OpenVLA: An Open-Source Vision-Language-Action Model

摘要

在互联网规模的视觉语言数据和多样化的机器人演示的基础上预先训练的大型策略有可能改变教授机器人新技能的方式：可以微调这种视觉语言动作 (VLA) 模型以获得稳健、可推广的视觉运动控制策略，而不是从头开始训练新行为。然而，机器人领域 VLA 的广泛采用一直具有挑战性，因为 1) 现有的 VLA 大部分是封闭的，无法向公众开放，2) 先前的工作未能探索有效微调 VLA 以完成新任务的方法，这是采用的关键要素。为了应对这些挑战，推出了 OpenVLA [17]，这是一个 7B 参数的开源 VLA，在 970k 个现实世界机器人演示的多样化集合上进行了训练。OpenVLA 建立在 Llama 2 语言模型的基础上，结合了融合了来自 DINOv2 和 SigLIP 的预训练特征的视觉编码器。作为增加的数据多样性和新模型组件的产物，OpenVLA 在通用操作方面表现出色，在 29 项任务和多个机器人实例中，绝对任务成功率比 RT-2-X (55B) 等封闭模型高出 16.5%，参数减少了 7 倍。进一步表明，可以有效地微调 OpenVLA 以适应新设置，在涉及多个对象和强大语言基础能力的多任务环境中，尤其具有强大的泛化效果，并且比 Diffusion Policy [7] 等富有表现力的从头开始的模仿学习方法高出 20.4%。并且还探索了计算效率；另外表明了 OpenVLA 可以通过低秩自适应方法在消费级 GPU 上进行微调，并通过量化高效地提供服务，而不会影响下游成功率。最后，发布了模型检查点、微调和 PyTorch 代码库，内置支持在 Open X-Embodiment 数据集上大规模训练 VLA。

关键词：具身智能；多模态大模型；

1 引言

OpenVLA [17] 旨在解决机器人操作策略无法超出训练数据的泛化问题。虽然现有的策略可以应对新的初始条件，如对象位置或照明的变化，但它们在面对场景干扰或新物体时缺乏鲁棒性，也难以执行未见过的任务指令。相比之下，像 CLIP [25]、SigLIP [31] 和 Llama 2 [28] 这样的视觉和语言基础模型，通过大规模的互联网数据训练，展现出更强的泛化能力。然而，机器人领域的大规模预训练仍然是一个挑战，目前最大的数据集只有 10 万到 100 万例。因此，将这些基础模型用于机器人策略的训练，能促进超越训练数据的对象、场景和任务的泛化。

现有的研究已开始将预训练的语言和视觉-语言模型整合到机器人表征学习中，或者作为模块化系统中的一部分来执行任务规划和执行。最新的趋势是直接学习视觉-语言-动作模型（VLA）来控制机器人。VLA 模型如 PaLI 被微调生成机器人控制动作，展示出对新物体和任

务的出色泛化能力，树立了通用机器人策略的新标准。然而，目前 VLA 的广泛使用受限于两个原因：一是模型的封闭性，缺乏对架构、训练过程和数据混合的透明性；二是缺乏将 VLA 部署和适应到新机器人、环境和任务的最佳实践，特别是在普通硬件上。因此，机器人领域需要类似于现有开源语言模型生态系统的开源通用 VLA，支持有效的微调和适应。

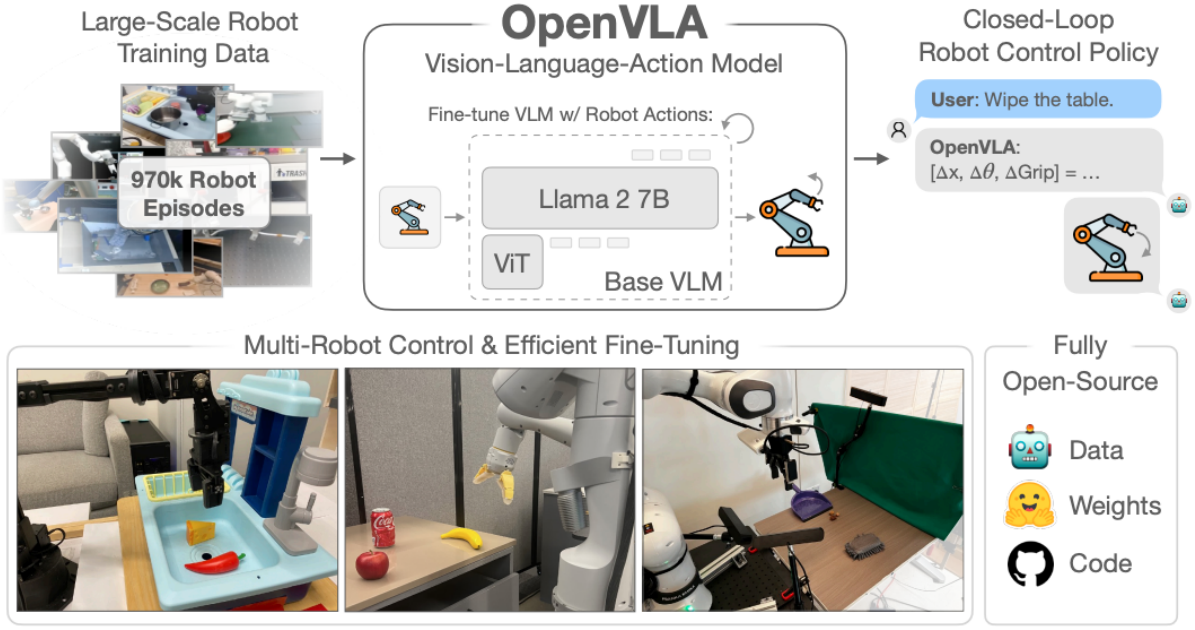

图 1. OpenVLA 是一个 7B 参数开源视觉-语言-动作模型 (VLA)，基于 Open X-Embodiment 数据集中的 97 万个机器人场景 [24] 进行训练。OpenVLA 为通用机器人操控策略树立了新标杆。它支持开箱即用地控制多个机器人，并且可以通过参数高效的微调快速适应新的机器人领域。OpenVLA 检查点和 PyTorch 训练管道是完全开源的，可以从 HuggingFace 下载和微调模型。

