FAST：面向视觉-语言-动作模型的高效动作词元化

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

Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, Sergey Levine

arXiv 2025 · arXiv:2501.09747v1
底本：https://arxiv.org/abs/2501.09747v1
许可：CC BY 4.0 · https://creativecommons.org/licenses/by/4.0/
整理与中文翻译：本地文献库 S16，2026-11-25

> 本册为完整中文改编译本，涵盖正文、附录、所有原图和表格；参考文献保留英文。翻译与重排为本次改动，原作者不为本译本背书。逐段英文对照、来源坐标与原始资产位于同目录“中英对照”。

阅读方式

先读问题定义与方法，再用实验设计校准结论。每段小字编号对应原文 HTML 锚点；可在 paper.md 与 source_map.json 中检索同一编号。数学公式以原文 MathML 排印，图表数据不作重算或改写。

英文原文 19 页。以下页码是中文重排页码，与原文页码不同；来源标记中的 p. 指英文 PDF 页码。

p.1 · abstract

自回归序列模型（如基于 Transformer 的视觉-语言-动作 VLA 策略）可以十分有效地刻画复杂且具有泛化能力的机器人行为。不过，这类模型要求我们为连续动作信号选择一种词元化方式，它决定模型预测的离散符号如何映射为连续机器人动作。我们发现，现有按动作维度、按时间步分别分箱的动作词元化方案，在从高频机器人数据中学习灵巧技能时通常表现不佳。为此，我们提出一种基于离散余弦变换、以压缩为核心的动作词元化方案。频域动作序列词元化（Frequency-space Action Sequence Tokenization，FAST）使我们能够在标准离散化方法完全失效的高灵巧性、高频任务上训练自回归 VLA。我们还发布 FAST+：一个在 100 万条真实机器人动作轨迹上训练的通用动作词元化器，可作为黑盒用于具有不同动作空间与控制频率的多种机器人动作序列。最后，结合 π_0 VLA 后，我们的方法可扩展至 1 万小时机器人数据的训练，达到扩散 VLA 的性能，同时将训练时间最多缩短 5 倍。

I 引言

p.1 · S1.p1

大容量 Transformer 模型无论从零训练 [8,69,51,6,20,62]，还是利用在互联网规模图文语料上进行下一词元预测预训练的模型 [10,39,63,7,65]，都能有效刻画复杂且可泛化的机器人行为。但必须选择连续动作信号的词元化方式，以确定预测的离散符号如何映射到连续动作 [64,34,41,12]。词元化的选择对序列模型性能至关重要 [55,57]。以往机器人策略通常按维度、按时间步独立分箱 [9,10,39]。我们发现，它们在高频控制的灵巧技能学习中表现不佳（图 2 右）。高频控制通常预测一段未来动作，即“动作块”；在这里，时间步间的相关性成为简单词元化的主要障碍。高度相关的动作词元削弱了自回归 VLA 下一词元预测目标的效力。直观地说，只需复制最近的动作词元，就可能取得较低的预测损失，使模型停留在较差的局部最优。

p.2 · S1.F2

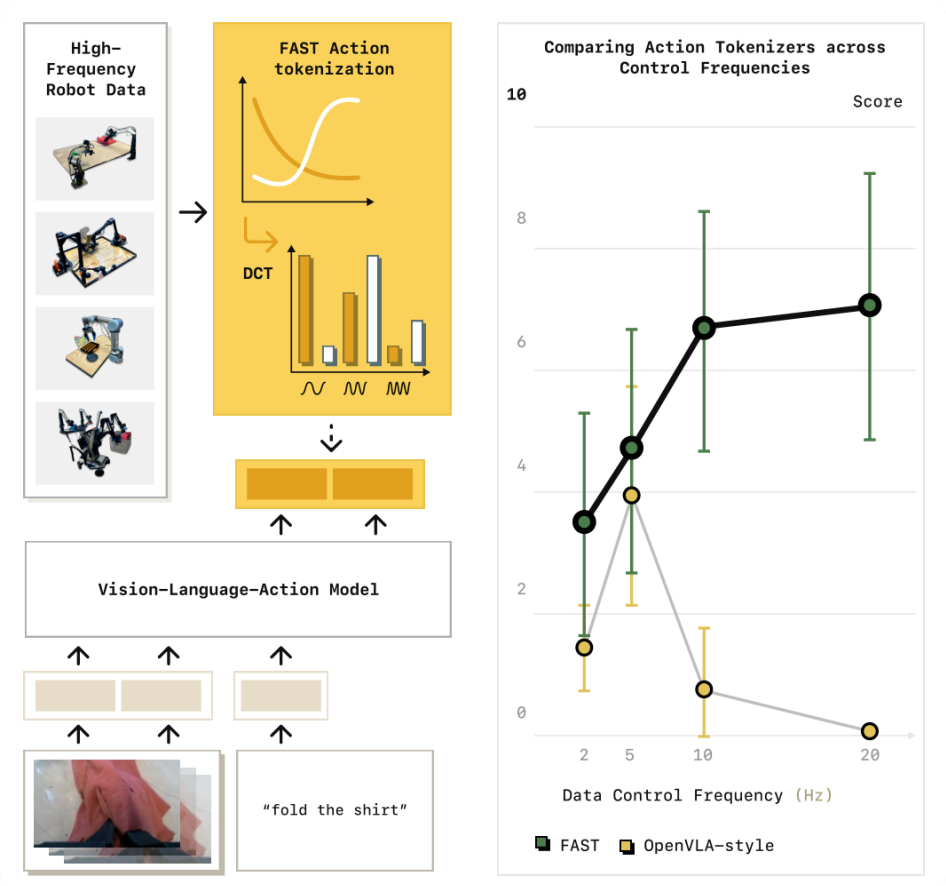


Fig. 2 : Left : FAST tokenization enables training of autoregressive Transformers for dexterous robot control via simple next token prediction. Right : FAST outperforms popular binning tokenization schemes, e.g., used in OpenVLA [39] , particularly for high-frequency robot data.

图 2：左：FAST 词元化使自回归 Transformer 能通过简单的下一词元预测学习灵巧机器人控制。右：FAST 优于 OpenVLA [39] 等使用的常见分箱方案，在高频机器人数据上尤其明显。

p.2 · S1.p2

本文从基本原理出发提出新的词元化策略。关键见解是：训练前应压缩机器人动作信号，以降低相邻词元的相关性。我们借鉴语言模型常用的字节对编码等基于压缩的词元化策略 [27,57]。但机器人动作连续，因此应采用相适应的压缩方式。我们以离散余弦变换（DCT）编码为基础；它广泛用于图像等连续信号的压缩，例如 JPEG。由此得到的频域动作序列词元化 FAST，只通过简单的下一词元预测（图 2 左），就能在标准离散化方法完全失败的高灵巧性、高频任务上训练自回归 VLA。此外，FAST 首次让近期提出的大规模、多任务、非受控环境机器人操作数据集 DROID [38] 上的 VLA 高效训练成为可能。得到的策略是首个仅凭自然语言提示、即可在未见环境中成功开展零样本评估的语言条件通用操作策略。

p.2 · S1.p3

基于 FAST，我们开发了通用机器人动作词元化器 FAST+。它在 100 万条真实动作轨迹上训练，覆盖多种机器人形态、动作空间与控制频率。FAST+ 能有效编码从单臂、双臂到移动机器人的广泛动作序列，是训练自回归 VLA 的良好即用型词元化器。与 π_0 VLA 集成后，基于 FAST 的自回归 VLA 可扩展至 1 万小时机器人数据，在多种任务上达到扩散 VLA 的性能，而训练时间最多减少 5 倍（图 1）。

p.1 · S1.F1

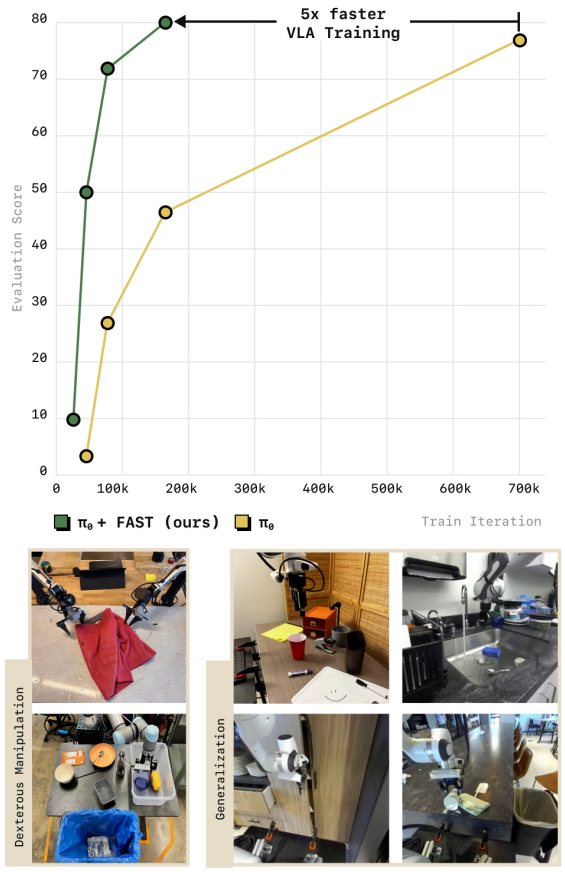


Fig. 1 : We propose FAST, a simple yet effective approach for tokenization of robot action trajectories via time-series compression. FAST enables training of autoregressive VLAs that solve complex dexterous manipulation tasks and generalize broadly to new scenes. We use it to train π_0 -FAST, a generalist 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).

图 1：FAST 通过时间序列压缩对机器人动作轨迹进行词元化，方法简单而有效。它支持训练可完成复杂灵巧操作、并能广泛泛化至新场景的自回归 VLA。我们据此训练了通用机器人策略 π_0-FAST；在灵巧操作与长时程操作任务上，其性能匹敌先进的 π_0 扩散 VLA，而训练速度快 5 倍（上图）。

II 相关工作

p.2 · S2.p1

语言、文本与音频的词元化。词元化是现代 Transformer 自回归序列模型训练流程的关键组成；其选择可显著影响训练与下游性能 [55]。尽管已有工作研究直接处理比特流的“无词元化”语言模型 [28,53]，多数语言模型仍在训练前进行文本词元化。常用的字节对编码 BPE [27,55] 通过将高频词元序列合并为新词元来压缩文本。对图像而言，学习式压缩是一条有效路径：预训练视觉编码器可将图像表示为“软词元” [44]；向量量化自编码器则支持完全自回归的图像输入输出 [22,59]。类似方法可扩展至视频 [66]。音频生成和语音合成与动作预测同具时间序列结构，先进模型通常将音频编码为频域声谱图 [29]，或采用学习到的向量量化器 [68]。

p.2 · S2.p2

视觉-语言-动作模型。近期多项研究发展了在不断扩大的机器人学习数据集 [52,38,60,24,47,35] 上训练的通用策略 [9,51,6,10,20,39,62,11]。一条有前景的路线是 VLA [10,17,39,67,7,63,73,71,13,11]：将在互联网规模图文上预训练的视觉语言模型微调用于机器人控制。数十亿参数的视觉语言骨干为拟合大型机器人数据集提供必要表达能力；复用互联网预训练权重，也有助于遵从多样语言命令，并泛化到新物体与新背景

[10,39,67,63,36]。目前多数 VLA 仍局限于相对简单的低频任务，最常见的自回归设计尤其如此 [10,39]。我们表明，这是其动作词元化方案的直接后果：它妨碍灵巧任务训练。新的方案让我们首次能够在灵巧、高频机器人数据上训练自回归 VLA。

p.3 · S2.p3

VLA 训练的动作表示。既有研究探索了多种机器人策略动作参数化。一路采用语言子任务 [21,2,4] 或关键点 [50,32,25,19] 等“语义”动作表示；它们通常只需少量样本，甚至无需机器人示范就能零样本执行任务 [50,32,25]，但依赖手工设计的低层控制器，限制了通用性。另一路直接训练 VLA，根据图像和语言指令输出低层控制命令。最常见设计把动作嵌入离散词元，以标准自回归序列模型生成。现有方案按维度、按时间步独立分箱 [9,10,39]，难以扩展到高频控制。我们以时间序列压缩为基础的新方案则支持高频数据。还有工作不使用词元化，而采用回归头或增加用于扩散解码的新权重 [20,7,41,63]。相比之下，我们不修改底层预训练 Transformer，可方便地用于任何预训练自回归 Transformer；在许多任务上与先进扩散 VLA [7] 性能相当，却显著降低训练计算成本。

p.3 · S2.p4

另一类相关工作采用向量量化动作表示 [41,3,49]，训练向量量化编码器-解码器网络；其重建质量对超参数和结构敏感 [66]。我们发现，这类方法在粗糙、低保真的重建任务上表现良好，却在需要细粒度控制的高频任务上失效。FAST 超参数很少，同时具有高精度动作重建与强压缩能力。

