arXiv:2601.01618v1 [cs.RO] 4 Jan 2026

Action-Sketcher：通过视觉草图实现从推理到动作的长时程机器人操作

Huajie Tan[1, 2, *, †], Peterson Co[1, 2, *], Yijie Xu[2, 3, *], Shanyu Rong[1,2], Yuheng Ji[2,4], Cheng Chi[2], Xiansheng Chen[2], Qiongyu Zhang[3], Zhongxia Zhao[1,2], Pengwei Wang[2, †], Zhongyuan Wang[2], Shanghang Zhang[1, 2, ✉]

[1] 多媒体信息处理国家重点实验室，计算机科学学院，北京大学
[2] 北京智源人工智能研究院，[3] 悉尼大学，[4] 自动化研究所，中国科学院

项目网站：https://action-sketcher.github.io

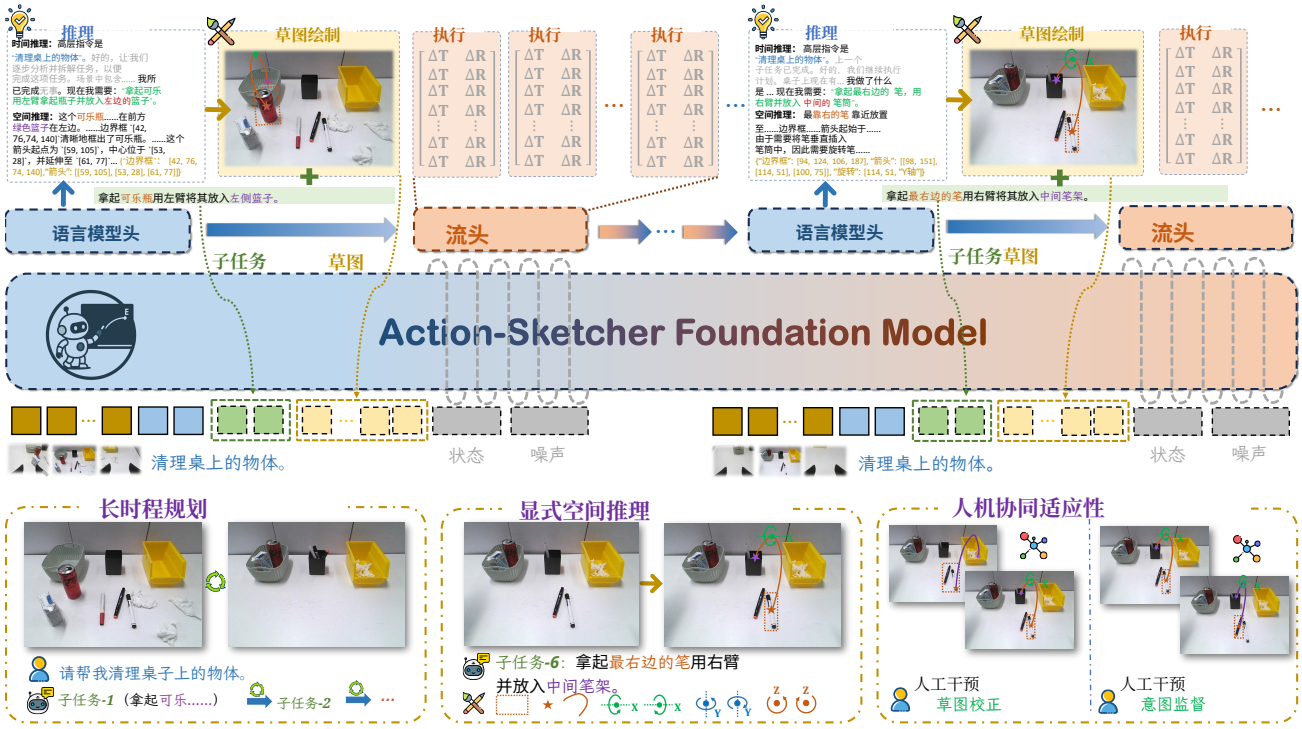


图 1. **Action-Sketcher 概览**。我们的框架采用看-想-画-做循环：基础模型首先进行时间和空间推理，将高层指令（例如“清理桌上的物体”）分解为一个子任务及其对应的视觉草图。草图由点、框和箭头等基本元素构成，作为明确且人类可读的计划，引导低层策略生成稳健的动作序列。该方法具备三项关键能力：（左下）通过任务分解实现长时程规划，（中下）通过将指令与场景几何进行指代定位来实现显式空间推理，以及（右下）通过直接校正草图和意图监督实现无缝的人在回路适应性。

摘要

长时程机器人操作对于真实世界部署日益重要，这要求进行空间消歧——复杂布局中的指代消歧能力，以及动态交互下的时间韧性。然而，现有的端到端和分层视觉-语言-动作（*VLA*）策略通常依赖纯文本线索，同时将计划意图隐而不显，这削弱了模型在杂乱或信息不充分的场景中的指称落地能力，妨碍其通过闭环交互有效分解长时程目标，并

* 贡献相同。
† 项目负责人。✉ 通讯作者：shanghang@pku.edu.cn

且由于行动选择背后的推理依据不明，因果解释能力也受到限制。为解决这些问题，我们首先提出视觉草图，这是一种合理的视觉中间表示，可在机器人当前视图中呈现点、框、箭头和类型化关系，以显化空间意图、将语言与场景几何联系起来，并在人类可验证的高层推理与底层控制之间搭建桥梁。在视觉草图的基础上，我们提出***Action-Sketcher***，一种按照循环式"观察 → 思考 → 绘制草图→ 行动"工作流运行的视觉-语言-动作框架，并通过自适应的令牌门控策略来协调推理触发、草图修订和动作发出，从而支持反应式修正与人机交互，同时保持实时动作预测。为实现可扩展的训练与评估，我们整理了一个多样化语料库，其中包含交错图像、文本、视觉草图监督和动作序列；并采用多阶段课程训练方案训练 *Action-Sketcher*，该方案结合了用于模态统一的交错序列对齐、用于精确语言落地的语言到草图一致性，以及通过草图到动作强化学习增强鲁棒性的模仿学习。在杂乱场景和多物体任务上进行的大量仿真与真实世界任务实验表明，该方法提升了长时域成功率，增强了对动态场景变化的鲁棒性，并借助可编辑草图和逐步计划提高了可解释性。项目网站：*Action-Sketcher*。

1. 引言

机器人操作正从短时单步原语迈向长时程、开放世界任务，在这些任务中，目标、布局和人类偏好会随时间变化。在此类情境下，智能体不仅必须感知并行动，还必须维持一条可解释的决策链，使其在空间关系和时间计划发生变化时依然可靠。

然而，要实现这种可靠性，仍受制于两个主要瓶颈。（***1***）在空间层面，从语言到动作的指代定位十分脆弱：自然语言指令往往含糊不清（例如，在有多个杯子的场景中"把茶倒进杯子里"），或信息不充分（例如，"把书放在杯子左边"，但没有明确具体目标姿态）。为消除指称不确定性，并将语言关系转化为可执行约束，我们提倡使用显式视觉草图（即锚定于场景几何的区域高亮、关键点和关系箭头）。（***2***）在时间层面，人在回路协调仍然薄弱：实时交互有限，且可解释规划过程产物鲜少呈现，导致小错误不断累积而未被察觉。通过逐步推理依据明确呈现时间结构，可以及早检测错误并高效恢复。总之，这些挑战表明，深度推理与显式视觉草图相结合，对于在动态环境中实现可靠操作至关重要。

近期的视觉-语言-动作（VLA）模型 [7–9, 19, 30, 52, 74] 在将观测和语言直接映射为动作方面取得了显著进展。然而，由于计划意图大多嵌入在潜在表征中，这些模型难以进行任务分解和因果解释，而这两者对于动态场景中的长时程运行都至关重要。分层视觉-语言-动作模型[61, 65, 77]尝试通过规划器和控制器系统应对这些挑战，但其推理往往局限于瞬时，缺乏对全局意图（例如，不断变化的人类目标、新出现的错误和先前状态）的持续建模；此外，它们对指称消歧的支持有限，导致杂乱环境中的空间指代无法明确。另一种互补思路是在视觉-语言-动作模型中引入先思考后行动推理，例如 EO-1[62]通过交错式视觉、思考与动作预训练，以及 OneTwoVLA [44] 通过自适应联合推理与行动；但当中间证据仍仅以文本形式呈现时，动作背后的空间指称消歧（接触点、接近方向、物体关系）仍然隐含，这既妨碍人工核验，也使控制器无法获得低熵的几何引导。

为解决这些问题，我们在语言到控制的接口处显式表达空间意图，引入一种视觉中间表征，称为视觉草图。它是在一个或多个当前视图上呈现的一组稀疏、面向动作的基本元素：点（关键接触位置或目标位置）、边界框（物体或区域选择），以及箭头（期望的运动方向或力的方向）。将语言与这些几何草图共同锚定，可以消除指称对象和空间约束方面的歧义；通过简单标注或自动标注提升可监督性；并增强可解释性和可调试性，因为人们可以在执行前阅读、批准或修改意图。实际上，草图充当高层推理与低层控制之间的可验证契约。

基于视觉草图，我们提出***Action-Sketcher***，这是一个运行于查看 →思考 →草图 → 动作循环中的视觉-语言-动作框架，如图1所示。给定多视角观测和任务指令，模型在以下三者之间交替进行：(1)思考，生成包含时间规划和空间构想的简洁推理依据；(2)草图，输出针对场景和子任务定制的视觉草图（即点、框和箭头）；以及 (3)动作，根据当前观测合成低延迟的动作块。为实现自适应切换和事件感知控制，自适应令牌门控策略负责协调该循环，例如，`<BOR>`（推理开始）启动或刷新全局推理，而 `<BOA>`（动作开始）触发动作块预测。策略根据观测状态、预测风险和用户反馈在这些模式之间进行切换，从而在保持实时

动作预测。为支持可扩展的训练和评估，我们构建了一个语料库，将交错图像、文本和动作序列与跨多种操作技能、杂乱桌面及多物体场景的视觉草图监督相结合。我们的多阶段课程训练方案包括：(i) 基础时空学习，用于模态统一；(ii)语言到草图的一致性，以确保短语到几何的精确绑定（例如，“抓住把手” → 把手关键点 + 视觉草图中的接近箭头）；以及 (iii) 结合草图到动作强化的模仿学习，以增强鲁棒性。本文的贡献总结如下：

- 我们将视觉草图形式化为一种共同锚定的显式空间意图接口，通过绘制点、框和箭头，明确在哪里以及如何执行动作，并作为高层推理与低层控制之间可由人类验证的契约。
- 我们提出***Action-Sketcher***，这是一种视觉-语言-动作框架，按照观察 → 思考 → 绘制草图 → 执行动作的循环运行，并由令牌门控状态进行编排，以便在推理、视觉草图生成/修订和动作合成之间自适应切换，从而实时处理打断、检测错误并在草图层面进行纠正，而不牺牲延迟表现。
- 我们构建了交错式语料库和训练方案，通过交错序列对齐、语言与视觉草图的一致性，以及结合草图到动作强化的模仿学习，使语言、视觉草图和动作相互对齐；实验表明，所得系统提升了长时域成功率、应对动态场景变化的鲁棒性和可解释性，并通过全面的消融实验和真实机器人研究进行了验证。

2. 相关工作

视觉-语言-动作模型（VLAs）。VLAs 基于多模态大语言模型骨干网络，将多模态感知与底层控制连接起来。端到端策略采用基于回归的 [10, 24, 30, 36, 39, 50], 基于扩散的 [9, 40–42, 52, 75], 或混合映射 [51]，将观测和语言映射为动作，虽在短时程任务中表现出色，却往往将计划意图保留为潜在信息，从而限制了任务分解和因果解释。分层 VLAs [8, 19, 49, 61, 65, 74, 85] 引入规划器-控制器拆分，以改善长时程行为；然而，推理往往是即时或局部的，缺乏对全局意图的持续建模（例如不断演变的人类目标、新出现的错误和先前状态），这妨碍了人在回路交互，以及动态场景中的指称消歧。近期的先思考再行动变体在统一的骨干网络中整合了显式推理，例如 EO-1 [62] 和 ONETWOVLA [44]。THINKACT [25]进一步将文本计划与强化的视觉计划潜变量结合起来，以此对下游动作模型进行条件化，改善了长时程规划；但计划最终仍被压缩为一种无法由人类编辑的潜变量，限制了执行过程中意图的持续性和可验证性。因此，我们需要一种能够使意图可观察、可执行的接口。

机器人视觉提示、草图与轨迹。视觉提示通过绘制在图像平面上的几何或图示线索来编码动作意图。在表征层面，LLARVA [58] 预测二维轨迹，以对齐视觉空间和动作空间；ROBOBRAIN [28,73] 则为下游策略提供明确的视觉引导（例如 指向、可供性、轨迹）。在策略层面，RT-TRAJECTORY [22] 以粗略的轨迹草图为条件，RT-SKETCH [69] 利用手绘目标草图来应对歧义和干扰项，TRACEVLA [84]注入视觉轨迹提示，而 ROVI [37]则系统化地运用以物体为中心的符号（箭头、圆圈、颜色、数字）。与此同时，MOLMOACT [34] 生成具备深度感知能力的标记，以及作为可编辑轨迹的中层空间规划，以引导动作。尽管这些工作展现了明确视觉线索的价值，但它们要么将提示固定为静态输入，要么将计划压缩为不可编辑的潜在表示。相比之下，我们的视觉草图具有持久性、可由人类编辑，并且每次仅针对一个子任务生成，从而提供有针对性且与场景对齐的引导。由于草图不再编码完整轨迹，因此能够表达更丰富、更精确的动作基元，例如接触关键点、旋转箭头和放置提示，而不是粗略的轨迹。

3. 方法

我们提出的框架 Action-Sketcher 旨在应对长时程操作任务中的空间歧义和时间脆弱性。通过引入显式的视觉中间表示，我们在高层推理与低层动作执行之间建立了一个可验证、可纠正的接口。本节首先在第 视觉草图节3.1中详细介绍所提出的 *Action–Sketcher*框架的结构与语义。随后，我们将在第3.2节介绍完整的 Action–Sketcher 框架，包括问题形式化、双模式推理流程以及多阶段课程学习策略。

3.1. 视觉草图：显式空间意图

在任务执行过程中，智能体会接收多视角观测和任务指令。为增强在不同平台和场景间的广泛适应性，我们采用机器人的自我视角摄像头作为规范参考坐标系，并在其中表达视觉草图；所有基本元素均定义在自我视角图像平面上。在每一步 t，视觉草图 都是由几何基元构成的稀疏元组，用于明确表示空间意图：

$$\mathcal{S}_t = \big(\mathcal{B}_t,\ \mathcal{P}_t,\ \mathcal{A}_t\big), \tag{1}$$

