https://robotics-transformer2.github.io

2023-8-1

RT-2:视觉-语言-动作模型将网络知识转移到机器人控制中

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu,

Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi,

Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, 和 Brianna Zitkovich

谷歌 DeepMind。作者按字母顺序排列，贡献列在附录 A。

我们研究了如何将基于互联网规模数据训练的视觉语言模型直接整合到端到端机器人控制中，以促进泛化并实现紧急语义推理。我们的目标是使单个端到端训练模型既能学习将机器人观察映射到动作，又能享受来自网络的语言和视觉语言数据的大规模预训练的好处。为此，我们建议在机器人轨迹数据和互联网规模的视觉语言任务(如视觉问答)上共同微调最先进的视觉语言模型。与其他方法相比，我们提出了一个简单、通用的方法来实现这一目标:为了将自然语言响应和机器人动作都拟合到相同的格式中，我们将动作表示为文本令牌，并以与自然语言令牌相同的方式将它们直接合并到模型的训练集中。我们将这类模型称为视觉-语言-动作模型(VLA)，并实例化了这种模型的一个例子，我们称之为 RT-2。我们的广泛评估(6k 次评估试验)表明，我们的方法导致了高性能的机器人策略，并使 RT-2 能够从互联网规模的培训中获得一系列应急能力。这包括显著提高对新对象的泛化能力，解释机器人训练数据中不存在的命令的能力(例如将对象放置在特定数字或图标上)，以及响应用户命令执行基本推理的能力(例如拿起最小或最大的对象，或最接近另一个对象的对象)。我们进一步表明，结合思维链推理可以让 RT-2 执行多阶段语义推理，例如，找出要拿起哪个物体作为临时锤子(石头)，或者哪种饮料最适合疲惫的人(能量饮料)。

1.介 绍

在广泛的网络规模数据集上预训练的大容量模型为广泛的下游任务提供了一个有效而强大的平台:大型语言模型不仅可以实现流畅的文本生成(Anil et al.， 2023;Brohan et al.， 2022;OpenAI, 2023)，但紧急问题解决(Cobbe 等人，2021;Lewkowycz 等人，2022;Polu et al.， 2022)和散文的创造性生成(Brown et al.， 2020;OpenAI, 2023)和代码(陈等人，2021)，而视觉语言模型实现开放词汇视觉识别(Kirillov 等人，2023;Minderer et al.， 2022;Radford et al.， 2021)，甚至可以对图像中的对象-代理交互做出复杂的推断(Alayrac et al.， 2022;陈等，2023a,b;Driess 等，2023;Hao 等，2022;黄等人，2023;Wang et al.， 2022)。这种语义推理、问题解决和视觉解释能力对于必须在现实世界环境中执行各种任务的通才机器人将非常有用。然而,

通讯作者:chebotar@google.com, tianheyu@google.com, karolhausman@google.com

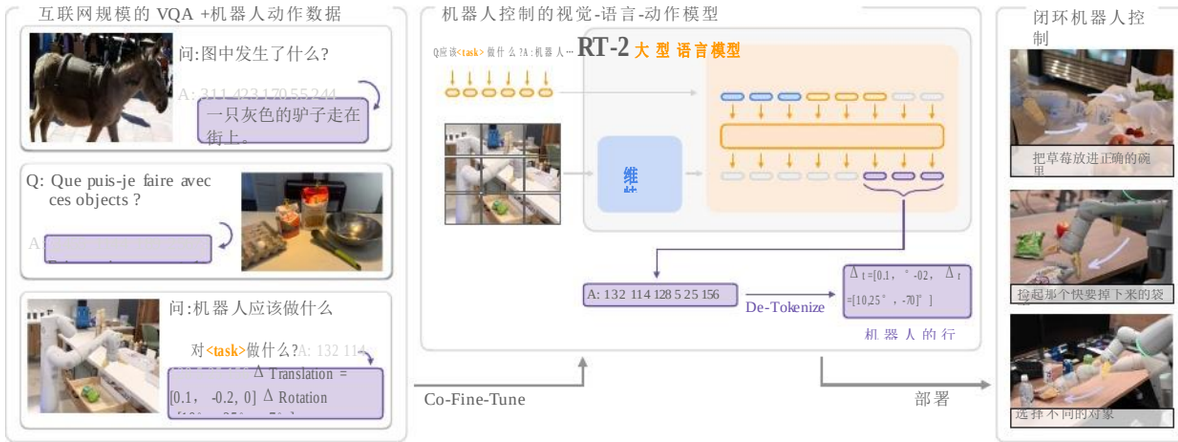


图 1 | RT-2 概述:我们将机器人动作表示为另一种语言，可以将其转换为文本令牌，并与互联网规模的视觉语言数据集一起训练。在推理过程中，文本令牌被去令牌化为机器人动作，从而实现闭环控制。这使我们能够在学习机器人策略时利用视觉语言模型的主干和预训练，将它们的一些泛化、语义理解和推理转移到机器人控制中。我们在项目网站 robotics-transformer2.github.io 上展示了 RT-2 执行的示例。

目前还不清楚机器人应该如何获得这样的能力。虽然蛮力方法可能需要收集数百万个机器人交互试验，但最强大的语言和视觉语言模型是在来自网络的数十亿个令牌和图像上进行训练的(Alayrac 等人，2022;Chen et al.， 2023a,b;

Huang et al.， 2023)——在不久的将来不太可能与机器人数据匹配。另一方面，将这样的模型直接应用到机器人任务中也很困难:这样的模型对语义、标签和文本提示进行推理，而机器人需要基于底层的动作，比如笛卡尔末端执行器命令。虽然最近的一些作品试图将语言模型(llm)和视觉语言模型(vlm)纳入机器人(Ahn et al.， 2022;Driess 等，2023;Vemprala 等人，2023)，这些方法通常只解决机器人规划的“高级”方面，本质上是充当状态机的角色，解释命令并将其解析为单个原语(例如拾取和放置对象)，然后由单独的低级控制器执行，这些低级控制器本身在训练期间不会受益于互联网规模模型的丰富语义知识。因此，在本文中我们提出了这样的问题:能否将大型预训练视觉语言模型直接集成到低级机器人控制中，以促进泛化并使紧急语义推理成为可能?

为此，我们探索了一种既简单又惊人有效的方法:我们直接训练为开放词汇视觉问答和视觉对话设计的视觉语言模型，以输出低级机器人动作，同时解决其他互联网规模的视觉语言任务。虽然这样的模型通常被训练为产生自然语言标记，但我们可以通过*将动作标记为文本标记*并创建“多模态句子”(Driess 等人，2023)来训练它们在机器人轨迹上，通过产生相应的动作来“响应”与相机观察配对的机器人指令。通过这种方式，我们可以直接训练视觉语言模型，使其按照机器人的策略充当指令。这种简单的方法与之前将 vlm 纳入机器人策略(Shridhar 等人，2022a)或从头开始设计新的视觉语言-动作架构(Reed 等人，2022)的替代方案形成对比:相反，预先存在的视觉语言模型，已经平摊了大量的计算投资，在没有任何新参数的情况下进行训练，以输出文本编码的动作。我们将这类模型称为视觉-语言-动作(VLA)模型。我们通过构建为 RT-1 提出的协议来实例化 VLA 模型(Brohan 等人，2022)，使用类似的数据集，但扩展模型以使用大型视觉语言主干。因此，我们将我们的模型称为 RT-2 (Robotics Transformer 2)。我们在图 1 中提供了概述。

我们观察到，基于这种视觉语言模型的机器人策略表现出一系列卓越的能力，将从机器人数据中学习到的物理运动与从网络数据中学习到的图像和文本的解释能力结合到一个单一的模型中。除了显著提高对新对象和语义变化指令的泛化的预期好处外，我们还观察到许多紧急能力。虽然该模型的物理技能仍然局限于机器人数据中看到的技能分布，但该模型通过使用从网络收集的知识来解释图像和语言命令，获得了以新方式部署这些技能的能力。图 2 中显示了一些示例亮点。该模型能够重新利用从机器人数据中学到的拾取和放置技能，将物体放置在语义指示的位置附近，例如特定的数字或图标，尽管这些线索不存在于机器人数据中。该模型还可以解释对象之间的关系，以确定要选择哪个对象以及将其放置在哪里，尽管机器人演示中没有提供这样的关系。此外，如果我们用思维链提示来增强命令，模型就能够做出更复杂的语义推断，比如弄清楚要拿起哪个物体作为临时锤子(石头)，或者哪种类型的饮料最适合疲惫的人(能量饮料)。

我们的主要贡献是 RT-2，这是一系列模型，源自对大型视觉语言模型的微调，这些模型是在网络规模的数据上训练的，可以直接充当可泛化和语义感知的机器人策略。我们的实验研究了在互联网数据上训练的多达 55B 个参数的模型，以及以前工作中带有指令注释的机器人轨迹(Brohan et al.， 2022)。在对 6k个机器人进行评估的过程中，我们发现 RT-2 能够显著提高对象、场景和指令的泛化能力，并展示了从网络规模的视觉语言预训练中继承的广泛的应急能力。

2.相关工作

视觉语言模型。 *视觉语言模型*(VLMs)有几个类别(Gan 等人， 2022)，其中可能有两个最相关的:(1)表示学习模型，例如 CLIP (Radford 等人， 2021)，它学习两种模式的常见嵌入，以及(2)形式为{vision, text}→{text}的视觉语言模型，它学习将视觉和语言作为输入并提供自由形式的文本。这两个类别都被用于为各种下游应用提供预训练，如对象分类(Radford 等人， 2021)、检测(Gu 等人，2021)和分割(Ghiasi 等人， 2021)。在这项工作中，我们关注的是后一个类别(Alayrac 等，2022;Chen et al.， 2023a,b;Driess 等， 2023;Hao 等， 2022;Li 等， 2023,2019;Lu et al.， 2019)。这些模型通常在许多不同的任务上进行训练，例如同时在多个数据集上进行图像字幕、视觉问答(VQA)和通用语言任务。虽然之前的工作研究 VLMs 用于广泛的问题和设置，包括机器人技术，但我们的重点是如何通过赋予 VLMs 预测机器人动作的能力，将 VLMs 的功能扩展到机器人闭环控制，从而利用 VLMs 中已经存在的知识来实现新的泛化水平。

机器人学习中的泛化。 开发能够在各种场景中广泛成功的机器人控制器是机器人研究的长期目标(Kaelbling, 2020;Smith and Coles, 1973)。实现机器人操作泛化的一种有前途的方法是从大型和多样化的数据集中学习(Dasari 等人， 2019;Levine 等人， 2018;平托和 Gupta， 2016)。通过这样做，之前的方法已经展示了机器人如何推广到新的对象实例(Finn 和 Levine, 2017;Levine 等人， 2018;Mahler 等人， 2017;Pinto and Gupta, 2016;Young 等人， 2021)，到涉及物体和技能新组合的任务(Dasari 和 Gupta, 2021;Finn et al.， 2017;James et al.， 2018;Jang 等， 2021;Yu 等人， 2018)，到新的目标或语言指令(Jang 等人， 2021;Jiang et al.， 2022;刘等人， 2022;Mees 等人， 2022;Nair 等， 2022a;Pong 等，