为此，OpenVLA 作为一个开源的 7B 参数 VLA 推出 [17]，设立了通用机器人操作策略的新标准。OpenVLA 包括一个预训练的视觉条件语言模型主干，捕捉多个粒度的视觉特征，并在包含 97 万条机器人操作轨迹的 Open-X Embodiment 数据集 [24] 上进行了微调。得益于数据的多样性和新的模型组件，OpenVLA 在 29 项评估任务中比之前的 55B 参数 RT-2-X 达到更优秀的效果。此外，OpenVLA 还研究了 VLA 的高效微调策略，在 7 项不同的操作任务中表现出色，显著优于如 Octo [23] 等预训练模型的微调策略。

2 相关工作

2.1 视觉条件语言模型

视觉条件语言模型通过在互联网规模的数据上训练，从输入图像和语言提示中生成自然语言，已被应用于视觉问答、物体定位等众多领域。近期 VLM 的关键进展在于其模型架构，结合了预训练的视觉编码器和语言模型的特征，利用计算机视觉和自然语言建模的进展，创造了强大的多模态模型。早期的研究探索了各种跨模态特征交互的架构 [2, 10, 13, 26]，而最新的开源 VLM 则趋向于一种更简单的 “补丁即标记” 方法，将来自预训练视觉变换器的补丁特

征视为标记，并将其投射到语言模型的输入空间。这种简化使得现有的大规模语言模型训练工具能够轻松适用于 VLM 的训练 [21, 22]。在工作中采用这些工具来扩展 VLA 的训练，特别是使用 Karamcheti 等人的 VLM [15] 作为预训练主干，这些模型通过多分辨率视觉特征训练，融合了 DINOv2 [6] 的低级空间信息和 SigLIP [31] 的高级语义，有助于视觉泛化。

2.2 通用机器人策略

近年来，机器人领域的趋势是训练能够执行多任务的“通用”机器人策略，这些策略基于大型多样化的机器人数据集 [16, 24, 29]，涵盖了许多不同的机器人形态 [1, 3, 9, 14]。特别是 Octo 模型，通过训练一个通用策略，实现了对多个机器人的直接控制，并且能够灵活地微调以适应新的机器人配置。这些方法与 OpenVLA 的关键区别在于模型架构。像 Octo [23] 这样的先前工作通常将预训练的组件（如语言嵌入或视觉编码器）与从零初始化的其他模型组件组合起来，通过策略训练过程学习如何将它们“拼接”在一起。相反，OpenVLA 采用了一种更加端到端的方式，直接微调视觉语言模型 (VLM) 来生成机器人动作，将这些动作视为语言模型词汇中的标记。实验评估表明，这种简单但可扩展的流程显著提高了性能和泛化能力，超越了现有的通用策略。

2.3 视觉语言动作模型

许多研究探讨了将视觉条件语言模型 (VLM) 应用于机器人领域，包括用于视觉状态表示、目标检测、高层规划和提供反馈信号。一些研究将 VLM 直接整合到端到端的视觉-运动操作策略中，但这些方法通常在策略架构中引入了显著的结构或需要校准的摄像头，限制了它们的适用性。近期的一些工作与该方法类似，直接微调大型预训练的 VLM 以预测机器人动作 [12, 18, 32]，这些模型通常被称为视觉-语言-动作模型 (VLA)，因为它们将机器人控制动作直接融合到 VLM 主干中。

这种方法有三个主要优势：(1) 在大规模的视觉-语言数据集上进行预训练的视觉和语言组件对齐；(2) 使用通用架构而非专为机器人控制设计的架构，使得能够利用现代 VLM 训练的可扩展基础设施，并通过最少的代码修改扩展到训练数十亿参数的策略；(3) 为机器人领域受益于 VLM 的快速改进提供了直接途径。现有的 VLA 研究要么专注于单一机器人或模拟环境中的训练和评估，缺乏通用性，要么是封闭的，不支持对新机器人设置的高效微调。

与之最相关的是 RT-2-X，它在 Open X-Embodiment [24] 数据集上训练了一个 55B 参数的 VLA 策略，并展示了最先进的通用操作策略性能。然而该工作在多个重要方面与 RT-2-X 不同：(1) OpenVLA 结合了强大的开放 VLM 主干和更丰富的机器人预训练数据集，在实验中性能优于 RT-2-X，同时模型规模小一个数量级；(2) 深入研究了 OpenVLA 模型在新目标设置中的微调，而 RT-2-X 未探索微调设置；(3) 首次展示了现代参数高效微调和量化方法对 VLA 的有效性；(4) OpenVLA 是第一个开源的通用 VLA，支持未来在 VLA 训练、数据混合、目标和推理方面的研究。