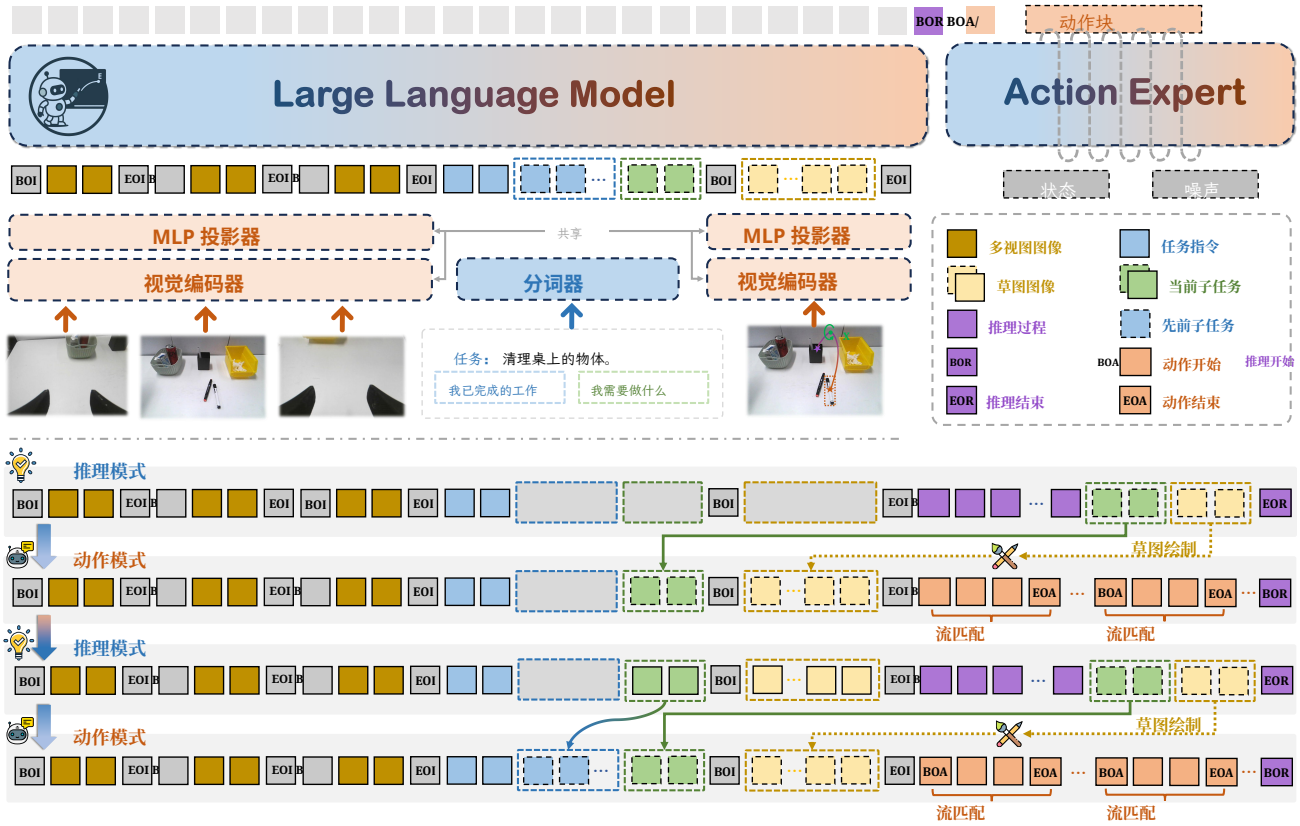


图 2. **Action–Sketcher 概览**。模型运行一个事件驱动循环：(i) 概述下一个子任务；(ii) 输出一个简洁的视觉草图（点、框、箭头、关系），以显式呈现空间意图；(iii) 根据该草图和机器人状态合成一个动作块。这种显式中间表示支持针对性监督、即时纠正，并能在单模型架构中实现可靠的长时程执行。

其中， $\mathcal{B}_t$ 通过物体级边界框指定目标区域， $\mathcal{P}_t$ 汇集对任务完成度至关重要的关键点，而 $\mathcal{A}_t$ 则通过在 $\mathcal{P}_t$中的关键点上定义的箭头来指定预期动作。每种基本元素的定义如下：

边界框($\mathcal{B}_t$)。边界框充当物体级的可供性线索，通过划定机器人可以操作或交互的目标物体区域，为任务提供指代定位。这能消除杂乱场景中的物体指代歧义，并在保留必要的尺度和位置关系信息的同时，抽象掉细粒度的外观信息：

$$\mathcal{B}_t = \{b_i\}_{i=1}^{N_b}, \qquad b_i = (x_{1,i}, y_{1,i}, x_{2,i}, y_{2,i}), \tag{2}$$

其参数由左上角（$x_{1,i}$,$y_{1,i}$）和右下角（$x_{2,i}$,$y_{2,i}$）的像素坐标确定。例如，给定指令“拿起离杯子最近的物品，”通过解析空间关系，可以生成边界框来明确指定目标物体（例如一个苹果）。

点 ($\mathcal{P}_t$)。我们使用关键点来指定机器人应与物体交互或相对于物体移动的精确位置。这些点可以表示部件级的可供性、运动航点或几何参考点。形式上，这组点表示为：

$$\mathcal{P}_t = \{p_i\}_{i=1}^{N_p}, \qquad p_i = (x_i, y_i), \tag{3}$$

其中（x_i, y_i）是自我视角图像平面上的像素坐标。以“倒茶”为例，我们可以用 p_{spout}表示茶壶壶嘴尖端，用p_{cup}表示杯子的中心，用 p_{handle}表示手柄上的一个稳定接触点。首先，为了“提起茶壶”，动作策略可以将p_{handle} 作为稳固的抓取锚点，表示部件级可供性。提起后，为了“将茶壶移向杯子”，点对（$p_{\text{spout}}, p_{\text{cup}}$）可作为用于对齐的几何参考，其中 p_{spout} 是起始路径点， p_{cup}是目标点。这些关键点锚点使模型能够理解并执行精确的几何子任务。

箭头 ($\mathcal{A}_t$)。箭头连接静态关键点与驱动，发挥动态纽带的作用。我们认为，复杂的 $SE(3)$ 操作可以有效分解为投影到二维图像平面上的平移轨迹和旋转线索。以关键点 $\mathcal{P}_t$ 为锚点，箭头集合 $\mathcal{A}_t$ 将这些预期运动编码为以下内容的并集：

两个子集： $\mathcal{A}_t = \mathcal{A}_t^{\text{trans}} \cup \mathcal{A}_t^{\text{rot}}$。 (*i*) 平移箭头编码末端执行器的预期轨迹：

$$\mathcal{A}_t^{\text{trans}} = \{a_i^{\text{trans}}\}_{i=1}^{N_{trans}}, \quad a_i^{\text{trans}} = \left(p_i^{\text{start}}, p_i^{\text{end}}\right), \quad (4)$$

其中每个箭头都是锚定在 $\mathcal{P}_t$ 中关键点上的有序序列。在“倒茶”示例中，主箭头（$p_{\text{spout}} \to p_{\text{cup}}$）引导茶壶壶嘴移向杯子中心。随后，如果观察到偏移，可以生成短的纠正箭头，使壶嘴重新居中。倒茶后，回撤箭头将壶嘴移开，以防滴漏。*(ii)* 旋转箭头指定绕规范轴的旋转：

$$\mathcal{A}_t^{\text{rot}} = \{a_i^{\text{rot}}\}_{i=1}^{N_{rot}}, \quad (5)$$

$$a_i^{\text{rot}} = \left(p_i,\ \text{axis} \in \{x, y, z\},\ \text{dir} \in \{\circlearrowleft, \circlearrowright\}\right), \quad (6)$$

其中每个箭头都定义了以 $p_i \in \mathcal{P}_t$ 为中心、绕规范轴（x， y， z）并沿指定方向（顺时针 $\circlearrowleft$ 或逆时针 $\circlearrowright$）进行的旋转。在我们的“倒茶”示例中，以 p_{handle} 附近为中心、绕自我视角 x轴（水平轴）的旋转箭头会指示倒茶所需的倾斜动作。随后进行反向旋转，使茶壶恢复水平。类似地，绕 z轴（垂直于图像平面）旋转，可以调整壶嘴的偏航方向。

3.2. 动作草图框架

我们的框架将长时程操作建模为混合输出空间（即离散和连续）上的序列建模任务。我们的目标是学习一种策略 π_θ，将上下文映射为联合输出分布，使智能体能够无缝交替进行离散推理（用于规划和绘制草图）与连续动作预测，并自主决定每个时间步采用何种模式。

3.2.1. 观察-思考-草图-执行流程

如图2所示，模型的输入上下文是一系列标记，表示多视图图像（例如，左腕、右腕、基座）、任务指令、已完成子任务的历史记录、当前子任务以及视觉草图图像。Action-Sketcher 以两种不同的模式运行，并可自适应切换： 推理模式和动作模式。这种双模式设计使模型能够动态平衡审慎思考的需求与实时执行的要求。模式切换由标记门控机制控制。

（***1***）推理模式。如果模型确定需要进行推理（例如完成一个子任务后、遇到错误或收到人工干预后），它会生成一个特殊标记 <BOR> （推理开始）。随后，模型以自回归方式进行时间推理和空间推理。时间推理：模型首先分析当前场景，结合整体任务指令和已完成子任务的历史记录，推断下一个合乎逻辑的子任务。空间推理：根据新确定的子任务，模型对场景中的物体布局及其关系进行空间推理。此过程将生成与该子任务对应的视觉草图文本元素（点、框、箭头）。此推理阶段以生成 <EOR> （推理结束）标记告终。随后，将生成的文本草图绘制到当前参考视图上，创建图像式视觉草图。接着，使用新的子任务和新的草图图像更新输入上下文。

(2) 动作模式。 相反，如果模型认为无需推理，例如在场景保持一致时例行执行子任务，它会生成一个 <BOA> （动作开始）标记。这会触发动作专家通过流匹配 [45]生成动作片段。

最初，已完成子任务的上下文标记、当前子任务和视觉草图图像均为空置。模型必须从推理模式 开始，以填充这些字段。随后，它可以在不同模式间灵活切换，以兼顾准确性和实时性能。

3.2.2. 模型结构

Action-Sketcher框架与模型无关，可与任何视觉-语言-动作模型集成。作为一个具体实例，我们采用 π_0 [9]作为骨干网络。模型经过训练，能够自回归地生成文本推理链、子任务计划和视觉草图的结构化描述，同时优化流匹配损失 [45] 以预测连续动作片段。模型结构的详细信息请参阅附录 A.1。

3.2.3. 训练策略

为实现“看-思考-草图-行动”流程，我们采用多阶段课程学习策略。该课程旨在首先建立广泛的时空理解基础（阶段 1），随后将这些知识专门用于使高层语言与结构化的视觉草图表示相匹配（阶段 2），最后联合训练动作策略和自适应模式切换机制，以实现稳健的真实世界执行（阶段 3）。下面将详细介绍各个阶段。

阶段 1：基础时空学习。 初始阶段着重于对模型进行预训练，使其具备通用的时空建模与推理能力。为此，我们使用了一个大规模数据集。在空间理解方面，训练数据包含来自多种来源的 340 万个样本，包括用于物体级边界框预测的视觉定位数据集 [23, 28, 33, 35] 、用于关键点预测的空间指向数据集 [17, 73, 79, 82, 86]、用于更广泛场景感知的场景理解数据集 [4,54, 78] ， 以及通用视觉问答（VQA）数据集 [47, 48, 48]。在时间学习方面，我们通过整合数据构建了一个包含 87 万个序列的数据集

，数据来自 EgoPlan [15], 、ShareRobot [28],和 AgiBot-World [11]。为促进显式推理，我们为该语料库中 20% 的数据添加了由 GPT-4o 生成的详细文本推理依据 [27]。这一阶段使 Action-Sketcher 具备了强大的时空推理能力和指令遵循能力。

阶段 2：推理到草图增强。在此阶段，模型学习推理模式所需的连贯推理过程。目标是分析场景，并在任务指令和执行历史的引导下，依次进行时间与空间推理，生成下一个子任务及其对应的文本式视觉草图。训练数据经过整理，包含我们在真实环境中收集的 2.6k 个多样化长时程操作任务回合（每个任务包含 2–16 个子任务，布局复杂，涉及 20 多种物体及其各种空间排列），涵盖整理桌面、倒茶和一般抓取放置等任务。我们还通过标注现有数据集中的 1.7k 条完整轨迹来补充数据，其中包括 LIBERO [46]和 RoboTwin 2.0 [14]。最终得到一个包含 21k 个样本的推理到草图数据集，用于微调 Action-Sketcher，使其掌握完整的推理模式流程。

阶段 3：草图到动作与模式适应。 最后一个阶段对动作策略和模式切换机制进行联合训练。训练数据有两个用途。首先，通过提供示例来训练自适应模式选择：在适当情况下（例如，子任务边界处或场景发生变化时）使用 a <BOR>token，而在常规执行过程中使用 <BOA>token。其次，利用动作标注数据训练动作专家。为增强鲁棒性，我们对与每个动作标签对应的视觉草图进行增强，以模拟推理时可能出现的不准确情况。具体而言，对于边界框，我们对真实标注框施加随机扰动，同时确保交并比（IoU）至少为 0.8。对于点，我们在真实标注点周围的小半径圆形区域内随机采样新的坐标 c。箭头根据点进行调整。这种增强确保动作专家能够抵御生成的草图中的细微误差。

此外，我们还解决了任务执行中固有的显著数据不平衡问题：动作模式步骤的数量远远超过推理模式步骤。为防止模型偏向出现频率更高的 <BOA> token，我们采用模式平衡采样策略。令 D_R 和D_A 分别表示推理模式和动作模式的训练样本集合。从合并数据集 $D = D_R \cup D_A$ 中选取样本 d的概率 $P(d)$ 定义如下：

$$P(d) = \begin{cases} \frac{1}{2|D_R|} & \text{if } d \in D_R \\ \frac{1}{2|D_A|} & \text{if } d \in D_A \end{cases} \tag{7}$$

该方法直接缓解了数据不平衡，防止模型形成模式偏差并促进从而具备更强的模式选择能力。

4. 实验

我们在模拟环境和真实世界的机器人平台上开展了大量实验，以验证 Action-Sketcher 在长时程复杂操作任务中的有效性。我们的评估围绕三个关键研究问题展开：

