OpenVLA
开源视觉语言动作模型

完整逐段中文精译

Moo Jin Kim；Karl Pertsch；Siddharth Karamcheti；Ted Xiao；Ashwin Balakrishna；Suraj Nair；Rafael Rafailov；Ethan Foster；Grace Lam；Pannag Sanketi；Quan Vuong；Thomas Kollar；Benjamin Burchfiel；Russ Tedrake；Dorsa Sadigh；Sergey Levine；Percy Liang；Chelsea Finn

CoRL 2024 · arXiv:2406.09246v3 · 2024-09-05
底本 37 页 · 译本制作 2026-11-25
CC BY 4.0 · https://creativecommons.org/licenses/by/4.0/

> 正文、致谢与附录 A–E 完整翻译；保留全部 11 幅图、12 张表和 117 条参考文献。表格数值及标准书目信息保留原文；附录逐项给出任务中文名称与评分条件。中英逐段定位见同目录 paper.md 与 source_map.json。译者补充以“译注”标明。

从模型接口读到部署边界

视觉语言预训练 → 动作离散化 → 数据混合与视觉适配 → 多机器人评估 → 低秩微调与量化 → 失效模式与完整实验附录。

术语：VLA 视觉语言动作模型；VLM 视觉语言模型；embodiment 机器人形态；grounding 语言落地；LoRA 低秩适配；rollout 策略运行；StdErr 标准误。

摘要。大型策略若同时在互联网规模的视觉语言数据和多样化机器人示范上预训练，可能改变我们教授机器人新技能的方式：不再从零训练新行为，而是微调这类视觉语言动作（VLA）模型，以获得稳健、可泛化的视觉运动控制策略。然而，VLA 在机器人领域的广泛使用仍面临两项障碍：（1）现有 VLA 大多封闭，公众无法获取；（2）已有工作没有研究如何为新任务高效微调 VLA，而这正是实际采用的关键。为此，我们提出 OpenVLA：一个在 97 万条真实世界机器人示范上训练的、具有 70 亿参数的开源 VLA。OpenVLA 以 Llama 2 语言模型为基础，结合融合 DINOv2 与 SigLIP 预训练特征的视觉编码器。由于数据多样性增加及新模型组件的引入，OpenVLA 在通用操作方面表现出色：在跨多个机器人形态的 29 项任务上，其任务成功率比 RT-2-X（550 亿参数）等封闭模型高 16.5 个百分点，而参数量约为其七分之一。我们进一步证明，可以把 OpenVLA 有效微调到新设置；它在包含多个物体、需要较强语言落地能力的多任务环境中具有尤其出色的泛化表现，比 Diffusion Policy 等表达能力强的从零模仿学习方法高 20.4 个百分点。我们还研究计算效率：作为另一项贡献，我们证明，可利用现代低秩适配方法在消费级 GPU 上微调 OpenVLA，并通过量化高效部署，而不损失下游任务成功率。最后，我们发布模型检查点、微调笔记本和 PyTorch 代码库，内置支持在 Open X-Embodiment 数据集上大规模训练 VLA。作者标注：* 表示同等贡献；通信邮箱为 moojink@stanford.edu、pertsch@berkeley.edu、skaramcheti@stanford.edu；机构依次包括斯坦福大学、加州大学伯克利分校、丰田研究院、Google DeepMind、Physical Intelligence 和麻省理工学院；† 表示部分工作在 Google DeepMind 完成。

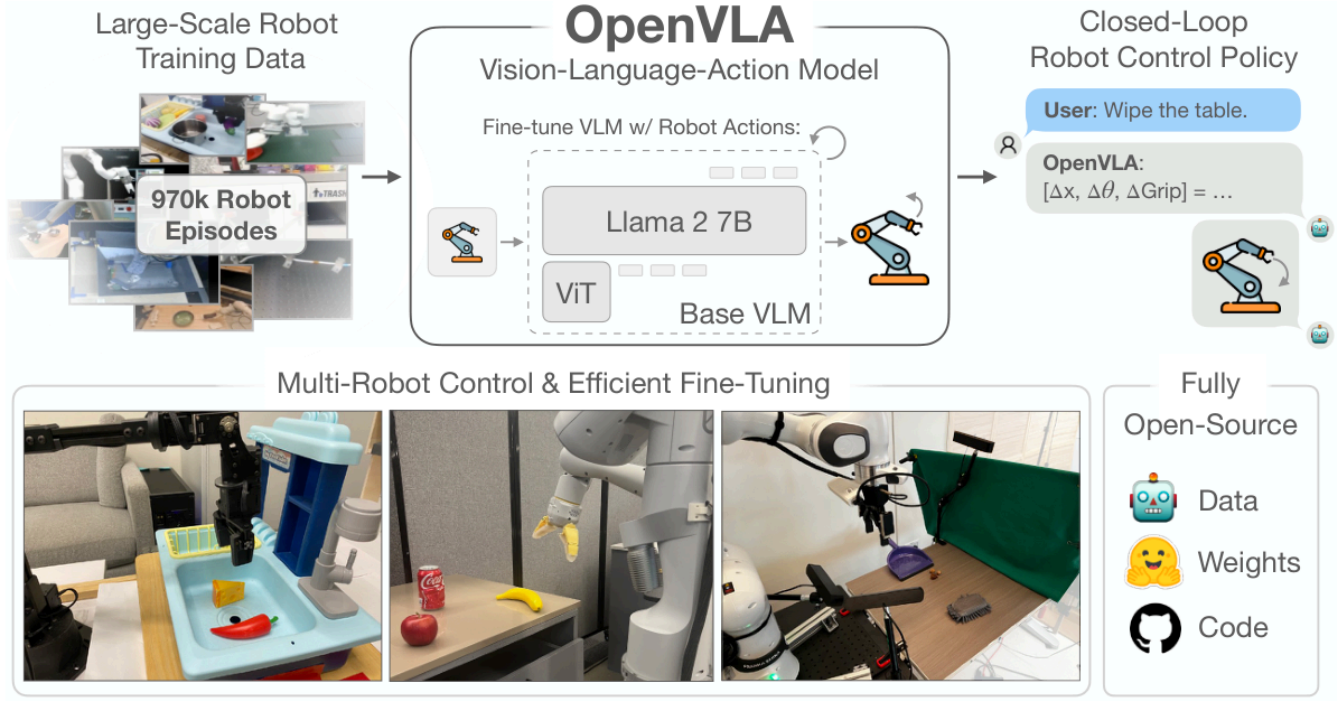

图 1：我们提出 OpenVLA，一个具有 70 亿参数的开源视觉语言动作模型（VLA），在 Open X-Embodiment 数据集 [1] 的 97 万段机器人回合上训练。OpenVLA 为通用机器人操作策略建立了新的最佳结果。它支持开箱即用地控制多种机器人，也能通过参数高效微调迅速适应新的机器人领域。OpenVLA 检查点与 PyTorch 训练流程完全开放，可从 Hugging Face 下载并微调模型。

1 引言

S003 · 原文 p.2

机器人操作学习策略的一个关键弱点是无法泛化到训练数据之外：尽管针对单个技能或语言指令训练的现有策略可以把行为外推到新的物体位置或光照等初始条件 [2,3]，它们对场景干扰物或新物体仍缺乏稳健性 [4,5]，也难以执行未见过的任务指令 [6,7]。但在机器人领域之外，CLIP [8]、SigLIP [9]、Llama 2 [10] 等现有视觉与语言基础模型，依靠互联网规模预训练数据捕获的先验，已经能够完成这些及更多类型的泛化。在机器人领域复现这种预训练规模仍是开放挑战——即使最大的机器人操作数据集 [1,11] 也只有 10 万至 100 万个样本——这种不均衡却提示了一个机会：把现有视觉语言基础模型作为训练机器人策略的核心构件，使策略能泛化到训练数据之外的物体、场景和任务。

S004 · 原文 p.2

为此，已有工作探索了把预训练语言模型和视觉语言模型用于机器人表征学习 [12,13,14]，或作为任务规划与执行模块化系统的组件 [15,16]。近期，它们又被用于直接学习控制用的视觉语言动作模型（VLA）[7,1,17,18]。VLA 是把预训练视觉语言基础模型用于机器人的直接方式：直接微调 PaLI [19,20] 等视觉条件语言模型（VLM），使其生成机器人控制动作。以互联网规模数据训练的强大基础模型为起点，RT-2 [7] 等 VLA 展现了令人印象深刻的稳健性以及对新物体、新任务的泛化能力，为通用机器人策略树立了新标准。然而，现有 VLA 难以广泛使用有两个关键原因：（1）当前模型 [7,1,17,18] 是封闭的，其架构、训练过程和数据混合方式缺乏透明度；（2）已有工作没有给出在新机器人、环境和任务上部署及适配 VLA 的最佳实践，特别是在消费级 GPU 等常见硬件上。我们认为，若要为后续研发建立丰富基础，机器人领域需要像开源语言模型生态 [21,22,23,24] 那样，拥有支持有效微调与适配的开源通用 VLA。

S005 · 原文 p.2

为此，我们提出 OpenVLA：一个具有 70 亿参数的开源 VLA，为通用机器人操作策略建立了新的最佳结果。脚注 1：OpenVLA 使用 SigLIP [9]、DINOv2 [25] 视觉编码器和 Llama 2 [10] 语言主干等多个预训练组件；三者均开放权重，但不开放其训练数据或代码。我们发布在这些组件之上复现 OpenVLA 所需的训练数据、代码和模型权重。OpenVLA 由能捕获多粒度视觉特征的预训练视觉条件语言主干构成，并在 Open X-Embodiment [1] 的 97 万条多样化机器人操作轨迹上微调；这些数据覆盖广泛的机器人形态、任务和场景。由于数据多样性增加及新组件的引入，OpenVLA 在 WidowX 与 Google Robot 两种形态的 29 项评估任务上，成功率比此前最先进的 550 亿参数 RT-2-X [7,1] 高 16.5 个百分点。我们还在涵盖物体抓放到清理桌面的 7 种不同操作任务上研究 VLA 高效微调策略，这是以往工作未探索的新贡献。我们发现，微调后的 OpenVLA 明显优于微调后的 Octo [5] 等预训练策略。与从零训练的扩散策略模仿学习 [3] 相比，在多物体、多任务且需要把语言对应到行为的场景中，微调后的 OpenVLA 显著更好。在此基础上，我们首次展示了低秩适配（LoRA）[26] 与模型量化 [27] 等计算高效微调方法的有效性，使 OpenVLA 可以在消费级 GPU 而非大型服务器节点上适配，同时保持性能。最后，我们开源全部模型、部署和微调笔记本，以及支持大规模 VLA 训练的 OpenVLA 代码库，希望帮助后续机器人 VLA 研究与适配。

2 相关工作

S006 · 原文 p.3

视觉条件语言模型（VLM）在互联网规模数据上训练，根据输入图像及语言提示生成自然语言，已用于从视觉问答 [28,29,30,31] 到物体定位 [32,33] 的众多应用。近期 VLM 的关键进展之一，是把预训练视觉编码器 [8,9,25] 的特征与预训练语言模型 [10,23,34,35,36] 连接起来的架构，直接利用计算机视觉与自然语言建模的进步构建强大多模态模型。早期工作探索了多种视觉与语言特征交叉注意力架构 [37,38,39,40,41]；新的开源 VLM [42,43,20,44] 则趋向更简单的“图像块即词元”方式：把预训练视觉 Transformer 的图像块特征作为词元，再投影到语言模型输入空间。这种简洁性使大规模语言模型训练工具容易复用于 VLM。我们利用这些工具扩展 VLA 训练，具体采用 Karamcheti 等 [44] 的 VLM 作为预训练主干；它融合 DINOv2 [25] 的低层空间信息和 SigLIP [9] 的高层语义，以多分辨率视觉特征促进视觉泛化。

S007 · 原文 p.3

机器人领域近期的一个趋势是在大型多样化机器人数据集 [50,51,45,52,53,54,55,6,2,56,49,11,1] 上训练多任务“通用”机器人策略 [45,46,47,6,2,48,49]，覆盖多种机器人形态 [57,53,58,59,60,61,62,63,64,65,1,5,66]。其中，Octo [5] 训练通用策略，能够开箱即用控制多种机器人，并灵活微调到新设置。这些方法与 OpenVLA 的关键区别是模型架构。Octo 等以往工作通常把语言嵌入或视觉编码器等预训练组件与从零初始化的额外组件组合 [6,2,5]，在策略训练中学会把它们“缝合”起来。OpenVLA 则采用更端到端的路线：把机器人动作当作语言模型词表中的词元，直接微调 VLM 生成动作。实验表明，这种简单而可扩展的流程显著提高了相对于既有通用策略的性能与泛化能力。

S008 · 原文 p.3

不少工作探索把 VLM 用于机器人，例如视觉状态表征 [12,13]、物体检测 [67]、高层规划 [16] 和提供反馈信号 [68,69,70,71]。另一些工作把 VLM 直接整合进端到端视觉运动操作策略 [14,15]，但在策略架构中加入大量专门结构或要求相机标定，限制了适用性。近期多项研究采用与我们类似的方案，直接微调大型预训练 VLM 来预测机器人动作 [7,1,72,73,17,18,74]。这类模型通常称为 VLA，因为它把机器人控制动作直接融入 VLM 主干。这有三项关键益处：（1）在大型互联网规模视觉语言数据集上对齐预训练视觉与语言组件；（2）采用非控制专用的通用架构，可以利用现代 VLM 训练的可扩展基础设施 [75,76,77]，以很少代码修改训练十亿参数级策略；（3）为机器人直接受益于 VLM 的快速进步提供路径。现有 VLA 工作要么局限于单机器人或仿真设置 [72,73,78,74]，缺乏通用性；要么封闭且不支持高效微调到新机器人设置 [7,1,17,18]。

S009 · 原文 p.3

最相关的 RT-2-X [1] 在 Open X-Embodiment 数据集上训练 550 亿参数 VLA，并展示最先进的通用操作性能。但本工作与 RT-2-X 在多个重要方面不同：（1）结合强大的开放 VLM 主干和更丰富的机器人预训练数据，OpenVLA 在实验中以约少一个数量级的规模超越 RT-2-X；（2）我们深入研究了 OpenVLA 对新目标设置的微调，而 RT-2-X 没有研究微调设置；（3）我们首次展示现代参数高效微调与量化方法对 VLA 有效；（4）OpenVLA 是首个开源通用 VLA，支持后续研究 VLA 训练、数据混合、目标函数和推理。

3 OpenVLA 模型

S010 · 原文 p.4

我们提出 OpenVLA，一个在 Open X-Embodiment [1] 的 97 万条机器人示范上训练的、具有 70 亿参数的视觉语言动作模型。开发 VLA 的最佳实践仍有许多问题尚未充分探索，例如应选择何种模型主干、数据集与训练超参数。下文详述我们的开发方法与主要经验：先概述构成 OpenVLA 主干的现代 VLM（3.1 节），再描述基本训练方案与数据（3.2、3.3 节），讨论关键设计决策（3.4 节），最后给出训练与推理基础设施细节（3.5 节）。

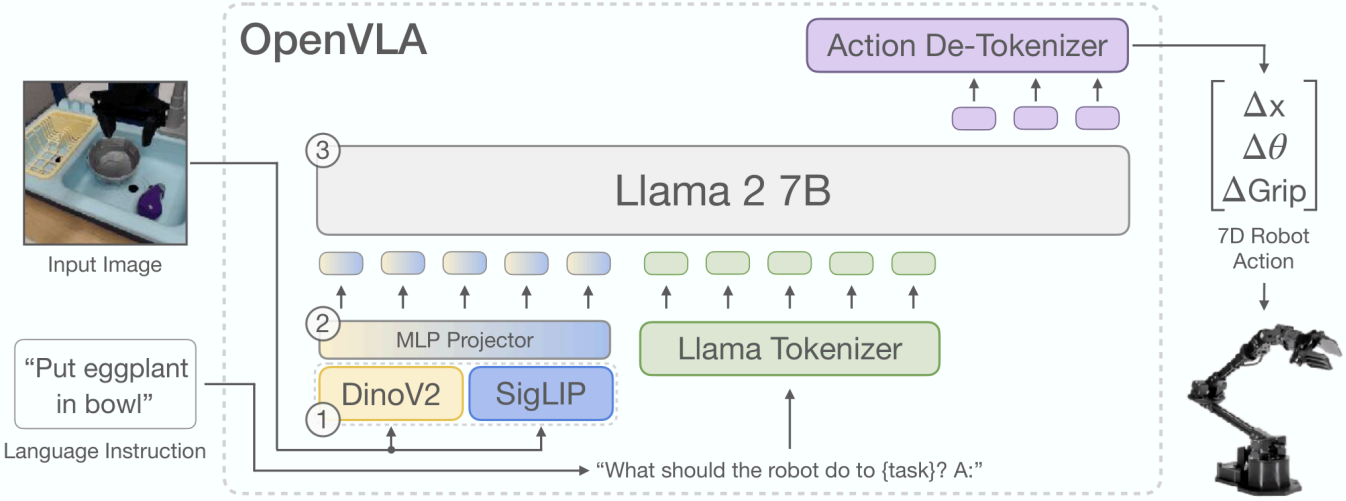

图 2：OpenVLA 架构。给定图像观测与语言指令，模型预测七维机器人控制动作。架构包含三个关键组件：（1）拼接 DINOv2 [25] 和 SigLIP [79] 特征的视觉编码器；（2）把视觉特征映射到语言嵌入空间的投影器；（3）Llama 2 的 70 亿参数大语言模型主干 [10]。

