arXiv:2602.01811v1 [cs.RO] 2 Feb 2026

从认知到精准行动：面向 VLA模型的通用自我纠正与终止框架

张文涛*†、*Aolan Sun*†‡、莫文涛*†、*Xiaoyang Qu*†、郑宇鑫*†、*Jianzong Wang*†‡

*清华大学深圳国际研究生院，清华大学，中国深圳，
†平安科技（深圳）有限公司，中国深圳

摘要

尽管面向具身智能体的视觉-语言-动作（VLA）模型整合了感知、推理与控制，但仍受两项关键弱点制约：首先，在抓取任务中，语言模型生成的动作令牌常与目标物体存在细微的空间偏差，导致抓取失败；其次，这些模型缺乏可靠识别任务是否完成的能力，因而会产生冗余动作并频繁出现超时错误。为应对这些挑战并增强鲁棒性，我们提出一种轻量级、无需训练的框架——**VLA-SCT**。该框架作为自我纠正控制循环运行，将数据驱动的动作优化与用于终止任务的条件逻辑相结合。因此，相较于基线方法，我们的方法在 LIBERO基准的所有数据集上均取得稳定提升，显著提高了精细操作任务的成功率，并确保任务准确完成，从而推动更可靠的视觉-语言-动作智能体在复杂、非结构化环境中的部署。

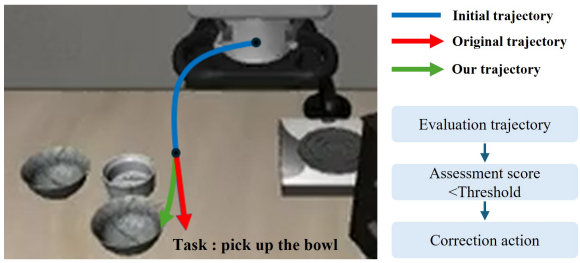
索引词— 视觉-语言-动作模型、具身智能、自我纠正、任务终止

1. 引言

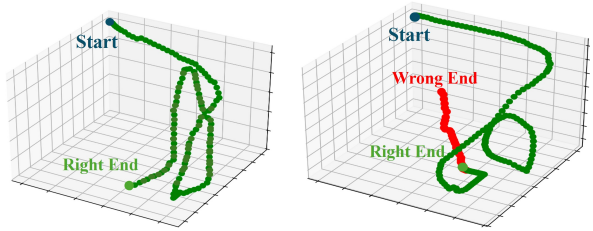
视觉-语言-动作（VLA）模型是具身人工智能领域的一项重要进展 [1, 2]。这些模型通过在统一的架构框架中整合视觉感知、自然语言处理和动作生成，实现了这一目标，使机器人智能体能够理解高层次的人类指令，并以相当强的适应性执行相应的物理任务 [3]。凭借精细的语义理解能力，VLA模型展现出强大的泛化能力，尤其是在复杂环境中导航和操控新颖物体方面。VLA领域的先驱性实现，例如 Google DeepMind 的 RT-1 [4] 和 RT-2 [5], , 证明了在大规模互联网数据与机器人轨迹数据集上进行联合训练，能够使这些模型具备泛化至未见物体和新颖指令的能力。该领域后续的发展包括 RT-X系列 [6], 、开源模型 OpenVLA [7], 以及较新的 π 系列，其中包括采用流匹配策略的 π_0 [8] 和 $\pi_{0.5}$ [9]。为提升这些模型的性能，研究人员采用了模型剪枝和令牌稀疏化等技术来加速推理，例如 SparseVLM [10]和 FastVLM [11]。

尽管这些方法在一定程度上可以提高推理速度，却大幅降低了模型在每次.....中的成功率

本研究获得深港联合资助项目（A 类）资助，项目编号为 SGDX20240115103359001。‡通讯作者为 Aolan Sun（aolansun@163.com）和 Jian-zong Wang（jzwang@188.com）。



