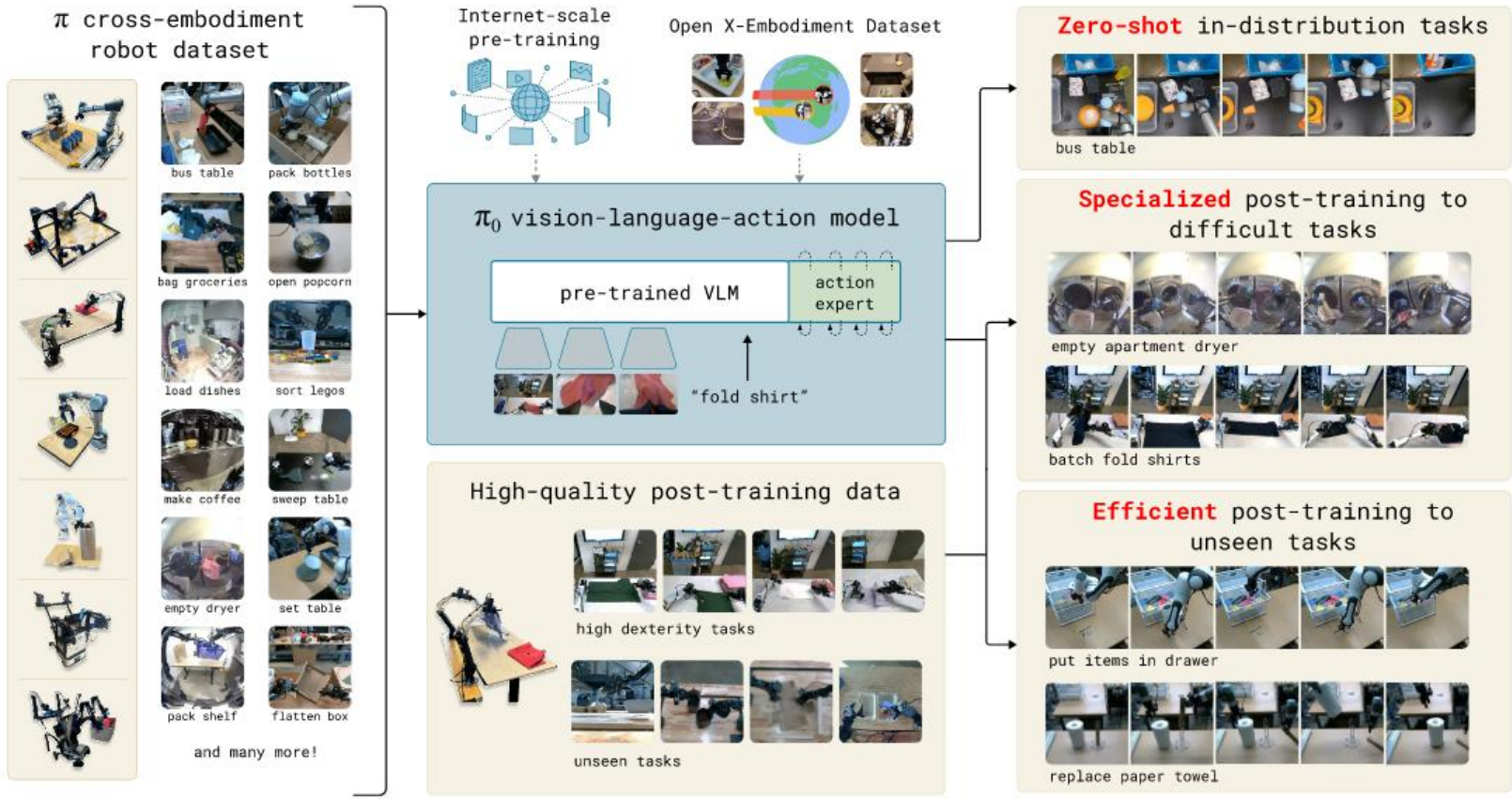
π0: A Vision-Language-Action Flow Model for General Robot Control

介绍人：李佩泽

2025.04.14

π_0: A Vision-Language-Action Flow Model for General Robot Control

Physical Intelligence
Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, Ury Zhilinsky
https://physicalintelligence.company/blog/pi0



引言

- 1. 灵活的通用机器人模型面临的三重挑战：
- 如何构建数据集、如何实现模型的泛用性、如何保证模型鲁棒性

- 2. 现有模型方法的局限性：
- **专用模型**：依赖于特定任务数据集，无法迁移至新的任务场景，不能处理多模态信息；
- **传统预训练策略**：在单一机器人平台、小规模数据集上难以获得通用物理规律，针对互联网语料训练的VLM缺乏对于具生信息的理解；
- **动作生成瓶颈**：现有的扩散模型难以对复杂任务场景产生深刻理解、生成长序动作。

引言

3. 研究动机与创新点

为解决上述问题，作者提出**π_0模型**，其核心创新包括：

- **大规模跨平台预训练**：整合来自7种机器人平台的10,000小时数据，覆盖68项任务，增强模型对多样物理交互的泛化能力；
- **新提出的视觉-语言-动作架构**：基于预训练的VLM注入语义理解能力，结合流匹配（Flow Matching）生成高精度连续动作，支持50Hz实时控制；
- **分层训练策略**：分“预训练-微调”两阶段，预训练学习通用技能，微调针对复杂任务（如叠衣服）优化，平衡效率与鲁棒性。