3.1 预备知识：视觉语言模型

S012 · 原文 p.4

近期大多数 VLM [42,43,20,44] 的架构包含三个主要部分（见图 2）：（1）把输入图像映射为若干“图像块嵌入”的视觉编码器；（2）把视觉编码器输出嵌入映射到语言模型输入空间的投影器；（3）大语言模型（LLM）主干。在 VLM 训练时，模型使用从多个互联网来源整理的配对或交错视觉语言数据，以预测下一个文本词元为目标进行端到端训练。

S013 · 原文 p.4

本工作以 Prismatic-7B VLM [44] 为基础。Prismatic 采用上述标准架构，含一个 6 亿参数视觉编码器、小型两层 MLP 投影器和 70 亿参数 Llama 2 语言主干 [10]。值得注意的是，其视觉编码器由预训练 SigLIP [79] 和 DINOv2 [25] 两部分组成：输入图像块分别经过两编码器，得到的特征向量再按通道拼接。与更常见的单独 CLIP [80] 或 SigLIP 编码器相比，加入 DINOv2 特征已被证明能改进空间推理 [44]，这对机器人控制可能特别有帮助。

S014 · 原文 p.4

SigLIP、DINOv2 和 Llama 2 没有公开训练数据细节；它们的训练数据很可能分别包含数万亿词元的互联网图文、纯图像和纯文本数据。Prismatic VLM 在这些组件之上使用 LLaVA 1.5 数据混合 [43] 微调；该混合来自开源数据集 [81,82,29,83,42]，共约 100 万条图文与纯文本样本。

3.2 OpenVLA 训练过程

S015 · 原文 p.5

为训练 OpenVLA，我们微调预训练 Prismatic-7B VLM 主干，使其预测机器人动作（图 2）。我们把动作预测表述为“视觉语言”任务：把输入观测图像和自然语言任务指令映射为一串预测机器人动作 [7]。为让 VLM 的语言主干预测动作，我们把连续动作映射为语言模型分词器使用的离散词元，使动作可以在 LLM 输出空间中表示。沿用 Brohan 等 [7]，我们分别把每个动作维度离散到 256 个箱之一。每维的箱宽均匀划分该维训练动作第 1 与第 99 百分位数之间的区间。与 Brohan 等 [7] 使用的最小最大边界相比，分位数可忽略异常动作，避免异常值大幅扩大离散区间、降低动作表示的有效精度。

S016 · 原文 p.5

这样，一个 N 维机器人动作变成 N 个取值在 [0,…,255] 内的离散整数。遗憾的是，OpenVLA 语言主干使用的 Llama 分词器 [10] 仅为微调时新增词元预留 100 个“特殊词元”，不足以表示 256 个动作词元。为保持简单，我们再次沿用 Brohan 等 [7] 的方法，用动作词元覆盖 Llama 词表中使用最少的 256 个词元（即末尾 256 个词元）。把动作处理为词元序列后，以标准下一词元预测目标训练 OpenVLA，但只对预测动作词元计算交叉熵损失。3.4 节讨论这一训练过程的关键设计选择，下面先介绍训练所用机器人数据。

3.3 训练数据

S017 · 原文 p.5

构建 OpenVLA 训练集的目标是覆盖丰富的机器人形态、场景与任务，使最终模型能开箱即用地控制多种机器人，并高效微调到新设置。我们以 Open X-Embodiment（OpenX）[1] 为基础整理训练数据。截至写作时，完整 OpenX 含超过 70 个独立机器人数据集与超过 200 万条轨迹；社区共同把它们汇总为一致且易用的数据格式。为使训练可行，我们对原始数据执行多步整理。

S018 · 原文 p.5

整理目标是：（1）在所有训练数据集间建立一致输入输出空间；（2）最终混合中的机器人形态、任务和场景保持平衡。脚注 2：Octo [5] 已展示跨异构感知输入数据集训练，很有前景；跨异构传感模态与动作空间的 VLA 训练留待未来研究。针对（1），沿用 [1,5]，我们仅保留至少有一个第三人称相机、采用单臂末端执行器控制的操作数据集。针对（2），对通过首轮筛选的数据集采用 Octo [5] 的混合权重。Octo 以启发式规则降低较不多样数据的权重或将其移除，提高任务与场景更丰富数据的权重；细节见 Octo Model Team 等 [5]。

S019 · 原文 p.5

我们也尝试把 Octo 发布后新增至 OpenX 的少量数据集加入混合，包括 DROID [11]，但仅赋予其保守的 10% 混合权重。实际发现 DROID 的动作词元准确率在整个训练中持续偏低，提示今后可能需要更大混合权重或模型来拟合其多样性。为不影响最终模型质量，我们在训练最后三分之一阶段从混合中移除了 DROID。所用数据集及权重完整列于附录 A。

3.4 OpenVLA 设计决策

S020 · 原文 p.5

开发 OpenVLA 时，我们先用较小规模实验探索设计选择，再启动最终模型训练。具体来说，为加快迭代、降低成本，初始实验只在 BridgeData V2 [6] 上训练和评估 OpenVLA，而非完整 OpenX 混合。以下总结主要经验。

S021 · 原文 p.6

VLM 主干。起初我们测试了多个 VLM：除 Prismatic [44] 外，还微调 IDEFICS-1 [84] 与 LLaVA [85] 预测机器人动作。在单物体场景中，LLaVA 与 IDEFICS-1 表现相当；但在包含多个物体、必须操作语言指令指定对象的任务中，LLaVA 的语言落地更好。具体而言，在 BridgeData V2 水槽环境的五个语言落地任务上，LLaVA 平均成功率比 IDEFICS-1 高 35 个百分点。微调后的 Prismatic 在简单单物体任务及多物体语言落地任务上又比 LLaVA 高约 10 个百分点。我们认为差异来自融合 SigLIP-DINOv2 主干所提供的更好空间推理（3.1 节）。此外 Prismatic 代码库模块化、易于使用，因此最终选其为 OpenVLA 主干。

S022 · 原文 p.6

图像分辨率。输入图像分辨率显著影响 VLA 训练计算需求：更高分辨率产生更多图像块词元，延长上下文，训练计算量随之二次增长。我们比较 224×224 与 384×384 像素输入，在评估中没有发现性能差异，后者训练时间却增加至三倍。因此最终 OpenVLA 使用 224×224 像素。许多 VLM 基准中提高分辨率确实有益 [44,86,87]，但我们尚未在 VLA 中观察到这种趋势。

S023 · 原文 p.6

微调视觉编码器。以往 VLM 工作发现冻结视觉编码器通常表现更好 [44]；直观上，冻结有助于保留互联网预训练学到的稳健特征。但我们发现 VLA 训练中微调视觉编码器对取得良好性能至关重要。我们的假设是，预训练视觉主干可能没有捕获场景关键部位足够精细的空间细节，因而不足以实现精确控制。

S024 · 原文 p.6

训练轮数。典型 LLM 或 VLM 训练最多遍历数据一两轮。相反，我们发现 VLA 训练需要显著更多遍历；直到训练动作词元准确率超过 95%，真实机器人性能仍持续提高。最终训练共遍历数据 27 轮。

S025 · 原文 p.6

学习率。我们在多个数量级内扫描 VLA 训练学习率，发现固定 2×10^{-5} 最好，与 VLM 预训练 [44] 使用的学习率相同。学习率预热没有带来收益。

3.5 训练与推理基础设施

S026 · 原文 p.6

最终 OpenVLA 在 64 张 A100 的集群上训练 14 天，合计约 21,500 A100 小时，批大小 2048。推理时，以 bfloat16（不量化）加载需约 15 GB GPU 显存，在单张 NVIDIA RTX 4090 上约 6 Hz，未使用编译、推测解码或其他加速技巧。量化还能进一步减少显存而不损害真实机器人任务性能，见 5.4 节；各消费级与服务器级 GPU 的推理速度见图 6。为方便使用，我们实现了远程 VLA 推理服务器，实时向机器人流式传输动作预测，使控制机器人不再要求本地拥有强计算设备。该远程推理方案随开源代码一并发布（第 4 节）。

4 OpenVLA 代码库

S027 · 原文 p.6

随模型一同发布的是模块化 PyTorch OpenVLA 代码库（https://openvla.github.io），用于训练 VLA。它从单 GPU 微调扩展到多节点集群上的十亿参数级训练，并支持自动混合精度（AMP，PyTorch [75]）、FlashAttention [76]、全分片数据并行（FSDP，Zhao 等 [77]）等现代大型 Transformer 技术。代码库直接完整支持 OpenX 数据训练、集成 Hugging Face [21] AutoModel 类，并支持 LoRA 微调 [26] 与量化推理 [88,27]。

5 实验

S028 · 原文 p.7

实验旨在检验 OpenVLA 能否作为开箱即用的强大多机器人控制策略，以及是否适合作为新机器人任务微调的初始化。具体回答以下问题：

S029 · 原文 p.7

在多种机器人及不同泛化类型上评估时，OpenVLA 与已有通用策略相比如何？

S030 · 原文 p.7

OpenVLA 能否有效微调到新的机器人设置与任务？与最先进的数据高效模仿学习方法相比如何？

S031 · 原文 p.7

能否用参数高效微调与量化降低 OpenVLA 训练和推理需求、提升可获取性？性能与计算的折中如何？

5.1 多机器人平台直接评估

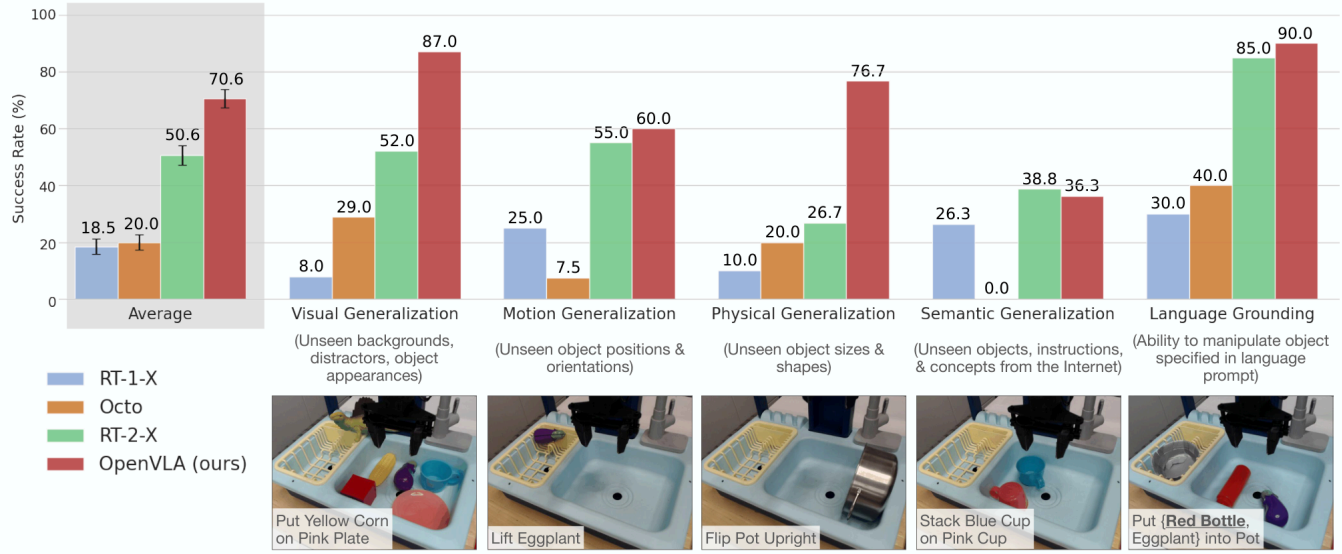

图 3：BridgeData V2 WidowX 评估任务与结果。我们在涵盖多个泛化维度以及专门测试语言条件能力的完整任务集上，评估 OpenVLA 与此前最先进的通用策略。OpenVLA 总体表现最好，除语义泛化外，在所有类别中均超越封闭 RT-2-X。平均成功率及标准误按每种方法共 170 次运行计算。详细结果见表 4。

S033 · 原文 p.7

机器人设置与任务。我们在两种形态上评估 OpenVLA 开箱即用性能：BridgeData V2 评估 [6] 的 WidowX（图 1 左），以及 RT-1、RT-2 评估 [2,7] 使用的移动操作机器人（“Google robot”，图 1 中）。两平台已广泛用于既有通用策略评估 [2,7,1,5]。每个环境都定义完整任务集，涵盖视觉（未见背景、干扰物、物体颜色与外观）、运动（未见物体位置与朝向）、物理（未见尺寸与形状）、语义（未见目标物、指令及来自互联网的概念）泛化。我们也在多物体场景测试语言条件能力，检验策略能否按用户提示操作正确目标。BridgeData V2 与 Google robot 的示例见图 3、4 底行。每种方法在 BridgeData V2 共运行 170 次（17 任务各 10 次），在 Google robot 共 60 次（12 任务各 5 次）。所有任务及其与训练数据的差异详见附录 B。为公平比较，本节及后续均采用 A/B 配对评估，即各方法使用同样任务与同组机器人及物体初态。

S034 · 原文 p.8

对比方法。我们比较 RT-1-X [1]、RT-2-X [1] 和 Octo [5] 三种既有通用操作策略。RT-1-X（3500 万参数）与 Octo（9300 万参数）是在 OpenX 子集上从零训练的 Transformer 策略；Octo 是当时开源操作策略中最先进的模型。RT-2-X（550 亿参数）是利用互联网预训练视觉语言主干的先进封闭 VLA。

S035 · 原文 p.8

BridgeData V2 结果汇总于图 3，Google robot 结果汇总于图 4，逐任务分解见附录表 4、6。RT-1-X 与 Octo 在这些任务上表现困难，尤其存在干扰物时常无法操作正确物体，有时机器人只会无目的地挥臂。需要注意，为挑战互联网预训练 VLA，我们的评估要求比这些方法原有评估更大程度的泛化，因此不具备互联网预训练的模型表现较低是预期之内。RT-2-X 明显优于 RT-1-X 和 Octo，展示大型预训练 VLM 对机器人的益处。