2019)，到具有新语义对象类别的任务(Shridhar 等人，2021;Stone 等人，2023)，以及看不见的环境(Cui 等人，2022;Du 等，2023a;Hansen et al.， 2020)。与大多数这些先前的工作不同，我们的目标是开发和研究一个单一的模型，它可以沿着所有这些轴推广到看不见的条件。我们方法的一个关键因素是利用预先训练的模型，这些模型已经暴露在比机器人看到的数据更广泛的数据中。

对机器人操作进行预训练。预训练在机器人学习中有着悠久的历史。大多数工作都集中在预训练的视觉表征上，这些视觉表征可用于初始化机器人相机观测的编码器，要么通过监督 ImageNet 分类(Shah 和 Kumar, 2021)，要么通过数据增强(Kostrikov 等人，2020;Laskin et al.， 2020a,b;Pari 等人，2021)或针对机器人控制量身定制的目标(Karamcheti 等人，2023;Ma et al.， 2022;Majumdar 等人，2023b;Nair 等，2022b;Xiao et al.， 2022b)。其他作品已经纳入了预训练的语言模型，通常作为指令编码器(Brohan et al.， 2022;Hill et al.， 2020;Jang 等人，2021;Jiang et al.， 2022;Lynch 和 Sermanet, 2020;Nair 等人，2022a;Shridhar 等人，2022b)或高层规划(Ahn 等人，2022;Driess 等，2023;Huang 等人，2022;Mu 等人，2023;Singh 等人，2023;Wu et al.， 2023)。而不是使用预训练视觉模型或预训练语言模型，我们特别考虑使用预训练的视觉语言模型(vlm)，它提供了丰富的，关于世界的基础知识。先前的工作已经研究了 vlm 在机器人中的应用(Driess 等人，2023;Du 等，2023b;Gadre 等人，2022;Karamcheti 等，2023;Shah 等，2023;Shridhar 等人，2021;Stone et al.，2023)，构成了这部作品的部分灵感来源。这些先前的方法使用 vlm 进行视觉状态表示(Karamcheti 等人，2023)，用于识别对象(Gadre 等人，2022;Stone 等人，2023)，用于高层规划(Driess 等人，2023)，或用于提供监督或成功检测(Du 等人，2023b;Ma et al.， 2023;Sumers 等人，2023;Xiao 等，2022a;Zhang et al.， 2023)。虽然 CLIPort (Shridhar 等人，2021)和 MOO (Stone 等人，2023)将预训练的 vlm 集成到端到端视觉运动操纵策略中，但两者都将重要的结构纳入了限制其适用性的策略中。值得注意的是，我们的工作不依赖于受限的 2D 动作空间，也不需要校准相机。此外，一个关键的区别是，与这些工作不同，我们利用生成语言的 vlm，并且我们公式的统一输出空间使模型权重能够在语言和操作任务之间完全共享，而不引入仅操作的模型层组件。

3.Vision-Language-Action 模型

在本节中，我们介绍了我们的模型族和设计选择，使训练 vlm 能够直接执行闭环机器人控制。首先，我们描述了我们模型的一般架构，以及它们如何从通常用于视觉语言任务的模型中派生出来。然后，我们介绍了微调大型 vlm 的方法和挑战，这些 vlm 在网络规模数据上进行预训练，以直接输出机器人动作，成为 VLA 模型。最后，我们描述了如何使这些模型适用于机器人任务，解决模型大小和推理速度方面的挑战，以实现实时控制。

3.1.预训练的视觉语言模型

视觉语言模型(Chen 等人，2023a;Driess 等人，2023)，我们在这项工作中建立的基础上，将一个或多个图像作为输入，并产生一系列符号，这些符号通常表示自然语言文本。这样的模型可以执行广泛的视觉解释和推理任务，从推断图像的组成到回答关于单个对象及其与其他对象的关系的问题(Alayrac et al.， 2022;Chen 等，2023a;Driess 等，2023;Huang et al.， 2023)。表示执行如此广泛的任务所需的知识

需要大型模型和网络规模的数据集。在这项工作中，我们采用了先前提出的两个 vlm 作为 VLA 模型:PaLI-X (Chen 等人，2023a)和 PaLM-E (Driess 等人，2023)。我们将这些模型的视觉语言动作版本称为 RT-2-PaLI-X 和 RT-2-PaLM-E。我们利用这些模型的实例化，其规模从数十亿到数百亿参数不等。我们在附录 D 中提供了这两个模型的架构的详细描述。

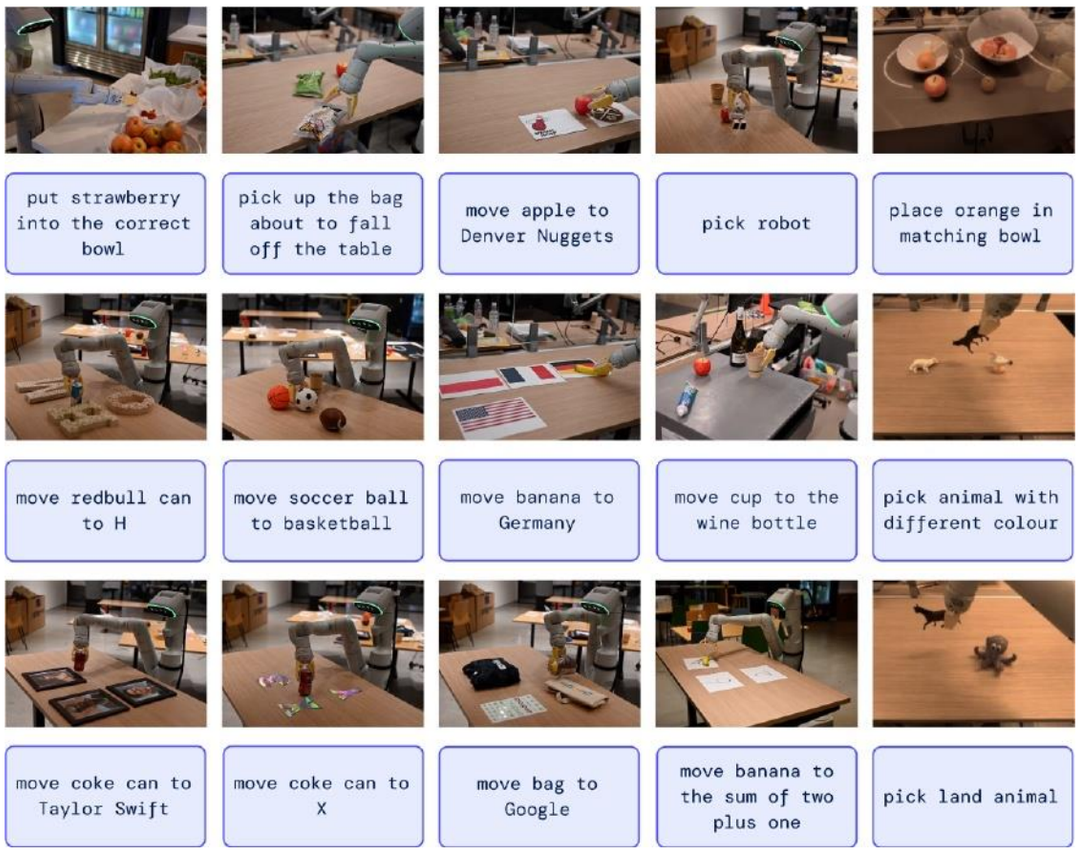


图 2 | RT-2 能够推广到各种需要推理、符号理解和人类识别的现实世界情况。我们将在第 4 节中详细研究这些具有挑战性的场景。

3.2.Robot-Action 微调

为了使视觉语言模型能够控制机器人，必须训练它们输出动作。我们对这个问题采取了直接的方法，将动作表示为模型输出中的标记，其处理方式与语言标记相同。我们的动作编码基于 Brohan 等人(2022)为 RT-1 模型提出的离散化方法。动作空间包括机器人末端执行器的 6-DoF 位置和旋转位移，以及机器人抓取器的延伸水平和用于终止事件的特殊离散命令，该命令应由策略触发以表示成功完成。连续维度(除离散终止命令外的所有维度)被统一离散成 256 个 bin。因此，机器人的动作可以用离散箱的序数表示为 8 个整数。为了使用这些离散动作将视觉语言微调为视觉语言*动作*模型，我们需要将模型*现有*的标记化中的标记与离散动作箱相关联。这需要

保留 256 个令牌作为动作令牌。选择哪些令牌取决于每个 VLM 使用的特定令牌化，我们将在本节后面讨论。为了定义 VLM 微调的目标，我们通过简单地将每个维度的动作标记与空格字符连接起来，将动作向量转换为单个字符串:

“终止 Δpos Δpos Δpos Δ 腐烂 Δ 腐烂 Δrot gripper_extension”

这样一个目标的可能实例化可以是:“1 128 91 241 5 101 127”。我们在实验中微调的两个 vlm, pal - x (Chen 等人，2023a)和 PaLM-E (Driess 等人，2023)，使用不同的标记化。对于 pal - x，每一个不超过 1000 的整数都有一个唯一的令牌，因此我们只需将动作箱与表示相应整数的令牌关联起来。对于 PaLM-E 模型，它不提供这种方便的数字表示，我们只需覆盖 256 个最不常用的符号来表示动作词汇表。值得注意的是，训练 vlm 用动作令牌覆盖现有令牌是符号调优的一种形式(Wei et al.，2023)，在之前的工作中，这已经被证明对 vlm 很有效。

采用上面描述的动作表示，我们将机器人数据转换为适合 VLM 模型微调的数据，其中我们的输入包括机器人相机图像和文本任务描述(使用标准 VQA 格式“Q:机器人应该对[任务指令]采取什么行动?”)A:”)，我们的输出被格式化为代表机器人动作的一串数字/最不常用的令牌。

Co-Fine-Tuning。正如我们将在实验中展示的那样，提高机器人性能的训练配方的一个关键技术细节是使用原始 web 数据对机器人数据*进行共同微调*，而不是仅对机器人数据进行 naïve 微调。我们注意到，协同微调导致了更一般化的策略，因为策略在微调期间暴露于来自 web 规模数据的抽象视觉概念和低级机器人动作，而不仅仅是机器人动作。在共微调期间，我们通过增加机器人数据集的采样权值来平衡每个训练批次中机器人和网络数据的比例。

输出约束。RT-2 和标准 vlm 之间的一个重要区别是，RT-2 需要输出有效的动作令牌，以便在真正的机器人上执行。因此，为了确保 RT-2 在解码过程中输出有效的动作令牌，当模型被提示机器人动作任务时，我们通过仅采样有效的动作令牌来约束其输出词汇表，而模型仍然被允许在标准视觉语言任务上输出全部自然语言令牌。

3.3.实时推理

现代 VLMs 的规模可以达到数百亿或数千亿个参数(Chen 等，2023a;Driess 等人，2023)。在这项工作中训练的最大的模型使用了 55B 参数。在标准的桌面式机器或通常用于实时机器人控制的机器人上的 gpu 上直接运行这样的模型是不可行的。据我们所知，我们的模型是有史以来最大的，超过一个数量级，用于直接闭环机器人控制，因此需要一套新的解决方案来实现高效的实时推理。我们开发了一个协议，允许我们在机器人上运行 RT-2 模型，将它们部署在一个多 tpu 云服务中，并通过网络查询该服务。通过这个解决方案，我们可以实现合适的控制频率，也可以使用相同的云服务为多个机器人提供服务。我们评估的最大的模型，55B 参数 rt -2- pal - x -55B 模型，可以在 1-3 Hz 的频率下运行。该模型的较小版本由 5B 个参数组成，可以以 5Hz 左右的频率运行。

4.实验

我们的实验专注于 RT-2 的现实泛化和应急能力，旨在回答以下问题:

1.RT-2 如何完成已知任务，更重要的是，如何在新的对象、背景和环境中进行泛化?