• **研究问题1：** 在时间和空间复杂度各异的任务中，Action-Sketcher 的性能与最先进的基线模型相比如何？

• **研究问题2：** 我们系统的主要失败模式是什么？通过视觉草图 界面进行人工介入，能够在多大程度上提高成功率？

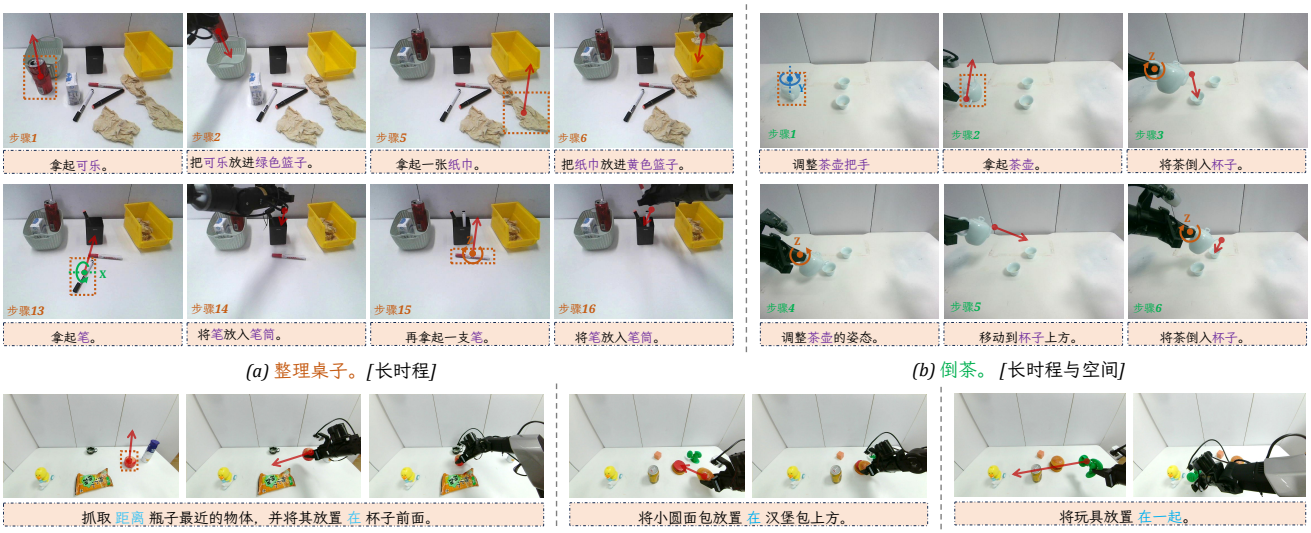
•**研究问题3：** Action-Sketcher 框架的核心组件如何影响其性能？

4.1. 任务性能评估（RQ1）

实验设置。 我们首先评估 Action-Sketcher 与一系列 VLA 基线模型相比的整体任务成功率。评估在两个仿真基准和一台真实世界机械臂上进行。在仿真中，我们使用 LIBERO [46]，（一个专注于终身技能的基准），以及 RoboTwin 2.0 [14] 的增强版本，该版本增加了物体杂乱程度和空间复杂度，包含“堆叠三个积木”、“悬挂马克杯”、“倒空杯子”和“将 *A2B* 放置在任意位置”等任务。在真实世界中，我们在三个长时程任务上进行评估：“整理杂乱的桌面”、“倒茶”和“通用抓取放置”，这些任务包含含糊指令，并基于 Agilex 和 Galaxea 双臂机器人平台构建。我们报告每个基准套件中所有任务的平均成功率。

基线。 我们将模型与几类视觉-语言-动作模型进行比较：端到端视觉-语言-动作模型（DP [16],Octo [75], OpenVLA [30]）、采用专用架构的视觉-语言-动作模型（SpatialVLA [63], π_0 [61], $\pi_{0.5}$ [61], OpenVLA-OFT [31]），以及融入视觉提示或中间表征的模型（TraceVLA [84],Molmo-ACT [34], PixelVLA [43]）。

结果。如表1所示，Action-Sketcher 在所有评估套件中均展现出优越的性能。值得注意的是，它在 LIBERO 基准的“Long”类别中优势尤为显著，这些类别专门用于测试长时程规划能力。这表明，与依赖潜在规划的方法相比，显式的看-思考-草图-行动循环具有显著优势。为了进一步分析长时程且空间结构复杂的任务，我们专门探究了模型在需要扩展规划和复杂空间推理的场景中的能力，并针对 RoboTwin 2. 环境中的四项具有挑战性的任务以及我们的三种真实世界设置进行了重点比较。这七项任务的选择是因为它们体现了我们的工作旨在解决的主要挑战0 [14]：长时程依赖，以及物体之间复杂且往往模糊的空间关系。



(a) 整理桌子。*[长时程]* *(b)* 倒茶。*[长时程与空间]*

(c) 通用抓取与放置。*[空间]*

图 3. **Action-Sketcher 的真实世界演示。** 长时程和空间操作任务的定性滚动执行结果。我们的方法生成明确的视觉草图（叠加的点、框和箭头），将高层推理落实为低层动作，并在杂乱环境中成功完成整理桌面、倒茶等任务。

表 1. LIBERO 基准上的成功率（%）。不同模型中的最佳结果以**粗体**突出显示，次佳结果以下划线标出。

方法	空间	物体	Goal	Long	Avg.
扩散策略 [16]	78.3	92.5	68.3	50.5	72.4
Octo [75]	78.9	85.7	84.6	51.1	75.1
OpenVLA [30]	84.7	88.4	79.2	53.7	76.5
SpatialVLA [63]	88.2	89.9	78.6	55.5	78.1
π_0 + FAST [60]	96.4	96.8	88.6	60.2	85.5
π_0 [61]	96.8	98.8	95.8	85.2	94.1
$\pi_{0.5}$ [61]	**98.8**	98.2	**98.0**	92.4	96.8
OpenVLA-OFT [31]	**97.6**	98.4	97.9	94.5	**97.1**
带视觉提示的视觉语言动作模型					
TraceVLA [84]	84.6	85.2	75.1	54.1	74.8
Molmo-ACT [34]	87.0	95.4	87.6	77.2	86.6
PixelVLA [43]	88.5	90.0	85.8	82.6	86.7
Action-Sketcher（我们的方法）	97.2	**99.6**	94.8	**96.0**	96.9

表2所示的结果毋庸置疑。尽管 $\pi_{0.5}$ [61] 和 OpenVLA-OFT [31]等强基线表现不俗，Action-Sketcher 仍在所有任务中展现出显著且稳定的性能优势。这一巨大的性能差距验证了我们的核心假设：对于计划跨度长且空间目标复杂的任务，仅在潜在空间中隐式嵌入意图是不够的。Action-Sketcher 显式的看-思考-草图-行动流程使智能体能够稳健地分解问题，将动作建立在可解释的空间表征之上，并以更高的精度执行

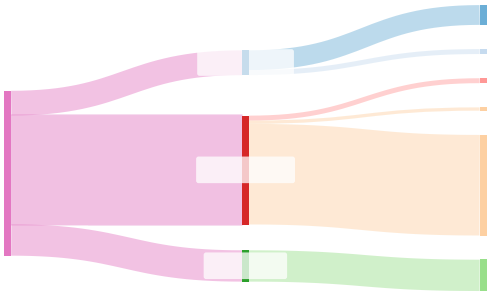


图4. Action-Sketcher 的失败分析。大多数错误发生在推理模式中，主要是由于视觉草图生成不准确。

确性，使其在这些具有挑战性的领域中表现出色。

4.2. 错误分析与人在回路（RQ2）

为更好地了解系统的局限性以及人类协作的潜力，我们分析了任务失败的根本原因，如图4所示。我们的分析发现，失败可归因于三个主要来源：模式切换错误（占失败总数的 12%）、推理模式中的错误（66%）以及动作模式中的失败（19%）。在推理模式中，绝大多数错误源于空间推理阶段，具体而言，是视觉草图的生成（占所有失败的 61%）。

这项分析表明，通过视觉草图 进行的空间定位不准确是主要瓶颈。然而，由于草图是一种明确且易于人类理解的交互界面，因此它为人工干预提供了一个自然的切入点。

表2。在选定的长时程和空间复杂任务上的成功率（%），将 Action-Sketcher 与顶尖基线方法进行比较。对于仿真得分，成功率基于任务完成度；而对于真实世界执行，成功率基于所有滚动执行中的平均子任务完成率。不同模型中的最佳结果以 **粗体** 突出显示，次优结果则加下划线。

方法	*RoboTwin 2.0*（仿真）					真实世界执行		
	堆叠积木	挂马克杯	倒空杯子	将A2B放置在左侧	将A2B放置在右侧	整理桌面	倒茶	抓取与放置
π_0 [9]	4.0	20.0	11.5	12.0	8.0	23.0	16.0	30.0
$\pi_{0.5}$ [61]	7.0	**27.0**	<u>16.0</u>	11.0	13.0	31.2	<u>20.0</u>	34.5
OpenVLA-OFT [31]	<u>12.4</u>	23.5	13.5	<u>21.0</u>	<u>15.0</u>	<u>36.0</u>	15.0	<u>52.5</u>
Action-Sketcher（我们的方法）	**34.5**	<u>25.0</u>	**28.0**	**43.0**	**28.0**	**52.0**	**27.6**	**67.0**

表3. 在各种长时程真实世界任务中，使用和不使用人在回路校正视觉草图时的子任务完成率（%）。

真实世界任务	原始 ↑(%)	+ 人在回路中 ↑ (%)	Δ ↑ (%)
整理桌面	52.0	75.0	+23.0
倒茶	27.6	44.0	+16.4
抓取与放置	67.0	85.5	+18.5

为验证这一点，我们开展了一项实验：在执行动作之前，人类监督者可以暂停执行，并对生成的草图进行少量修改。如表 3所示，这种草图级校正机制显著提升了性能，使我们最具挑战性的真实世界任务的成功率接近完美。这表明，Action-Sketcher 不仅能够自主出色地完成任务，也是一种有效的人机协作框架。

4.3. 消融研究（研究问题3）

为评估各系统组件的贡献，我们从三个维度进行了消融研究，详见表4。***(1)*** 框架组件。移除空间推理会使成功率大幅降至 13.8%，而完全移除视觉草图则会使其进一步降至 9.8%（仿真）和 15.0%（真实世界）。这证实草图并非仅仅是辅助可视化手段，而是将语言指代定位为可执行动作的基础桥梁。***(2)*** 视觉基元。尽管所有基本元素协同发挥作用，关键点最为关键（26.6%），因为它能提供精确的坐标落地。箭头和边界框对于动态引导和消歧也至关重要；省略其中任一项，都会使可靠性明显低于完整模型（34.5%）。***(3)*** 训练课程。结果验证了我们的多阶段策略。跳过基础预训练（阶段 1）会导致成功率适度下降（29.2%），而省略推理微调（阶段 2）则会因草图缺乏连贯性而导致成功率大幅下降（18.1%）。至关重要的是，移除最终适配（阶段 3）会导致彻底失败（0.0%），这证明阶段 3 对策略将草图解读为动作不可或缺。

表 4. 框架组件、视觉原语和训练策略的消融研究。我们报告了代表性任务的结果：堆叠方块（仿真）和整理桌面（真实环境）。成功率表示仿真中的任务完成情况，而完成度则衡量真实环境中已完成子任务的百分比。

模型变体	成功率 ↑ (%)	完成度 ↑(%)
Action-Sketcher	**34.5**	**52.0**
框架组件消融		
无空间推理	13.8	23.9
无视觉草图	9.8	15.0
视觉草图元素消融		
无边界框	31.2	49.0
无关键点	26.6	43.6
无箭头	29.9	48.2
训练阶段消融实验		
不含阶段 1	29.2	39.7
不含阶段 2	18.1	21.9
不含阶段 3	0.0	0.0

5. 结论

本文提出了 Action-Sketcher，这是一种视觉-语言-动作框架，旨在解决长时程机器人操作中的空间歧义和时间脆弱性。为此，我们引入了 视觉草图，这是一种显式空间表征，通过结构化的看-思考-草图-行动循环将推理过程付诸实践。我们的方法会在生成可解释视觉计划的推理与基于这些计划执行动作之间自适应切换，因而优于最先进方法，尤其是在需要复杂空间定位和多步序列规划的任务中。大量实验和消融实验进一步表明，视觉草图 是关键组成部分，在高层推理与底层控制之间搭建了稳健的桥梁。其显式特性不仅提升了自主性能，还支持有效的人在回路校正，为打造更透明、更具协作性的机器人智能体铺平了道路。附录中详细介绍了未来可能的研究方向。F。