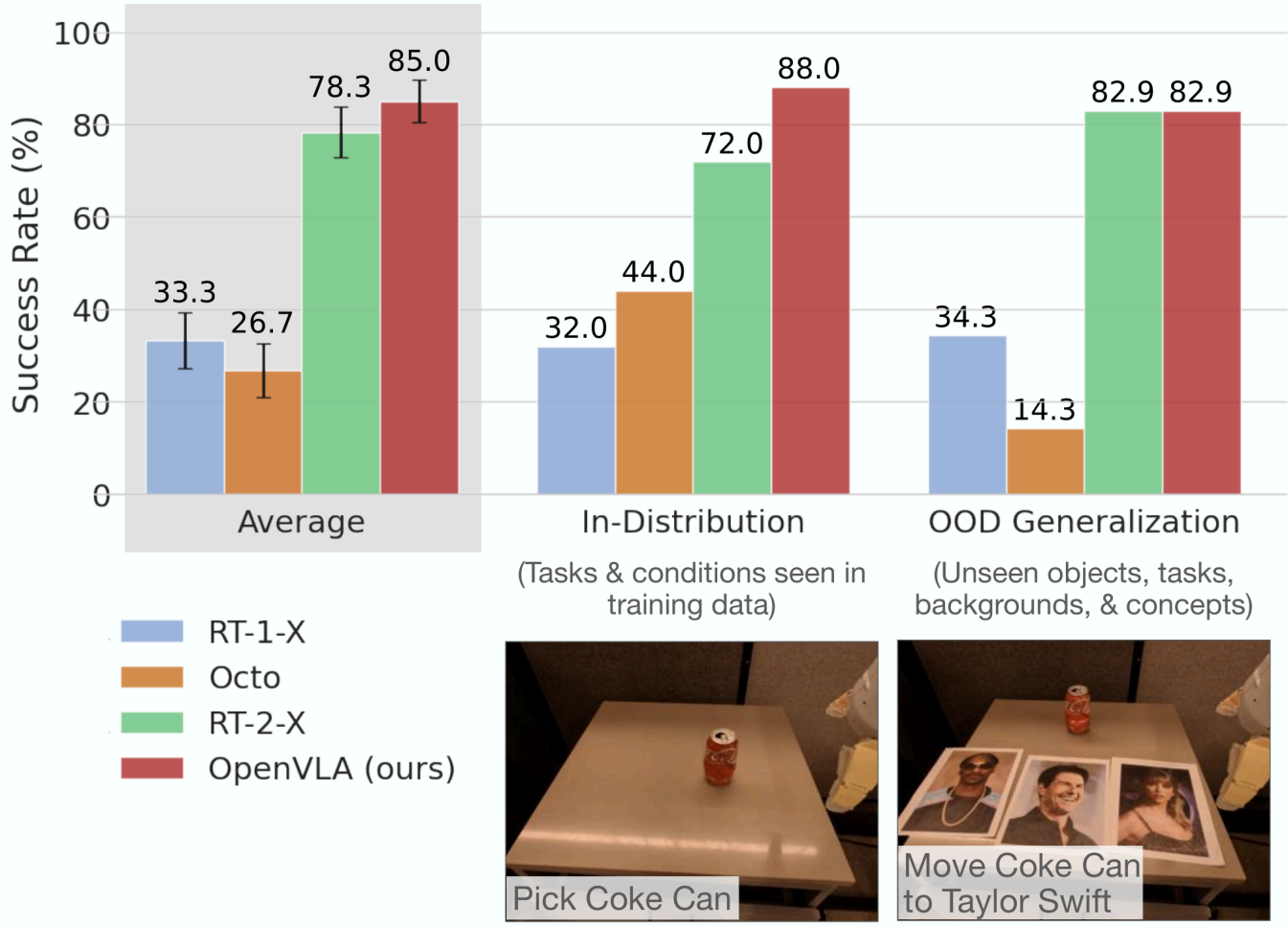

图 4：Google robot 评估结果。我们在 RT-1、RT-2 [2,7] 使用的移动操作机器人上，以分布内和分布外（OOD）任务评估通用策略。OpenVLA 与 RT-2-X 表现相当，总体显著优于 RT-1-X 与 Octo。每方法共 60 次运行，报告平均成功率及标准误。详见表 6。

S037 · 原文 p.8

值得注意的是，尽管规模约小一个数量级（70 亿对 550 亿参数），OpenVLA 在 Google robot 上与 RT-2-X 相当，在 BridgeData V2 上显著更好。定性观察表明，RT-2-X 与 OpenVLA 都比其他模型更稳健，例如有干扰物时仍接近正确物体、使末端朝向与物体对齐，甚至能从抓握不稳等错误中恢复；示例见项目网站。图 3 中 RT-2-X 的语义泛化更高，这符合预期：其互联网预训练规模更大，并把机器人动作与互联网数据共同微调，以更好保存预训练知识，而 OpenVLA 仅在机器人数据上微调。但在两环境的其他所有类别

中，OpenVLA 都相当或更好。差异可能来自多重因素：OpenVLA 使用更大训练集（97 万轨迹，对 RT-2-X 的 35 万）；进行了更细致清洗，例如滤除 Bridge 的全零动作（附录 C 详述）；并融合预训练语义与空间特征。组件消融见附录 D。

5.2 对新机器人设置的数据高效适配

S038 · 原文 p.8

以往工作主要直接评估 VLA 开箱即用性能 [16,7,1]，而对新任务、新机器人设置的有效微调很少探索，却是广泛使用的关键。本节研究 OpenVLA 能否快速适配新的真实机器人设置；仿真微调见附录 E。

S039 · 原文 p.8

机器人设置与任务。我们测试简单微调方案：用目标任务的 10–150 条示范全参数微调 OpenVLA（图 5；参数高效方法见 5.3 节）。测试两个设置：Franka-Tabletop 是固定桌面的 Franka Emika Panda 七自由度机械臂；Franka-DROID 是 DROID 数据集 [11] 的 Franka 机械臂配置，安装于可移动升降桌。两者分别采用 5 Hz 和 15 Hz 非阻塞控制器。Franka 在机器人学习社区广泛使用，因此很可能成为微调目标；不同控制频率用于检验 OpenVLA 在多类用例的适用性。

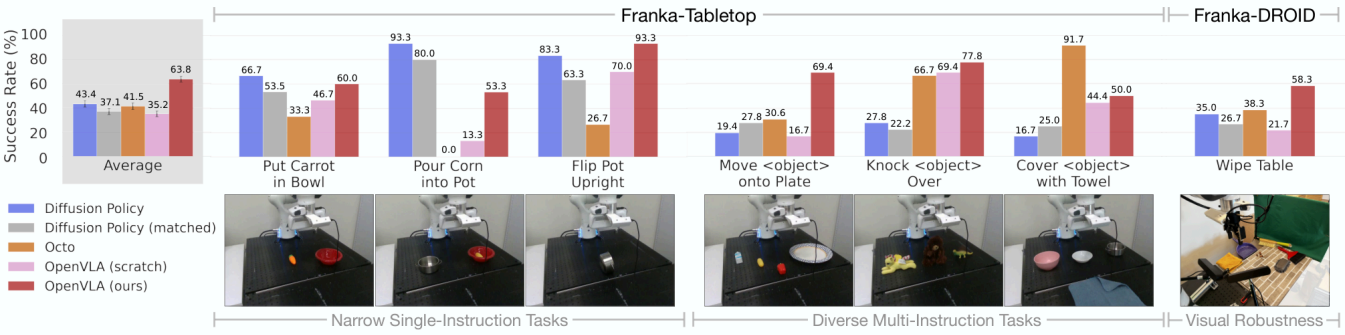

图 5：适配新机器人设置。我们在七项 Franka Emika Panda 任务（各 10–150 条示范）上，评估从零训练的先进 Diffusion Policy，以及在同样数据上微调的 Octo 与 OpenVLA。Diffusion Policy 在范围较窄的单指令任务上很强；Octo 和 OpenVLA 在涉及多指令与干扰物的多样化任务上更好。OpenVLA 在两个设置上的汇总表现最高，提示它是下游学习策略的有效默认选择。每方法共 129 次运行，其中 Franka-Tabletop 99 次、Franka-DROID 30 次，报告平均成功率及标准误；详见表 7。

S041 · 原文 p.9

对比方法。我们比较从零训练的先进数据高效模仿学习方法 Diffusion Policy [3]，及输入输出规范与 OpenVLA 相同的 Diffusion Policy（matched）。脚注 3：完整 Diffusion Policy 使用两步历史图像及本体状态，采用滚动时域控制，预测未来 T 步动作块并开环执行前 X 步后再预测；15 Hz 时按 DROID [11] 设置 T=16、X=8，5 Hz 时减为 T=8、X=3。它也是 5.2 节中唯一预测绝对笛卡尔坐标的方法，其他方法都用相对位置控制。matched 版只有单张图像输入，无本体状态或历史，预测单个相对位置动作，不做动作分块。此外，比较在目标集上微调的 Octo [5]，因为它是当时支持微调的最佳通用策略（RT-2-X 推理 API 不支持微调）；同样数据微调后的 OpenVLA 简记为 OpenVLA。最后，作为消融，我们比较 OpenVLA（scratch）：直接把底层 Prismatic VLM 微调到目标设置，而不是微调在 OpenX 预训练的 OpenVLA，以检验大规模机器人预训练的益处。

S042 · 原文 p.9

结果见图 5，逐任务数据见附录表 7。在“把胡萝卜放入碗中”“把玉米倒进锅中”等较窄的单指令任务上，两种 Diffusion Policy 都能与 Octo、OpenVLA 竞争或超越它们；但在多物体、需要语言条件的更复杂微调任务中，预训练通用策略更好。OpenX 预训练使 Octo 和 OpenVLA 更好地适应语言落地重要的多样

任务；OpenVLA（scratch）的较差表现为此提供证据。

S043 · 原文 p.9

总体上 OpenVLA 平均表现最高。以往大多方法仅在较窄单指令或多样多指令任务之一表现强，因此成功率变化很大。OpenVLA 是唯一在所有测试任务上至少取得 50% 成功率的方法，提示其是模仿学习、尤其多样语言指令任务的强默认选项。不过，对范围窄但高度灵巧的任务，Diffusion Policy 的轨迹仍更平滑精确；引入其动作分块与时间平滑可能帮助 OpenVLA 达到同等灵巧性，是未来方向（当前局限见第 6 节）。

5.3 参数高效微调

S044 · 原文 p.10

上一节全参数微调每任务需 8 张 A100 运行 5–15 小时，取决于数据规模。尽管远少于 VLA 预训练的计算量，本节仍探索计算与参数更高效的微调方法并检验有效性。

Strategy	Success Rate	Train Params (× 10 6)	VRAM (batch 16)
Full FT	**69.7 ± 7.2 %**	7,188.1	163.3 GB*
Last layer only	30.3 ± 6.1 %	465.1	51.4 GB
Frozen vision	47.0 ± 6.9 %	6,760.4	156.2 GB*
Sandwich	62.1 ± 7.9 %	914.2	64.0 GB
LoRA, rank=32	**68.2 ± 7.5%**	**97.6**	**59.7 GB**
rank=64	**68.2 ± 7.8%**	195.2	60.5 GB

表 1：参数高效微调评估。LoRA 在性能和计算间取得最佳折中，只训练 1.4% 参数却匹配全参数微调。平均成功率及标准误按选定 Franka-Tabletop 任务上每方法 33 次运行计算（详见表 8）。∗：用 FSDP [77] 分片到两张 GPU。

S046 · 原文 p.10

具体比较以下方法：全参数微调按 5.2 节更新全部权重；仅最后一层只微调 Transformer 主干最后一层和词元嵌入矩阵；冻结视觉保留视觉编码器不动、微调其他权重；三明治微调解冻视觉编码器、词元嵌入矩阵及最后一层；LoRA 则以多个秩 r，把 Hu 等 [26] 的低秩适配用于模型所有线性层。

S047 · 原文 p.10

表 1 报告多个 Franka-Tabletop 任务的微调成功率、可训练参数数和显存需求。脚注 4：5.3、5.4 节使用较小机器人混合（与 Octo 相同 OpenX 混合）预训练的 OpenVLA 版本；其架构也略小，仅有 SigLIP [79] 视觉主干，不是融合 DINOv2-SigLIP。该简化架构在微调及开箱即用任务中仍表现强。仅微调最后一层或冻结视觉效果不佳，说明视觉特征对目标场景的继续适配很重要。三明治微调因更新视觉编码器而更好，又因不更新完整 LLM 而更省显存。LoRA 则实现最佳性能/训练显存折中，优于三明治，并以仅 1.4% 参数更新匹配全微调。秩对性能影响很小，故推荐默认 r=32。LoRA 可在单张 A100 上用 10–15 小时适配新任务，比全微调计算量减少八倍。

5.4 量化实现显存高效推理

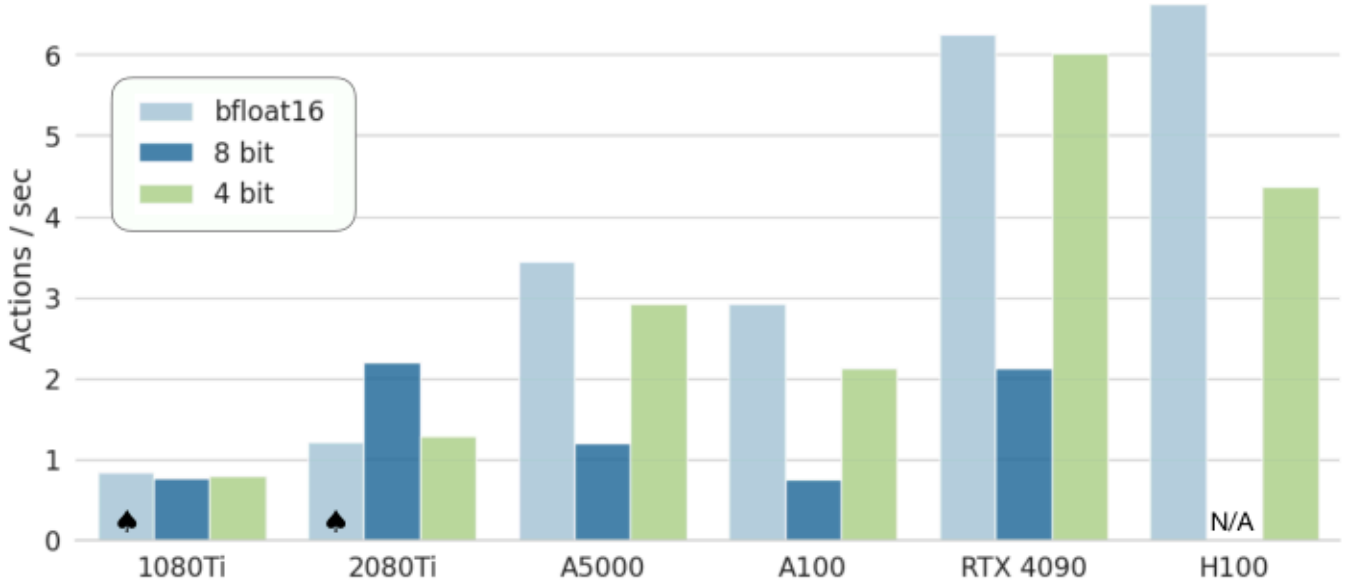

图 6：不同 GPU 上的 OpenVLA 推理速度。bfloat16 与 int4 量化均有较高吞吐，尤其是采用 Ada Lovelace 架构的 GPU（原文此处列 RTX 4090、H100）。TensorRT-LLM [89] 等现代推理框架可能进一步加速。♠：模型需分片到两张 GPU 才能容纳。译注：H100 实属 Hopper 架构，此处保留原文列举并指出该硬件分类笔误。

S049 · 原文 p.10

精度 / Bridge 成功率 / 显存：bfloat16 为 71.3±4.8%、16.8 GB；int8 为 58.1±5.1%、10.2 GB；int4 为 71.9±4.7%、7.0 GB。

精度 Precision	Bridge 成功率	显存 VRAM
bfloat16	71.3 ± 4.8%	16.8 GB
int8	58.1 ± 5.1%	10.2 GB
int4	71.9 ± 4.7%	7.0 GB

表 2：量化推理性能。四位量化的表现匹配默认 bfloat16 推理，而显存占用减少一半以上。成功率均值及标准误由 8 个代表性 BridgeData V2 任务 [6]、每方法共 80 次运行计算；详见表 5。

S051 · 原文 p.10

70 亿参数 OpenVLA 推理占用比不足一亿参数的 Octo 等既有开源通用策略更大。按 LLM 服务最佳实践，默认以 bfloat16 保存和加载模型，使占用减半，可在仅 16 GB 显存的 GPU 上部署。本节测试能否利用现代 LLM 量化技术 [88,27] 进一步降低策略推理显存、提高可获取性。这些方法以较低精度加载权重，减少内存需求，但可能牺牲速度和准确率。

S052 · 原文 p.11

具体在 8 个代表性 BridgeData V2 任务上研究八位、四位精度部署。表 2 报告显存和运行表现，图 6 报告各消费级与服务器 GPU 控制频率。八位量化因新增量化操作开销，在大多数 GPU 上反而较慢；四位因减少 GPU 内存传输，抵消量化开销，吞吐更高。

S053 · 原文 p.11

推理变慢导致八位量化表现明显下降：评估所用 A5000 上仅达 1.2 Hz，显著改变了相对于 BridgeData V2 数据采集时 5 Hz 非阻塞控制器的系统动力学。脚注 5：我们把损失归因于速度，因为在训练数据离线评估时，八位和四位词元准确率均与 bfloat16 相当；支持证据见 D.4 节。值得注意的是，四位虽需不到一半显存，却与 bfloat16 表现相近；A5000 上四位可达 3 Hz，更接近采集时动力学。

6 讨论与局限

S054 · 原文 p.11

本工作提出 OpenVLA：一个最先进的开源视觉语言动作模型，能开箱即用地取得较强跨机器人形态控制表现，也可通过参数高效微调轻松适配新设置。

S055 · 原文 p.11

当前模型有若干局限。首先，仅支持单图像观测；现实机器人设置却是异构的，可能包含多种传感输入 [5]。支持多图像、本体感知和观测历史是重要方向，探索以交错图文预训练的 VLM 可能有助于这种灵活输入适配。

S056 · 原文 p.11

其次，提高吞吐对 ALOHA [90] 等 50 Hz 高频控制至关重要，也能让 VLA 在比本工作更灵巧的双臂任务上接受检验。动作分块或推测解码 [91] 等推理优化是潜在方案。

S057 · 原文 p.11

此外，性能仍有提升空间。尽管优于既有通用策略，OpenVLA 在所测任务上通常仍低于 90% 成功率，尚不能提供很高可靠性。

S058 · 原文 p.11

最后，受计算资源限制，许多设计问题探索不足：基础 VLM 大小怎样影响性能？机器人动作数据与互联网规模视觉语言数据联合训练能否显著改善性能？哪种视觉特征最适合 VLA？希望发布模型与代码能让社区共同研究。

致谢

S059 · 原文 p.11

致谢。感谢丰田研究院提供完成研究所需的大量经费和计算资源，斯坦福基础模型研究中心提供额外计算资源，Google DeepMind 为评估提供 RT-2-X API 的早期访问。另获 Volkswagen、Physical Intelligence、ONR 项目 N00014-22-1-2621 与 N00014-22-1-2293、NSF IIS-2246811 及 DARPA ANSR 支持。

参考文献

[1] Open X-Embodiment Collaboration, A. Padalkar, A. Pooley, A. Jain, A. Bewley, A. Herzog, A. Irpan, A. Khazatsky, A. Rai, A. Singh, A. Brohan, A. Raffin, A. Wahid, B. Burgess-Limerick, B. Kim, B. Schölkopf, B. Ichter, C. Lu, C. Xu, C. Finn, C. Xu, C. Chi, C. Huang, C. Chan, C. Pan, C. Fu, C. Devin, D. Driess, D. Pathak, D. Shah, D. Büchler, D. Kalashnikov, D. Sadigh, E. Johns, F. Ceola, F. Xia, F. Stulp, G. Zhou, G. S. Sukhatme, G. Salhotra, G. Yan, G. Schiavi, H. Su, H.-S. Fang, H. Shi, H. B. Amor, H. I. Christensen, H. Furuta, H. Walke, H. Fang, I. Mordatch, I. Radosavovic, I. Leal, J. Liang, J. Kim, J. Schneider, J. Hsu, J. Bohg, J. Bingham, J. Wu, J. Wu, J. Luo, J. Gu, J. Tan, J. Oh, J. Malik, J. Tompson, J. Yang, J. J. Lim, J. Silvério, J. Han, K. Rao, K. Pertsch, K. Hausman, K. Go, K. Gopalakrishnan, K. Goldberg, K. Byrne, K. Oslund, K. Kawaharazuka, K. Zhang, K. Majd, K. Rana, K. Srinivasan, L. Y. Chen, L.