III 预备知识

p.3 · S3.p1

问题定义。目标是训练策略 $\pi(a_{1:H} \mid o)$，将观测 o 映射为未来机器人动作序列 $a_{1:H}$。假设策略输出“动作块” [69,40]，即 H 个动作组成的序列 [15,7,69]，有利于生成时间一致的动作并减少误差累积。动作词元化需定义映射 $\mathcal{T}_a : a_{1:H} \to [T_1, ..., T_n]$，把维数为 $|\mathcal{A}|$ 的连续动作序列 $a_{1:H}$ 映射为大小 $|\mathcal{V}|$ 的词表中的 n 个离散词元 $T \in |\mathcal{V}|$。词元数 n 可以随动作序列改变，正如同长度句子也可能分成不同数量的文本词元。

p.3 · S3.p2

分箱动作词元化。最常见方案是简单的分箱离散化 [8,10,39,72,56]。对动作 a，各维独立处理，将训练数据中的取值范围划分为 N 个等宽箱，常取 $N = 256$。对 D 维动作序列 $a_{1:H}$，逐时间步应用后得到词元序列 $\mathcal{T}_a\left(a_{1:H}\right) = [T_{1,1}, ..., T_{1,D}, ..., T_{H,1}, ..., T_{H,D}]$。对高频数据而言，这并不理想：每个动作块很容易产生数百个词元，导致训练困难、推理缓慢。

IV 案例研究：词元化如何影响 VLA 训练?

p.3 · S4.p1

为说明现有词元化下自回归策略训练的困难，我们从简单的教学示例出发，构造时间序列数据集：预测穿过四个随机点的三次样条（图 3 下）。它反映了高频动作块策略的挑战，即给定条件信息预测连续动作序列。我们采用以往 VLA 的朴素方案，将序列的每个元素分别离散为 256 个箱之一（第 III 节），再训练小型自回归 Transformer 根据条件点预测词元化信号。保持底层数据集不变，把每条目标序列的采样点数从 25 改到 800 并重复实验，以模拟不同采集频率下的策略训练。

p.3 · S4.F3

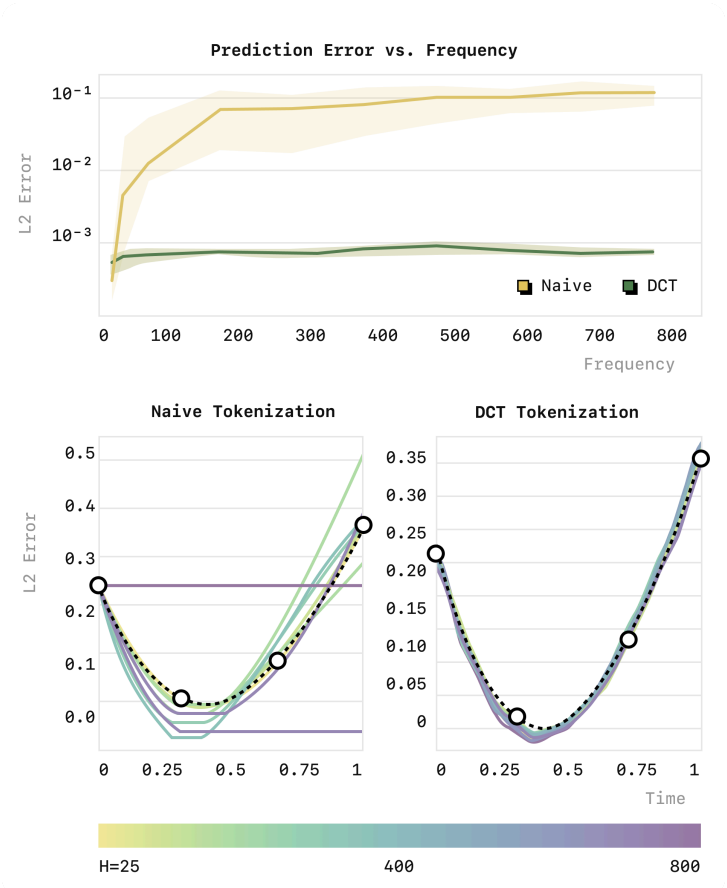


Fig. 3 : Effect of sampling rate on prediction performance. We train a small autoregressive transformer model on a didactic interpolation task, in which the network must predict the black dashed curve given the four circles. We find that models trained with the binning tokenization approach used in prior VLAs [10 , 39] produce increasingly poor predictions as we increase the sampling frequency of the underlying signal, due to strong correlation between consecutive tokens at high frequencies. Our FAST tokenization approach, based on the discrete cosine transform (DCT), addresses the problem and leads to high-quality predictions across all sampling rates.

图 3：采样率对预测性能的影响。小型自回归 Transformer 在教学性插值任务中，需要根据四个圆点预测黑色虚线。采用以往 VLA [10,39] 分箱词元化时，底层信号采样率越高，相邻词元相关性越强，预测越差。基于 DCT 的 FAST 解决了这个问题，在所有采样率上均能高质量预测。

p.4 · S4.p2

图 3 上方“naive”曲线给出各频率模型的平均预测均方误差 MSE。分箱模型在低采样率时预测良好，但随采样率升高，误差陡增，最终只复制首个动作（图 3 左下）。这不能归因于数据：底层分布复杂度没有变化，因此相同容量、相同训练步数的模型理应在各采样率上得到相近性能。那么发生了什么？

p.4 · S4.p3

需要考察学习目标本身。自回归模型根据此前全部词元预测下一词元，其学习信号与给定 $T_{1:i-1}$ 时 T_i 的边际信息量成正比。关键在于，逐时间步朴素词元化下，这一信息量随控制频率提高趋近于零：对平滑信号，时间步越短，每步变化按比例减小。这会大幅拖慢训练收敛，使拟合复杂高频数据困难。既有工作也有类似观察：OpenVLA 在低频 BridgeV2、RT-1 上表现良好，却难以拟合频率更高的 DROID [39]。本案例强调了为机器人动作设计更好词元化的必要性。

V 通过时间序列压缩实现高效动作词元化

p.4 · S5.p1

上一节说明，高频动作轨迹中的冗余导致每个词元边际信息量低、训练效果差。因此需要把高度冗余的动作信号压缩为少量高信息词元。本节先介绍连续时间序列的简单压缩方法（V-A），据此设计动作词元化算法（V-B），再说明通用机器人动作词元化器的训练（V-C）。

V-A 用离散余弦变换压缩时间序列

p.4 · S5.SS1.p1

连续时间序列的高效压缩已有大量研究，包括变换到频域后压缩 [18,1,61]，以及基于向量量化等学习式方法 [59,48]。本文的一点重要结论是：只要足够有效，任何用于动作目标的压缩方法都有望改善 VLA 训练速度。实际中仍需考虑词元化器训练复杂度、编码与解码效率等因素，以选择合适算法。

p.4 · S5.SS1.p2

本文采用基于 DCT [1] 的压缩算法。DCT 将连续信号表示为不同频率余弦分量之和：低频描述整体形状，高频反映突变。它简单、计算高效，对实际图像有良好压缩性，因此常用于 JPEG 等压缩算法 [61]。像素通常平滑变化，因而少量系数往往可表征输入的大部分信息；丢弃权重很低的频率分量即可压缩信号。与基于向量量化的学习式压缩不同，DCT 压缩是解析方法，极其简单且快速。

V-B FAST 词元化算法

p.4 · S5.SS2.p1

我们用 DCT 设计快速有效的 FAST。原始动作到词元的步骤见图 4。先归一化各动作维，使训练数据第 1 与第 99 百分位映射到 $[-1, \ldots, 1]$。这既把数据变到指定范围，也方便对动作尺度不同的跨形态数据编码；使用分位数可抵抗大数据集中偶发的离群动作。随后对每个动作维独立进行 DCT。为压缩转换后的信号，只需省略不显著系数；实现方式是缩放后取整，缩放系数作为超参数，控制词元化损失与压缩率的权衡。

p.5 · S5.F4

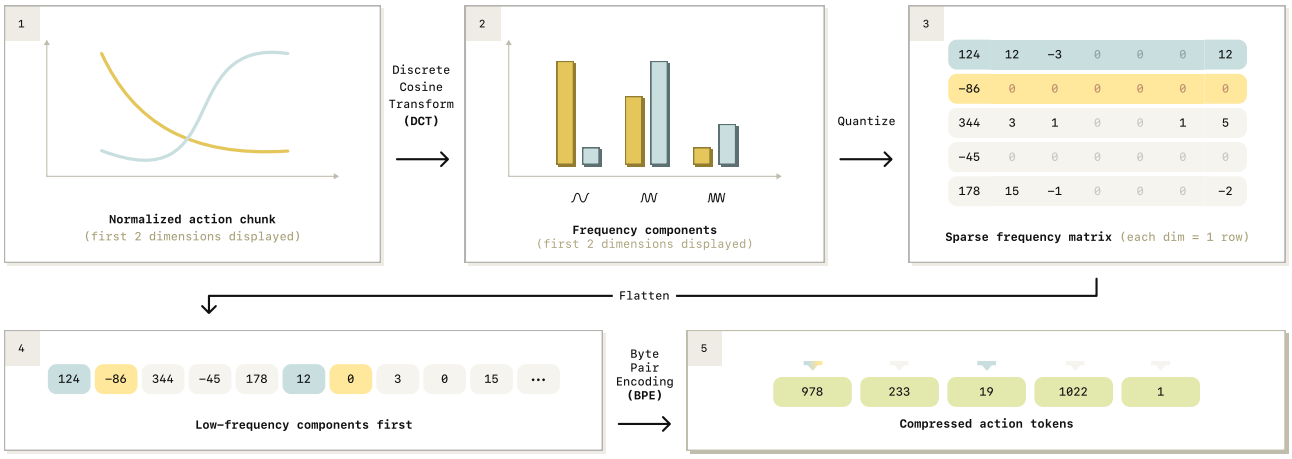


Fig. 4 : Overview of the FAST action tokenization pipeline. Given a normalized chunk of actions, we apply discrete cosine transform (DCT) to convert the signal to the frequency domain. We then quantize the DCT coefficients and use byte-pair encoding (BPE) to compress the flattened sequence of per-dimension DCT coefficients into the final action token sequence. See Section V-B for a detailed description.

图 4：FAST 动作词元化流程。对归一化动作块逐维施加 DCT，转到频域；量化系数后，将逐维 DCT 系数展平，再用 BPE 压缩为最终动作词元序列。详见第 V-B 节。

p.4 · S5.SS2.p2

取整后 DCT 系数矩阵通常稀疏，多数条目为零，每个动作维仅保留少数显著系数。要实际压缩，需把稀疏矩阵转成稠密词元序列。我们先按低频优先、动作维交错的顺序，将矩阵展为整数向量，再训练 BPE [27] 将其无损压缩成稠密动作词元。BPE 把零值分量“挤合”，并合并不同动作维间常见的系数组合。选择 BPE 是因

为高效实现丰富，且输出词表大小固定，易于集成到 VLM 现有词表中。也可采用 Huffman [33] 或 Lempel-Ziv [75]（gzip 的基础）等无损算法，本文留待未来研究。

p.5 · S5.SS2.p3

在 BPE 前展平 $|A| \times H$ DCT 系数矩阵的顺序，会显著影响策略训练。有两种选择：按列，即先拼接各维最低频分量；或按行，即先拼接一个动作维的所有频率分量。我们采用前者，因为自回归预测时先预测描述输出整体形状的低频分量，能获得更稳定的策略回放。

p.5 · alg1

scale γ, (for inference) BPE dictionary Φ
procedure FASTTokenizer($a_{1:H}$)
$C_j^i \leftarrow \mathrm{DCT}\left(a_{1:H}^i\right)$ ▷ Compute DCT coefficients
$\bar{C}_j^i \leftarrow \mathrm{round}\left(\gamma \cdot C_j^i\right)$ ▷ Quantize coefficients
$[T_k] \leftarrow \left[\bar{C}_1^1, \bar{C}_1^2, \ldots, C_2^1, \ldots, C_H^n\right]$ ▷ Flatten tokens
$\phi \leftarrow \mathrm{TrainBPE}(\mathcal{D} := \{[T_k]\})$ BPE Training:
Tokenization:
$\left[T_1, \ldots, T_{\hat{k}}\right] \leftarrow \mathrm{BPE}\left([T_1, \ldots, T_k], \phi\right)$
return action_tokens

Algorithm 1 FAST Tokenizer

算法 1：FAST 词元化器。输入缩放量 γ；推理时还需 BPE 词典 Φ。FASTTokenizer($a_{1:h}$)：①沿动作维分别计算 DCT 系数；②乘缩放量后取整以量化系数；③将词元按低频优先顺序展平；④训练时在展平序列上学习 BPE，词元化时用已训练词典执行 BPE；⑤返回 action_tokens。各行的精确赋值与符号按下方原算法保留。

p.5 · S5.SS2.p4

流程中的操作都容易求逆，可快速解码预测动作。词元化器只有两个超参数：取整前 DCT 系数的缩放量，以及 BPE 词表大小。两者并不十分敏感，所有单数据集实验均使用取整缩放量 10、BPE 词表大小 1024。相比之下，基于向量量化的端到端学习压缩模块 [59] 常训练繁琐，需针对数据集细调超参数才能得到良好重建 [66,48]。实验表明，我们更简单、易调的 DCT 方案可训练出性能更好的策略。

p.5 · S5.SS2.p5

我们在第 IV 节玩具示例中实证 DCT 词元化的收益。图 3 表明，在 DCT 压缩目标词元上训练的自回归模型，在广泛采样频率下持续保持低误差。算法 1 总结流程；第 VI 节在机器人控制问题上检验 FAST 的效果。