（a）针对精度问题的动作自我纠正



（b）因未能终止而产生的多余动作可视化。

图 1：VLA-SCT 框架应对视觉-语言-动作模型中的两项根本挑战。（a）针对运动精度问题，纠正偏离的轨迹。（b）针对任务终止问题，防止任务完成后继续执行多余动作。

这阻碍了它们在现实场景中的可靠部署 [12]。当前 VLA 模型的核心问题是知识-动作鸿沟。首先，这一鸿沟的主要体现是模型在精细操作任务中缺乏精确性。这一缺陷通常源于其执行策略主要采用开环方式，缺少在任务执行过程中纠正细微偏差所必需的实时反馈机制。因此，灵巧抓取或放置等要求高精度控制的任务容易出现轨迹偏差、抓取不稳乃至执行失败 [13]。这一局限在医疗和工业自动化等应用中尤为明显，因为这些领域通常要求亚毫米级的精度 [1]。如图 1a 所示，基线视觉-语言-动作模型的开环特性可能导致执行偏差，例如未能抓中目标碗（红线）。我们提出的自我纠正机制通过先评估轨迹，再采取纠正动作（绿线）来解决这一问题，确保任务精确完成。其次，一个普遍存在的局限是缺乏稳健的任务终止机制，导致许多 VLA 模型无法识别既定目标何时已经达成。因此，这些模型常常表现出非终止行为，不断执行多余动作，直到达到预设的步数上限，从而导致

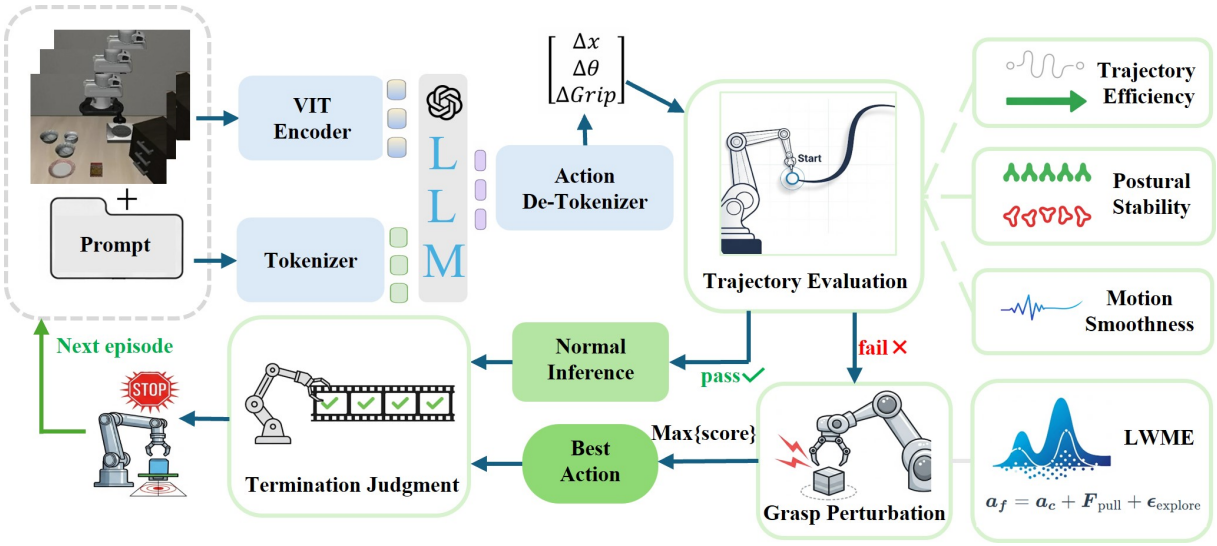


图 2：VLA-SCT 框架概览。我们的方法通过引入用于动作自我纠正的条件控制回路，以及用于可靠任务终止的视觉判断模块，增强了标准 VLA 模型。

任务被错误地判定为失败 [14]。这种行为严重损害运行效率和系统可靠性，带来重大安全风险，也构成了在开放世界环境中部署自主智能体的一项根本挑战 [15, 16]。图 1b 清楚展示了这一失败模式：机械臂尽管已成功完成任务，却仍持续运动，直至超过最大步数，最终导致超时失败。

