



面向资源受限具身智能的视觉语言动作(VLA)模型综述

周毅 孙煜坤 石磊 程玉华 秦家虎

Vision-Language-Action Models for Resource-constrained Embodied Intelligence: A Survey

ZHOU Yi, SUN Yu-Kun, SHI Lei, CHENG Yu-Hua, QIN Jia-Hu

在线阅读 View online:

<http://www.aas.net.cn/article/current/85e91bf2c08829fc81c11b5a8a12b9e3de37f02408322693bf379760dd0329b4>

您可能感兴趣的其他文章

面向具身操作的视觉语言动作模型综述

Survey of Vision-Language-Action Models for Embodied Manipulation

自动化学报. 2026, 52(1): 18–51 <https://doi.org/10.16383/j.aas.c250394>

视觉语言动作模型综述: 从前史到前沿

Vision-Language-Action Models: From the Early Foundations to the State-of-the-Art

自动化学报. 2025, 51(9): 1922–1950 <https://doi.org/10.16383/j.aas.c250417>

基于大模型的具身智能系统综述

Embodied Intelligence Systems Based on Large Models: A Survey

自动化学报. 2025, 51(1): 1–19 <https://doi.org/10.16383/j.aas.c240542>

具身智能研究的关键问题: 自主感知、行动与进化

Key Problems of Embodied Intelligence Research: Autonomous Perception, Action, and Evolution

自动化学报. 2025, 51(1): 43–71 <https://doi.org/10.16383/j.aas.c240364>

具身智能自主无人系统技术

Embodied Intelligence Autonomous Unmanned Systems Technology

自动化学报. 2025, 51(4): 762–777 <https://doi.org/10.16383/j.aas.c240456>

大语言模型的工具使用综述

Survey of Tool Use in Large Language Models

自动化学报. 2025, 51(11): 2371–2386 <https://doi.org/10.16383/j.aas.c240793>

面向资源受限具身智能的视觉-语言-动作 (VLA) 模型综述

周毅¹ 孙煜坤¹ 石磊¹ 程玉华² 秦家虎³

摘要 视觉-语言-动作 (VLA) 模型是实现具身智能的重要技术路径. 针对现有方法普遍依赖大规模数据与高算力平台, 难以适应边缘侧等资源受限场景的瓶颈, 提出覆盖“数据-模型-训练-部署”全周期的优化框架. 在数据层面, 剖析真实采集、仿真生成与视频迁移三类路径的协同优化机制; 在设计层面, 梳理轻量化架构、状态空间模型、分层双系统、条件计算及流匹配等高效模型技术; 在系统层面, 重点探讨输入侧剪枝、模型压缩、动态推理路径及软硬件协同优化等部署加速策略. 最后, 对资源受限 VLA 模型的未来研究方向与应用前景进行展望.

关键词 具身智能; 视觉-语言-动作模型; 资源受限; 高效模型设计; 模型部署优化

引用格式 周毅, 孙煜坤, 石磊, 程玉华, 秦家虎. 面向资源受限具身智能的视觉-语言-动作 (VLA) 模型综述. 自动化学报, xxxx, xx(x): x-xx

DOI 10.16383/j.aas.c260051 **CSTR** 32138.14.j.aas.c260051

Vision-Language-Action Models for Resource-constrained Embodied Intelligence: A Survey

ZHOU Yi¹ SUN Yu-Kun¹ SHI Lei¹ CHENG Yu-Hua² QIN Jia-Hu³

Abstract Vision-language-action (VLA) models constitute a fundamental technical pathway toward embodied intelligence. However, existing approaches are hindered by their heavy reliance on large-scale data and high-performance computing platforms, which limits deployment in resource-constrained settings such as edge devices. To address this, we propose a comprehensive optimization framework that spans the entire development cycle: Data, model, training, and deployment. At the data level, we analyze synergistic optimization mechanisms across three primary pathways: Real-world collection, simulation generation, and video transfer; At the design level, we systematically review efficient model techniques including lightweight architectures, state-space models, hierarchical dual systems, conditional computation, and flow matching; At the system level, key deployment acceleration strategies are explored, such as input pruning, model compression, dynamic inference paths, and hardware-software co-optimization. Finally, we discuss future research directions and application prospects for resource-constrained VLA models.