V-C 通用机器人动作词元化器

p.5 · S5.SS3.p1

唯一需要学习的组件是 BPE 编码器词表，原本每个新数据集都要重新训练。虽然通常仅需几分钟，仍会增加使用阻力。因此我们希望训练一个能编码任意机器人动作块的通用词元化器。为此，用上述流程在大型跨形态动作数据上训练：约 100 万个 1 秒动作块，来自单臂、双臂和移动操作机器人，覆盖关节与末端执行器控制空间、多种控制频率。详细混合比例见附录 A。训练后的 FAST+ 可作为黑盒，处理任意机器人设置的 1 秒动作序列；评估表明其可媲美针对各数据集单独调优的词元化器。

p.6 · S5.SS3.p2

代码发布。我们以 HuggingFace AutoProcessor 类发布预训练通用动作词元化器 FAST+，只需三行代码即可处理新动作块：

p.6 · S5.SS3.p3

```
from transformers import AutoProcessor
tokenizer = AutoProcessor.from_pretrained("physical-intelligence/fast", trust_remote_code=True)
tokens = tokenizer(action_chunk)
```

```
from transformers import AutoProcessor

tokenizer = AutoProcessor.from_pretrained(
 "physical-intelligence/fast",
 trust_remote_code=True
)
tokens = tokenizer(action_chunk)
```

p.6 · S5.SS3.p4

为获得最佳压缩效果，建议按 V-B 的分位数归一化，将动作映射到 $[-1, ..., 1]$，每次编码一个 1 秒动作块。模块也可方便地训练新 FAST 词元化器：先用 AutoProcessor.from_pretrained("physical-intelligence/fast", trust_remote_code=True) 加载，再调用 new_tokenizer = tokenizer.fit(action_dataset)。

```
from transformers import AutoProcessor

tokenizer = AutoProcessor.from_pretrained(
 "physical-intelligence/fast",
 trust_remote_code=True
)
new_tokenizer = tokenizer.fit(action_dataset)
```

VI 实验

p.6 · S6.p1

我们采用两个 VLA 骨干检验 FAST：π_0 [7] 与 OpenVLA [39]。先比较其他动作词元化方案并消融关键设计，再把 FAST 训练的 π_0 与先进的 π_0 流匹配（扩散）VLA 比较，最后检验 FAST 自回归训练扩展至 1 万小时跨形态灵巧操作数据的能力。

VI-A 实验设置

p.6 · S6.SS1.p1

策略实现。我们在常用 VLA 骨干上测试不同自回归动作词元化方案。多数实验采用基于 PaliGemma-3B [5] 的 π_0 [7]，也测试基于 Prismatic 7B [37] 的 OpenVLA [39]。训练时编码 1 秒动作块，遵循以往 VLA [10,39] 的做法，以动作词元覆盖 VLM 词表中最少使用的词元。微调时不冻结权重。更多训练设置见附录 C。

p.6 · S6.SS1.p2

评估任务。7 项任务（6 个真实、1 个仿真；图 5）同时检验高灵巧性与零样本泛化。① Libero：在 Libero [43] 的 Spatial、Object、Goal、10 四个套件上报告平均性能。② 清理餐桌 [7]（20 Hz）：UR5 单臂将 12 个物体分类放入垃圾桶或餐具收纳箱，要求准确抓取多样物体。③ 叠 T 恤 [7]（50 Hz）：ARX 双臂将初始平铺于桌面的不同衬衣折叠，要求精确抓取与移动。④ 杂货装袋 [7]（20 Hz）：UR5 把桌上 7 个物体装袋，不能使袋子倾倒或破裂，要求抓取不同物体并小心放入。⑤ 从烤面包机取吐司 [7]（50 Hz）：Trossen Viper-X 双臂取出两片面包并放到盘中，要求精确抓取与放置。⑥ 从衣篮取衣折叠 [7]（50 Hz）：ARX 双臂取出上衣、短裤，展开、铺平、折叠并叠放。这是最灵巧的任务，涉及精确抓取、动态展开、缠结后的重试修正，以及准确叠放；按单件衣物报告成功率。⑦ DROID 零样本桌面操作 [38]（15 Hz）：在完整 DROID 上训练，测试抓放、擦拭、开关抽屉等。测试环境完全未见，包括新桌面布置、背景、物体、视角和桌高。据我们所知，这是首次不经共训练或微调、仅用自然语言提示，在完全未见环境中零样本评估 DROID 策略。遵循 Black 等 [7]，杂货装袋、吐司机和衣篮折叠仅用于 VI-F 最强通用 VLA 的评估。数据与任务细节见附录 E。

p.6 · S6.F5

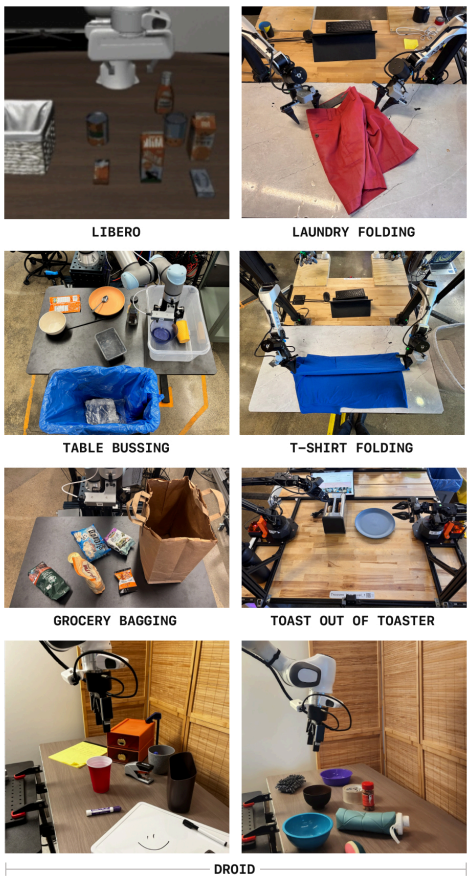


Fig. 5 : Evaluation environments. We test FAST across 7 evaluation environments: 6 real-robot tasks and 1 simulation environment. The tasks are designed to test VLA performance on highly dexterous tasks, like folding cloths from a laundry basket (“Laundry Folding”), and generalization, e.g., zero-shot table-top manipulation in unseen environments (“DROID”).

图 5：评估环境。7 个环境包括 6 个真实机器人任务和 1 个仿真环境，检验从衣篮取衣叠放等高灵巧性操作，以及在未见环境中零样本桌面操作（DROID）等泛化能力。

p.7 · S6.SS1.p3

比较对象。FAST 为各评估数据集分别训练的 DCT 词元化器；FAST+ 为在 100 万动作序列上训练的通用 DCT 词元化器。通用词元化器采用我们可汇集的最多样真实机器人数据，其中包含真实机器人评估任务的数据。二者均与 RT-2 [10]、RT-2-X [52]、OpenVLA [39] 等逐维分箱的“朴素词元化”比较；后者逐时间步编码动作块后拼接。我们也测试另一种压缩方案：训练模型通过有限标量量化 FSQ [48]（VQ-VAE [59] 的较简单替代）生成量化动作块表示。它曾用于高维图像编码 [48,66]，可视为本文基于压缩路线的消融：同样使用压缩表示，但以更复杂的学习式方法替代简单 DCT。

VI-B VLA 训练中的动作词元化比较

p.7 · S6.T1

数据集	Action Dimension	Control Frequency	平均词元数		压缩比
			朴素分箱	FAST	
BridgeV2	7	5 Hz	35	20	1.75
DROID	7	15 Hz	105	29	3.6
清理餐桌	7	20 Hz	140	28	5.0
叠 T 恤	14	50 Hz	700	53	13.2

TABLE I : Comparison of the average token count per action chunk for naïve tokenization and FAST. We use 1-second chunks in all datasets. With our method, each chunk requires many fewer tokens, particularly for high-frequency domains such as the T-shirt folding task, indicating that it is more effective at removing redundancy.

表 I：朴素词元化与 FAST 每动作块平均词元数。所有数据集均取 1 秒动作块。FAST 所需词元显著更少，叠 T 恤等高频领域尤其明显，表明它能更有效地消除冗余。

p.7 · S6.SS2.p1

表 I 先比较 FAST 与朴素分箱的压缩率。数据集覆盖不同动作维数、控制频率，均采用 1 秒动作块。两方法使用默认超参数，词元化误差相近。FAST 在所有数据集上都显著压缩输入，高频数据的收益特别明显。有趣的是，在每个领域，FAST 每只机械臂每块都约生成 30 个动作词元（双臂约 60 个），说明它找到的表示近似反映底层动作信号复杂度，而基本不依赖数据频率。

p.7 · S6.SS2.p2

这并非完全无损压缩：算法 1 的缩放参数 γ 控制压缩比与重建精度的权衡。表 I 数值在相近重建精度下比较；各词元化器压缩与保真关系曲线见附录 B。

p.7 · S6.SS2.p3

接着使用 VI-A 的策略架构与词元化方式训练，结果见图 6。

p.8 · S6.F6

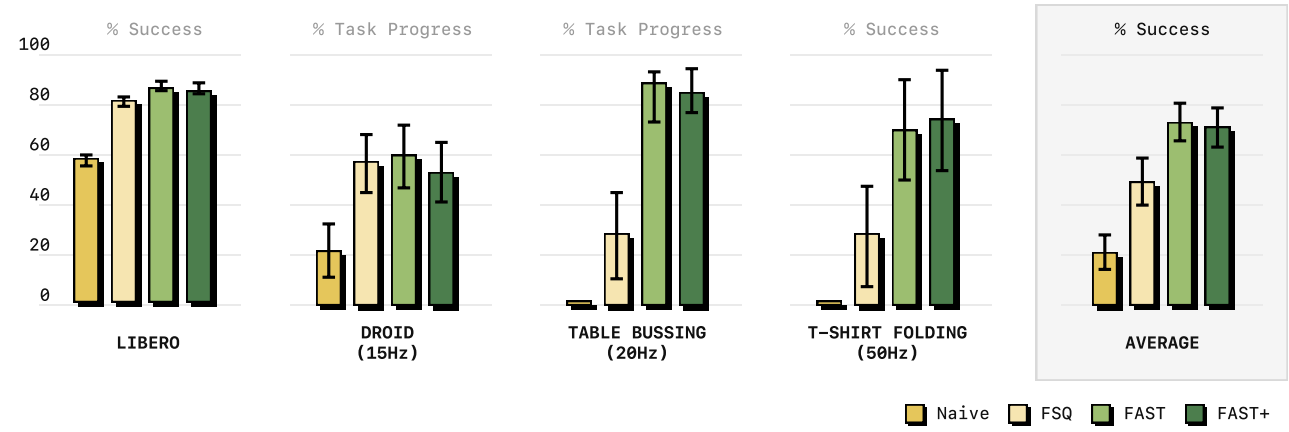


Fig. 6 : Comparison of policy performance using different tokenization approaches. We find that tokenization approaches that compress action targets (FAST, FSQ) lead to substantially more efficient training than the naïve binning tokenization used in prior VLAs. Overall, we find that FAST leads to more effective policy training than FSQ, particularly on dexterous real-robot tasks. Our universal tokenizer, FAST+, matches the performance of dataset-specific tokenizers. We report mean and 95% CI.