为解决这些问题，我们提出了一个通用的**自**我**纠**正与**终**止框架，即**VLA-SCT** 框架。该框架通过整合自我纠正和终止检测机制，系统地提升了 VLA 模型的稳健性和可靠性。我们的主要贡献总结如下：

- 我们提出并实现了一个通用、轻量的 VLA-SCT 框架，旨在增强现有的 VLA 模型。该框架在 VLA 模型之外构建了一个模块化智能控制层，从而降低了动作执行和任务终止过程中出现故障的风险。
- 我们提出了两种数据驱动机制：一种基于局部加权矩估计（LWME）的在线自校准模块，以及一种基于视觉特征匹配的非参数终止决策模块。
- 大量实验表明，我们的 VLA-SCT 框架不仅取得了最高的平均成功率，还显著提高了推理效率，展现出该框架的高效性和巨大潜力。

2. 方法

VLA-SCT 框架是一种模块化智能控制层，可增强现有的 VLA 模型，如图 2 所示。该框架引入了一个评估—纠正—终止反馈回路，通过处理基线模型生成的初始运动规划，提高执行的鲁棒性。以下小节将详细介绍该回路的三个核心组件：轨迹评估模块、抓取扰动模块和终止检测模块。

2.1. 轨迹评估

在任务初始阶段，我们设计了一种轻量级轨迹质量评估器，用于分析机器人早期运动轨迹的效率、平滑度和稳定性，从而预测任务失败的风险。

轨迹效率。 我们将轨迹效率 S_{eff} 定义为一个归一化得分，与路径的总曲率和总挠率成反比，其定义如下：

$$S_{eff} = \frac{1}{1 + a\int_{\gamma}\kappa(s)ds + b\int_{\gamma}|\tau(s)|ds} \times 100\%, \qquad (1)$$

其中， γ 表示机器人末端执行器在评估窗口内三维空间轨迹的一段， a， b为参数， s 为轨迹的弧长参数， $\kappa(s)$表示轨迹在点 s处的曲率， $\tau(s)$表示轨迹在点 s处的挠率。

姿态稳定性。 为评估轨迹的姿态稳定性，我们采用微分几何中的方法，并将每个机器人姿态建模为 SO(3) 流形上的一个点。我们的姿态稳定性得分定义为沿轨迹累积的测地距离呈指数衰减的函数，其定义如下：

$$S_{sta} = \exp\left(-k\sum_{t}\arccos\left(\frac{\mathrm{Tr}(\mathbf{R_t}\mathbf{R_{t-1}^T}) - 1}{2}\right)\right), \qquad (2)$$

其中$\mathbf{R_t}$ 是一个 3×3 旋转矩阵，表示机器人末端执行器在离散时间步 t的三维空间方位， k 为平滑系数。该指标通过累积每对连续姿态之间最短的旋转角，准确反映轨迹的时间连续性。

运动平滑度。 我们的评估指标基于量化一条轨迹的总加加速度。该方法遵循最小加加速度原理，这是生物力学和机器人学中描述平滑、自然运动的黄金标准，即将加速度的变化率最小化。给定一条由位姿点组成的机器人轨迹 $\{ \boldsymbol{p_1}\ \boldsymbol{q_1}\ \cdots\ \boldsymbol{p_N}\ \boldsymbol{q_N}\ \}$ N $\boldsymbol{p_i}(.),(.),$

其中

是位置向量， $\boldsymbol{q_i}$ 是表示姿态的四元数，采样时间间隔为 Δt，该分数定义为

$$S_{smo} = \exp(-\mu \sum_i \left(\|\boldsymbol{j_{p,i}}\|^2 + w_{rot} \cdot \|\boldsymbol{\zeta_{q,i}}\|^2 \right) \Delta t), \quad (3)$$

其中，$\boldsymbol{j_{p,\,i}}$ 为平移加加速度，$\boldsymbol{\zeta_{q,\,i}}$ 为角加加速度， μ 为敏感度系数， w_{rot} 为旋转权重。