3 本文方法

OpenVLA 代表了开放领域具身智能的重大飞跃，将多模态学习与灵活的推理能力相结合（图 1）。该模型旨在通过结合视觉和语言信息来处理各种任务。

3.1 本文方法概述

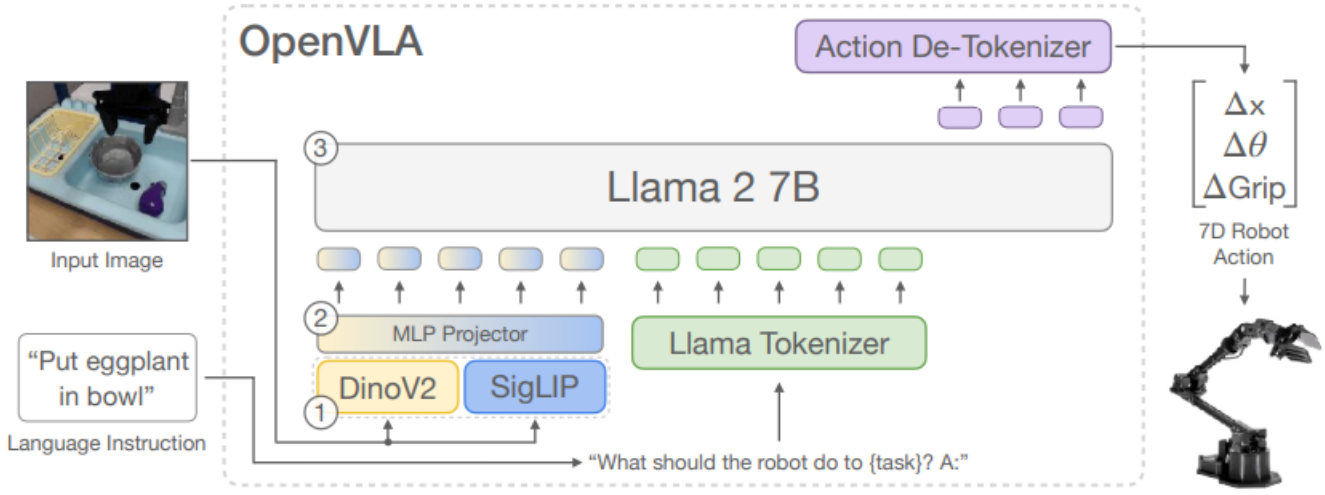

图 2. OpenVLA 模型架构。给定一个图像观察和一个语言指令，该模型预测 7 维机器人控制动作。该架构由三个关键组件组成：(1) 连接 Dino V2 [6] 和 SigLIP [31] 特征的视觉编码器，(2) 将视觉特征映射到语言嵌入空间的投影仪，以及 (3) LLM 主干，Llama 2 [28] 7B 参数大型语言模型。

OpenVLA，一个 7B 参数的开源 VLA，为通用机器人操作策略建立了新的先进水平。

1) OpenVLA [17] 由预训练的视觉条件语言模型主干组成，该模型主干可捕获多个粒度的视觉特征，并根据来自 Open-X 实施数据集的 970k 机器人操作轨迹的大型多样化数据集进行微调 - 该数据集涵盖了广泛的机器人实施、任务和场景。

2) 作为数据多样性增加和新模型组件的产物，OpenVLA 在 WidowX [29] 和 Google Robot 实施的 29 个评估任务中，绝对成功率比 55B 参数的 RT-2-X [24] 模型（之前最先进的 VLA）高出 16.5%。

3) 此外，研究了 VLA 的有效微调策略，这是一项在先前工作中未探索的新贡献，涵盖了 7 种不同的操作任务，涵盖从拾取和放置物体到清洁桌子的行为。

4) 并发现微调的 OpenVLA 策略明显优于微调的预训练策略（例如 Octo [23]）。与使用扩散策略的从头开始的模仿学习相比，微调的 OpenVLA 在涉及在具有多个对象的多任务设置中将语言与行为联系起来的任务上显示出显着的改进。

根据这些结果，它首次展示了利用低秩自适应和模型量化的计算高效微调方法 [8] 的有效性，以便在不影响性能的情况下在消费级 GPU 而不是大型服务器节点上调整 OpenVLA 模型。作为最后的贡献，作者开源了所有模型、部署和微调笔记本以及用于大规模训练 VLA 的 OpenVLA 代码库，希望这些资源能够支持未来探索和调整 VLA 用于机器人技术的工作。

3.2 视觉语言模型

在这项工作中，以 Prismatic-7B VLM [15] 为基础。Prismatic 遵循上述相同的标准架构，具有 600M 参数的视觉编码器、小型 2 层 MLP 投影仪和 7B 参数的 Llama 2 语言模

型主干 [28]。值得注意的是，Prismatic 使用两部分视觉编码器，由预训练的 Dino V2 [6] 和 SigLIP [31] 模型组成。输入图像块分别通过两个编码器，并将生成的特征向量按通道连接起来。与更常用的视觉编码器（如 CLIP [25] 或仅 SigLIP 编码器）相比，添加 DinoV2 特征已被证明有助于改进空间推理 [15]，这对机器人控制特别有帮助。SigLIP、DinoV2 和 Llama 2 并未公布有关其训练数据的详细信息，这些数据可能分别由来自互联网的图像文本、纯图像和纯文本数据组成，数量达数万亿。Prismatic VLM 在这些组件的基础上使用 LLaVA 1.5 数据混合 [21] 进行了微调，其中包含来自开源数据集的总共约 100 万个图像文本和纯文本数据样本。

3.3 训练过程

构建 OpenVLA 训练数据集的目的是捕捉大量不同的机器人实例、场景和任务。这使得最终模型能够开箱即用地控制各种机器人，并允许对新的机器人设置进行有效的微调。作者利用 Open X-Embodiment 数据集 (OpenX) [24] 作为基础来整理训练数据集。在撰写本文时，完整的 OpenX 数据集包含 70 多个单独的机器人数据集，拥有超过 200 万条机器人轨迹，这些数据集在社区的共同努力下被整合成一种连贯且易于使用的数据格式。为了使这些数据上的训练变得切实可行，作者对原始数据集应用了多个数据整理步骤。此次整理的目标是确保 (1) 所有训练数据集的输入和输出空间一致，以及 (2) 最终训练混合中实施方案、任务和场景的均衡组合。为了解决 (1)，将训练数据集限制为仅包含至少具有一个第三人称摄像机的操作数据集，并使用单臂末端执行器控制。对于 (2)，作者利用 Octo [23] 的数据混合权重来处理通过第一轮过滤的所有数据集。Octo 启发式地降低或删除多样性较低的数据集，并增加任务和场景多样性较大的数据集的权重。

3.4 训练数据

OpenVLA [17] 训练数据集的目标是捕获多样化的机器人形态、场景和任务，使最终模型能够直接控制各种机器人，并能够高效地微调以适应新的机器人设置。作者基于 Open X-Embodiment 数据集 (OpenX) 来构建训练数据集。OpenX 数据集包含超过 70 个单独的机器人数据集和 200 万条机器人轨迹，由社区共同努力整合成一个一致、易用的数据格式。