参考文献

[1] Ali Amin、Raichelle Aniceto、Ashwin Balakrishna、Kevin Black、Ken Conley、Grace Connors、James Darpinian、Karan Dhabalia、Jared DiCarlo、Danny Driess 等。pi*0.6：一种从经验中学习的视觉-语言-动作模型。arXiv 预印本 arXiv:2511.14759，2025。20[2]Xiang An、Yin Xie、Kaicheng Yang、Wenkang Zhang、Xiuwei Zhao、Zheng Cheng、Yirui Wang、Songcen Xu、Changrui Chen、Didi Zhu 等。Llava-onevision-1.5：用于普及多模态训练的完全开放框架。arXiv 预印本 arXiv:2509.23661，2025。15[3] Anthropic。Claude Sonnet 4。2025。18, 19[4] Daichi Azuma、Taiki Miyanishi、Shuhei Kurita 和 Motoaki Kawanabe。ScanQA：用于空间场景理解的三维问答。载于 CVPR，第 19129–19139 页，2022。5, 15[5] Alisson Azzolini、Hannah Brandon、Prithvijit Chattopadhyay、Huayu Chen、Jinju Chu、Yin Cui、Jenna Diamond、Yifan Ding、Francesco Ferroni、Rama Govindaraju 等。Cosmos-Reason1：从物理常识到具身推理。arXiv 预印本 arXiv:2503.15558，2025。18, 19[6] Shuai Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wenbin Ge、Sibo Song、Kai Dang、Peng Wang、Shijie Wang、Jun Tang、Humen Zhong、Yuanzhi Zhu、Mingkun Yang、Zhaohai Li、Jianqiang Wan、Pengfei Wang、Wei Ding、Zheren Fu、Yiheng Xu、Jiabo Ye、Xi Zhang、Tianbao Xie、Zesen Cheng、Hang Zhang、Zhibo Yang、Haiyang Xu 和 Junyang Lin。Qwen2.5-VL 技术报告。arXiv 预印本 arXiv:2502.13923，2025。18, 19[7] Shuanghao Bai、Wenxuan Song、Jiayi Chen、Yuheng Ji、Zhide Zhong、Jin Yang、Han Zhao、Wanqi Zhou、Wei Zhao、Zhe Li 等。迈向对机器人操作的统一理解：一项综合综述。arXiv 预印本 arXiv:2510.10903，2025。2[8]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。3[9] Kevin Black、Noah Brown、Danny Driess、Adnan Esmail、Michael Equi、Chelsea Finn、Niccolo Fusai、Lachy Groom、Karol Hausman、Brian Ichter 等。π 0：一种用于通用机器人控制的视觉-语言-动作流模型。arXiv 预印本 arXiv:2410.24164，2024。2, 3, 5, 8, 13[10] Anthony Brohan、Noah Brown、Justice Carbajal、Yevgen Chebotar、Xi Chen、Krzysztof Choromanski、Tianli Ding、D˜anny Driess、Avinava Dubey、Chelsea Finn 等。RT-2：视觉-语言-动作模型将网络知识迁移至机器人控制。arXiv 预印本 arXiv:2307.15818，2023。3[11] Qingwen Bu、Jisong Cai、Li Chen、Xiuqi Cui、Yan Ding、Siyuan Feng、Shenyuan Gao、Xindong He、Xu Huang、Shu Jiang 等。Agibot World Colosseo：用于可扩展智能具身系统的大规模操作平台。arXiv 预印本 arXiv:2503.06669，2025。6, 15[12] Jierun Chen、Fangyun Wei、Jinjing Zhao、Sizhe Song、Bohai Wu、Zhuoxuan Peng、S-H Gary Chan 和 Hongyang Zhang。《大型多模态模型时代的指代表达理解评估再探》。arXiv 预印本 arXiv:2406.16866，2024。15[13] Sixiang Chen、Jiaming Liu、Siyuan Qian、Han Jiang、Lily Li、Renrui Zhang、Zhuoyang Liu、Chenyang Gu、Chengkai Hou、Pengwei Wang 等。Ac-dit：用于移动操作的自适应协调扩散 Transformer。arXiv 预印本 arXiv:2507.01961，2025。20[14] Tianxing Chen、Zanxin Chen、Baijun Chen、Zijian Cai、Yibin Liu、Zixuan Li、Qiwei Liang、Xianliang Lin、Yiheng Ge、Zhenyu Gu 等。RoboTwin 2.0：一种可扩展的数据生成器和基准，借助强领域随机化实现稳健的双臂机器人操作。arXiv 预印本 arXiv:2506.18088，2025。6, 15, 16[15] Yi Chen、Yuying Ge、Yixiao Ge、Mingyu Ding、Bohao Li、Rui Wang、Ruifeng Xu、Ying Shan 和 Xihui Liu。EgoPlan-bench：使用多模态大语言模型对自我中心具身规划进行基准测试。CoRR，2023。6, 15[16] Cheng Chi、Zhenjia Xu、Siyuan Feng、Eric Cousineau、Yilun Du、Benjamin Burchfiel、Russ Tedrake 和 Shuran Song。扩散策略：通过动作扩散实现视觉运动策略学习。《国际机器人研究杂志》，页码 02783649241273668，2023。6, 7[17] 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：为最先进的多模态模型提供开放权重和开放数据。arXiv 预印本 arXiv:2409.17146，2024。5, 15[18] Mengfei Du、Binhao Wu、Zejun Li、Xuan-Jing Huang 和 Zhongyu Wei。EmbSpatial-Bench：使用大型视觉-语言模型对具身任务的空间理解进行基准测试。发表于计算语言学协会年会论文集（ACL），2024。18[19] Figure AI。Helix：用于通用人形机器人控制的视觉-语言-动作模型。https://www.figure.ai/news/helix，2025。访问日期：2025年4月18日。2, 3[20] Xingyu Fu、Yushi Hu、Bangzheng Li、Yu Feng、Haoyu Wang、Xudong Lin、Dan Roth、Noah A Smith、Wei-Chiu Ma 和 Ranjay Krishna。BLINK：多模态大语言模型能看见，却无法感知。发表于欧洲计算机视觉会议（ECCV）论文集，2024。18[21] 谷歌。Gemini 2.5 Pro Preview：编程表现更上一层楼。https://developers.googleblog.com/en/gemini-2-5-pro-io-improved-coding-performance/，2025。访问日期：2025年5月6日。18, 19[22] Jiayuan Gu、Sean Kirmani、Paul Wohlhart、Yao Lu、Montserrat Gonzalez Arenas、Kanishka Rao、Wenhao Yu、Chuyuan Fu、Keerthana Gopalakrishnan、Zhuo Xu 等。RT-TRAJECTORY：通过事后轨迹草图实现机器人任务泛化。arXiv 预印本 arXiv:2311.01977，2023。3[23] Agrim Gupta、Piotr Dollar 和 Ross Girshick。LVIS：用于大词汇量实例分割的数据集。发表于IEEE/CVF计算机视觉与模式识别会议论文集，第5356–5364页，2019。5, 15[24] Peng Hao、Chaofan Zhang、Dingzhe Li、Xiaoge Cao、Xiaoshuai Hao、Shaowei Cui 和 Shuo Wang。TLA：触觉-

语言-动作模型，用于高接触操作。arXiv 预印本 arXiv:2503.08548，2025。3[25] Chi-Pin Huang、Yueh-Hua Wu、Min-Hung Chen、Yu-Chiang Frank Wang 和 Fu-En Yang。THINKACT：通过强化视觉潜在规划实现视觉-语言-动作推理。arXiv 预印本 arXiv:2507.16815，2025。3[26] 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。18, 19[27] 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。6, 15[28] Yuheng Ji、Huajie Tan、Jiayu Shi、Xiaoshuai Hao、Yuan Zhang、Hengyuan Zhang、Pengwei Wang、Mengdi Zhao、Yao Mu、Pengju An 等。RoboBrain：从抽象到具体、用于机器人操作的统一大脑模型。arXiv 预印本 arXiv:2502.21257，2025。3, 5, 6, 15, 19[29] Yuheng Ji、Yipu Wang、Yuyang Liu、Xiaoshuai Hao、Yue Liu、Yuting Zhao、Huaihai Lyu 和 Xiaolong Zheng。VisualTrans：真实世界视觉变换推理基准。arXiv 预印本 arXiv:2508.04043，2025。20[30] Moo Jin Kim、Karl Pertsch、Siddharth Karamcheti、Ted Xiao、Ashwin Balakrishna、Suraj Nair、Rafael Rafailov、Ethan Foster、Grace Lam、Pannag Sanketi 等。OpenVLA：开源视觉-语言-动作模型。arXiv 预印本 arXiv:2406.09246，2024。2, 3, 6, 7[31] Moo Jin Kim、Chelsea Finn 和 Percy Liang。视觉-语言-动作模型微调：优化速度和成功率。arXiv 预印本 arXiv:2502.19645，2025。6, 7, 8[32] Alexander Kirillov、Eric Mintun、Nikhila Ravi、Hanzi Mao、Chloe Rolland、Laura Gustafson、Tete Xiao、Spencer Whitehead、Alexander C. Berg、Wan-Yen Lo、Piotr Doll´ar 和 Ross Girshick。Segment Anything。发表于 IEEE/CVF国际计算机视觉会议（ICCV）论文集，第4015–4026页，2023。15, 16[33] Ranjay Krishna、Yuke Zhu、Oliver Groth、Justin Johnson、Kenji Hata、Joshua Kravitz、Stephanie Chen、Yannis Kalanidis、Li-Jia Li、David A Shamma 等。Visual Genome：利用众包密集图像标注连接语言与视觉。《国际计算机视觉杂志》，123：32–73，2017。5, 15[34] 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。3, 6, 7[35] Bo Li、Yuanhan Zhang、Dong Guo、Renrui Zhang、Feng Li、Hao Zhang、Kaichen Zhang、Peiyuan Zhang、Yanwei Li、Ziwei Liu 等。LLaVA-OneVision：轻松实现视觉任务迁移。arXiv 预印本 arXiv:2408.03326，2024。5, 15[36] Xinghang Li、Minghuan Liu、Hanbo Zhang、Cunjun Yu、Jie Xu、Hongtao Wu、Chilam Cheang、Ya Jing、Weinan Zhang、Huaping Liu 等。视觉-语言基础模型作为有效的机器人模仿者。arXiv 预印本 arXiv:2311 01378，2023 3。

[37] Yanbang Li、Ziyang Gong、Haoyang Li、Xiaoqi Huang、Haolan Kang、Guangping Bai 和 Xianzheng Ma。机器人视觉指令。载于《计算机视觉与模式识别会议论文集》，第 12155–12165 页，2025 年。3、20 – 。[38] Yunfei Li、Xiao Ma、Jiafeng Xu、Yu Cui、Zhongren Cui、Zhi- gang Han、Liqun Huang、Tao Kong、Yuxiao Liu、Hao Niu 等。Gr-rl：迈向灵巧、精准的长时程机器人操作。arXiv 预印本 arXiv:2512 01801，2025 年。20。[39] Zhe Li、Weihao Yuan、Yisheng He、Lingteng Qiu、Shen- hao Zhu、Xiaodong Gu、Weichao Shen、Yuan Dong、Zi- long Dong 和 Laurence T Yang。Lamp：用于运动生成、检索和描述的语言-运动预训练。arXiv 预印本 arXiv:2410.07093，2024 年。3[40] Zhe Li、Cheng Chi、Yangyang Wei、Boan Zhu、Tao Huang、Zhenguo Sun、Yibo Peng、Pengwei Wang、Zhongyuan Wang、Fangzhou Liu 等。Robomirror：先理解再模仿，实现从视频到人形机器人运动。arXiv 预印本 arXiv:2512.23649，2025 年。3[41] Zhe Li、Cheng Chi、Yangyang Wei、Boan Zhu、Tao Huang、Zhenguo Sun、Yibo Peng、Pengwei Wang、Zhongyuan Wang、Fangzhou Liu 等。你会自由式吗？通过音频控制实现富有表现力的人形机器人运动。arXiv 预印本 arXiv:2512.23650，2025 年。[42] Zhe Li、Cheng Chi、Yangyang Wei、Boan Zhu、Yibo Peng、Tao Huang、Pengwei Wang、Zhongyuan Wang、Shanghang Zhang 和 Chang Xu。从语言到运动：通过运动潜在引导实现无重定向人形机器人控制。arXiv 预印本 arXiv:2510.14952，2025 年。3[43] Wenqi Liang、Gan Sun、Yao He、Jiahua Dong、Suyan Dai、Ivan Laptev、Salman Khan 和 Yang Cong。Pixelvla：推进视觉-语言-动作模型中的像素级理解。arXiv 预印本 arXiv:2511.01571，2025 年。6、7[44] Fanqi Lin、Ruiqian Nai、Yingdong Hu、Jiacheng You、Jun- ming Zhao 和 Yang Gao。Onetwovla：具备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。2、3[45] Yaron Lipman、Ricky TQ Chen、Heli Ben-Hamu、Maximil- ian Nickel 和 Matt Le。生成建模的流匹配。arXiv 预印本 arXiv:2210.02747，2022 年。5[46] Bo Liu、Yifeng Zhu、Chongkai Gao、Yihao Feng、Qiang Liu、Yuke Zhu 和 Peter Stone。Libero：终身机器人学习知识迁移基准。载于《神经信息处理系统进展》，第 36 卷，第 44776–44791 页，2023 年。6、15、16[47] Fuxiao Liu、Kevin Lin、Linjie Li、Jianfeng Wang、Yaser Ya- coob 和 Lijuan Wang。通过鲁棒指令调优缓解大型多模态模型中的幻觉。arXiv 预印本 arXiv:2306.14565，2023 年。5、15[48] Haotian Liu、Chunyuan Li、Yuheng Li 和 Yong Jae Lee。通过视觉指令调优改进基线。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第 26296–26306 页，2024 年。5、15[49]Jiaming Liu、Chenxuan Li、Guanqun Wang、Lily Lee、Kaichen Zhou、Sixiang Chen、Chuyan Xiong、Jiaxin Ge、Renrui Zhang 和 Shanghang Zhang。用于端到端机器人操