图 6：不同词元化下的策略性能。压缩动作目标的 FAST、FSQ 比朴素分箱训练效率显著更高。整体上 FAST 优于 FSQ，尤其在真实机器人灵巧任务上。通用 FAST+ 与数据集专用词元化器性能相当。报告均值与 95% 置信区间。

p.7 · S6.SS2.p4

总体而言，以往工作的朴素词元化难以在高频数据上学得有效策略。在本文高频任务清理餐桌（20 Hz）、叠 T 恤（50 Hz）上尤其明显：朴素词元化策略均无法推动任务进展。

p.7 · S6.SS2.p5

相比之下，基于压缩的词元化能有效训练。FAST 与 FSQ 相当，有时更好，特别是灵巧、高频任务；同时它更简单，不需额外训练神经网络。

p.8 · S6.SS2.p6

值得注意的是，FAST 首次成功在 DROID [38] 上训练出强通用策略，无需微调，仅以自然语言提示便可在未见环境零样本评估。包括 DROID 原论文 [38] 与 OpenVLA [39] 在内的既有工作，都未展示零样本结果，而仅进行共训练或微调评估。我们在三所大学校区的环境中测试多种桌面任务（图 7），展示通用性。无需适

配，策略即可跨场景与视角完成抓放、开关柜门、开水龙头等简单操作。失败回合也呈现合理行为，例如接近微波炉或洗碗机门把手，虽然最终未能拉开。网站提供成功与失败视频。策略远非完美，但其通用与稳健水平显著超过以往 DROID 策略。

p.8 · S6.F7

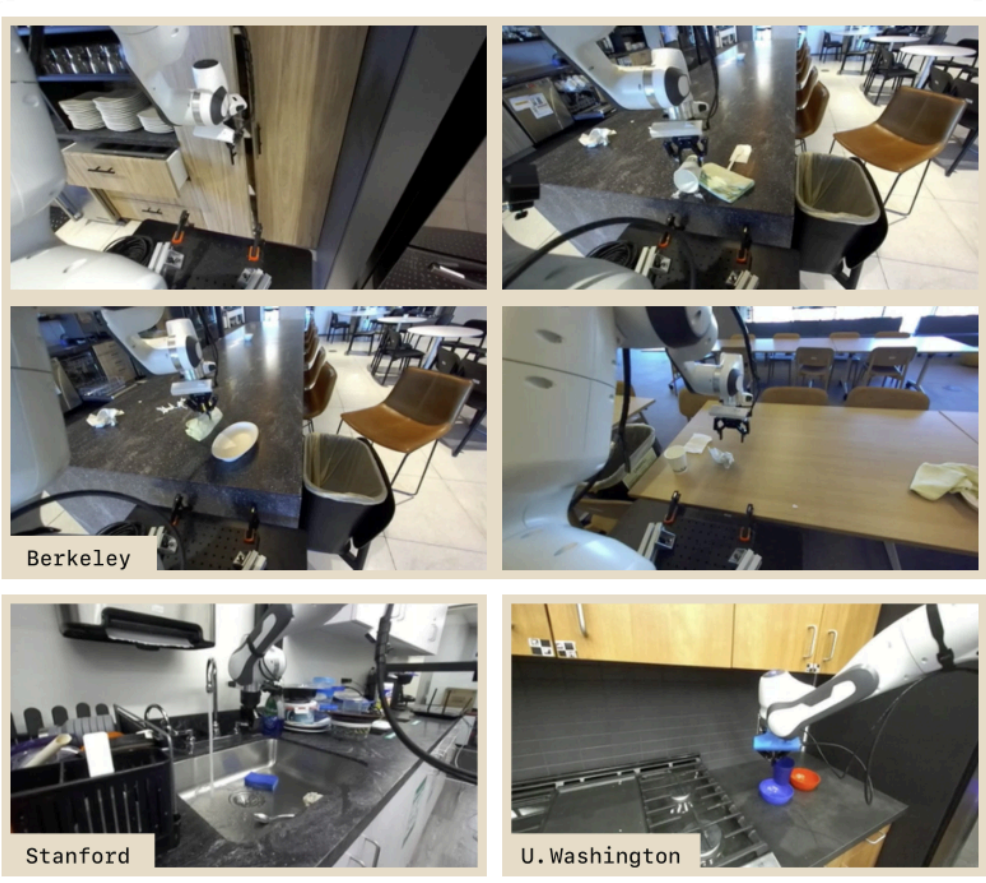


Fig. 7 : Evaluation environments of FAST policy trained on DROID [38] . We find that the same policy checkpoint generalizes robustly, and performs various simple table-top tasks zero-shot across three university campuses.

图 7：DROID [38] 训练的 FAST 策略评估环境。同一个策略检查点可稳健泛化，在三所大学校区零样本执行不同的简单桌面任务。

VI-C 通用动作词元化器

p.8 · S6.SS3.p1

本节评估在 100 万真实动作序列上训练的 FAST+（V-C）。为检验通用性，汇集覆盖广泛机器人形态、动作空间与控制频率的小型测试数据集（图 8，完整列表见表 III）。这些数据均不在词元化器训练集中，因此测试的是直接应用于全新机器人设置、无需重新学习词元化的情形。FAST+ 在广泛数据上均取得良好压缩，所有数据集的词元数至少约减少 2 倍，部分更多。

p.8 · S6.F8

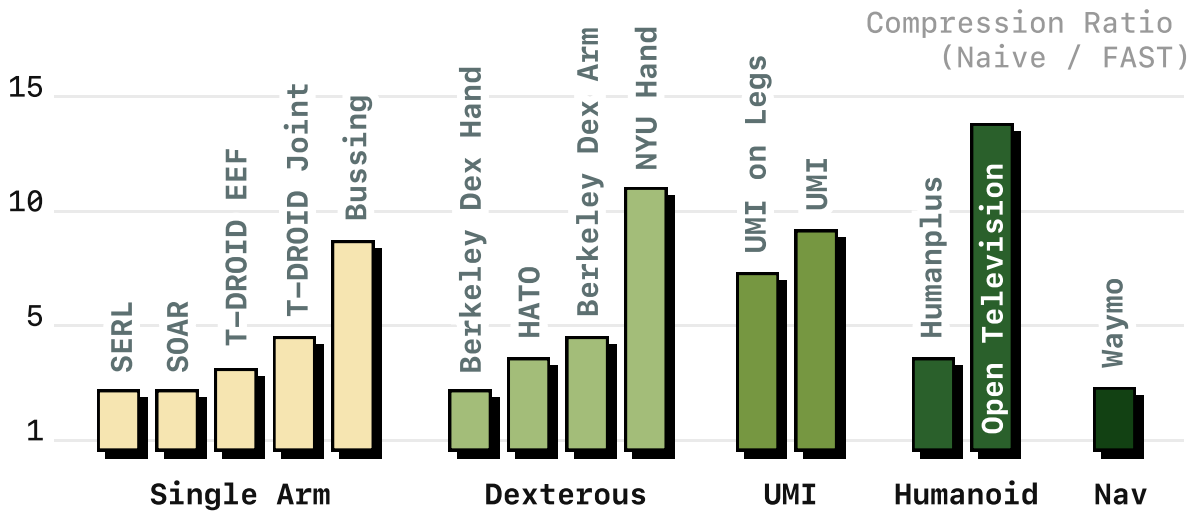


Fig. 8 : Universal tokenizer. We test the compression rate achieved by our FAST+ tokenizer vs. naïve tokenization across diverse robot datasets, unseen during tokenizer training. We find that FAST is effective across a wide range of robot morphologies, action spaces and control frequencies.

图 8：通用词元化器。比较 FAST+ 与朴素词元化在训练未见的多种机器人数据集上的压缩率；它可有效覆盖不同形态、动作空间与控制频率。

p.9 · S6.SS3.p2

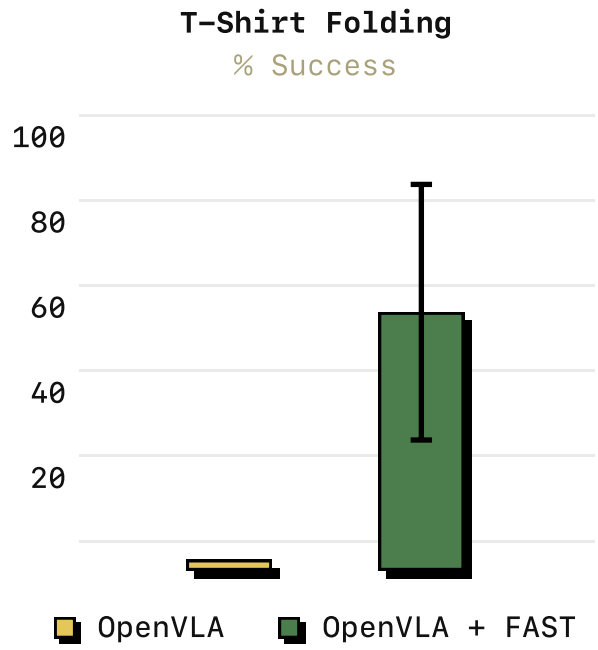
我们还检验通用词元化器用于策略训练的性能，与逐数据集词元化器一起报告于图 6。所有任务中，FAST+ 均紧贴专用 FAST 的性能，说明它可作为动作词元化的强默认选择。

VI-D 消融研究

p.9 · S6.SS4.p1

分析两项关键问题：（1）FAST 是否独立于底层 VLA 骨干？（2）作为流程唯一学习组件的 BPE 压缩步骤有多重要?

p.9 · S6.SS4.fig1

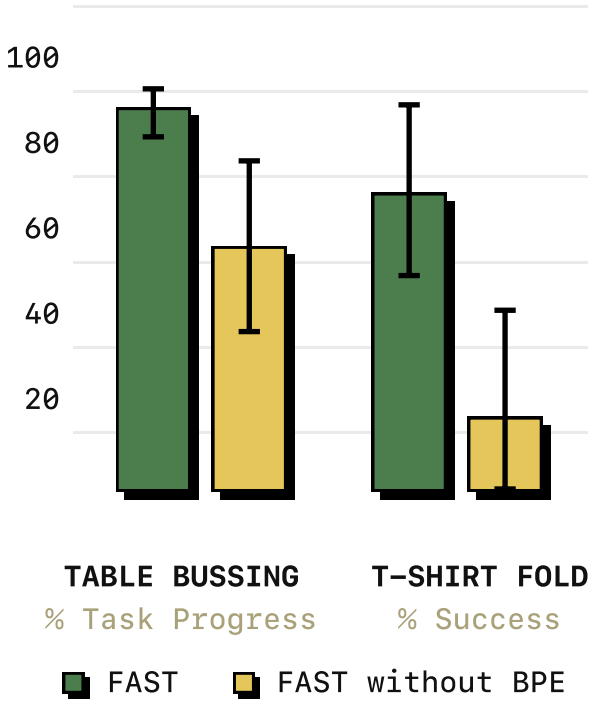


无编号消融图：OpenVLA 在叠 T 恤任务上的朴素词元化与 FAST+ 比较；读图对应下段说明，原图未单独提供图注。

p.9 · S6.SS4.p2

针对第一个问题，在高频叠 T 恤数据上训练 OpenVLA [39]，比较原有朴素词元化与 FAST+。为适配任务，修改 OpenVLA 代码使之接受多张输入图像、预测 1 秒动作块。右图表明，FAST 显著提升 OpenVLA，使其能够有效学习高频操作，说明该词元化不依赖具体骨干，可容易地用于多种预训练自回归 Transformer。

p.9 · S6.SS4.fig2



无编号消融图：清理餐桌与叠 T 恤任务中移除 BPE 的影响；读图对应下段说明，原图未单独提供图注。

p.9 · S6.SS4.p3

其次，在清理餐桌、叠 T 恤任务移除 BPE。右图中，无 BPE 的策略回放更差，但仍优于朴素词元化。直观上，DCT 已把多数信息集中到少量词元，改善学习信号；但没有 BPE 时仍有大量重复零词元，稀释学习信号。模型还需逐个预测数百词元，显著拖慢推理，最终使策略性能下降。

VI-E FAST 与扩散的比较

p.9 · S6.SS5.p1

本节比较先进扩散 VLA π_0 与结合 FAST、自回归解码的 π_0，任务同 VI-B。

p.9 · S6.SS5.p2

图 9 显示，小数据集（Libero、叠 T 恤；$<$ 50 小时）上两种 VLA 相当；在清理餐桌等大数据集上，FAST VLA 的训练步数仅需扩散 π_0 的约三分之一即可达高性能。FAST 自回归 π_0 也更好地遵循语言：DROID 评估中，扩散 π_0 常忽略指令，因此得分较低。对两类 VLA 的语言遵循能力的详细研究留待未来。

p.9 · S6.F9

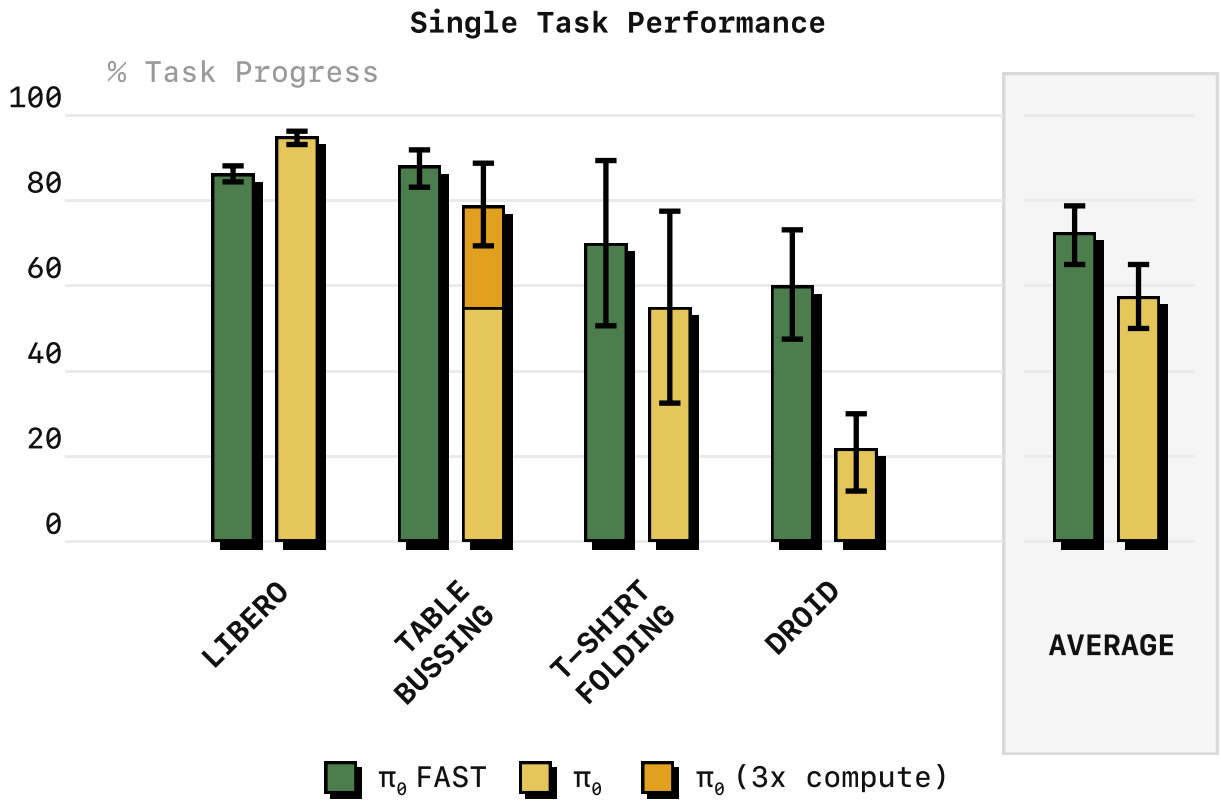


Fig. 9 : Comparison of diffusion π_0 [7] to our π_0 model with FAST decoding on single-task training. On small datasets (Libero, T-Shirt Folding), both perform comparably. On large datasets (Table Bussing), FAST converges faster. In DROID, we find that FAST follows language instructions better. We report mean and 95% CI.

