）。

标准视觉-语言数据协同训练。(1) 如图 4A 和 5A 所示，使用标准视觉-语言数据进行协同训练，可显著提升模型对分布偏移的鲁棒性、泛化能力、对未见任务的泛化能力以及语言跟随能力，同时分布内性能（已见任务）没有发生统计学显著变化。(2) 使用我们整理的视觉-语言数据对预训练视觉-语言模型进行微调，可增强其在机器人操作任务上的表示能力；从两阶段仅第一阶段协同训练相较于基线的性能提升中可见这一点（图 4A、5A）。(3) 在第二阶段继续使用视觉-语言数据进行协同训练，可进一步提升模型在未见任务上的表现以及语言跟随能力（尤其是涉及未见物体时）。我们推测，持续接触这些数据能让模型保留视觉-语言语料库中丰富且可泛化的知识，而这些知识在机器人数据中并不存在，从而避免灾难性遗忘。

机器人数据的密集语言标注协同训练。(1) 使用脚本标注（图 4B、5B）和基于视觉语言模型的标注（图 4C、5C）对机器人数据进行协同训练，可提升模型应对分布偏移的鲁棒性、对未见任务的泛化能力以及语言跟随能力，同时不会显著改变分布内性能。(2) 由于基于视觉语言模型的标注具有更高的语言多样性，并且对物体与环境交互的描述更加丰富，因此与脚本标注相比，它能更显著地提升模型在未见任务上的表现和语言跟随能力（图 4B-C、5B-C）。(3) 在两阶段协同训练中，在第二阶段加入这些标注并无额外收益，这表明仅在第一阶段使用它们效果最佳（图 4B-C、5B-C）。我们认为，由于这些标注描述的是与机器人动作数据相同的物理轨迹，它们的主要作用是在第一阶段训练期间引导语言-动作对齐，而不是在第二阶段引入新信息。

跨具身机器人数据协同训练。（1）如图4D和5D所示，与跨具身机器人数据协同训练可提高对分布偏移的鲁棒性、对未见任务的泛化以及语言跟随，而分布内性能无统计学显著变化。（2）跨具身机器人数据在仅限于两阶段协同训练的第一阶段时最为有效，带来最大增益，尤其是在未见任务泛化和分布偏移下的鲁棒性方面（图4D）。当在两阶段完全协同训练的第二阶段训练中纳入时，它对语言跟随几乎没有额外收益。我们假设，来自跨具身数据的多样形态和操作策略，对于在第一阶段训练中学习可泛化的视觉与行为表征最为有价值。在第二阶段，模型受益于向目标具身的专门化，此时继续接触其他本体的价值有限。

人类视频协同训练。*A*）潜在动作： 在单阶段协同训练中，模型同时学习 TRI-Ramen 的连续动作，以及从所有视频数据（TRI-Ramen、OXE-Ramen 和人类视频）中提取的离散潜在动作标记。对于双阶段方法，模型在第一阶段从所有视频数据中学习潜在动作。在两阶段完全协同训练中，模型在第二阶段同时从所有视频数据中学习 TRI-Ramen 的连续动作和潜在动作。

单阶段协同训练结合潜在动作并未带来任何改进——

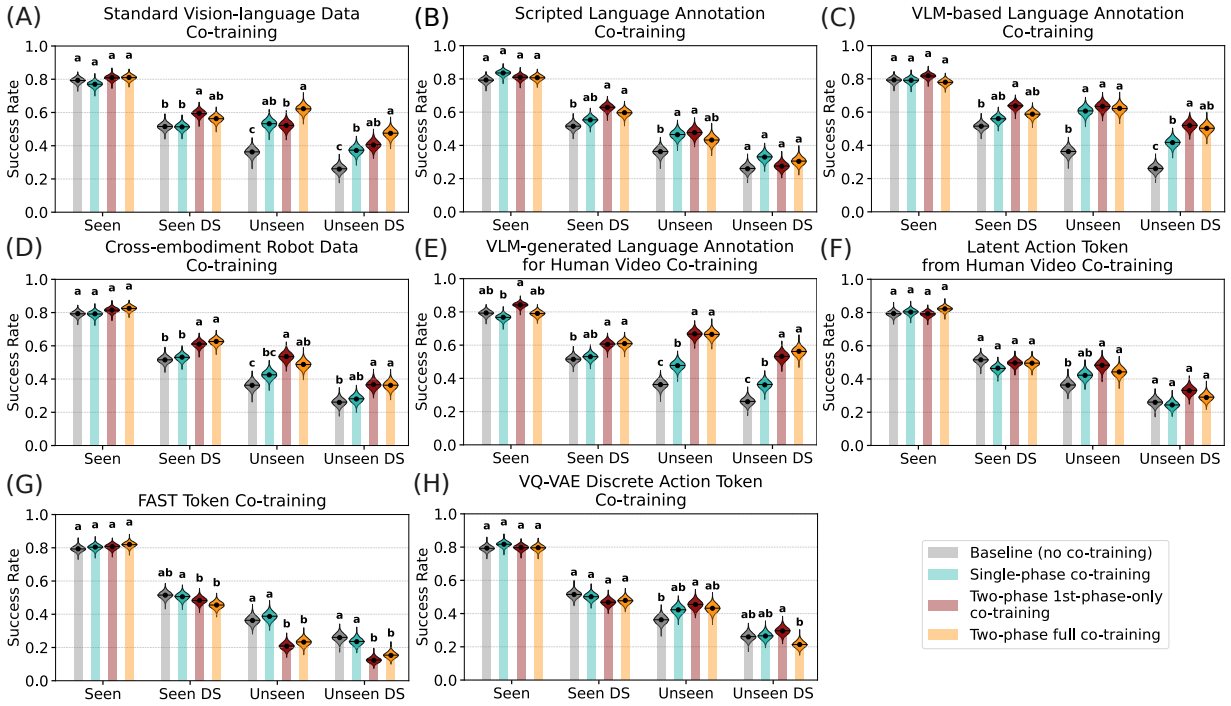


图 4. **协同训练数据与策略的仿真消融实验**。无协同训练基线与在连续训练阶段使用单一数据模态进行协同训练的策略之比较。评估这些策略在常规条件和分布偏移条件下对已见任务和未见任务的表现（A–H 表示数据模态）。

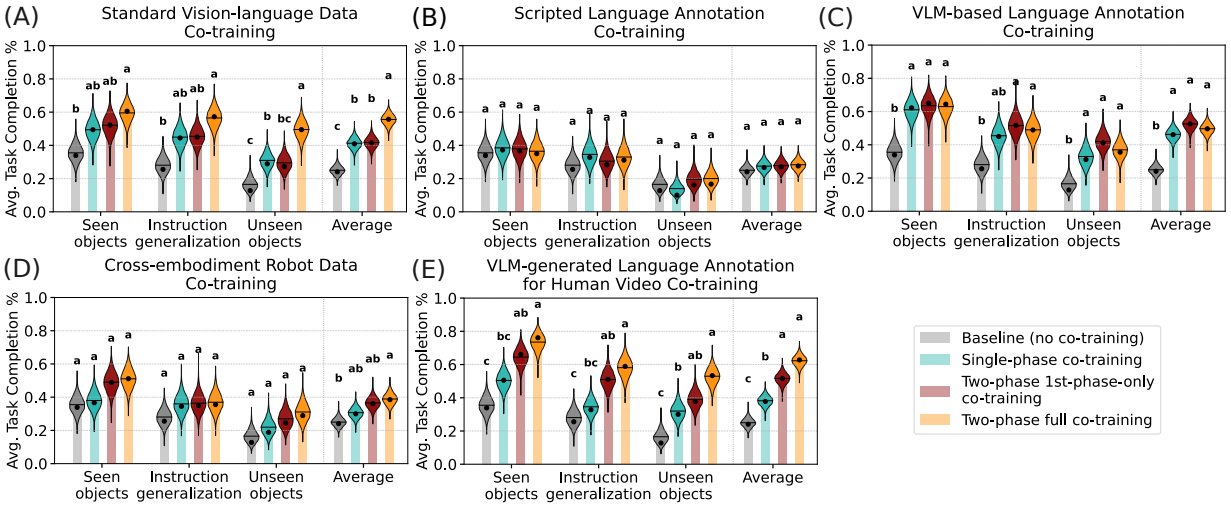


图 5. **协同训练数据与策略的真实场景消融实验**。无协同训练基线与在不同训练阶段使用单一数据模态进行协同训练的策略的表现；评估内容包括使用已见物体进行语言跟随、指令泛化以及未见物体（A–E 表示数据模态）。

相较于基线（图 4F），而单阶段协同训练 结合其他有效模态（例如标准视觉语言数据）则始终能提升性能。另一方面，对于两阶段协同训练：(1) 潜在动作第一阶段训练能提升模型在未见任务上的性能；(2) 在第二阶段纳入潜在动作并无额外益处。这些结果表明，潜在动作第一阶段训练在两阶段之外带来的益处，可能源于计算量增加，而非真正的知识迁移。为探究这一点，我们设计了一个数据和计算资源充足的设置，对比两种策略：(a) 跨具身协同训练基线：第一阶段使用所有机器人数据（TRI-Ramen 和 OXE-Ramen）进行训练，第二阶段则使用 TRI-Ramen 进行训练；该基线等同于使用跨具身数据、但不进行潜在动作协同训练的仅第一阶段两阶段协同训练。(b) 潜在动作三阶段协同训练：(i) 使用所有视频数据中的潜在动作进行训练；(ii) 使用所有机器人数据训练连续动作；(iii) 使用 TRI-Ramen 进行训练。值得注意的是，这两种方法的唯一区别在于，后者增加了一个使用所有视频数据进行初始训练的阶段。如图 6 所示，在这种数据和计算资源丰富的设置下，新增的初始潜在动作训练阶段并无益处，

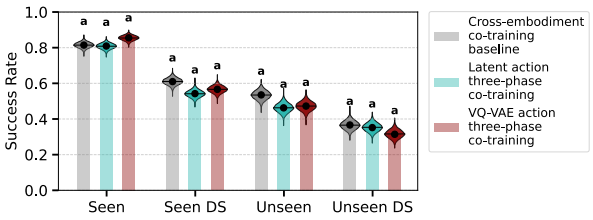


图 6. **在数据和算力充足的设置下，对潜在动作和 VQ-VAE 动作协同训练进行消融**。 仿真结果表明，与跨具身基线相比，这些模态并未带来性能提升。

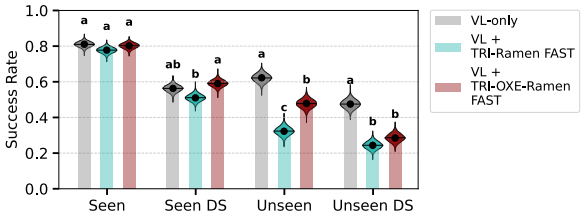


图7. **在机器人数据和标准视觉-语言（VL）数据的广泛混合数据上，对 FAST 令牌协同训练进行消融**。 与仅使用 VL 数据的协同训练基线相比，FAST 协同训练结合 VL 数据未能提升整体性能，且降低了模型在未见任务上的泛化能力。

资源丰富的设置。

鉴于先前研究 [10, 5, 13] 指出，在目标机器人数据规模较小时，潜在动作预训练具有实用价值，我们进一步探究其在不同机器人数据规模下的效果，范围从单任务到完整机器人数据集。图 8 显示，潜在动作第一阶段训练提升了目标机器人数据规模较小时的性能，但随着微调机器人数据量增加，这些收益逐渐减弱。

B) 视觉语言模型生成的标注： (1) 如图 4E 和 5E 所示，针对人类视频使用视觉语言模型生成的标注进行协同训练，可提升模型对 DS 的鲁棒性、对未见任务的泛化能力以及语言跟随能力，同时不会对分布内性能造成统计显著的影响。(2) 在两阶段协同训练中（图 5E），第二阶段继续纳入这些标注，可进一步提升语言跟随性能，尤其是在未见物体上。我们认为，这一优势源于人类视频中丰富多样的动作、物体和环境，而 TRI-Ramen 中缺乏这些内容。在第二阶段进行联合训练，使模型能够保留更广泛的世界知识，而不是逐渐局限于目标机器人数据的分布。

离散机器人动作令牌协同训练。 *A) FAST* 令牌：我们的结果表明，FAST 令牌协同训练未能在所有维度上提升性能，且降低了模型在未见任务上的泛化能力（图4G）。先前研究 [33, 22]表明，在大量混合机器人数据和标准视觉-语言数据上进行预训练时，FAST 令牌协同训练能够提升性能。为检验这一说法，我们比较了三种方法：(a) 视觉-语言 + TRI-OXE-Ramen FAST：使用 FAST 令牌在所有机器人数据上进行第一阶段训练，同时使用视觉-语言数据。(b) 视觉-语言+ TRI-Ramen FAST：仅使用 TRI-Ramen 数据进行第一阶段训练，使用 FAST 令牌，同时使用视觉-语言数据。(c) 仅视觉-语言：仅使用视觉-语言数据进行第一阶段训练。所有方法均采用相同的第二阶段：通过流匹配在 TRI-Ramen 上进行连续动作学习，同时使用标准视觉-语言数据进行协同训练。

如图 7 所示，FAST 令牌协同训练未能提升整体性能，反而降低了在未见任务上的泛化能力。然而，在第一阶段纳入 OXE-Ramen 的效果显著优于仅使用 TRI-Ramen 进行训练，这表明，当扩展到规模大得多的机器人数据集时，FAST 协同训练可能会带来益处，尽管在当前数据规模下仍然无效。我们认为，这源于 FAST 令牌近乎无损的动作表征特性：使用 FAST 令牌进行协同训练，可能会使视觉语言模型骨干网络偏向于学习精确的动作映射，而非可泛化特征。

B) VQ-VAE 离散动作令牌：（1）VQ-VAE 离散动作令牌协同训练在未见任务上带来了些许提升，但在 DS 下的性能略有下降（图 4H）。（2）在两阶段协同训练的第二阶段引入 VQ-VAE 令牌不会带来额外收益。

与潜在动作协同训练类似，我们通过将跨具身协同训练基线与一种 VQ-VAE 离散动作三阶段协同训练策略进行比较，考察其在数据与算力充足的设置下的有效性。该策略在第一阶段先使用所有机器人数据学习 VQ-VAE 离散动作令牌。如图 6 所示，在这一设置下，VQ-VAE 协同训练并未带来任何提升。

总结。(1) 通过多样化的视觉-语言数据和跨具身机器人数据进行协同训练，能够显著提升模型对分布偏移、未见任务的泛化能力以及语言跟随能力。值得注意的是，由于信息丰富，使用标准视觉-语言数据和人类视频的语言标注进行协同训练，对第一阶段和第二阶段协同训练均有益；而机器人轨迹的语言标注和跨具身数据则主要在两阶段协同训练的第一阶段发挥作用。(2) 在所有有效的协同训练数据模态中，标准视觉-语言数据、基于视觉语言模型的机器人数据语言标注，以及人类视频的语言标注最为有益。这三种特定模态均属于多样化的视觉-语言数据，表明增强视觉语言模型骨干网络的视觉-语言理解能力，有助于提升机器人策略。(3) 离散动作词元（包括从视频中提取的潜在动作、FAST 令牌，以及通过 VQ-VAE 学习的动作标记）协同训练并未在我们的实验中带来具有统计显著性的性能提升。具体而言，使用 FAST 令牌进行协同训练会降低泛化能力；而视频中的潜在动作仅在目标机器人数据规模较小时带来收益，且随着机器人数据占比增加，收益逐渐减弱。(4) 在所考察的所有协同训练模态中，我们均未观察到对分布内性能具有统计显著性的影响。

图 9 比较了每种有用协同训练模态的最佳训练策略。具体而言，对于标准视觉-语言以及为人类视频生成的基于视觉语言模型的语言标注，最佳模型对应于两阶段完全协同训练。对于脚本控制与基于视觉语言模型的语言标注以及跨具身机器人数据，最佳模型对应于两阶段仅第一阶段协同训练。使用标准视觉-语言数据、基于视觉语言模型

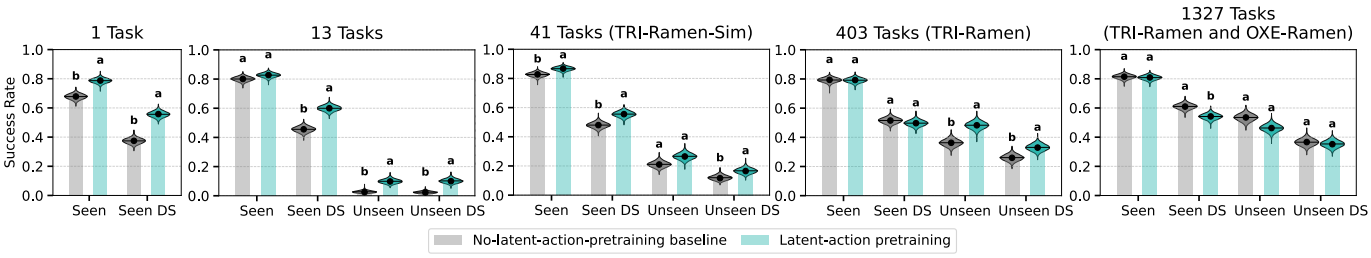


图 8。**潜在动作协同训练在增加数据量时的消融实验**。在多阶段训练中，第一阶段引入潜在动作可在目标机器人数据量较少时带来性能提升；随着微调机器人数据量增加，其收益逐渐减小。