作的自我纠正多模态大语言模型。arXiv 预印本 arXiv:2405.17418，2024 年。3[50]Jiaming Liu、Mengzhen Liu、Zhenyu Wang、Lily Lee、Kaichen Zhou、Pengju An、Senqiao Yang、Renrui Zhang、Yandong Guo 和 Shanghang Zhang。Robomamba：用于高效机器人推理与操作的多模态状态空间模型。arXiv 预印本 arXiv:2406.04339，2024 年。3[51] Jiaming Liu、Hao Chen、Pengju An、Zhuoyang Liu、Renrui Zhang、Chenyang Gu、Xiaoqi Li、Ziyu Guo、Sixiang Chen、Mengzhen Liu 等。Hybridvla：统一视觉-语言-动作模型中的协同扩散与自回归。arXiv 预印本 arXiv:2503.10631，2025 年。3[52] 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。2、3[53] Gen Luo、Ganlin Yang、Ziyang Gong、Guanzhou Chen、Hao- nan Duan、Erfei Cui、Ronglei Tong、Zhi Hou、Tianyi Zhang、Zhe Chen 等。视觉具身大脑：让多模态大语言模型在空间中观察、思考和控制。arXiv 预印本 arXiv:2506.00123，2025 年。19[54] Ruiyuan Lyu、Tai Wang、Jingli Lin、Shuai Yang、Xiaohan Mao、Yilun Chen、Runsen Xu、Haifeng Huang、Chenming Zhu、Dahua Lin 和 Jiangmiao Pang。Mmscan：包含分层场景落地语言标注的多模态三维场景数据集。arXiv 预印本 arXiv:2406.09401，2024 年。5、15[55] Xiaojian Ma、Silong Yong、Zilong Zheng、Qing Li、Yitao Liang、Song-Chun Zhu 和 Siyuan Huang。Sqa3d：三维场景中的情境问答。载于 ICLR，2023 年。15[56] Junhua Mao、Jonathan Huang、Alexander Toshev、Oana Camburu、Alan L Yuille 和 Kevin Murphy。无歧义物体描述的生成与理解。载于《IEEE计算机视觉与模式识别会议论文集》，第 11–20 页，2016 年。15[57]Soroush Nasiriany、Abhiram Maddukuri、Lance Zhang、Adeet Parikh、Aaron Lo、Abhishek Joshi、Ajay Man- dlekar 和 Yuke Zhu。Robocasa：面向通用机器人的日常任务大规模仿真。arXiv 预印本 arXiv:2406.02523，2024 年。20[58]Dantong Niu 、 Yuvan Sharma、Giscard Biamby、Jerome Quenum、Yutong Bai、Baifeng Shi、Trevor Darrell 和 Roei Herzig。LLARVA：视觉-动作指令调优提升机器人学习能力。arXiv 预印本 arXiv:2406.11815，2024 年。3[59] OpenAI。GPT-4V(ision) 系统卡。https://openai. com / index / introducing - o3 - and - o4 - mini/，2025 年。访问日期：2025-04-16。19[60] 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。7。[61] Physical Intelligence。π0. 5：具备开放世界泛化能力的视觉-语言-动作模型。https://www. physicalintelligence . company / blog / pi05，2025 年 2025 04 25 2、3、6、7、8。访问日期：- - 。[62] Delin Qu、Haoming Song、Qizhi Chen、Zhaoqing Chen、Xi- anqiang Gao、Xinyi Ye、Qi Lv、Modi Shi、Guanghui Ren、Cheng Ruan 等. Embodiedonevision：用于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。2, 3[63] Delin Qu、Haoming Song、Qizhi Chen、Yuanqi Yao、Xinyi Ye、Yan Ding、Zhigang Wang、Jiafei Duan、Bin Zhao、Dong Wang 等. SpatialVLA：探索视觉-语言-动作模型的空间表征。arXiv 预印本 arXiv:2501.15830，2025。6, 7[64] Vignesh Ramanathan、Anmol Kalia、Vladan Petrovic、Yi Wen、Baixue Zheng、Baishan Guo、Rui Wang、Aaron Marquez、Rama Kovvuri、Abhishek Kadian 等. Paco：常见物体的部件与属性。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第7141–7151页，2023。15[65] Lucy Xiaoyang Shi、Brian Ichter、Michael Equi、Liyiming Ke、Karl Pertsch、Quan Vuong、James Tanner、Anna Walling、Haohuan Wang、Niccolo Fusai 等. Hi robot：通过分层视觉-语言-动作模型实现开放式指令跟随。arXiv 预印本 arXiv:2502 19417，2025 2, 3 . . [66] Chan Hee Song、Valts Blukis、Jonathan Tremblay、Stephen Tyree、Yu Su 和 Stan Birchfield. RoboSpatial：为机器人学中的二维和三维视觉语言模型传授空间理解能力。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第15768–15780页，2025。18 [67] Zirui Song、Guangxian Ouyang、Mingzhe Li、Yuheng Ji、Chenxi Wang、Zixiang Xu、Zeyu Zhang、Xiaoqing Zhang、Qian Jiang、Zhenhao Chen 等. Maniplvm-r1：利用大型视觉-语言模型，通过强化学习实现具身操作推理。arXiv 预印本 arXiv:2505.16517，2025。20[68] starVLA Community. Starvla：用于开发视觉-语言-动作模型的乐高式代码库，2025。20[69] Priya Sundaresan、Quan Vuong、Jiayuan Gu、Peng Xu、Ted Xiao、Sean Kirmani、Tianhe Yu、Michael Stark、Ajinkya Jain、Karol Hausman 等. RT-SKETCH：从手绘草图进行目标条件模仿学习。载于第八届机器人学习年会，2024。3[70] Huajie Tan、Sixiang Chen、Yijie Xu、Zixiao Wang、Yuheng Ji、Cheng Chi、Yaoxu Lyu、Zhongxia Zhao、Xiansheng Chen、Peterson Co 等. Robo-dopamine：用于高精度机器人操作的通用过程奖励建模。arXiv 预印本 arXiv:2512.23703，2025。20[71]Huajie Tan、Cheng Chi、Xiansheng Chen、Yuheng Ji、Zhongxia Zhao、Xiaoshuai Hao、Yaoxu Lyu、Mingyu Cao、Junkai Zhao、Huaihai Lyu 等. Roboos-next：一种用于终身化、可扩展且稳健的多机器人协作的统一记忆框架。arXiv 预印本 arXiv:2510 26536，2025 20 . . [72] Huajie Tan、Yuheng Ji、Xiaoshuai Hao、Xiansheng Chen、Pengwei Wang、Zhongyuan Wang 和 Shanghang Zhang. Reason-rft：用于视觉语言模型视觉推理的强化微调。载于第三十九届神经信息处理系统年会，2025。20[73] BAAI RoboBrain Team、Mingyu Cao、Huajie Tan、Yuheng Ji、Minglan Lin、Zhiyu Li、Zhou Cao、Pengwei Wang、En-

shen Zhou、Yi Han 等. Robobrain 2.0 技术报告。arXiv 预印本 arXiv:2507.02029，2025。3, 5, 19[74]Gemini Robotics Team、Saminda Abeyruwan、Joshua Ainslie、Jean-Baptiste Alayrac、Montserrat Gonzalez Arenas、Travis Armstrong、Ashwin Balakrishna、Robert Baruch、Maria Bauza、Michiel Blokzijl 等. Gemini robotics：将人工智能引入物理世界。arXiv 预印本 arXiv:2503.20020，2025。2, 3[75]Octo Model Team、Dibya Ghosh、Homer Walke、Karl Pertsch、Kevin Black、Oier Mees、Sudeep Dasari、Joey Hejna、Tobias Kreiman、Charles Xu 等. Octo：一种开源通用机器人策略。arXiv 预印本 arXiv:2405.12213，2024。3, 6, 7[76] Peter Tong、Ellis Brown、Penghao Wu、Sanghyun Woo、Adithya Jairam Vedagiri、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完全开放、以视觉为中心的探索。载于神经信息处理系统进展（NeurIPS），2024。18[77] Sai H Vemprala、Rogerio Bonatti、Arthur Bucker 和 Ashish Kapoor. 面向机器人学的 ChatGPT：设计原则与模型能力。IEEE Access，12:55682–55696，2024。2[78] Johanna Wald、Armen Avetisyan、Nassir Navab、Federico Tombari 和 Matthias Nießner. Rio：变化室内环境中的三维物体实例重新定位。载于 ICCV，第7658–7667页，2019。5, 15[79] Yipu Wang、Yuheng Ji、Yuyang Liu、Enshen Zhou、Ziqiang Yang、Yuxuan Tian、Ziheng Qin、Yue Liu、Huajie Tan、Cheng Chi 等. 迈向视觉语言模型中的跨视角点对应。arXiv 预印本 arXiv:2512 04686，2025 5, 20 . . [80] Jianwei Yang、Reuben Tan、Qianhui Wu、Ruijie Zheng、Baolin Peng、Yongyuan Liang、Yu Gu、Mu Cai、Seonghyeon Ye、Joel Jang 等. Magma：一种面向多模态人工智能智能体的基础模型。载于计算机视觉与模式识别会议论文集，第14203–14214页，2025。19[81] Licheng Yu、Patrick Poirson、Shan Yang、Alexander C Berg 和 Tamara L Berg. 对指代表达中的上下文进行建模。载于《计算机视觉—ECCV 2016：第14届欧洲会议》，荷兰阿姆斯特丹，2016年10月11日至14日，第14卷，第69–85页，2016年论文集，第二部分。施普林格，2016。15[82] Wentao Yuan、Jiafei Duan、Valts Blukis、Wilbert Pumacay、Ranjay Krishna、Adithyavairavan Murali、Arsalan Mousavian 和 Dieter Fox. RoboPoint：一种用于机器人空间可供性预测的视觉语言模型。arXiv 预印本 arXiv:2406.10721，2024。5, 15[83] Xiaohua Zhai、Basil Mustafa、Alexander Kolesnikov 和 Lucas Beyer. 用于语言图像预训练的 Sigmoid损失。载于 IEEE/CVF国际计算机视觉会议论文集，第11975–11986页，2023。13[84] Ruijie Zheng、Yongyuan Liang、Shuaiyi Huang、Jianfeng Gao、Hal Daume´e III、Andrey Kolobov、Furong Huang 和 Jianwei Yang. TraceVLA：视觉轨迹提示增强通用机器人策略的时空感知能力。arXiv 预印本 arXiv:2412.10345，2024。3, 6, 7

[85] Yifan Zhong、Xuchuan Huang、Ruochong Li、Ceyao Zhang、Zhang Chen、Tianrui Guan、Fanlian Zeng、Ka Num Lui、Yuyao Ye、Yitao Liang 等。Dexgraspvla：面向通用灵巧抓取的视觉-语言-动作框架。*arXiv* 预印本 *arXiv:2502.20900*，2025年。3[86] 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带推理的空间指代。*arXiv* 预印本 *arXiv:2506.04308*，2025年。5，15，18[87] Enshen Zhou、Cheng Chi、Yibo Li、Jingkun An、Jiayuan Zhang、Shanyu Rong、Yi Han、Yuheng Ji、Mengzhen Liu、Pengwei Wang 等。Robotracer：掌握面向机器人学视觉语言模型的空间轨迹推理。*arXiv* 预印本 *arXiv:2512.13660*，2025年。20

Action-Sketcher：通过视觉草图实现长时程机器人操作中的推理到动作

补充材料

附录

本补充材料提供了所提出方法的更多细节，以及因篇幅限制而未在论文中呈现的实验结果。第 A 节介绍模型、训练策略和评估设置的更多细节。第 B 节介绍训练数据集的详情。第C节介绍数据构建流程。第 D节补充更多实验结果与分析。第 E节展示更多可视化结果，以证明 Action-Sketcher 的有效性。第 F 节讨论 Action-Sketcher 潜在的未来研究方向和局限性。

A. 模型、训练与评估详情

本节将全面介绍 Action-Sketcher 框架的实现细节。首先，我们将详细说明基于 π_0视觉-语言-动作策略的具体模型架构，并定义有助于实现循环式看-思考-草图-行动流程的不同输入输出路径。随后，我们将说明三阶段课程训练中使用的详细超参数、计算配置和优化策略。

A.1. 模型结构

Action-Sketcher框架与模型无关，可与任何视觉-语言-动作（VLA）模型集成。在我们的实现和实验中，我们采用 π_0 [9] 作为基础VLA策略。 π_0 是一种通用机器人策略，它利用视觉语言模型主干网络进行语义理解，并通过流匹配策略生成连续动作，由三个主要组件构成：

• 视觉编码器：一个经过预训练的Vision Transformer（即SigLIP-SO400M [83]），用于处理视觉输入。该编码器将来自多个摄像机视图（例如，腕部视图和自我中心视角）的高分辨率图像，以及渲染后的视觉草图图像，映射为一系列视觉词元。

• 自回归大语言模型主干网络：核心推理单元是一个基于Transformer的大语言模型（即，Gemma-3B）。它以拼接后的词元序列作为输入，其中包括视觉词元、经词元化的任务指令，以及基于文本的历史子任务词元。该骨干网络以自回归方式运行，执行高层推理，并在推理模式下生成基于文本的草图参数和子任务描述。

• 流匹配动作专家：对于底层控制（即动作模式），模型采用流匹配头。该头不会以自回归方式将动作离散化，而是作为条件生成模型运行。它以机器人的本体感知状态和来自大语言模型骨干网络的潜在嵌入作为条件，通过求解常微分方程（ODE）生成连续、高频的动作片段。

A.2. 输入与输出

基于视觉-语言-动作模型，Action-Sketcher 以循环方式运行，并根据当前激活的模式（即推理模式和动作模式）采用不同的输入输出流程。输入和输出如下：

• 推理模式输入：模型接收：(1) 多视角 RGB 图像 $I^{1:n}_t$用于捕捉当前场景（即左腕视角、左腕视角和自我中心视角）；(2) 任务指令 ℓ（例如“清理桌上的物体”）；(3) 已完成子任务的文本历史记录。(4) 如果当前子任务不是该回合中的第一个子任务，则还会接收紧接在前一个子任务中生成的视觉草图图像。

•推理模式输出：大语言模型由 <BOR>（推理开始）标记触发，以自回归方式生成时空推理依据，随后生成子任务以及视觉草图的结构化定义，其中包括边界框、关键点和箭头的坐标（即 {"bbox":[...],"points":[...],"arrow":[...] }）。此模式以 <EOR> 终止，随后将纹理视觉草图渲染为图像。

•动作模式输入：生成并渲染子任务和视觉草图后，模型的输入会更新为以下内容：(1) 当前观测图像$I^{1:n}_t$；(2) 渲染后的视觉草图图像，其中将空间意图明确叠加在参考视图上；(3) 机器人的当前本体感知状态 s_t（例如，关节角度或末端执行器位姿）。

• 动作模式输出：收到来自大语言模型的 <BOA>（动作开始）令牌后，动作专家会预测一系列未来动作（一个动作块）A_t。具体而言，它会预测流向量场，将随机噪声样本去噪为有效的末端执行器位姿/关节角度和夹爪状态轨迹。

A.3. 训练设置

在论文正文中，我们采用了分阶段训练策略，完整设置见表5。在整个训练阶段，我们在一个服务器集群上进行了所有实验，每台服务器均配备 8×H100 GPU。我们

表 5。**Action-Sketcher 各训练阶段的详细配置**。该表展示了三阶段课程中的训练统计数据和超参数。阶段 1 侧重于基础时空表征，阶段 2 侧重于推理到草图生成，阶段 3 侧重于动作策略对齐和模式适应。

		阶段 1 （基础时空学习）	**阶段 2** （推理到草图）	**阶段 3** （草图到动作 / 模式适应）
Data	**数据集来源**	指代定位 / 指向 / 规划	推理到草图数据集	推理-草图-动作语料库
	#样本	∼3.4M（空间）+ 870k（时间）	∼21k 样本	∼6k 回合（2.8M 样本）
	数据模态	图像 + 文本	图像 + 文本 → 草图	图像 + 文本 + 草图 →动作
Model	**基础架构**	π_0 (PaliGemma-3B)	π_0 (PaliGemma-3B)	π_0 (PaliGemma-3B)
	可训练部分	视觉编码器 + 大语言模型	视觉编码器 + 大语言模型	大语言模型 + 动作专家
	动作头	冻结	冻结	流匹配
Training	**全局批次大小**	256	32	32
	梯度累积	4	2	2
	学习率： $\{\psi_v^{\mathbf{VIT}}, \phi_v^{\mathbf{LLM}}, \phi_v^{\mathbf{AE}}\}$	$\{1\times10^{-6}, 1\times10^{-4}, N/A\}$	$\{1\times10^{-6}, 5\times10^{-5}, N/A\}$	$\{N/A, 1\times10^{-5}, 1\times10^{-4}\}$
	学习率调度	余弦衰减	余弦衰减	带预热的恒定学习率
	预热比例	0.03	0.03	0.05
	优化器	AdamW	AdamW	AdamW
	权重衰减	0.05	0.05	0.05
	图像分辨率	224×224	224×224	224×224
	最大序列长度	4096	8192	8192
	精度	bfloat16	bfloat16	bfloat16
	训练轮数	1	1	20
	硬件	32 × H100	8 × H100	8 × H100

使用 PyTorch π_0 代码库实现我们的模型。在所有阶段，我们均采用 AdamW 优化器，使用余弦衰减学习率调度，权重衰减设为 0.05。为确保在高分辨率输入（224×224）下训练的稳定性和效率，我们采用 bfloat16 混合精度训练。具体的课程安排如下：

• **阶段 1：基础时空学习**。为处理大规模预训练语料库（合计 ∼4.3M 个样本），我们将训练扩展至 32 张 H100 GPU。我们采用 256 的全局批次大小，以高效处理多样的视觉定位和时间规划数据。在此阶段，动作头尚未初始化；我们专注于调整视觉编码器（LR 1×10^{-6}）和大语言模型骨干网络（LR 1×10^{-4}），使其掌握在 4096 个 token 上下文长度内进行时空推理的能力。