Keywords embodied intelligence; vision-language-action models; resource-constrained; efficient model design; model deployment optimization

Citation Zhou Yi, Sun Yu-Kun, Shi Lei, Cheng Yu-Hua, Qin Jia-Hu. Vision-language-action models for resource-constrained embodied intelligence: A survey. *Acta Automatica Sinica*, xxxx, xx(x): x-xx

早期机器人系统普遍采用模块化设计, 即将感知、理解与控制等功能解耦为多个独立模块, 通过预定义接口实现交互与协作. 该架构在工程实现方

面虽有利于分工开发与系统集成, 但随着应用场景扩展到开放、动态的真实环境, 其局限性日益显现: 模块间的信息传递易产生语义损失, 导致机器人在非结构化场景下的泛化能力及控制稳定性不足. 为此, 研究者开始探索将感知、推理与执行紧密耦合的一体化架构.

近年来, 以 CLIP^[1]、BLIP-2^[2]、LLaVA^[3] 等为代表的大规模视觉-语言模型 (vision-language model, VLM) 在海量图文数据驱动下, 展现出强大的跨模态语义理解与迁移能力. 然而, 此类模型的推理过程与物理环境脱节, 缺乏实时交互与闭环反馈能力, 难以直接输出可执行的动作序列, 因此多被视为“离身智能”系统, 无法满足具身智能对感知-决策-执行一体化的需求.

在这一背景下, 构建基础模型 (foundation

收稿日期 2026-01-20 录用日期 2026-06-15

Manuscript received January 20, 2026; accepted June 15, 2026
河南省重大科技专项 (251000220200, 251000210300), 河南省重点研发专项 (251111211100, 261111241100) 资助

Supported by Science and Technology Major Project of Henan Province (251000220200, 251000210300) and Key Research and Development Project of Henan Province (251111211100, 261111241100)

本文责任编辑 刘华平

Recommended by Associate Editor LIU Hua-Ping

1. 河南大学人工智能学院 郑州 450046 2. 电子科技大学自动化工程学院 成都 611731 3. 中国科学技术大学自动化系 合肥 230027

1. School of Artificial Intelligence, Henan University, Zhengzhou 450046 2. School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu 611731 3. Department of Automation, University of Science and Technology of China, Hefei 230027

models, FMs) 驱动的新一代具身智能系统成为重要研究方向. 具身智能强调智能体通过与物理环境的实时交互来感知、学习和决策. 研究者尝试将大模型强大的语言与多模态推理能力嵌入机器人的感知-决策回路, 以提升其在开放环境中的泛化与交互能力. 作为实现上述构想的一种代表性路径, 视觉-语言-动作 (vision-language-action, VLA) 模型通过端到端学习框架, 实现从多模态指令到机器人动作序列的映射, 在提升语义理解能力的同时, 增强了机器人在真实场景中的交互与执行能力.

然而, VLA 模型在真实场景中的部署面临严峻的资源约束挑战: 它需在同一框架中融合视觉、语言与动作三类异构模态, 并完成从高层语义理解到底层运动控制的端到端映射. 这对训练数据的规模与质量、模型架构的计算效率以及系统部署的资源适配, 均提出极高要求.

在数据层面, 稀缺性构成制约模型性能的首要瓶颈. 与易于获取的互联网图文数据不同, 具身数据需依赖实体机器人在真实环境中采集, 成本高、周期长、可扩展性差. 尽管已出现如 Open X-Embodiment^[4] 等大规模数据集, 但数据分布仍存在明显局限: 人工示教多集中于常见操作, 导致数据呈现长尾分布, 且缺乏任务失败、异常恢复等关键场景样本. 此外, 现有数据多采集于实验室结构化环境, 任务单一, 难以支撑模型在开放环境中的泛化^[4-5]. 尽管仿真环境可提供低成本的数据扩展路径, 但其物理动力学与感知建模存在仿真到现实的差距, 导致数据可迁移性普遍受限.