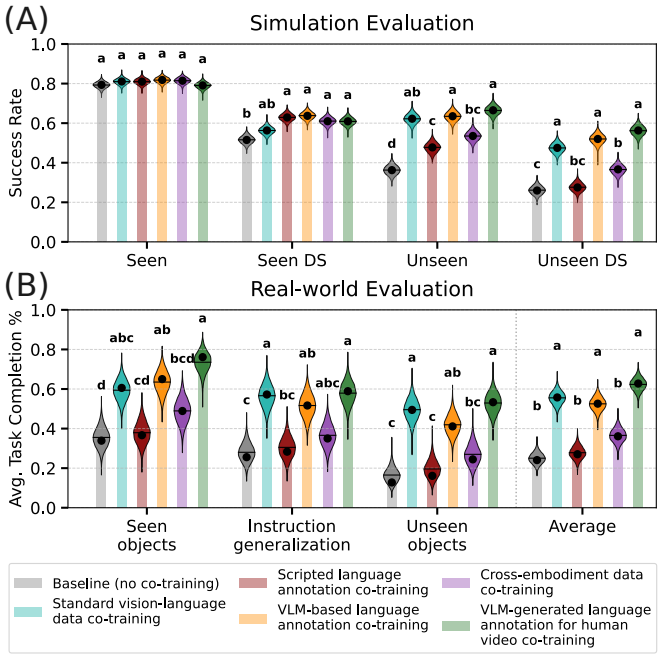


图 9。**采用针对有效协同训练数据模态的最佳协同训练策略训练的策略性能**。A）在名义条件和分布偏移条件下，已见任务和未见任务的仿真结果。B）真实世界中，在已见物体、指令泛化和未见物体方面的语言跟随性能。多样化的视觉-语言数据和跨具身机器人数据显著增强了模型对分布偏移、未见任务和语言跟随的泛化能力，同时不影响分布内性能。

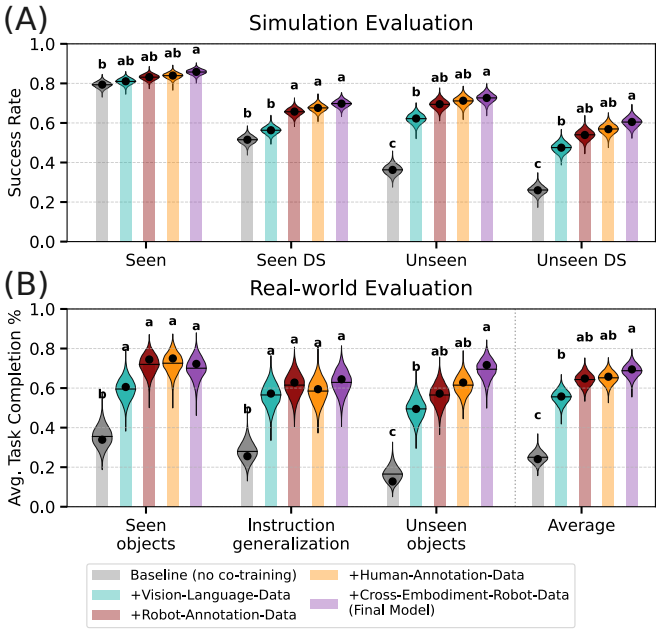


图 10。**以加性方式组合有效数据模态进行协同训练的策略性能**。A）在名义条件和分布偏移条件下，已见任务和未见任务的仿真结果。B）真实世界评估：在已见物体、指令泛化和未见物体设置下的语言跟随性能。组合有效的协同训练数据模态可使策略性能获得累积性提升。

的标注（针对机器人数据）以及人类视频的标注，最为显著地提升了在未见任务和语言跟随方面的性能。此外，得益于机器人演示数据中所缺乏的丰富信息，与针对机器人数据的基于视觉语言模型的标注相比，使用标准视觉-语言数据以及人类视频的标注进行协同训练，能更有效地使模型识别未见物体。

C. 结合有效的协同训练模态能否带来累积的性能提升？

在确定了有效的协同训练数据模态及其最佳训练策略后，我们进一步探究组合这些模态能否带来累积收益。我们进行了一项消融研究，逐步加入每种有效数据源（训练细节见表 S1）：(1) 不进行协同训练的基线，(2)+视觉语言数据：仅进行标准视觉-语言数据协同训练，(3) +机器人标注数据：为机器人数据添加密集的语言标注（包括脚本控制生成和基于视觉语言模型生成的标注），(4)+人类视频标注数据：为人类视频添加由视觉语言模型生成的标注，(5) +跨具身机器人数据（即最终模型）：添加跨具身机器人数据。

如图 10 所示，结合有效的协同训练模态，在所有评估维度上都能持续提升累积性能。我们的最终模型在所有设置下均表现出色，在仿真中的未见任务上取得了 72.6% 的经验成功率（较基线提升 36.4% ），在真实世界的语言跟随任务中，平均任务完成率达到 69.4%（较基线提升 45.3%[1] ）。

[1]改进率根据经验均值计算。

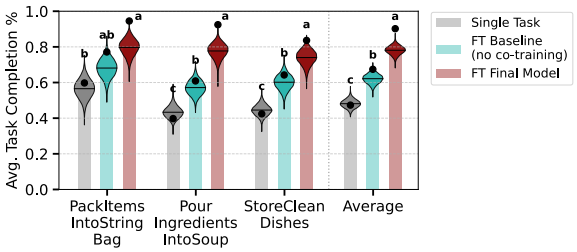


图 11. **通过微调适应未见的长时程灵巧任务**。对从头训练的单任务策略、仅使用机器人数据进行预训练的微调（FT）基线，以及使用所有有效协同训练模态进行预训练的微调最终模型进行真实世界评估。微调最终模型取得了更高的平均任务完成率，凸显了协同训练在学习可迁移的高质量表示方面的作用。

D. 协同训练能否提升学习到的表征质量，从而实现对未见的长时程灵巧任务的快速适应？

先前研究 [1, 6, 33] 已证明，高质量预训练能够使策略通过微调快速适应下游未见任务。我们通过在一组未见的长时程灵巧操作任务上进行微调（FT），研究协同训练是否能够提升所学表示的质量。我们比较了三种方法：（1）微调最终模型：对第 III-C 节中的最终模型进行微调（该模型使用所有有效协同训练模态进行训练）；（2）微调基线：对无协同训练基线模型进行微调（该模型仅使用 TRI-Ramen 进行训练）；（3）单任务：不进行预训练的单任务策略。

如图 11 所示，得益于我们精心整理的协同训练数据和协同训练策略，我们的最终模型通过微调迅速习得新技能，仅需 200 条示范数据，平均任务完成率便达到 90.2 %——比微调基线高 22.8%，比单任务策略高 42.9%。这表明，有效的协同训练能够显著提升表示质量，从而使模型在下游任务中能够学习更精细的动作。我们观察到，微调基线和单任务策略的失败主要源于操作精度不足：它们在*PackItemsIntoStringBag*任务中经常无法将盖子对准并固定在瓶子上，在*PourIngredientsIntoSoup*任务中无法将锅铲准确对准以便抓取，并且难以抓取透明杯子*StoreCleanDishes*。相比之下，微调最终模型始终能够高精度地执行这些精细操作。

E. 有效的协同训练模态如何塑造视觉语言模型骨干网络？

为研究协同训练期间视觉语言模型骨干网络如何变化，我们使用 VLMEvalKit [23]，在一组标准视觉语言基准测试上，对从训练后的 VLA 策略中提取的视觉语言模型进行评测。我们评估了使用各个有效协同训练模态训练的策略，以及将所有有效模态相加结合后训练的策略，评测涵盖三个互补维度：语义理解与推理（MMBench、MME、SeedBench）、空间推理（RealWorldQA、GQA、SpatialEval），以及规划与长期推理（LEGO）。此外，我们还报告了 PaliGemma2-PT（用于初始化我们策略的预训练视觉语言模型）和 PaliGemma2-Mix（一种经过进一步微调以增强指令遵循能力的骨干网络版本）的结果。这些基线有助于对比 VLA 训练和协同训练相对于预训练版本及指令微调版本对骨干网络的影响。图 12 展示了各基准测试上的归一化性能，表 S4 则列出了未归一化分数。我们观察到以下趋势：(1) 无协同训练基线在几乎所有基准测试上的表现都很差。与 PaliGemma2-PT 和 PaliGemma2-Mix 相比，其性能显著退化，完全丧失了生成语言的能力，这表明仅使用机器人数据进行训练会削弱视觉语言模型骨干网络从预训练中继承的视觉语言理解能力。(2) 与 PaliGemma2-PT 基线相比，使用标准视觉语言数据进行协同训练，在大多数基准测试上都带来了显著提升，尤其是在空间推理和现实世界问答方面。相比之下，单独使用其他模态并未带来任何提升。(3) 将有效的协同训练模态相加结合后，视觉语言模型骨干网络在各方面均稳定提升，表现优于无协同训练基线和预训练的 PaliGemma2-PT 模型。组合模型在空间、推理和感知基准测试上均取得均衡提升，达到或接近 PaliGemma2-Mix 的表现，表明其多模态表示更加稳健、全面。(4) 无协同训练基线的视觉语言模型基准测试表现有所退化，在机器人任务上的泛化能力也最弱；而使用有效协同训练模态组合进行协同训练的策略，不仅提升了视觉语言模型基准测试分数，在分布偏移下及未见任务上的泛化能力也更强。

F. 显式地以从协同训练数据中学到的思维链（*CoT*）为条件进行动作生成，会如何影响策略性能？

我们考察生成中间思维链轨迹以条件化动作生成，是否比标准训练更有优势。具体而言，我们评估了三种自然产生类似思维链中间内容的协同训练数据类型：（1）机器人数据的脚本标注；（2）机器人数据的基于视觉语言模型的标注；以及（3）视频中的潜在动作。

对于每种协同训练数据类型，我们还训练了三个额外的策略，它们采用不同的思维链条件策略。正如第 II-A3 节所述，这三种方法都遵循相同的第一阶段训练流程，并使用相应的协同训练数据。在第二阶段，所有方法都联合使用连续机器人动作和协同训练数据进行学习。具体而言：（1）仅训练时使用50%思维链：在第一阶段，模型以50%的概率以思维链为条件生成动作。推理时，模型直接生成动作，不使用思维链。（2）推理时使用50%思维链：训练流程与（1）相同。推理时，模型先生成思维链，再以其为条件生成后续动作。（3）100%思维链：在第二阶段，模型始终以思维链为条件生成动作（概率为100%）。推理时，模型生成思维链，并以其为条件生成动作。

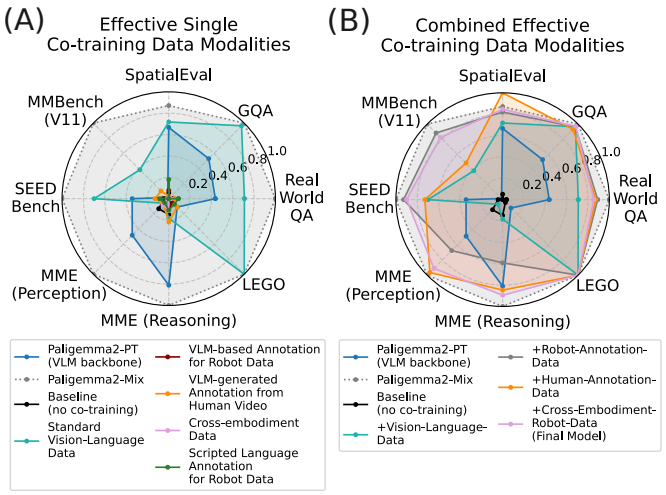


图 12. **采用有效协同训练数据模态训练的策略的视觉语言模型骨干网络基准测试**。A) 将采用各个有效数据模态进行协同训练的视觉语言模型骨干网络的归一化性能，与无协同训练基线、作为我们所有策略骨干网络的预训练 PaliGemma2-PT，以及指令微调模型 PaliGemma2-Mix 进行比较；评估基于标准视觉-语言、空间、推理和感知基准测试。B) 组合有效协同训练模态时，视觉语言模型骨干网络的性能。组合模态能在所有基准测试上带来稳定且累积的提升，这与下游机器人策略泛化能力的提升相呼应。

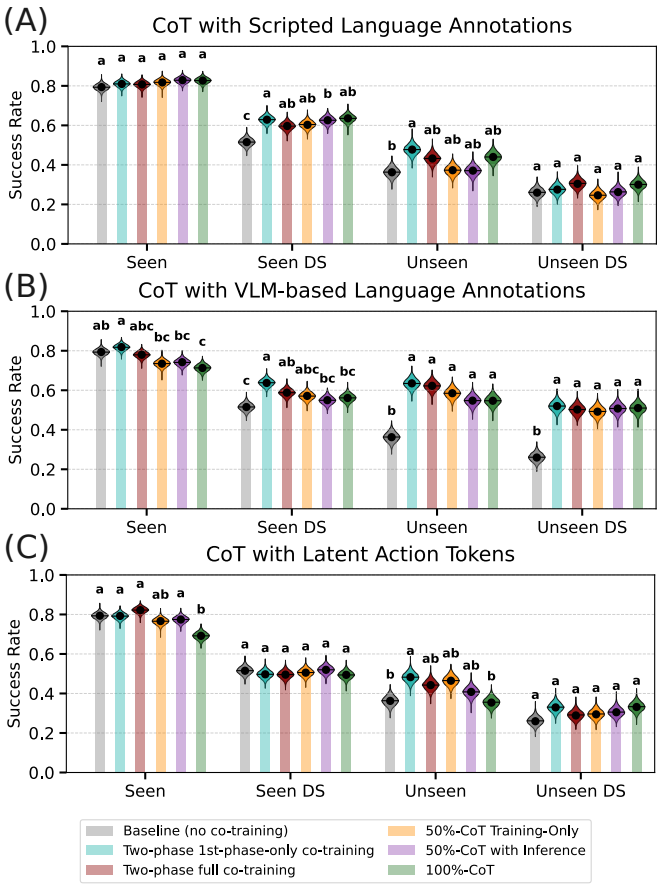


图 13. **思维链条件化动作生成的仿真消融实验**。评估通过以下思维链轨迹显式条件化动作生成而训练的策略：A）机器人数据的脚本语言标注，B）机器人数据的基于视觉语言模型的语言标注，以及 C）视频中的潜在动作标记。以协同训练数据中的思维链轨迹为条件进行动作生成，并未提升我们的仿真基准表现。

我们将这些显式思维链策略与第 II-A3 节中讨论的隐式两阶段协同训练方法，以及未经过协同训练的基线策略进行比较。如图 13 所示，尽管显式思维链条件化在某些设置下优于基线（例如，在未见任务上使用基于视觉语言模型的机器人数据标注进行协同训练），但在所有设置下，它始终未能优于隐式两阶段协同训练方法；使用基于视觉语言模型的标注和潜在动作作为思维链来源时，性能还出现了明显下降。这种改进有限的情况可归因于评估任务的性质。先前研究 [44, 93, 67] 中，显式思维链生成通常在需要多步规划或复杂语义推理的任务上进行评估。相比之下，我们的仿真基准侧重于目标明确且能立即获得视觉反馈的操作任务，其中从观测到动作的映射相对直接。在此类设置下，协同训练过程中习得的隐式推理似乎已足够，因此显式思维链生成显得多余。

观察到的性能下降，尤其是在使用基于视觉语言模型的标注和潜在动作时，可能源于两个因素。首先，这些协同训练数据来源可能存在固有的不准确性——视觉语言模型生成的标注可能会误解视觉场景或动作，而潜在动作模型可能捕捉到背景变化等无关的视觉变化，而非真正的动作语义。其次，与脚本标注相比，这两个来源会生成内容更丰富、复杂度更高的思维链内容。当

显式地将动作生成条件化于此类思维链时，任何生成的思维链中的误差或不精确性都会直接传播到后续的动作预测，从而加剧初始不准确性，并导致操作行为不够精确。

IV. 相关工作

A. 大型行为模型

开发能够感知、理解并在物理世界中采取行动的通用策略，是机器人学的一项核心目标。早期的机器人学习系统依赖于在有限的机器人示范数据上训练的特定任务策略 [42, 24, 83], , 这限制了它们超越狭窄训练分布进行泛化的能力。相比之下，大型行为模型在模型容量和数据集规模方面扩大了模仿学习的规模，并在灵巧任务上展现出令人瞩目的表现 [1]。大型行为模型中的一个重要类别是视觉-语言-动作模型 [37, 100, 21, 6, 33, 22, 68, 67, 5, 80, 79], , 它们集成了预训练视觉-语言模型 [4, 65, 47, 73]骨干网络。典型例子包括像 [6, 33, 63] 那样将视觉语言模型与动作头配对的模型，以及完全自回归的模型 [39, 58, 26]。尽管不断扩大数据规模，视觉-语言-动作模型对 [1, 94, 97] 的泛化能力仍然有限，

与视觉语言模型等非具身基础模型相比，它们对新物体、新环境和新指令的泛化能力较弱。这种泛化差距主要源于训练数据可用量的巨大差异 [25, 55]：机器人数据集的规模仍比视觉语言模型所使用的互联网规模文本与图像语料小几个数量级。