2.我们能观察和测量 RT-2 的应急能力吗?

3.泛化是如何随着参数计数和其他设计决策而变化的?

4.RT-2 能表现出与视觉语言模型类似的思维链推理吗?

我们在各种条件下用大约 6000 条评估轨迹评估了我们的方法和几个基线，我们将在以下章节中描述。除非另有说明，否则我们使用具有第 3.2 节中描述的动作空间的 7DoF 移动机械臂。我们还在项目网站 robotics-transformer2.github.io 上展示了 RT-2 执行的示例。我们训练了利用预训练 VLMs 的 RT-2 的两个特定实例:(1)**RT-2-PaLI-X** 由 5B 和 55B PaLI-X 构建(Chen 等人，2023a)，(2) **RT-2-PaLM-E** 由 12B PaLM-E 构建(Driess 等人，2023)。

对于训练，我们利用来自 Chen 等人(2023a)和 Driess 等人(2023)的原始网络规模数据，该数据由视觉问答、字幕和非结构化交织的图像和文本示例组成。我们将其与 Brohan 等人(2022)的机器人演示数据结合起来，该数据是在办公室厨房环境中用 13 个机器人在 17 个月内收集的。每个机器人演示轨迹都用描述执行任务的自然语言指令进行注释，该指令由描述技能的动词(例如，“pick”，“open”，“place into”)和描述操作对象的一个或多个名词(例如，“7up can”，“drawer”，“napkin”)组成(有关使用数据集的更多详细信息，请参阅附录 B)。对于所有 RT-2 训练运行，我们采用原始 PaLI-X (Chen 等人，2023a)和 PaLM-E (Driess 等人，2023)论文中的超参数，包括学习率时间表和正则化。更多的训练细节可以在附录 E 中找到。

基线。我们将我们的方法与多个最先进的基线进行比较，这些基线挑战了我们方法的不同方面。所有的基线都使用完全相同的机器人数据。为了与最先进的策略进行比较，我们使用了 **RT-1**(Brohan 等人，2022)，这是一个基于 35M 参数变压器的模型。为了与最先进的预训练表征进行比较，我们使用 VC-**1**(Majumdar 等人，2023a)和 **R3M**(Nair 等人，2022b)，并通过训练 RT-1 骨干来实现将其表征作为输入的策略。为了与使用 VLM 的其他架构进行比较，我们使用 **MOO**(Stone 等人，2023)，它使用 VLM 为语义映射创建额外的图像通道，然后将其馈送到 RT-1 主干。更多信息请参见附录 C。

4.1.RT-2 如何完成已知任务，更重要的是，如何在新的对象、背景和环境中进行泛化?

图 3 |图 4 和 6b 以及表 4 和表 6 中用于评估的示例泛化场景。

为了评估分布内性能和泛化能力，我们将 RT-2- pal - x 和 RT-2-PaLM-E 模型与前面章节中列出的四个基线进行了比较。对于视觉*任务*类别，我们使用与 RT-1 相同的视觉指令套件(Brohan 等人，2022)，其中包括 200 多个任务:36 个用于拾取物体，35 个用于敲开物体，35 个用于放置物体，48 个用于移动物体，18 个用于打开和关闭各种抽屉，36 个用于将物体取出并放入抽屉。然而，请注意，这些“分布中”评估仍然会改变物体的放置位置和诸如一天中的时间和机器人位置等因素，这需要技能来概括环境中的现实可变性。

图 3 显示了示例泛化评估，它被分为*看不见*的类别(*对象*、*背景*和*环境*)，并且另外分为简单和困难的案例。对于看不见的对象，难案例包括更难以掌握和更独特的对象(如玩具)。对于看不见的背景，硬案例包括更多样的背景和新奇的物体。最后，对于看不见的环境，硬壳对应的是一个视觉上更鲜明的办公桌面环境，有显示器和配件，而更容易的环境是厨房水槽。这些评估包括超过 280 个任务，主要侧重于在许多不同的场景中挑选和放置技能。未见类别的说明列表见附录 F.2。

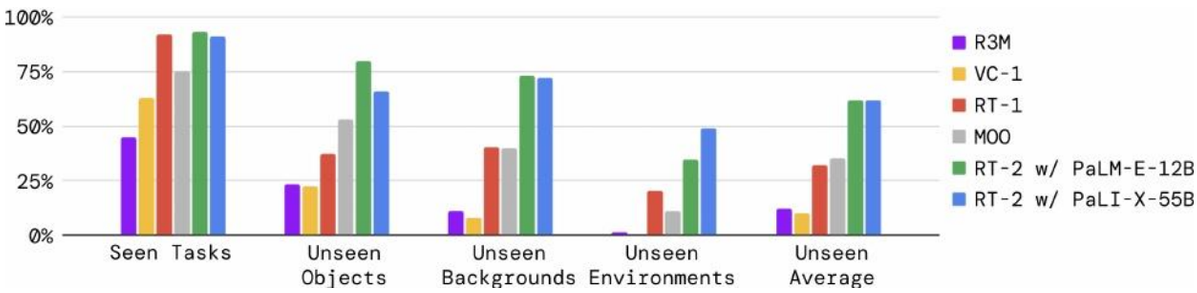


图 4 | RT-2 的两个实例的总体性能和基线跨越可见的训练任务，以及测量对新对象、新背景和新环境的泛化的未见评估。附录表 4 详细列出了完整的结果。

评价结果见图 4和附录表 4。RT-2模型和 RT-1 模型在所看到的任务上的性能相似，其他基线的成功率较低。RT-2 模型与基线之间的差异在各种泛化实验中最为明显，这表明视觉-语言-动作模型的优势在于从其互联网规模的预训练数据中转移更多可泛化的视觉和语义概念。在这里，平均而言，RT-2 的两个实例的表现相似，导致比接下来的两个基线(RT-1 和 MOO)提高约 2 倍，比其他基线好约 6倍。PaLM-E 版本的 RT-2 似乎在较难的泛化场景中比 RT-2- pal - x表现得更好，而在较容易的泛化场景中表现不佳，导致平均性能相似。

开源语言表基准。为了使用开源基线和环境提供额外的比较点，我们利用 Lynch 等人(2022)的开源语言表模拟环境。我们针对语言表数据集的几个预测任务(包括域内 VQA 任务)对一个较小的 PaLI 3B 模型进行了共同微调，并在模拟中评估了结果策略。对于动作预测任务，我们将动作离散化并编码为“X Y”格式的文本，其中 X 和 Y 的范围在{-10，-9，…， +9， +10}，并表示末端执行器的 2D 直角坐标 delta 设定值。由于其减小的尺寸，所得到的模型可以以与其他基线相似的速率(5 Hz)运行推理。本实验的结果如表 1 所示。当使用我们的模型时，与基线相比，我们观察到显著的性能提升，这表明基于 vmm 的预训练以及大型 PaLI 模型的表达能力在其他场景中是有益的，在这种情况下，使用不同的机器人进行模拟。我们还在图 5 中展示了定性的真实世界的 out-distribution 行为，展示了在这种环境中从未见过的新颖的推送任务和目标对象。关于语言表实验的更多细节可以在附录 B 和 D 中找到。

4.2.我们能观察和测量 RT-2 的应急能力吗?

除了评估视觉-语言-动作模型的泛化能力外，我们还旨在评估这些模型能够在多大程度上实现超出所演示的新功能

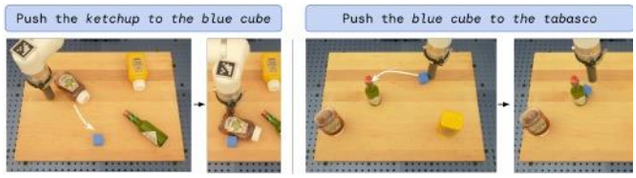


图 5 |语言表环境中真实的分布外行为。采用与 rt -2- pal - 3b 模型相同的检查点，见表 1。

模型	语言表
BC-Zero (Jang et al., 2021)	72±3
RT-1 (Brohan 等，2022)	74±13
熔岩(Lynch et al., 2022)	77±4
RT-2-PaLI-3B(我们的)	**90±10**

表 1 |在模拟 Language-Table 任务上的表现(Lynch and Ser- manet, 2020)。

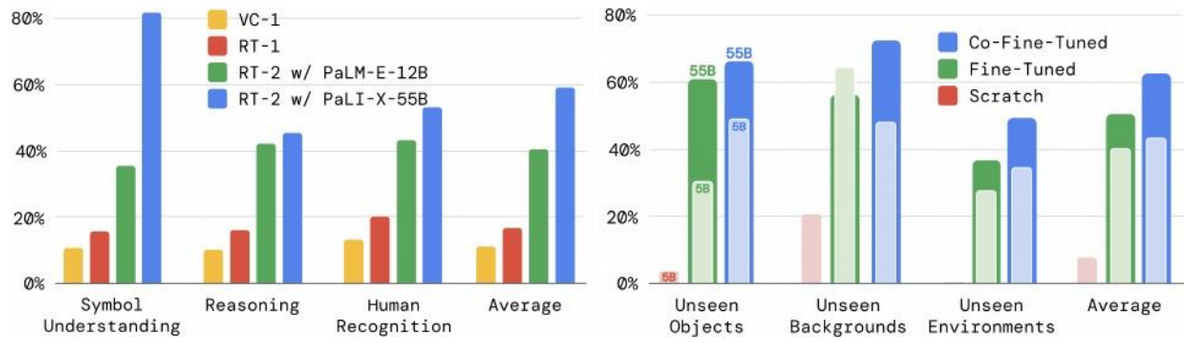
在机器人数据中通过网络传递知识。我们将这种能力称为*涌现*能力，因为它们是通过转移互联网规模的预训练而出现的。我们不期望这种转移能够实现新的机器人*运动*，但我们确实期望语义和视觉概念(包括关系和名词)能够有效地转移，即使在机器人数据中没有看到这些概念的情况下也是如此。

定性评估。首先，我们用我们的 RT-2- pal - x 模型进行实验，以确定从视觉语言概念转移的各种紧急能力。我们在图 2 中展示了此类交互的一些示例。通过我们的探索，我们发现 RT-2 在场景语境中的语义理解和基本推理方面继承了新的能力。例如，完成“把草莓放进正确的碗里”这个任务，不仅需要对草莓和碗是什么有细致入微的理解，还需要在场景的上下文中进行推理，知道草莓应该和类似的水果搭配。对于“捡起即将从桌子上掉下来的袋子”的任务，RT-2 展示了物理理解能力，以消除两个袋子之间的歧义，并识别不稳定放置的物体。在这些场景中测试的所有交互都从未在机器人数据中出现过，这表明了视觉语言数据中语义知识的转移。