Pi0：整体架构

- 整体架构：由两组参数组成的单个Transformer

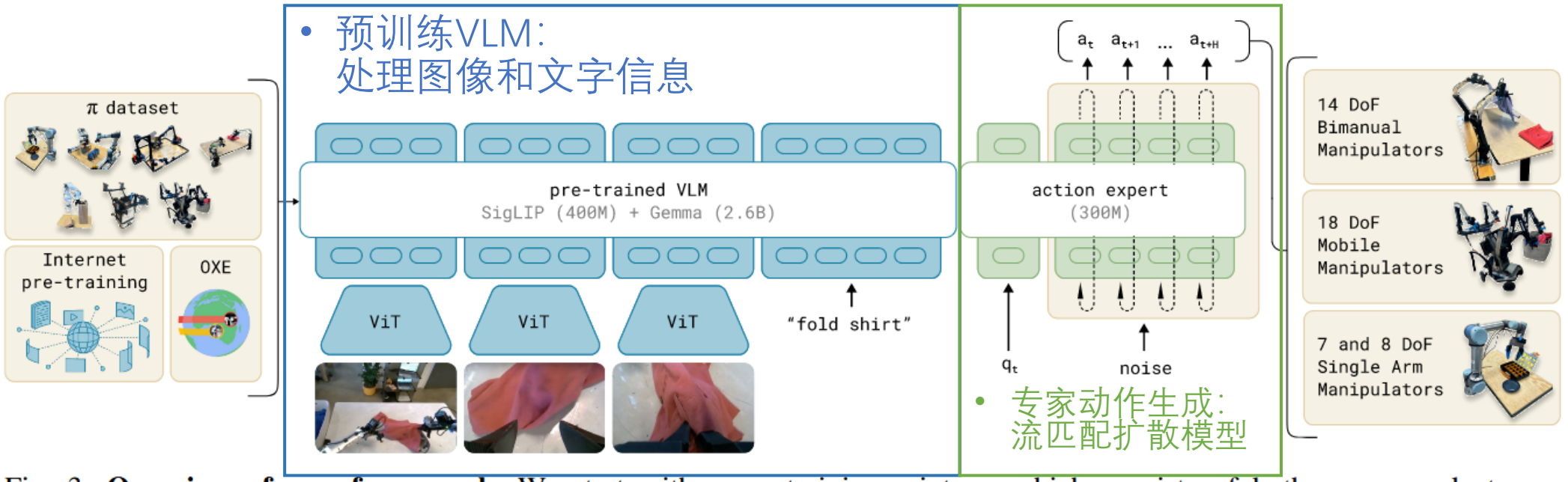
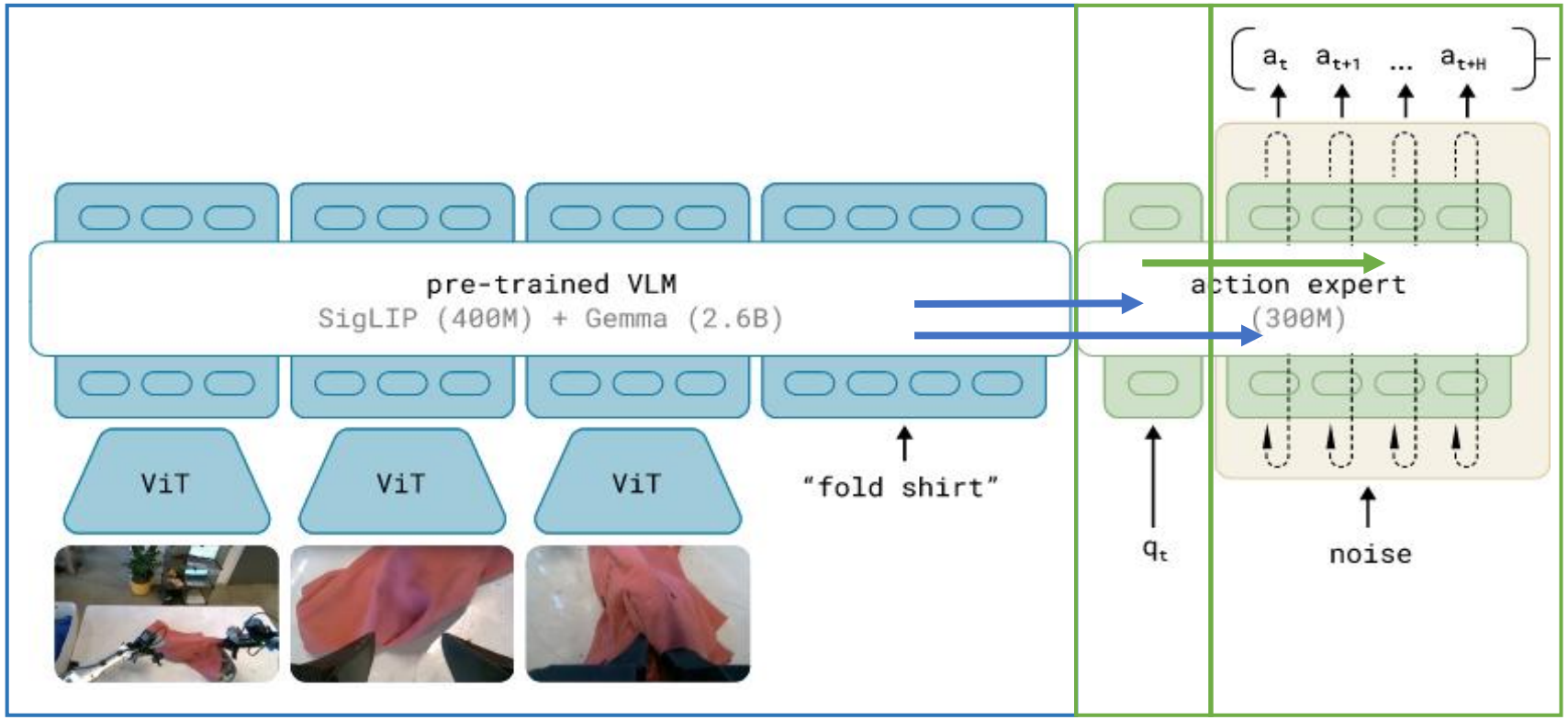


Fig. 3: **Overview of our framework.** We start with a pre-training mixture, which consists of both our own dexterous manipulation datasets and open-source data. We use this mixture to train our flow matching VLA model, which consists of a larger VLM backbone and a smaller *action expert* for processing robot states and actions. The VLM backbone weights are initialized from PaliGemma [5], providing representations learned from large-scale Internet pre-training. The resulting π_0 model can be used to control multiple robot embodiments with differing action spaces to accomplish a wide variety of tasks.