Pinto, L. Tan, L. Ott, L. Lee, M. Tomizuka, M. Du, M. Ahn, M. Zhang, M. Ding, M. K. Srirama, M. Sharma, M. J. Kim, N. Kanazawa, N. Hansen, N. Heess, N. J. Joshi, N. Suenderhauf, N. D. Palo, N. M. M. Shafiullah, O. Mees, O. Kroemer, P. R. Sanketi, P. Wohlhart, P. Xu, P. Sermanet, P. Sundaresan, Q. Vuong, R. Rafailov, R. Tian, R. Doshi, R. Martín-Martín, R. Mendonca, R. Shah, R. Hoque, R. Julian, S. Bustamante, S. Kirmani, S. Levine, S. Moore, S. Bahl, S. Dass, S. Song, S. Xu, S. Haldar, S. Adebola, S. Guist, S. Nasiriany, S. Schaal, S. Welker, S. Tian, S. Dasari, S. Belkhale, T. Osa, T. Harada, T. Matsushima, T. Xiao, T. Yu, T. Ding, T. Davchev, T. Z. Zhao, T. Armstrong, T. Darrell, V. Jain, V. Vanhoucke, W. Zhan, W. Zhou, W. Burgard, X. Chen, X. Wang, X. Zhu, X. Li, Y. Lu, Y. Chebotar, Y. Zhou, Y. Zhu, Y. Xu, Y. Wang, Y. Bisk, Y. Cho, Y. Lee, Y. Cui, Y. hua Wu, Y. Tang, Y. Zhu, Y. Li, Y. Iwasawa, Y. Matsuo, Z. Xu, and Z. J. Cui. Open X-Embodiment: Robotic learning datasets and RT-X models. https://arxiv.org/abs/2310.08864 , 2023.

[2] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, T. Jackson, S. Jesmonth, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, K.-H. Lee, S. Levine, Y. Lu, U. Malla, D. Manjunath, I. Mordatch, O. Nachum, C. Parada, J. Peralta, E. Perez, K. Pertsch, J. Quiambao, K. Rao, M. Ryoo, G. Salazar, P. Sanketi, K. Sayed, J. Singh, S. Sontakke, A. Stone, C. Tan, H. Tran, V. Vanhoucke, S. Vega, Q. Vuong, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich. Rt-1: Robotics transformer for real-world control at scale. In arXiv preprint arXiv:2212.06817 , 2022.

[3] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. In Proceedings of Robotics: Science and Systems (RSS) , 2023.

[4] A. Xie, L. Lee, T. Xiao, and C. Finn. Decomposing the generalization gap in imitation learning for visual robotic manipulation. arXiv preprint arXiv:2307.03659 , 2023.

[5] Octo Model Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, C. Xu, J. Luo, T. Kreiman, Y. Tan, D. Sadigh, C. Finn, and S. Levine. Octo: An open-source generalist robot policy. https://octo-models.github.io , 2023.

[6] H. Walke, K. Black, A. Lee, M. J. Kim, M. Du, C. Zheng, T. Zhao, P. Hansen-Estruch, Q. Vuong, A. He, V. Myers, K. Fang, C. Finn, and S. Levine. Bridgedata v2: A dataset for robot learning at scale, 2023.

[7] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, P. Florence, C. Fu, M. G. Arenas, K. Gopalakrishnan, K. Han, K. Hausman, A. Herzog, J. Hsu, B. Ichter, A. Irpan, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, L. Lee, T.-W. E. Lee, S. Levine, Y. Lu, H. Michalewski, I. Mordatch, K. Pertsch, K. Rao, K. Reymann, M. Ryoo, G. Salazar, P. Sanketi, P. Sermanet, J. Singh, A. Singh, R. Soricut, H. Tran, V. Vanhoucke, Q. Vuong, A. Wahid, S. Welker, P. Wohlhart, J. Wu, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In arXiv preprint arXiv:2307.15818 , 2023.

[8] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (ICML) , volume 139, pages 8748–8763, 2021.

[9] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. Sigmoid loss for language image pre-training. In International Conference on Computer Vision (ICCV) , 2023.

[10] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 , 2023.

[11] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis, P. D. Fagan, J. Hejna, M. Itkina, M. Lepert, Y. J. Ma, P. T. Miller, J. Wu, S. Belkhale, S. Dass, H. Ha, A. Jain, A. Lee, Y. Lee, M. Memmel, S. Park, I. Radosavovic, K. Wang, A. Zhan, K. Black, C. Chi, K. B. Hatch, S. Lin, J. Lu, J. Mercat, A. Rehman, P. R. Sanketi, A. Sharma, C. Simpson, Q. Vuong, H. R. Walke, B. Wulfe, T. Xiao, J. H. Yang, A. Yavary, T. Z. Zhao, C. Agia, R. Baijal, M. G. Castro, D. Chen, Q. Chen, T. Chung, J. Drake, E. P. Foster, J. Gao, D. A. Herrera, M. Heo, K. Hsu, J. Hu, D. Jackson, C. Le, Y. Li, K. Lin, R. Lin, Z. Ma, A. Maddukuri, S. Mirchandani, D. Morton, T. Nguyen, A. O'Neill, R. Scalise, D. Seale, V. Son, S. Tian, E. Tran, A. E. Wang, Y. Wu, A. Xie, J. Yang, P. Yin, Y. Zhang, O. Bastani, G. Berseth, J. Bohg, K. Goldberg, A. Gupta, A. Gupta, D. Jayaraman, J. J. Lim, J. Malik, R. Martín-Martín, S. Ramamoorthy, D. Sadigh, S. Song, J. Wu, M. C. Yip, Y. Zhu, T. Kollar, S. Levine, and C. Finn. Droid: A large-scale in-the-wild robot manipulation dataset. 2024.

[12] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta. R3m: A universal visual representation for robot manipulation. In CoRL , 2022.

[13] S. Karamcheti, S. Nair, A. S. Chen, T. Kollar, C. Finn, D. Sadigh, and P. Liang. Language-driven representation learning for robotics. ArXiv , abs/2302.12766, 2023. URL https://api.semanticscholar.org/CorpusID:257205716 .

[14] M. Shridhar, L. Manuelli, and D. Fox. Cliport: What and where pathways for robotic manipulation. In Conference on robot learning , pages 894–906. PMLR, 2022.

[15] A. Stone, T. Xiao, Y. Lu, K. Gopalakrishnan, K.-H. Lee, Q. Vuong, P. Wohlhart, B. Zitkovich, F. Xia, C. Finn, et al. Open-world object manipulation using pre-trained vision-language models. arXiv preprint arXiv:2303.00905 , 2023.

[16] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, et al. Palm-e: An embodied multimodal language model. arXiv preprint arXiv:2303.03378 , 2023.

[17] A. S. et al. Introducing rfm-1: Giving robots human-like reasoning capabilities, 2024. URL https://covariant.ai/insights/introducing-rfm-1-giving-robots-human-like-reasoning-capabilities/ .

[18] Wayve. Lingo-2: Driving with natural language. 2024. URL https://wayve.ai/thinking/lingo-2-driving-with-language/ .

[19] X. Chen, X. Wang, S. Changpinyo, A. J. Piergiovanni, P. Padlewski, D. M. Salz, S. Goodman, A. Grycner, B. Mustafa, L. Beyer, A. Kolesnikov, J. Puigcerver, N. Ding, K. Rong, H. Akbari, G. Mishra, L. Xue, A. V. Thapliyal, J. Bradbury, W. Kuo, M. Seyedhosseini, C. Jia, B. K. Ayan, C. Riquelme, A. Steiner, A. Angelova, X. Zhai, N. Houlsby, and R. Soricut. Pali: A jointly-scaled multilingual language-image model. ArXiv , abs/2209.06794, 2022. URL https://api.semanticscholar.org/CorpusID:252222320 .

[20] X. Chen, X. Wang, L. Beyer, A. Kolesnikov, J. Wu, P. Voigtlaender, B. Mustafa, S. Goodman, I. M. Alabdulmohsin, P. Padlewski, D. M. Salz, X. Xiong, D. Vlasic, F. Pavetic, K. Rong, T. Yu, D. Keysers, X.-Q. Zhai, and R. Soricut. PaLI-3 vision language models: Smaller, faster, stronger. arXiv preprint arXiv:2310.09199 , 2023.

[21] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, and … Transformers: State-of-the-art natural language processing. In Proceedings of the 6th International Conference on Learning Representations , 2020. URL https://arxiv.org/abs/1910.03771 .

[22] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 , 2023.

[23] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al. Mistral 7b. arXiv preprint arXiv:2310.06825 , 2023.

[24] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, et al. Gemma: Open models based on gemini research and technology. arXiv preprint arXiv:2403.08295 , 2024.

[25] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 , 2023.

[26] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 , 2021.

[27] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. Qlora: Efficient finetuning of quantized llms. Advances in Neural Information Processing Systems , 36, 2024.

[28] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In Computer Vision and Pattern Recognition (CVPR) , 2017.

[29] D. A. Hudson and C. D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In Computer Vision and Pattern Recognition (CVPR) , 2019.

[30] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, and M. Rohrbach. Towards VQA models that can read. In Computer Vision and Pattern Recognition (CVPR) , 2019.

[31] J. P. Bigham, C. Jayant, H. Ji, G. Little, A. Miller, R. C. Miller, R. Miller, A. Tatarowicz, B. White, S. White, and T. Yeh. VizWiz: nearly real-time answers to visual questions. In User Interface Software and Technology (UIST) , pages 333–342, 2010.

[32] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg. ReferItGame: Referring to objects in photographs of natural scenes. In Empirical Methods in Natural Language Processing (EMNLP) , pages 787–798, 2014.

[33] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg. Modeling context in referring expressions. In European Conference on Computer Vision (ECCV) , 2016.

[34] T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, P. Tafti, L. Hussenot, P. G. Sessa, A. Chowdhery, A. Roberts, A. Barua, A. Botev, A. Castro-Ros, A. Slone, A. Héliou, A. Tacchetti, A. Bulanova, A. Paterson, B. Tsai, B. Shahriari, C. L. Lan, C. A. Choquette-Choo, C. Crepy, D. Cer, D. Ippolito, D. Reid, E. Buchatskaya, E. Ni, E. Noland, G. Yan, G. Tucker, G.-C. Muraru, G. Rozhdestvenskiy, H. Michalewski, I. Tenney, I. Grishchenko, J. Austin, J. Keeling, J. Labanowski, J.-B. Lespiau, J. Stanway, J. Brennan, J. Chen, J. Ferret, J. Chiu, J. Mao-Jones, K. Lee, K. Yu, K. Millican, L. L. Sjoesund, L. Lee, L. Dixon, M. Reid, M. Mikuła, M. Wirth, M. Sharman, N. Chinaev, N. Thain, O. Bachem, O. Chang, O. Wahltinez, P. Bailey, P. Michel, P.

Yotov, R. Chaabouni, R. Comanescu, R. Jana, R. Anil, R. McIlroy, R. Liu, R. Mullins, S. L. Smith, S. Borgeaud, S. Girgin, S. Douglas, S. Pandya, S. Shakeri, S. De, T. Klimenko, T. Hennigan, V. Feinberg, W. Stokowiec, Y. hui Chen, Z. Ahmed, Z. Gong, T. Warkentin, L. Peran, M. Giang, C. Farabet, O. Vinyals, J. Dean, K. Kavukcuoglu, D. Hassabis, Z. Ghahramani, D. Eck, J. Barral, F. Pereira, E. Collins, A. Joulin, N. Fiedel, E. Senter, A. Andreev, and K. Kenealy. Gemma: Open models based on gemini research and technology. arXiv preprint arXiv:2403.08295 , 2024.

[35] Y. Li, S. Bubeck, R. Eldan, A. D. Giorno, S. Gunasekar, and Y. T. Lee. Textbooks are all you need ii: phi-1.5 technical report. arXiv preprint arXiv:2309.05463 , 2023.

[36] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, et al. Qwen technical report. arXiv preprint arXiv:2309.16609 , 2023.

[37] J. Li, D. Li, C. Xiong, and S. C. H. Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In International Conference on Machine Learning (ICML) , 2022.

[38] J. Li, D. Li, S. Savarese, and S. C. H. Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In International Conference on Machine Learning (ICML) , 2023.

[39] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. A. Li, P. Fung, and S. C. H. Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. arXiv preprint arXiv:2305.06500 , 2023.

[40] H. H. Tan and M. Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In Empirical Methods in Natural Language Processing (EMNLP) , 2019.

[41] H. Laurençon, L. Saulnier, L. Tronchon, S. Bekman, A. Singh, A. Lozhkov, T. Wang, S. Karamcheti, A. M. Rush, D. Kiela, M. Cord, and V. Sanh. OBELICS: An open web-scale filtered dataset of interleaved image-text documents. In Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks) , 2023.

[42] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. In Advances in Neural Information Processing Systems (NeurIPS) , 2023a.

[43] H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 , 2023b.

[44] S. Karamcheti, S. Nair, A. Balakrishna, P. Liang, T. Kollar, and D. Sadigh. Prismatic vlms: Investigating the design space of visually-conditioned language models. arXiv preprint arXiv:2402.07865 , 2024.

[45] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, et al. QT-Opt: Scalable deep reinforcement learning for vision-based robotic manipulation. arXiv preprint arXiv:1806.10293 , 2018.

[46] D. Kalashnkov, J. Varley, Y. Chebotar, B. Swanson, R. Jonschkowski, C. Finn, S. Levine, and K. Hausman. Mt-opt: Continuous multi-task robotic reinforcement learning at scale. arXiv , 2021.

[47] F. Ebert, Y. Yang, K. Schmeckpeper, B. Bucher, G. Georgakis, K. Daniilidis, C. Finn, and S. Levine. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. arXiv preprint arXiv:2109.13396 , 2021.

[48] K. Ehsani, T. Gupta, R. Hendrix, J. Salvador, L. Weihs, K.-H. Zeng, K. P. Singh, Y. Kim, W. Han, A. Herrasti, et al. Imitating shortest paths in simulation enables effective navigation and manipulation in the real world. arXiv preprint arXiv:2312.02976 , 2023.

[49] H. Bharadhwaj, J. Vakil, M. Sharma, A. Gupta, S. Tulsiani, and V. Kumar. Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking. arXiv preprint arXiv:2309.01918 , 2023.