定量评估。为了量化这些应急能力，我们从之前的评估中选取了前两个基线，RT-1 和 VC-1，并将它们与我们的两个模型:RT-2- pal - x 和 RT-2-PaLM-E 进行比较。为了减少这些实验的方差，我们使用 A/B 测试框架(Fisher, 1936)对所有方法进行评估，其中在完全相同的条件下对所有四种模型进行逐一评估。

我们将 RT-2 的紧急能力分为三类，包括推理和语义理解(每种能力的示例见附录图 8)。第一类是*符号理解*，它明确测试 RT-2 策略是否从视觉语言预训练中转移语义知识，而这些知识不存在于任何机器人数据中。这一类别中的示例指令是“将苹果移到 3”或“将可乐罐推到心脏顶部”。第二类我们称之为*推理*，它展示了将底层 VLM 推理的各个方面应用于控制任务的能力。这些任务需要视觉推理(“把苹果移到颜色相同的杯子里”)、数学(“把 X 移到接近二加一的和的地方”)和多语言理解(“mueve la manzana al vaso verde”)。我们将最后一类任务称为*人类识别*任务，其中包括“将可乐罐移到戴眼镜的人那里”等任务，以展示以人类为中心的理解和识别。用于此评估的完整指令列表在附录 F.2 中有详细说明。

我们在图 6a 中给出了这个实验的结果，所有的数值结果都在附录 H.2 中。我们观察到我们的 VLA 模型在所有类别中都明显优于基线，我们最好的 RT-2- pal - x 模型比下一个最好的基线(RT-1)实现了 3 倍以上的平均成功率。我们还注意到，虽然较大的基于 pali - x 的模型在平均意义上具有更好的符号理解、推理和人物识别性能，但较小的基于 palm 的模型在涉及数学推理的任务上具有优势。我们将这一有趣的结果归因于 PaLM-E 中使用的不同预训练混合物，这使得模型在数学计算方面比大多数视觉预训练的 PaLI-X 更有能力。



(a)各种紧急技能评估的性能比较- (b) RT-2- pal - x的消融，显示 RT-2 和两个基线之间参数的影响(图 8)。Eter计数和泛化的训练策略。

图 6 | RT-2 在(6a)紧急技能和(6b)大小和训练消融方面的定量表现。附录表 5 和表 6 详细列出了完整的数值结果。

4.3.泛化是如何随着参数计数和其他设计决策而变化的？

对于这个比较，我们使用 RT-2-PaLI-X 模型，因为它在模型大小方面具有灵活性(由于 PaLM-E 的性质，RT-2-PaLM-E 仅限于 PaLM 和 ViT 模型的特定大小)。特别地，我们比较了两种不同的模型大小，5B 和 55B，以及三种不同的训练例程:从头开始训练模型，不使用来自 VLM 预训练的任何权重;只使用机器人动作数据对预训练模型进行微调;以及协同微调(协同训练与微调)，这是本工作中使用的主要方法，我们同时使用原始 VLM 训练数据和机器人数据进行 VLM 微调。由于我们最感兴趣的是这些模型的泛化方面，因此我们从这组实验中删除了*看到的任务*评估。

消去的结果如图 6b 和附录表 6 所示。首先，我们观察到，即使对于 5B 模型，从头开始训练一个非常大的模型也会导致非常差的性能。鉴于此结果，我们决定跳过评估一个更大的 55B pal - x 模型时，从零开始训练。其次，我们注意到，与简单地用机器人数据进行微调相比，对模型进行共同微调(无论其大小)会产生更好的泛化性能。我们将其归因于这样一个事实，即在训练的微调部分保留原始数据，使模型不会忘记在 VLM 训练期间学习到的先前概念。最后，不出所料，我们注意到模型大小的增加导致了更好的泛化性能。

4.4.RT-2 能表现出与视觉语言模型类似的思维链推理吗？

受法学硕士中的思维链提示方法(Wei et al.， 2022)的启发，我们对 PaLM-E 的 RT-2 变体进行了几百个梯度步骤的微调，以提高其同时利用语言和动作的能力，希望它能引发更复杂的推理行为。我们增强了数据，包括一个额外的“计划”步骤，该步骤首先用自然语言描述机器人即将采取的动作的目的，然后是实际的动作标记，例如“指令:我饿了。计划:挑选 rxbar 巧克力。行动:1 128 124 136 121 158 111 255。”这个数据增强方案充当了 VQA 数据集(视觉推理)和操作数据集(生成动作)之间的桥梁。

我们定性地观察到，具有思维链推理的 RT-2 能够回答更复杂的命令，因为它首先在自然语言中有一个地方来计划它的行动。这是一个很有前途的方向，它提供了一些初步证据，证明使用法学硕士或法学硕士作为规划者(Ahn et al.， 2022;Driess et al.， 2023)可以在单个 VLA 模型中与低级策略相结合。采用思维链推理的 RT-2 的部署如图 7 和附录 1 所示。

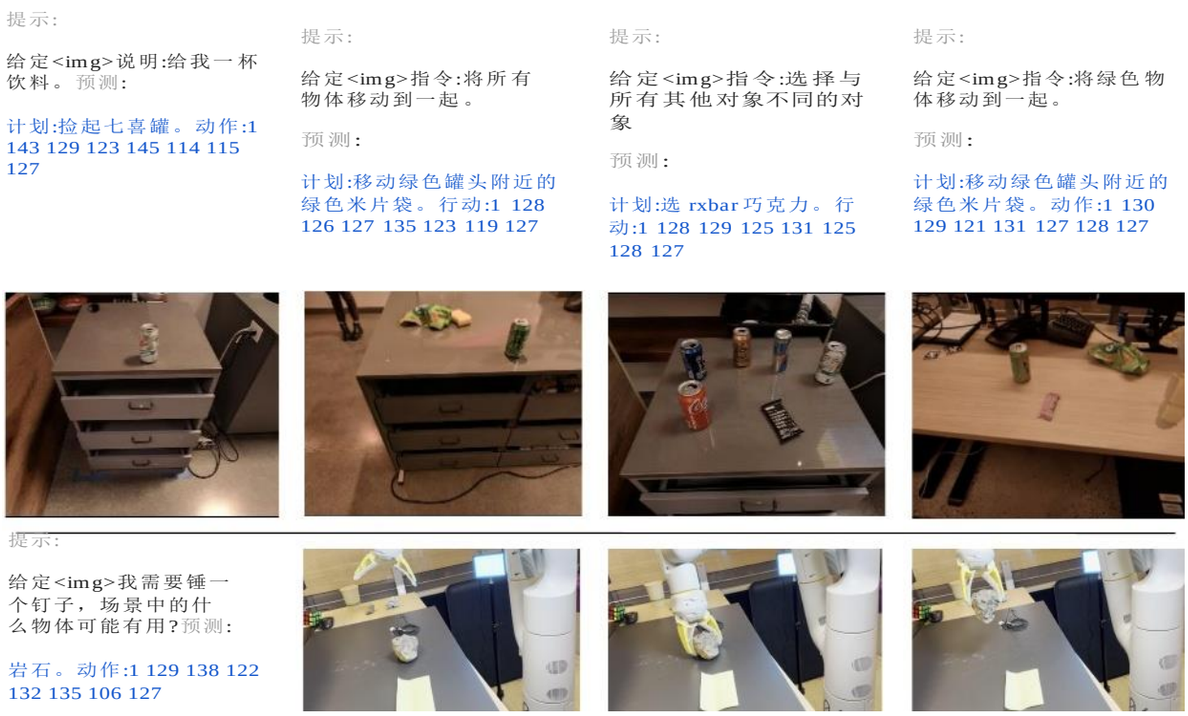


图 7 |采用思维链推理的 RT-2 布局，其中 RT-2 生成计划和行动。**5.限制**

尽管 RT-2 显示出很有希望的泛化特性，但这种方法有许多局限性。首先，尽管我们表明，通过 VLMs 进行网络规模的预训练可以促进语义和视觉概念的泛化，但机器人并没有通过包括这种额外的经验来获得执行新*动作*的任何能力。该模型的物理技能仍然局限于机器人数据中看到的技能分布(参见附录 G)，但它学会了以新的方式部署这些技能。我们认为，这是数据集没有沿着技能轴进行足够变化的结果。未来工作的一个令人兴奋的方向是研究如何通过新的数据收集范式(如人类视频)获得新技能。

其次，尽管我们表明我们可以实时运行大型 VLA 模型，但这些模型的计算成本很高，并且由于这些方法应用于需要高频控制的设置，实时推断可能成为主要瓶颈。未来研究的一个令人兴奋的方向是探索量化和蒸馏技术，这些技术可能使此类模型以更高的速率或在更低成本的硬件上运行。这也与另一个当前的限制有关，因为只有少数通常可用的 VLM 模型可用于创建 RT-2。我们希望有更多的开源模型可用(例如 https://llava-vl.github.io/)，而专有模型将开放它们的微调 api，这是构建 VLA 模型的充分要求。

6.结 论

在本文中，我们描述了如何将视觉语言模型(VLM)预训练与机器人数据相结合来训练视觉语言动作(VLA)模型。然后，我们提出了两个基于 PaLM-E 和 PaLI-X 的 vla 实例，我们称之为 RT-2-PaLM-E 和 RT-2-PaLI-X。这些模型与机器人轨迹数据进行共微调，以输出机器人动作，这些动作被表示为文本令牌。我们表明，我们的方法产生了非常高性能的机器人策略，更重要的是，导致了明显更好的泛化性能和从

网络规模的视觉语言预训练。我们相信，这种简单而通用的方法显示了机器人技术直接受益于更好的视觉语言模型的前景，这使得机器人学习领域处于一个战略位置，随着其他领域的进步，机器人学习领域将进一步提高。

致谢

我们要感谢 Fred Alcober、Jodi Lynn Andres、Carolina Parada、Joseph 达比斯、Rochelle Dela Cruz、Jessica Gomez、Gavin Gonzalez、John Guilyard、Tomas Jackson、Jie Tan、Scott Lehrer、Dee M、Utsav Malla、Sarah Nguyen、Jane Park、Emily Perez、Elio Prado、Jornell Quiambao、Clayton Tan、Jodexty Therlonge、Eleanor Tomlinson、Wenxuan Zhou 以及更大的 Google DeepMind 团队的反馈和贡献。

参考文献

M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog 等。做我能做的，而不是我说的:机器人启示中的基础语言。*arXiv 预印本 arXiv:2204.01691*，2022。

j - b。Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds 等。Flamingo:一种用于少镜头学习的视觉语言模型。*arXiv 预印本 arXiv:2204.14198*,2022。

R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen，等。palm2 技术报告。*arXiv 预印本 arXiv:2305.10403*，2023。

A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. 达比斯, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu，等。Rt-1:用于大规模现实世界控制的机器人变压器。*arXiv 预印本 arXiv:2212.06817*,2022。

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell 等。语言模型是几次学习器。*神经信息处理系统进展*，33:1877-1901,2020。

dr . Cer，杨勇，孔树清，华宁，N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S.袁，C. Tar, Y. Sung, B. Strope, R. Kurzweil。通用句子编码器。*CoRR*, abs/1803.11175, 2018。URL http://arxiv.org/abs/1803.11175。