B. 机器人学习的协同训练

为弥合有限的机器人数据与互联网规模多模态资源之间的差距，众多研究采用了结合多种数据模态的协同训练。

公开视觉语言数据集 [14, 71, 89, 19, 91, 96]，富含常识知识，并且天然适配VLA架构，因此被广泛用于协同训练机器人策略 [33, 99, 98, 39]。除公开数据集外，近期研究 [44, 84, 61, 93] 直接基于机器人轨迹构建具身推理视觉语言数据集，其中融入了丰富的规划和空间信息。这些工作表明，与多样化的视觉语言数据进行协同训练能够增强泛化能力 [33, 44]，并改善VLM针对操作任务所学习的表征 [11, 84] 。

一些研究探索了使用跨具身机器人数据进行协同训练。部分工作 [2, 20, 85, 86, 7] 直接在多种机器人具身数据上训练单一策略，使统一模型能够适应不同的形态。[6, 33, 22, 50, 48] 在预训练期间引入跨具身数据，以学习具有泛化能力且与具身无关的表征，随后将其适配到特定的目标机器人。 [69, 1, 37, 38, 62]利用Open X-Embodiment 数据集 [55], ，这是对来自众多机器人平台的大规模示范数据的汇总，旨在提升鲁棒性和泛化能力。

与通常需要通过遥操作来采集的机器人数据相比，人类视频提供了更具可扩展性的数据源。许多研究 [41, 52, 36, 40, 51] 会从视频中明确提取动作标签（例如手部姿态），用于策略协同训练；然而，获取准确标签往往需要额外的传感模态，例如虚拟现实设备 [60, 90] 或可穿戴外骨骼 [81, 66]。另一些研究 [9, 88, 13,8, 12, 5]则探索潜在动作表征，采用 VQ-VAE [72], 等方法从视频帧中提取离散动作词元，作为可跨不同本体的统一表示，以编码运动信息。然而，这些方法仅在目标机器人数据量较少的情况下得到验证。

将连续机器人动作离散化为令牌，使策略能够将动作生成视为序列建模问题。除了朴素的逐维分箱方法 [37, 100]，之外， [58, 76] 先进方法还采用基于频率的技术（例如 FAST 分词器）或向量量化（例如 VQ-VAE）来压缩动作空间。值得注意的是，依赖离散令牌进行低层控制通常会导致精度有限 [22] 且推理速度较慢 [58, 39]。为缓解这一问题，近期方法 [33, 22, 35, 93] 仅将这些令牌用于预训练或协同训练目标，同时保留动作头以进行连续动作生成。这种策略已被证明能够提升样本效率和泛化能力。

C. 机器人控制的思维链

受思维链在处理复杂任务时为语言模型带来显著益处的启发 [29, 34, 77,95], ，近期研究 [92, 33, 11, 43] 开始探索将思维链应用于机器人控制。具体而言，这些方法会先生成中间内容，再以此为条件进行动作生成。中间内容可以是语言性的，例如子任务分解或视觉定位信息（如物体位置） [44, 93, 67, 30, 87], ；也可以以控制为中心，例如末端执行器的运动 [39, 3] 和潜在动作 [10, 13]。尽管已有研究表明，思维链有益于长时程任务或需要复杂推理的任务 [44, 93, 67], ，但目前缺乏实证证据来比较：（i）明确以预测的思维链轨迹为条件的策略；与（ii）仅将相同的思维链内容用作辅助协同训练目标。这方面的证据对于目标明确且定义清晰的操作任务而言尤为不足。

V. 讨论、局限性与未来工作

我们开展了一项大规模实证研究，系统剖析了多样化协同训练数据和策略对大型行为模型性能的影响。研究结果表明，使用视觉语言数据和跨具身机器人数据进行协同训练，能够显著提升模型对分布偏移、未见任务的泛化能力以及语言跟随能力，而离散动作令牌变体则未带来统计上显著的收益。此外，我们还表明，结合有效模态能够累积提升性能，并通过微调实现对灵巧型长时程任务的快速适应。

值得注意的是，在所有有用的协同训练模态中，多样化的视觉语言数据（包括标准数据集，以及对机器人与人类视频的丰富标注）带来了最显著的改进。这一观察与良好调节器定理 [17], 相呼应，该定理指出，系统必须包含一个关于其运行环境的内部模型（隐式或显式），才能有效地对其进行调节。在我们的设定中，强大的基础模型（例如视觉语言模型）恰好提供了此类内部模型，它们具备对物理世界丰富的语义与空间理解。我们的结果表明，朝着真正通用的机器人策略迈进，与这些基础模型的进步有着内在联系。具体而言，视觉语言模型基准测试结果印证了这一解释：与有效数据模态进行协同训练不仅能提升下游机器人性能，还能保留视觉语言模型骨干网络自身的视觉语言推理、空间理解和感知能力。尽管我们评估的模型都呈现出这一模式，但仍需进一步研究，以更严格地刻画骨干网络的视觉语言理解与策略泛化之间的关系。有趣的是，我们发现，对于具有明确即时目标的操作任务，显式地以从协同训练数据中学习到的思维链为动作生成提供条件并无益处，而这些任务具有

清晰的即时目标，这表明，在此类设定中，协同训练期间学到的隐式推理已足够。

尽管我们的研究提供了颇有前景的见解，但仍有若干局限性值得指出。首先，虽然我们考察了多种视觉-语言数据来源，但并未按照任务分类体系（例如视觉问答、图像描述、目标检测、空间推理）系统分析其影响。了解不同视觉-语言任务类别如何影响特定的策略能力，将有助于更有针对性、且样本效率更高的数据整理。其次，我们仅通过潜在动作和语言标注，探索了人类视频的粗粒度表征。随着手部姿态估计技术的进步以及灵巧机器人手的持续发展，从人类视频中显式提取精细的灵巧动作，或将成为一种有价值的协同训练信号。第三，我们对思维链的探索仅限于协同训练数据中自然产生的形式，主要是缺乏高层级规划或复杂推理的低层级动作抽象。未来的工作可以探索更丰富的思维链形式，例如历史与反思轨迹，或针对需要复杂决策的任务制定层级计划。最后，我们的研究只关注模仿学习；在其他学习范式中探索协同训练，例如世界建模或强化学习，仍是开发可扩展通用策略的一项开放课题。

致谢

感谢 Muhammad Zubair Irshad、Vitor Campagnolo Guizilini、Sedrick Keh 和 Swati Gupta 审阅稿件；感谢 Kosei Tanada、Takuto Ito、Phoebe Horgan、Gordon Richardson 和 Aykut Onol 负责评估质量保证；感谢 Sam Shereda、Mariah Smith-Jones 和 Matthew Tran 负责评估和数据收集；感谢 Alejandro Castro 和 Joseph Masterjohn 提供仿真支持，并感谢 TRI 的 PROPS 团队协助搭建实验环境。

参考文献

[1]Jose Barreiros、Andrew Beaulieu、Aditya Bhat、Rick Cory、Eric Cousineau、Hongkai Dai、Ching-Hsin Fang、Kunimatsu Hashimoto、Muhammad Zubair Irshad、Masha Itkina 等人。对用于多任务灵巧操作的大型行为模型进行细致考察。*arXiv* 预印本 *arXiv:2507.05331*，2025年。[2] Erik Bauer、Elvis Nava 和 Robert K Katzschmann。用于跨具身操作的潜在动作扩散。*arXiv* 预印本 *arXiv:2506.14608*，2025年。[3]Suneel Belkhale、Tianli Ding、Ted Xiao、Pierre Sermanet、Quon Vuong、Jonathan Tompson、Yevgen Chebotar、Debidatta Dwibedi 和 Dorsa Sadigh。RT-H：利用语言构建动作层级。*arXiv* 预印本 *arXiv:2403.01823*，2024年。[4]Lucas Beyer、Andreas Steiner、Andr´e Susano Pinto、Alexander Kolesnikov、Xiao Wang、Daniel Salz、Maxim Neumann、IbrahimAlabdulmohsin、MichaelTschannen、Emanuele Bugliarello 等人。PaliGemma：一种多功能用于迁移的 3b 视觉语言模型。*arXiv* 预印本 *arXiv:2407.07726*，2024年。[5] Johan Bjorck、Fernando Casta˜neda、Nikita Cherniadev、Xingye Da、Runyu Ding、Linxi Fan、Yu Fang、Dieter Fox、Fengyuan Hu、Spencer Huang 等。Gr00t n1：面向通用人形机器人的开放基础模型。*arXiv* 预印本 *arXiv:2503.14734*，2025年。[6] Kevin Black、Noah Brown、Danny Driess、Adnan Esmail、Michael Equi、Chelsea Finn、Niccolo Fusai、Lachy Groom、Karol Hausman、Brian Ichter、Szy mon Jakubczak、Tim Jones、Liyiming Ke、Sergey Levine、Adrian Li-Bell、Mohith Mothukuri、Suraj Nair、Karl Pertsch、Lucy Xiaoyang Shi、James Tanner、Quan Vuong、Anna Walling、Haohuan Wang 和 Ury Zhilin- sky。π_0：用于通用机器人控制的视觉-语言-动作流模型，2024年。URL https://arxiv.org/abs/ 2410.24164。[7] Nico Bohlinger、Grzegorz Czechmanowski、Maciej Krupka、Piotr Kicki、Krzysztof Walas、Jan Peters 和 Davide Tateo。一个策略运行所有机器人：一种面向多具身形态运动的端到端学习方法。*arXiv* 预印本 *arXiv:2409.06366*，2024年。[8] Qingwen Bu、Jisong Cai、Li Chen、Xiuqi Cui、Yan Ding、Siyuan Feng、Shenyuan Gao、Xindong He、Xuan Hu、Xu Huang 等。Agibot world colosseo：面向可扩展且智能的具身系统的大规模操作平台。*arXiv* 预印本 *arXiv:2503.06669*，2025年。[9] Qingwen Bu、Yanting Yang、Jisong Cai、Shenyuan Gao、Guanghui Ren、Maoqing Yao、Ping Luo 和 Hongyang Li。学习在任何地方执行任务：以任务为中心的潜在动作。*arXiv* 预印本 *arXiv:2502.14420*，2025年。[10] Qingwen Bu、Yanting Yang、Jisong Cai、Shenyuan Gao、Guanghui Ren、Maoqing Yao、Ping Luo 和 Hongyang Li。Univla：学习在任何地方执行任务：以任务为中心的潜在动作。*arXiv* 预印本 *arXiv:2505.06111*，2025年。[11]William Chen、Suneel Belkhale、Suvir Mirchandani、Oier Mees、Danny Driess、Karl Pertsch 和 Sergey Levine。高效具身推理的训练策略。*arXiv* 预印本 *arXiv:2505.08243*，2025年。[12] Xiaoyu Chen、Junliang Guo、Tianyu He、Chuheng Zhang、Pushi Zhang、Derek Cathera Yang、Li Zhao 和 Jiang Bian。Igor：图像目标表征是具身人工智能基础模型的原子控制单元。*arXiv* 预印本 *arXiv:2411.00785*，2024年。[13] Xiaoyu Chen、Hangxing Wei、Pushi Zhang、Chuheng Zhang、Kaixin Wang、Yanjiang Guo、Rushuai Yang、Yucen Wang、Xinquan Xiao、Li Zhao 等。Villa-x：增强视觉-语言-动作模型中的潜在动作建模。*arXiv* 预印本 *arXiv:2507.23682*，2025年。[14]Xinlei Chen、Hao Fang、Tsung-Yi Lin、Ramakr- ishna Vedantam、Saurabh Gupta、Piotr Doll´ar 和 C Lawrence Zitnick。Microsoft coco captions：数据收集与评估服务器。*arXiv* 预印本 *arXiv:1504.00325*，2015年。[15] Cheng Chi、Zhenjia Xu、Chuer Pan、Eric Cousineau、

Benjamin Burchfiel、Siyuan Feng、Russ Tedrake 和 Shuran Song。通用操作接口：无需野外机器人即可进行真实环境中的机器人教学。*arXiv* 预印本 *arXiv:2402.10329*，2024年。[16] Gheorghe Comanici、Eric Bieber、Mike Schaekermann、Ice Pasupat、Noveen Sachdeva、Inderjit Dhillon、Marcel Blistein、Ori Ram、Dan Zhang、Evan Rosen 等。Gem- ini 2.5：凭借高级推理、多模态、长上下文和新一代智能体能力推动前沿。*arXiv* 预印本 *arXiv:2507.06261*，2025年。[17]Roger C Conant 和 W Ross Ashby。系统的每个优秀调节器都必须是该系统的模型。《国际系统科学杂志》，1(2):89–97，1970年。[18]Dima Damen、Hazel Doughty、Giovanni Maria Farinella、Sanja Fidler、Antonino Furnari、Evangelos Kazakos、Davide Moltisanti、Jonathan Munro、Toby Perrett、Will Price 等。EPIC-KITCHENS 数据集：收集、挑战与基线。《*IEEE*模式分析与机器智能汇刊》，43(11): 4125–4141，2020年。[19]Matt Deitke、Christopher Clark、Sangho Lee、Rohun Tripathi、Yue Yang、Jae Sung Park、Mohammadreza Salehi、Niklas Muennighoff、Kyle Lo、Luca Soldaini 等。Molmo 和 PixMo：面向最先进多模态模型的开放权重与开放数据。*arXiv* 电子论文，页码 arXiv–2409，2024年。[20] Ria Doshi、Homer Walke、Oier Mees、Sudeep Dasari 和 Sergey Levine。扩展跨具身学习：一个策略用于操作、导航、运动和航空。*arXiv* 预印本 *arXiv:2408.11812*，2024年。[21]Danny Driess、Fei Xia、Mehdi SM Sajjadi、Corey Lynch、Aakanksha Chowdhery、Ayzaan Wahid、Jonathan Tompson、Quan Vuong、Tianhe Yu、Wenlong Huang 等。PaLM-E：具身多模态语言模型。*ICML23*：《第*40*届国际机器学习会议论文集》，2023年。[22] Danny Driess、Jost Tobias Springenberg、Brian Ichter、Lili Yu、Adrian Li-Bell、Karl Pertsch、Allen Z Ren、Homer Walke、Quan Vuong、Lucy Xiaoyang Shi 等。知识隔离型视觉-语言-动作模型：训练快、运行快、泛化更佳。*arXiv* 预印本 *arXiv:2505.23705*，2025年。[23] Haodong Duan、Junming Yang、Yuxuan Qiao、Xinyu Fang、Lin Chen、Yuan Liu、Xiaoyi Dong、Yuhang Zang、Pan Zhang、Jiaqi Wang 等。VLMEvalKit：用于评估大型多模态模型的开源工具包。载于 第*32*届 *ACM* 国际多媒体会议论文集，页码 11198–11201，2024年。[24] Chelsea Finn、Tianhe Yu、Tianhao Zhang、Pieter Abbeel 和 Sergey Levine。通过元学习进行单样本视觉模仿学习。载于 机器人学习会议，页码 357–368。PMLR，2017年。[25]Ken Goldberg。优秀的传统工程可以弥合机器人学中长达 10 万年的“数据鸿沟”，2025年。[26]Ankit Goyal、Hugo Hadfield、Xuning Yang、Valts Blukis 和 Fabio Ramos。Vla-0：无需修改即可构建最先进的视觉-语言-动作模型。*arXiv* 预印本 *arXiv:2510.13054*，2025年。[27]Raghav Goyal、Samira Ebrahimi Kahou、Vincent Michalski、Joanna Materzynska、Susanne Westphal、He-una Kim、Valentin Haenel、Ingo Fr uend、Peter Yianilos、Moritz Mueller-Freitag 等。《用于学习和评估视觉常识的“something-something”视频数据库》。载于*IEEE*国际计算机视觉会议论文集，第5842–5850页，2017年。[28]Kristen Grauman、Andrew Westbury、Eugene Byrne、Zachary Chavis、Antonino Furnari、Rohit Girdhar、Jackson Hamburger、Hao Jiang、Miao Liu、Xingyu Liu 等。Ego4D：遍及全球的3000小时第一视角视频。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第18995–19012页，2022年。[29] Daya Guo、Dejian Yang、Haowei Zhang、Junxiao Song、Ruoyu Zhang、Runxin Xu、Qihao Zhu、Shirong Ma、Peiyi Wang、Xiao Bi 等。DeepSeek-R1：通过强化学习激励大语言模型的推理能力。*arXiv* 预印本 *arXiv:2501.12948*，2025年。[30]Asher J Hancock、Xindi Wu、Lihan Zha、Olga Russakovsky 和 Anirudha Majumdar。将动作作为语言：将视觉语言模型微调为视觉-语言-动作模型，同时避免灾难性遗忘。*arXiv* 预印本 *arXiv:2509.22195*，2025年。[31] Ryan Hoque、Peide Huang、David J Yoon、Mouli Sivapurapu 和 Jian Zhang。EgoDex：从大规模自我中心视频中学习灵巧操作。*arXiv* 预印本 *arXiv:2505.11709*，2025年。[32]Aaron Hurst、Adam Lerer、Adam P Goucher、Adam Perelman、Aditya Ramesh、Aidan Clark、AJ Ostrow、Akila Welihinda、Alan Hayes、Alec Radford 等。GPT-4o系统卡。*arXiv* 预印本 *arXiv:2410.21276*，2024年。[33] Physical Intelligence、Kevin Black、Noah Brown、James Darpinian、Karan Dhabalia、Danny Driess、Adnan Esmail、Michael Equi、Chelsea Finn、Niccolo Fusai、Manuel Y. Galliker、Dibya Ghosh、Lachy Groom、Karol Hausman、Brian Ichter、Szymon Jakubczak、Tim Jones、Liyiming Ke、Devin LeBlanc、Sergey Levine、Adrian Li-Bell、Mohith Mothukuri、Suraj Nair、Karl Pertsch、Allen Z. Ren、Lucy Xiaoyang Shi、Laura Smith、Jost Tobias Springenberg、Kyle Stachowicz、James Tanner、Quan Vuong、Homer Walke、Anna Walling、Haohuan Wang、Lili Yu 和 Ury Zhilinsky。$\pi_{0.5}$：一种具备开放世界泛化能力的视觉-语言-动作模型，2025年。URL https://arxiv.org/abs/2504.16054。[34] Aaron Jaech、Adam Kalai、Adam Lerer、Adam Richardson、Ahmed El-Kishky、Aiden Low、Alec Helyar、Aleksander Madry、Alex Beutel、Alex Carney 等。OpenAI o1系统卡。*arXiv* 预印本 *arXiv:2412.16720*，2024年。[35] Tao Jiang、Tianyuan Yuan、Yicheng Liu、Chenhao Lu、Jianning Cui、Xiao Liu、Shuiqi Cheng、Jiyang Gao、Huazhe Xu 和 Hang Zhao。Galaxea开放世界数据集与G0双系统视觉-语言-动作模型 . *arXiv*预印本