[50] L. Pinto and A. Gupta. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. In 2016 IEEE international conference on robotics and automation (ICRA) , pages 3406–3413. IEEE, 2016.

[51] A. Mandlekar, Y. Zhu, A. Garg, J. Booher, M. Spero, A. Tung, J. Gao, J. Emmons, A. Gupta, E. Orbay, et al. Roboturk: A crowdsourcing platform for robotic skill learning through imitation. In Conference on Robot Learning , pages 879–893. PMLR, 2018.

[52] A. Gupta, A. Murali, D. P. Gandhi, and L. Pinto. Robot learning in homes: Improving generalization and reducing dataset bias. Advances in neural information processing systems , 31, 2018.

[53] S. Dasari, F. Ebert, S. Tian, S. Nair, B. Bucher, K. Schmeckpeper, S. Singh, S. Levine, and C. Finn. Robonet: Large-scale multi-robot learning. CoRL , 2019.

[54] S. Cabi, S. G. Colmenarejo, A. Novikov, K. Konyushkova, S. Reed, R. Jeong, K. Zolna, Y. Aytar, D. Budden, M. Vecerik, O. Sushkov, D. Barker, J. Scholz, M. Denil, N. de Freitas, and Z. Wang. Scaling data-driven robotics with reward sketching and batch reinforcement learning. RSS , 2019.

[55] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In Conference on Robot Learning , pages 991–1002. PMLR, 2022.

[56] H.-S. Fang, H. Fang, Z. Tang, J. Liu, C. Wang, J. Wang, H. Zhu, and C. Lu. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot. Towards Generalist Robots: Learning Paradigms for Scalable Skill Acquisition@ CoRL2023 , 3:5, 2023.

[57] C. Devin, A. Gupta, T. Darrell, P. Abbeel, and S. Levine. Learning modular neural network policies for multi-task and multi-robot transfer. In Proceedings of IEEE International Conference on Robotics and Automation , 2017.

[58] E. S. Hu, K. Huang, O. Rybkin, and D. Jayaraman. Know thyself: Transferable visual control policies through robot-awareness. In International Conference on Learning Representations , 2022.

[59] J. H. Yang, D. Sadigh, and C. Finn. Polybot: Training one policy across robots while embracing variability. In 7th Annual Conference on Robot Learning , 2023. URL https://openreview.net/forum?id=HEIRj51IcS .

[60] S. Reed, K. Zolna, E. Parisotto, S. G. Colmenarejo, A. Novikov, G. Barth-maron, M. Giménez, Y. Sulsky, J. Kay, J. T. Springenberg, T. Eccles, J. Bruce, A. Razavi, A. Edwards, N. Heess, Y. Chen, R. Hadsell, O. Vinyals, M. Bordbar, and N. de Freitas. A generalist agent. Transactions on Machine Learning Research , 2022. ISSN 2835-8856.

[61] G. Salhotra, I.-C. A. Liu, and G. Sukhatme. Bridging action space mismatch in learning from demonstrations. arXiv preprint arXiv:2304.03833 , 2023.

[62] I. Radosavovic, B. Shi, L. Fu, K. Goldberg, T. Darrell, and J. Malik. Robot learning with sensorimotor pre-training. In Conference on Robot Learning , 2023.

[63] D. Shah, A. Sridhar, A. Bhorkar, N. Hirose, and S. Levine. Gnm: A general navigation model to drive any robot. In 2023 IEEE International Conference on Robotics and Automation (ICRA) , pages 7226–7233. IEEE, 2023.

[64] K. Bousmalis, G. Vezzani, D. Rao, C. Devin, A. X. Lee, M. Bauza, T. Davchev, Y. Zhou, A. Gupta, A. Raju, et al. Robocat: A self-improving foundation agent for robotic manipulation. arXiv preprint arXiv:2306.11706 , 2023.

[65] D. Shah, A. Sridhar, N. Dashora, K. Stachowicz, K. Black, N. Hirose, and S. Levine. ViNT: A foundation model for visual navigation. In 7th Annual Conference on Robot Learning , 2023. URL https://arxiv.org/abs/2306.14846 .

[66] J. Yang, C. Glossop, A. Bhorkar, D. Shah, Q. Vuong, C. Finn, D. Sadigh, and S. Levine. Pushing the limits of cross-embodiment learning for manipulation and navigation. arXiv preprint arXiv:2402.19432 , 2024.

[67] S. Y. Gadre, M. Wortsman, G. Ilharco, L. Schmidt, and S. Song. Cows on pasture: Baselines and benchmarks for language-driven zero-shot object navigation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition , pages 23171–23181, 2023.

[68] Y. Du, K. Konyushkova, M. Denil, A. Raju, J. Landon, F. Hill, N. de Freitas, and S. Cabi. Vision-language models as success detectors. arXiv preprint arXiv:2303.07280 , 2023.

[69] Y. J. Ma, V. Kumar, A. Zhang, O. Bastani, and D. Jayaraman. Liv: Language-image representations and rewards for robotic control. In International Conference on Machine Learning , pages 23301–23320. PMLR, 2023.

[70] X. Zhang, Y. Ding, S. Amiri, H. Yang, A. Kaminski, C. Esselink, and S. Zhang. Grounding classical task planners via vision-language models. arXiv preprint arXiv:2304.08587 , 2023.

[71] S. Sontakke, J. Zhang, S. Arnold, K. Pertsch, E. Bıyık, D. Sadigh, C. Finn, and L. Itti. Roboclip: One demonstration is enough to learn robot policies. Advances in Neural Information Processing Systems , 36, 2024.

[72] J. Huang, S. Yong, X. Ma, X. Linghu, P. Li, Y. Wang, Q. Li, S.-C. Zhu, B. Jia, and S. Huang. An embodied generalist agent in 3d world. In Proceedings of the International Conference on Machine Learning (ICML) , 2024.

[73] X. Li, M. Liu, H. Zhang, C. Yu, J. Xu, H. Wu, C. Cheang, Y. Jing, W. Zhang, H. Liu, et al. Vision-language foundation models as effective robot imitators. arXiv preprint arXiv:2311.01378 , 2023.

[74] H. Zhen, X. Qiu, P. Chen, J. Yang, X. Yan, Y. Du, Y. Hong, and C. Gan. 3d-vla: 3d vision-language-action generative world model. arXiv preprint arXiv:2403.09631 , 2024.

[75] PyTorch. Automatic mixed precision. URL https://pytorch.org/docs/stable/amp.html .

[76] T. Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. arXiv preprint arXiv:2307.08691 , 2023.

[77] Y. Zhao, A. Gu, R. Varma, L. Luo, C.-C. Huang, M. Xu, L. Wright, H. Shojanazeri, M. Ott, S. Shleifer, et al. Pytorch fsdp: experiences on scaling fully sharded data parallel. arXiv preprint arXiv:2304.11277 , 2023.

[78] N. Dorka, C. Huang, T. Welschehold, and W. Burgard. What matters in employing vision language models for tokenizing actions in robot control? In First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024 .

[79] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. Sigmoid loss for language image pre-training. In Proceedings of the IEEE/CVF International Conference on Computer Vision , pages 11975–11986, 2023.

[80] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning , pages 8748–8763. PMLR, 2021.

[81] P. Sharma, N. Ding, S. Goodman, and R. Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , pages 2556–2565, 2018.

[82] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 , 2021.

[83] O. Sidorov, R. Hu, M. Rohrbach, and A. Singh. Textcaps: a dataset for image captioning with reading comprehension. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16 , pages 742–758. Springer, 2020.

[84] H. Face. Introducing idefics: An open reproduction of state-of-the-art visual langage model. Hugging Face Blog , 2024.

[85] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. Advances in neural information processing systems , 36, 2024.

[86] B. McKinzie, Z. Gan, J.-P. Fauconnier, S. Dodge, B. Zhang, P. Dufter, D. Shah, X. Du, F. Peng, F. Weers, et al. Mm1: Methods, analysis & insights from multimodal llm pre-training. arXiv preprint arXiv:2403.09611 , 2024.

[87] J. Lin, H. Yin, W. Ping, Y. Lu, P. Molchanov, A. Tao, H. Mao, J. Kautz, M. Shoeybi, and S. Han. Vila: On pre-training for visual language models. arXiv preprint arXiv:2312.07533 , 2023.

[88] T. Dettmers, M. Lewis, Y. Belkada, and L. Zettlemoyer. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. Advances in Neural Information Processing Systems , 35:30318–30332, 2022.

[89] NVIDIA. Tensorrt-llm. URL https://github.com/NVIDIA/TensorRT-LLM .

[90] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn. Learning fine-grained bimanual manipulation with low-cost hardware. arXiv preprint arXiv:2304.13705 , 2023.

[91] Y. Leviathan, M. Kalman, and Y. Matias. Fast inference from transformers via speculative decoding. In International Conference on Machine Learning , pages 19274–19286. PMLR, 2023.

[92] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. arXiv preprint arXiv:2212.06817 , 2022.

[93] E. Rosete-Beas, O. Mees, G. Kalweit, J. Boedecker, and W. Burgard. Latent plans for task agnostic offline reinforcement learning. In Proceedings of the 6th Conference on Robot Learning (CoRL) , 2022.

[94] O. Mees, J. Borja-Diaz, and W. Burgard. Grounding language with visual affordances over unstructured data. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) , London, UK, 2023.

[95] S. Dass, J. Yapeter, J. Zhang, J. Zhang, K. Pertsch, S. Nikolaidis, and J. J. Lim. CLVR jaco play dataset, 2023. URL https://github.com/clvrai/clvr_jaco_play_dataset .

[96] J. Luo, C. Xu, X. Geng, G. Feng, K. Fang, L. Tan, S. Schaal, and S. Levine. Multi-stage cable routing through hierarchical imitation learning. arXiv preprint arXiv:2307.08927 , 2023.

[97] A. Mandlekar, Y. Zhu, A. Garg, J. Booher, M. Spero, A. Tung, J. Gao, J. Emmons, A. Gupta, E. Orbay, S. Savarese, and L. Fei-Fei. RoboTurk: A crowdsourcing platform for robotic skill learning through imitation. CoRR , abs/1811.02790, 2018. URL http://arxiv.org/abs/1811.02790 .

[98] Y. Zhu, A. Joshi, P. Stone, and Y. Zhu. Viola: Imitation learning for vision-based manipulation with object proposal priors, 2023.

[99] L. Y. Chen, S. Adebola, and K. Goldberg. Berkeley UR5 demonstration dataset. https://sites.google.com/view/berkeley-ur5/home .

[100] G. Zhou, V. Dean, M. K. Srirama, A. Rajeswaran, J. Pari, K. Hatch, A. Jain, T. Yu, P. Abbeel, L. Pinto, C. Finn, and A. Gupta. Train offline, test online: A real robot learning benchmark, 2023.

[101] C. Lynch, A. Wahid, J. Tompson, T. Ding, J. Betker, R. Baruch, T. Armstrong, and P. Florence. Interactive language: Talking to robots in real time. IEEE Robotics and Automation Letters , 2023.

[102] S. Belkhale, Y. Cui, and D. Sadigh. Hydra: Hybrid robot actions for imitation learning. arxiv , 2023.

[103] Y. Zhu, P. Stone, and Y. Zhu. Bottom-up skill discovery from unsegmented demonstrations for long-horizon robot manipulation. IEEE Robotics and Automation Letters , 7(2):4126–4133, 2022.

[104] Z. J. Cui, Y. Wang, N. M. M. Shafiullah, and L. Pinto. From play to policy: Conditional behavior generation from uncurated robot data. arXiv preprint arXiv:2210.10047 , 2022.

[105] M. Heo, Y. Lee, D. Lee, and J. J. Lim. Furniturebench: Reproducible real-world benchmark for long-horizon complex manipulation. In Robotics: Science and Systems , 2023.

[106] G. Yan, K. Wu, and X. Wang. ucsd kitchens Dataset. August 2023.

[107] S. Nasiriany, T. Gao, A. Mandlekar, and Y. Zhu. Learning and retrieval from prior data for skill-based imitation learning. In Conference on Robot Learning (CoRL) , 2022.

[108] H. Liu, S. Nasiriany, L. Zhang, Z. Bao, and Y. Zhu. Robot learning on the job: Human-in-the-loop autonomy and learning during deployment. In Robotics: Science and Systems (RSS) , 2023.

[109] G. Quere, A. Hagengruber, M. Iskandar, S. Bustamante, D. Leidner, F. Stulp, and J. Vogel. Shared Control Templates for Assistive Robotics. In 2020 IEEE International Conference on Robotics and Automation (ICRA) , page 7, Paris, France, 2020.

[110] S. Saxena, M. Sharma, and O. Kroemer. Multi-resolution sensing for real-time control with vision-language models. In 7th Annual Conference on Robot Learning , 2023. URL https://openreview.net/forum?id=WuBv9-IGDUA .

[111] R. Shah, R. Martín-Martín, and Y. Zhu. MUTEX: Learning unified policies from multimodal task specifications. In 7th Annual Conference on Robot Learning , 2023. URL https://openreview.net/forum?id=PwqiqaaEzJ .

[112] X. Zhu, R. Tian, C. Xu, M. Ding, W. Zhan, and M. Tomizuka. Fanuc manipulation: A dataset for learning-based manipulation with fanuc mate 200id robot. 2023.

[113] R. Mendonca, S. Bahl, and D. Pathak. Structured world models from human videos. CoRL , 2023.

[114] J. Luo, C. Xu, F. Liu, L. Tan, Z. Lin, J. Wu, P. Abbeel, and S. Levine. Fmb: a functional manipulation benchmark for generalizable robotic learning. arXiv preprint arXiv:2401.08553 , 2024.

[115] N. M. M. Shafiullah, A. Rai, H. Etukuru, Y. Liu, I. Misra, S. Chintala, and L. Pinto. On bringing robots home, 2023.

[116] B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. Advances in Neural Information Processing Systems , 36, 2024.

[117] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 , 2019.

附录 A 数据混合细节

表 3 列出所用数据混合，主要沿用 [5]，另加入少量数据集。

OpenVLA Training Dataset Mixture	
Fractal *[92]*	12.7%
Kuka *[45]*	12.7%
Bridge*[47, 6]*	13.3%
Taco Play *[93, 94]*	3.0%
Jaco Play *[95]*	0.4%
Berkeley Cable Routing *[96]*	0.2%
Roboturk *[97]*	2.3%
Viola *[98]*	0.9%
Berkeley Autolab UR5 *[99]*	1.2%
Toto *[100]*	2.0%
Language Table *[101]*	4.4%
Stanford Hydra Dataset *[102]*	4.4%
Austin Buds Dataset *[103]*	0.2%
NYU Franka Play Dataset *[104]*	0.8%
Furniture Bench Dataset *[105]*	2.4%
UCSD Kitchen Dataset *[106]*	<0.1%
Austin Sailor Dataset *[107]*	2.2%
Austin Sirius Dataset *[108]*	1.7%
DLR EDAN Shared Control *[109]*	<0.1%
IAMLab CMU Pickup Insert *[110]*	0.9%
UTAustin Mutex *[111]*	2.2%
Berkeley Fanuc Manipulation *[112]*	0.7%
CMU Stretch *[113]*	0.2%
BC-Z *[55]*	7.5%
FMB Dataset *[114]*	7.1%
DobbE *[115]*	1.4%
DROID *[11]*	10.0%[66] 6 We remove DROID for the last third of training due to slow learning progress (see Section 3.3) and re-distribute its mixture weights across all other datasets.