- **注：这一框架与Encoder-Decoder模型存在本质性差异**

Pi0：整体架构

- Blockwise Casual Attention Mask



- Block内部：
  Full Bidirectional Attention
- Block之间：
  Block1 仅可见自己
- Block2 可见自己与Block1
- Block3可见自己与Block1 2
- K V 缓存
- Block1 2计算k,v后缓存，供block3运算调用；
- 在Transformer的内部实现了不同频率的分层控制，提高了计算效率。

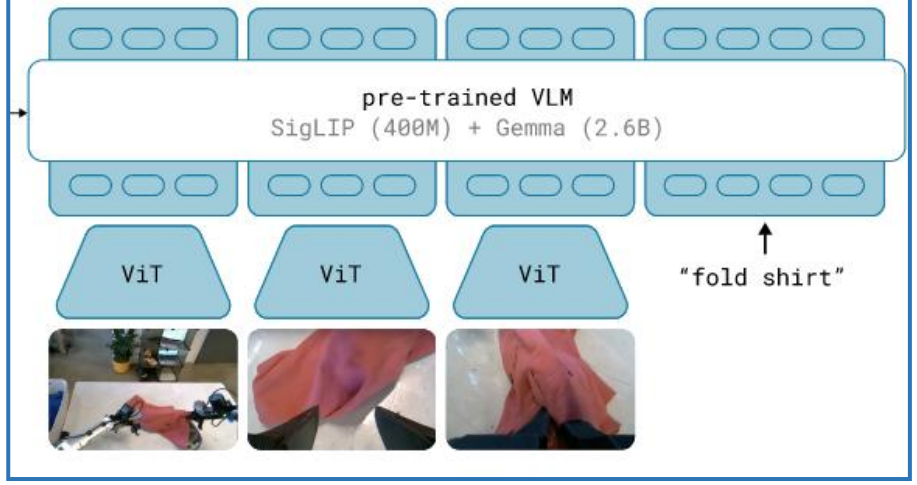
- 【Block1】~1Hz
- 图像、文字信息

- 【Block2】~1Hz
- 机器人本体信息

- 【Block3】~50Hz
- 去噪生成动作

Pi0：预训练VLM PeliGemma

- 预训练VLM：
  处理图像和文字信息



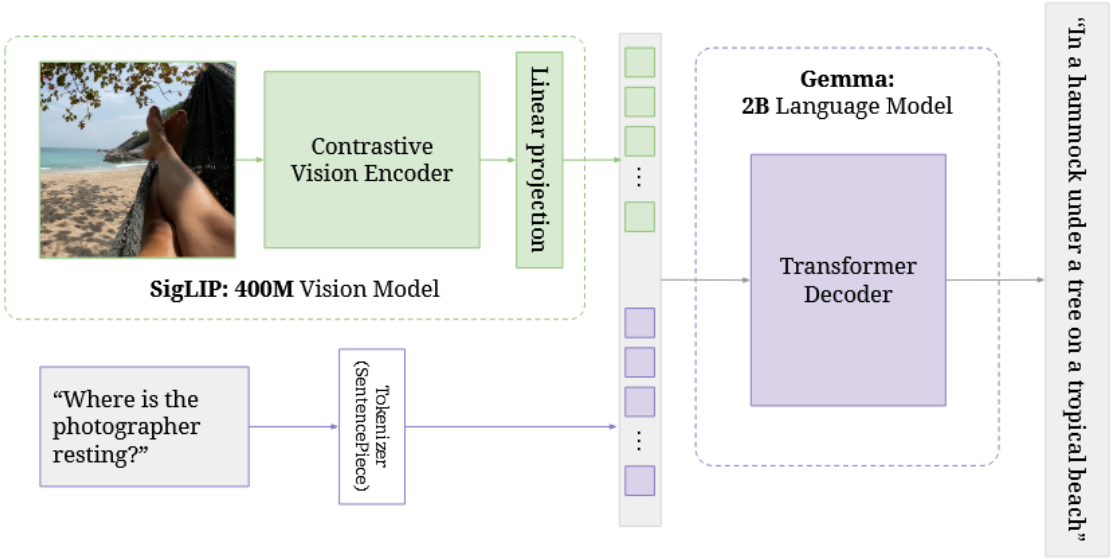
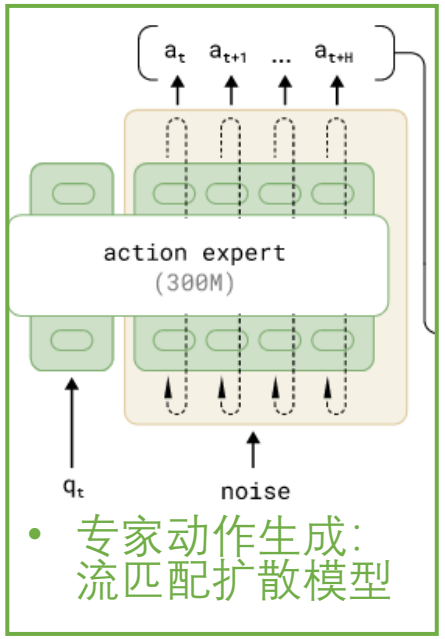


Figure 1 | PaliGemma's architecture: a SigLIP image encoder feeds into a Gemma decoder LM.