在模型层面, 高复杂度带来沉重计算负担. 为保持强大的语义理解能力, VLA 模型通常以大规模 VLM 为骨干网络. 典型架构一般包含视觉编码器 (如 DINOv2^[6]、SigLIP^[7])、语言模型以及动作解码器, 参数量可达数十亿, 推理延迟常达数百毫秒, 带来巨大的计算与存储开销. 若为追求效率而进行模型裁剪, 则易导致泛化性能显著下降; 若维持大规模参数以保证能力, 又难以满足实时控制对低延

迟的严苛要求. 因此, 在模型规模、泛化能力与计算效率间取得平衡, 是核心挑战之一.

在系统层面, 巨大计算开销使 VLA 模型难以直接部署于资源受限的机器人平台. 受尺寸、重量与功耗 (size, weight, and power, SWaP) 的限制, 移动机器人通常无法搭载高性能计算单元; 同时, 不同平台的硬件接口、通信协议与坐标系差异, 也增加了模型与底层控制系统集成的复杂性. 云边端协同架构虽可提供额外算力支持, 但其网络延迟、数据安全及基础设施依赖性, 使其难以适用于高实时、高可靠的控制场景.

综上所述, 数据稀缺与计算受限已成为制约 VLA 模型在边缘侧部署的核心瓶颈, 并引发广泛研究. 近期已出现多篇综述从不同角度梳理此领域. 为明确本文定位, 表 1 对比代表性综述工作.

从表 1 可见, 尽管近期已有研究从处理流程^[11]或生命周期框架^[12]等角度对 VLA 模型的效率优化进行系统性综述, 但现有工作大多集中于算法层面的计算效率提升, 即如何从理论层面降低模型的浮点运算量或参数量. 现有综述虽涉及训练与推理优化, 却往往未能深入探讨边缘侧机器人所面临的严格物理约束. 与上述工作相比, 本文与已有综述存在明显区别: 本文提出一套覆盖从数据到部署全周期的技术体系, 并以资源受限为核心脉络进行组织, 系统探讨在严格物理约束下实现 VLA 模型部署的可行技术路径, 涵盖面向资源受限环境的多源数据获取与高效利用 (第 1 节)、原生高效模型设计与训练优化 (第 2 节)、部署阶段的模型适配与压缩以及软硬件协同优化机制 (第 3 节). 上述内容的整体组织框架如图 1 所示. 本文旨在为 VLA 模型在资源受限环境中的系统设计与工程部署, 提供一个更为深入和系统的分析框架.

本文的贡献如下:

1) 构建“数据-计算”双约束分析框架, 系统梳理资源受限条件下 VLA 模型的核心挑战与效率优化的关键技术路径;

表 1 VLA 模型相关综述对比
Table 1 Comparison of VLA model related surveys

综述	年份	研究主题	框架视角	主要贡献
Zhong 等 ^[8]	2025	动作表示机制	动作 Token 分类	提出 VLA 模型的动作 Token 分类框架
张慧等 ^[9]	2025	VLA 技术演进	历史与范式演化	回顾前 VLA 时代技术, 梳理三大建模范式
王文晟等 ^[10]	2025	大模型具身智能	四级控制层级	构建需求-任务-规划-动作四级控制框架
Guan 等 ^[11]	2025	高效 VLA 模型	处理流程视角	从处理流程角度系统分类高效 VLA 技术
Yu 等 ^[12]	2025	高效 VLA 模型	数据-模型-训练	构建数据-模型-训练的高效 VLA 框架
Ma 等 ^[13]	2026	具身智能应用	组件-策略-规划分类	建立了涵盖基础组件、低层控制与高层规划的分类体系
本文	2026	资源受限条件	模型全生命周期	聚焦资源受限条件下的具身智能, 涵盖数据-模型-训练-部署全周期技术体系