2.2. 抓取扰动

当轨迹评估模块判定初始动作规划质量较低时，VLA-SCT 框架会激活抓取扰动模块。该模块并非简单地替换原始动作，而是作为一种基于经验的在线自适应校正机制运行。其核心原理是利用根据当前视觉环境动态构建的概率分布生成有针对性的扰动——该分布由历史上成功的动作构成——从而引导可能失败的动作转向成功概率更高的区域。我们使用径向基函数（RBF）核来衡量当前视觉特征向量 $\boldsymbol{v_c}$ 与记忆库中每个成功历史经验的视觉特征 $\boldsymbol{v_i}$ 之间的相似度。该相似度值用作权重 w_i:

$$w_i = \exp(-\gamma \|\boldsymbol{v_c} - \boldsymbol{v_i}\|^2), \quad (4)$$

其中， γ 是控制衰减率的超参数。权重 w_i 表示第 i 个历史成功动作 $\boldsymbol{a_i}$与当前情境的相关性。利用计算得到的权重 w_i，我们对相关的历史成功动作 a_i 进行加权统计，以估计成功动作分布的中心和形状。首先，我们计算加权平均值 $\boldsymbol{a_\mu}$，它表示这一视觉场景中成功动作分布的密度中心:

$$\boldsymbol{a_\mu} = \frac{\sum_i w_i \boldsymbol{a_i}}{\sum_i w_i}. \quad (5)$$

接下来，我们计算加权协方差矩阵，用以描述七维动作空间中成功动作的分布形状、方向以及各维度之间的相关性:

$$\mathbf{Q} = \frac{\sum_i w_i (\boldsymbol{a_i} - \boldsymbol{a_\mu})(\boldsymbol{a_i} - \boldsymbol{a_\mu})^T}{\sum_i w_i}. \quad (6)$$

协方差矩阵 $\mathbf{Q}$ 是生成结构化随机扰动的关键。为确保数值稳定性，我们将其正则化为 $\mathbf{Q_{reg}} = \mathbf{Q} + \lambda \mathbf{I}$，其中 λ 是避免矩阵奇异性的最小常数。我们设计的最终扰动由三部分组成：将动作拉向成功区域的确定性重力项、用于探索成功分布模式的各向异性噪声项，以及促进泛化的各向同性噪声项，如下所示:

$$\boldsymbol{a_f} = \boldsymbol{a_c} + \alpha(\boldsymbol{a_\mu} - \boldsymbol{a_c}) + \sqrt{T}\mathcal{N}(0, \beta \mathbf{Q}_{\mathrm{reg}}) + \mathcal{N}(0, \sigma^2 \mathbf{I}), \quad (7)$$

其中， α 和 β 是可控参数， T 是采样温度。在实现层面，除了对协方差矩阵进行正则化外，我们还需要对生成的运动向量（例如夹爪开合幅度）进行裁剪，以确保所有输出值都在机器人的物理限制范围内，从而保证算法的稳定性和安全性。

2.3. 终止检测

我们的终止检测模块通过将当前视觉状态与动态更新的存储库 $\mathcal{M} = \{\boldsymbol{I_1}, \boldsymbol{I_2}, \ldots, \boldsymbol{I_N}\}$进行比较来判断任务是否完成，其中 $\boldsymbol{I_i}$ 表示过去成功时的视觉状态。这种非参数方法避免了硬编码规则，因此具有很强的泛化能力和适应性。在每一步中，该模块都会将当前相机图像 $\boldsymbol{I_0}$ 与存储库中的每张图像 $\boldsymbol{I_i}$ 进行比较。我们对两张图像进行预处理并将其展平为一维向量 $\boldsymbol{X} = \{x_1, x_2, \ldots, x_n\}$ 和$\boldsymbol{Y} = \{y_1, y_2, \ldots, y_n\}$，其中 n 是像素总数。我们使用皮尔逊相关系数 r 作为核心相似度指标，因为它对亮度和对比度的变化具有较强的鲁棒性。我们不直接进行逐像素比较，而是将 r 计算为两个向量之间的标准化协方差: 对所有像素处的中心化偏差$(x_i - \mu_x)$和$(y_i - \mu_y)$的乘积求和，再除以两个向量的标准差之积，其中 μ_x 和 μ_y 分别表示向量$\boldsymbol{X}$和 $\boldsymbol{Y}$中元素的均值。随后，该系数会被线性转换为直观的相似度分数 $S = \frac{r+1}{2}$，其范围为0到1。模块会计算当前视图与所有存储图像之间的最高相似度分数 $S_{\max}$。如果该分数超过预设的终止阈值 τ，则认为任务已完成，并发出停止信号；否则，继续执行。阈值 τ 是一个可调超参数，可用于将系统的终止策略从保守调整为激进。