为了使这些数据的训练变得实际可行，作者对原始数据集进行了多步骤的筛选。筛选的目标是确保所有训练数据集具有一致的输入和输出空间，并在最终的训练混合中保持形态、任务和场景的平衡。为实现第一个目标，仅保留包含至少一个第三人称摄像头的操控数据集，并使用单臂末端执行器控制。为实现第二个目标，利用 Octo 的数据混合权重，对通过第一轮筛选的数据集进行加权，增加任务和场景多样性较大的数据集的权重，减少多样性较低的数据集的权重。

同时还尝试将一些在 Octo 发布后新增的 OpenX 数据集（如 DROID 数据集）[16] 纳入训练混合中，初始权重为 10%。实践中，作者发现 DROID 数据集的动作标记准确率在训练过程中始终较低，表明可能需要更大的混合权重或模型才能适应其多样性。为了不影响最终模型的质量，在最后三分之一的训练中将 DROID 数据集从混合中移除。

3.5 OpenVLA 设计决策

VLM 主干 LLaVA 在 BridgeData V2 [29] 接收器环境中的五项语言基础任务中，绝对成功率比 IDEFICS-1 高出 35%。经过微调的 Prismatic VLM [15] 策略取得了进一步的改进，在简单的单对象任务和多对象语言基础任务中，绝对成功率比 LLaVA 策略高出约 10%。将此性能差异归因于融合的 SigLIP-DinoV2 主干所提供的改进的空间推理能力。除了性能增强之外，Prismatic 还提供了模块化且易于使用的代码库，因此作者最终选择它作为 OpenVLA 模型的主干。为了训练 OpenVLA，作者对预训练的 Prismatic-7B VLM 主干进行了微调，以进行机器人动作预测。作者将动作预测问题表述为“视觉-语言”任务，其中输入的观察图像和自然语言任务指令被映射到一串预测的机器人动作 [4]。为了使 VLM 的语言模型主干能够预测机器人动作，并通过将连续的机器人动作映射到语言模型的标记器使用的离散标记来表示 LLM 输出空间中的动作。按照 Brohan 等人 [4] 的方法，作者将机器人动作的每个维度分别离散化为 256 个箱中的一个。对于每个动作维度，将箱宽设置为均匀划分训练数据中动作的第 1 分位数和第 99 分位数之间的间隔。使用分位数代替使用的最小-最大边界可以忽略数据中的异常动作，否则这些异常动作可能会大幅扩大离散化间隔并降低动作离散化的有效粒度。

微调视觉编码器先前对 VLM 的研究发现，在 VLM 训练期间冻结视觉编码器通常会带来更高的性能 [21]。直观地说，冻结的视觉编码器可以更好地保留从其互联网规模预训练中学习到的稳健特征。然而发现在 VLA 训练期间微调视觉编码器对于良好的 VLA 性能至关重要。假设预训练的视觉主干可能无法捕获有关场景重要部分的足够细粒度空间细节，从而无法实现精确的机器人控制。

机器人设置和任务作者在两个机器人实例上“开箱即用”地评估了 OpenVLA 的性能：BridgeData V2 评估中的 WidowX 机器人和 RT-1 [5] 和 RT-2 [4] 评估中的移动操作机器人。这两个平台在之前的研究中都被广泛用于评估通用机器人策略。作者在每个环境中定义了一套全面的评估任务，涵盖了各种泛化轴，例如视觉（看不见的背景、干扰对象、对象的颜色/外观）；运动（看不见的对象位置/方向）；物理（看不见的对象大小/形状）；和语义（看不见的目标对象、来自互联网的指令和概念）泛化。作者还评估了具有多个对象的场景中的语言调节能力，测试策略是否可以操纵用户提示中指定的正确目标对象。总体而言，作者在 BridgeData V2 实验的 170 次部署（17 个任务，每次 10 次试验）中评估了每种方法，并在 Google 机器人实验的 60 次部署（12 个任务，每次 5 次试验）中评估了每种方法。

训练和推理基础配置最终的 OpenVLA 模型在由 64 个 A100 GPU 组成的集群上训练了 14 天，总共 21,500 个 A100 小时，使用 2048 的批处理大小 [17]。在推理过程中，OpenVLA 在以 bfloat16 精度加载时需要 15GB 的 GPU 内存（即没有量化），并且在一个 NVIDIA RTX 4090 GPU 上以大约 6Hz 的速度运行（没有编译、推测解码或其他推理加速技巧）。同时可以通过量化进一步减少 OpenVLA 在推理过程中的内存占用，而不会影响现实世界机器人任务的性能。

4 复现细节

4.1 与已有开源代码对比

复现工作主要参考了 OpenVLA 的开源代码，该代码开源在https://openvla.github.io/，包括部署、训练、微调等脚本以及基础模型的架构。原论文采用的方式是安排相关的现实场景，使用 Google 机器人、WidowX 机器人等实机来验证上述实验任务，实验要求高，成本高。在复现过程中，我尝试在仿真环境中实现实验任务，希望避免实际情况中机器人的缺失。首先验证 OpenVLA 的单图推理能力，得到相应的单帧决策。同时，该工作还参考了 Simpler-Env，链接为https://github.com/simpler-env/SimplerEnv，该仓库的代码基于 SAPIEN [30] 模拟器和基于 CPU 的 ManiSkill2 [11] 基准测试。基于仓库代码修改 OpenVLA 的策略，调整相应的任务场景，使其能够在仿真平台上运行。同时，在 Libero [20] 仿真基准中对不同任务数据集进行微调得到的模型也在 Libero 仿真环境中进行了验证，包括 LIBERO-Spatial、LIBERO-Object 等任务套件。在该环境中执行了具体任务并得到了运行视频和结果。Simpler [19] 和 Libero [20] 中的具体验证结果将在后面的文字和附带的视频中再次提及。