Pi0：动作专家模型：流匹配模型



- 专家动作生成：流匹配扩散模型

- 1. 训练目标Loss

$$L^\tau(\theta) = \mathbb{E}_{p(\mathbf{A}_t|\mathbf{o}_t),q(\mathbf{A}_t^\tau|\mathbf{A}_t)}||\mathbf{v}_\theta(\mathbf{A}_t^\tau, \mathbf{o}_t) - \mathbf{u}(\mathbf{A}_t^\tau|\mathbf{A}_t)||^2$$

- t　物理时间步
- τ　为流匹配模型去噪引入的虚时间轴 τ ∈ [0,1)
- A_t　来自数据集的动作Bunch 包含H个a_t动作矢量
- A_t^τ　根据A_t的构造的噪声动作组

$$\mathbf{A}_t^\tau = \tau\mathbf{A}_t + (1-\tau)\epsilon,\ \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

- v_θ　模型生成的速度，用于去噪生成动作组;
- u　去噪场分布
- o_t　在t时刻获得的状态观测 $\mathbf{o}_t = [\mathbf{I}_t^1, ..., \mathbf{I}_t^n, \ell_t, \mathbf{q}_t]$

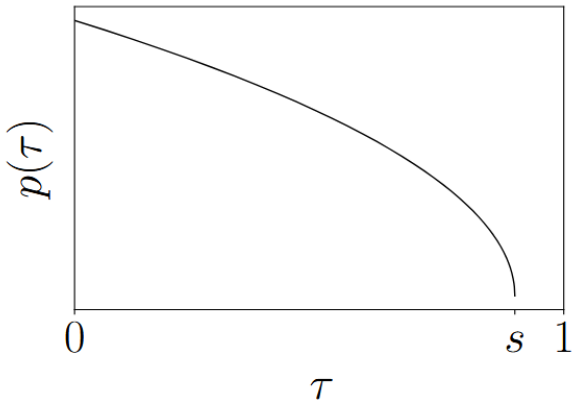
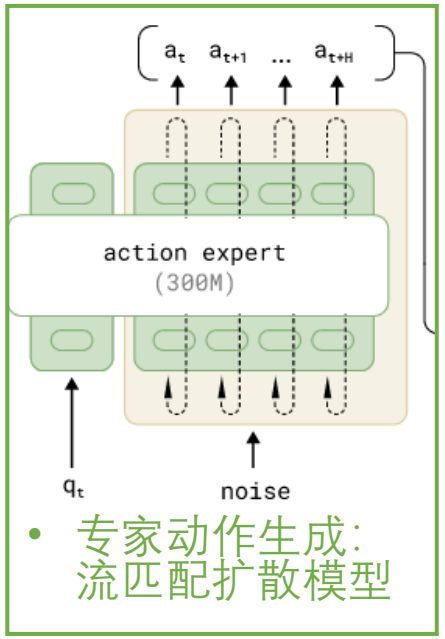


Fig. 14: **Flow matching timestep sampling distribution.** We sample τ from a shifted beta distribution that emphasizes lower timesteps (corresponding to noisier actions), and does not sample timesteps at all above a cutoff value s. We use $s = 0.999$ in our experiments.

Pi0：动作专家模型：流匹配模型



- 专家动作生成：流匹配扩散模型

- 2. 去噪过程
- 前向积分过程：

$$W_3 \cdot \text{swish}(W_2 \cdot \text{concat}(W_1 \cdot \mathbf{a}_{t'}^{\tau}, \phi(\tau)))$$

$$\mathbf{A}_t^{\tau+\delta} = \mathbf{A}_t^{\tau} + \delta \mathbf{v}_\theta(\mathbf{A}_t^{\tau}, \mathbf{o}_t)$$

- δ　沿虚时间轴积分步长
- 如 对于10步去噪，有$\delta = 0.1$

- 每一次完成去噪获得一组后续动作：

$$[a_t, a_{t+1}, a_{t+2}, \dots, a_{t+H}]$$

- 保证了关节运动的实时性和连续性

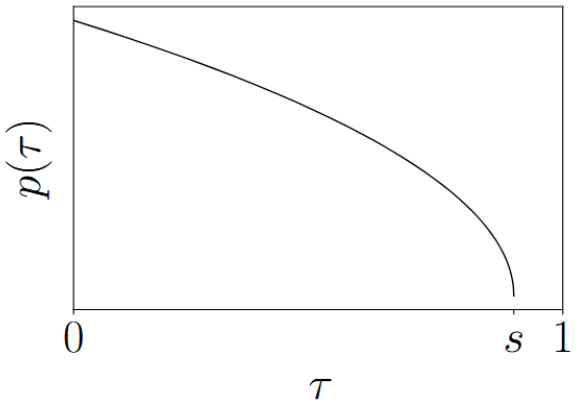
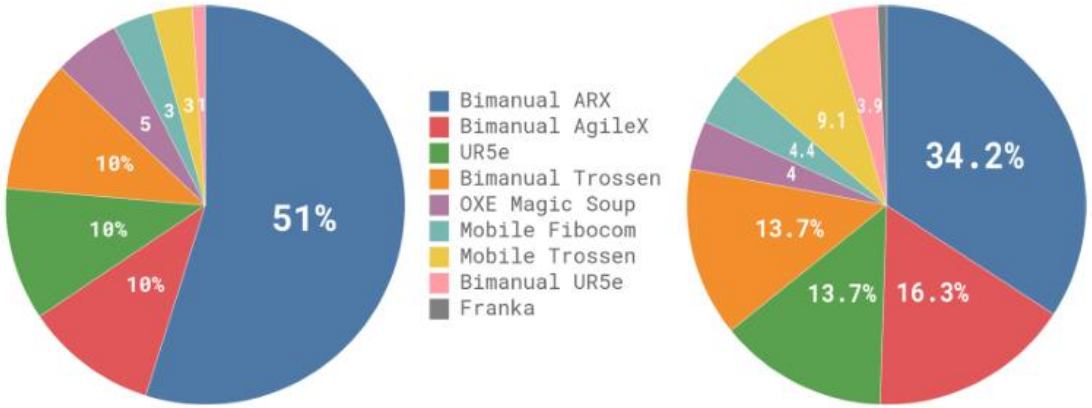