arXiv:2509.00576，2025年。[36]Simar Kareer、Dhruv Patel、Ryan Punamiya、Pranay Mathur、Shuo Cheng、Chen Wang、Judy Hoffman 和 Danfei Xu。Egomimic：通过第一视角视频扩展模仿学习。收录于*2025*年 *IEEE* 国际机器人与自动化会议（*ICRA*），第 13226–13233 页。IEEE，2025年。[37]Moo Jin Kim、Karl Pertsch、Siddharth Karamcheti、Ted Xiao、Ashwin Balakrishna、Suraj Nair、Rafael Rafailov、Ethan Foster、Grace Lam、Pannag Sanketi 等。Openvla：一种开源视觉-语言-动作模型。*arXiv*预印本 *arXiv:2406.09246*，2024年。[38]Moo Jin Kim、Chelsea Finn 和 Percy Liang。视觉-语言-动作模型微调：优化速度与成功率，2025年。网址 *https://arxiv.org/abs/2502.19645*，2025年。[39] Jason Lee、Jiafei Duan、Haoquan Fang、Yuquan Deng、Shuo Liu、Boyang Li、Bohan Fang、Jieyu Zhang、Yi Ru Wang、Sangho Lee 等。Molmoact：能够进行空间推理的动作推理模型。*arXiv*预印本 *arXiv:2508.07917*，2025年。[40] Marion Lepert、Jiaying Fang 和 Jeannette Bohg。Masquerade：利用数据编辑从野外人类视频中学习。*arXiv* 预印本 *arXiv:2508.09976*，2025年。[41] Marion Lepert、Jiaying Fang 和 Jeannette Bohg。Phantom：仅使用人类视频、无需机器人即可训练机器人。*arXiv*预印本 *arXiv:2503.00779*，2025年。[42] Sergey Levine、Chelsea Finn、Trevor Darrell 和 Pieter Abbeel。深度视觉运动策略的端到端训练。《机器学习研究期刊》，17(39)：1–40，2016年。[43] Yi Li、Yuquan Deng、Jesse Zhang、Joel Jang、Marius Memmel、Raymond Yu、Caelan Reed Garrett、Fabio Ramos、Dieter Fox、Anqi Li 等。Hamster：用于开放世界机器人操作的分层动作模型。*arXiv* 预印本 *arXiv:2502.05485*，2025年。[44] Fanqi Lin、Ruiqian Nai、Yingdong Hu、Jiacheng You、Junming Zhao 和 Yang Gao。Onetwovla：具有自适应推理能力的统一视觉-语言-动作模型。*arXiv*预印本 *arXiv:2505.11917*，2025年。[45]Juyi Lin、Amir Taherin、Arash Akbari、Arman Akbari、Lei Lu、Guangyu Chen、Taskin Padir、Xiaomeng Yang、Weiwei Chen、Yiqian Li、Xue Lin、David Kaeli、Pu Zhao 和 Yanzhi Wang。Vote：通过轨迹集成投票优化视觉-语言-动作。*arXiv*预印本 *arXiv:2507.05116*，2025年。网址：https://arxiv.org/abs/2507.05116。[46] Yaron Lipman、Ricky TQ Chen、Heli Ben-Hamu、Maximilian Nickel 和 Matt Le。用于生成建模的流匹配。*arXiv*预印本 *arXiv:2210.02747*，2022年。[47] Haotian Liu、Chunyuan Li、Yuheng Li 和 Yong Jae Lee。通过视觉指令调优改进基线。在*IEEE/CVF*计算机视觉与模式识别会议论文集，第 26296–26306 页

[48]Jiaming Liu, Hao Chen, Pengju An, Zhuoyang Liu, Renrui Zhang, Chenyang Gu, Xiaoqi Li, Ziyu Guo, Sixiang Chen, Mengzhen Liu 等。Hybridvla：在统一的视觉-语言-动作模型中协作使用扩散与自回归。*arXiv*预印本 *arXiv:2503.10631*，2025年。[49]Qiang Liu。Rectified flow：一种保持边缘分布的最优传输方法。*arXiv*预印本 *arXiv:2209.14577*，2022年。[50] Songming Liu、Lingxuan Wu、Bangguo Li、Hengkai Tan、Huayu Chen、Zhengyi Wang、Ke Xu、Hang Su 和 Jun Zhu。Rdt-1b：用于双手操作的扩散基础模型。*arXiv*预印本 *arXiv:2410.07864*，2024年。[51] Yangcen Liu、Woo Chul Shin、Yunhai Han、Zhenyang Chen、Harish Ravichandar 和 Danfei Xu。Im-mimic：通过映射与插值，从人类视频中进行跨域模仿。*arXiv*预印本 *arXiv:2509.10952*，2025年。[52]Hao Luo、Yicheng Feng、Wanpeng Zhang、Sipeng Zheng、Ye Wang、Haoqi Yuan、Jiazheng Liu、Chaoyi Xu、Qin Jin 和 Zongqing Lu。BEING-H0：基于大规模人类视频的视觉-语言-动作预训练。*arXiv*预印本 *arXiv:2507.15597*，2025年。[53] OpenAI。GPT-5 发布，2025年。网址：https://openai.com/gpt-5/。[54] Maxime Oquab、Timoth´ee Darcet、Th´eo Moutakanni、Huy Vo、Marc Szafraniec、Vasil Khalidov、Pierre Fernandez、Daniel Haziza、Francisco Massa、Alaaeldin El-Nouby 等。DINOv2：无需监督地学习鲁棒视觉特征。*arXiv*预印本 *arXiv:2304.07193*，2023年。[55]Abby O'Neill、Abdul Rehman、Abhiram Maddukuri、Abhishek Gupta、Abhishek Padalkar、Abraham Lee、Acorn Pooley、Agrim Gupta、Ajay Mandlekar、Ajinkya Jain 等。Open X-Embodiment：机器人学习数据集与 RT-X 模型：Open X-Embodiment 协作项目 0。载于*2024 IEEE*国际机器人与自动化会议（*ICRA*），第 6892–6903 页。IEEE，2024年。[56] William Peebles 和 Saining Xie。基于 Transformer 的可扩展扩散模型。载于*IEEE/CVF*国际计算机视觉会议论文集，第 4195–4205 页，2023年。[57] Baoqi Pei、Yifei Huang、Jilan Xu、Guo Chen、Yuping He、Lijin Yang、Yali Wang、Weidi Xie、Yu Qiao、Fei Wu 等。为第一视角视频表征学习建模细粒度手物体动态。*arXiv*预印本 *arXiv:2503.00986*，2025年。[58]Karl Pertsch、Kyle Stachowicz、Brian Ichter、Danny Driess、Suraj Nair、Quan Vuong、Oier Mees、Chelsea Finn 和 Sergey Levine。FAST：面向视觉-语言-动作模型的高效动作词元化。*arXiv*预印本 *arXiv:2501.09747*，2025年。[59]Hans-Peter Piepho。一种用字母表示所有

，2024.

-成对比较

·

期刊

计算与图形统计，13(2):456– 466，2004年。[60] Ri-Zhao Qiu、Shiqi Yang、Xuxin Cheng、Chaitanya Chawla、Jialong Li、Tairan He、Ge Yan、David J Yoon、Ryan Hoque、Lars Paulsen 等。人形机器人策略~人类策略。*arXiv* 预印本 *arXiv:2503.13441*，2025年。[61] Delin Qu、Haoming Song、Qizhi Chen、Zhaoqing Chen、Xianqiang Gao、Xinyi Ye、Qi Lv、Modi Shi、Guanghui Ren、Cheng Ruan 等。Eo-1：面向通用机器人控制的交错式视觉-文本-动作预训练。*arXiv* 预印本 *arXiv:2508.21112*，2025年。[62] Lucy Xiaoyang Shi、Brian Ichter、Michael Equi、Liy- iming Ke、Karl Pertsch、Quan Vuong、James Tanner、Anna Walling、Haohuan Wang、Niccolo Fusai 等。嗨，机器人：借助分层视觉-语言-动作模型实现开放式指令遵循。*arXiv* 预印本 *arXiv:2502.19417*，2025年。[63] Mustafa Shukor、Dana Aubakirova、Francesco Capuano、Pepijn Kooijmans、Steven Palma、Adil Zouitine、Michel Aractingi、Caroline Pascal、Martino Russi、Andres Marafioti 等。Smolvla：一种经济高效的视觉-语言-动作模型，适用于机器人学。*arXiv* 预印本 *arXiv:2506.01844*，2025年。[64] David Snyder、Asher James Hancock、Apurva Badithela、Emma Dixon、Patrick Miller、Rares Andrei Ambrus、Anirudha Majumdar、Masha Itkina 和 Haruki Nishimura。你的模仿学习策略比我的好吗？通过近似最优停止进行策略比较。载于机器人学：科学与系统会议（*RSS*）第二十一届会议论文集，2025年。[65] Andreas Steiner、Andr´e Susano Pinto、Michael Tschan- nen、Daniel Keysers、Xiao Wang、Yonatan Bitton、Alexey Gritsenko、Matthias Minderer、Anthony Sher- bondy、Shangbang Long 等。Paligemma 2：面向迁移的多功能视觉语言模型系列。*arXiv* 预印本 *arXiv:2412.03555*，2024年。[66] Tony Tao、Mohan Kumar Srirama、Jason Jingzhou Liu、Kenneth Shaw 和 Deepak Pathak。Dexwild：面向真实环境机器人策略的灵巧人类交互。*arXiv* 预印本 *arXiv:2505.07813*，2025年。[67] 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：通过先进的具身推理、思考和运动迁移，推动通用型机器人的发展前沿。*arXiv* 预印本 *arXiv:2510.03342*，2025年。[68]Gemini Robotics 团队、S Abeyruwan、J Ainslie、JB Alayrac、MG Arenas、T Armstrong、A Balakrishna、R Baruch、M Bauza、M Blokzijl 等。Gemini robotics：将人工智能带入物理世界，2025年。网址 *https://arxiv.org/abs/2503.20020*，2025年。[69] Octo Model 团队、Dibya Ghosh、Homer Walke、Karl Pertsch、Kevin Black、Oier Mees、Sudeep Dasari、Joey Hejna、Tobias Kreiman、Charles Xu 等。Octo：一开源通用机器人策略。*arXiv* 预印本 *arXiv:2405.12213*，2024年。[70] Russ Tedrake 等。Drake：面向机器人学的基于模型的设计与验证，2019年。[71] Peter Tong、Ellis Brown、Penghao Wu、Sanghyun Woo、Adithya Jairam Vedagiri IYER、Sai Charitha Akula、Shusheng Yang、Jihan Yang、Manoj Middepogu、Ziteng Wang 等。Cambrian-1：对多模态大语言模型进行全面开放、以视觉为中心的探索。神经信息处理系统大会进展，37:87310–87356，2024年。[72]Aaron Van Den Oord、Oriol Vinyals 等。神经离散表示学习。神经信息处理系统大会进展，30，2017年。[73] Peng Wang、Shuai Bai、Sinan Tan、Shijie Wang、Zhi- hao Fan、Jinze Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wenbin Ge 等。Qwen2-vl：增强视觉-语言模型对世界的感知能力，支持任意分辨率。*arXiv* 预印本 *arXiv:2409.12191*，2024 年。[74] Xiaofeng Wang 、Kang Zhao、Feng Liu、Jiayu Wang、Gu- osheng Zhao、Xiaoyi Bao、Zheng Zhu、Yingya Zhang 和 Xingang Wang。Egovid-5m：用于第一人称视频生成的大规模视频-动作数据集。*arXiv* 预印本 *arXiv:2411.08380*，2024年。[75] Xin Wang、Taein Kwon、Mahdi Rad、Bowen Pan、Is- hani Chakraborty、Sean Andrist、Dan Bohus、Ashley Feniello、Bugra Tekin、Felipe Vieira Frujeri 等。Holoassist：面向真实世界交互式人工智能助手的第一人称人类交互数据集。载于*IEEE/CVF*国际计算机视觉会议论文集，第20270–20281页，2023年。[76] Yating Wang、Haoyi Zhu、Mingyu Liu、Jiange Yang、Hao-Shu Fang 和 Tong He。Vq-vla：通过扩展向量量化动作词元化器，改进视觉-语言-动作模型。*arXiv* 预印本 *arXiv:2507.01016*，2025年。[77] Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Fei Xia、Ed Chi、Quoc V Le、Denny Zhou 等。思维链提示可激发大语言模型中的推理能力。神经信息处理系统大会进展，35:24824–24837，2022年。[78]Bernard L Welch。当涉及多个不同总体方差时“学生”问题的推广。*Biometrika*，34(1-2):28–35，1947年。[79]Junjie Wen、Yichen Zhu、Jinming Li、Zhibin Tang、Chaomin Shen 和 Feifei Feng。Dexvla：采用即插即用扩散专家实现通用机器人控制的视觉-语言模型。*arXiv* 预印本 *arXiv:2502.05855*，2025年。[80]Junjie Wen、Yichen Zhu、Jinming Li、Minjie Zhu、Zhibin Tang、Kun Wu、Zhiyuan Xu、Ning Liu、Ran Cheng、Chaomin Shen 等。Tinyvla：迈向用于机器人操作的快速、数据高效视觉-语言-动作模型。*IEEE*机器人与自动化快报，2025年。[81] Mengda Xu、Han Zhang、Yifan Hou、Zhenjia Xu、Linxi Fan、Manuela Veloso 和 Shuran Song。Dexumi：