4.2 实验环境搭建

4.2.1 Libero

终身学习为构建一个在其生命周期内学习和适应的通才代理提供了一个有发展前景的范例。与图像和文本领域的传统终身学习问题不同，后者主要涉及实体和概念的陈述性知识的转移，而决策中的终身学习 (LLDM) [27] 还需要转移程序性知识，例如动作和行为。为了推进 LLDM 的研究，作者引入了 LIBERO，这是机器人操作终身学习的新基准。具体来说，LIBERO 强调了 LLDM 中的五个关键研究主题：1) 如何有效地转移陈述性知识、程序性知识或两者的混合；2) 如何设计有效的策略架构和 3) LLDM 的有效算法；4) 终身学习者在任务排序方面的稳健性；5) 模型预训练对 LLDM 的影响。作者开发了一个可扩展的程序生成管道，原则上可以生成无限多的任务。为了进行基准测试，作者创建了四个任务套件（总共 130 个任务）[20]，用于研究上述研究主题。为了支持样本高效学习，为所有任务提供了高质量的人机远程演示数据。作者进行了广泛的实验，得出了几个有见地甚至出乎意料的发现：顺序微调在前向迁移中优于现有的终身学习方法，没有一种视觉编码器架构在所有类型的知识迁移方面都表现出色，而简单的监督预训练可能会阻碍代理在后续 LLDM 中的表现。

4.2.2 Simpler

机器人领域在通用机器人操纵策略方面取得了重大进展。然而，对此类策略的真实世界评估是不可扩展的，并且面临着可重复性的挑战，随着策略扩大其可执行的任务范围，这些挑战可能会恶化。在这项工作中，证明了基于模拟的评估可以成为现实世界评估的可扩展、可重复和可靠的代理。将真实环境和模拟环境之间的控制和视觉差异确定为可靠模拟评估的关键挑战，并提出了缓解这些差距的方法，而无需制作真实环境的全保真数字孪生。然后采用这些方法创建 SIMPLER [19]，这是一组模拟环境，用于在常见的真实机器人设置上进行操纵策略评估。通过对操纵策略进行配对的模拟和真实评估，证明了 SIMPLER 环境和现实世界中的策略性能之间存在很强的相关性。此外发现了 SIMPLER 评估准确反映了现实世界的政

策行为模式，例如对各种分布变化的敏感性。开源所有 SIMPLER 环境以及用于创建新环境的工作流程，以促进对通用操纵策略和模拟评估框架的研究。

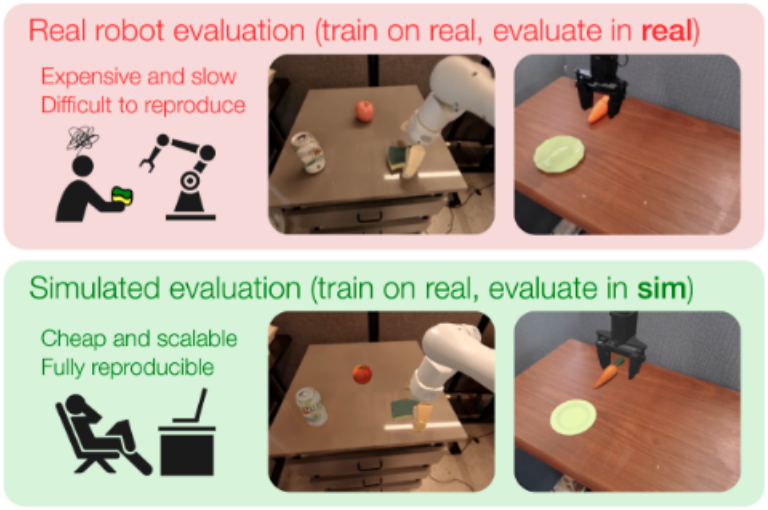

图 3. Simulated Manipulation Policy Evaluation Environments for Real Robot Setups

4.2.3 具体环境

实验环境	具体信息
Python	3.10.8
torch	2.1.2+cu121
torchvision	0.16.2+cu121
transformers	4.40.1
robosuite	1.4.0

表 1. 实验环境具体配置

```
+-----------------------------------------------------------------------------------------+
| NVIDIA-SMI 560.35.03              Driver Version: 560.35.03      CUDA Version: 12.6     |
|-----------------------------------------+------------------------+----------------------+
| GPU  Name                 Persistence-M | Bus-Id          Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |           Memory-Usage | GPU-Util  Compute M. |
|                                         |                        |               MIG M. |
|=========================================+========================+======================|
|   0  NVIDIA GeForce RTX 4090 D      On  |   00000000:B8:00.0 Off |                  Off |
| 30%   32C    P8             17W /  425W |       1MiB /  24564MiB |      0%      Default |
|                                         |                        |                  N/A |
+-----------------------------------------+------------------------+----------------------+

+-----------------------------------------------------------------------------------------+
| Processes:                                                                              |
|  GPU   GI   CI        PID   Type   Process name                              GPU Memory |
|        ID   ID                                                               Usage      |
|=========================================================================================|
|  No running processes found                                                             |
+-----------------------------------------------------------------------------------------+
```

图 4. 实验环境 GPU 配置

实验环境的搭建如图 4和表 1所示。

5 实验结果分析

OpenVLA (Open Vision-Language Action) 在 Google Robot、WidowX Robot 和 Libero 上各自展示了多模态智能体在不同环境中的任务执行能力。在 Google Robot 上，OpenVLA 被用于复杂环境中的任务执行，如拾取和放置物体、导航和交互，通过视觉和语言模型的结合，使机器人能够理解和执行自然语言指令，提升人机交互体验。在 WidowX Robot 上，OpenVLA 用于桌面环境的任务，如装配或物体操作，展示如何通过语言指令控制机械臂执行精确动作。而在 Libero 平台上，OpenVLA 被用于灵活的任务规划和执行，使机器人能够自主理解和执行复杂的任务链，适应性极强。这些应用展示了 OpenVLA 在不同复杂性和环境下的灵活性，推动了机器人与多模态模型的融合研究。在图 5中，展示了四个基准提供的任务，分别是 Pick Coke Can，Move Can near Orange，Move near Can 和 Place Coke Can Upright，图中是模型推理后的结果在以上任务场景下都能实现。