Fig. 14: **Flow matching timestep sampling distribution.** We sample τ from a shifted beta distribution that emphasizes lower timesteps (corresponding to noisier actions), and does not sample timesteps at all above a cutoff value s. We use $s = 0.999$ in our experiments.

预训练数据集构建:

数据集	类型	数据量	说明
开源数据集 OXE , Bridge v2, DROID		总量9.1%	包含有多种机器人执行多种任务
自主构建数据集	总量	903M timesteps	共包含68种不同动作
	单臂	106M timesteps	
	双臂	797M timesteps	



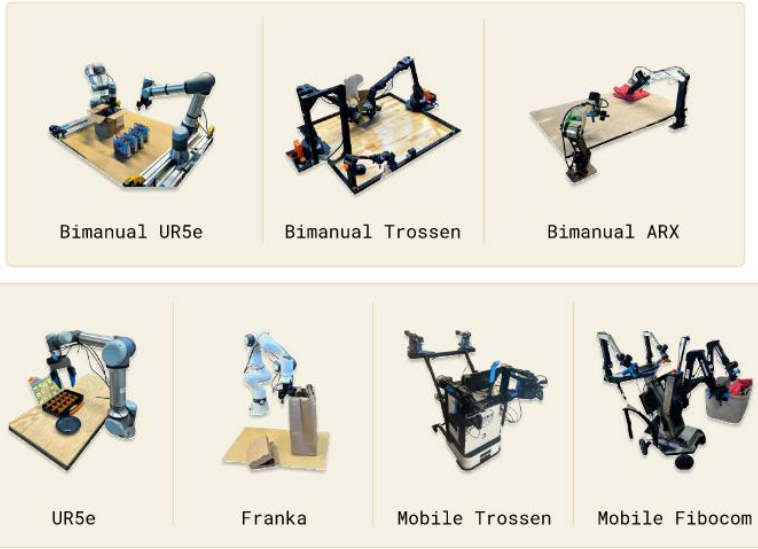
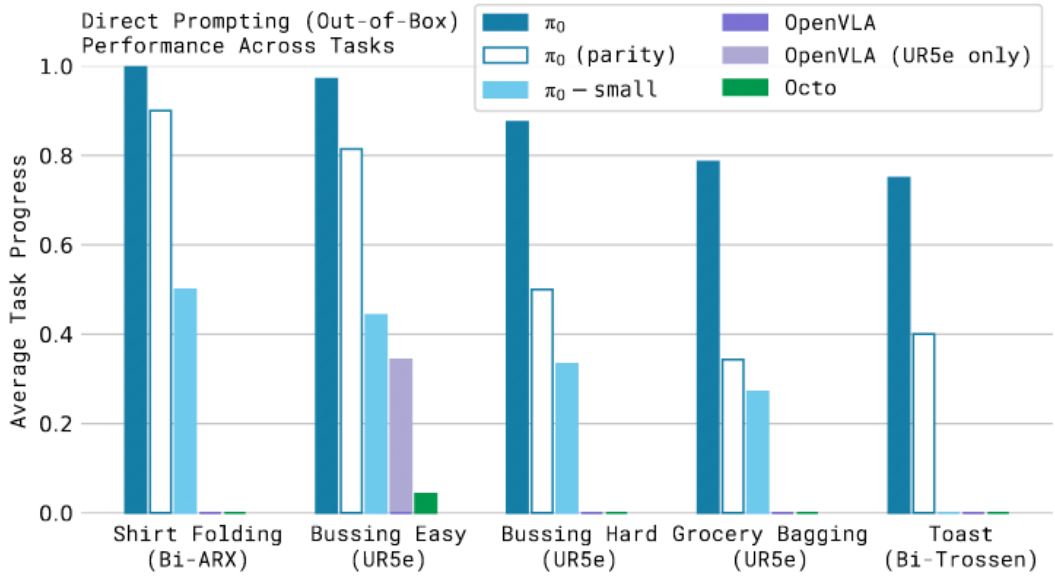


Fig. 5: **The robots used in our experiments.** These include single and dual-arm manipulators with 6-DoF and 7-DoF arms, as well as holonomic and nonholonomic mobile manipulators. π_0 is trained jointly on all of these platforms.