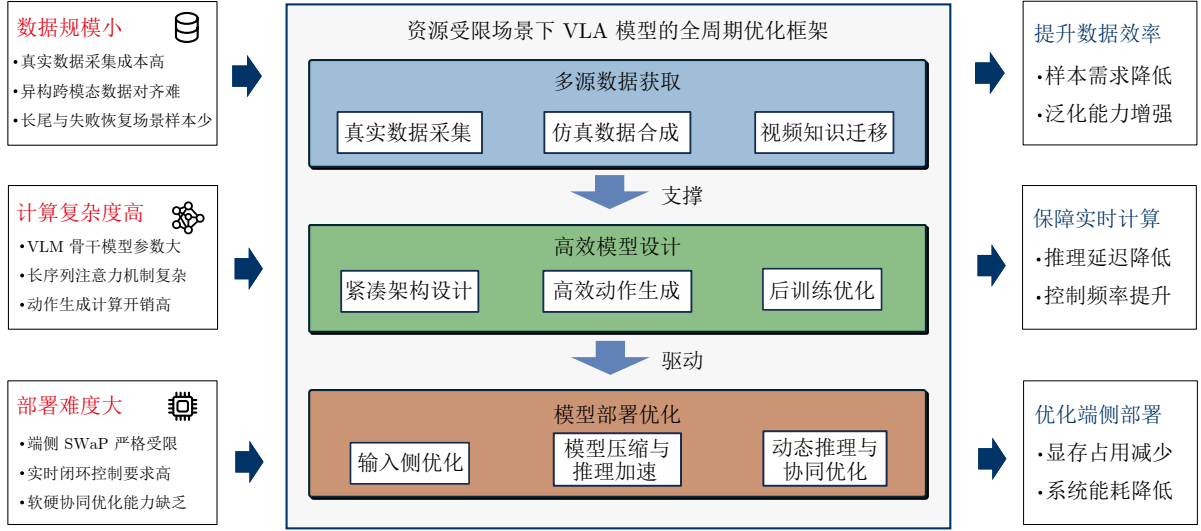


图1 资源受限场景下 VLA 模型的全周期优化框架

Fig.1 Full-lifecycle optimization framework for VLA models under resource-constrained settings

2) 提出覆盖从数据到部署的全周期技术体系, 深入阐释各环节核心方法的内在关联、协同机制及不同技术路线的适用边界;

3) 展望 VLA 模型发展的关键未来方向, 为资源受限条件下的后续研究与应用部署提供参考。

1 多源数据获取与高效利用

VLA 模型的训练依赖于结构化、多模态的具身交互数据, 这类数据需同步包含视觉观测、语言指令与机器人动作, 并在时序上严格对齐。与易于从互联网获取的图文数据不同, 具身数据的采集依赖于在真实或仿真环境中的主动物理交互, 导致其获取成本高昂、可扩展性受限, 构成制约模型性能的关键瓶颈。

VLA 模型面临的数据挑战, 可归纳为数据规模不足与数据结构异构两方面。从数据规模看, 大规模、高质量的具身数据极为稀缺。数据采集依赖实体机器人平台与复杂任务流程, 涉及平台部署运维、高质量人工示教以及多模态数据的精准对齐, 属于典型的资源密集型任务。从数据结构看, 数据异构性增加了利用难度。一方面, 不同机器人平台在硬件构型、传感器配置与坐标系定义等方面的差异, 使得跨平台数据难以统一; 另一方面, 自然语言指令与连续动作序列之间缺乏统一标注规范, 导致高层任务语义与底层控制信号之间的映射关系模糊。

面对上述问题, 现有研究从多方位拓展训练数据的来源。当前, 主要形成了三条互为补充的技术路径: 构建大规模真实交互数据集、利用高保真物理仿真生成合成数据以及从海量互联网视频中迁移