图 9：单任务训练中，扩散 π_0 [7] 与 FAST 解码 π_0 的比较。小数据集（Libero、叠 T 恤）上两者相当；大数据集（清理餐桌）上 FAST 收敛更快。在 DROID 中 FAST 更好地遵循语言指令。报告均值与 95% 置信区间。

p.9 · S6.SS5.p3

自回归 VLA 当前的限制是推理速度：扩散 π_0 通常在 NVIDIA 4090 上用不到 100 ms 预测 1 秒动作块；FAST 词元化的 π_0 每块约需 750 ms，因为需更多解码步骤（通常 30–60 个动作词元，而扩散 π_0 只有 10 步），还需用完整 20 亿参数语言骨干逐词元解码，而扩散 π_0 使用 3 亿参数“动作专家”。较慢推理未损害本文静态操作任务的性能，但明显延长评估时间。未来可采用语言模型领域的推测解码、量化、定制内核等加速方法，本文不作展开。

VI-F 扩展到大型机器人数据集

p.10 · S6.SS6.p1

FAST 对单独数据集有效，但能否扩展至灵巧通用策略？我们在 π_0 [7] 使用的跨形态混合数据上训练前节 π_0-FAST。这是当时最大的灵巧机器人操作数据，包含自有数据的 9.03 亿时间步；混合数据中另有 9.1% 来自 BRIDGE v2 [60]、DROID [38]、OXE [52] 开源数据集。

p.10 · S6.SS6.p2

图 11 在 Black 等 [7] 的任务上比较与扩散 π_0 的零样本性能。总体上，自回归 π_0-FAST 与扩散 π_0 相当，包括最难的衣篮折叠，而训练计算显著更少。图 10 展示 π_0-FAST 折叠衣物的定性示例，更多视频见网站。

p.10 · S6.F10

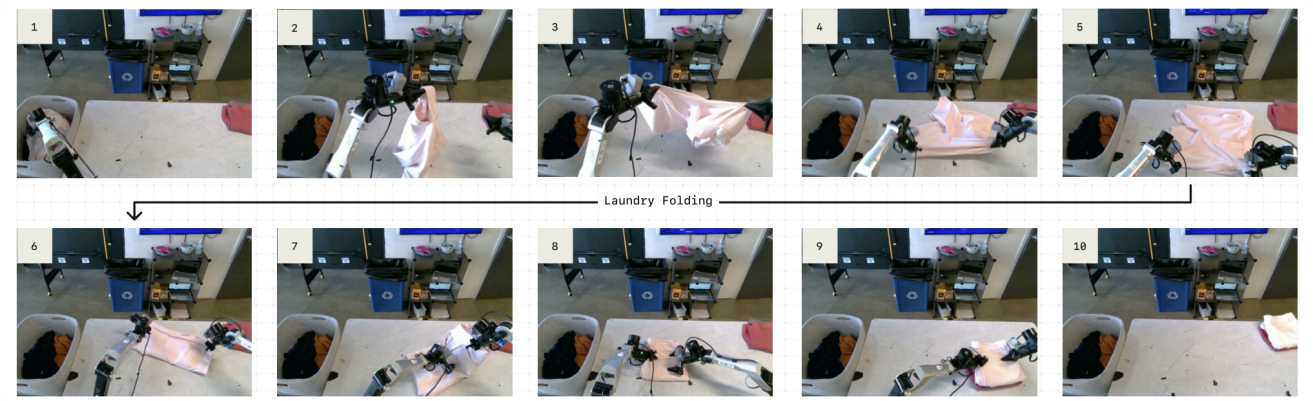


Fig. 10 : Rollout of π_0 -FAST on the laundry folding task. FAST tokenization enables autoregressive VLAs to perform complex, long-horizon, and dexterous tasks that were impossible with previous tokenization schemes.

图 10：π_0-FAST 在衣篮折叠任务中的回放。FAST 使自回归 VLA 能完成以往词元化无法实现的复杂、长时程灵巧任务。

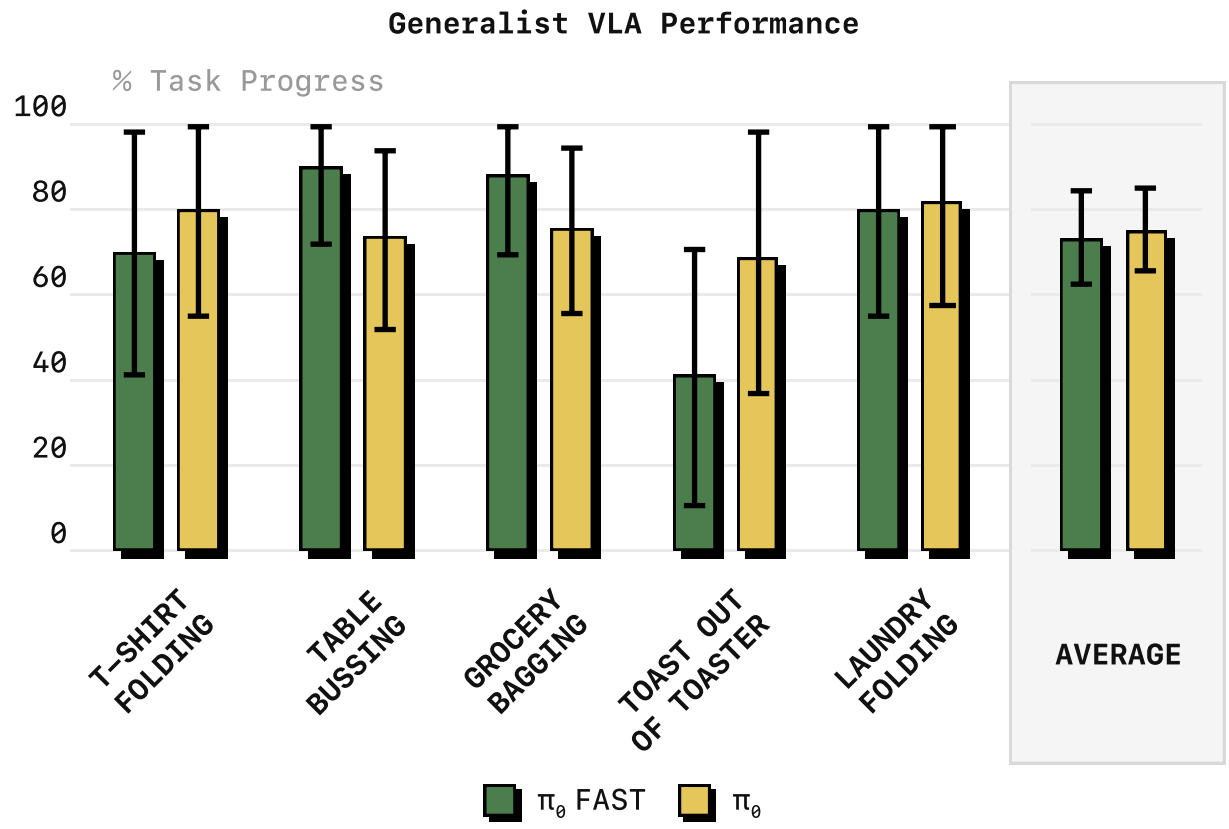


Fig. 11 : Comparison of π_0 -FAST and diffusion π_0 [7] generalist policies. π_0 -FAST matches the performance of diffusion π_0 while requiring significantly less compute for training. Reported: mean and 95% CI.

图 11：π_0-FAST 与扩散 π_0 [7] 通用策略比较。π_0-FAST 用明显更少训练算力匹敌扩散 π_0。报告均值与 95% 置信区间。

重要的是，π_0-FAST 收敛显著快于扩散 π_0：上述模型所需 GPU 小时是 Black 等 [7] 的 π_0 的五分之一。图 1 展示训练中多个检查点的机器人评估，取清理餐桌与叠 T 恤两项代表任务平均，清楚表明 π_0-FAST 更快达到高性能。先进 VLA 训练常需数千 GPU 小时，因此节省 5 倍计算意义重大。附录图 15 还给出所有任务上与计算量匹配的 π_0 的比较；结论一致，因收敛更快，π_0-FAST 明显胜过同计算量 π_0。

p.18 · A0.F15

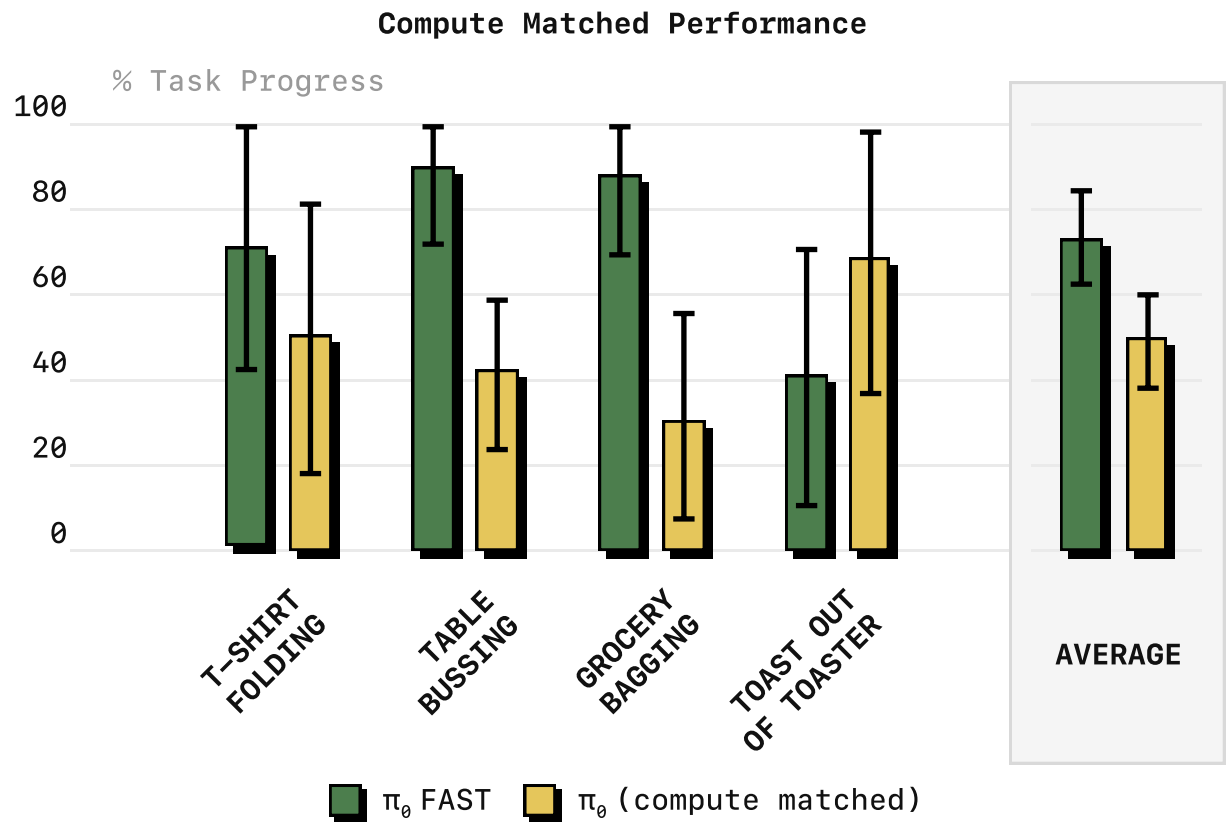


Fig. 15 : Comparison of π_0 -FAST and compute-matched diffusion π_0 [7] generalist policies. π_0 -FAST clearly outperforms the diffusion VLA when trained with the same amount of training compute, due to its faster convergence. Reported: mean and 95% CI.