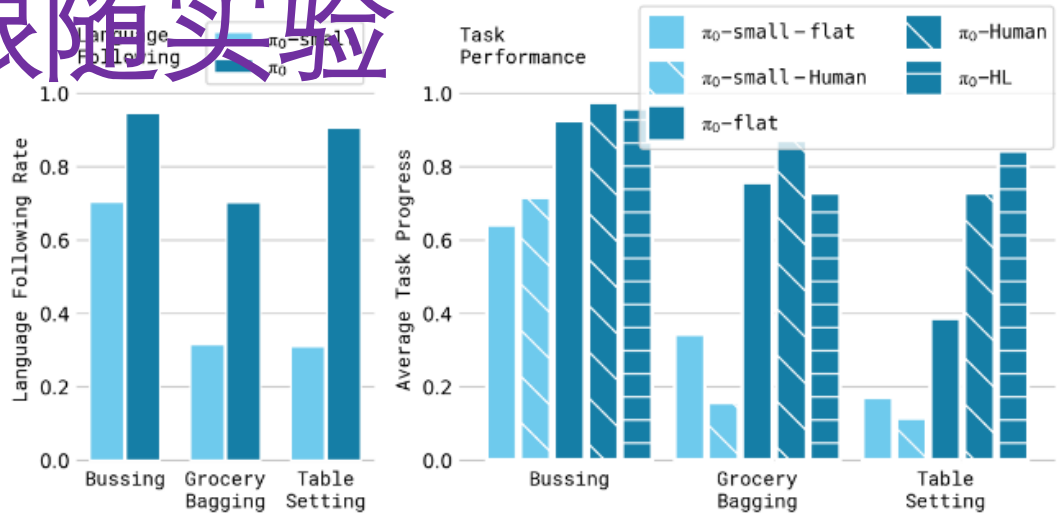
Since the datasets are somewhat imbalanced in size (e.g., the more difficult laundry folding tasks are overrepresented), we weight each task-robot combination by $n^{0.43}$, where n is the number of samples for that combination, such that over-represented combinations are down-weighted.

实验与结果：A-基础模型实验



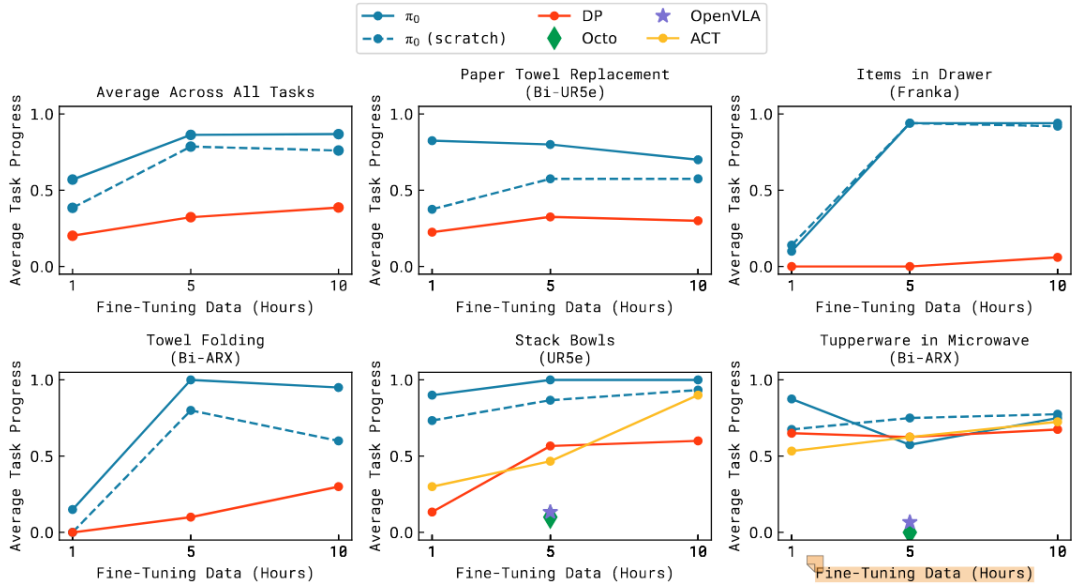
模型	说明	实验目的
Pi0	所提出的完整模型	验证所提出模型的部署应用效果。
Pi0 (parity)	减少了预训练样本的Pi0	训练样本数量与OpenVLA等模型基本一致，横向比较模型架构的合理性。
Pi0-small	用随机初始化的小尺寸VLM代替了Pi0中的PaliGemma	验证VLM中的互联网知识在操作任务中所起到的作用。