人类操作知识。如图 2 所示, 这三类数据在物理真实性与采集成本上各具优势, 形成互补。本节后续将依次对这三种路径展开深入剖析。

1.1 大规模真实数据采集

构建大规模真实交互数据集是应对 VLA 数据瓶颈的最直接路径。尽管采集成本高、周期长, 但真实数据在物理真实性、感知细节和交互复杂度方面具有不可替代的价值, 能够为模型提供最接近部署环境的训练样本。为此, 学术界与工业界通过制定标准化采集协议、开发自动化标注工具、建立跨机构共享机制, 逐步构建起多个具有代表性的大规模数据集。这些数据集在规模、任务多样性、场景覆盖度与标注质量上均实现了显著突破, 为 VLA 模型的训练与评测提供了不可或缺的样本基础。

早期工作以 RT-1^[14] 为代表。该项目基于 13 台同构移动机械臂, 在 17 个月内采集约 13 万条轨迹, 系统性验证了 Transformer^[15] 架构用于大规模真实世界机器人控制的可行性。但该数据集受限于一一的机器人硬件与高度结构化的室内场景, 导致以此训练的模型在跨平台与开放场景下的泛化能力受限。随后的 RH20T^[16] 则针对富接触复杂操作任务, 构建了规模超 11 万条的真实机器人操作序列数据集。该数据集同步采集了视觉、力觉、音频等多模态信号, 旨在为跨模态对齐及基于人类演示的技能迁移研究开辟新途径。不过, 该数据集的采集环境仍以受控的室内台面场景为主, 其场景多样性有待进一步扩展。此类由单一机构主导构建的数据集, 在数据规模与多模态采集上取得了显著进展。然而, 这类数据集在跨平台数据的标准化与可复用性方面

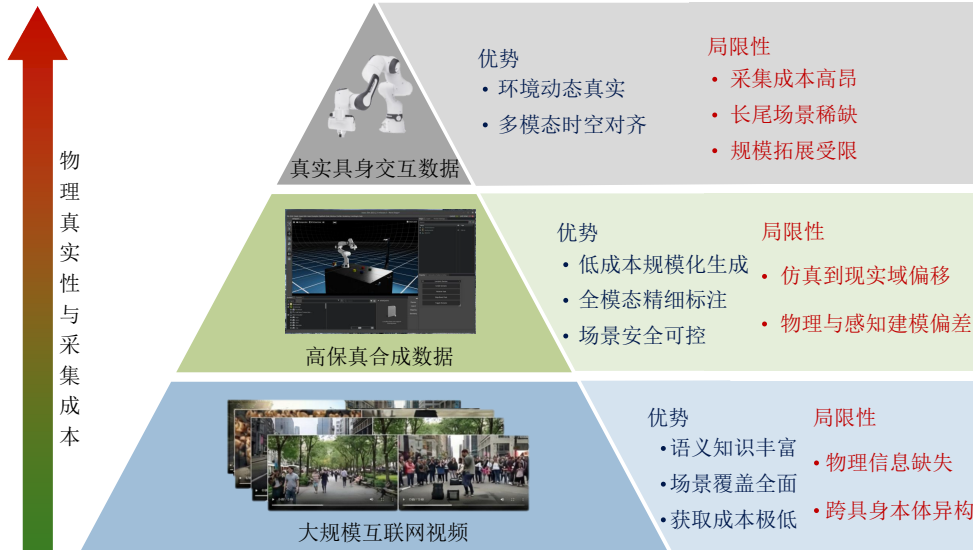


图 2 VLA 模型训练数据的三类来源及其特性对比

Fig.2 Comparison of three types of training data sources and their characteristics for VLA models

存在局限,难以支撑模型实现大规模、鲁棒的跨机器人本体泛化。

为此,研究界转向更开放的协作模式以实现规模化扩展,主要沿两条路径推进:一是通过跨机构协作整合异构平台数据;二是通过开源软硬件生态激发分布式社区采集. Open X-Embodiment^[4]是多机构、跨平台整合的代表性工作.该项目联合 21 家机构的 34 个实验室,整合来自 22 种异构机器人平台的超 100 万条轨迹,涵盖 500 余项技能,其关键贡献在于提出了统一的数据格式与标准化接口,使得来源各异的数据能够在单一模型中进行联合训练.基于此数据集训练的 RT-X 模型,相比仅在单一平台数据上训练的模型,在未见过的机器人平台上实现了 50% 的零样本成功率提升,有力地证明了跨平台数据整合对于具身智能体知识迁移的促进作用。