3. 实验

3.1. 数据集与指标

我们的所有实验均在 LIBERO基准 [17]上进行。LIBERO 是视觉-语言-动作和机器人学习领域广泛使用的基准。我们根据任务成功率和推理速度对模型进行了评估。任务成功率是评估框架任务执行能力的核心指标，直接反映了我们提出的自我校正与终止模块在复杂操作场景中的实际有效性和稳健性。

3.2. 实现细节

我们的实验使用公开可用的 OpenVLA-7B [7] 模型作为基线。我们提出的 VLA-SCT 框架是一个仅用于推理的增强模块，无需训练。这一设计与其他领域近期采用的高效方法一致，例如语音识别纠错 [18]。因此，所有实验中预训练的 OpenVLA模型权重均保持冻结。所有输入图像均处理为 OpenVLA 标准的 224×224 像素。实验在单台 NVIDIA RTX 4090 GPU 上进行，模型以 bfloat16精度加载，约需 15GB 的 CUDA显存。

3.3. 实验结果

与其他方法的比较。 如表1所示，我们全面的实验结果清楚地表明，所提出的 VLA-SCT 框架在具有挑战性的 LIBERO基准上达到了新的最先进水平。我们的方法取得了 81.55% 的平均成功率，较 OpenVLA 基线（75.45%）绝对提升了 6.1%，并优于所有对比方法。这一优势在四类任务中始终如一: VLA-SCT 在目标（82.00%）、物体（92.80%）、空间（91.20%）以及具有挑战性的长时任务（60.20%）中均取得最佳表现。值得注意的是，VLA-SCT

表1：在 LIBERO 基准上，我们提出的方法与最先进方法的性能比较，其中“+R”表示保留相对令牌位置，“+S”表示融入 Canny 边缘信息。

方法	成功率（%，↑）/ 加速比（↑）				
	Goal	物体	空间	Long	平均值
OpenVLA [7]	76.40 / 1.00	88.80 / 1.00	83.20 / 1.00	53.40 / 1.00	75.45 / 1.00
SparseVLM [10]	74.20 / 1.33	84.00 / 1.33	83.40 / 1.33	52.80 / 1.33	73.60 / 1.33
Fopru + R [19]	59.80 / 1.29	81.20 / 1.29	71.60 / 1.30	26.20 / 1.35	59.70 / 1.31
PruMerge + R [20]	0.00 / 1.54	0.00 / 1.32	0.00 / 1.27	0.00 / 1.32	0.00 / 1.36
FastVLM +R +S [11]	73.20 / 1.21	77.00 / 1.11	79.80 / 1.12	36.60 / 1.20	66.65 / 1.16
VisionZip + R +S [21]	46.00 / 1.20	47.40 / 1.23	34.20 / 1.19	4.60 / 1.22	33.05 / 1.21
SP-VLA [22]	75.40 / 1.46	85.60 / 1.30	84.40 / 1.30	54.20 / 1.32	74.90 / 1.35
VLA-SCT（我们的方法）	**82.00** / 1.22	**92.80** / 1.09	**91.20** / 1.10	**60.20** / 1.06	**81.55** /1.12