将人手用作灵巧操作的通用操作接口。*arXiv* 预印本 *arXiv:2505.21864*，2025年。[82] Mingxing Xu、Wenrui Dai、Chunmiao Liu、Xing Gao、Weiyao Lin、Guo-Jun Qi 和 Hongkai Xiong。用于交通流预测的时空变换器网络。*arXiv* 预印本 *arXiv:2001.02908*，2020年。[83] Zhenjia Xu、Zhou Xian、Xingyu Lin、Cheng Chi、Zhiao Huang、Chuang Gan 和 Shuran Song。Roboninja：为多材料物体学习自适应切割策略。*arXiv* 预印本 *arXiv:2302.11553*，2023年。[84] GanlinYang、TianyiZhang、HaoranHao、Weiyun Wang、Yibin Liu、Dehui Wang、Guanzhou Chen、Zijian Cai、Junting Chen、Weijie Su 等。Vlaser：具备协同具身推理能力的视觉-语言-动作模型。*arXiv* 预印本 *arXiv:2510.11027*，2025年。[85] Jonathan Yang、Dorsa Sadigh 和 Chelsea Finn。Poly-bot：在拥抱可变性的同时，为不同机器人训练统一策略。*arXiv* 预印本 *arXiv:2307.03719*，2023年。[86]Jonathan Yang、Catherine Glossop、Arjun Bhorkar、Dhruv Shah、Quan Vuong、Chelsea Finn、Dorsa Sadigh 和 Sergey Levine。突破跨具身学习在操作与导航方面的极限。*arXiv* 预印本 *arXiv:2402.19432*，2024年。[87] Shuai Yang、Hao Li、Yilun Chen、Bin Wang、Yang Tian、Tai Wang、Hanqing Wang、Feng Zhao、Yiyi Liao 和 Jiangmiao Pang。Instructvla：从理解到操作的视觉-语言-动作指令微调。*arXiv* 预印本 *arXiv:2507.17520*，2025年。[88]Seonghyeon Ye、Joel Jang、Byeongguk Jeon、Sejune Joo、Jianwei Yang、Baolin Peng、Ajay Mandlekar、Reuben Tan、Yu-Wei Chao、Bill Yuchen Lin 等。基于视频的潜在动作预训练。*arXiv* 预印本 *arXiv:2410.11758*，2024年。[89] Qiying Yu、Quan Sun、Xiaosong Zhang、Yufeng Cui、Fan Zhang、Yue Cao、Xinlong Wang 和 Jingjing Liu。Capsfusion：重新思考大规模图文数据。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第14022–14032页，2024年。[90]Chengbo Yuan、Rui Zhou、Mengzhen Liu、Yingdong Hu、Shengjie Wang、Li Yi、Chuan Wen、Shanghang Zhang 和 Yang Gao。Motiontrans：人类虚拟现实数据助力机器人操作策略的运动层级学习。*arXiv* 预印本 *arXiv:2509.17759*，2025年。[91]Wentao Yuan、Jiafei Duan、Valts Blukis、Wilbert Pumacay、Ranjay Krishna、Adithyavairavan Murali、Ar- salan Mousavian 和 Dieter Fox。Robopoint：用于机器人学空间可供性预测的视觉-语言模型。*arXiv* 预印本 *arXiv:2406.10721*，2024年。[92]Michał Zawalski、William Chen、Karl Pertsch、Oier Mees、Chelsea Finn 和 Sergey Levine。通过具身思维链推理实现机器人控制。*arXiv* 预印本 *arXiv:2407.08693*，2024年。[93] Andy Zhai、Brae Liu、Bruno Fang、Chalse Cai、Ellie Ma、Ethan Yin、Hao Wang、Hugo Zhou、James Wang, Lights Shi 等。点燃视觉语言模型，迈向具身空间。*arXiv* 预印本 *arXiv:2509.11766*，2025年。[94] Shiduo Zhang、Zhe Xu、Peiju Liu、Xiaopeng Yu、Yuan Li、Qinghui Gao、Zhaoye Fei、Zhangyue Yin、Zuxuan Wu、Yu-Gang Jiang 等。Vlabench：用于语言条件机器人操作的大规模基准测试，涵盖长时程推理任务。收录于*IEEE/CVF* 国际计算机视觉会议论文集，第 11142–11152 页，2025年。[95] Zhuosheng Zhang、Aston Zhang、Mu Li 和 Alex Smola。大语言模型中的自动思维链提示。*arXiv* 预印本 *arXiv:2210.03493*，2022年。[96] Enshen Zhou、Jingkun An、Cheng Chi、Yi Han、Shanyu Rong、Chi Zhang、Pengwei Wang、Zhongyuan Wang、Tiejun Huang、Lu Sheng 等。Roborefer：面向机器人学视觉语言模型中结合推理的空间指代。*arXiv* 预印本 *arXiv:2506.04308*，2025年。[97] Xueyang Zhou、Yangming Xu、Guiyao Tie、Yongchao Chen、Guowen Zhang、Duanfeng Chu、Pan Zhou 和 Lichao Sun。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年。[98] Zhongyi Zhou、Yichen Zhu、Junjie Wen、Chaomin Shen 和 Yi Xu。ChatVLA-2：基于预训练知识进行开放世界具身推理的视觉-语言-动作模型。*arXiv* 预印本 *arXiv:2505.21906*，第 3 卷，2025年。[99] Zhongyi Zhou、Yichen Zhu、Minjie Zhu、Junjie Wen、Ning Liu、Zhiyuan Xu、Weibin Meng、Yaxin Peng、Chaomin Shen、Feifei Feng 等。ChatVLA：利用视觉-语言-动作模型实现统一的多模态理解与机器人控制。收录于*2025*年自然语言处理经验方法会议论文集，第 5377–5395 页，2025年。[100] Brianna Zitkovich、Tianhe Yu、Sichun Xu、Peng Xu、Ted Xiao、Fei Xia、Jialin Wu、Paul Wohlhart、Stefan Welker、Ayzaan Wahid 等。RT-2：视觉-语言-动作模型将网络知识迁移至机器人控制。收录于机器人学习会议，第 2165–2183 页。PMLR，2023年。

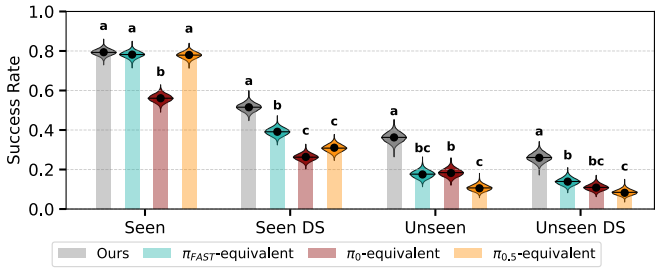


图 S1. 在仿真中进行的架构消融：比较我们的模型与等效于 π_0、$\pi_{0.5}$ 和 π_{FAST} 的变体在已见任务和未见任务上的表现，并考察有无分布偏移的情况。

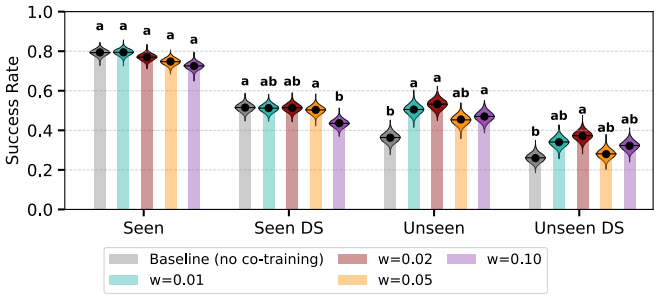


图 S2. 在使用标准视觉-语言数据进行单阶段协同训练时，对交叉熵损失权重 w 的消融。

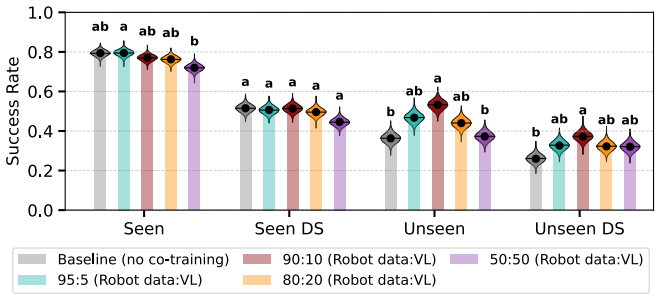


图 S3. 在单阶段协同训练中，采用标准视觉-语言数据时，每批次机器人数据与协同训练数据比例的消融实验。

1. 模型架构与超参数消融

A. 模型架构消融。

我们展示了四种模型架构的消融结果：(1) 我们的方法：由一个视觉语言模型和一个流变换器动作头（ActionFT）组成；(2) π_{FAST}等效模型 [58]：仅由一个视觉语言模型构成，该模型以自回归方式生成 FAST 令牌，随后将其解码为连续机器人动作。(3)π_0等效模型：由一个视觉语言模型和一个动作专家组成，其形式见 [6]。视觉语言模型各层中所有令牌的注意力键和值均用作动作专家的视觉语言表征。(4) $\pi_{0.5}$等效模型 [33]：与采用视觉语言模型和动作专家的 π_0等效模型类似，但流匹配时间步嵌入会注入动作专家的每一层；而在 π_0等效模型中，时间步嵌入仅在第一层与加噪动作一同注入。四种架构均采用相同的预训练视觉-语言模型（PaliGemma2-PT），并使用相同的超参数在 TRI-Ramen数据上训练。例外情况是，由于内存消耗较高， π_0等效模型和 $\pi_{0.5}$等效模型的批量大小设为 112，训练 230k 步；而我们的方法和 π_{FAST} 等效模型的批量大小设为 128，训练 200k 步。如图 S1 所示，尽管π_{FAST}等效模型和 $\pi_{0.5}$等效模型取得了与我们的方法相当的分布内表现，但它们在分布偏移和未见任务上的表现明显下降，表明我们更紧凑的表示能够提升泛化能力。此外， π_0等效模型在分布内任务上的表现显著较差，说明在动作网络的每一层注入流匹配时间步嵌入至关重要。

B. 超参数消融。

我们研究了两个对有效训练至关重要的超参数：用于平衡流匹配损失和交叉熵损失的加权因子，以及每个批次中目标机器人数据与协同训练数据的比例。

我们在视觉语言数据单阶段协同训练设置下分析这些超参数。如图 S2 和 S3 所示，提高加权因子或协同训练数据比例都会降低分布内性能，而将这些参数设置得过低则会削弱协同训练带来的泛化收益。我们采用 $w = 0.02$ 和对于样本未与目标机器人数据一一配对的模态（即标准视觉语言数据、跨具身机器人数据和人类视频），采用机器人数据与协同训练数据 9:1 的比例进行协同训练。这一配置在保持分布内性能的同时，最大限度地提升了泛化效果。

对于与目标机器人数据配对的模态，例如机器人数据的语言标注或离散动作词元，模型会同时从连续动作和相应的协同训练样本中学习。跨具身机器人数据是一种特殊情况，因为学习目标仍然是对机器人动作进行流匹配；在这种设置下，第一阶段训练期间，目标机器人数据与跨具身数据按 6:4 的批次比例混合。

2. 训练详情

多任务 VLA 策略训练的每个阶段运行 200k 步，默认批量大小为 128，并采用余弦学习率调度，在最后 60k 步内将学习率从 2e-5 衰减至 2e-6。在此默认设置下，每个训练阶段使用 16 个 H100 GPU 约需 64 小时。机器人数据与协同训练数据的比例遵循附录 1B。例外情况包括：（1）由于视频数据量较大，所有视频数据上的潜在动作第一阶段训练使用 256 的批量大小；（2）对于第 III-C 节中的多模态协同训练，为确保每种模态的有效样本数与单模态协同训练策略保持一致，我们在某些训练阶段增大了批量大小。这些特殊 VLA 策略在各训练阶段使用的数据模态、数据比例和批量大小列于表 S1。对于在单一任务上训练或微调的 VLA 策略

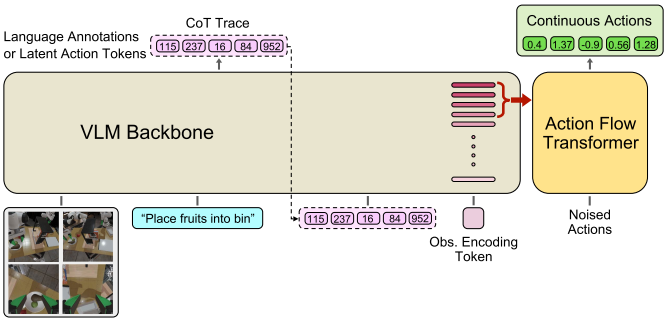


图 S4. 思维链条件策略的训练与推理流程示意图。

我们训练 30k 步，并在最后 24k 步内采用余弦调度，将学习率从 2e-5 衰减至 2e-6。所有其他共享训练超参数列于表 S2。

3. 数据处理与整理详情

A. 机器人轨迹的密集语言标注

我们通过比较机器人当前步骤与未来 16 步（与动作块时域相匹配）时的末端执行器状态，得出脚本标注。对于每个夹爪，这种比较可得出其沿 xyz 轴的移动、滚转、俯仰和偏航方向的旋转变化，以及夹爪宽度的变化。我们为左侧和右侧夹爪设计了以下模板： [右侧/左侧] 夹爪向 [前方/后方][右方/左方] [上方/下方], 旋转 [滚转正向/负向] [俯仰正向/负向] [偏航正向/负向], [张开/闭合]我们将 xyz 方向移动、旋转和夹爪宽度变化的阈值分别设为 2.5 厘米、20 度和 1.3 厘米。如果某项变化未超过相应阈值，我们就会在标注中省略对应的描述块。对于基于视觉语言模型的标注，我们在图 S5 中给出了提供给 GPT-5 [53] 的完整提示。此外，图 S6 展示了脚本标注和基于视觉语言模型的标注示例。

B. 人类视频

1) 潜在动作：**人类视频数据整理**。 我们使用五个公开可用的第一人称人类视频数据集：Ego4D [28], EgoDex [31],Something-Something V2 [27],Epic Kitchen [18], 和 HoloAssist [75]。对于 Ego4D、Epic Kitchen 和 HoloAssist，我们利用所提供的语言标注，并依据启发式规则筛选视频（例如，最大回合时长，以及根据动作动词如 看、走、笑 移除与操作无关的回合）。对于 EgoDex 和 Something-Something V2，它们已组织为带有精确指令的高质量短视频片段，我们直接使用，无需额外筛选。我们还利用 EgoDex 视频子集中可用的反向播放标注。各数据集的总时长见表 S3。

潜在动作模型（LAM）的实现。

IDM 是一个 12 层的时空变换器 [82]。向量量化模块（码本大小为 32）将 IDM 的连续输出映射为每个时间步 8 个离散潜在令牌。视觉 FDM 是一个 12 层的空间变换器 [82]。ActionFDM 是一个轻量级卷积解码器，可将令牌转换为机器人动作块。我们使用 16 张 H100 GPU，以 1024 的批量大小训练潜在动作模型，训练 300k 步（约 68 小时），学习率为 1e-4，并采用 1k 步线性预热和 20k 步余弦衰减。

2) 视觉语言模型生成的标注：我们使用三个用于标注的人类视频数据集：Ego4D、EgoDex 和 Something-Something V2。代表性标注示例见图 S7，已标注数据样本数量汇总于表 S3。对于 Ego4D，我们首先通过计算两个已筛选子集的视频唯一标识符（唯一标识符）交集来筛选数据集：EgoVid [74] 和 EgoHOD [57], ，从而确保仅纳入同时被两条筛选流程保留的视频。随后，我们按照 EgoVid 的划分将筛选后的视频切分为 4 秒片段，并以 1 秒间隔对每个视频片段进行下采样。此外，Ego4D 提供细粒度动作描述，我们将其与每一帧配对作为动作先验。因此，我们对 GPT-5 的提示针对每个视频片段包含：(i) 以 1 秒间隔下采样的帧，(ii) 每一帧的动作先验，(iii) 指令，以及 (iv) 一幅描绘世界坐标系三轴的参考图像。完整提示见图 S8。由此得到一组帧级配对，每一帧对应一条语言标注。为获得准确的视频片段级指令，我们执行第二步提示：将每个视频片段的首帧和末帧以及生成标注的完整列表提供给 GPT-5，以生成清晰、精确的视频片段摘要（提示见图 S9）。

对于 EgoDex 和 Something-Something V2，由于这两个数据集均未提供细粒度动作描述，我们在第一步提示中省略了帧级动作先验。此外，这两个数据集已为每个视频包含高质量的视频片段级指令，因此我们跳过第二步提示。

C. 离散机器人动作令牌

FAST 令牌。我们直接将现成的 FAST 分词器 [58]应用于 TRI-Ramen，而不进行微调，得到的平均令牌长度为 42.1，重建误差为 2.26e-4。我们还尝试在 TRI-Ramen 上对该分词器进行微调，得到的平均令牌长度为 36.8，重建误差为 2.26e-4。微调并未带来显著改进。因此，鉴于现成的分词器是在更大且更多样的机器人数据集上训练的，能够提供更具泛化性的表示，我们选择在不进行微调的情况下使用它。

VQ-VAE 离散动作令牌。 我们采用轻量级 VQ-VAE 架构，将动作块压缩为离散令牌。编码器和解码器都是紧凑型网络：编码器使用一系列 1D 卷积层和残差块，将动作块

表 S1 特殊策略各训练阶段所使用的数据模态、批次数据比例和批次大小（BS）。

策略	第一阶段	第二阶段	第三阶段
三阶段潜在动作协同训练	TRI-Ramen : OXE-Ramen : 人类视频 = 3 : 3 : 4; 批次大小 = 256	TRI-Ramen : OXE-Ramen = 6 : 4; 批次大小 = 128	TRI-Ramen：OXE-Ramen = 9 : 1；批次大小 = 128
视觉-语言 + TRI-OXE-Ramen FAST	TRI-Ramen (FAST)：标准视觉-语言数据：OXE-Ramen (FAST) = 5 : 2: 3；批次大小 = 128	TRI-Ramen：标准视觉-语言 数据 = 9: 1；批次大小 = 128	—
+机器人标注数据	标准视觉语言数据：语言 TRI-Ramen 的语言标注 = 1 : 1；批次大小 = 256	TRI-Ramen：标准视觉-语言 数据 = 9: 1；批次大小 = 128	—
+人工标注数据	标准视觉语言数据：语言 TRI-Ramen 的标注：语言-言标注用于人类视频 = 1 : 1 : 1；批次大小 = 384	TRI-Ramen：标准视觉-语言 数据：人类的语言标注 视频 = 9 : 0.5 : 0.5；批次大小 = 128	—
+跨具身机器人-数据（最终模型）	标准视觉语言数据：语言 TRI-Ramen 的语言标-注人类视频 = 1 : 1: 1；批次大小 = 384	TRI-Ramen：标准视觉语言 数据：OXE-Ramen：语言标注 人类视频 = 4 : 1 : 4 : 1；批次大小 = 256	TRI-Ramen：标准视觉-语言 数据：人类视频的语言标注 视频 = 9 : 0.5 : 0.5；批次大小 = 128

表 S2 所有策略共用的训练超参数。

超参数	值
图像观测时域	1
相机数量（TRI-Ramen）	4（2 个场景相机，1 个腕部相机每个机械臂）
图像增强	随机裁剪（针对 TRI-Ramen，$256 \times 342 \rightarrow 224 \times 224$); 颜色抖动（亮度 $= 0.3$，对比度 $=$ 0.4，饱和度 $= 0.5$，色调 $= 0.05$)
学习率预热步数	1000
学习率预热调度器	线性
学习率衰减调度器	余弦衰减