实验与结果：B-语言跟随实验



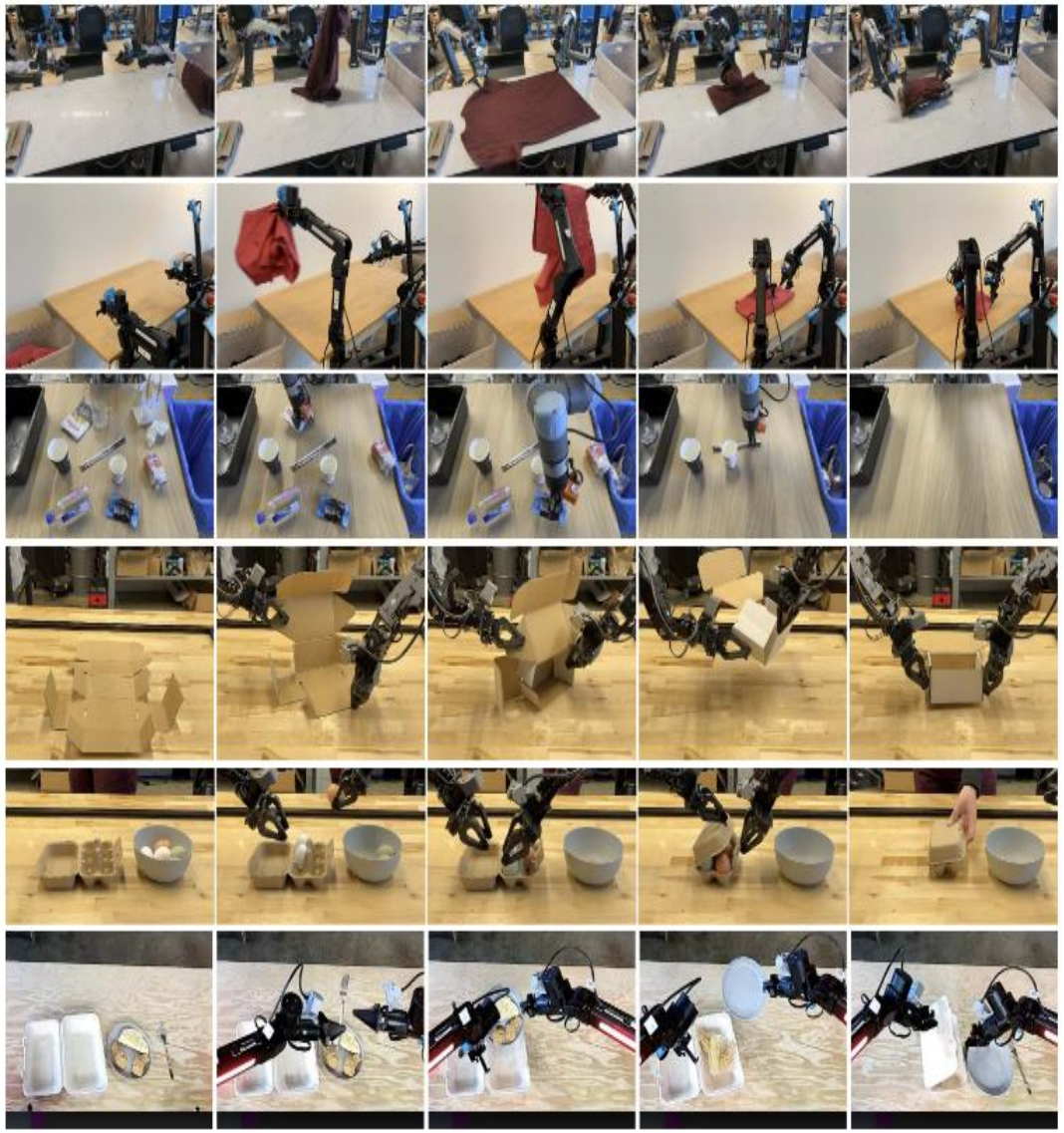
模型	模型	任务文字描述	实验目的
Pi0-small-flat	用随机初始化的小尺寸VLM代替了Pi0中的PaliGemma	仅提供任务文字描述 不做动作拆解	探究去除VLM部分参数中的互联网知识，对Pi0模型的影响。
Pi0-small-Human	用随机初始化的小尺寸VLM代替了Pi0中的PaliGemma	人工对任务进行动作拆解 提供文字描述	
Pi0-flat	所提出完整模型	仅提供任务文字描述 不做动作拆解	探究任务描述拆解为动作描述对Pi0模型的影响。
Pi0-Human	所提出完整模型	人工对任务进行动作拆解 提供文字描述	
Pi0-HL	所提出完整模型	使用VLM模型SayCan进行任务拆解	探究更高层任务拆分的可能性

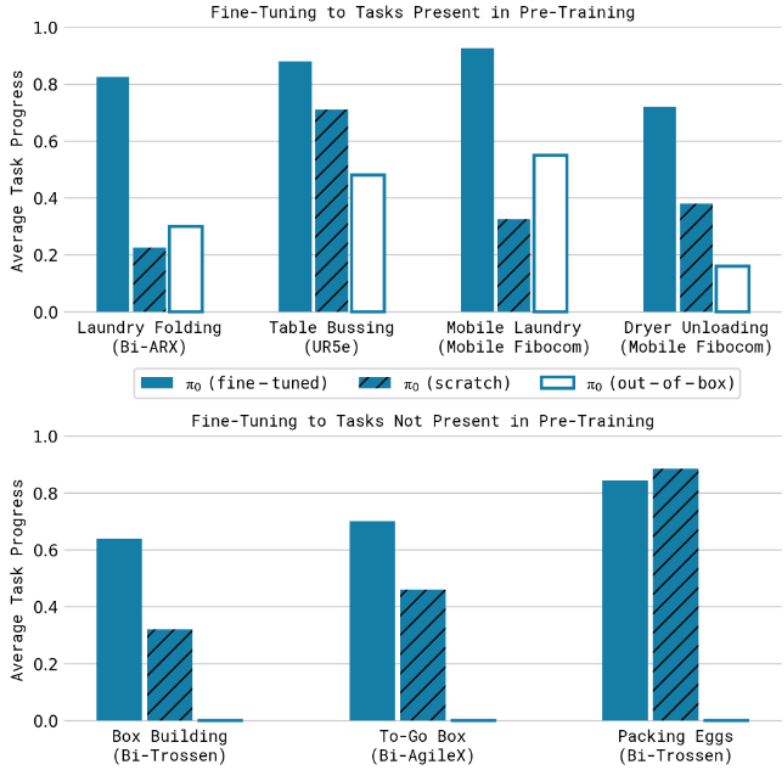
实验与结果：C-新操作任务学习



模型	说明	实验目的
Pi0	所提出的完整模型	探究后训练数据集大小和微调效果的关系
Pi0 (scratch)	不进行预训练仅进行微调	探究预训练对完成后训练后执行特定任务效果的影响

实验与结果：D-处理复杂多阶段任务





模型	预训练	后训练	实验目的
Pi0 (fine-tuned)	√	√	探究预处理和后处理对处理复杂多阶段任务的影响。
Pi0 (scratch)	×	√	
Pi0 (out-of-box)	√	×	

结论与评价

- 1. 创新性的视觉语言动作架构
  - 在单个Transformer中结合了两组参数，实现了分层控制；
  - 相较于OpenVLA等工作，引入流匹配模型的动作组生成增强了稳定性和实时性；
  - 相较于DP等工作，和VLM模型的紧密捆绑极大增强了对场景的理解能力；
- 2. 构建了巨大的操作任务数据集
- 3. 严密的实验论证