表2：对各个模块有效性的消融研究。模块A为轨迹评估；模块B为抓取扰动；模块C为终止检测。

模型	模块		成功率（%）
	A+B	C	
VLA-SCT	✓	✓	**91.20**
-	✓	×	87.40
-	×	✓	85.60
基线	×	×	83.20

在不牺牲效率的前提下实现了较高的任务成功率。SP-VLA 虽然以牺牲准确率为代价获得了更高的加速比，但我们的方法实现了最佳的综合权衡，在达到最先进的成功率的同时，相较 OpenVLA 实现了 1.12× 倍加速。这表明，VLA-SCT 框架不仅能够纠正错误动作，还能使模型在完成任务后及时终止。这种兼具卓越准确率与实用效率的独特优势，证明了我们的框架通过提升模型内在的决策质量而非牺牲性能来换取速度，从而彰显了其在现实世界机器人应用中的稳健性与价值。

消融研究。 为量化各核心组件的贡献，我们在 LIBERO空间套件上开展了消融研究，结果汇总于表2。表1已验证我们框架的整体效能，而选择这一子集是因为其准确性要求严格，能够细致分析各个模块的贡献。我们的分析从基线模型开始，该模型在未使用任何所提模块时的成功率为 83.20%。仅加入自我纠正机制（模块 A+B）便将成功率提升至 87.40%，验证了轨迹评估和抓取扰动模块在提高操作精度方面的效能。同样，仅部署模块 C 就通过减少超时导致的失败，将成功率提升至 85.60 %。尤为重要的是，在完整的 VLA-SCT 框架中结合所有组件后，模型取得了最高的成功率 91.20%，证实自我纠正机制和终止机制带来了显著且互补的收益。本研究验证了，提升动作精度并可靠识别任务完成状态，对于框架整体达到最先进水平至关重要。

敏感性分析。 为探究关键超参数的影响，

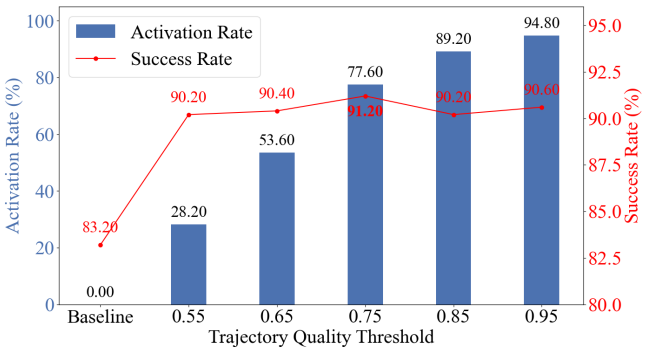


图 3：轨迹质量阈值的敏感性分析。

我们对轨迹质量阈值进行了敏感性分析，如图3所示。该阈值是核心把关机制，决定是否激活抓取扰动模块。随着阈值收紧，激活率单调上升（阈值从0.55升至0.95时，激活率从28.20%升至94.80%），符合干预次数增加的预期。值得注意的是，成功率呈现出明显的峰值，在阈值为0.75时达到最高值91.20%。然而，当阈值进一步提高至0.85和0.95时，尽管激活率持续上升，成功率却略有下降。这一趋势凸显了其中的权衡：阈值过于宽松会遗漏部分错误，而阈值过于严格则会导致过度纠正，对有效轨迹进行不必要的干预可能引入不稳定性。因此，我们确定0.75为最佳阈值，既能最大限度地提高成功率，又能避免不必要的干预。

4. 结论

本文提出了 VLA-SCT，这是一种用于增强预训练 VLA 模型的创新通用框架。我们总结了现有 VLA 模型在精细操作和可靠完成任务方面的局限性，并通过集成轨迹评估、抓取扰动和终止检测模块，提高了操作精度和任务可靠性。在 LIBERO 上开展的对比研究和消融研究验证了我们达到了最先进性能，推动了 VLA 自主性的提升。未来工作的一个重要方向是研究该框架在不同 VLA架构中的适用性，以改善跨模型泛化。