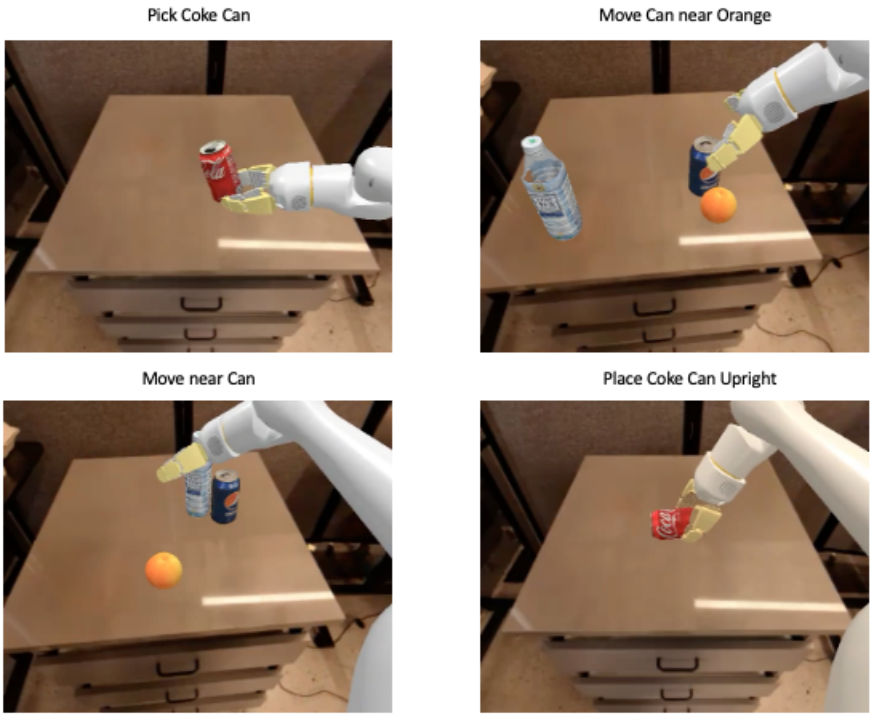

图 5. Google robots simulation evaluation tasks

在图 6中，展示了部分 WidowX 机器人的仿真效果。由于原文是在实机中进行实验，因此在仿真环境中出现一定的误差，在大部分任务可以实现正确推理，但在部分任务中会存在偏移、寻找不到目标或实现效果较差。分析仿真环境的质量与训练环境的质量存在一定的偏差，导致推理过程中准确率和效率下降。

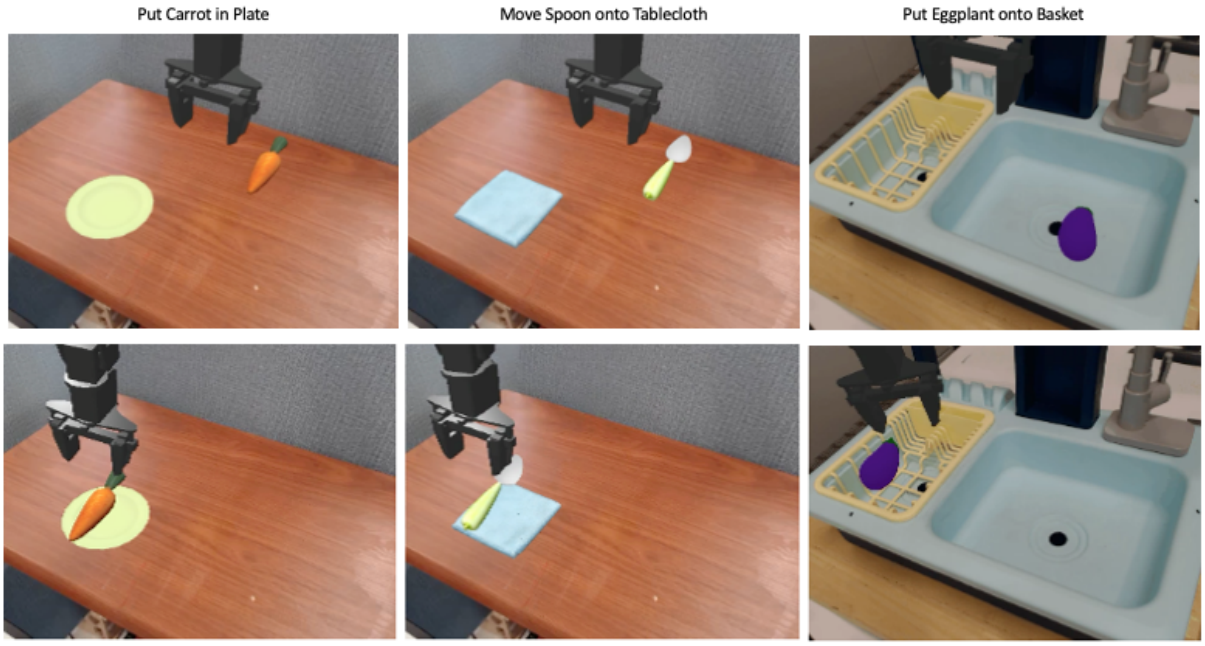

图 6. Widowx robots simulation evaluation tasks

LIBERO 基准包含四个任务套件，旨在研究机器人操作中的终身学习，因此原始论文研究了向各种任务的前向和后向迁移 [20]。在实验中，关注目标任务套件的监督微调，测量通过行为克隆训练的各种策略在成功演示任务上的表现。使用了以下四个任务套件进行实验，每个任务套件包含 10 个任务，每个任务有 50 个人类遥控演示：LIBERO-Spatial 由同一组对象但布局不同组成，并测试模型对空间关系的理解；LIBERO-Object 由相同的场景布局但不同的对象组成，并测试模型对对象类型的理解；LIBERO-Goal 由相同的对象和布局但不同的任务目标组成，并测试模型对不同面向任务的行为的了解；LIBERO-Long（也称为 LIBERO-10）由具有多样化对象、布局和任务的长期任务组成；如图 7和图 8中所示，分别展示了不同任务下的最初环境以及推理结果，由此可见在实际部署中，可以通过具体任务中做高效微调