表 S3 人类视频数据的时长及视觉语言模型生成的标注数据样本数量汇总。

数据集	视频时长	标注数据样本
Ego4D	774.5 小时	5.2M
EgoDex	744.4 小时	3.0M
反向 EgoDex	455.7 小时	–
Something-Something V2	155.8 小时	0.8M
Epic Kitchen	60.4 小时	–
HoloAssist	80.8 小时	–
总计	**2271.6 小时**	**9.0M**

（时域长度为 16，维度为 20）映射到低维潜在空间；解码器则采用与编码器相对应的架构，从量化后的嵌入中重建原始动作块。向量量化器将每个潜在向量映射到学习得到的、大小为 32 的码本中与之最近的条目，从而将连续动作空间离散化为每个动作块 8 个令牌。模型使用重建损失、量化损失和承诺损失的组合进行训练，具体如 [72] 所述。

4. 策略部署

我们采用与 [1] 相同的双臂Franka机器人平台。在仿真运行轨迹中，策略在每个推理步骤预测一个 16 步动作块，并以开环方式执行前 8 步，然后重新计算动作。在真实世界部署过程中，我们观察到，预测动作块之间的不连续会导致动作生硬。为解决这一问题，我们采用时间集成技术。具体而言，策略通过连续推理（平均延迟为 0.146 秒）生成在时间上相互重叠的动作块。在每个时间步，实际执行的动作由最近预测的四个动作块中对应的动作取均匀平均计算得出。

5. 统计分析

为降低过度解读微小实证差异的风险，我们采用 [1] 的统计分析框架，以：(1) 通过 A/B测试严格比较多种协同训练策略；(2) 使用贝叶斯分析量化认知不确定性。对于基于平均任务完成百分比的比较，我们使用 Welch t检验 [78]。对于成功率的比较，我们使用 STEP检验 [64], ，这是一种用于二元结果 A/B测试的最先进方法。

比较 k 策略时，我们进行所有 k（$k-1$）/2成对检验，并使用邦费罗尼校正将族错误率控制在 5%。最终结果通过紧凑字母显示（CLD） [59],进行汇总，这是一种基于标签的 A/B 测试表示。两个不共享任何英文字母的紧凑字母显示表示至少以 95% 的置信度在统计上存在差异。例如，若策略 A、B 和 C 分别对应a、ab和b 这几个紧凑字母显示，则只有策略对（A，C）的性能差异显著。紧凑字母显示的字典序排列通常与实测性能相符（例如本例中的 A >B > C），但当某些成对差异在统计上不显著时，这一点并无保证。示例请见图4G（“已见DS”列）。

除了假设检验外，我们还报告了各项策略的贝叶斯不确定性估计。对于成功率，我们使用均匀贝塔分布

先验和观测到的成功次数来计算后验分布。对于平均任务完成百分比，我们对任务进度值采用均匀狄利克雷先验。正文中的条形图叠加了小提琴图，以可视化后验不确定性（圆点和水平线分别表示经验均值和后验均值）；此外，为完整起见，图 S10、S11、S12、S13 展示了原始经验分布（仅以圆点表示经验均值）。

6. 仿真实验详情

我们在图 S14 中展示了仿真基准下的13个已见任务和8个未见任务在常规条件和分布偏移条件下的情况，每个任务均以一张展示初始配置的快照表示。对于初始条件和成功标准，我们采用与 [1]相同的规范。具体而言，在初始条件方面，每次 rollout 都根据仿真随机种子从预定义分布中确定性地采样。我们使用相同的随机种子集评估所有策略，以确保比较公平。从这些快照（图 S14）可以看出，场景在视觉上模棱两可；因此，策略无法仅凭视觉外观判断任务，需要借助语言来消除歧义。在成功标准方面，每个任务都分配了固定时间预算；执行期间，我们监测模拟器状态中的一组任务特定谓词，以确定策略是否在分配的时间预算内成功完成任务。

7. 真实世界实验详情

A. 语言跟随。

我们在图 S15 和图 S16 中展示了已见物体、指令泛化和未见物体设置下的所有初始条件及相应指令。任务完成率依据四个里程碑进行评估：

1）机器人朝目标物体移动以进行抓取，并在距离目标物体 10 厘米以内到达。2）机器人成功抓取目标物体。3）机器人将抓取的物体移动到距离目标位置 10 厘米以内。4）根据指令，将抓取的物体正确放置在目标位置：

- **位于内部/上方：** 抓取的物体位于目标位置内部或上方。
- **相邻：** 抓取的物体位于目标位置 10 厘米范围内。

B. 长时程灵巧操作。

我们在图 S17 中展示了三个任务的样本初始条件。图 S18 、 S19 和 S20 展示了任务执行过程的视频帧拼图。每个任务的任务完成里程碑定义如下：

PackItemsIntoStringBag

- 左臂抓取抽绳袋的边缘。
- 右臂抓取对侧边缘并打开袋子。
- 机器人拿起瓶子。
- 机器人拿起瓶盖。
- 机器人牢牢盖紧瓶子。
- 机器人将瓶子放置在袋子内。
- 机器人拿起球。
- 机器人将球放置在袋子内。
- 机器人拿起太阳镜。
- 机器人折叠太阳镜的第一根镜腿。
- 机器人折叠太阳镜的第二根镜腿。
- 机器人将折叠好的太阳镜放入袋子。
- 机器人抓取两根抽绳。
- 机器人通过拉动抽绳收紧袋子。

PourIngredientsIntoSoup

- 机器人抓取汤锅的两个把手。
- 机器人将锅放置到炉子左方的炉灶上。
- 右臂拿起锅盖。
- 右臂将锅盖放在桌子上。
- 左臂拿起胡萝卜碗，并将其移到锅上方。
- 右臂抓取锅铲。
- 机器人用锅铲将胡萝卜倒入锅中。
- 右臂将锅铲放置在桌子上。
- 左臂将空碗放置在桌子上。
- 左臂拿起蘑菇碗，并将其移至锅的上方。
- 左臂将蘑菇倒入锅中，并将空碗放置在桌子上。
- 左臂拿起黄瓜碗，将其移到锅上方。
- 左臂将黄瓜倒入锅中，并将空碗放在桌子上。
- 右臂拿起锅盖。
- 右臂将锅盖盖在锅上。

StoreCleanDishes

- 机器人打开橱柜门。
- 机器人拿起碗。
- 机器人将碗叠放在橱柜内的碗上。
- 机器人拿起盘子。
- 机器人将盘子叠放在橱柜内的盘子上。
- 机器人拿起杯子。
- 右臂将杯子递给左臂。
- 机器人将杯子叠放到橱柜上的杯子中。
- 机器人拿起玻璃杯。
- 右臂将玻璃杯递给左臂。
- 机器人将玻璃杯倒扣插入架子。

C. 实验流程

所有真实世界评估均遵循标准化且受控的实验流程，以确保不同策略之间的公平比较。对于每种评估设置，我们首先生成待评估的策略检查点列表，并按顺序从列表中选取检查点。对于每次评估，我们根据任务规范采样一个初始条件（IC），其中包括物体及其姿态。

随后手动配置桌子及周围的工作空间，使其尽可能符合该初始条件。

选定的策略检查点在实体机器人上执行一次，之后实验人员按照评分标准记录任务结果。随后，评估人员使用相同的初始条件，继续评估列表中的下一个检查点。重复此过程，直到在该初始条件下评估完所有检查点，然后抽取一个新的初始条件并重新开始。

在短时间内使用相同的初始条件评估所有检查点，这一方案可确保性能差异归因于策略行为，而非硬件、光照或环境布置的变化。实际上，所有检查点都在尽可能近乎相同的物理条件下进行评估。此外，任何缓慢且难以察觉的条件变化都会同等影响所有检查点。

D. 评分细则质量保证

与 [1] 类似，我们开展一轮质量保证（QA），以估计在真实世界运行轨迹的评分细则评估中，由人为错误或偏差造成潜在差异的频率。我们抽取了 895 段评估运行（占总数的 31.6%），并请一组与原评估人员不同的审阅人员审查运行视频及其评分标准。为减少偏差，审阅人员并不知道每段视频评估的是哪个检查点。QA 的总体任务完成度差异（按所有回合和里程碑取平均值计算）为 2.01%。

表 S4 视觉语言模型骨干网络在视觉语言基准上的评估结果。

模型	RealWorldQA	GQA	SpatialEval	MMBench	SEED	MME-P	MME-R	LEGO
PaliGemma2-PT（视觉语言动作主干网络）	0.24	31.68	0.26	0.01	0.24	701.56	242.50	0.03
PaliGemma2-Mix	0.55	59.81	0.34	0.68	0.70	1449.43	299.29	0.27
基线（无协同训练）	0.02	0.06	0.02	0.00	0.03	193.75	44.29	0.01
标准视觉语言数据协同训练（+视觉语言数据）	0.39	57.69	0.28	0.26	0.49	88.24	54.64	0.27
脚本化语言标注协同训练	0.05	0.00	0.07	0.00	0.06	80.06	32.14	0.02
基于视觉语言模型的语言标注协同训练	0.02	0.06	0.03	0.00	0.04	92.00	31.07	0.01
跨具身机器人数据协同训练	0.01	0.01	0.01	0.00	0.01	114.02	33.21	0.00
用于人类视频协同训练的视觉语言模型生成的语言标注	0.03	0.02	0.01	0.07	0.09	105.43	65.36	0.03
+机器人标注数据	0.49	57.58	0.32	0.60	0.65	977.79	177.14	0.27
+人工标注数据	0.48	55.78	0.39	0.33	0.51	1397.43	253.57	0.27
+跨具身机器人数据（最终模型）	0.47	58.02	0.33	0.56	0.63	1319.75	267.86	0.27

你是一名严谨的双手机器人视频帧标注员。

尝试执行的任务是：*[任务指令]*

你将收到：
• 一组*{K}*视频帧；每帧后面紧接着一条针对该帧的弱启发式提示。
• 末尾还有一张额外的参考图像，展示了定义前方、后方、右方、左方、上方和下方的世界坐标三轴。

你的任务：
对于每一帧*{K}*视频帧（忽略最后一张参考图像），请为该帧中最显著的操作动作撰写一条简短的祈使式描述。

每条标注都必须包含以下内容：1）每
ach acti条描述都需要说明是针对左夹爪还是右侧夹爪。当两个夹爪同时移 ve
同时的 时，需要分别描述每个夹爪的动作。
2）在可见时提及被操纵物体（例如“红色方块”“马克杯”“螺丝”“锅盖”“工具”）。如果不确定，请使用通用名词，例如“物体”“方块”“手柄”或“容器”。

3) 有运动时，请注明相对于参考三轴的运动方向：前方/后方/左方/右方/上方/下方；旋转时使用“旋转”（偏航为正表示逆时针，为负表示顺时针）或“倾斜”（不要说滚转、俯仰、偏航）。4) 不必在每一帧都提到旋转；认为旋转明显时再提，有时也可以相对于某个物体描述旋转，例如朝杯子倾斜以倒水。

将启发式提示用作先验：补充遗漏的物体名称。依据参考三轴（世界坐标系）判断左方/右方，而非相机视角。

风格与限制：
• 优先使用以下动词：接近、对齐、张开夹爪、闭合夹爪、抓取、旋转、倾斜、释放、抬起、放下、向前方/后方/左方/右方/上方/下方移动、推动、拉动、滑动、插入、收回、放置、保持。
• 不要提及参考图像、帧、相机或任何元说明。
• 不要为条目编号，不要添加评论，也不要使用代码围栏。

输出格式：
仅返回一个 Python 列表，其中恰好包含 *{K}* 个字符串（每个非参考坐标系对应一个），使用单引号和逗号，例如：['标注1‘，’标注2‘，.....]。

如果数量与 *{K}*不符，请在返回前进行调整。

[图像 1] [脚本化标注 1][图像 2] [脚本化标注 2]
.
.
.
[图像 K] [脚本化标注 K]

这里给出了参考坐标系，展示参考坐标系，并标明前方、后方、右方、左方、上方和下方的方向。你必须严格遵循该参考坐标系。

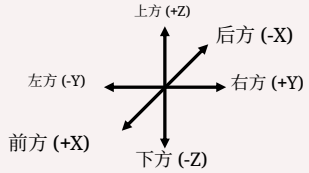


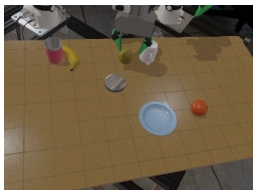
图 S5. 用于生成机器人数据的基于视觉语言模型的语言标注的提示词。

机器人数据语言标注将青苹果放在碟子上



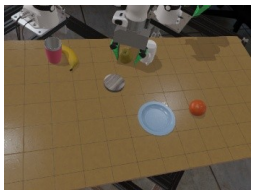
脚本控制： 左夹爪移动
前方 右方 下方，旋转 偏航
负向。

基于视觉语言模型： 左夹爪向
前右方和下方移动，旋转
顺时针对齐至青
苹果。



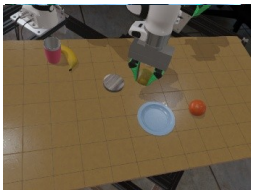
脚本控制： 左夹爪向
后方和下方移动。

基于视觉语言模型： 左夹爪向
后方和下方移动以进行定位
围绕青苹果。



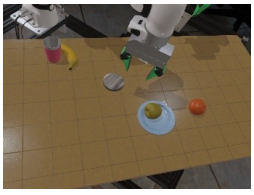
脚本控制： 左夹爪移动
向前上方移动，然后闭合。

基于视觉语言模型： 左夹爪靠近
抓取青苹果并移动
向前并向上。



脚本控制： 左夹爪向上方移动，
张开。

基于视觉语言模型： 左夹爪向上方移动
到蓝色茶碟上方并张开以
释放苹果。



脚本控制： 无

基于视觉语言模型： 左夹爪收回
放置苹果后从蓝色茶碟向后方
移动。

将桌上的杯子挂到杯架上



脚本控制： 左夹爪向前方、右
方、下方移动，偏航负向旋转。

基于视觉语言模型： 左夹爪继续向右前
方和下方移动，对齐至灰色马克杯上方。



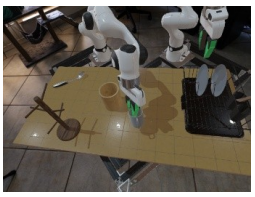
脚本控制： 左夹爪向
下方移动。

基于视觉语言模型： 左夹爪移动
向下移至灰色马克杯上。



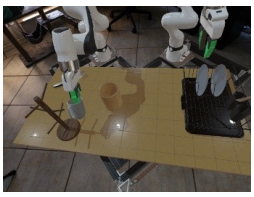
脚本控制： 左夹爪移动
向前、向左、向上移动，然后闭合。

基于视觉语言模型： 左夹爪合拢，准备
抓取灰色马克杯，并稍微抬起
前方-左方并向上。



脚本控制： 左夹爪向左方移动
上方，偏航向负方向旋转。

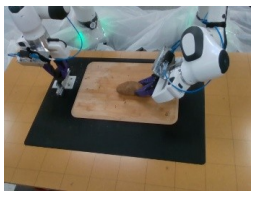
基于视觉语言模型： 左夹爪携带
马克杯向左方并向上朝树形支架
移动，同时顺时针旋转。



脚本控制： 左夹爪向左移动
下方。

基于视觉语言模型： 左夹爪向左移动
并向下移动，使马克杯把手对齐
与支架销钉。

剥土豆



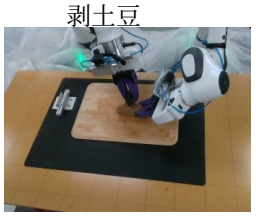
脚本控制： 左夹爪向右移动
上方，向正方向滚转，闭合。

基于视觉语言模型： 左夹爪抬起
削皮器并向右方移动，同时
稍微倾斜；右侧夹爪保持
土豆稳定。



脚本控制： 左夹爪向右方移动
下方。

基于视觉语言模型： 左夹爪将
削皮器向右方和下方移向
土豆，同时右侧夹爪继续
保持。



脚本控制： 左夹爪向左方移动。

基于视觉语言模型： 左夹爪拉动
削皮器向左划过土豆，进行一次
削皮动作，同时右侧夹爪
将其保持在下方。



脚本控制： 左夹爪向
前方右侧移动。

基于视觉语言模型： 左夹爪移动
削皮器向前右方移动以复位
为下一次动作做准备；右侧夹爪
保持夹持。



脚本控制： 左夹爪向前移动
并向左移动，俯仰角正向旋转。

基于视觉语言模型： 左夹爪扫动
削皮器向前偏左移动，略微
倾斜着削皮，同时右侧夹爪
保持土豆不动。

在意大利面上涂奶酪



脚本控制： 左夹爪移动
向下移动，俯仰角负向旋转。

基于视觉语言模型： 左夹爪
向下接近
圆柱形奶酪容器，并倾斜
向下。



脚本控制： 左夹爪移动
向前向下，闭合。

基于视觉语言模型： 左夹爪移动
向前并向下抓取
奶酪容器，闭合
夹爪。



脚本控制： 左夹爪移动
向后上方移动，沿滚转方向正向旋转
俯仰为正。

基于视觉语言模型： 左夹爪抬起
奶酪容器，使其向上并向后移动，
向上倾斜至直立。



脚本控制： 左夹爪移动
向前、向右、向下，俯仰旋转
偏航角为负。

基于视觉语言模型： 左夹爪将
奶酪容器向前偏右并
向下移至意大利面碗上方，倾斜
向下倾倒。



脚本控制： 左夹爪移动
向后移动，正向旋转偏航角。

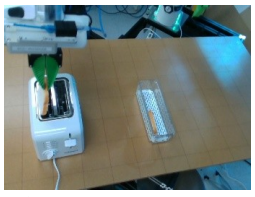
基于视觉语言模型： 左夹爪旋转
逆时针，同时移动
向后移动以将奶酪撒在
意大利面上。

将吐司插入烤面包机



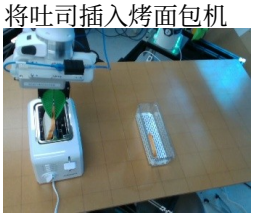
脚本控制： 左夹爪向左移动，
俯仰角向负方向旋转。

基于视觉语言模型： 左夹爪向左移动
并倾斜，使面包与烤面包机插槽对齐。



脚本控制： 左夹爪向下移动。

基于视觉语言模型： 左夹爪向
下放置，将面包插入烤面包机。



脚本控制： 左夹爪向前、向右、
向下移动，然后张开。

基于视觉语言模型： 左夹爪张开以释放
面包，同时略微向前、向右、向下移动。



脚本控制： 左夹爪向右下方移动。

基于视觉语言模型： 左夹爪朝架子
方向向右下方移动，以取另一片面包。



脚本控制： 左夹爪闭合。

基于视觉语言模型： 左夹爪闭合，
从架子上抓取第二片面包。

图 S6。 示例 机器人数据的脚本控制和基于视觉语言模型的语言标注。每一行对应一个情节的片段 e.