表 3：OpenVLA 采用 Open X-Embodiment [1] 数据集的训练混合，主要沿用 [5]，并作少量补充。

附录 B 评估任务及详细结果

S062 · 原文 p.21

本节补充 5.1 节的 BridgeData V2 WidowX 与 Google robot 评估，以及 5.2 节的 Franka-Tabletop 与 Franka-DROID 微调评估细节。

B.1 BridgeData V2 WidowX 评估细节

S063 · 原文 p.21

这里专门讨论 5.1 节中的 BridgeData V2 评估。

B.1.1 BridgeData V2 评估任务

S064 · 原文 p.21

如 5.1 节所述，每个通用操作策略在 17 任务上各试 10 次。本节详述任务类别和各个任务。

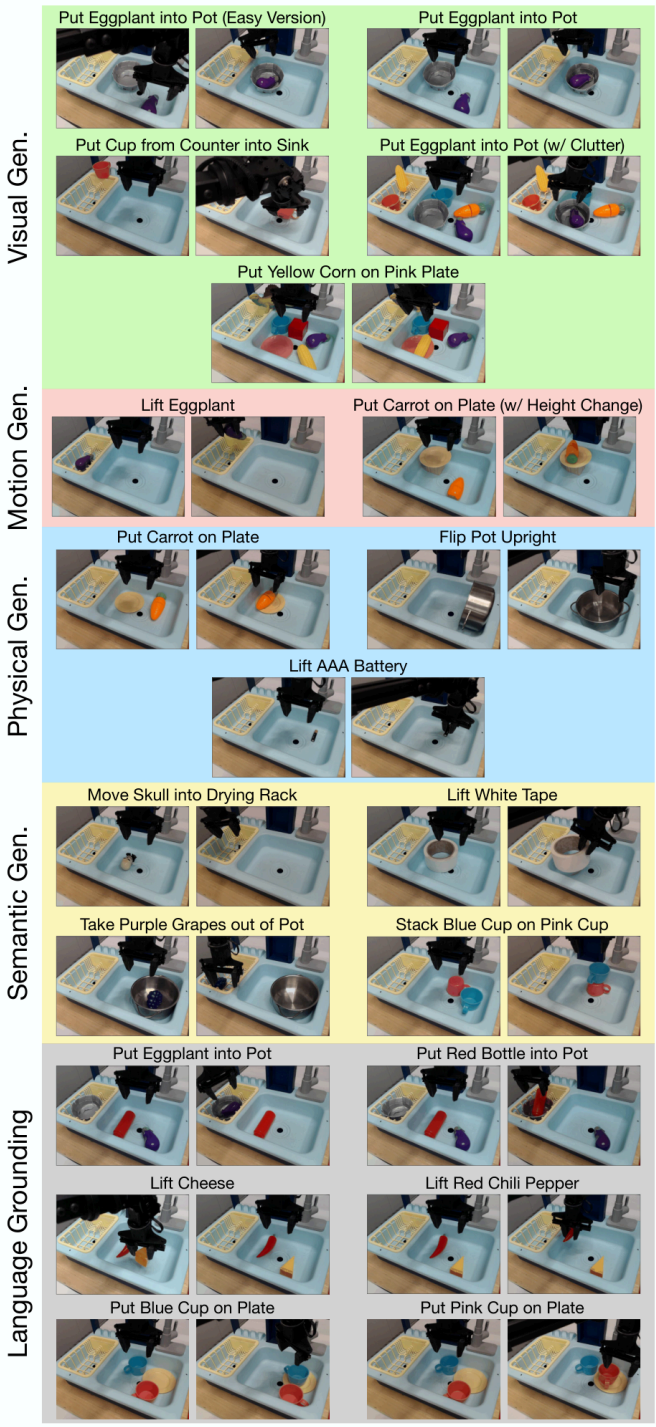

图 7：BridgeData V2 WidowX 评估任务。每个策略接受视觉、运动、物理、语义四类分布外泛化测试（定义见 5.1 节）。每对图给出初态与任务完成后的示例终态。底部三行还严格评估语言落地：固定初态、改变提示，测试策略能否接近正确目标。

S066 · 原文 p.21

共评估 5 个视觉、2 个运动、3 个物理、4 个语义泛化任务及 3 个语言落地任务。由于无法取得原始数据中的完全相同物体，所有测试都引入某种分布偏移；在不同地点重建真实环境也自然产生偏移，详见 B.1.2。全部 17 任务见图 7。每次运行记失败 0 或成功 1；部分较难任务允许部分成功 0.5，下文说明条件。

S067 · 原文 p.23

下面按图 7 的顺序描述 17 个任务：

S068 · 原文 p.23

1. 把茄子放进锅中（简单版）。机器人需抓起茄子并放进锅。因无法获取原来的锅，我们使用外观不同的手工纸锅，所以属于视觉泛化。与其余 16 任务不同，本任务运行前把末端直接置于茄子上方，故称“简单版”。

S069 · 原文 p.23

2. 把茄子放进锅中。与前项相同，但末端并非起始于茄子正上方，而在各次运行一致的固定位置，因此必须先水平伸向茄子再操作；以下所有其他任务亦如此。出于相同原因，它属于视觉泛化。

S070 · 原文 p.23

3. 把台面上的杯子放进水槽。机器人从厨房台面或沥水架抓起粉色杯子，放入右侧水槽。使用粉色而非原数据的蓝色杯子，因此属于视觉泛化；我们发现各方法都不能可靠操作蓝杯，最可能是杯子颜色与水槽混在一起。

S071 · 原文 p.23

4. 把茄子放进锅中（杂乱场景）。任务同普通版，但多个干扰物增加难度。除普通版的外观差异外，还有未见干扰物，因此属于视觉泛化。朝正确目标移动即可获 0.5 分。

S072 · 原文 p.23

5. 把黄玉米放到粉色盘子上。机器人需抓起黄玉米并放到粉盘。场景存在未见干扰物，例如水槽后部台面的绿色恐龙，因此属于视觉泛化。朝正确目标移动可获 0.5 分。

S073 · 原文 p.23

6. 抬起茄子。目标是抓住并举起茄子。其初始位置或朝向未见，机器人须超出训练分布，常要长距离伸臂，故属运动泛化。原环境训练示范未展示长距离伸臂，见 B.1.2。看似简单，许多策略实际难以完成。接触到茄子可获 0.5 分。

S074 · 原文 p.23

7. 把胡萝卜放到盘子上（高度改变）。目标是抓起胡萝卜放到黄盘。盘子高于通常的槽底位置，机器人必须调整轨迹，把胡萝卜放在抬高平台且不撞倒盘子，故属运动泛化。抓住胡萝卜并使其碰到盘子可获 0.5 分。

S075 · 原文 p.23

8. 把胡萝卜放到盘子上。与前项相同，但盘子位于水槽底部或沥水架等通常位置。胡萝卜比原训练数据所用的更大、更宽，因此属物理泛化。严格来说，上一任务也因同样胡萝卜而属于物理泛化，但其重点在运动，故归运动类。

S076 · 原文 p.24

9. 把锅翻正。目标是在回合结束时使锅在水槽中正立。所用锅比原示范中的更宽、更矮，故属物理泛化。

S077 · 原文 p.24

10. 抬起 AAA 电池。目标是抓起并举起电池。其尺寸比该环境训练示范目标小得多、细得多，故归物理泛化，详见 B.1.2。该物体也未见于该环境，技术上兼具语义泛化，但重点在物理性质。

S078 · 原文 p.24

11. 把骷髅移入沥水架。目标是抓起骷髅发条玩具，放入水槽左侧黄色沥水架。骷髅未见于 BridgeData V2 训练示范，故属语义泛化。

S079 · 原文 p.24

12. 抬起白胶带。目标是抓住并举起白色胶带卷。该目标未见于训练示范，故属语义泛化。它的形状也不同于此环境训练物体，可视为物理泛化；多数策略难以抓取环状结构，常把末端直接移入中心空洞。

S080 · 原文 p.24

13. 把紫葡萄从锅中取出。目标是抓起钢锅里的紫葡萄，将其举出或放到锅外任意位置。该语言指令在原 BridgeData V2 训练中未见，故属语义泛化。

S081 · 原文 p.24

14. 把蓝杯叠到粉杯上。目标是抓起蓝杯，稳放在粉杯上。此环境训练中未见该指令，故属语义泛化。抓住蓝杯并用它碰到粉杯可获 0.5 分。

S082 · 原文 p.24

15. 把{茄子，红瓶}放进锅中。语言落地任务：两者同时存在，机器人需将指定目标放入锅。采用配对评估，同一初态下先提示茄子、下一回合提示红瓶；每种方法对相同五个初态分别各试五次茄子和五次红瓶。朝正确目标移动可获 0.5 分。

S083 · 原文 p.24

16. 抬起{奶酪，红辣椒}。语言落地任务，目标是抓起并举起指令指定物体；按上一项方式配对评估。朝正确目标移动可获 0.5 分。

S084 · 原文 p.24

17. 把{蓝杯，粉杯}放到盘上。语言落地任务，目标是抓起指定杯子放到盘上，按其他语言任务配对评估。朝正确目标移动可获 0.5 分。

B.1.2 评估与原始训练数据比较

S085 · 原文 p.24

我们在原 BridgeData V2 [6] 使用过的水槽环境开展评估，大致复现机器人相对水槽、相机相对场景的位置。由于原始数据没有精确位置测量，无法完全重现；机器人、水槽和相机位置的细微变化自然造成分布偏移。评估地点又不同于示范采集地点，光照及水槽后方可见区域等背景也必然变化。我们也无法取得完全相同物体，因此训练与测试物体之间仍有分布偏移。

S086 · 原文 p.25

尽管存在这些挑战，OpenVLA、RT-2-X 等部分通用策略仍可开箱即用地较可靠执行多类任务。RT-1-X 和 Octo 也能完成某些任务，但在 BridgeData V2 更难的泛化测试中表现吃力。

S087 · 原文 p.25

原 BridgeData V2 在这个水槽环境中包含七项任务示范：“把锅翻正”“把胡萝卜放到盘上”“把台面（或沥水架）上的杯子放进水槽”“把茄子放进锅中”“把刀放到砧板上”“把勺子放进锅中”“把拨杆从竖直转向前方”。图 8 展示原数据示例。注意此环境所有训练示范都让末端起始于目标物正上方；但 BridgeData V2 的其他环境并非如此，某些环境初态距离目标更远，需要先水平伸向物体。



图 8：原 BridgeData V2 水槽环境任务。原数据的示范图像表明，此环境所有示范都把末端初始化在目标物正上方。这不同于图 7 的评估初态；我们的评估总把末端置于水槽上方的固定位置，而不是直接位于目标上方，唯一例外是“把茄子放进锅中（简单版）”。

S089 · 原文 p.25

BridgeData V2 评估只有“把茄子放进锅中（简单版）”让末端直接悬在目标上方，其余 16 项都从水槽上方固定位置出发，要求水平伸向目标。该初态与各类 OOD 分布偏移结合，对策略提出挑战，需要高度稳健性。因此 RT-1-X、Octo 的成功率比既有工作报告更低；但 RT-2-X、OpenVLA 在这些偏移和挑战下仍相对较强。

B.1.3 BridgeData V2 详细结果

S090 · 原文 p.26

表 4 给出全部 WidowX 结果：每个 17 任务均列出各方法 10 次试验中的成功数。OpenVLA 在大多数任务及通用策略总体成功率上最高。RT-2-X 也较好，优于 RT-1-X、Octo，但不及 OpenVLA。RT-1-X 与 Octo 普遍难以应对这些泛化任务。

Category	Task	# Trials	RT-1-X # Successes	Octo # Successes	RT-2-X # Successes	OpenVLA (ours) # Successes
Visual gen	Put Eggplant into Pot (Easy Version)	10	1	5	7	**10**
Visual gen	Put Eggplant into Pot	10	0	1	5	**10**
Visual gen	Put Cup from Counter into Sink	10	1	1	0	**7**
Visual gen	Put Eggplant into Pot (w/ Clutter)	10	1	3.5	6	**7.5**
Visual gen	Put Yellow Corn on Pink Plate	10	1	4	8	**9**
Motion gen	Lift Eggplant	10	3	0.5	6.5	**7.5**
Motion gen	Put Carrot on Plate (w/ Height Change)	10	2	1	**4.5**	**4.5**
Physical gen	Put Carrot on Plate	10	1	0	1	**8**
Physical gen	Flip Pot Upright	10	2	6	5	**8**
Physical gen	Lift AAA Battery	10	0	0	2	**7**
Semantic gen	Move Skull into Drying Rack	10	1	0	**5**	**5**
Semantic gen	Lift White Tape	10	**3**	0	0	1
Semantic gen	Take Purple Grapes out of Pot	10	**6**	0	5	4
Semantic gen	Stack Blue Cup on Pink Cup	10	0.5	0	**5.5**	4.5
Language grounding	Put {Eggplant, Red Bottle} into Pot	10	2.5	4	**8.5**	7.5
Language grounding	Lift {Cheese, Red Chili Pepper}	10	1.5	2.5	8.5	**10**
Language grounding	Put {Blue Cup, Pink Cup} on Plate	10	5	5.5	8.5	**9.5**
		Mean Success Rate	18.5±2.7%	20.0±2.6%	50.6±3.5%	**70.6±3.2%**

表 4：BridgeData V2 WidowX 详细评估。完整 17 任务（5.1 节）涵盖视觉、运动、物理、语义泛化及语言落地。某些任务允许 0.5 分部分成功，详见 B.1.1。OpenVLA 多数任务及总体最好，RT-2-X 次之。RT-1-X 和 Octo 较困难，多个任务只成功 0–2 次。任务图示见图 7。

S092 · 原文 p.26

此外，表 5 给出表 2 所汇总量化推理实验的完整结果。测试包含完整评估所有类别中的 8 个代表性 BridgeData V2 任务。

Category	Task	# Trials	bfloat16 # Successes	int8 # Successes	int4 # Successes
Visual gen	Put Eggplant into Pot (Easy Version)	10	**9**	7	**9**
Visual gen	Put Eggplant into Pot	10	**7**	**7**	**7**
Visual gen	Put Cup from Counter into Sink	10	5	3	**7**
Motion gen	Lift Eggplant	10	6	4	**7.5**
Physical gen	Put Carrot on Plate	10	6	5	**7**
Physical gen	Lift AAA Battery	10	**7**	5	3
Semantic gen	Take Purple Grapes out of Pot	10	8	8	**9**
Language grounding	Put {Eggplant, Red Bottle} into Pot	10	**9**	7.5	8
		Mean Success Rate	**71.3 ± 4.8%**	58.1 ± 5.1%	**71.9 ± 4.7%**

表 5：完整量化推理结果，是表 2 的详细版本。

B.2 Google robot 评估细节

S094 · 原文 p.26

本节补充 5.1 节 Google robot 评估细节。

B.2.1 Google robot 任务



图 9：Google robot 评估任务。每个通用策略均接受分布内及分布外泛化测试；OOD 涉及未见背景、目标物、指令 / 物体关系和语义概念，例如机器人动作数据中不存在的互联网照片。

S096 · 原文 p.26

Google robot 上每个通用策略在 12 项任务各运行 5 次，共 60 次。前五项是分布内条件，后七项是更难的 OOD 条件。全部任务见图 9，每次记失败 0 或成功 1。

S097 · 原文 p.23

下面描述 12 个任务：

S098 · 原文 p.26

1. 抓起可乐罐（分布内）。机器人位于放有可口可乐罐的平台前，目标是抓住并举起罐子。

S099 · 原文 p.27

2. 把苹果移到绿罐附近（分布内）。平台上放着苹果和绿色汽水罐，机器人需抓起苹果，将其移到绿罐旁。

S100 · 原文 p.27

3. 把蓝薯片袋移到苹果附近（分布内）。平台上放着蓝色薯片袋和苹果，机器人需抓起蓝袋并将其移近苹果。

S101 · 原文 p.27

4. 把可乐罐立正（分布内）。平台上的可乐罐横躺，机器人需抓住它并摆成竖直。

S102 · 原文 p.27

5. 打开中间抽屉（分布内）。机器人位于三个抽屉前，需抓住中间抽屉把手并拉开。

S103 · 原文 p.27

6. 把橙子移到棕薯片袋附近（OOD）。平台上放棕色薯片袋和橙子，物体下方铺蓝天白云桌布。机器人需抓橙子并移到袋旁。橙子是训练未见物体、桌布是未见背景，故属 OOD。脚注 7：Google robot OOD 条件完整列表见 Brohan 等 [7] 附录。

S104 · 原文 p.28

7. 抓起百事罐（OOD）。平台上放百事罐，其下铺亮黄 / 棕色图案桌布。机器人需抓住并举起罐子。百事罐及桌布背景都未见，故属 OOD。

S105 · 原文 p.28

8. 抓起香蕉（OOD）。平台上有苹果、可乐罐和香蕉，目标是抓起并举起香蕉；香蕉是未见目标，故属 OOD。

S106 · 原文 p.28

9. 抓起绿杯（OOD）。平台上有香蕉、百事罐和绿杯，目标是抓起并举起绿杯。所有物体都未见于训练，故属 OOD。

S107 · 原文 p.28

10. 把苹果放到盘上（OOD）。平台上有盘子和苹果，机器人需抓苹果移到盘上。训练只展示把苹果移近盘子，而不是放在盘上；指令表达了未见物体关系，故属 OOD。

S108 · 原文 p.28

11. 把香蕉放入平底锅（OOD）。平台上有平底锅和香蕉，机器人需抓香蕉放入锅。香蕉是未见目标，且指令描述未见关系，故属 OOD。

S109 · 原文 p.28

12. 把可乐罐移到 Taylor Swift 处（OOD）。平台上有可乐罐及三位名人的照片，其中包含 Taylor Swift。机器人需抓罐并移到她的照片旁；这些名人照片未见于机器人交互数据，故属 OOD。

