arXiv:2603.02271v1 [cs.PF] 1 Mar 2026

视觉-语言-动作模型特征分析：识别边缘人工智能架构的动作生成瓶颈

Manoj Vishwanathan[1,2],Suvinay Subramanian[1], 和 Anand Raghunathan[2]1谷歌，美国加利福尼亚州山景城2普渡大学，美国印第安纳州西拉法叶

摘要

视觉-语言-动作（VLA） [1] 模型是一类新兴工作负载，对边缘端的机器人学和具身智能至关重要。随着模型规模不断扩大，其能力显著提升，但为了满足实时应用严格的延迟要求，必须在本地部署。本文分析了两代边缘硬件上的 VLA 性能，即Nvidia Jetson Orin [6]和 Thor [7]平台。我们采用先进的 VLA 模型 MolmoAct-7B[4], ，发现了一个主要执行瓶颈：端到端延迟中，高达 75% 消耗在内存受限的动作生成阶段。通过分析建模与仿真，我们预测了扩展至千亿参数模型所需的硬件配置。我们还探讨了高带宽内存技术和存内处理（PIM）[3] 作为具身智能边缘系统未来发展方向的潜力。

1 问题陈述

视觉-语言-动作（VLA）模型代表了具身智能的一次变革性转变，使机器人能够从僵化的预编程程序转向通用的语义推理。与大语言模型（LLMs）一样，这些模型的发展日益受到神经网络缩放定律的支配；近期研究表明，机器人任务性能以幂律速率提升 [8]。为了在复杂的现实环境和任务中实现真正的通用效用，模型必须扩展到 10—100B 个参数，从而有效地将世界知识与精确的感觉运动控制相结合。

然而，这种扩展需求在实时部署中面临关键的硬件障碍。在物理环境中进行安全、动态的操作，需要至少10–20 Hz的稳定控制频率。当代边缘加速器在结构上并不适合视觉-语言-动作模型动作生成所需的稀疏、内存受限的自回归处理。尽管Gemini Robotics 1.5 [2] 等最先进模型已通过“双脑”架构展现出出色的推理能力，但这些方案依赖于将能力较强的模型卸载到强大的板外服务器，同时仅在本地维持 rudimentary 控制回路。此类混合配置对演示有效，但在低延迟、完全自主的边缘系统场景中仍远不可行。

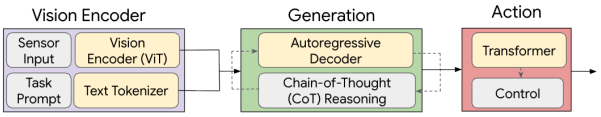


图1。VLA 系统架构。该工作负载由视觉编码器、生成引擎（自回归解码）和动作变换器组成。

本文对视觉-语言-动作模型工作负载进行了系统性表征，以指导下一代边缘系统设计。我们在当代Nvidia Jetson Orin与Thor硬件平台上对MolmoAct-7B模型进行性能剖析，以建立基线并识别关键执行瓶颈。在这些测量结果的基础上，我们利用内部模拟器来预测未来视觉-语言-动作模型在当前及假设的未来硬件配置上的性能。

2 背景与相关工作：视觉-语言-动作（VLA）模型

视觉-语言-动作（VLA）模型是一类多模态基础模型，融合了视觉感知、自然语言理解和物理动作。如图1所示，VLA 的计算架构由三个主要子系统构成：

- 视觉编码器（感知核心）：此阶段将原始像素数据处理为结构化特征嵌入。为同时捕捉语义上下文和空间几何细节，现代视觉-语言-动作模型通常采用 SigLIP 和 DINOv2 等融合骨干网络。这些高维特征通过投影器模块（通常为多层感知机）映射到推理引擎的嵌入空间中。
- 生成（推理引擎）：模型的核心是一个仅解码器 Transformer，它处理由视觉和文本词元拼接而成的序列。在此阶段，模型执行跨模态推理，并可能生成中间输出，例如“思维链”（CoT）推理或空间航点，以便将高级指令分解为可执行计划。
- 动作Transformer：最后阶段将模型的内部表示转换为电机命令。在离散动作词元化中，机器人的