M.陈，J. Tworek, H. Jun, Q.袁，H. P. d. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman，等。评估在代码上训练的大型语言模型。*arXiv 预印本 arXiv:2107.03374*,2021。

陈，J. Djolonga，P. Padlewski，B. Mustafa，S. Changpinyo，J. Wu, C. R. Ruiz，S. Goodman，王新，Y. Tay, S. Shakeri, M. Dehghani, D. Salz, M. Lucic, M. Tschannen, A. Nagrani, H. Hu, M. Joshi，B. Pang, C. Montgomery，P. Pietrzyk，M. Ritter，A. Piergiovanni，M. Minderer，F. Pavetic，A. Waters，G. Li, I. Alabdulmohsin，L. Beyer, J. Amelot，K. Lee, A. P. Steiner，Y. Li, D. Keysers，A. Arnab, Y. Xu，A. Kolesnikov，M. Seyedhosseini，A. Angelova，X. Zhai, N. Houlsby，R. Soricut。Pali-x:关于扩大多语言视觉和语言模型，2023a。

X.陈，X. Wang, S. Changpinyo, A. Piergiovanni, P. Padlewski, D. Salz, S. Goodman, A. Grycner, B. Mustafa, L. Beyer, A. Kolesnikov, J. Puigcerver, N. Ding, K. Rong, H. Akbari, G. Mishra, L. Xue, A. Thapliyal, J. Bradbury, W. Kuo, M. Seyedhosseini, C. Jia, B. K. Ayan, C. Riquelme, A. Steiner, A. Angelova, X. Zhai, N. Houlsby 和 R. Soricut。巴利文:一种联合尺度的多语言语言图像模型，2023b。

K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano 等。训练验证器来解决数学单词问题。*arXiv 预印本 arXiv:2110.14168*，2021。

崔彧，王，N. Muhammad, L. Pinto，等。从游戏到政策:从未经整理的机器人数据生成条件行为。*arXiv 预印本 arXiv: 2210.10047*,2022。

S. Dasari 和 A. Gupta。一次性视觉模仿的《变形金刚》。*《Conference on Robot Learning》*，2071-2084 页。PMLR，2021 年。

S. Dasari, F. Ebert, S. Tian, S. Nair, B. Bucher, K. Schmeckpeper, S. Singh, S. Levine 和 C. Finn。Robonet:大规模多机器人学习。*Conference on Robot Learning*, 2019。

M. Dehghani, J. Djolonga, B. Mustafa, P. Padlewski, J. Heek, J. Gilmer, A. Steiner, M. Caron, R. Geirhos, I. Alabdulmohsin, R. Jenatton, L. Beyer, M. Tschannen, A. Arnab, X.王， C. Riquelme, M. Minderer, J. Puigcerver, U. Evci, M. Kumar, S. van Steenkiste, G. F. Elsayed, A. Mahendran, F. 余, A. Oliver, F. Huot, J. Bastings, M. P. Collier, A. Gritsenko, V. Birodkar, C. Vasconcelos, Y. Tay, T. Kipf, M. luk, X. Zhai, D. Keysers, J. Harmsen，和 N.霍尔斯比。2023年，将视觉变压器缩放到220亿个参数。

D. Driess, F. Xia, M. S. sajjjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. thompson, Q. Vuong, T. 余，等。Palm-e:一种具身的多模态语言模型。*arXiv 预印本 arXiv:2303.03378*，2023。

M.杜，S.奈尔，D.萨迪，和 C.芬恩。行为检索:通过查询未标记数据集进行的少量模仿学习。*arXiv 预印本 arXiv:2304.08742*，2023a。

Y.杜，K. Konyushkova, M. Denil, A. Raju, J. Landon, F. Hill, N. de Freitas, S. Cabi。作为成功检测器的视觉语言模型。*arXiv 预印本 arXiv: 2303.07280*，2023b。

C.芬恩和 S.莱文。机器人运动规划的深度视觉预见。*2017 年 IEEE 机器人与自动化国际会议(ICRA)*，第2786-2793页。IEEE，2017。

C. Finn, T.余，T.张，P. Abbeel和 S. Levine。通过元学习的一次性视觉模仿学习。*《Conference on robot learning》*，357-368页。PMLR，2017年。

r.a.费雪。实验设计。*英国医学杂志*，1(3923):554,1936。

S. Y. Gadre, M. Wortsman, G. Ilharco, L. Schmidt, S. Song。车轮上的夹子:作为对象定位和探索的零射击对象导航。*arXiv 预印本 arXiv:2203.10421*，2022。

甘志刚，李丽，李丽，王，刘志刚，高军，等。视觉语言预训练:基础、最新进展和未来趋势。*计算机图形学与视觉的基础与趋势®*，14(3-4):163-352,2022。

贾思，顾晓明，崔彧，李志强。林。开放词汇图像分割。*arXiv 预印本 arXiv:2112.12143*，2021。

K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M.刘，X.刘，M. Martin, T. Nagarajan, I. Radosavovic, S. K. Ramakrishnan, F. Ryan, J. Sharma, M. Wray, M. Xu, E. Z. Xu, C. Batra, V. lier, S. Do, M. Doulaty, A. Erapalli, C. Feichtenhofer, A. Fragomeni, Q. Fu, A. gebrreselasie, C. Gonzalez, J. Hillis, X.黄，Y.黄，W. Jia, W. Khoo, F. Landini, C. Li, Y. Li, Z. Mangalam，R. Modhugu, J. Munro, T. Murrell, T. Nishiyasu, W. Price, P. R. Puentes, M. Ramazanova, L. Sari, K. Somasundaram, A. Southerland, Y. Sugano, R. Tao, M. Vo, Y. Wang, Y. Wu, T. Yagi，赵正昭，Y. Zhu, P. Arbelaez, D. Crandall, D. Damen, G. M. Farinella, C. V. Jawahar, H. Joo, K. Kitani, H. Ghanem, V. K. Ithapu, C. V. Jawahar, J. M. Rehg, Y. Sato, J. Newcombe, A. Oliva, J. M. Rehg, J. Shi, M. Shou, A. Torralba, L. Torresani, M. Yan和 J. Malik。《Ego4d:用 3000 小时的自我中心视频环游世界》，2022年。

顾旭东，吴廷玉。林，郭伟，崔。基于视觉和语言知识蒸馏的开放词汇对象检测。*arXiv 预印本 arXiv:2104.13921*，2021。

N. Hansen, R. Jangir, Y. Sun, G. aleny, P. Abbeel, A. A. Efros, L. Pinto，和 X.王伟。部署过程中自我监督的政策适应。*arXiv 预印本 arXiv: 2007.04309*，2020。

郝彦，宋辉，董林，黄，迟志智，王伟，马索，魏峰。语言模型是通用接口。*arXiv 预印本 arXiv:2206.06336*，2022。

F. Hill, S. Mokra, N. Wong 和 T. Harley。通过文本迁移学习的深度强化学习的人类指令遵循。*arXiv 预印本 arXiv:2005.09382*, 2020。

黄，董林，王伟，郝彦，S. Singhal, 马索，吕涛，崔，O. K. Mohammed，刘，等。语言不是你所需要的全部:将感知与语言模型对齐。*arXiv 预印本 arXiv:2302.14045*, 2023。

黄伟，P. Abbeel, D. Pathak 和 I. Mordatch。语言模型作为零射击计划者:为具身代理提取可操作的知识。*机器学习国际会议*，918 - 9147 页。PMLR，2022 年。

S. James, M. Bloesch 和 A. J. Davison。少量模仿学习的任务嵌入式控制网络。*《Conference on robot learning》*，783-795 页。PMLR, 2018 年。

E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine，和 C. Finn。Bc-z:机器人模仿学习的零射击任务泛化。*《Conference on Robot Learning》*，991-1002 页。PMLR, 2021 年。

蒋勇，A. Gupta，张志强，王国强，窦勇，陈勇，李飞飞，A. Anandkumar，朱勇，范磊。Vima:具有多模态提示符的通用机器人操作。*arXiv 预印本 arXiv:2210.03094*，2022。

L. P. Kaelbling。高效机器人学习的基础。*科学*，369(6506):915-916,2020。

S. Karamcheti, S. Nair, A. S. Chen, T. Kollar, C. Finn, D. Sadigh 和 P. Liang。机器人学的语言驱动表示学习。*arXiv 预印本 arXiv:2302.12766*, 2023。

罗，等。段任何东西。*arXiv 预印本 arXiv: 2304.02643*,2023。

I. Kostrikov, D. Yarats 和 R. Fergus。图像增强就是你所需要的:从像素中正则化深度强化学习。*arXiv 预印本 arXiv:2004.13649*, 2020。

M. Laskin, K. Lee, A. stoke, L. Pinto, P. Abbeel 和 A. Srinivas。增强数据的强化学习。*神经信息处理系统进展*，33:19884-19895,2020a。M. Laskin, A. Srinivas 和 P. Abbeel。Curl:用于强化学习的对比无监督表示。In *International Conference on Machine Learning*, 5639-5650 页。PMLR 2020 b。

S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz 和 D. Quillen。通过深度学习和大规模数据收集学习机器人抓取的手眼协调。*国际机器人研究杂志*，37(4-5):421-436,2018。

A. Lewkowycz, A. Andreassen, D. Dohan, E. Dyer, H. Michalewski, V. Ramasesh, A. Slone, C. Anil, I. Schlag, T. Gutman-Solo 等。用语言模型解决定量推理问题。*arXiv 预印本 arXiv:2206.14858*, 2022。

李家杰，李德华，李国强，李国强。Blip-2:使用冻结图像编码器和大型语言模型的 Bootstrapping 语言-图像预训练。*arXiv 预印本 arXiv:2301.12597*, 2023。

李丽华，叶思明，尹德华，陈志军。谢家华、k.w。张。Visualbert:一个简单而高效的视觉和语言基线。*arXiv 预印本 arXiv:1908.03557*，2019。

刘汉，李莉，李凯，和 P. Abbeel。联合预训练视觉语言模型的指令跟随代理。*arXiv 预印本 arXiv:2210.13431*, 2022。

J. Lu, D. Batra, D. Parikh 和 S. Lee。Vilbert:针对视觉和语言任务的任务不可知论视觉语言表征的预训练。*神经信息处理系统进展*，32,2019。

C. Lynch 和 P. Sermanet。非结构化数据上的语言条件模仿学习。*arXiv 预印本 arXiv:2005.07648*，2020。

C. Lynch, A. Wahid, J. thompson, T. Ding, J. Betker, R. Baruch, T. Armstrong 和 P. Florence。互动语言:与机器人实时对话。*arXiv 预印本 arXiv:2210.06407*，2022。

马英杰，S. Sodhani, D. Jayaraman, O. Bastani, V. Kumar, A. Zhang。Vip:通过价值内隐预训练走向通用视觉奖励和表征。*arXiv 预印本 arXiv: 2210.00030*,2022。马英杰，梁伟，V. Som, V. Kumar, A. Zhang, O. Bastani 和 D. Jayaraman。Liv:机器人控制的语言-图像表征和奖励。*arXiv 预印本 arXiv:2306.00958*，2023。