图 15：π_0-FAST 与计算量匹配的扩散 π_0 [7] 通用策略比较。相同训练计算量下，因收敛更快，π_0-FAST 明显优于扩散 VLA。报告均值与 95% 置信区间。

p.10 · S6.SS6.p4

综上，FAST 可在以往词元化完全失败的复杂灵巧任务上训练自回归 VLA；结合 π_0 等先进 VLA，还能扩展至通用跨形态策略，匹敌最佳扩散 VLA，而训练显著更快。

VII 讨论与未来工作

p.10 · S7.p1

本文提出高频控制数据的高效动作词元化器 FAST：用 DCT 后接 BPE 压缩动作块，在多类机器人领域比现有词元化器压缩更好。真实与仿真 VLA 实验表明，FAST 相比朴素动作离散化大幅提升性能，并优于更复杂的向量量化学习式方法。通用词元化器 FAST+ 可作为任意动作序列的强默认方案。我们据此训练灵巧通用策略 π_0 -FAST，用显著更高训练效率达到先进扩散 VLA 的性能。

p.11 · S7.p2

未来有许多值得研究的方向：

p.11 · S7.p3

动作词元化器。FAST 是通用动作词元化的重要一步，但仍有许多问题。本文实际策略测试限于固定式操作臂；离线实验表明 FAST+ 在移动机器人、灵巧手、类人机器人等形态上有良好压缩潜力，仍需检验这些平台的实际策略。还可研究其他压缩方法，以及把压缩动作编码与扩散 [7] 等非自回归解码结合。

p.11 · S7.p4

VLA 架构。本文初步研究自回归与扩散解码两类 VLA 的权衡，但尚无定论。未来应细致比较训练速度、语言落地能力和表达性。

p.11 · S7.p5

推理速度。π_0-FAST 整体性能匹敌扩散 π_0，但推理更慢（VI-E）。该速度在本文静态任务可接受，未来仍需加速以解决高度动态任务。语言模型已有丰富推理优化，可直接用于自回归 VLA。

致谢

p.11 · Sx1.p1

感谢 Ury Zhilinsky、Kevin Black 协助搭建数据与训练基础设施；Pranav Atreya、Haohuan Wang、Lucy Shi、Arhan Jain、Andy Yun 协助在 UC Berkeley、Stanford、华盛顿大学评估 DROID；Will Chen 测试、调试 FAST+ 开源实现。感谢 Noah Brown、Szymon Jakubczak、Adnan Esmail、Tim Jones、Mohith Mothukuri、James Tanner 维护机器人；Anna Walling 协助机器人、数据与评估运营；Physical Intelligence 全体机器人操作员对采集与评估的大量贡献；Claudio Guglieri、Lachy Groom、Karol Hausman 协助论文及网站可视化。

参考文献（保留英文原书目）

[1] Nasir Ahmed, T_ Natarajan, and Kamisetty R Rao. Discrete cosine transform. *IEEE transactions on Computers*, 100(1):90–93, 1974.

[2] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do as i can and not as i say: Grounding language in robotic affordances. In *arXiv preprint arXiv:2204.01691*, 2022.

[3] Suneel Belkhale and Dorsa Sadigh. Minivla: A better vla with a smaller footprint, 2024. URL https://github.com/Stanford-ILIAD/openvla-mini.

[4] Suneel Belkhale, Tianli Ding, Ted Xiao, Pierre Sermanet, Quon Vuong, Jonathan Tompson, Yevgen Chebotar, Debidatta Dwibedi, and Dorsa Sadigh. Rt-h: Action hierarchies using language, 2024. URL https://arxiv.org/abs/2403.01823.

[5] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

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

- [7] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. *pi_0*: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.
- [8] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-1: Robotics transformer for real-world control at scale. In *arXiv preprint arXiv:2212.06817*, 2022a.
- [9] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022b.

[10] 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, Alex 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, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *arXiv preprint arXiv:2307.15818*, 2023.

[11] Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, Hanbo Zhang, and Minzhao Zhu. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *arXiv preprint arXiv:2410.06158*, 2024.

[12] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. Beats: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058*, 2022.

[13] An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Xueyan Zou, Jan Kautz, Erdem Biyik, Hongxu Yin, Sifei Liu, and Xiaolong Wang. NaVILA: Legged Robot Vision-Language-Action Model for Navigation. *arXiv preprint arXiv:2412.04453*, 2024a.

[14] Xuxin Cheng, Jialong Li, Shiqi Yang, Ge Yang, and Xiaolong Wang. Open-television: Teleoperation with immersive active visual feedback. *arXiv preprint arXiv:2407.01512*, 2024b.

[15] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[16] Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[17] OX-Embodiment Collaboration, A Padalkar, A Pooley, A Jain, A Bewley, A Herzog, A Irpan, A Khazatsky, A Rai, A Singh, et al. Open X-Embodiment: Robotic learning datasets and RT-X models. *arXiv preprint arXiv:2310.08864*, 1(2), 2023.

[18] James W Cooley and John W Tukey. An algorithm for the machine calculation of complex fourier series. *Mathematics of computation*, 19(90):297–301, 1965.

[19] Norman Di Palo and Edward Johns. Keypoint action tokens enable in-context imitation learning in robotics. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[20] Ria Doshi, Homer Walke, Oier Mees, Sudeep Dasari, and Sergey Levine. Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation. In *Conference on Robot Learning*, 2024.

[21] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

- [22] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2020.
- [23] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R. Qi, Yin Zhou, Zoey Yang, Aur'elien Chouard, Pei Sun, Jiquan Ngiam, Vijay Vasudevan, Alexander McCauley, Jonathon Shlens, and Dragomir Anguelov. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9710–9719, October 2021.
- [24] Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu, and Cewu Lu. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 653–660. IEEE, 2024a.
- [25] Kuan Fang, Fangchen Liu, Pieter Abbeel, and Sergey Levine. Moka: Open-world robotic manipulation through mark-based visual prompting. *Robotics: Science and Systems (RSS)*, 2024b.
- [26] Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. Humanplus: Humanoid shadowing and imitation from humans. In *Conference on Robot Learning (CoRL)*, 2024.
- [27] Philip Gage. A new algorithm for data compression. *The C Users Journal*, 12(2):23–38, 1994.
- [28] Dan Gillick, Cliff Brunk, Oriol Vinyals, and Amarnag Subramanya. Multilingual language processing from bytes, 2016. URL https://arxiv.org/abs/1512.00103.
- [29] Yuan Gong, Yu-An Chung, and James Glass. AST: Audio Spectrogram Transformer. In *Proc. Interspeech 2021*, pages 571–575, 2021. doi: 10.21437/Interspeech.2021-698.
- [30] Irmak Guzey, Yinlong Dai, Georgy Savva, Raunaq Bhirangi, and Lerrel Pinto. Bridging the human to robot dexterity gap through object-oriented rewards, 2024. URL https://arxiv.org/abs/2410.23289.
- [31] Huy Ha, Yihuai Gao, Zipeng Fu, Jie Tan, and Shuran Song. UMI on legs: Making manipulation policies mobile with manipulation-centric whole-body controllers. In *Proceedings of the 2024 Conference on Robot Learning*, 2024.
- [32] Wenlong Huang, Chen Wang, Yunzhu Li, Ruohan Zhang, and Li Fei-Fei. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. *arXiv preprint arXiv:2409.01652*, 2024.
- [33] David A. Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101, 1952. doi: 10.1109/JRPROC.1952.273898.
- [34] Huiwon Jang, Sihyun Yu, Jinwoo Shin, Pieter Abbeel, and Younggyo Seo. Efficient long video tokenization via coordinated-based patch reconstruction. *arXiv preprint arXiv:2411.14762*, 2024.
- [35] Zhenyu Jiang, Yuqi Xie, Kevin Lin, Zhenjia Xu, Weikang Wan, Ajay Mandlekar, Linxi Fan, and Yuke Zhu. Dexmimicgen: Automated data generation for bimanual dexterous manipulation via imitation learning. *arXiv preprint arXiv:2410.24185*, 2024.
- [36] Joshua Jones, Oier Mees, Carmelo Sferrazza, Kyle Stachowicz, Pieter Abbeel, and Sergey Levine. Beyond sight: Finetuning generalist robot policies with heterogeneous sensors via language grounding. *arXiv preprint arXiv:2501.04693*, 2025.

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

[38] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan Lin, Jingpei Lu, Jean Mercat, Abdul Rehman, Pannag R Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Z. Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, David Antonio Herrera, Minho Heo, Kyle Hsu, Jiaheng Hu, Donovon Jackson, Charlotte Le, Yunshuang Li, Kevin Lin, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abigail O'Neill, Rosario Scalise, Derick Seale, Victor Son, Stephen Tian, Emi Tran, Andrew E. Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J Lim, Jitendra Malik, Roberto Martín-Martín, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Michael C. Yip, Yuke Zhu, Thomas Kollar, Sergey Levine, and Chelsea Finn. Droid: A large-scale in-the-wild robot manipulation dataset. In *Proceedings of Robotics: Science and Systems*, 2024.

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

[40] Lucy Lai, Ann ZX Huang, and Samuel J Gershman. Action chunking as conditional policy compression.

[41] Seungjae Lee, Yibin Wang, Haritheja Etukuru, H. Jin Kim, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto. Behavior generation with latent actions. *arXiv preprint arXiv:2403.03181*, 2024.

[42] Toru Lin, Yu Zhang, Qiyang Li, Haozhi Qi, Brent Yi, Sergey Levine, and Jitendra Malik. Learning visuotactile skills with two multifingered hands. *arXiv:2404.16823*, 2024.

[43] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[44] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[45] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[46] Jianlan Luo, Zheyuan Hu, Charles Xu, You Liang Tan, Jacob Berg, Archit Sharma, Stefan Schaal, Chelsea Finn, Abhishek Gupta, and Sergey Levine. Serl: A software suite for sample-efficient robotic reinforcement learning, 2024.

- [47] Ajay Mandlekar, Yuke Zhu, Animesh Garg, Jonathan Booher, Max Spero, Albert Tung, Julian Gao, John Emmons, Anchit Gupta, Emre Orbay, et al. Roboturk: A crowdsourcing platform for robotic skill learning through imitation. In *Conference on Robot Learning*, pages 879–893. PMLR, 2018.

- [48] Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple, 2023. URL https://arxiv.org/abs/2309.15505.

- [49] Atharva Mete, Haotian Xue, Albert Wilcox, Yongxin Chen, and Animesh Garg. Quest: Self-supervised skill abstractions for learning continuous control, 2024. URL https://arxiv.org/abs/2407.15840.

- [50] Soroush Nasiriany, Fei Xia, Wenhao Yu, Ted Xiao, Jacky Liang, Ishita Dasgupta, Annie Xie, Danny Driess, Ayzaan Wahid, Zhuo Xu, et al. Pivot: Iterative visual prompting elicits actionable knowledge for vlms. In *Forty-first International Conference on Machine Learning*, 2024.

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

[52] Open X-Embodiment Collaboration, Abhishek Padalkar, Acorn Pooley, Ajinkya Jain, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anikait Singh, Anthony Brohan, Antonin Raffin, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Brian Ichter, Cewu Lu, Charles Xu, Chelsea Finn, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Chuer Pan, Chuyuan Fu, Coline Devin, Danny Driess, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Federico Ceola, Fei Xia, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Giulio Schiavi, Hao Su, Hao-Shu Fang, Haochen Shi, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homer Walke, Hongjie Fang, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jaehyung Kim, Jan Schneider, Jasmine Hsu, Jeannette Bohg, Jeffrey Bingham, Jiajun Wu, Jialin Wu, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jitendra Malik, Jonathan Tompson, Jonathan Yang, Joseph J. Lim, João Silvério, Junhyek Han, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Zhang, Keyvan Majd, Krishan Rana, Krishnan Srinivasan, Lawrence Yunliang Chen, Lerrel Pinto, Liam Tan, Lionel Ott, Lisa Lee, Masayoshi Tomizuka, Maximilian Du, Michael Ahn, Mingtong Zhang, Mingyu Ding, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J Joshi, Niko Suenderhauf, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Pannag R Sanketi, Paul Wohlhart, Peng Xu, Pierre Sermanet, Priya Sundaresan, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Martín-Martín, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Sherry Moore, Shikhar Bahl, Shivin Dass, Shuran Song, Sichun Xu, Siddhant Haldar, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Sudeep Dasari, Suneel Belkhale, Takayuki Osa, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Vidhi Jain, Vincent Vanhoucke, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiaolong Wang, Xinghao Zhu, Xuanlin Li, Yao Lu, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yueh hua Wu, Yujin Tang, Yuke Zhu, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zhuo Xu, and Zichen Jeff Cui. Open X-Embodiment: Robotic learning datasets and RT-X models. https://arxiv.org/abs/2310.08864, 2023.