VLM-generated Annotations for Human Videos

Instruction: Fill the green cap from the detergent bottle and pour it into the open detergent drawer.



Left hand tilt the detergent bottle down to pour into the green cap held steady by the right hand.

Right hand tilt the filled green cap down over the detergent drawer and pour, while the left hand holds the bottle.

Right hand keep pouring detergent down into the open drawer while moving slightly forward; left hand hold the detergent bottle.

Right hand continue pouring down into the drawer; left hand hold the bottle near the sink.

Right hand retract backward from the drawer after the pour; left hand hold and slightly lift the detergent bottle up.

Instruction: Hold the bowl and slice the food with the knife.



Left hand reach forward to grasp the rim of the bowl.

Left hand hold the bowl steady; right hand press the knife down to cut the food.

Left hand hold the bowl; right hand pull the knife backward to slice the food.

Left hand steady the bowl; right hand move the knife right to reposition above the bowl.

Left hand keep holding the bowl; right hand lower the knife down onto the cutting board to place it.

Instruction: Fold a small T-shirt on a green tablecloth while sitting.



Left hand pull the left side of the t-shirt leftward while the right hand pull the right side rightward to straighten.

Right hand move forward to pinch the right sleeve of the t-shirt.

Right hand grasp the top-right edge of the t-shirt and lift slightly up.

Right hand pull the right side of the t-shirt leftward to fold toward the center.

Left hand hold the bottom hem while the right hand press the folded panel down.

Instruction: Insert the plush, and close the box on a wooden table with a white background while sitting.



Right hand grasp the pink plush and move it forward toward the box; left hand move forward to steady the box front edge.

Right hand insert the plush down into the box; left hand hold the box rim.

Left hand and right hand move up to grasp the upright box lid at the sides.

Left hand and right hand rotate the lid down and forward to close the box.

Left hand press the top of the lid down; right hand push the front flap backward to seat it.

Instruction: Pushing mobile phone with marker pen.



Right hand approach the phone with the marker pen moving left.

Right hand bring the marker pen left to align with the phone's right edge.

Right hand press the marker pen against the phone and push it left.

Right hand continue pushing the phone left with the marker pen.

Right hand slide the phone farther left using the marker pen.

Instruction: Right hand insert the yellow toothbrush down into the sand.



Right hand insert the yellow toothbrush down into the sand.

Right hand push the toothbrush further down to plant it.

Right hand pull sand forward to cover the toothbrush base.

Right hand press the sand down over the toothbrush.

Right hand sweep sand right to smooth and hide the toothbrush.

Fig。S7. 人类视频的视觉语言模型生成标注示例。每行对应视频片段中的一个片段。

你是一名精准的人类 egovideo 视频帧标注员。

尝试执行的任务是：*[任务指令]*

你将收到：
• 一组*{K}*视频帧；每帧后面紧接着是针对该帧的弱动作提示。
• 最后还有一张额外的参考图像，展示定义前方的世界坐标三轴 ，后方 ，右方，left, 上方和下方。

你的任务：
对于 *{K}*视频帧中的每一帧（忽略最后一张参考图像），为该帧中可见的最显著操作动作生成一条简短的祈使式描述。

每条标注的必需内容：1) E
ach action needs to 说明是左手还是右手。当双手同时移动时，
y你需要分别描述 每只手的动作。
2）当被操纵物体可见时，请注明物体名称（例如“红色方块”“马克杯”“螺丝”“锅盖”“工具”）。如果不确定，请使用“物体”“方块”“手柄”或“容器”等通用名词。
3）发生运动时，请注明相对于参考三轴的运动方向：前方/后方/左方/右方/上方/下方；描述旋转时，请使用“旋转”“倾斜”“转动”“自转”“扭转”“枢转”或类似说法（不要说滚转、俯仰、偏航）。
4）不必在每一帧都提到旋转；当你认为旋转很明显时再说明。有时也可以参照某个物体描述旋转，例如倾斜至杯子一侧以便倒水。

将动作提示作为先验：补充缺失的物体名称。根据参考三轴（世界坐标系）判断左方/右方，而非相机视角。

风格与限制：
• 优先使用动词：优先使用动词：接近、对齐、张开手、合拢手、抓取、旋转、倾斜、释放、抬起、放下、向前/后方/左方/右方/上方/下方移动、推动、拉动、滑动、插入、收回、放置、保持。
• 不要提及参考图像、帧、相机或任何元说明。
• 不要给条目编号，不要添加评论，也不要使用代码围栏。

输出格式：
仅返回一个 Python 列表，其中恰好包含 *{K}* 个字符串（每个非参考坐标系对应一个），使用单引号和逗号，例如：['标注1', '标注2', ...]。

如果数量与 *{K}*不符，请在返回前进行调整。

[图像 1] [动作提示 1][图像 2] [动作提示 2]
⋮
[图像 K] [动作提示 K]

这里还有一个参考坐标系，用来展示坐标系并指示前方、后方、右方、左方、上方和下方的方向。你需要严格遵循此参考坐标系。

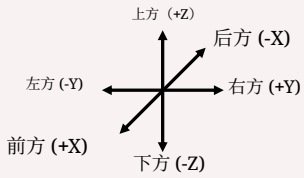


图 S8. 用于为 Ego4D 数据集生成视觉语言模型标注的提示。

你是一位精准的人类 egovideo 视频标注员。

尝试执行的任务是：*[任务指令]*

你将收到：
• 视频片段的起始帧和结束帧。
• 一系列视频标注。

你的任务：
为视频中展示的任务编写一条简短的祈使句指令。

每条标注都必须包含以下内容：
1) 在被操纵物体可见时提及它（例如“红色方块”“马克杯”“螺丝”“锅盖”“工具”）。如果不确定，请使用通用名词，例如“物体”“方块”“手柄”或“容器”。
2) 使用一到两个动词描述动作。
3) 无需提及执行者（例如人类、操作员）。

将视频标注作为先验。

风格和限制：
• 不要对条目编号，不要添加评论，也不要使用代码围栏。

输出格式：
仅返回一个字符串，并使用单引号，例如，'instruction'。

[起始帧]
[结束帧]

[视觉语言模型生成的标注 1]
[视觉语言模型生成的标注 2]
.
.
.
[视觉语言模型生成的标注 K]

图 S9. 用于为 Ego4D 数据集中的视频片段生成片段指令的提示词。

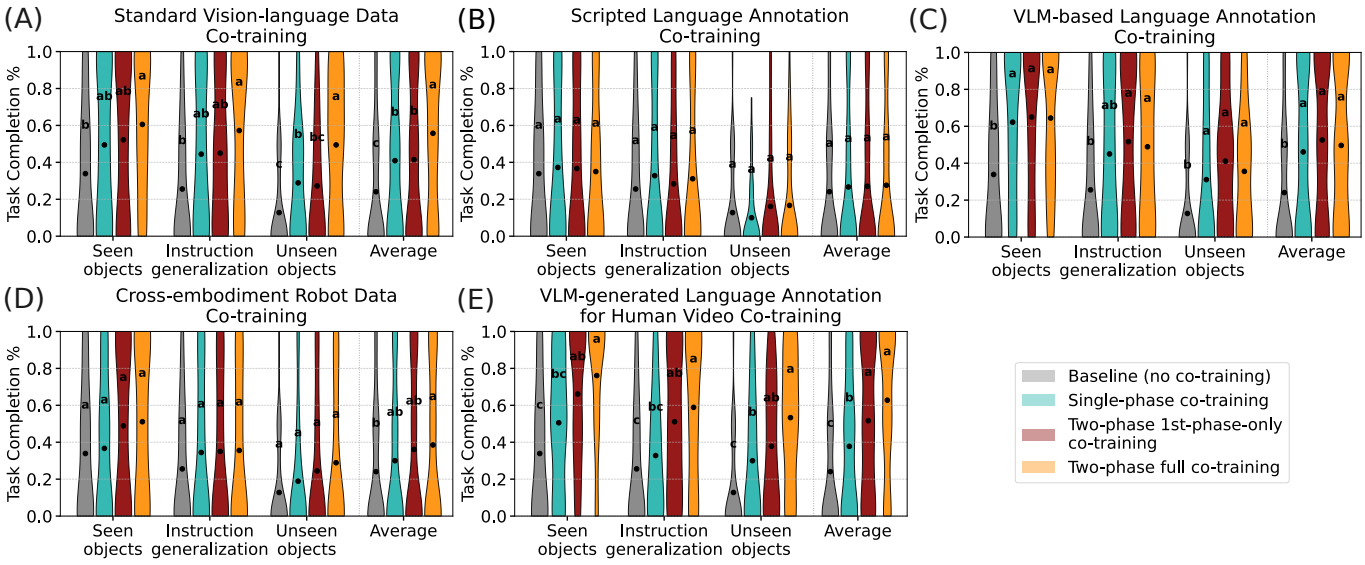


图 S10. 真实世界中协同训练数据来源和策略消融的完整分布图（图 5）。

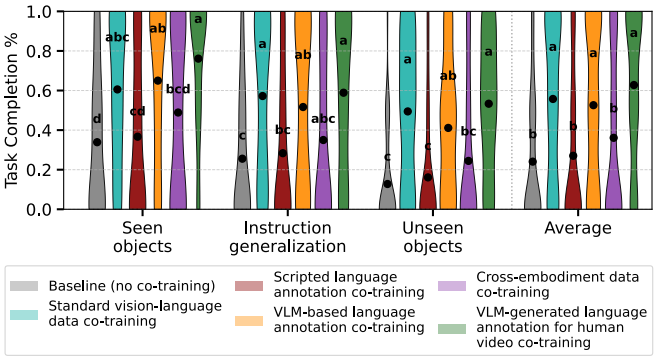


图 S11. 使用适用于所有有效协同训练数据模态的最佳协同训练策略训练的策略在真实世界中的性能完整分布图（图 9B）。

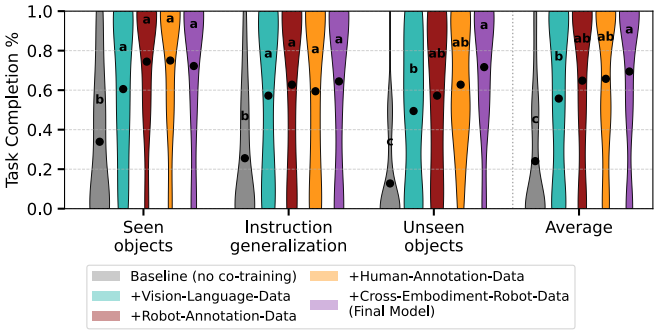


图 S12. 将所有有效数据模态结合进行协同训练的策略在真实世界中的性能完整分布图（图 10B）。

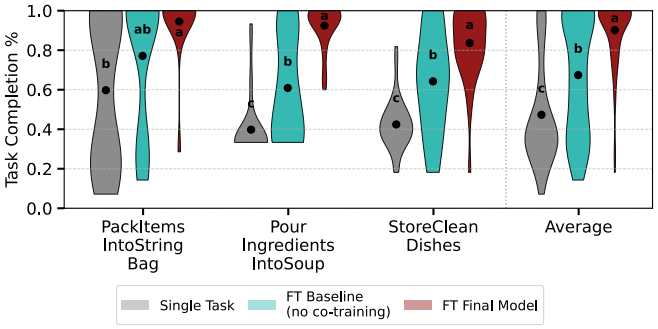


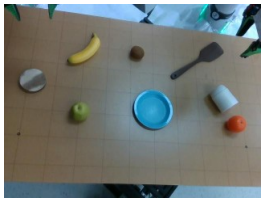
图 S13. 通过微调适应未见过的长时程灵巧操作任务的完整分布图（图 11）。



图 S14. 我们在仿真基准中用于评估的 13个已见任务和 8个未见任务示意图，涵盖标称条件和分布偏移条件。

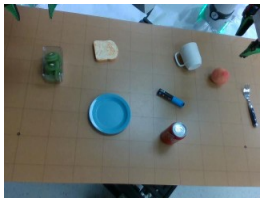
语言跟随

已见物体与指令泛化



已见物体1.
pick up the banana an d pl ace it on the plate 2. 拿起苹果并将其放在杯垫上 3. 拿起猕猴桃并将其放在锅铲旁边

指令泛化： 1. 抬起黄色水果并将其放到盘子上 2. 拿起青苹果，然后将其放在杯垫上 3. 拿起猕猴桃并将其放在锅铲旁边



已见物体1.
拿起猕猴桃片并将其放到盘子上 2. 拿起笔并将其放在叉子旁边 3. 拿起马克杯并将其放在面包旁边

指令泛化： 1.
move the sli物体o f 绿色 f ruit并将它们放在盘子上 2. 拿起书写工具，然后将其放在叉子旁边 3. 移动马克杯，使其靠近面包



已见物体1.
拿起粉色杯子并将其放在胡萝卜片盒上 2. 拿起牛油果并将其放进垃圾桶 3. 拿起梨并将其放在烤盘上

指令泛化： 1.
retrieve th杯子 that is pink
并将其放到胡萝卜盒上。2. 拿起看起来过熟的黑色水果，把它放到垃圾桶里。3. 烘焙前，拿起梨并将其放在平底锅上



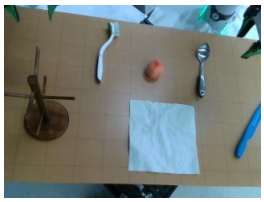
已见物体1.
拿起酒杯并将其放在麦片盒上。2. 拿起罐头并将其放入垃圾桶。3. 拿起清洁工具并将其放在面包旁边

指令泛化： 1. t
ake the fragile object used
葡萄酒，并将其搁在麦片容器上方 2. 拿起小型金属圆柱形物体，将它丢进垃圾桶 3. 拿起马桶刷，将它放在烘焙食品旁边



已见物体1.
pick up th
胡萝卜片，并将它们放入碗中 2. 拿起婴儿杯，将它放在烘焙纸上 3. 拿起奶油，将它放在清洁工具旁边

指令泛化： 1. 把蔬菜拿起来放进碗里 2. 拿起儿童用小容器并放在烹饪纸上 3. 保持拿着软管，并把它放在清洁工具旁边



已见物体1.
pick up the peac h and place it
在沥水架旁边 2. 拿起刷子并将其放在毛巾上 3. 拿起勺子并将其放在刀旁边

指令泛化： 1. 抬起水果并将其放在木制结构旁边 2. 抓取用于擦洗的工具并将其放在毛巾上方 3. 抓取用于舀取的器具并将其放在切割工具旁边



已见物体1.
pick up the t
ong 并将其放在刷子旁边 2. 拿起红甜椒并将其放入碗中 3.拿起胡椒瓶并将其放入垃圾桶

指令泛化： 1. 抬起金属夹持工具，并将其放在带刷毛的物品旁边 2. 将食物移入碗中 3. 拿起香料罐，并将其妥善丢进垃圾桶



已见物体1.
pick up the k
刀并将它放在厨房用具上 2. 拿起草莓并将它放在毛巾上 3. 拿起胡萝卜并将它放在清洁工具旁边

指令泛化： 1. 抬起 th
e cutting tool and rest
将它放在餐具上方。2. 保持柔软的红色浆果，并将其放在毛巾上晾干。3. 拿起尖端渐细的蔬菜，放在清洁设备旁边



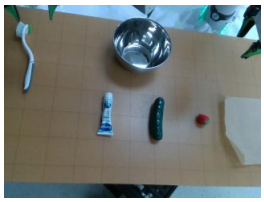
已见物体1.
pick up th
拿起土豆并将其放在毛巾上。2. 拿起吸管并将其放在绳子旁边。3. 拿起奶酪容器并将其放在刀旁边