B.2.2 Google robot 详细结果

Category	Task	# Trials	RT-1-X # Successes	Octo # Successes	RT-2-X # Successes	OpenVLA (ours) # Successes
In-distribution	Pick Coke Can	5	**5**	1	**5**	**5**
In-distribution	Move Apple near Green Can	5	3	3	3	**5**
In-distribution	Move Blue Chip Bag near Apple	5	0	3	4	**5**
In-distribution	Place Coke Can Upright	5	0	0	**4**	**4**
In-distribution	Open Middle Drawer	5	0	**4**	2	3
OOD	Move Orange near Brown Chip Bag	5	1	2	**5**	**5**
OOD	Pick Pepsi Can	5	3	0	**5**	4
OOD	Pick Banana	5	**5**	3	**5**	**5**
OOD	Pick Green Cup	5	1	0	**5**	**5**
OOD	Place Apple on Plate	5	0	0	**4**	**4**
OOD	Place Banana in Pan	5	0	0	2	**4**
OOD	Move Coke Can near Taylor Swift	5	2	0	**3**	2
		Mean Success Rate	33.3±6.1%	26.7±5.8%	**78.3±5.4%**	**85.0±4.6%**

表 6：Google robot 详细评估。完整报告 5.1 节所述 12 任务、每策略 60 次运行，涵盖分布内与 OOD。底行是平均成功率及标准误。OpenVLA 和 RT-2-X 总体均显著优于 RT-1-X 与 Octo；因前两者误差棒重叠，均值均以粗体标注。任务图示见图 9。

S111 · 原文 p.28

完整 Google robot 结果见表 6。RT-1-X 与 Octo 在评估中较困难，多个任务五次试验往往一次也不能成功。RT-2-X 与 OpenVLA 表现强，每项任务至少成功两次；二者在这套评估中相当。

B.3 数据高效适配实验细节

S112 · 原文 p.28

本节补充 5.2 节的数据高效适配实验，研究在 Franka-Tabletop 和 Franka-DROID 等新设置上微调 OpenVLA 的效果。

B.3.1 Franka-Tabletop 与 Franka-DROID 任务

S113 · 原文 p.29

七项任务各收集 10–150 条示范。前六项对应桌面固定 Franka Emika Panda（Franka-Tabletop），最后一项对应 Franka-DROID。

S114 · 原文 p.29

Franka-Tabletop 六任务中，前三项是范围较窄的单指令任务；后三项为多指令任务，场景含多个物体，机器人必须按语言选择正确目标。

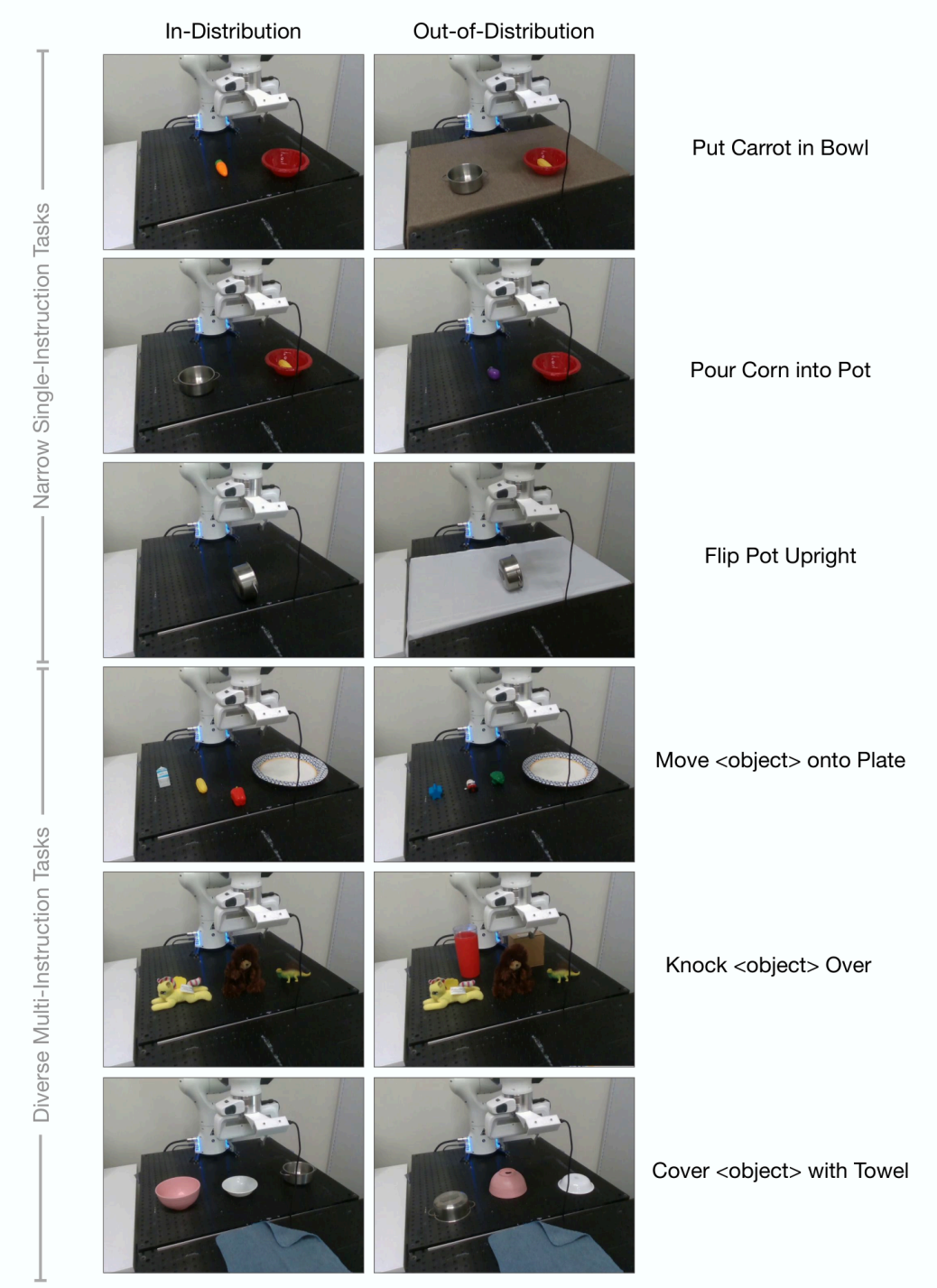

图 10：Franka-Tabletop 微调任务。上图展示 5.2 节数据高效适配中使用的任务，并在本附录详述。前三行是单指令任务，后三行是多物体、多指令任务，指令指定目标物或目标位置。第一列是匹配训练分布的初态，第二列为 OOD 初态，例如未见背景、目标、干扰物及位置 / 朝向。5.2 节每策略在分布内运行 10–12 次，OOD 运行 5–6 次。译注：原图注把“详细描述”指回图 10 自身，实际说明位于 B.3.1。

S116 · 原文 p.30

下面描述图 10 中的六个 Franka-Tabletop 任务：

S117 · 原文 p.30

1. 把胡萝卜放入碗（单指令）。机器人需抓胡萝卜放入碗。训练采集 50 条示范，每回合随机改变胡萝卜与碗在桌面的摆放位置，胡萝卜总在碗左侧。评估记成功 1 或失败 0，无部分分。

S118 · 原文 p.30

2. 把玉米倒进锅（单指令）。机器人需抓住红碗、移向钢锅并把其中黄玉米倒入。训练采集 50 条示范，每回合随机改变碗和锅位置，碗总在锅右边。评估记成功 1 或失败 0，无部分分。

S119 · 原文 p.30

3. 把锅翻正（单指令）。机器人需抓住初始竖放的钢锅，将其旋转成正常直立并放回桌面。训练仅收集 10 条示范，在桌面小区域随机改变锅位置。评估记成功 1、失败 0 或部分成功 0.5；抓住但没翻正，或撞成正立却未仔细引导，均算部分成功。满分还要求回合末释放锅。

S120 · 原文 p.30

4. 把〈物体〉移到盘上（多指令）。机器人按语言指令，从三个物体中选一个抓起，放到桌子右侧盘上。训练采集 150 条示范，随机摆不同三物体组合并选目标；盘总在右侧。评估为 1、0 或 0.5；若首次接触的物体是指令目标，但未完成任务，记部分成功。

S121 · 原文 p.30

5. 推倒〈物体〉（多指令）。机器人按指令接近三物体中的目标并推倒。训练采集 70 条示范，随机改变三物体组合并选目标。评估记 1、0 或 0.5；首次接触正确物体但未完成记部分成功。

S122 · 原文 p.30

6. 用毛巾盖住〈物体〉（多指令）。机器人需抓蓝毛巾，放到语言指定的三个目标之一上。训练采集 45 条示范，随机摆不同三物体组合。评估记 1、0 或 0.5；毛巾首次碰到正确目标但未完成，例如掉在桌上而非目标上，记部分成功。只要毛巾任一部分覆盖目标顶面即满分，无需完全盖住。

S123 · 原文 p.30

每个 Franka-Tabletop 任务，每方法评估 10–12 次分布内试验和 5–6 次 OOD 泛化试验。两类条件见图 10，其中第二列是 OOD。

S124 · 原文 p.30

六项任务的 OOD 条件如下：

S125 · 原文 p.31

1. 把胡萝卜放入碗（OOD）：用未见物体茄子替换胡萝卜。

S126 · 原文 p.31

2. 把玉米倒进锅（OOD）：桌面铺未见过的棕桌布。

S127 · 原文 p.31

3. 把锅翻正（OOD）：桌面铺未见过的白桌布。

S128 · 原文 p.31

4. 把〈物体〉移到盘上（OOD）：桌面放三个未见物体。

S129 · 原文 p.31

5. 推倒〈物体〉（OOD）：在三个见过的目标后方放未见干扰物——红塑料杯与棕盒。

S130 · 原文 p.31

6. 用毛巾盖住〈物体〉（OOD）：三个物体倒置，且放在未见位置。

S131 · 原文 p.31

最后，在 Franka-DROID 环境中测试“清扫桌面”及其变体（图 11）。机器人需抓刷子，把三个棕色小物体全扫进簸箕。训练采集 70 条示范，改变所有物体的位置。

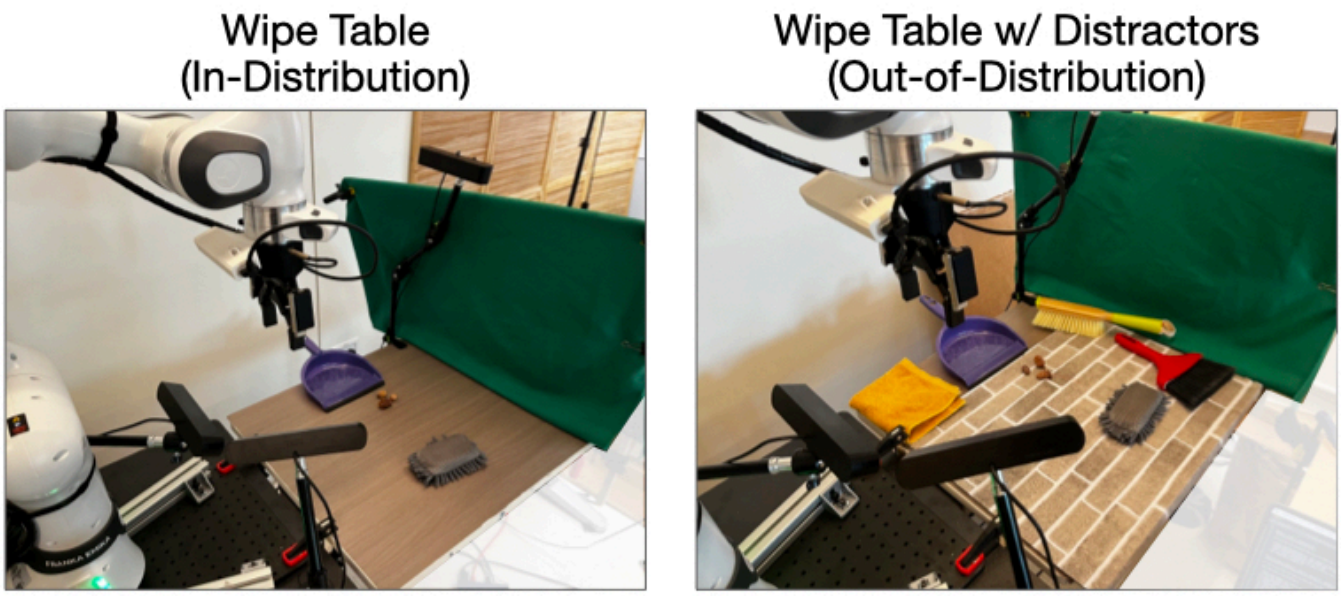

图 11：Franka-DROID 微调任务。这里的“清扫桌面”是 5.2 节数据高效适配的最后任务。左图是分布内初态，右图是桌面存在未见干扰物的 OOD。完整成功要求抓刷子，把三个物体全扫进簸箕。

S133 · 原文 p.31

测试既有匹配训练数据的分布内条件（图 11 左），也有桌面含干扰物的 OOD（右）。由于结果多样，采用以下评分：每次最高 2 分；三个物体全扫入得 2 分，扫入一或两个得 1 分，否则 0 分。每策略进行 18 次分布内和 12 次 OOD，合计满分 60 分。

B.3.2 Franka 详细评估结果

S134 · 原文 p.31

Franka-Tabletop、Franka-DROID 完整结果见表 7，评估方法见 5.2 节。Diffusion Policy 在“胡萝卜入碗”“玉米倒锅”等单指令桌面任务上很强并优于其他方法；但在“物体上盘”“推倒物体”“毛巾盖物体”等多样多指令任务中，OpenVLA 和 Octo 更好。Franka-DROID 上 OpenVLA 最佳。总体，OpenVLA 在两个设置的平均表现最高。

		# trials	Diffusion Policy	Diffusion Policy (matched)	Octo	OpenVLA (scratch)	OpenVLA (ours)
Franka-Tabletop (5Hz)	“Put Carrot in Bowl” (in-distribution)	10	**90.0%**	80.0%	40.0%	70.0%	70.0%
	“Put Carrot in Bowl” (OOD)	5	20.0%	0.0%	20.0%	0.0%	**40.0%**
	“Pour Corn into Pot” (in-distribution)	10	**100.0%**	90.0%	0.0%	10.0%	50.0%
	“Pour Corn into Pot” (OOD)	5	**80.0%**	60.0%	0.0%	20.0%	60.0%
	“Flip Pot Upright” (in-distribution)	10	**100.0%**	85.0%	40.0%	85.0%	**100.0%**
	“Flip Pot Upright” (OOD)	5	50.0%	20.0%	0.0%	40.0%	**80.0%**
	“Move <object> onto Plate” (in-distribution)	12	25.0%	25.0%	41.7%	8.3%	**75.0%**
	“Move <object> onto Plate” (OOD)	6	8.3%	33.3%	8.3%	33.3%	**58.3%**
	“Knock <object> Over” (in-distribution)	12	33.3%	25.0%	**83.3%**	75.0%	75.0%
	“Knock <object> Over” (OOD)	6	16.7%	16.7%	33.3%	58.3%	**83.3%**
	“Cover <object> with Towel” (in-distribution)	12	16.7%	20.8%	**91.7%**	41.7%	50.0%
	“Cover <object> with Towel” (OOD)	6	16.7%	33.3%	**91.7%**	50.0%	50.0%
	Average		48.5±4.9%	43.4±4.7%	43.4±4.4%	43.4±4.6%	**67.2±4.0%**
Franka-DROID (15Hz)	“Wipe Table” (in-distribution)	18	50.0%	27.8%	52.8%	25.0%	55.6%
	“Wipe Table” + Distractors (OOD)	12	12.5%	25.0%	16.7%	16.7%	62.5%
	Average		35.0±8.0%	26.7±7.5%	38.3±8.5%	21.7±6.6%	**58.3±7.2%**

表 7：数据高效适配详细结果，对应图 5 的完整分解。报告在新任务上从零训练的 Diffusion Policy，以及同样数据微调的通用策略；各策略都测分布内与 OOD，桌面条件见图 10、DROID 条件见图 11。没有一种策略所有任务都最优：Diffusion Policy 擅长单指令，OpenVLA 与 Octo 擅长多样多指令。汇总两个环境后 OpenVLA 平均成功率最高。

S136 · 原文 p.31

此外，表 8 是表 1 参数高效微调结果的详细版本。使用两项代表性桌面任务及其分布内、OOD 变体：较窄单指令“胡萝卜入碗”与多样多指令“物体上盘”。示范数与 5.2 节一致，分别 50 和 150，见 B.3.1。