连续动作空间被量化为多个分箱，使模型能够在现有词汇表中将动作预测为离散词元。在连续动作生成中，则使用扩散 Transformer（DiT）等专用解码器，输出平滑、高频的关节轨迹或末端执行器轨迹。

3 方法论

我们的评估采用混合方法，将实证硬件性能分析与高保真模拟预测相结合。这种方法使我们能够利用当代边缘平台建立基准，并在此基础上扩展研究结果，以评估更大规模模型在假设的未来硬件架构上的性能。

3.1 真实硬件特性分析

为建立实证基准，我们对 MolmoAct-7B 在 NVIDIA Jetson AGX Orin (64GB) 和 NVIDIA Jetson Thor (128GB) 平台上的性能进行特性分析。我们使用 NVIDIA Nsight Compute 对 PyTorch 运行时进行插桩 [5]，以捕获内核级执行跟踪。这项性能分析使我们能够将端到端延迟分解为具体阶段：视觉编码、生成（自回归解码）和动作Transformer。

3.2 性能预测与模拟

对于未来架构和更大模型规模的建模，我们采用了内部开发的高保真度 XPU 模拟器。该模拟器包含经过多个量产级加速器（GPU、TPU）验证的详细解析性能模型，在多个量产级模型（x−−xxx B 参数量的 LLM）上实现了 70% 至 90% 的准确率。该模拟器将视觉-语言-动作模型分解为以下组成阶段：视觉编码、自回归解码和动作生成。每个阶段均建模为多层 Transformer 主干网络，其中每一层又进一步解析为一系列算子，主要是高维爱因斯坦求和。该建模框架的主要特性包括：

- 微架构保真度：成本模型纳入了具体的硬件细节，包括流式多处理器（SM）的数量、分块策略，以及 XPU 矩阵引擎在不同维度上的非对称带宽特性。
- 解析屋顶线：使用屋顶线模型计算各个算子的性能，同时考虑计算和内存带宽限制。
- 跨算子优化：该框架跨越算子边界进行优化，以实现有效的预取。这对于内存受限操作尤为重要，因为它可以提前通过内存层次结构移动操作数，从而最大限度地减少停顿。

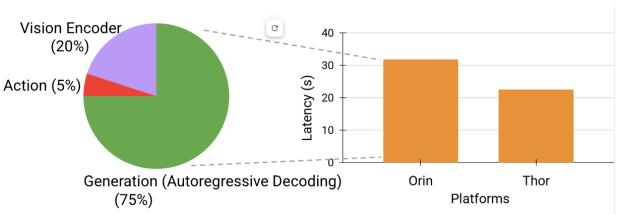


图 2. 当前边缘平台上的性能。MolmoAct-7B 在 Jetson Orin 和 JetsonThor 上的延迟。

商用硬件 内存		带宽 (GB/s)	BF16 TFLOPS
Orin	LPDDR5	203	100
Thor	LPDDR5X	273	500
假设变体 内存		**带宽 (GB/s)**	**BF16 TFLOPS**
Orin+LPDDR5X	LPDDR5X	273	100
Orin+GDDR7	GDDR7	1000	100
Orin+存内处理	LPDDR6X 存内处理	2180	1074
Thor+GDDR7	GDDR7	1000	500
Thor+存内处理	LPDDR6X 存内处理	2180	3993

表 1. 商用边缘平台和假设的硬件系统用于我们的实验。对于采用存内处理的系统，计算吞吐量（BF16 TFLOPS）包括片上系统和存内处理两部分。

4 评估与分析

本节展示并分析了我们的测量和仿真结果。

4.1 真实硬件特性分析

图 2总结了我们的性能分析结果，并表明：(i) 延迟 ~ 200 – 300× 高于实时（10Hz）运行所需的水平；(ii) 生成阶段（包含推理的自回归解码）是视觉-语言-动作模型的主要瓶颈，占（完整模型）单步延迟的 ≈ 75%；(iii) 生成阶段主要受内存带宽限制——尽管 Thor 的算力是 Orin 的 5 倍，端到端延迟却仅提升了 1.4 倍。