[53] Artidoro Pagnoni, Ram Pasunuru, Pedro Rodriguez, John Nguyen, Benjamin Muller, Margaret Li, Chunting Zhou, Lili Yu, Jason Weston, Luke Zettlemoyer, Gargi Ghosh, Mike Lewis, Ari Holtzman†, and Srinivasan Iyer. Byte latent transformer: Patches scale better than tokens. 2024. URL https://github.com/facebookresearch/blt.

[54] Haozhi Qi, Ashish Kumar, Roberto Calandra, Yi Ma, and Jitendra Malik. In-hand object rotation via rapid motor adaptation, 2022. URL https://arxiv.org/abs/2210.04887.

[55] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[56] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gómez Colmenarejo, Alexander Novikov, Gabriel Barth-maron, Mai Giménez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *Transactions on Machine Learning Research*, 2022.

[57] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

- [58] Himanshu Gaurav Singh, Antonio Loquercio, Carmelo Sferrazza, Jane Wu, Haozhi Qi, Pieter Abbeel, and Jitendra Malik. Hand-object interaction pretraining from videos, 2024. URL https://arxiv.org/abs/2409.08273.
- [59] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning, 2018. URL https://arxiv.org/abs/1711.00937.
- [60] Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. BridgeData v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pages 1723–1736. PMLR, 2023.
- [61] Gregory K Wallace. The jpeg still picture compression standard. *IEEE transactions on consumer electronics*, 38(1):xviii–xxxiv, 1992.
- [62] Lirui Wang, Xinlei Chen, Jialiang Zhao, and Kaiming He. Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [63] Junjie Wen, Yichen Zhu, Jinming Li, Minjie Zhu, Kun Wu, Zhiyuan Xu, Ning Liu, Ran Cheng, Chaomin Shen, Yaxin Peng, Feifei Feng, and Jian Tang. Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation. *arXiv preprint arXiv:2409.12514*, 2024.
- [64] Wilson Yan, Matei Zaharia, Volodymyr Mnih, Pieter Abbeel, Aleksandra Faust, and Hao Liu. Elastictok: Adaptive tokenization for image and video. *arXiv preprint arXiv:2410.08368*, 2024.
- [65] Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, et al. Latent action pretraining from videos. *arXiv preprint arXiv:2410.11758*, 2024.
- [66] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G. Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, and Lu Jiang. Magvit: Masked generative video transformer, 2023. URL https://arxiv.org/abs/2212.05199.
- [67] Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. Robotic control via embodied chain-of-thought reasoning. In *Conference on Robot Learning*, 2024.
- [68] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec, 2021. URL https://arxiv.org/abs/2107.03312.
- [69] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.
- [70] Tony Z Zhao, Jonathan Tompson, Danny Driess, Pete Florence, Kamyar Ghasemipour, Chelsea Finn, and Ayzaan Wahid. Aloha unleashed: A simple recipe for robot dexterity. *arXiv preprint arXiv:2410.13126*, 2024.
- [71] Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3d-vla: A 3d vision-language-action generative world model. *arXiv preprint arXiv:2403.09631*, 2024a.
- [72] Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3d-vla: 3d vision-language-action generative world model. *arXiv preprint arXiv:2403.09631*, 2024b.

[73] Ruijie Zheng, Yongyuan Liang, Shuaiyi Huang, Jianfeng Gao, Hal Daumé III, Andrey Kolobov, Furong Huang, and Jianwei Yang. Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies. *arXiv preprint arXiv:2412.10345*, 2024.

- [74] Zhiyuan Zhou, Pranav Atreya, Abraham Lee, Homer Walke, Oier Mees, and Sergey Levine. Autonomous improvement of instruction following skills via foundation models. In *Conference on Robot Learning*, 2024.
- [75] Jacob Ziv and Abraham Lempel. Compression of individual sequences via variable-rate coding. *IEEE transactions on Information Theory*, 24(5):530–536, 1978.

附录 A 通用词元化器的训练混合数据

p.16 · A0.SS1.p1

通用词元化器训练混合主要来自 VI-F 描述的 π_0 [7] 数据。许多数据集同时包含关节空间、末端世界坐标系、末端相机坐标系等动作参数化版本，以保证通用性。Open X-Embodiment [52]、DROID [38]、Bridge V2 [60] 按原始形式纳入。编码前，各动作补零至 32 维，以容纳不同维数动作空间。

p.16 · A0.SS1.tab1

数据集名称	机器人形态	动作空间	Control Frequency (Hz)	Mixture Weight (%)
ARX	双臂	关节	50	7.2
AgileX	双臂	关节	50	1.8
Fibocom	移动	关节	50	2.9
Franka FR3	单臂	关节	20	3.7
Mobile Trossen	移动	关节	50	2.5
Trossen Biarm	双臂	关节	50	4.3
UR5 single	单臂	关节	20	10.3
UR5 biarm	双臂	关节	20	2.4
ARX slate mobile	移动	关节	50	2.5
ARX EE	双臂	末端执行器	50	3.6
AgileX EE	双臂	末端执行器	50	0.9
Fibocom EE	移动	末端执行器	50	1.4
Franka FR3 EE	单臂	末端执行器	20	1.9
Mobile Trossen EE	移动	末端执行器	50	1.2
Trossen Biarm EE	双臂	末端执行器	50	2.1
UR5 single EE	单臂	末端执行器	20	5.2
UR5 biarm EE	双臂	末端执行器	20	1.2
ARX slate mobile EE	移动	末端执行器	50	1.2
ARX Cam	双臂	相机坐标系	50	3.6
AgileX Cam	双臂	相机坐标系	50	0.9
Fibocom Cam	移动	相机坐标系	50	1.4
Franka FR3 Cam	单臂	相机坐标系	20	1.9
Mobile Trossen Cam	移动	相机坐标系	50	1.2
Trossen Biarm Cam	双臂	相机坐标系	50	2.1
UR5 single Cam	单臂	相机坐标系	20	5.2
UR5 biarm Cam	双臂	相机坐标系	20	1.2
ARX slate mobile Cam	移动	相机坐标系	50	1.2
ALOHA *[69]*	双臂	关节	50	5.0

DROID *[38]*	单臂	关节	15	11.2
Bridge V2 *[60]*	单臂	末端执行器	5	5.0
OpenX *[52]*	单臂	末端执行器	混合	3.8

附录训练数据表：列出数据集、机器人形态、动作空间与控制频率。保留原表所有数据与英文名称；Joint 为关节空间，EEF 为末端执行器，World/Camera 分别为世界/相机坐标系，Mobile 为移动底盘。

附录 B 压缩与重建的权衡

p.16 · A0.F12

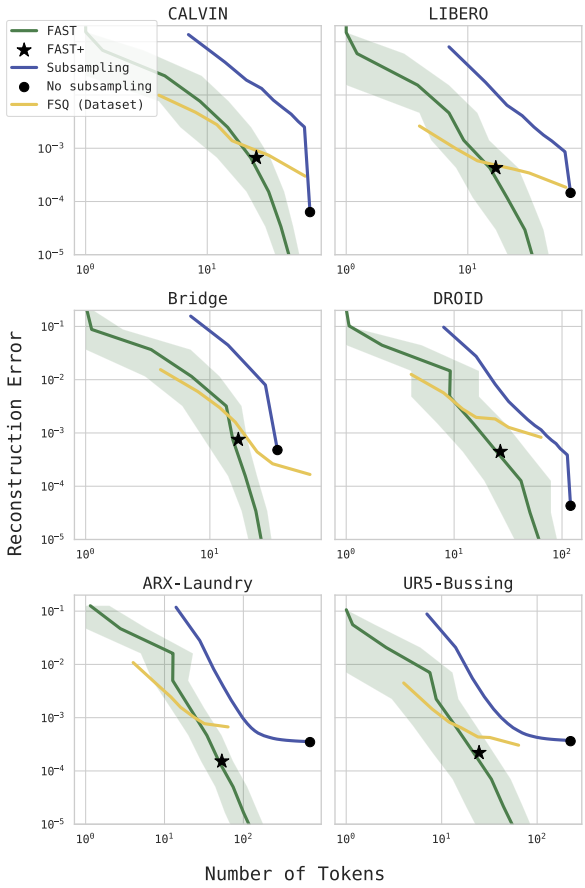


Fig. 12 : Comparison of compression-reconstruction tradeoff on six training datsets. Any discretization method includes some hyperparameter that controls the tradeoff between reconstruction fidelity and compression level, represented here as number of tokens in the output (vocab size is held constant across all tokenizers). We sweep this hyperparameter (FAST: rounding scale; naïve tokenization: subsampling frequency; FSQ: number of latent tokens) and find that FAST performs well across a wide range of scales. In particular, although it is less efficient than VQ-based tokenizers at low fidelities, it exhibits much better scaling to higher reconstruction fidelity, making FAST much more applicable to fine-grained control problems. Specific instantiations of each tokenizer (FAST+, and naïve tokenization without subsampling) are also shown.

图 12：六个训练数据集的压缩-重建权衡。各离散化方法均有控制保真与压缩程度的超参数，这里用输出词元数衡量压缩，词表大小在各方法间固定。扫描 FAST 的取整缩放量、朴素方案的下采样频率、FSQ 的潜在词元数，发现 FAST 在广泛尺度表现良好。低保真时它不如 VQ 词元化器高效，但向高重建保真扩展明显更好，更适合细粒度控制。同时显示 FAST+ 和不下采样朴素方案的具体实例。

附录 C 策略训练

p.16 · A0.SS3.p1

采用 π_0 [7]、OpenVLA [39] 骨干。按任务输入两或三张 224×224 图像：一个第三人称相机，以及每臂一个腕部相机。预训练视觉编码器分别编码图像，再拼接词元。另以自然语言指令、本体感知状态作为条件；二者都视作字符串，用 LLM 语言词元化器处理。本体状态先按类似 RT-2 动作编码 [10] 的方式分成 256 个箱，再将整数编码为文本输入序列。本体状态是策略输入而非输出，因此简单分箱已足够；动作输出才需要本文实验所示的先进词元化。

p.16 · A0.SS3.p2

所有策略先线性预热学习率 1000 步，再保持 5×10^{-5}。采用 AdamW [45]（$b1 = 0.9$，$b2 = 0.95$），无权重衰减，梯度范数裁剪至 1，以 0.999 权重计算网络参数 EMA。

p.16 · A0.SS3.p3

推理用贪心自回归解码，但双臂任务（叠 T 恤、取吐司、衣篮折叠）例外：较小温度 $\beta = 0.7$ 有助于策略离开初始位置，因为训练回合开头部分数据含有在初始位置悬停的静止动作块。

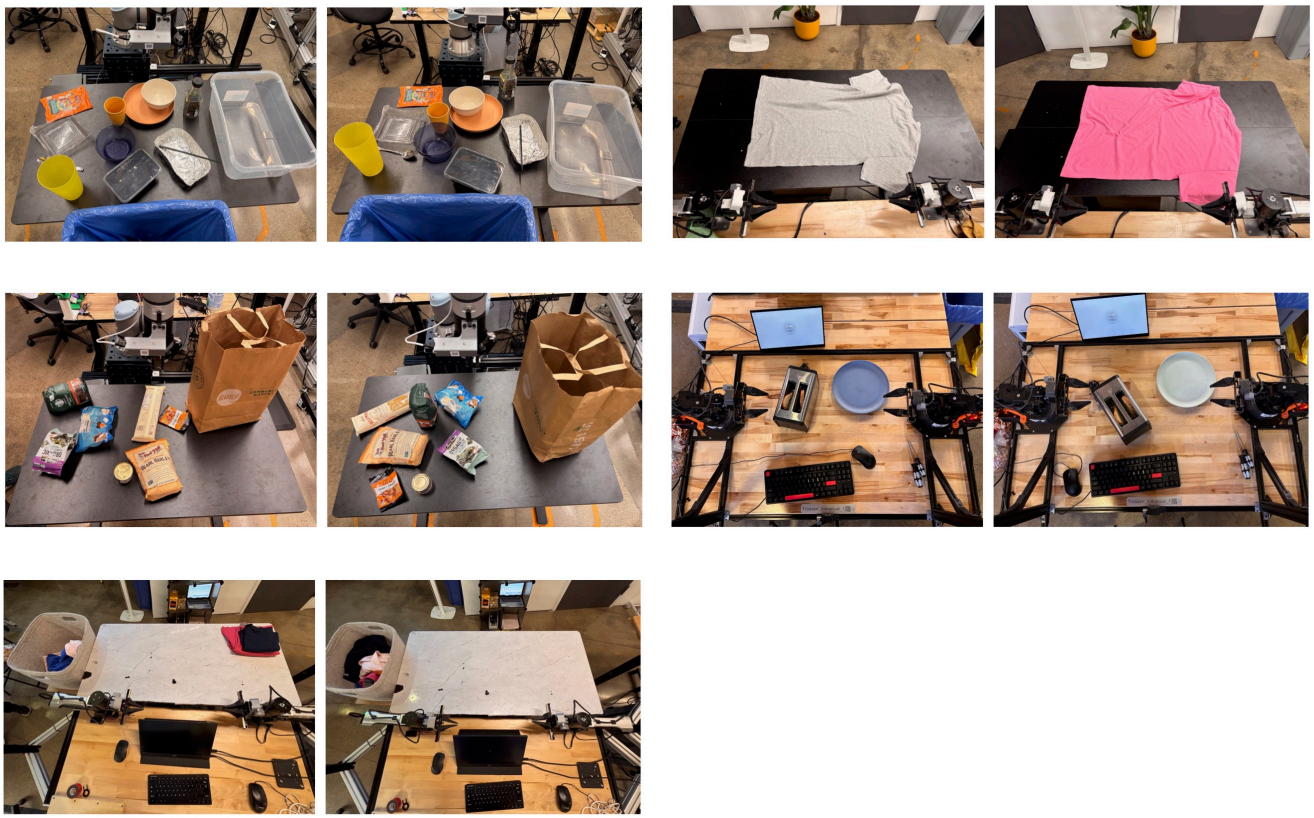
附录 D DROID 策略设置

p.17 · A0.SS4.p1

为便于复现，补充 DROID 训练细节。策略条件包括单个第三人称视图与腕部视图；每回合有两个外部视图，训练时随机选一。每回合还有三条语言标注，训练时同样随机选择。不使用相机标定信息，因此可直接测试新视角，无需标定。动作空间为关节速度与绝对夹爪位置，预测 15 步动作块；推理时开环执行 8 或 15 步。仅做轻量筛选：只用标为成功的 7.5 万回合，并过滤全零动作的空闲时间步，后者常是遥操员复位 VR 控制器的时段。除此之外完整数据训练效果良好，精细筛选仍可能提升性能。训练 3 个 epoch（batch 256，24 万次迭代），本文 30 亿参数 VLA 在 8 块 H100 上约需 4 天。