		# trials	Full FT	Last layer only	Frozen vision	Sandwich	LoRA, r=32	LoRA, r=64
Franka-Tabletop (5Hz)	“Put Carrot in Bowl” (in-distribution)	10	90.0	40.0	40.0	90.0	60.0	90.0
	“Put Carrot in Bowl” (OOD)	5	40.0	0.0	40.0	0.0	60.0	40.0
	“Move <object> onto Plate” (in-distribution)	12	79.2	33.3	50.0	75.0	75.0	62.5
	“Move <object> onto Plate” (OOD)	6	41.7	33.3	58.3	41.7	75.0	66.7
	Average		**69.7±7.2%**	30.3±6.1%	47.0±6.9%	62.1±7.9%	**68.2±7.5%**	**68.2±7.8%**

表 8：参数高效微调的详细任务结果，对应表 1。

附录 C RT-2-X 与 OpenVLA 的 BridgeData V2 比较

S138 · 原文 p.32

本节补充 5.1 节 BridgeData V2 中 RT-2-X 与 OpenVLA 的比较。除此前提到 OpenVLA 使用更多 OpenX 数据和融合 SigLIP-DINOv2 而非单视觉编码器外，我们认为其在 BridgeData V2 上显著优于 RT-2-X（图 3），也源于更细致的 Bridge 预处理。

S139 · 原文 p.32

开发时发现原 BridgeData V2 有大量全零（无操作）动作转移。例如每条示范首时间步都把全零动作记为真值。不预处理就训练表达能力很强的 VLA，会导致策略频繁预测全零、评估时冻结。因此训练 OpenVLA 时，我们简单滤除每条示范首转移，已足以缓解大多数冻结。

S140 · 原文 p.32

RT-2-X 训练没有这种预处理，因此若原样部署且不修改查询程序，常会冻结，严重损害运行效果。它是专有模型，无法由我们用处理后数据重训，故用简单办法缓解：总查询第二高概率动作，因为第一高往往全零而第二高不是。这也是 RT-2-X 开发者在 Open X-Embodiment [1] 的 BridgeData V2 评估中采用的方案。它显著提高表现，但我们认为仍不及在预处理数据上重训。

S141 · 原文 p.32

我们也试过动态查询：先取最高概率动作，若全零再取第二高。但实验发现，它比始终取第二高更差。我们假设原因是动态查询改变机器人动力学：查询链路时延不可忽略，在轨迹中途暂停、重新查询会轻微打断运动，造成性能下降。因此按 Open X-Embodiment [1] 做法，报告始终查询第二高动作的 RT-2-X 性能。

附录 D 额外实验与消融

S142 · 原文 p.33

本节增加实验分析 OpenVLA 架构与训练各组件，并为前文结论提供量化证据，具体回答以下问题：

S143 · 原文 p.1

以下四个问题分别在 D.1–D.4 节展开。

S144 · 原文 p.33

1. OpenX 训练有多重要，怎样影响 OpenVLA 性能（D.1）？

S145 · 原文 p.33

2. 相比单 SigLIP，融合 SigLIP-DINOv2 编码器怎样影响性能（D.2）？

S146 · 原文 p.33

3. 应微调还是冻结 OpenVLA 视觉编码器（D.3）？

S147 · 原文 p.33

4. 把策略质量与推理速度分开后，前述量化结果如何变化（D.4）？译注：原文指向 5.3 节，量化主实验实际在 5.4 节。

S148 · 原文 p.33

下文依次讨论上述每个问题的实验设置与结果。

D.1 OpenX 训练数据消融

S149 · 原文 p.33

如 3.3 节所述，OpenVLA 在 OpenX [1] 的多形态、多场景、多任务大数据集上训练。本节消融 OpenX 混合，只在一个机器人数据集上训练 VLA，检验 OpenX 对性能的影响。5.2 节 OpenVLA（Scratch）已在微调设置中显示去掉 OpenX 的负面影响；这里增加另一形态的实验，提供进一步证据。

S150 · 原文 p.33

实验设置与任务。比较原 OpenVLA 和 OpenVLA-Bridge：后者从同一预训练 Prismatic VLM [44] 出发，仅在 BridgeData V2 [6] 微调，不使用附录 A 的完整 OpenX。两模型在 B.1.1 WidowX 任务中选出的 8 项代表任务上评估，列表见表 9。

S151 · 原文 p.33

结果。OpenX 混合消融见表 9。OpenVLA 相比 OpenVLA-Bridge，去掉 OpenX 后成功率大降 30 个百分点，说明其对最终性能重要。语言落地未受影响，但所有泛化类别均下降。这提示 OpenX 中场景、物体和任务的大幅多样化，对解锁更强泛化能力至关重要。

Category	Task	# Trials	OpenVLA # Successes	OpenVLA-Bridge # Successes	OpenVLA-Bridge-SigLIP # Successes
Visual gen	Put Eggplant into Pot (Easy Version)	10	10	8	8
Visual gen	Put Eggplant into Pot	10	10	2	3
Visual gen	Put Cup from Counter into Sink	10	7	4	2
Motion gen	Lift Eggplant	10	7.5	5.5	6.5
Physical gen	Put Carrot on Plate	10	8	4	1
Physical gen	Lift AAA Battery	10	7	2	2
Semantic gen	Take Purple Grapes out of Pot	10	4	3	3
Language grounding	Put {Eggplant, Red Bottle} into Pot	10	7.5	8	7
		Mean Success Rate	76.3 ± 4.8%	45.6 ± 5.6%	40.6 ± 5.5%

表 9：BridgeData V2 WidowX 消融。在 8 项代表任务上评估架构与训练组件。OpenVLA-Bridge 只训练 BridgeData V2、不用 OpenX；OpenVLA-Bridge-SigLIP 进一步移除 DINOv2，仅保留 SigLIP 视觉主干。OpenX 与融合视觉编码器都改善性能，但前者影响远大于后者。

D.2 双编码器与单编码器

S153 · 原文 p.33

OpenVLA 视觉主干融合 SigLIP [9] 和 DINOv2 [25]。本节去掉 DINOv2，检验双视觉编码器重要性。

S154 · 原文 p.33

实验设置与任务。OpenVLA-Bridge-SigLIP 仅用 BridgeData V2，视觉主干仅 SigLIP。与 D.1 的 OpenVLA-Bridge 比较；后者与原模型同架构，也只用 Bridge 数据。因此两者唯一差别是前者不含 DINOv2。沿用上一节相同 8 任务。

S155 · 原文 p.34

结果。双编码器消融见表 9。从 OpenVLA-Bridge 到 SigLIP 单编码器的下降说明加入 DINOv2 有益，但下降 5 个百分点远小于去掉 OpenX 的 30 个百分点。DINOv2 的低层空间特征似乎只在部分情形促进泛化。

D.3 微调与冻结视觉编码器

S156 · 原文 p.34

3.4 节提到，既有 VLM 工作冻结视觉编码器比微调更好 [44]。但训练 OpenVLA 时，我们更新包括 SigLIP-DINOv2 主干在内的全部 70 亿参数；因为开发早期发现微调视觉会带来更强 VLA，且这一发现跨多种预训练 VLM 和架构成立。下文给出细节。

S157 · 原文 p.34

实验设置与任务。我们把 Prismatic VLM [44] 仓库的两个预训练模型微调至 BridgeData V2，报告得到的两种策略。预训练模型名为 SigLIP ViT-SO 224px 和 LLaVa v1.5 7B（Reproduction）；架构和混合细节见 Karamcheti 等 [44]。任务见表 10。注意这里评估配置不同于前文 Bridge 评估，不能直接比较类似实验的数值。

S158 · 原文 p.34

结果。微调与冻结结果见表 10。两种 VLA 中，微调视觉均在多任务上显著提高成功率。定性上，有时冻结视觉策略导致明显次优且不稳定的机器人行为。因此开发早期就决定不再继续冻结视觉实验。

		SigLIP ViT-SO 224px		LLaVa v1.5 7B (Reproduction)	
Task	# Trials	Frozen Vision # Successes	Fine-Tuned # Successes	Frozen Vision # Successes	Fine-Tuned # Successes
Put Eggplant into Pot	10	7	10	5	9
Put Corn on Plate	10	10	9	0	9
	Mean Success Rate	85	**95**	25	**90**
Put { Eggplant, Red Bottle } into Pot	4	2	4	–	3
Put { Blue Cup, Pink Cup } on Plate	4	0	0	–	0
Lift { Cheese, Red Chili Pepper }	4	0	3	–	2
Put { Strawberry, Lime } into Pot	4	1	0	–	3
Move { Sushi, Grapes }	4	3	4	–	3
	Mean Success Rate	30	**55**	–	55

表 10：微调对冻结视觉实验。比较基于 Prismatic VLM [44] 仓库两个预训练 VLM 构建的 VLA，视觉更新（Fine-Tuned）与冻结（Frozen Vision）的表现。任务在同一 Bridge 水槽环境，但初态不同，因为实验发生于项目早期。微调视觉对良好性能至关重要。某些冻结实验因性能近零、行为不稳而提前停止。对两方法都完成的评估，微调平均成功率 80.0%，冻结为 46.7%。

D.4 分离策略质量与推理速度

S160 · 原文 p.34

前述量化推理比较 bfloat16、八位和四位。八位 BridgeData V2 表现较差，我们假设是其量化操作导致推理变慢。本节开展实验检验。译注：原文写 5.3 节，实际量化对比在 5.4 节。

S161 · 原文 p.34

具体再次评估三种精度，但采用阻塞控制：机器人完整执行一个动作后，策略才预测下一个并交给控制器。这样可在时延不同的方法间控制系统动力学，从而独立于预测速度测试动作质量。实际上，把更快的 bfloat16 和四位也迫使放慢，使其匹配八位部署时动力学。因此预期阻塞控制下八位应与另两者相当。

S162 · 原文 p.35

实验设置与任务。使用 D.1、D.2 同样 8 个 BridgeData V2 任务，报告阻塞控制和量化推理下的表现。

S163 · 原文 p.35

结果。表 11 给出阻塞控制结果。不同于表 2 中八位因速度慢而最差，这里排除推理速度影响后，八位与 bfloat16、四位表现相当。这支持了此前非阻塞控制下速度导致八位性能下降的假设。与前述结果一致，最低四位也未出现明显性能损失。

Category	Task	# Trials	bfloat16 # Successes	int8 # Successes	int4 # Successes
Visual gen	Put Eggplant into Pot (Easy Version)	10	10	10	10
Visual gen	Put Eggplant into Pot	10	9	10	10
Visual gen	Put Cup from Counter into Sink	10	5	5	3
Motion gen	Lift Eggplant	10	8	7	7.5
Physical gen	Put Carrot on Plate	10	10	10	10
Physical gen	Lift AAA Battery	10	3	6	4
Semantic gen	Take Purple Grapes out of Pot	10	2	2	2
Language grounding	Put {Eggplant, Red Bottle} into Pot	10	9	9.5	8.5
		Mean Success Rate	70.0 ± 5.1%	74.4 ± 4.9%	68.8 ± 5.2%

表 11：阻塞控制下量化推理。报告多个 WidowX 任务上，bfloat16（默认）、int8、int4 推理的成功率及标准误。所有平均值误差棒重叠，提示各方法表现相当。

附录 E LIBERO 仿真实验

S165 · 原文 p.36

5.2、5.3 节讨论适配新的真实机器人设置和任务。本节探索适配仿真，具体使用 LIBERO 基准 [116]。仿真实验有两项主要益处：

S166 · 原文 p.36

1. 展示通用性。尽管 OpenVLA 仅在真实机器人数据预训练，也能有效适配仿真，克服真实与仿真环境及动力学的潜在差异。

S167 · 原文 p.36

2. 提高可获取性与可复现性。将模型集成到公共仿真平台，使尤其没有机器人硬件的研究者更容易使用；仿真实验也比真实实验更容易复现。

S168 · 原文 p.36

E.1 节讨论设置，E.2 节给结果；复现材料随代码库发布。

E.1 LIBERO 实验设置

S169 · 原文 p.36

仿真设置与任务。LIBERO [116] 包含四套任务，旨在研究机器人操作终身学习，其原论文考察各种任务的前向与后向迁移。我们的实验仅关注目标套件上的监督微调，以成功示范做行为克隆，测量所得策略表现。

S170 · 原文 p.36

我们在下列四套任务上实验，每套各含 10 任务，每任务 50 条人工遥操作示范：

S171 · 原文 p.1

四套任务分别为：

S172 · 原文 p.36

LIBERO-Spatial：物体相同但布局不同，测试空间关系理解。

S173 · 原文 p.36

LIBERO-Object：场景布局相同但物体不同，测试物体类别理解。

S174 · 原文 p.36

LIBERO-Goal：物体和布局相同但目标不同，测试不同任务导向行为知识。

S175 · 原文 p.36

LIBERO-Long（又名 LIBERO-10）：物体、布局与任务多样的长时域任务。

S176 · 原文 p.36

我们对各训练数据集作如下修改：

S177 · 原文 p.36

1. 为适应需 256×256 或 224×224 像素输入的方法，我们以更高的 256×256 重新生成所有示范。原基准图像是 128×128，直接放大到 256×256 会画质很差，故先获得高分辨率图像，再按需缩小。生成方法是按提供的人类示范动作逐步运行仿真，保存模拟器渲染图像。

S178 · 原文 p.36

2. 滤除所有无操作动作，即平移、旋转量近零且不改变夹爪状态的动作。对 OpenVLA 等高表达力单步策略，这一简单清洗至关重要，否则会模仿无操作，在评估某些状态无限冻结。

S179 · 原文 p.36

3. 训练与测试的所有第三人称图像均旋转 180°，因为 LIBERO 在我们硬件上返回的图像上下颠倒。

S180 · 原文 p.37

4. 模仿学习要求示范成功，因此重放所有示范，按环境成功标准滤除未完成者。由此分别从各 500 条示范中去掉 LIBERO-Spatial 的 68 条、Object 的 46 条、Goal 的 72 条、Long 的 121 条。

S181 · 原文 p.37

5. 为公平比较，全部方法只用静态第三人称相机，不用原数据另提供的腕部相机；OpenVLA 的视觉输入仅含第三人称图像。

S182 · 原文 p.37

对比方法包括从零训练的 Diffusion Policy [3]、目标数据微调的 Octo [5]，以及按 5.3 节以 LoRA（r=32）微调的 OpenVLA。脚注 8：Diffusion Policy 使用 DROID 论文 [11] 的实现，以任务标签的 DistilBERT [117] 语言嵌入作为动作生成条件。每种策略在四套任务上分别独立训练，而不是一策略混合四套。各方法用相同示范，均受益于上述清洗。

S183 · 原文 p.37

评估细节。为减小方差，各方法每套任务评估 500 次，报告三个随机种子的平均成功率，每项统计合计 1500 次。虽修改训练集，但不改测试环境，沿用原 LIBERO 初始环境配置。

E.2 LIBERO 实验结果

S184 · 原文 p.37

LIBERO 结果见表 12。OpenVLA 适配后平均成功率与排名最高，说明它能有效适配仿真任务。但其与其他方法的差距比 5.2 节真实微调更小。我们认为原因是预训练只有真实机器人数据、没有仿真，真实与仿真的外观及动力学域差使仿真微调不如真实微调有效。另一种以大量真实数据预训练的 Octo，相比从零训练的简

单强基线 Diffusion Policy 也只小幅提高，支持此解释。我们预期在预训练混合中加入仿真数据，预训练再微调方法可获更大收益。

	LIBERO-Spatial		LIBERO-Object		LIBERO-Goal		LIBERO-Long		Average	
	SR (↑)	Rank (↓)	SR (↑)	Rank (↓)	SR (↑)	Rank (↓)	SR (↑)	Rank (↓)	SR (↑)	Rank (↓)
Diffusion Policy from scratch	78.3 ± 1.1%	3	**92.5 ± 0.7%**	1	68.3 ± 1.2%	3	50.5 ± 1.3%	3	72.4 ± 0.7%	2.5
Octo fine-tuned	78.9 ± 1.0%	2	85.7 ± 0.9%	3	**84.6 ± 0.9%**	1	51.1 ± 1.3%	2	75.1 ± 0.6%	2
OpenVLA fine-tuned (ours)	**84.7 ± 0.9%**	1	88.4 ± 0.8%	2	79.2 ± 1.0%	2	**53.7 ± 1.3%**	1	**76.5 ± 0.6%**	**1.5**

表 12：LIBERO 仿真基准。报告四套任务中各方法成功率（SR）及标准误，按三个随机种子、各 500 次运行平均，并报告每套内排名，1 最强、3 最弱。平均排名值得关注，因为它提示哪种方法更适合作为多任务默认选项；比未按套件难度归一化的平均成功率更有信息。总体上，微调 OpenVLA 的平均成功率和排名最高，其次微调 Octo，再次从零训练的 Diffusion Policy。