• **阶段 2：推理到草图对齐**。我们使用精选的 21k 样本数据集对模型进行微调，使其专门用于生成视觉草图。上下文窗口扩展至 8192 个 token，以容纳长时程推理历史和详细的草图定义。我们将全局批次大小降至 32，并将大语言模型的学习率降低至 5×10^{-5}，以保留预训练知识，同时使文本输出与特定的基于 JSON 的草图格式对齐。

• **阶段 3：草图到动作和模式适应**。在最后阶段，我们联合训练完整模型，包括此前冻结的组件和流匹配动作头。为了支持从草图到连续动作的复杂多模态映射，我们采用差异化学习率策略：视觉编码器保持冻结，以维持特征稳定性；大语言模型以较低的学习率（1×10^{-5}）进行微调；动作专家则以较高的学习率（1×10^{-4}）进行训练。该阶段运行 20 个周期，以确保动作策略在草图条件输入下充分收敛。

A.4. 评估详情

我们详细介绍了仿真和真实环境中的评估流程、指标及实验设置。对于真实世界实验，我们将训练好的策略部署在两种机器人平台上：Aloha AgileX 和 Galaxea R1。

• **RoboTwin 2.0（仿真）**。我们在 RoboTwin 2.0 仿真环境中的五项自定义操作任务上进行评估。每项任务均使用相同的随机种子，在所有模型上进行 100 次滚动执行，以确保公平比较。成功与否由仿真器内置的成功条件判定，该条件会检查回合是否符合特定任务的完成标准。

• **长时域任务（真实环境）**。对于两个长时程操作任务（整理桌面和倒茶），我们对每个模型进行 10 次滚动执行评估。为测试鲁棒性，我们每进行 2 次滚动执行便修改一次场景，包括改变物体位置并引入新的物体类型。所有模型均在同一场景下进行评估，

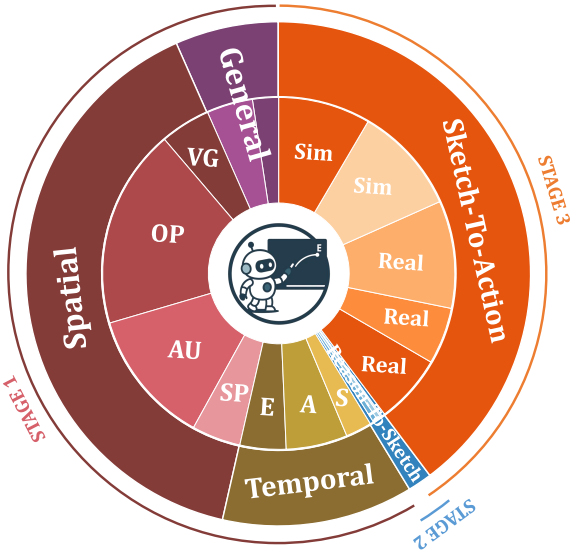


图5。整个训练数据集的分布。

完成后再切换到下一个场景配置。

• **通用抓取与放置（真实环境）**。 我们对每个模型进行了30次滚动执行评估，每完成5次滚动执行便更改一次场景，以测试模型在不同空间布局和干扰物体下的泛化能力。与长时域任务相同，所有模型均须在场景变更前完成当前场景的评估。

B. 训练数据集详情

在论文正文中，我们强调了训练数据的重要性以及机器人数据所占的比例。本节将详细介绍训练数据及其来源。整个训练数据集的分布如图所示。5。

• **通用多模态大语言模型视觉问答（47万）**。 为了奠定稳健的多模态理解基础，我们构建了一个包含47万个高质量样本的通用视觉问答数据集。该子集主要来源于两个数据集：LLaVA-665k [48] 和LRV-400k [47], , 涵盖标准视觉问答、基于OCR的查询以及视觉对话等多种任务。

• **视觉定位（327k）**。我们利用 LVIS [23], Ref-L4 [12], OV-VG [2, 35], Ref- COCO [56, 81], 和 Visual Genome [33]的大量标注，增强物体级定位能力。在此子集中，我们纳入了 327k 张高分辨率图像，并采用标准化的边界框坐标（x_1，y_1，x_2，y_2）。

• **对象指向（130 万）**。该模块专注于在具身场景中精确识别坐标。我们通过两步处理 Pixmo-Points [17] 数据集过滤机制，即，限制点的数量，并利用 GPT-4o [27] 筛选与室内环境相关的场景，最终得到 19 万组问答对，以减少杂乱信息。此外，我们纳入 RoboPoint [82]提供的 34.7 万个对象指代样本，并利用空间指代表达数据集 [86]（80.2 万个样本）扩展模型与物理世界交互的能力，确保模型能够泛化到各种指向任务。

• **可供性理解（881k）**。我们研究功能可供性和空间可供性，以支持物体交互。在功能推理方面，我们利用 PACO-LVIS [64] 的部件级标注，生成了 561k 组关于物体功能和部件用途的问答对。在空间可供性方面，我们整合了来自 RoboPoint 的 320k 组区域指代对，并优化坐标标注，帮助模型在真实世界场景中识别有效的放置区域和空间关系。

• **空间感知（318k）**。为实现细粒度的三维环境理解，我们整合了多个以三维为中心的数据集：用于分割和描述的 MMScan-224k [54]、用于语义重建的 3RScan-43k [78]，以及用于空间指代定位问答的 ScanQA-25k [4] 和 SQA3D- 26k [55]。

• **时间规划（870k）**。 为赋予模型稳健的时间推理与执行能力，我们构建了一个包含 870k 个样本的统一规划数据集。该数据集整合了三个主要来源中的多样化具身场景：用于第一人称视角任务规划的 EgoPlan-IT [15]、支持精细操作的 ShareRobot [28]，以及来自多样家庭环境的连续视觉序列 AgiBot-World [11]。该子集有助于进行长时程任务分解与序列推理。

• **推理到视觉（21k）**。 为连接高层语义推理与精确的视觉定位，我们构建了一个包含 21k 个高质量指令到草图配对的专用数据集。数据生成采用双流程策略：自动流程利用 SAM [32] 和 GPT-4o 进行可扩展的标注，人在回路流程则使用定制界面来确保几何精度。每个样本都将一条文本子任务指令与结构化视觉草图（边界框、关键点和方向箭头）配对，并由 GPT-4o 生成的推理链加以增强，用于说明动作在时间上的必要性以及视觉线索背后的空间推理依据。

• **视觉到动作（280 万）**。为了将视觉规划转化为可执行的运动控制，我们构建了一个大规模轨迹数据集，将视觉草图与低层动作序列对齐。该语料库涵盖 6100 个回合，共包含 280 万个状态-动作对，数据来自 LIBERO [46]、改良后的复杂版本 RoboTwin 2.0 [14], 以及真实世界的遥操作数据。每个样本都整合了

多视角观测、生成的视觉草图和推理上下文，以预测精确的动作片段，确保模型学会有效地依据所提供的视觉引导来确定其操作策略。

C. 数据构建细节

对于阶段 2，为使模型具备下游任务推理能力，并能够生成语义扎实的视觉草图，我们搭建了三条相互独立的标注流水线：(i) 自动视觉草图流水线，(ii) 人工视觉草图流水线，以及 (iii) 基于前两条流水线生成的视觉草图数据构建的自动推理生成流水线。因此，本阶段分为两个互补的小节：视觉草图生成（第 C.1 节）和推理生成（第 C.2 节）。对于阶段 3，为构建完整的推理-草图-动作训练语料库，我们将阶段 2 中的视觉草图和推理标注，与演示轨迹中对应的动作序列相结合（第 C.3 节）。

C.1. 视觉草图生成

• **自动标注流水线**。 通过检测夹爪连续的打开/闭合状态转换，将每段演示按时间划分为多个子任务。将 SAM [32] 应用于每个片段的起始帧和结束帧，以生成实例掩码；将质心位移最大的掩码对应物指定为目标物体。随后，将自我中心视角图像对、目标物体的边界框及其二维起始和结束坐标输入 GPT-4o，并使用确定性提示词要求其生成 (i) 简洁的子任务指令，以及 (ii) 符合 JSON 格式的视觉草图记录，其中包含边界框、关键点和箭头。

• **人工标注流程**。我们实现了一个 PyQt 界面（图 6），可串流读取 HDF5、MP4、JPG 或 PNG 序列。标注人员可以导航至任意帧，叠加相同的 JSON 基本图元，并验证或重写子任务描述；该界面可确保架构一致性，平均每分钟完成 2.3 条子任务标注。该工具也可选择性地部署在人类参与的推理环节中，以便实时修正模型生成的草图。

质量评估。第D 节中的消融实验表明，在仿真中，两种流程的成功率均高于基线，但人工整理数据集带来了进一步的 +7.6% 绝对提升，这主要归因于不准确草图更少。因此，所有主要结果中的模型均在人工精修划分上训练，而自动化流程将作为开源工具发布，以支持未来的规模化扩展。

C.2. 推理生成

推理数据沿两个互补维度生成：时间维度和空间维度。两个分支都利用以视觉草图标注为条件的 GPT-4o。

• **时间推理**。给定一个回合级子任务列表，以及当前关键帧的视觉草图图像，我们提示 GPT-4o 生成简洁的思维链，其中 (i) 回顾高层指令，(ii) 总结场景状态，(iii) 陈述此前已完成的子任务，以及 (iv) 说明当前子任务为何是接下来必要的动作。完整的提示模板见图 8。

• **空间推理**。对于每个关键帧，我们还会额外询问 GPT-4o，解释为什么真实标注视觉提示是完成当前子任务所需的最小集合。提示以子任务文本和 JSON 编码的草图为条件；完整模板见图9。

C.3. 草图-动作生成

对于每个回合，我们将生成的子任务和视觉草图与相应的动作序列对齐。关键在于，属于同一子任务的所有动作样本都共享同一张参考视觉草图，从而确保子任务执行期间空间定位的一致性。由此得到约 280 万个动作预测样本，涵盖 6100 个回合；每个样本包含多视角观测、当前子任务的渲染视觉草图、推理文本以及真实标注动作块。我们的训练语料库结合了仿真和真实世界数据，以确保稳健的泛化能力：

• **LIBERO [46]**。我们采用了四个任务套件的开源训练集：LIBERO-Spatial、LIBERO-Object、LIBERO-Goal 和 LIBERO-Long。这些训练集涵盖多样的操作场景，对空间推理和长时程规划的要求各不相同。我们使用第三人称视角图像、腕部相机图像、机器人本体感知状态以及语言指令。

• **RoboTwin2.0 [14].** 我们从 RoboTwin 2.0 中选取了 5 项任务，并通过引入随机物体杂乱、提高相机高度以缩小物体尺寸，以及延长回合时长等方式增加复杂度。这些改动确保任务足够长（≥4子任务），并用于测试空间理解。任务详情汇总于表6。所有任务均使用 Aloha Agile-X 机器人构型，配备第一人称视角和左右腕部视角，并通过绝对关节位置进行控制。

• **真实世界数据**。为验证真实世界迁移，我们在实验室环境中收集了两个长时程操作任务和一个空间理解任务。详情见表 6。所有任务均采用绝对关节位置控制，并使用自我中心视图和腕部摄像机视图。

表6. RoboTwin 2.0 和真实世界任务规格。

Task	描述	机器人	#回合
RoboTwin 2.0 任务			
堆叠积木	按不同颜色顺序（RGB/GBR）堆叠三个积木	AgileX Piper	550
悬挂马克杯	抓取马克杯并将其悬挂在架子上	AgileX Piper	200
放置杯子	将两个杯子分别放到随机摆放的盘子中最近的盘子旁	AgileX Piper	200
放置 A2B	将物体放在另一个物体的左侧或右侧	AgileX Piper	550
现实世界任务			
整理桌面	将不同的物体放入容器中（16 个子任务）	AgileX Piper	342
倒茶	将茶倒入两个茶杯中（7–8 个子任务）	AgileX Piper	749
抓取与放置	在干扰物中放置物体	Galaxea R1	1,500

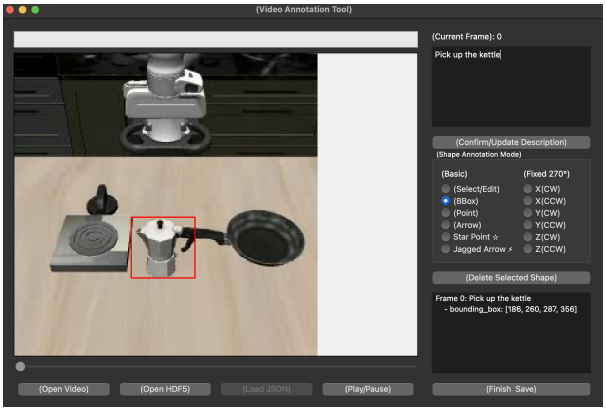


图6. **人工标注界面**。基于 PyQt 的标注工具截图，展示了用于创建视觉草图标注的界面。标注人员可以浏览帧、叠加边界框、关键点和方向箭头，并实时编辑子任务描述。

D. 补充实验

在本节中，我们将介绍由于篇幅限制而未在正文中呈现的完整实验和结果。我们还将介绍补充实验，这些实验为我们的方法提供了进一步的见解，但因篇幅限制而未在正文中呈现。

D.1. 消融实验详述

在论文正文中，我们进行了多项消融研究，以验证 Action-Sketcher 的关键设计选择。这里我们将进一步介绍这些实验的细节，并进行更深入的分析。为验证 Action-Sketcher 框架中各组成部分的必要性，我们通过移除或修改特定模块进行了受控消融，同时保持其他所有训练和评估流程不变。

不含空间推理。我们训练了一个模型，使其无需中间的空间推理步骤即可直接生成视觉草图。具体而言，模型在时间推理后直接预测 JSON 编码的草图参数，跳过关于为何需要特定视觉提示的思维链解释。这项消融实验旨在单独考察显式空间推理对生成准确落地的视觉草图的影响。

不使用视觉草图。我们训练了一个模型，使其执行与原始 Action-Sketcher 相同的推理步骤（同时生成时间推理和空间推理），但在动作执行期间不渲染视觉草图，也不将其叠加到参考图像上。相反，仅向动作预测模块提供原始观测图像。此消融实验旨在隔离使用视觉草图作为动作执行期间明确空间参考的影响，并将其与仅提供额外参考图像的做法进行对比。

结果表明，空间推理和视觉草图条件化均显著提升了任务性能，而完整的 Action-Sketcher 框架在所有评估指标上均取得了最佳结果。

D.2. 子任务级与任务级视觉草图绘制

我们的第一个补充实验研究了视觉草图绘制粒度的影响。我们比较了任务级草图（为整个任务提供一张视觉草图）与子任务级草图（为每个经过标注的子任务分别提供一张视觉草图，即 Action-Sketcher）。这两种方案均在仿真中的相同 RoboTwin 2.0 任务上进行评估，除草图绘制层级外，训练和评估协议完全相同。结果见表 7。