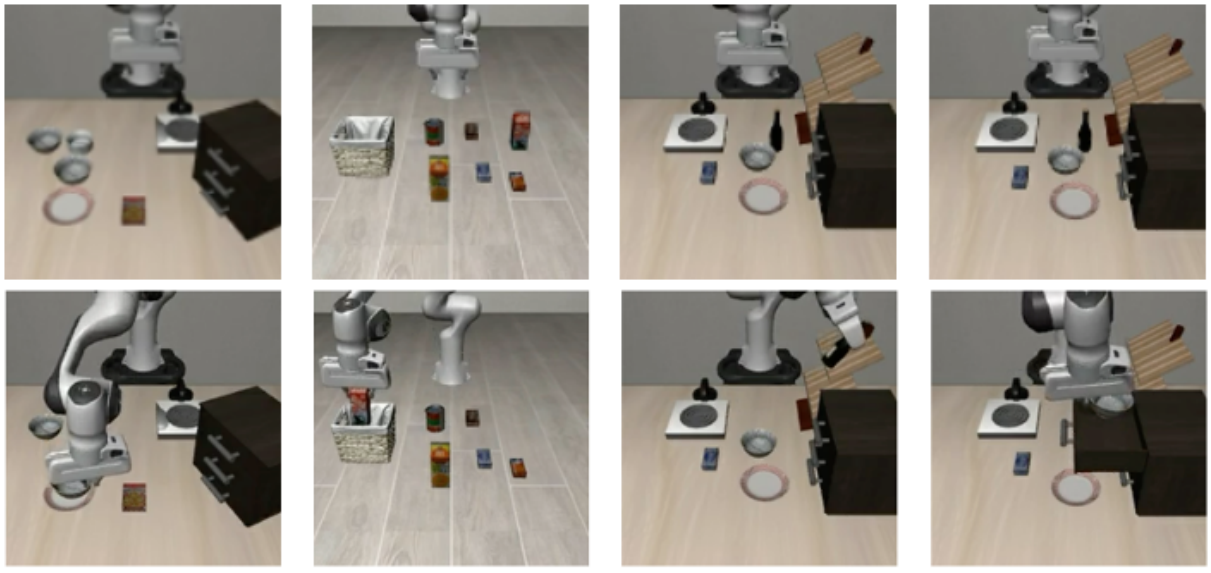

图 7. LIBERO-spatial and LIBERO-object simulation evaluation tasks

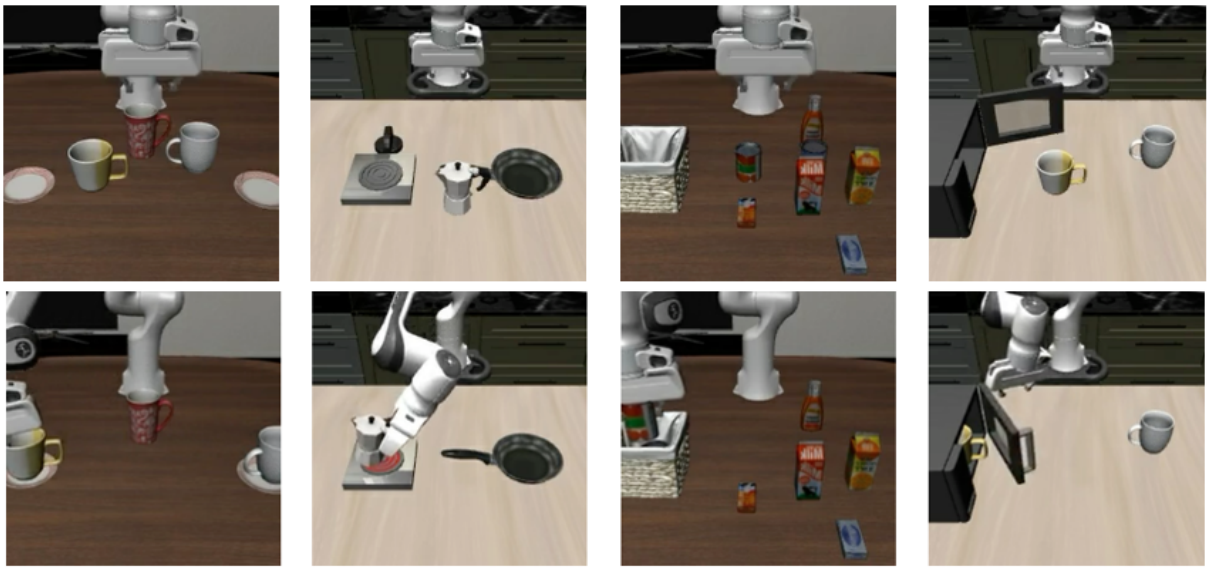

图 8. LIBERO-Long simulation evaluation tasks

LIBERO 模拟基准测试结果中报告了 LIBERO 基准测试中四个任务套件的每种方法的成功率 (SR) 和标准误差，这些结果取自三个随机种子的平均值，每个种子进行了 500 次试验。此外显示了每个任务套件中每种方法的排名，其中排名 1 表示套件中最强的方法，排名 3 表示最弱的方法。(平均排名很重要，因为它可以告知哪种方法最适合用作各种任务的默认方法；它比平均成功率更具参考价值，平均成功率并未根据单个任务套件难度进行标准化。) 总体而言发现经过微调的 OpenVLA 实现了最高的平均成功率和排名，其次是经过微调的 OCTO，然后是从头开始训练的扩散策略。复现工作中对于基准进行测试，得到了如表 2所示的复现效果，基本与原文平均效果一致。