  从基础任务执行、语言跟随能力、新任务微调能力、多阶段复杂任务处理能力4个方面针对不同侧重展开消融实验，验证了模型的有效性和各部分的必要性。
- 4. 关注到：

  在Pi0-HL中引入了更高层的VLM进行任务规划，为之后RoboAgent的研究搭建了良好的基座。

后续工作

[摘要]

自回归序列模型（如基于Transformer的VLA策略）在捕捉复杂且具有泛化能力的机器人行为方面非常有效。然而，这类模型需要我们对连续的动作信号进行离散化（即令牌化），这决定了模型预测的离散符号如何映射为连续的机器人动作。我们发现，当前对机器人动作进行令牌化的方法通常采用简单的按维度、按时间步划分区间的方式，但在从高频机器人数据中学习灵巧技能时，这些方法往往表现不佳。

为了解决这一挑战，我们提出了一种基于**离散余弦变换（DCT）的机器人动作压缩式令牌化方案。我们的方法称为频域动作序列令牌化（FAST）**，它使我们能够训练适用于高度灵巧且高频率任务的自回归VLA模型，而标准的离散化方法在此类任务中完全失效。

在FAST的基础上，我们发布了**FAST+**，一个通用的机器人动作令牌器。它在100万个真实机器人动作轨迹上进行了训练，可作为黑盒工具用于各种机器人动作序列，适配多样的动作空间和控制频率。

最后，我们展示了，当FAST与**π0 VLA**结合时，该方法能够扩展至1万小时的机器人数据训练，**在性能上可与扩散模型VLA媲美，同时将训练时间最多减少5倍。**

FAST: Efficient Action Tokenization for Vision-Language-Action Models

Karl Pertsch[*,1,2,3], Kyle Stachowicz[*,2],
Brian Ichter[1], Danny Driess[1], Suraj Nair[1], Quan Vuong[1], Oier Mees[2], Chelsea Finn[1,3], Sergey Levine[1,2]
[1]Physical Intelligence, [2]UC Berkeley, [3]Stanford
https://pi.website/research/fast

***Abstract*—Autoregressive sequence models, such as Transformer-based vision-language action (VLA) policies, can be tremendously effective for capturing complex and generalizable robotic behaviors. However, such models require us to choose a tokenization of our continuous action signals, which determines how the discrete symbols predicted by the model map to continuous robot actions. We find that current approaches for robot action tokenization, based on simple per-dimension, per-timestep binning schemes, typically perform poorly when learning dexterous skills from high-frequency robot data. To address this challenge, we propose a new compression-based tokenization scheme for robot actions, based on the discrete cosine transform. Our tokenization approach, Frequency-space Action Sequence Tokenization (FAST), enables us to train autoregressive VLAs for highly dexterous and high-frequency tasks where standard discretization methods fail completely. Based on FAST, we release FAST+, a *universal* robot action tokenizer, trained on 1M real robot action trajectories. It can be used as a black-box tokenizer for a wide range of robot action sequences, with diverse action spaces and control frequencies. Finally, we show that, when combined with the π_0 VLA, our method can scale to training on 10k hours of robot data and match the performance of diffusion VLAs, while reducing training time by up to 5x.**

I. Introduction

Large, high-capacity Transformer models can be tremendously effective for capturing complex and generalizable robotic behaviors both from scratch [8, 69, 51, 6, 20, 62] and using models pre-trained for next-token prediction on Internet-scale image-text corpora [10, 39, 63, 7, 65]. However, these models require choosing a tokenization of the continuous action signal, which determines how the discrete symbols predicted by the model map to continuous robot actions [64, 34, 41, 12]. It is widely known that a good choice of tokenization can be critical to the performance of sequence models [55, 57]. Prior robotic policies of this sort typically use naïve tokenization strategies based on a per-dimension, per-timestep binning scheme [9, 10, 39]. We find that such methods perform poorly when learning dexterous skills with high-frequency control (see Figure 2, right). We observe that correlations between time steps are a major challenge for naïve tokenization strategies when predicting sequences of

*: Core contributors
Correspondence to: research@physicalintelligence.company

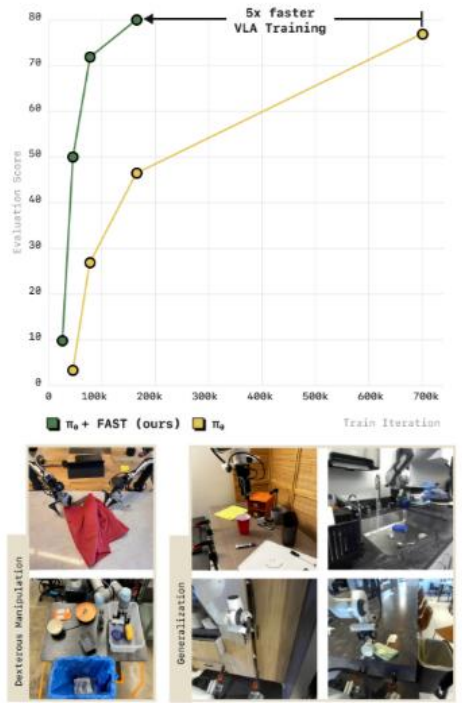


Fig. 1: We propose FAST, a simple yet effective approach for tokenization of robot action trajectories via time-series compression. FAST enables training of autoregressive VLAs that solve complex dexterous manipulation tasks and generalize broadly to new scenes. We use it to train π_0-FAST, a generalist robot policy that matches the performance of the state-of-the-art π_0 diffusion VLA on dexterous and long-horizon manipulation tasks, while training 5x faster (**top**).