分析。结果表明，在所有任务中，按子任务级别绘制草图都具有明显优势；与按任务级别绘制草图相比，我们的方法平均提升了 11.5 个百分点。值得注意的是，性能差距会因任务的复杂度和长度而显著变化。堆叠方块任务需要按照严格的空间排序约束，依次精确地进行操作

表7. RoboTwin 2.0 仿真任务中的子任务级与任务级视觉草图绘制（成功率 %）。

方法	堆叠	悬挂	空置	A2B-L	A2B-R	平均值
任务级	9.5	20.0	15.0	35.5	21.0	20.2
子任务（我们的方法）	**34.5**	**25.0**	**28.0**	**43.0**	**28.0**	**31.7**
Δ 提升	+25.0	+5.0	+13.0	+7.5	+7.0	+11.5

，并涵盖 6 个子任务，其提升幅度最大（+25.0%）；而具有冗余性的悬挂杯子任务提升幅度较小，但仍相当显著（+5.0%）。

我们假设，随着任务变得更长、包含更多子任务，一次性推理完整轨迹并生成一张全面的视觉草图，反而会引入更多噪声并降低性能。任务级草图绘制要求模型同时规划多个操作阶段，导致对单个动作的空间定位不够精确，并可能使彼此重叠的视觉提示产生冲突。相比之下，我们的子任务级方法将复杂任务分解为易于处理的推理步骤，使模型能够专注于为每个当前的操作目标生成准确的视觉草图。这表明，我们的方法不仅在整体性能上具有优势，在更长、更复杂的任务上尤其如此，因为分解式推理和渐进式草图生成对成功至关重要。

D.3. 自动化与人工标注

第二项补充实验比较了使用人工标注的视觉草图训练的策略与使用由我们的流程全自动生成的草图训练的策略的性能。人工标注流程始终取得更高的成功率，而我们的自动化流程仍优于基线。结果汇总于表 8。

表 8. RoboTwin 2.0 仿真任务中的人工与自动化标注（成功率 %）。

方法	堆叠	悬挂	空置	A2B-L	A2B-R	平均值
自动化	21.5	21.0	23.0	32.0	23.0	24.1
人工	**34.5**	**25.0**	**28.0**	**43.0**	**28.0**	**31.7**
Δ提升	+13.0	+4.0	+5.0	+11.0	+5.0	+7.6

D.4. 推理速度分析

为评估 Action-Sketcher 的计算效率，我们分析不同组件的推理速度，并与基线模型进行比较。我们测量视觉草图生成时间和回合执行时间

在仿真和真实世界环境中的代表性任务上进行评估。仿真和真实世界评估的所有测量均在单块 NVIDIA RTX 4090 GPU 上完成。对于每项任务，我们计算到达子任务帧的平均时间，以及完成推理和草图生成所需的时间，取 5 个回合的平均值。结果见表9。

表9. 推理速度与成功率分析。草图时间衡量每个子任务的推理/草图生成开销，执行时间为平均回合执行时间，成功率/时间为按总时间归一化的成功率（$\times 10^{-3}$）。

	堆叠方块		挂杯子		整理桌面	
指标	π_0	Ours	π_0	Ours	π_0	Ours
草图（秒）	–	3.8	–	3.5	–	6.4
动作（秒）	9.0	9.4	13.1	13.0	26.0	24.5
总耗时（秒）	9.0	13.2	13.1	16.5	26.0	30.9
成功率（%）	4.0	**34.5**	20.0	**25.0**	23.0	**52.0**
成功率/时间	0.44	**2.61**	**1.53**	1.52	0.88	**1.68**

分析。 尽管每个子任务的视觉草图生成会带来 3.5–6.4 秒的额外开销，Action-Sketcher 仍保持了具有竞争力的推理效率。按总时间进行归一化后，Action-Sketcher 在所有任务上都实现了更优或相当的效率（成功率/时间），在堆叠方块（5.9×）和整理桌面（1.9×）等复杂任务上更是取得了显著提升。这表明，显式空间推理带来的性能增益超过了其计算开销。

D.5. 具身基准测试上的性能

为验证通过我们的训练课程培养的空间推理能力能否迁移到操作任务之外，我们在一组多样化的具身推理基准测试上评估 Action-Sketcher。这些基准测试评估空间理解的各个方面，包括深度感知、空间关系推理、可供性理解和轨迹预测。

评估设置。 我们在五个成熟的基准测试上进行评估，涵盖合成场景和真实世界场景：BLINK [20], CV-Bench [76], EmbSpatial-Bench [18],RoboSpatial [66], 和RefSpatial-Bench [86]。这些基准测试共同评估了多种模态下的空间推理能力，包括视觉问答、坐标预测、可供性推理和轨迹预测。

基线比较。 我们将 Action-Sketcher 与两类基线进行比较：（1）通用视觉语言模型：包括 Gemini- 2.5-Pro [21], GPT-4o [26], Claude-Sonnet-4 [3], 和 Qwen2.5-VL [6] 在内的先进视觉语言模型；（2）具身模型：包括 Cosmos-Reason1 [5], 在内的专用于机器人推理的模型

表10。**在五个空间推理基准上的表现**。不同模型中的最佳结果以**粗体**突出显示，次佳结果则加下划线。

模型 / 指标	BLINK			CV-Bench	EmbSpatial	RoboSpatial	RefSpatial-Bench		
	Dep.	Spa.	全部 ↑	全部 ↑	全部 ↑	全部 ↑	Loc.	Pla.	全部 ↑
通用基线									
Gemini-2.5-Pro-preview-05-06 [21]	79.03	84.62	81.83	84.59	**78.74**	59.87	44.58	31.73	38.16
Gemini-2.5-Flash-preview-04-17 [21]	77.42	79.02	78.22	84.03	74.75	54.10	37.50	23.00	30.25
GPT-o4-mini-2025-05-16 [59]	79.03	**88.11**	83.57	85.21	78.29	51.25	15.00	19.58	17.29
GPT-4o-2024-11-20 [26]	72.58	83.22	77.90	78.63	71.92	44.42	8.00	9.55	8.78
Claude-Sonnet-4-2025-05-14 [3]	75.81	80.42	78.12	78.43	64.26	51.26	5.00	10.37	7.69
Qwen2.5-VL-32B-Instruct [6]	77.42	85.31	81.37	81.59	74.45	52.16	16.83	10.60	13.72
Qwen2.5-VL-72B-Instruct [6]	74.19	78.32	76.26	82.68	73.30	48.33	23.50	15.83	19.67
具身基线									
Cosmos-Reason1-7B [5]	63.71	73.43	68.57	74.71	65.22	38.81	9.84	1.04	5.44
VeBrain-8B [53]	78.23	81.12	79.68	78.57	70.52	42.48	0.03	0.57	0.30
Magma-8B [80]	65.32	66.43	65.88	60.98	64.59	33.71	1.00	8.00	4.50
RoboBrain-7B-1.0 [28]	75.81	78.32	77.07	76.22	68.13	51.53	14.43	5.41	9.92
RoboBrain-7B-2.0	**84.68**	83.22	**83.95**	**85.75**	76.32	54.23	36.00	29.00	32.50
RoboBrain-32B-2.0	79.84	87.41	83.63	83.92	78.57	**72.43**	**54.00**	**54.00**	**54.00**
Action-Sketcher-3B（我们的方法）	82.30	74.67	82.00	82.89	72.65	62.28	44.50	41.00	42.50

VeBrain [53], Magma [80],RoboBrain [28, 73]。

结果与分析。 表。 10 展示了所有基准测试的综合结果。Action-Sketcher 表现出色，在多个基准测试中取得了具有竞争力或达到最先进水平的结果。值得注意的是，我们的模型在强调空间定位和坐标预测的基准测试（如 RoboSpatial、RefSpatial-Bench）上表现尤为出色，验证了我们的视觉草图训练方法能够有效培养细粒度空间推理能力。

结果表明，Action-Sketcher 的显式空间推理训练能够有效迁移到操作之外的通用空间推理任务。在该训练中，模型学习生成带有精确坐标、有依据的视觉提示。这表明，我们的框架不仅提升了操作性能，还培养了更稳健的空间理解能力，使其能够泛化到多种具身推理场景。

以下是主要发现：

• Action-Sketcher在空间基准测试（RoboSpatial、RefSpatial-Bench）上的表现优于专门的具身模型，证明了视觉草图训练有助于实现精准空间推理。

• 我们的模型参数量少得多，且专为具身任务优化，同时取得了与规模大得多的通用视觉语言模型相媲美的表现。

• 在需要坐标级空间理解的任务上，性能提升尤为显著，这验证了我们选择显式训练模型生成有依据的视觉提示是正确的。

E. 更多定性结果

本节提供了更多可视化结果，以展示 Action-Sketcher 在各种操作任务中的推理和视觉草图生成过程的有效性。

E.1. 推理与视觉草图生成

图 10-13 展示了多个具有代表性的操作任务中，完整的看-思考-草图-行动循环的核心推理部分。这些可视化结果展示了 Action-Sketcher 如何将高级指令分解为子任务，进行空间和时间推理，生成有依据的视觉草图，并据此执行动作。每幅图都展示了一个不同任务的推理过程：(a) 倒茶展示了各子任务中特定的空间推理；(b) 积木堆叠展示了具有精确空间顺序的连续操作；(c) 通用物体抓取展示了对空间关系的理解；(d) 长时程操作展示了结合上下文推理的多步骤任务分解。对于每个步骤，模型都会生成详细的场景描述，回顾已完成的动作，并明确推理接下来需要执行的子任务。视觉草图通过标示目标物体的边界框、指示运动方向的箭头、标记当前位置的点，以及标示目标位置的星形点，精准地编码空间信息。

E.2. 人在回路能力

Action-Sketcher 的显式推理与草图生成流程的一大优势，是能够自然地整合

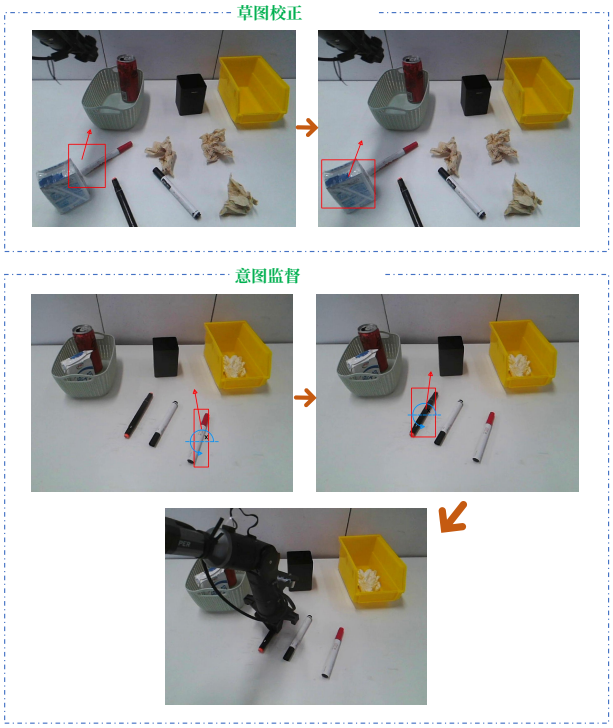


图7. **通过视觉草图进行人工干预**。 这种显式的中间表示充当交互式界面。*(上)* 草图校正： 用户识别出空间位置不准确的草图（左），并修正边界框和箭头（右），以避免执行失败。*(下)* 意图监督： 尽管模型提出了有效的计划（左），用户仍会修改草图，以贯彻特定偏好或安全约束（右），从而实时调整机器人的行为。

人工监督与干预。与以黑箱方式运行的端到端策略不同，我们框架中可解释的视觉草图使人们能够实时检查、纠正并引导机器人的空间理解和预期动作。

草图校正。 在推理过程中，用户可以在每个推理步骤后、动作执行开始前，检查并完善模型生成的视觉草图。图 7 的第一个子图展示了这样一种情况：模型初始生成的视觉草图存在空间位置偏差（上），可能导致任务失败。借助基于 PyQt 的标注界面，人工操作员可以迅速发现这些错误，并通过调整坐标、重新定位关键点或重绘箭头来提供修正后的草图（下）。这种校正机制每个子任务只需 3–5 秒的检查时间，同时确保操作准确。

意图监督。 除了纠正空间偏差之外，我们的框架还支持意图监督。即使模型生成的视觉草图在空间上准确，人工操作员仍可能希望修改机器人的预期行为，使其更符合自身偏好或安全约束。图 7第二个子图展示了用户如何在每个推理步骤后修改视觉草图，从而引导模型的行为。模型最初生成了一种可行的接近方式（上方），但人工操作员重新绘制视觉草图，以指定另一种策略（下方）。通过视觉修改而非复杂的策略调整，这种方式实现了透明直观的行为控制。

F. 未来工作

在这项工作中，我们展示了显式视觉定位在长时程操作中的有效性。展望未来，我们计划通过五项关键举措进一步推进 Action-Sketcher:

- **稳健泛化与基准测试**。 我们将广泛验证观察-思考-草图-行动框架在更多样的仿真基准和真实世界场景中的有效性 [29, 57, 79, 87],，确保其在领域内分布之外的多样化、非结构化环境中也能稳健运行。我们还将进一步探索采用强化微调（RFT） [67, 72] 或策略强化学习（RL） [1,38, 70] 等技术来提升模型的泛化能力。

- **架构的缩放特性**。 我们计划采用参数量显著更大且模型结构各异的骨干网络 [68]，研究本方法的缩放定律。我们的目标是评估，更大的模型容量或不同的模型范式是否能够提升视觉草图生成器预测结果的空间精度和语义保真度。

- **视觉草图设计空间**。 我们计划系统地探索不同视觉属性（例如基本元素的形状、配色方案和线条样式）如何影响策略注意力与鲁棒性。这项消融研究将找出能够最大限度提升推理与执行之间信息传递的最佳视觉语言 [37]。

- **扩展至移动操作**。 我们希望突破静态桌面环境的限制，将 Action-Sketcher 扩展至移动操作任务 [13]。这需要调整视觉草图，以协调底盘移动与机械臂操作；例如，在地面上绘制导航航点，或在大规模环境中定义远距离交互目标。

- **先进的人在回路范式**。最后，我们将探索新颖的交互模态，将人的反馈融入智能体系统 [71]。我们的目标是开发直观的界面，以便实时修正草图和修改意图，从而提升动态具身场景中的决策安全性。

时间推理和子任务生成提示

系统_提示 = """ 你是一位擅长时间分析和任务规划的机器人助手，目前正在执行一项任务。你将作为标注模型，生成真实标注推理。

你的任务是分析场景的当前状态，回顾计划，报告刚刚完成的任务，并说明你接下来将采取的动作。你的推理应体现出你正在继续已经开始的流程。

为了帮助你识别与子任务相关的正确物体，我们还会提供一张包含目标对象视觉提示的图像。请将其用作指引，但不要提及或承认这张引导图像的存在；你必须表现得仿佛只能参考场景的纯净图像。