	LIBERO-Spatial		LIBERO-Object		LIBERO-Goal		LIBERO-Long	
	SR (↑)	Rank (↓)	SR (↑)	Rank (↓)	SR (↑)	Rank (↓)	SR (↑)	Rank (↓)
Diffusion Policy from scratch	78.3 ± 1.1%	4	92.5 ± 0.7%	1	68.3 ± 1.2%	4	50.5 ± 1.3%	4
Octo fine-tuned	78.9 ± 1.0%	3	85.7 ± 0.9%	4	84.6 ± 0.9%	1	51.1 ± 1.3%	3
Original	84.7 ± 0.9%	2	88.4 ± 0.8%	2	79.2 ± 1.0%	2	53.7 ± 1.3%	2
Reproduction	85.0	1	88.0	3	79.0	2	62.0	1

表 2. Libero 实验结果对比

6 总结与展望

作者介绍了 OpenVLA，这是一种最先进的开源视觉-语言-动作模型，它开箱即用，在物理机器人控制中表现出色，并且 OpenVLA 可以通过参数高效的微调技术轻松适应新的机器人设置。它是在 Open-X 实施数据集中由 970k 条机器人操作轨迹组成的大型多样化数据集上进行微调的，该数据集涵盖了广泛的机器人实施、任务和场景。该模型将动作预测问题表述为“视觉-语言”任务，其中输入的观察图像和自然语言任务指令被映射到一系列预测的机器人动作。为了使 VLM 的语言模型主干能够预测机器人动作，LLM 输出空间中的动作通过将连续的机器人动作映射到语言模型的标记器使用的离散标记来表示。

在我的复现工作中，为了降低实验成本并提高可重复性，转向在模拟环境中验证实验任务。首先，在单帧图像上验证了 OpenVLA 的推理能力，确保其在仿真环境中能够正确输出决策。然后参考基于 SAPIEN 模拟器和 ManiSkill2 基准的 Simpler-Env 项目，调整 OpenVLA 的策略和任务场景，使其适应仿真平台的运行环境。此外在 Libero 仿真基准中使用不同的任务数据集对模型进行微调，并在 LIBERO-Spatial 和 LIBERO-Object 等任务套件中验证了其性能。通过这些努力，在仿真环境中成功验证了 OpenVLA 模型的推理和任务执行能力，进一步证明了该模型在多种任务场景中的适应性和有效性。

在未来的工作中，计划进一步扩展和优化 OpenVLA 模型的应用场景，以提升其实验环境的质量和模型性能。首先，计划将模型迁移到其他仿真平台，例如 NVIDIA 的 Isaac Sim。Isaac Sim 凭借其高质量的物理仿真和逼真的环境建模能力，将提供更加真实的实验环境，从而更好地模拟现实世界中的任务执行。其次计划将机械臂系统迁移到 Isaac Sim 平台并进行适配，使其能够在新的仿真环境中执行推理任务。这个迁移和适配过程将包括重新配置机械臂的控制参数，以确保其能够在新的仿真环境中平稳运行并执行复杂的视觉-语言-运动任务。最后将尝试进一步优化 OpenVLA 模型，提高其在各项任务中的表现以及泛化能力。这包括模型架构的改进、训练策略的优化以及在更多任务数据集上进行微调。通过这些优化工作，期望显著提高模型的推理速度和准确率，使其在复杂的任务场景中表现得更加出色。些未来的工作将进一步增强 OpenVLA 的适应性和实用性，使其在更多的仿真平台和任务环境中发挥更大的作用。

参考文献

[1] Homanga Bharadhwaj, Jay Vakil, Mohit Sharma, Abhinav Gupta, Shubham Tulsiani, and Vikash Kumar. Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4788–4795. IEEE, 2024.

[2] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342, 2010.

[3] Cristian Bodnar, Adrian Li, Karol Hausman, Peter Pastor, and Mrinal Kalakrishnan. Quantile qt-opt for risk-aware vision-based robotic grasping. *arXiv preprint arXiv:1910.02787*, 2019.

[4] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

[5] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1:

Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[7] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 2024.

[8] Shilpa Devalal and A Karthikeyan. Lora technology-an overview. In *2018 second international conference on electronics, communication and aerospace technology (ICECA)*, pages 284–290. IEEE, 2018.

[9] Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu, and Cewu Lu. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 653–660. IEEE, 2024.

[10] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[11] Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhan Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yunchao Yao, et al. Maniskill2: A unified benchmark for generalizable manipulation skills. *arXiv preprint arXiv:2302.04659*, 2023.

[12] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*, 2023.

[13] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[14] Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning*, pages 991–1002. PMLR, 2022.

[15] Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. Prismatic vlms: Investigating the design space of visually-conditioned language models. In *International Conference on Machine Learning (ICML)*, 2024.

[16] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

[17] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[18] Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, et al. Vision-language foundation models as effective robot imitators. *arXiv preprint arXiv:2311.01378*, 2023.

[19] Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su, Quan Vuong, and Ted Xiao. Evaluating real-world robot manipulation policies in simulation. *arXiv preprint arXiv:2405.05941*, 2024.

[20] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *arXiv preprint arXiv:2306.03310*, 2023.

[21] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

[22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[23] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Charles Xu, Jianlan Luo, Tobias Kreiman, You Liang Tan, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024.

[24] Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023.

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[26] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.

[27] Sebastian Thrun and Tom M Mitchell. Lifelong robot learning. *Robotics and autonomous systems*, 15(1-2):25–46, 1995.

[28] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[29] Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Vuong, Andre He, Vivek Myers, Kuan Fang, Chelsea Finn, and Sergey Levine. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning (CoRL)*, 2023.

[30] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, et al. Sapien: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11097–11107, 2020.

[31] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023.

[32] Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3d-vla: A 3d vision-language-action generative world model. *arXiv preprint arXiv:2403.09631*, 2024.