指令泛化： 1. 抬起蔬菜并将其放在布上 2. 拿起细塑料管并将其放在绳子附近 3. 将塑料盒移到刀刃旁边



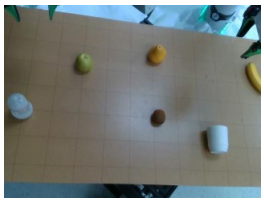
已见物体1.
拿起土豆并将其放在麦片盒上 2. 拿起香料罐并将其放在纸上 3. 拿起刷子并将其放在笔旁边

指令泛化： 1. 拿起土豆形物品并将其放在玉米片盒上 2. 拿起瓶子后，小心地将它放在纸上 3. 抬起用于刷东西的物体，并将它放在书写工具旁边



已见物体1.
pick up the b
拿起它并放在奶油旁边 2. 拿起黄瓜并将它放进碗里 3. 拿起草莓并将它放在烘焙纸上

指令泛化： 1. 抓取带有软毛的手柄，并将其放在奶油管旁边 2. 拿起新鲜蔬菜并将其移入碗中 3. 保持多汁红浆果，并将其放在羊皮纸层上



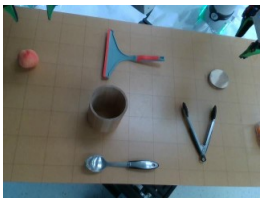
已见物体1.
拿起婴儿杯并将其放在苹果旁边 2. 拿起猕猴桃并将其放在梨旁边 3. 拿起马克杯并将其放在香蕉旁边

指令泛化： 1. 拿起婴儿用杯并放在苹果旁边 2. 拿起猕猴桃形状物品并放在梨旁边 3. 拿起饮水杯并移到香蕉旁边



已见物体1.
拿起牛油果并放入碗中 2. 拿起刀并放在面包片上 3. 拿起胡椒瓶并放入平底锅中

指令泛化： 1.
pick up the overr ipe fruit
并将其丢进白色容器 2. 保持切片工具，并将其放在面包片上方 3. 抓取香料容器，并将其放在金属长方形物体中



已见物体1.
拿起桃子并将其放入厨具筒中 2. 拿起夹子并将其放在清洁工具旁边 3. 拿起胡萝卜片并将它们放在杯垫上

指令泛化： 1.
pick up the swee t fruit with
粉红色外皮并将其放入餐具罐 2. 取起类似钳子的物体并将其放在清洁设备旁边 3. 将装有橙色薄片食物的容器移到饮料垫上



已见物体1.
拿起猕猴桃片并将其放在毛巾上 2. 拿起餐具并将其放在吸管旁边 3. 拿起罐头并将其放在粉色杯子旁边

指令泛化： 1. 抬起细长的绿色片状物，并将其放置在织物下方 2. 拿起锅铲，将它放在细长中空管旁边 3. 抓起洋水罐，将它移到粉色饮水容器附近

图 S15。用于真实世界语言实验、已见物体和指令泛化设置的所有初始条件与指令。对于每个初始条件，编号标签（例如 1.）表示已见物体评估中使用的对应指令，以及其在指令泛化评估中使用的替代表述。

语言跟随

未见物体

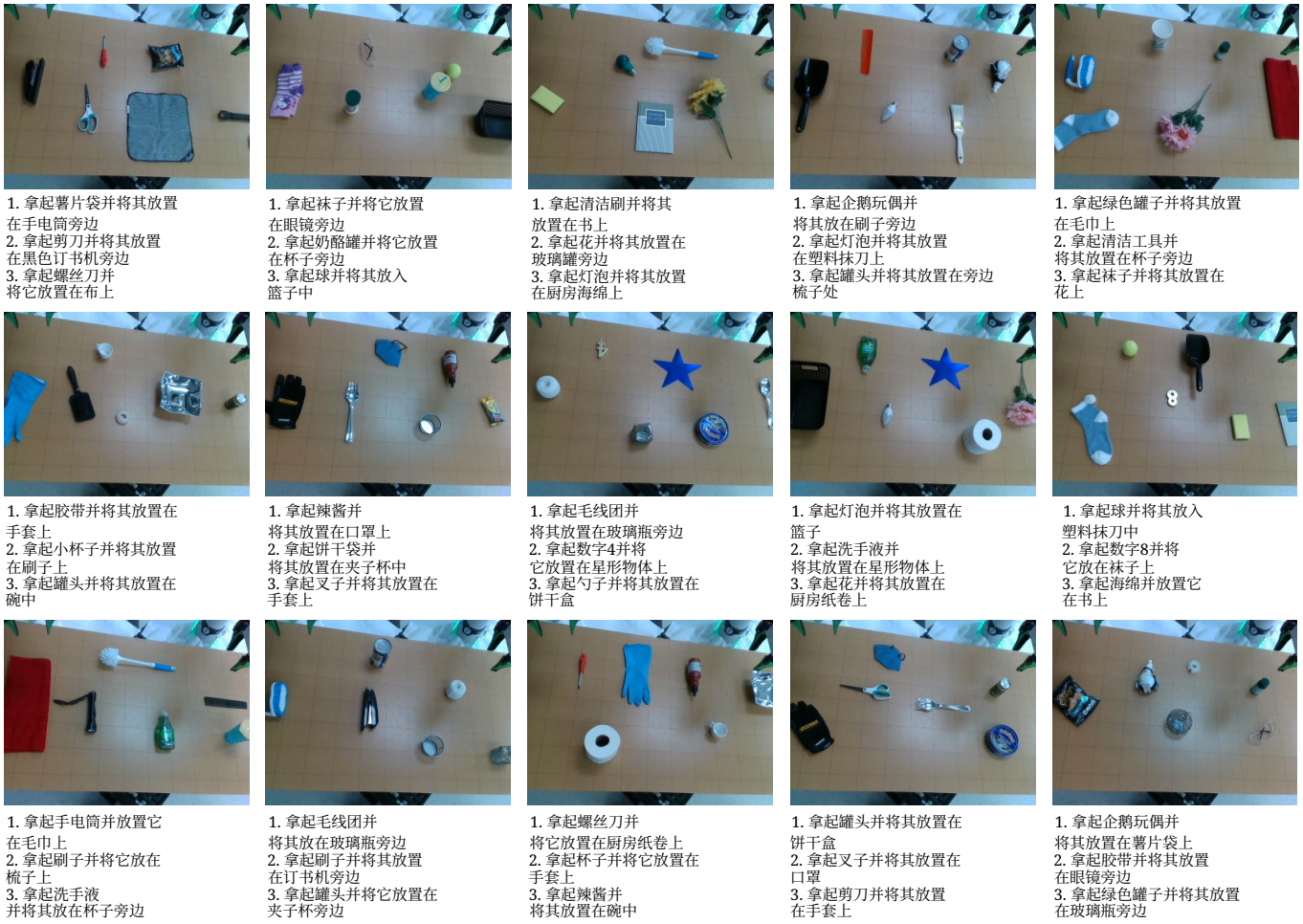


图 S16。 所有初 始条件和指令，用于未见物体设置下的真实世界语言跟随实验 g.

长时域灵巧任务

PackItemsIntoStringBag



PourIngredientsIntoSoup



StoreCleanDishes



图 S17。三个真实世界未见物体、长时程和灵巧任务的初始条件示例。

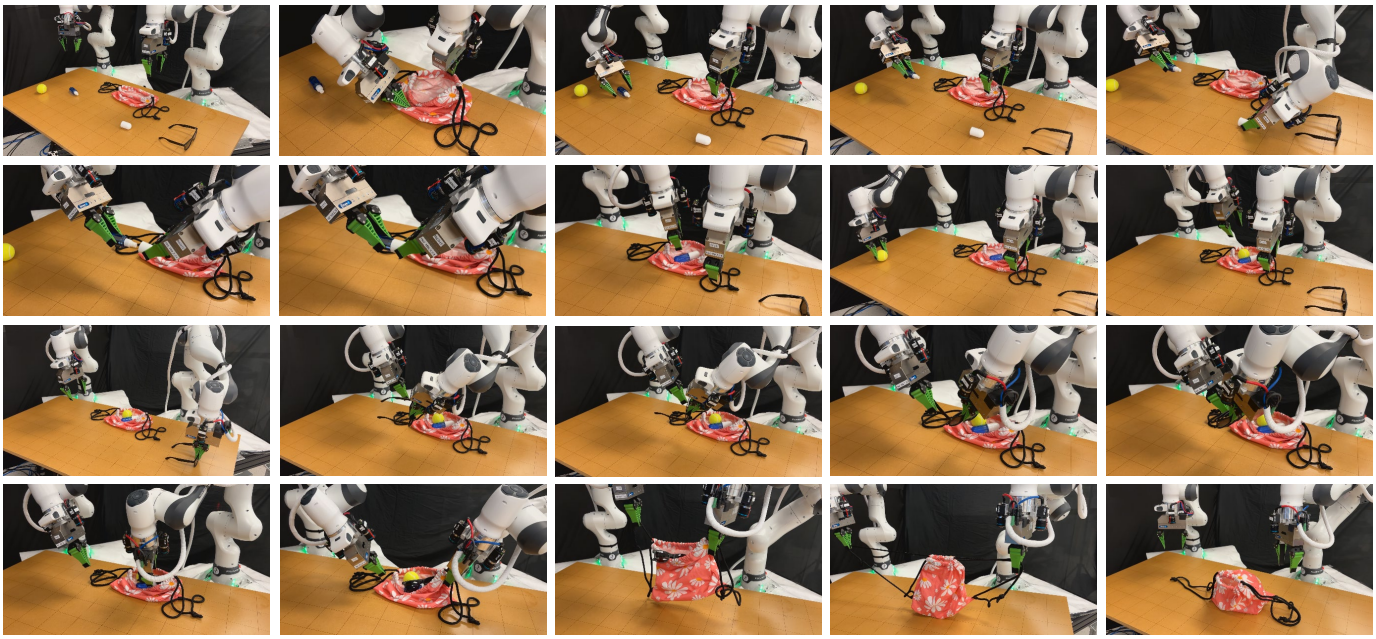


图 S18. 展示机器人执行 *PackItemsIntoStringBag* 任务的视频截图。目标是打开抽绳袋，将所有物品（瓶子、网球和太阳镜）装入袋中，并拉动袋口抽绳将其封好。机器人需要先给瓶子盖上瓶盖，并折叠太阳镜，然后再将它们装入袋中。

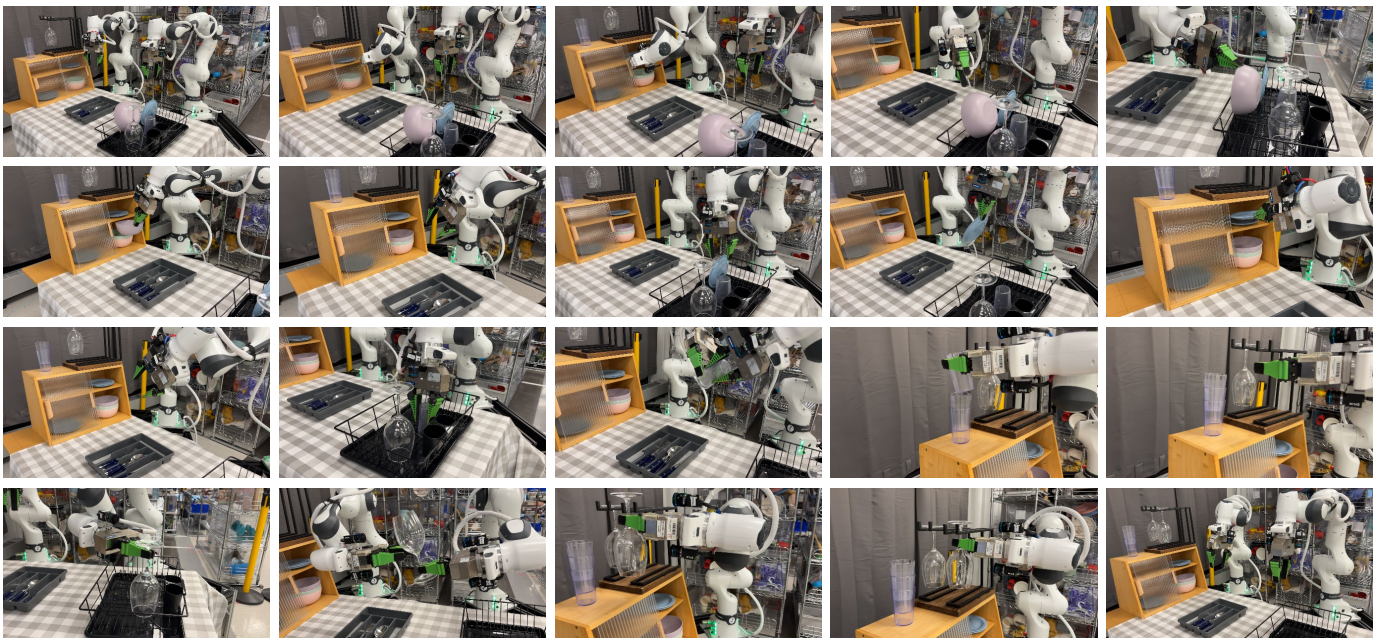


图 S19. 展示机器人执行 *StoreCleanDishes* 任务的视频截图。目标是将沥水架上的所有物品（碗、盘子、杯子和酒杯）收进餐具柜。机器人必须将碗放置在碗堆上，将盘子放置在盘子堆上，将杯子叠放在其他杯子上方，并将酒杯插入玻璃杯架。

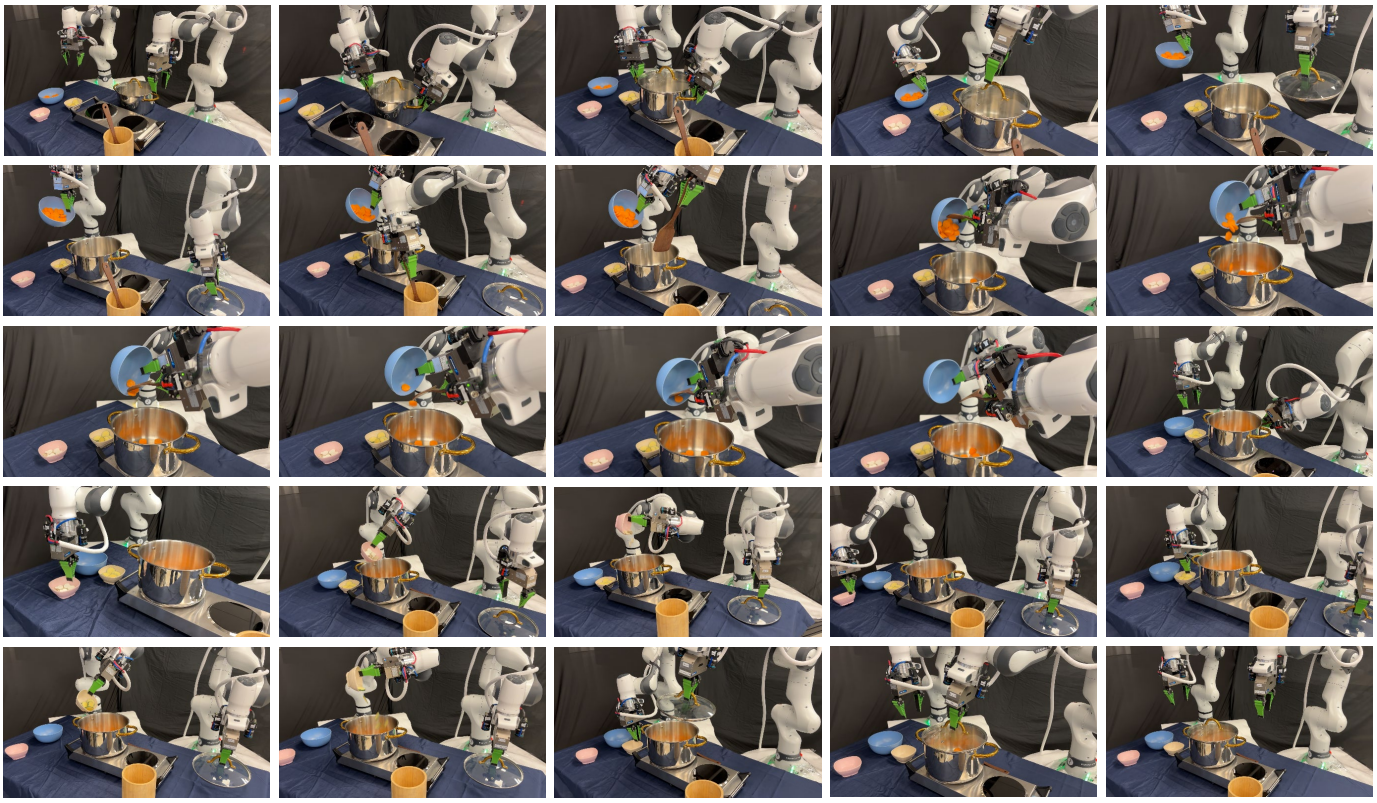


图 S20. 展示机器人执行*PourIngredientsIntoSoup* 任务的视频截图。目标是将所有食材（切碎的胡萝卜、蘑菇和黄瓜）倒入汤锅中。倒入之前，机器人必须先将锅放到炉子上并取下锅盖。机器人必须用锅铲舀起所有胡萝卜片。