5. 参考文献

[1] Zhaoyuan Gu、Junheng Li、Wenlan Shen、Wenhao Yu、Zhaoming Xie、Stephen McCrory、Xianyi Cheng、Abdulaziz Shamsah、Robert Griffin、C Karen Liu 等人，“人形机器人运动与操作：控制、规划和学习的当前进展与挑战”，*IEEEASME* 机电学报，2025年。[2] Songming Liu、Lingxuan Wu、Bangguo Li、Hengkai Tan、Huayu Chen、Zhengyi Wang、Ke Xu、Hang Su 和 Jun Zhu，“RDT-1B：用于双臂操作的扩散基础模型”，发表于第十三届国际学习表征会议，2025年。[3] Ziqi Jia、Junjie Li、Xiaoyang Qu 和 Jianzong Wang，“通过结合基于大语言模型的规划器和基于图的策略的强化学习增强多智能体系统”，发表于*2025*年 *IEEE* 国际机器人与自动化会议（*ICRA*），2025年，第1240–1246页。

[4] Anthony Brohan、Noah Brown、Justice Carbajal、Yevgen Cheb-otar、Joseph Dabis、Chelsea Finn、Keerthana Gopalakrishnan、Karol Hausman、Alexander Herzog、Jasmine Hsu 等，《RT-1：面向大规模真实世界控制的机器人Transformer》，机器人：科学与系统第十九届会议，2023年。

[5] Brianna Zitkovich、Tianhe Yu、Sichun Xu、Peng Xu、Ted Xiao、Fei Xia、Jialin Wu、Paul Wohlhart、Stefan Welker、Ayzaan Wahid 等，《RT-2：视觉-语言-动作模型将网络知识迁移至机器人控制》，载于机器人学习会议。PMLR，2023年，第2165–2183页。

[6] Abby O’Neill、Abdul Rehman、Abhiram Maddukuri、Ab-hishek Gupta、Abhishek Padalkar、Abraham Lee、Acorn Poo-ley、Agrim Gupta、Ajay Mandlekar、Ajinkya Jain 等，《Open X-Embodiment：机器人学习数据集与RT-X模型：Open X-Embodiment协作项目》，载于*2024*年*IEEE*国际机器人与自动化会议（*ICRA*）。IEEE，2024年，第6892–6903页。

[7] Moo Jin Kim、Karl Pertsch、Siddharth Karamcheti、Ted Xiao、Ashwin Balakrishna、Suraj Nair、Rafael Rafailov、Ethan P Fos- ter、Pannag R Sanketi、Quan Vuong 等人，“Openvla：一种开源视觉-语言-动作模型”，载于机器人学习会议。PMLR，2025年，第2679–2713页。

[8] Kevin Black、Noah Brown、Danny Driess、Adnan Esmail、Michael Equi、Chelsea Finn、Niccolo Fusai、Lachy Groom、Karol Hausman、Brian Ichter 等人，“π_0：一种用于通用机器人控制的视觉-语言-动作流模型”，*arXiv*预印本 *arXiv.2410.24164*，2024年。

[9] Kevin Black、Noah Brown、James Darpinian、Karan Dhabalia、Danny Driess、Adnan Esmail、Michael Equi、Chelsea Finn、Niccolo Fusai、Manuel Y Galliker 等人，“$\pi_{0.5}$ 一种具有开放世界泛化能力的视觉-语言-动作模型”，*arXiv*预印本 *arXiv:2504.16054*，2025年。

[10] Yuan Zhang、Chun-Kai Fan、Junpeng Ma、Wenzhao Zheng、Tao Huang、Kuan Cheng、Denis A Gudovskiy、Tomoyuki Okuno、Yohei Nakata、KurtKeutzer 等，“SparseVLM：面向高效视觉-语言模型推理的视觉令牌稀疏化”，载于第四十二届国际机器学习会议，2025年。