J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, K. Goldberg。Dex-net 2.0:利用合成点云和分析抓取度量来规划稳健抓取的深度学习。*arXiv 预印本 arXiv:1703.09312*，2017。

A. Majumdar, K. Yadav, S. Arnaud, Y. J. Ma, C. Chen, S. Silwal, A. Jain, v . p。Berges, P. Abbeel, J. Malik，等。在为具身智能寻找人工视觉皮层的过程中，我们到了什么地步?*arXiv 预印本 arXiv:2303.18240*，2023a。

A. Majumdar, K. Yadav, S. Arnaud, Y. J. Ma, C. Chen, S. Silwal, A. Jain, v . p。Berges, P. Abbeel, J. Malik，等。在为具身智能寻找人工视觉皮层的过程中，我们到了什么地步?*arXiv 预印本 arXiv:2303.18240*，2023b。

O. Mees, L. Hermann 和 W. Burgard。语言条件下机器人模仿学习对非结构化数据的影响。*IEEE 机器人与自动化学报*，7(4):11205 - 11212,2022。M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen 等。使用视觉变形器的简单开放词汇对象检测。*arXiv 预印本 arXiv:2205.06230*，2022。

穆艳，张琪，胡敏，王伟，丁敏，金俊，王宝宝，戴杰，乔艳，罗鹏。Embodiedgpt:通过具身思维链的视觉语言预训练。*arXiv 预印本 arXiv: 2305.15021*,2023。

S.奈尔，E. Mitchell, K. Chen, S. Savarese, C.芬恩，等。从离线数据和众包注释中学习语言条件机器人行为。*机器人学习会议*，1303-1315 页。PMLR，2022。

S.奈尔，A.拉杰斯瓦兰，V.库马尔，C.芬恩和 A.古普塔。R3m:机器人操作的通用视觉表示。*arXiv 预印本 arXiv:2203.12601*，2022b。

OpenAI。Gpt-4 技术报告，2023 年。

J. Pari, N. M. Shafiullah, S. P. am 和 L.平托。表征学习在视觉模仿中的惊人效果。*arXiv 预印本 arXiv:2112.01511*，2021。

L.平托和 A.古普塔。超大规模的自我监督:从 5 万次尝试和 700 个机器人小时中学习掌握。*2016 年 IEEE 机器人与自动化国际会议(ICRA)*，3406-3413 页。IEEE，2016。

S. Polu, J. M. Han, K. Zheng, M. Baksys, I. Babuschkin, I. Sutskever。形式数学陈述课程学习。*arXiv 预印本 arXiv:2202.01344*，2022。

V. H. Pong, M. Dalal, S. Lin, A.奈尔，S. Bahl，和 S.莱文。Skew-fit:覆盖状态的自监督强化学习。*arXiv 预印 arXiv:1903.03698*，2019。

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark 等。从自然语言监督中学习可转移的视觉模型。*国际机器学习会议*，8748-8763 页。PMLR, 2021 年。

S. Reed, K. Zolna, E. Parisotto, S. G. Colmenarejo, A. Novikov, G. Barth-Maron, M. Gimenez, Y. Sulsky, J. Kay, J. T. springberg，等。通才型代理人。*arXiv 预印本 arXiv:2205.06175*，2022。

M. Ryoo, A. Piergiovanni, A. Arnab, M. Dehghani, A. Angelova。Tokenlearner:视频自适应时空标记化(Tokenlearner)。*神经信息处理系统进展*，34(4):12786 - 12797,2021。D.沙阿，B. Osiński, B.希特，和 S.莱文。Lm-nav:具有语言、视觉和动作的大型预训练模型的机器人导航。K. Liu, D. Kulic 和 J. Ichnowski，编辑，*第六届机器人学习会议论文集，《机器学习*研究论文集》第 205 卷，第 492 - 504 页。PMLR, 2023 年 12 月 14-18 日。URL https://proceedings.mlr.press/v205/shah23b.html。R.沙阿和 V.库马尔。Rrl: Resnet 作为强化学习的表示。*arXiv 预印本 arXiv:2107.03380*，2021。

M. Shridhar, L. Manuelli 和 D. Fox。Cliport:机器人操纵的途径是什么和在哪里。*第五届机器人学习会议(CoRL)论文集*，2021。

M. Shridhar, L. Manuelli 和 D. Fox。Cliport:机器人操纵的途径是什么和在哪里。*机器人学习会议*，894-906 页。PMLR，2022。

M. Shridhar, L. Manuelli 和 D. Fox。感知者-行动者:用于机器人操作的多任务转换器。*arXiv 预印本 arXiv:2209.05451*, 2022b。

I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, A. Garg。Progprompt:使用大型语言模型生成情境机器人任务计划。在 *ICRA*，2023。M. H. Smith 和 L. S. Coles。低成本、通用机器人的设计。见 *IJCAI*，324-336 页，1973 年。

A. Stone, T. Xiao, Y. Lu, K. Gopalakrishnan, K.- h。李，王琪，P. Wohlhart, B. Zitkovich，夏峰，C. Finn，等。使用预训练的视觉语言模型的开放世界对象操纵。*arXiv 预印本 arXiv: 2303.00905*,2023。

T. Sumers, K. Marino, A. Ahuja, R. Fergus, I. Dasgupta。将互联网规模的视觉语言模型提炼成具身代理(embodied agents)。*arXiv 预印本 arXiv: 2301.12507*,2023。

Y. Tay, M. Dehghani, V. Q. Tran, X. Garcia, J. Wei, X. Wang, H. W. Chung, S. Shakeri, D. Bahri, T. Schuster, H. S. Zheng, D. Zhou, N. Houlsby 和 D. Metzler。Ul2:统一语言学习范式，2023。

S. Vemprala, R. Bonatti, A. Bucker, and A. Kapoor。机器人的 Chatgpt:设计原则和模型能力。*微软 Auton。系统。机器人。Res*，2023 年 2:20。

王建军，杨振华，胡晓明，李丽丽，林凯，甘志刚，刘振华，刘春春，王丽丽。Git:一种用于视觉和语言的生成式图像到文本转换器。*arXiv 预印本 arXiv:2205.14100*，2022。

魏建军，王晓霞，D. Schuurmans, M. Bosma，池 E.， Le Q.，周 D.。思维链提示在大型语言模型中引出推理。*arXiv 预印本 arXiv: 2201.1193*,2022。

魏建军，侯丽玲，A. Lampinen，陈晓，黄迪，Tay，陈晓，陆勇，周道德，马涛，乐庆伟。符号调优改进语言模型中的语境学习，2023。

J.吴，R. Antonova, A. Kan, M. Lepert, A. Zeng, S. Song, J. Bohg, S. Rusinkiewicz, T. Funkhouser。Tidybot: 使用大型语言模型的个性化机器人辅助。*arXiv 预印本 arXiv:2305.05658*，2023。

T.肖，H.陈，P. Sermanet, A.瓦希德，A.布罗汉，K.豪斯曼，S.莱文和 J.汤普森。基于视觉语言模型的指令增强机器人技能习得。*arXiv 预印本 arXiv: 2211.11736*,2022a。

T.肖，I. Radosavovic, T.达雷尔，J.马利克。运动控制的蒙面视觉预训练。*arXiv 预印本 arXiv:2203.06173*,2022b。

S. Young, D. Gandhi, S. Tulsiani, A. Gupta, P. Abbeel, L. Pinto。视觉模仿变得容易。*《Conference on Robot Learning》*，1992-2005 页。PMLR，2021 年。

K.-T。余 t，M. Bauza, N. Fazeli 和 A. Rodriguez。超过一百万种被推的方式。平面推的高保真实验数据集。*2016 年 IEEE/RSJ 智能机器人与系统国际会议(IROS)*， 30-37 页。IEEE，2016。

余 t，芬 C.，谢 a .， S. Dasari，张 t .， P. Abbeel, S. Levine。通过领域自适应元学习观察人类的一次性模仿。*arXiv 预印本 arXiv:1802.01557*，2018。

X. Zhai, A. Kolesnikov, N. Houlsby, L. Beyer。缩放视觉变压器。*IEEE/CVF 计算机视觉和模式识别会议论文集*，12104-12113 页，2022。X. Zhang, Y. Ding, S. Amiri, H. Yang, A. Kaminski, C. Esselink, S. Zhang。基于视觉语言模型的经典任务规划器。*arXiv 预印本 arXiv:2304.08587*，2023。

答:贡献

- **训练和评估(设计和执行训练模型的程序，在模拟和现实世界中评估模型，为算法设计选择运行烧烧)**: 叶夫根·切波塔、克日什托夫·乔罗曼斯基、丁天利、丹尼·德里斯、阿维纳瓦·杜贝、皮特·弗洛伦斯、傅楚远、蒙特塞·冈萨雷斯·阿雷纳斯、基尔塔纳·戈帕拉克里希南、韩克航、亚历山大·赫尔佐格、布莱恩·伊切特、亚历克斯·伊尔潘、伊莎贝尔·莱阿尔、丽莎·李、路遥、亨利克·米哈莱夫斯基、伊戈尔·莫达奇、卡尔·珀尔奇、迈克尔·柳格、阿尼凯特·辛格、王全、艾扎安·瓦希德、保罗·沃尔哈特、夏飞、肖泰特和于天河。
- **Network Architecture (designing and implementing model network modules, working on tokenization of actions, enabling inference of the model networks during experiments)**: Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Danny Driess, Pete Florence, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Brian Ichter, Alex Irpan, Isabel Leal, Lisa Lee, Henryk Michalewski, Igor Mordatch, Kanishka Rao, Michael Ryoo, Anikait Singh, Quan Vuong, Ayzaan Wahid, Jialin Wu, Fei Xia, Ted Xiao, and Tianhe Yu.
- **数据收集(收集真实机器人的数据，运行真实机器人评估，执行运行真实机器人所需的操作)**:Noah Brown, Justice Carbajal, Tianli Ding, Krista Reymann, Grecia Salazar, Pierre Sermanet, Jaspiar Singh, Huong Tran, Stefan Welker 和 Sichun 徐。
- **领导(领导项目工作，管理项目人员，为项目方向提供建议)**:Yevgen Chebotar, Chelsea Finn, 卡罗尔·豪斯曼, Brian Ichter, Sergey Levine, Yao Lu, Igor Mordatch, Kanishka Rao, Pannag Sanketi, Radu Soricut, Vincent Vanhoucke, Tianhe Yu。
- **论文(处理纸质手稿，设计纸质可视化和图形)**:Yevgen Chebotar、丹尼·德里斯、Chelsea Finn、皮特·弗洛伦斯、卡罗尔·豪斯曼、Brian Ichter、Lisa Lee、Sergey Levine、伊戈尔·莫达奇、卡尔·珀奇、王泉、夏飞、特德·肖、于天和。
- **基础设施(致力于训练模型、运行实验、存储和访问数据所需的基础设施和代码库主干)**:Anthony Brohan、Yevgen Chebo- tar、丹尼·德里斯、Kehang Han、Jasmine Hsu、Brian Ichter、Alex Irpan、Nikhil Joshi、Ryan Julian、Dmitry Kalashnikov、kuyuheng、Isabel Leal、Lisa Lee、Tsang-Wei Edward Lee、Yao Lu、Igor Mordatch、Quan Vuong、Ayzaan Wahid、Fei Xia、Ted Xiao、Peng Xu 和 Tianhe Yu。

b .数据集

视觉语言数据集基于 Chen 等人(2023b)和 Driess 等人(2023)的数据集混合。该数据的大部分由 webi 数据集组成，该数据集包含 109 种语言的大约 10B 个图像-文本对，过滤出得分最高的 10%的跨模态相似性示例，以提供 1B 个训练示例。许多其他字幕和视觉问答数据集也包括在内，关于数据集混合的更多信息可以在 Chen 等人(2023b)的 rt -2- pal - x 和 Driess 等人(2023)的 RT-2-PaLM-E 中找到。在对 rt -2- pal - x 进行共微调时，我们没有使用 Chen 等人(2023a)描述的 Episodic webi 数据集。