附录 E 评估任务与训练数据集

p.17 · A0.F13



(a) Table Bussing

图 13：评估任务场景：（a）清理餐桌；（b）叠 T 恤；（c）杂货装袋；（d）从吐司机取吐司；（e）衣篮折叠。原 HTML 将多子图附在同一图对象内，本册保留所有子图。

p.17 · A0.SS5.p1

以下描述所有评估任务、训练数据，详列初始条件分布与评分标准。

p.17 · A0.SS5.p2

Libero。遵循 Liu 等 [43] 的训练评估设置，在 Spatial、Object、Goal、Long 套件测试，并用作者提供的对应数据训练。合并为含 27 万样本的数据集，联合训练一个策略，以减少所需策略数量。各策略训练 4 万次迭代（≈ 40 个 epoch），采用 Kim 等 [39] 重渲染的数据集。逐回合二值判定成功。

p.17 · A0.SS5.p3

清理餐桌。单臂 UR5e 把垃圾、盘子和餐具分类放到垃圾桶或收纳箱。训练示范在随机场景中使用约 70 种物体。图 13(a) 的测试场景在桌面新布局放置 12 个物体，故意让餐具压在垃圾上、物体相互遮挡，并加入筷子、透明塑料、反光容器等难物体。总分为正确丢弃或收纳的物体比例。

p.17 · A0.SS5.p4

叠 T 恤。ARX 双臂折叠 T 恤。训练包含约 150 件不同尺寸、颜色、款式的衣物示范。图 13(b) 测试循环使用 5 件已见、颜色尺寸不同的 T 恤，均从平铺开始；人工评判成功折叠比例。

p.18 · A0.SS5.p5

杂货装袋。UR5e 单臂将杂货装袋。直接评估以 Black 等 [7] 完整混合数据预训练的模型。图 13(c) 含 7 件形状、尺寸、材质、重量不同的物体与一个大纸袋；按成功入袋比例评分。

p.18 · A0.SS5.p6

取吐司。与 ALOHA [70] 配置相仿的 Trossen ViperX 双臂从吐司机取出两片吐司，放在盘上。直接评估以 Black 等 [7] 完整混合数据预训练的模型。场景见图 13(d)。每取出一片记 1 分，放到盘上再记 1 分，总分 4 分。

p.18 · A0.SS5.p7

衣篮折叠。ARX 双臂从篮中取出短裤或 T 恤并折叠；缠结衣物需多次展开与铺平，任务很难。遵循 Black 等 [7]，先用完整 π_0 混合数据预训练，再用少量高质量任务专用数据微调。图 13(e) 的衣篮随机放 5 件衣物；人工判定成功折叠并叠放的比例。

p.18 · A0.SS5.p8

DROID。在全部成功回合上训练（7.5 万回合、2100 万样本），共 24 万次迭代（原文写作 ≈ 3 episodes，结合附录 D 应理解为约 3 个 epoch；保留该原文差异），轻量筛选见附录 D。训练后在新背景、新相机角度、新物体场景零样本部署。定量评估含 16 项任务，每策略共 44 次试验（表 II），按任务进展评分，例如拿对物体 1 分、放到正确容器再 1 分。图 14 展示定量场景。此外在三所大学校区做定性测试（图 7），不统计成功率，而提供大量成功与失败视频展示能力。

p.18 · A0.F14



Fig. 14 : Setups used for quantitative DROID evaluation.

图 14：用于 DROID 定量评估的场景。

p.18 · A0.T2

任务	试验数
把勺子放入沥水架	4
把胡萝卜放入碗	4
把盘子放入沥水架	2
擦桌子	2
把盘子放在桌上	2
清理桌面	2
关抽屉	4
把订书机放在笔记本上	2
把订书机放入抽屉	4

擦白板	2
把记号笔放入杯子	4
把黑海绵放入蓝碗	2
把红瓶放入黑碗	2
把西瓜放入紫碗	2
把西瓜从紫碗移入蓝碗	2
把胶带放入紫碗	2
把水瓶放在桌面左侧	2
总计	44

TABLE II : DROID evaluation tasks.

表 II：DROID 评估任务。列为任务与试验数；任务指令逐行提供中文对照，原有试验数完整保留。

p.19 · A0.T3

机器人形态	数据集名称	平台	动作空间	动作维数	控制频率	任务
单臂	SOAR *[74]*	WidowX	EEF	7	5	抓取/放置
	DROID-Eval EEF *[38]*	Franka	EEF	7	15	抓取/放置
	DROID-Eval Joint *[38]*	Franka	关节	8	15	抓取/放置
	SERL *[46]*	Franka	EEF	7	10	插入
	π Table Bussing *[7]*	UR5	关节	8	20	抓取/放置
灵巧手	NYU DexHand *[30]*	ALLEGRO	Joint+EEF	30	16	灵巧操作
	Berkeley DexHand *[54]*	ALLEGRO	关节	16	20	手内操作
	Berkeley DexArm *[58]*	xArm+ALLEGRO	关节	23	20	灵巧抓放
	HATO *[42]*	UR5+Psyonic Hand	EEF+Joint	24	10	灵巧抓放
UMI	UMI *[16]*	UMI	EEF	7	20	抓取/放置
	UMI on Legs *[31]*	UMI	EEF	7	20	全身操作
类人	HumanPlus *[26]*	Unitree H1	关节	40	50	全身操作
	UCSD TeleVision *[14]*	Unitree H1 w/Neck	关节	28	60	操作+主动感知
导航	Waymo *[23]*	Waymo Car	二维增量	2	10	自动驾驶

TABLE III : Universal Tokenizer Evaluation Datasets.

表 III：通用词元化器评估数据集。列为数据集、机器人形态、动作空间、控制频率；专名保留英文。

读者校注与工程映射（本册原创）

先区分三个不同的“快”

FAST 的核心收益是改善动作目标的统计表示与训练收敛，不是把自回归 VLA 变成最快的动作生成器。原文 VI-E 的 RTX 4090 对照中，FAST 每个 1 秒动作块约 750 ms，扩散 π_0 约 100 ms；VI-F 的“最多 5 倍”针对训练 GPU 小时。它们不是同一计时指标。

DCT 量化误差来自哪里

令 $A \in \mathbb{R}^{H \times d}$ 为归一化动作块，C 为正交 DCT 矩阵，系数 $X=CA$。FAST 先计算整数 $Q=\mathrm{round}(\gamma X)$，再按低频优先展平，并用 BPE 无损压缩整数序列。恢复为 $\hat{A}=C^{T}(Q/\gamma)$。在不额外截断、不裁剪、使用正交归一化 DCT 的条件下，每个系数量化误差绝对值至多 $1/(2\gamma)$，故 $\|\hat{A}-A\|_F=\|Q/\gamma-X\|_F \le \sqrt{(Hd)}/(2\gamma)$。这给出可检查的量化上界；实际机器人任务成功率仍取决于控制敏感度和闭环状态。BPE 只压缩整数符号，不再新增数值误差。

本文参数 $\gamma=10$ 是实验选择，不是任意动作单位下的精度保证。改变归一化分位数、DCT 的 norm 约定或夹爪通道处理，会改变误差的物理意义。工程复现应一起保存归一化参数、γ、BPE 词典、动作维数和块长度。

低交叉熵为什么可能掩盖坏动作

自回归训练最小化逐词元负对数似然。当高频平滑信号被逐时间步分箱，前一词元往往已经足够预测后一词元；训练可通过复制取得低损失，而未学好条件图像与动作整体形状的关系。DCT 集中能量、BPE 移除重复零模式后，每个预测更有信息。原文的证据是图 3 采样率扫描和图 6 控制评估；不能只看训练 loss 就判断是否改善。

表 I 可复核的压缩尺度

数据	朴素词元数	FAST 词元数	原表压缩比
BridgeV2	35	20	1.75
DROID	105	29	3.6
清理餐桌	140	28	5.0
叠 T 恤	700	53	13.2

这些是一秒动作块、近似重建保真下的平均数。不能把词元数下降 13.2 倍直接换成端到端推理加速 13.2 倍：视觉编码、语言骨干、自回归缓存和控制调度仍在路径中。

与本机四条实证线如何接通

S3（V:/lerobot-flow-matching-v2/DELIVERY_REPORT.md）：4,200 条配对闭环回合显示，单步动作头很快，接触恢复仍失败。教师首动作误差在第 32 tick 从专家观测的约 17.8 mm 增至自身闭环的 234.3 mm。对 FAST 的直接启示是：动作块重建精度与接触维持是两个门禁。应先检查失接触状态是否在数据支持内，再比较压缩策略，不能把更低词元交叉熵视作恢复控制已学会。

S7（V:/rectified-flow-distill-bench/DELIVERY_REPORT.md）：同容量 MicroDiT 的 1 NFE Shortcut Feature SWD 0.1819，普通 1-RF 为 0.3287；这是光学图像生成对照，不是 FAST 的机器人结果。它与 FAST 的可比问题是“降低采样成本还是重塑训练目标”；连续场采用步长条件，FAST 则重塑离散预测序列。指标、数据集与网络不能混报。

S8（V:/event-neural-flow/DELIVERY_REPORT.md）：设备内原始事件到体素再到光流的 CUDA Graph 路径为 1.213 ms、824.5 FPS，排除采集、磁盘、主机上传。若尝试用频域压缩光流序列，应额外评估涡量等导数；小的速度重建误差不必然带来小的空间导数误差。该应用是本册工程建议，不是 FAST 已验证实验。

S9（V:/embodied-3d-affordance-policy/DELIVERY_REPORT.md）：45° 相机变化时，准确标定且带表面配准的世界点云 Flow 为 59/64，固定像素映射图像 Flow 为 0/64。编码动作前先明确世界系、工具系或相机系。FAST 可压缩不同坐标动作，但压缩本身不会修复错误外参。测试新增障碍的 64/64 无碰撞来自步内 SDF 路径修正，不来自动作词元化。

原文差异与复现时应保留的条件

附录 E 文字称 DROID 定量有 16 项任务，而表 II 列出 17 条任务指令，试验数合计 44；本册保留两处原值。附录 E 写“约 3 episodes”，而附录 D 明确三轮数据训练（3 epochs）；译文在该段标注差异。FAST+ 的离线通用压缩测试覆盖训练未见机器人，但策略在线控制测试主要是固定操作臂；这两种泛化证据应分开读取。

Terminology ledger

English	中文与约定
action chunk	动作块，一段未来动作，不等于实际全部开环执行
tokenization	词元化，统一不混用“分词”描述机器人动作
discrete cosine transform, DCT	离散余弦变换
byte-pair encoding, BPE	字节对编码
frequency-space	频域
reconstruction fidelity	重建保真度
autoregressive	自回归
FAST+	通用预训练动作词元化器，保留专名

English engineering takeaway

FAST changes the action representation, not the closed-loop data distribution. Its DCT quantizer determines numerical distortion; BPE then compresses the integer sequence losslessly. Audit normalization and coordinate frames before comparing token counts. Keep training compute, chunk inference latency, reconstruction error and robot success as separate measurements. The local S3 contact failures and S9 registration controls are especially useful counterchecks when transferring this tokenizer to a new policy.