[11] Pavan Kumar Anasosalu Vasu、Fartash Faghri、Chun-Liang Li、CemKoc、NateTrue、Albert Antony、GokulaSanthanam、James Gabriel、Peter Grasch、Oncel Tuzel 等，“Fastvlm：面向视觉语言模型的高效视觉编码”，载于计算机视觉与模式识别会议论文集，2025年，第19769–19780页。[12] Yifan Zhong、Fengshuo Bai、Shaofei Cai、Xuchuan Huang、Zhang Chen、Xiaowei Zhang、Yuanfei Wang、Shaoyang Guo、Tianrui Guan、Ka Nam Lui 等，“视觉-语言-动作模型综述：动作令牌化视角”，*arXiv*预印本 *arXiv:2507.01925*，2025年。[13] Karl Pertsch、Kyle Stachowicz、Brian Ichter、Danny Driess、Suraj Nair、Quan Vuong、Oier Mees、Chelsea Finn 和 Sergey Levine，“FAST：面向视觉-语言-动作模型的高效动作令牌化”，载于机器人学：科学与系统会议论文集，美国加利福尼亚州洛杉矶，2025年六月。

[14] Junjie Li、Nan Zhang、Xiaoyang Qu、Kai Lu、Guokuan Li、Jiguang Wan 和 Jianzong Wang，“RATE-nav：面向视觉语言模型零样本物体导航的区域感知终止增强”，载于计算语言学协会发现：*ACL 2025*，2025年，第6564–6574页。

[15] Ziqi Jia、Anmin Wang、Xiaoyang Qu、Xiaowen Yang 和 Jianzong Wang，“面向具身持续学习的层次任务感知的增量式 LoRA 专家多模态混合”，载于第*63*届计算语言学协会年会论文集，2025年，第28415–28427页。

[16] Hui Wei、Hao Tang、Xuemei Jia、Zhixiang Wang、Hanxun Yu、Zhubo Li、Shin’ichi Satoh、Luc Van Gool 和 Zheng Wang，“物理对抗攻击遇上计算机视觉：十年综述”，*IEEE*模式分析与机器智能汇刊，第46卷，第12期，页9797–9817，2024年。

[17] Bo Liu、Yifeng Zhu、Chongkai Gao、Yihao Feng、Qiang Liu、Yuke Zhu 和 Peter Stone，“Libero：终身机器人学习的知识迁移基准测试”，神经信息处理系统进展，第36卷，44791，2023年，44776页–。

[18] Chao-Han Huck Yang、Yile Gu、Yi-Chieh Liu、Shalini Ghosh、Ivan Bulyko 和 Andreas Stolcke，“利用大语言模型和任务激活提示进行生成式语音识别错误纠正”，载于*2023*年 *IEEE* 自动语音识别与理解研讨会（*ASRU*）。IEEE，2023年，页1–8。

[19] Lei Jiang、Weizhe Huang、Tongxuan Liu、Yuting Zeng、Jing Li、Lechao Cheng 和 Xiaohua Xu，《Fopru：面向高效大型视觉语言模型的焦点剪枝》，*arXiv* 预印本 *arXiv:2411.14164*，2024年。

[20] Yuzhang Shang、Mu Cai、Bingxin Xu、Yong Jae Lee 和 Yan Yan，《Llava-prumerge：面向高效大型多模态模型的自适应令牌缩减》，载于*IEEE/CVF* 计算机视觉国际会议论文集，2025年，第22857–22867页。

[21] Senqiao Yang、Yukang Chen、Zhuotao Tian、Chengyao Wang、Jingyao Li、Bei Yu 和 Jiaya Jia，《VisionZip：视觉语言模型中，更长并不意味着更好》，载于计算机视觉与模式识别会议论文集，2025年，第19792–19802页。

[22] Ye Li、Yuan Meng、Zewen Sun、Kangye Ji、Chen Tang、Jiajun Fan、Xinzhu Ma、Shutao Xia、Zhi Wang 和 Wenwu Zhu，《SP-VLA：一种联合模型调度与令牌剪枝的 VLA 模型加速方法》，*arXiv* 预印本 *arXiv:2506 12723*，2025年。。