机器人数据集基于 Brohan 等人(2022)的数据集。这包括用移动操作机器人收集的演示集。每个演示都有一个自然的语言说明，说明了七种技能中的一种:“从容器中取出物体”、“移动物体附近的物体”、“将物体直立”、“撞倒物体”、“打开抽屉”、“关闭抽屉”、“将物体放入容器”和“从容器中取出物体并放在柜台上”。进一步的细节可以在 Brohan 等人(2022)中找到。

rt -2- pal - x 对机器人数据集进行加权，使其占训练混合数据的 50%左右

co-fine-tuning。RT-2-PaLM-E 对机器人数据集的权重约为训练混合物的 66%。

对于表 1中 Language-Table 的结果，我们的模型是在 Lynch 等人(2022)的 Language-Table 数据集上训练的。我们的模型在几个预测任务上进行了共同微调:(1)预测动作，给定两个连续的图像帧和一个文本指令;(2)预测指令，给定图像帧;(3)预测机器人手臂位置，给定图像帧;(4)预测给定图像帧之间的时间步数;(5)在给定图像帧和指令的情况下预测任务是否成功。

c .基线

我们将我们的方法与多个最先进的基线进行比较，这些基线挑战了我们方法的不同方面。所有的基线都使用完全相同的机器人数据。

- **RT-1**:机器人变压器 1 Brohan 等人(2022)是一种基于变压器的模型，在发布时在类似的任务套件上实现了最先进的性能。该模型没有使用基于 vlm 的预训练，因此它提供了一个重要的数据点来证明基于 vlm 的预训练是否重要。
- **VC-1**: VC-1 Majumdar 等人(2023a)是一种视觉基础模型，使用专门为机器人任务设计的预训练视觉表示。我们使用来自 VC-1 ViT-L 模型的预训练表示。由于 VC-1 不包括语言条件反射，我们通过通用句子编码器 Cer 等人(2018)单独嵌入语言命令来添加语言条件反射，以便与我们的方法进行比较。特别是，我们将生成的语言嵌入令牌连接到 VC-1 生成的图像令牌，并通过令牌学习器 Ryoo 等人(2021)传递连接的令牌序列。令牌学习器产生的令牌序列然后由 RT-1 解码器转换器模型消耗，以预测机器人动作令牌。我们端到端训练 VC-1 基线，并在训练期间解冻 VC-1 权重，因为这比使用冻结的 VC-1 权重产生更好的结果。

R3M: R3M Nair 等人(2022b)是一种与 VC-1 类似的方法，R3M 使用预训练的视觉语言表示来改进策略训练。在这种情况下，作者使用人类活动的 Ego4D 数据集 Grauman et al.(2022)来学习策略使用的表示。VC-1 和 R3M 都测试了不同的最先进的表示学习方法，作为使用 VLM 的替代方案。为了从 R3M 预训练的表示中获得语言条件策略，我们遵循与上面描述的 VC-1 相同的过程，除了我们使用 R3M ResNet50 模型来获得图像令牌，并在训练期间解冻它。

- **MOO**: MOO Stone 等人(2023)是一种以对象为中心的方法，其中首先使用 VLM 以原始图像中的单个彩色像素的形式指定感兴趣的对象。然后用端到端策略训练这个经过像素修改的图像，以完成一组操作任务。该基线对应于这样一种情况，即 VLM 被用作增强感知的单独模块，但其表示不用于策略学习。

D.用于 RT-2 的 vlm

pal - x 模型架构由一个 ViT-22B Dehghani 等人(2023)来处理图像，它可以接受图像序列 ，导致每张图像有 x 个令牌，其中 是每张图像的补丁数量。通过投影层的图像令牌然后由 32B 参数和 50 层的编码器-解码器主干消耗，类似于 UL2 Tay 等人(2023)，它联合处理文本和图像作为嵌入，以自动回归的方式生成输出令牌。文本

输入通常由任务类型和任何额外的上下文组成(例如，对于标题任务，“⟨ lang⟩ 生成标题”或对于 VQA 任务，“⟨ lang⟩ 的答案:问题”)。

在语言表(Language-Table)上训练的 pal - 3b 模型(表 1)使用较小的 viti - g /14 (Zhai et al.， 2022) (2B 参数)来处理图像，UL2-3B (Tay 等， 2023)用于编解码器网络。

PaLM-E 模型基于仅解码器的 LLM，它将机器人数据(如图像和文本)投射到语言标记空间中，并输出文本(如高级计划)。在使用 PaLM-E-12B 的情况下，用于将图像投影到语言嵌入空间的视觉模型是 viti - 4b, Chen 等(2023b)。连续变量与文本输入的连接允许 PaLM-E 完全是多模态的，接受各种各样的输入，如多传感器模式、以对象为中心的表示、场景表示和对象实体引用。

E.训练细节

我们对来自 pal - x (Chen 等人， 2023a) 5B 和 55B 模型、PaLI (Chen 等人， 2023b) 3B 模型和 PaLM-E (Driess 等人， 2023)12B 模型的预训练模型进行了共微调。对于 RT-2-PaLI-X-55B，我们使用学习率 1e-3 和批量大小 2048，并对模型进行 80K 梯度步长共同微调，而对于 RT-2-PaLI-X-5B，我们使用相同的学习率和批量大小，并对模型进行 270K 梯度步长共同微调。对于 RT-2-PaLM-E-12B，我们使用学习率 4e-4 和批大小 512 对模型进行 1M 梯度步长的共微调。两个模型都使用下一个令牌预测目标进行训练，这对应于机器人学习中的行为克隆损失。对于表 1 中用于语言表结果的 rt -2- pal - 3b 模型，我们使用学习率 1e-3 和批大小 128 来对模型进行 300K 梯度步长的共微调。

F.评估细节

F.1。评估场景

为了定量研究 RT-2 的应急能力，我们研究了各种具有挑战性的语义评估场景，旨在衡量推理、符号理解和人类识别等能力。图 8 提供了这些场景的一个子集的视觉概述，表 3 显示了用于定量评估的完整指令列表。

F.2。评估说明

表 2 列出了在模型评估中用于看不见的物体、背景和环境的自然语言指令。每条指令运行 1-5 次，具体取决于该评估集中总指令的数量。表 3 列出了用于评估定量紧急评估的自然语言指令。每条指令运行 5 次。

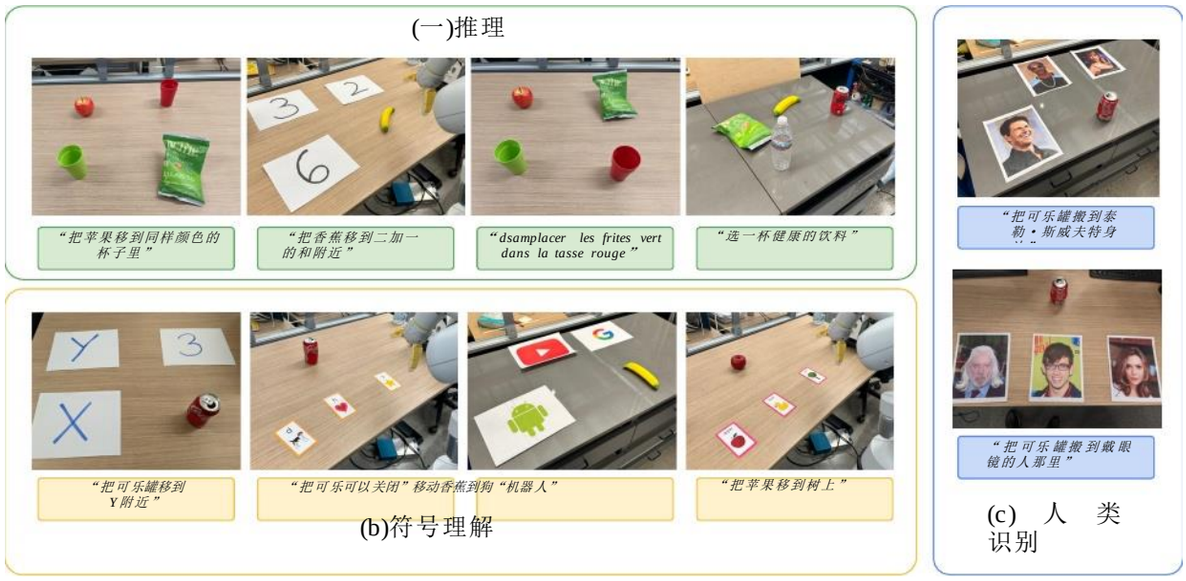


图 8 |用于研究 RT-2 应急能力的一些评估场景概述。它们侧重于三大类，分别是(a)推理、(b)符号理解和(c)人类识别。可视化指令是完整指令的一个子集，完整指令列在附录 F.2 中。

任务组	任务
符号理解:符号 1	移动 X 附近的焦炭罐，移动 3 附近的焦炭罐，移动 Y 附近的焦炭罐
符号理解:符号 2	将苹果移到树上，将苹果移到鸭子上，将苹果移到苹果上，将苹果移到匹配牌上
符号理解:符号 3	把可乐罐靠近狗，把可乐罐推到心脏上方，把可乐罐放在星星上方
推理:数学	把香蕉移到 2，把香蕉移到 2 加 1 的和附近，把香蕉移到 3 乘以 2 的答案附近，把香蕉移到最小的数字附近
推理:标志	把杯子搬到谷歌，把杯子搬到安卓，把杯子搬到 youtube，把杯子搬到搜索引擎，把杯子搬到手机
推理:营养	给我买健康零食，挑健康饮料，挑甜饮料，把健康零食移到健康饮料，挑咸零食
理由:颜色和多语言	将苹果移到相同颜色的杯子，将苹果移到不同颜色的杯子，将绿色薯片移到匹配颜色的杯子，将苹果移到 vaso verde, Bewegen Sie den Apfel in die rote Tasse，将绿色薯片移到 vaso rojo, mueve la manzana al vaso verde, dsamplacer les frites vert dans la Tasse rouge
人物辨识度:名人	把可乐罐递给泰勒·斯威夫特，把可乐罐递给汤姆·克鲁斯，把可乐罐递给史努比狗
人 识别: Celeb A	把可乐罐递给戴眼镜的人，把可乐罐递给白发男子，把可乐罐递给黑发女子