4.2 未来架构上的扩展模型

我们按照 [1, 8]中的缩放定律，将视觉-语言-动作模型扩展至 100B 个参数。我们还在内部模拟器中，针对一系列当前及假设系统评估这些模型。假设系统基于标准的现成边缘硬件构建，并为其配备了更强大的内存系统（规格见表 1）。图 3显示，GDDR7 和存内处理内存带来的带宽提升显著改善了性能；然而，其性能仍远低于实际部署所需的 10–20 Hz 目标。

5 结论

空间推理 VLA 工作负载的主要瓶颈在于受内存限制的动作生成阶段。标准的内存扩展方案不足以在交互速率下处理 10–100B 参数模型。未来的研究必须探索全面的系统优化——涵盖硬件和软件，以弥合具身智能的延迟差距。

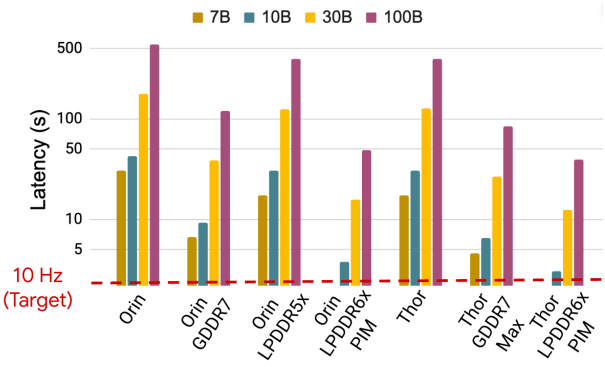


图3. 不同边缘系统配置下的控制频率 更高的内存带宽和存内处理可提高控制频率，但要在更大的模型规模下实现长时域动作生成的 10 Hz 目标，还需要新的创新以及算法与系统协同设计。

AI 使用声明

作者使用 Gemini（谷歌）起草问题陈述和相关工作部分，调试 MolmoAct-7B 模型的 PyTorch 性能分析脚本，并对稿件的结构提供反馈。所有由 AI 生成的作者对相关建议进行了审查和核实，并对结果的技术完整性承担全部责任。

参考文献

[1] Anthony Brohan等人。2023。RT-2：视觉语言动作模型将网络知识迁移至机器人控制。*arXiv*预印本 *arXiv:2307.15818* abs/2307.15818（2023），1–15。[2] 谷歌DeepMind。2025。Gemini Robotics 1.5：借助先进的具身推理、思考与运动迁移，推动通用型机器人的前沿。*arXiv*预印本 *arXiv:2510.03342* abs/2510.03342（2025），1–1。[3] S. Lee等人。2021。基于商用DRAM技术的存内处理硬件架构与软件栈。载于*ISCA*。IEEE，西班牙瓦伦西亚，43–56。[4] S. Lee等人。2025。MolmoAct：能够在空间中进行推理的动作推理模型。*arXiv*预印本 *arXiv:2508.07917*（2025），1–15。[5] 英伟达公司。 [无日期]。NVIDIA Nsight Compute。https: //developer.nvidia.com/nsight-compute。访问日期：2026-02-04。[6] 英伟达公司。2022。*NVIDIA Jetson AGX Orin*技术简报。英伟达，加利福尼亚州圣克拉拉。[7] 英伟达公司。2024。*NVIDIA Drive Thor*：集中式计算的下一代。GTC主题演讲，加利福尼亚州圣何塞。[8] Neil Thompson、Sebastian Sartor。2024。机器人学中的神经网络缩放定律。*arXiv*预印本 *arXiv:2405.14005* abs/2405.14005（2024），1–1。