你的标注模型角色：
我们会向你提供完整的真实标注计划、你刚刚完成的子任务，以及当前关键帧对应的具体正确子任务。

这些信息是你的隐藏知识。你的目标不是制定新计划或猜测下一个动作，而是运用这些真实标注，构建一段完美、合乎逻辑的思维链，解释*为什么*在完成上一个子任务后，给定的当前子任务是正确且必要的下一步。

你的推理过程：.
1
确认指令：说明你正在执行的高层指令。2. **分析场景**：简要描述当前场景，并指出你上一步动作造成的任何变化。3. **说明进度**：说明你刚刚完成的子任务，即你已完成的动作。4. **宣布当前动作**：说明你即将执行的具体子任务。

输出结构
你的完整回复必须使用 XML 风格标签括起来。

首先，构建推理块，并将其放在 `<think></think>` 标签中。推理块应以高层指令开头（暂时留空，因为稍后会插入），然后接上带有标题的必需部分：

<think>高层指令是“{{}}”。好的，继续执行计划。**场景描述：**[对当前场景的简要分析，并指出任何变化]。**我已完成的工作：** *{previous_子任务}* **
现在我需要：** *{current_子任务}*</think>

最后，在闭合的 </think> 标签之后另起一行，添加一个 <answer> 标签，其中仅包含当前子任务。

<答案>{当前_子任务}</答案>"""

用户_提示 = """
<图像>
<图像>
<图像>
<图像> *(基于图像的视觉草图生成器)*
这是我之前的推理：*"{previous_推理_输出}"*

根据新的图像，继续执行任务。子任务的完整计划如下：*{all_子任务_已排序}*。 我之前完成的子任务是：*"{previous_子任务}"*。 当前子任务是：*"{current_子任务}"*。

请根据这些输入生成下一条推理输出。"""

图 8. 时间推理提示。

空间推理与视觉草图生成提示

系统_提示 = """ 你是机器人学和计算机视觉领域的专家，专长是为机器人操作任务生成视觉提示。你将作为标注模型，生成真实标注推理。

你的任务是分析机器人的子任务及其对应场景，然后针对引导机器人动作所需的特定视觉提示（标注）提供详细推理。我会提供所需输出格式的示例。

你作为标注模型的角色：你将收到

当前关键帧。 当前关键帧的完整真实标注视觉提示（例如边界框、箭头及其坐标）

这些信息是你的隐藏知识。
你的目标不是臆造新的提示，而是构建一条严谨、合乎逻辑的思维链，解释*为什么*给定的视觉提示是完成该子任务最有效且必要的标注。

你的推理过程：
1. **分解子任务**：分析给定的子任务字符串。识别核心的**动作**（例如“拿起”“放置”）、**目标物体**（例如“可乐瓶”“纸团”），以及指定的**机械臂**等其他关键信息。 2. **分析场景**：查看图像，确定目标物体、机器人夹爪以及相关目标位置（例如“篮子”“黄色卡车”）。评估物体的朝向及其相对于机器人的位置。 3. **说明视觉提示的作用**：针对提供的每个真实标注提示，结合子任务和场景说明其用途。 4. **综合分析**：将分析整合成一段连贯的文字，说明这组完整的视觉提示为何能为机器人提供最佳引导。

--- 示例 ---###
用户输入：
<图像>要执行的子任务是：“用左臂拿起可乐瓶”。真实标注的视觉提示为：{"bbox": [[250, 186, 335, 325]], "箭头": [[296, 266, 238, 80]]}。请根据这些输入生成空间推理结果。

正确输出：
<思考>
首先，我将分解当前子任务：“用左臂拿起可乐瓶”。我需要执行的**动作**是“拿起”，我的**目标物体**是“可乐瓶”，并且必须使用我的**左臂**。接下来，我将分析场景。我的摄像机视图显示，可乐瓶正直立在桌子中央。我可以看到其他物体，但我的注意力集中在瓶子上。我的左夹爪已经就位，准备靠近它。

要正确完成这一操作，我需要视觉引导。位于 `[250, 186, 335, 325]` 的**边界框（`bbox`）**至关重要；它精确定位了可乐瓶，并界定了我需要抓取的确切范围。**箭头（`arrow`）**向我显示了“拿起”动作所需的运动向量。它起始于瓶子中心附近（`[296, 266]`），并指向 `[238, 80]`，引导我的左臂靠近瓶子并完成初始抬升。

结合用于识别的 `边界框` 和用于表示轨迹的 `箭头`，我可以获得清晰完整的指令，从而用左臂成功拿起可乐瓶。</think><answer>{"bbox": [[250, 186, 335, 325]], "arrow": [[296, 266, 238, 80]]}</answer>***### --- END EXAMPLE ---###***

输出结构你的完整回复必须遵循上方示例所示的格式。首先，以第一人称构建详细推理，并将其放在 `<think></think>` 标签内。**`<think>` 块中的推理不得超过 200 个单词。**最后，在闭合的 </think> 标签之后另起一行，添加一个 <answer> 标签。在其中，将提供的（经过预筛选的）真实标注视觉提示格式化为单行 JSON 字符串。"""

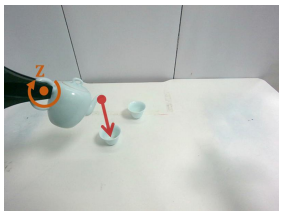
用户_提示词 = """
<i图像>
要执行的子任务是： *"{subtask_描述}"*。 真实标注视觉提示为： *{ground_truth_prompts_json}*。

请根据这些输入生成你的空间推理输出。"""

图 9. 空间推理提示。



参考图像



视觉草图器图像

提示： 高层指令是“把茶倒进两个茶杯里”。现在我需要执行子任务“将茶壶移到上方的茶杯旁”。“为了指导我完成这个子任务，我将使用以下视觉提示 {{"点": [[111, 59]], "星形_点": [[171, 59]], "锯齿形_箭头": [[110, 57, 170, 58]]}。”请观察我们是否已完成这个子任务。如果已完成，请考虑接下来要执行哪个子任务，以完成高层指令。如果尚未完成，请继续执行动作。

时间推理：

<思考>高层指令是“将茶倒入两个茶杯中”。好，我们继续执行计划。

场景描述： 机器人机械臂正用左夹爪拿着茶壶，并将其置于上方茶杯上方。茶杯是空的，准备盛茶。桌子上有一些散落的水滴，另一个茶杯也如预期般仍然是空的。

我已完成的工作： 将茶壶移到上方茶杯处。 **现在我需要：** 将茶倒入茶杯中</思考>

空间推理：

<思考>

我首先分解这个子任务：“将茶倒入茶杯”。这里的**动作**是“倒入”，**源物体**是当前由机械臂抓住的茶容器，**目标位置**是位于我面前桌子上的茶杯。

在这个场景中，机器人的夹爪正握着位于画面左上方的茶容器（水壶/茶壶），而目标茶杯位于下方的桌面上，且稍偏右。该空间关系要求进行受控的倾斜动作，以实现正确倒茶。

`[167, 56]` 处的**点**标记了茶容器的壶嘴尖端，液体将从此处流出。`[155, 49]` 处的**旋转_z_逆时针**指示器规定了倾斜茶容器并开始倒茶所需的逆时针旋转轴。`[170,74]` 处的**星形_点**突出显示了茶杯杯沿，即茶流精准落入的目标区域。

这些高度详细的标注为机器人提供了连贯的计划，使其能够平稳、准确地执行这项精密任务。最后，连接 `[167, 54]` 到 `[171, 76]` 的 **锯齿状_箭头**标示了从茶壶到茶杯的倾倒轨迹，确保液体沿着理想路径流动。\n\n这些高度详细的标注为机器人提供了连贯的计划，使其能够平稳、准确地执行这项精密任务。

</思考>

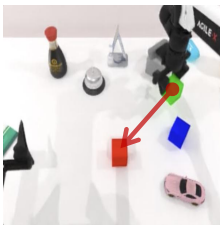
子任务： <答案>将茶倒入茶杯中</答案>

视觉草图： <答案> {"点": [[167, 56]], "旋转_z_逆时针": [[155, 49]], "星形_点": [[170, 74]], "锯齿状_箭头": [[167, 54, 171, 76]]}</答案>

图 10. 子任务与视觉草图推理的可视化。



参考图像



视觉草图器图像

提示： 高级 指令是“按以下顺序堆叠积木：红色、绿色、蓝色”。现在我需要执行子任务“拿起绿色方块”。“为了指导我完成这一子任务，我将使用以下视觉提示 {"bbox": [[172, 83, 202,121]], "arrow": [[187, 93, 188, 23]]}。”
请观察我们是否已完成这个子任务。如果已完成，请思考接下来要完成哪个子任务，才能实现高层指令。如果尚未完成，请继续执行动作。

时间推理：

<思考>高层指令是“按以下顺序堆叠积木：红色、绿色、蓝色”。好的，让我们继续执行计划。

场景描述： 场景显示了从俯视角度观察到的工作空间，桌面上散落着多个不同颜色的积木。红色方块位于工作空间的下方中央区域，蓝色积木则在它的右侧。绿色方块是我们之前的目标，已被机器人的左夹爪成功抓起，目前正悬在桌面上方。我之前的动作已使机器人的左臂成功抓起绿色方块，它现在被稳稳地夹持着。红色方块静置在桌面上，等待下一步动作。

我已完成的工作： 用左臂抓起绿色方块。 **现在我需要：** 将绿色方块放在红色方块上方。 </思考>

空间推理：

<思考>

首先，我会分解当前子任务：将绿色方块放在红色方块上面。

绿色方块当前由夹爪夹持，位于画面右上区域的大约坐标 [188, 83] 处。目标红色方块位于下方中央区域的 [120, 154] 处，需要大致向下移动70像素、向左移动68像素。

位于 `[188, 83]` 的**点**标记了由夹爪稳固夹持的绿色方块当前位置。位于 `[120, 154]`的**星形_点**表示红色方块的顶部中央精确表面，这里需要精准放置以实现稳定堆叠。从 `[181, 95]` 到 `[131, 147]` 的**锯齿形_箭头**描绘了最佳对角线轨迹路径，引导夹爪向左下方移动，同时保持受控下降，从而实现方块之间的正确对齐。

这些视觉标注共同消除了定位和移动方面的歧义，从而能够准确完成子任务。

</think>

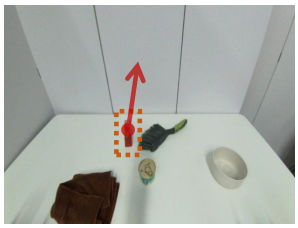
子任务： <答案>将绿色方块放在红色方块上面</答案>

Visual 草图: <答案>{"point": [[188, 83]], "star_point": [[120, 154]], "jagged_arrow": [[181, 95, 131, 147]]}</答案 >

图11。子任务与视觉草图推理的可视化。



参考图像



视觉草图器图像

提示： 用右臂把桌子上的牛奶放到棕色碗上。

时间推理：

<思考>高层　　级指令是“用右臂把桌上的牛奶放到棕色碗上面”。好，我们继续　　ue
执行计划。

场景描述： 场景包含一个摆放着多个物体的平面（桌子）。左侧是一叠棕色布料，中间是一个类似牛奶容器的罐子，再往右是一个带绿色手柄的黑色簸箕刷，最右侧是一个白碗。刷子附近还放置着一个小红色物体（容器）。

我已完成的工作： 尚未完成任何工作。**现在我需要：** 使用右臂拿起牛奶 </think>

空间推理：

<思考>

首先，我会分解当前子任务：“使用右臂拿起牛奶”。**动作**是“拿起”，**目标物体**是牛奶容器，所需的**机械臂**是右臂。\n\n现在分析场景，摄像机视图显示牛奶容器直立放置在前景中，

现在我需要确定视觉提示。牛奶容器清晰可见，位于图像的中左部。我将提供：
1. **边界框**：为突出牛奶容器的位置，我将在其周围绘制一个框，坐标大约为 [[90, 114, 102, 146]]
2. **箭头提示**：为指示夹爪的接近方向，我将添加一个向下指向牛奶顶部的箭头
容器，坐标为 [[94, 121, 94, 88]]

这些视觉提示共同构成了最佳组合，以确保机器人准确识别并执行拿起牛奶容器的动作。

</think>

子任务： <答案>用右臂拿起牛奶</答案>

视觉草图： <答案>{\"边界框\": [[90, 114, 102, 146]], \"箭头\": [[94, 121, 94, 88]]}</答案>

图12。子任务与视觉草图推理的可视化。



参考图像



视觉草图器图像

提示： 高层指令是“清理桌上的物体”。现在我需要完成子任务“把纸团放到黄色卡车上”。“为了指导我完成这个子任务，我将使用以下视觉提示 {\"点\": [[176, 29]],\"星形_点\": [[175, 87]], \" 锯齿形 _ 箭头 \": [[175, 30, 175,86]]}”

请观察我们是否已完成此子任务。如果已经完成，请思考下一项子任务，以实现高层指令；如果尚未完成，请继续执行动作

时间推理：

<思考>高层指令是“清理桌上的物体”。好，我们继续执行计划。

场景描述： 当前场景显示，两个纸团已根据上一个子任务放入黄色卡车内。桌上还剩下几个物体，包括钢笔、其他纸团和其他物品。我的夹爪目前没有拿着任何东西，说明它已准备好执行下一步动作

我已　　完成的操作：将纸团放到黄色卡车上。**现在我需要：**用右臂拿起纸团 </thin　　k>

空间推理：

<思考>

子任务是\"用右臂拿起纸团\"。所需的**动作**是“拿起”，**目标物体**是“纸团”，指定的**机械臂**是右臂。

在场景中，纸团位于前景，略偏中心左侧。它皱缩且形状不规则，因此需要精确定位才能抓取。机器人的右夹爪出现在图像右上角，正准备靠近目标。

所提供的视觉提示能够有效促进任务完成。位于 `[73, 184, 113, 223]` 的**边界框（`bbox`）**清晰界定了纸团的空间边界，将其与附近的物体（如钢笔或记号笔）区分开来。这确保机器人能够识别出要拾取的正确物体。**箭头（`arrow`）**起始于 `[95, 202]`，即纸团中心附近，并指向`[116, 137]`。这为右夹爪的轨迹提供指引，说明如何移动和对齐，以成功完成抓取。

用于识别的边界框与用于指示运动的箭头共同构成精准而简洁的指导，确保右臂能够正确抓取纸团

</think>

子任务：　<答案>用右臂拿起纸团</答案>

视觉草图：　<答案>{\"边界框\": [[73, 184, 113, 223]], \"箭头\": [[95, 202, 116, 137]]}</答案>

图13。子任务与 Visual-Sketcher 推理的可视化。