表 3 |用于定量紧急评估的自然语言指令。

G.故障案例示例

在图 9 中，我们提供了语言表设置中一个显著类型的失败案例的示例，其中 RT-2 模型没有推广到看*不见的对象动态*。在这些情况下，尽管模型能够正确地关注语言指令并移动到第一个正确的对象，但它无法控制这些对象的挑战性动态，这些对象与 Lynch 等人(2022)在此环境中看到的一小组块对象明显不同。然后，钢笔就会从桌子上滚下来(图 9，左)，而香蕉的质心离机器人接触的地方很远(图 9，右)。我们注意到，推动是出了名的难以预测和控制 Yu et al.(2016)。我们假设，通过在不同的环境和对象中进一步缩放数据集，可以在机器人环境交互动力学中实现更大的泛化——例如，在这种情况下，包含类似类型的更多样化的推动动力学的数据集 Dasari 等人(2019)。

此外，尽管 RT-2 在定性和定量紧急评估中在现实世界的操作任务中表现良好，但我们仍然发现了许多值得注意的失败案例。例如,

在目前的训练数据集组成和训练方法下，RT-2 似乎在以下方面表现不佳
:

- 通过特定部位抓取物体，比如手柄
- 超越机器人数据中看到的新颖动作，例如用毛巾擦拭或使用工具

灵巧的或精确的动作，如叠毛巾

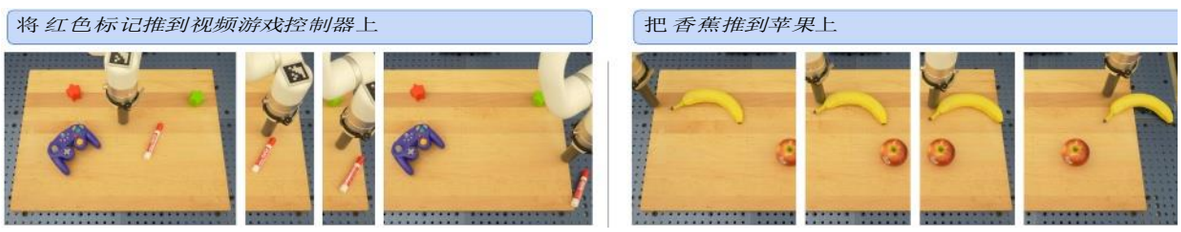


图 9 |现实世界中无法推广到*看不见的物体动态*的定性例子失败案例。

H.定量实验结果

1。总体性能，参见第 4.1 节

表 4 列出了我们的定量总体评价结果。我们发现，RT-2 在可见任务上的表现与基线一样好，甚至更好，并且在对未见物体、背景和环境的泛化方面明显优于基线。

模型	见过任务	看不见的对象		看不见的背景		看不见的环境		看不见的平均
		容易	硬	容易	硬	容易	硬	
R3M (Nair et al.， 2022b)	45	32	14	13	9	0	2	12
VC-1 (Majumdar et al.， 2023a)	63	34	10	13	3.	0	0	10
RT-1 (Brohan et al.， 2022)	**92**	31	43	71	9	26	14	32
MOO (Stone et al.， 2023)	75	58	48	38	41	19	3.	35
RT-2-PaLI-X-55B(我们的)	**91**	**70**	**62**	**96**	**48**	**63**	**35**	**62**
RT-2-PaLM-E-12B1(我们的)	**93**	**84**	**76**	**75**	**71**	**36**	**33**	**62**

表 4 | RT-2 的两个实例的总体性能和基线跨越可见的训练任务，以及衡量对新对象、新背景和新环境的泛化的未见评估。

H.2。紧急评估，参见第 4.2 节

表 5 列出了我们所有的定量应急评估结果。我们发现 RT-2 在这些新指令上的表现比 RT-1 好 2 到 3 倍，而无需任何额外的机器人演示。这展示了我们的方法如何允许我们利用网络规模视觉语言数据集上的预训练功能。

模型	符号的理解				推理					人识别			平均
	标志 1	象征 2	象征 3	平均	数学	标志	营养颜色/多语言		平均	名人 CelebA		平均	
VC-1 (Majumdar et al., 2023a)	7	25	0	11	0	8	20.	13	10	20.	7	13	11
RT-1 (Brohan et al., 2022)	27	20.	0	16	5	0	32	28	16	20.	20.	20.	17
RT-2-PaLI-X-55B(我们的)	**93**	**60**	**93**	**82**	25	52	**48**	**58**	**46**	**53**	**53**	**53**	**60**
RT-2-PaLM-E-12B(我们的)	67	20.	20.	36	**35**	**56**	44	35	43	33	**53**	43	40

表 5 | RT-2 的表现和定量应急评估的基线。

H.3。尺寸和训练消融，见第 4.3 节

表 6 详细说明了不同模型尺寸和训练方法的消融的定量结果。在每一种方法中，我们都看到模型大小在性能中起着重要作用，并且共同微调优于微调，后者优于从头开始训练。

模型	大小	培训	看不见的对象		看不见的背景		看不见的环境		平均
			容易	硬	容易	硬	容易	硬	
RT-2-PaLI-X	5 b	从头开始	0	10	46	0	0	0	9
RT-2-PaLI-X	5 b	微调	24	38	79	50	36	23	42
RT-2-PaLI-X	5 b	co-fine-tuning	60	38	67	29	44	24	44
RT-2-PaLI-X	55 个 b	微调	60	62	75	38	57	19	52
RT-2-PaLI-X	55 个 b	co-fine-tuning	70	62	96	48	63	35	63

表 6 |参数计数和训练策略对 RT-2 泛化的影响

1 .附加的思维链推理结果

我们提供了使用 RT-2-PaLM- E 完成的思维链推理部署的其他示例，如图 10 中第 4.4 节所述。

1 PaLM-E-12B 中使用的原始预训练数据混合(如 Driess 等人(2023)所述)包括用于高级 VQA 规划任务的机器人图像，这些图像可能类似于泛化场景中遇到的图像。然而，这些训练样例都不包括本实验中评估的低级动作。

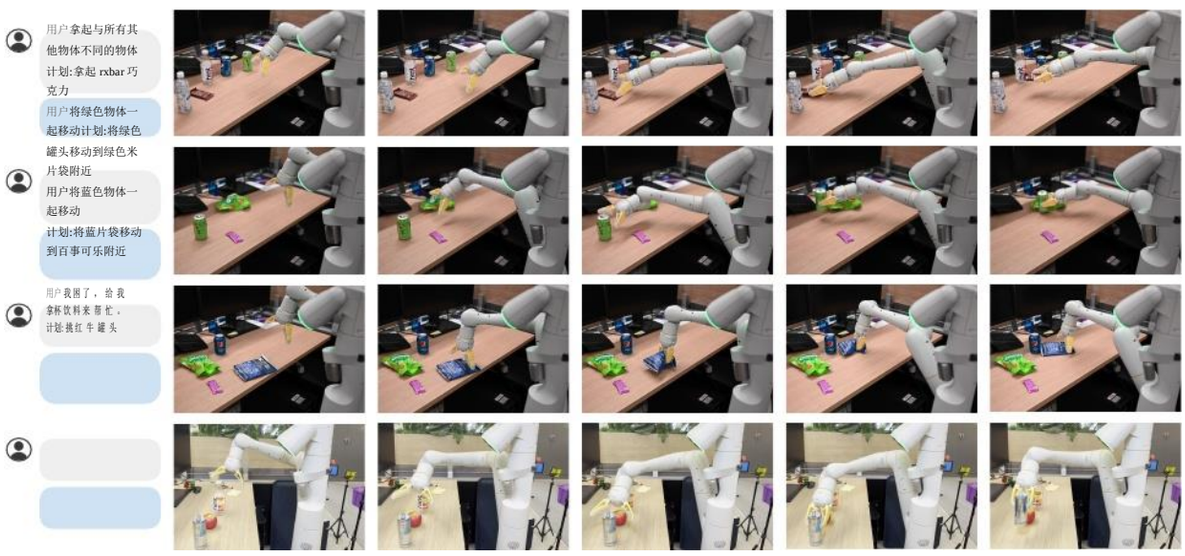


图 10 |采用思维链推理的 RT-2 的附加示例

任务组	任务
看不见的物体 (简单)	摘香蕉，把香蕉移到可乐罐附近，把橘子罐移到香蕉附近，摘奥利奥，把奥利奥移到苹果附近，把红牛罐移到奥利奥附近，摘梨，把椰子水移到椰子水附近，把百事可乐罐移到梨附近
看不见的物体 (Hard)	挑冷萃罐，挑大橙盘，挑嚼玩具，挑大十—— Nis 球，挑鸟饰品，挑鱼玩具，挑姜柠檬康普茶， 挑蛋分离器，挑腕表，挑青翠雪碧罐，挑青翠 挑超细纤维布，挑黄梨，挑椒盐脆饼薯片袋，挑消毒液 湿巾，挑菠萝提示水，挑绿杯，挑泡菜零食，挑 小蓝盘，挑小橘子擀面杖，挑章鱼玩具，挑 猫薄荷玩 具
看不见的背景 (Easy)	挑绿色墨西哥胡椒薯片袋，挑橙子罐，挑百事可乐罐，挑 7up 罐，挑苹果，挑蓝色薯片袋，挑橙子，挑 7up 罐，移橙子靠近水池，挑可乐罐，挑海绵，挑 rxbar 蓝莓
看不见的背景 (Hard)	挑腕表，挑蛋分离器，挑绿色雪碧罐，挑蓝色 挑超细纤维布，挑黄梨，挑椒盐脆饼薯片袋，挑消毒液 湿巾，挑菠萝提示水，挑青菜杯，挑泡菜点心，挑青菜 小蓝盘，挑小橘子擀面杖，挑章鱼玩具，挑 猫薄荷玩具，挑瑞典鱼袋，挑大绿色擀杖，挑黑色 太阳镜
看不见的环境 (Easy)	摘可乐罐，摘苹果，摘 rxbar 蓝莓，把苹果移到可乐罐附近， 把 rxbar 蓝莓移到苹果旁边，把可乐罐移到 rxbar 蓝莓旁边， 挑蓝色塑料瓶，挑海绵，挑蓝筹袋，移海绵 靠近蓝色塑料瓶，移动蓝筹袋靠近海绵，移动蓝色 塑料瓶靠近蓝筹袋，移动可乐罐靠近白色马克杯，移动 海绵靠近白色马克杯，移动可乐罐靠近黄色碗，移动海绵 靠近黄碗，移动可乐罐靠近绿布，靠近海绵 绿布，移动焦炭罐靠近盘子，移动海绵靠近盘子，移动 可乐罐靠近勺子，移动海绵靠近勺子，移动可乐罐靠近 橙色杯子，移动海绵靠近橙色杯子，挑白色杯子，挑黄色杯子 碗，挑绿色的布，把白色的马克杯移到海绵附近，移到黄色的碗 靠近海绵，移动靠近海绵的绿布，挑盘子，挑勺子，挑 橙色杯子，移盘子靠近海绵，移勺子靠近海绵，移动 橙色杯子靠近海绵，将可乐罐放入水槽，将可乐罐放入水槽， 将可乐罐推入水槽，将海绵放入水槽，将海绵放入水槽， 将海绵推入水槽，将绿布放入水槽，将绿布放入 水槽，把绿布推入水槽
看不见的环境 (Hard)	摘可乐罐，摘苹果，摘 rxbar 蓝莓，把苹果移到可乐罐附近，

如图 3 所示。