相较于研究机构主导的大规模整合,社区驱动的开源硬件与软件生态形成了另一条规模化路径.该路径通过大幅降低数据采集的软硬件门槛,推动了机器人数据生态的去中心化发展.在硬件方面,以 ALOHA^[17]与 UMI^[18]为代表的开源系统,利用商用元件与 3D 打印技术构建了低成本机器人平台,使得研究者和开发者得以在家庭、办公室等多样真实场景中便捷采集高质量交互数据.在软件与协作生态方面,LeRobot 项目通过提供标准化的工具链与数据规范,依托开源平台 Hugging Face 构建了社区数据仓库,极大促进了分布式数据集的共享与整合.基于这些公开数据训练并开源的高效模型 SmolVLA^[19](如图 3),成功验证了“开源硬件驱动

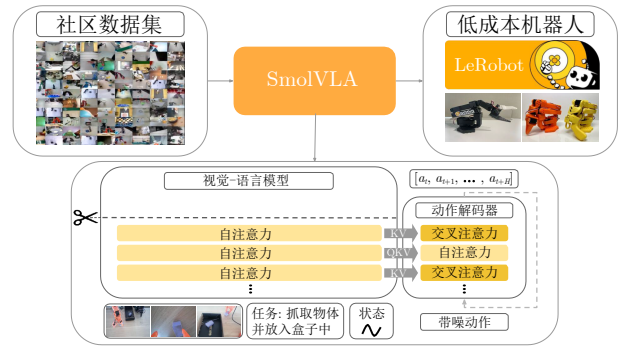
图 3 SmolVLA^[19] 开源生态与模型架构,经许可转载自文献 [19], ©arXiv, 2025

Fig.3 The open-source ecosystem and model architecture of SmolVLA, reproduced with permission from reference [19], ©arXiv, 2025

采集、社区平台汇聚数据、高效模型反哺应用”的协同范式,为资源受限下的研究提供了可扩展的新路径。

同时,数据集的构建正呈现出纵向深化的趋势,其核心在于提升数据的复杂度、感知精度与评估的系统性.代表性工作主要沿三个方向展开:其一,基于统一硬件的系统性长程采集,如 Galaxea Open-World^[20]与 AgiBot World^[21]. Galaxea Open-World 采集了 500 h 演示,覆盖 150 余项任务与 1600 多个物体;AgiBot World 规模更大,包含超百万轨迹、200 余项任务及失败恢复样本,为长程规划与错误恢复研究提供了系统性数据.其二,感知模态的精细化,如 Kaiwu^[22]同步采集手部运动、肌电、眼动等十一维信号,构建了万余精细实例,旨在探究

感知-动作协调机制. 其三, 面向特定能力的专项基准, 如 RoboCerebra^[23] 以平均近 3 000 步的超长轨迹评估规划与记忆; REASSEMBLE^[24] 则提供视觉-力觉-语言多模态装配数据, 建立接触式操作的专用评估标准. 当前, 真实数据虽已形成协同共建的多元生态, 但其采集受制于物理条件、经济成本与时间周期, 规模增长难以匹配 VLA 模型的海量需求, 且在长程决策、错误恢复等长尾场景的系统性覆盖上仍显不足.

然而, 真实数据在物理交互真实性、环境噪声表征及任务执行稳定性等方面具有不可替代的核心价值. 尤其是在工业装配、精密操作等对物理接触和动力学精度要求极高的场景中, 仅依靠合成数据往往难以建模真实的摩擦、形变与动态耦合等复杂效应. 因此, 构建高质量的真实数据集, 始终是 VLA 模型在关键场景中可靠落地的基石.

1.2 高保真仿真数据生成

尽管真实数据具有上述核心价值, 其采集成本高昂且规模扩展受限, 促使研究者积极探索替代性数据来源. 物理仿真为此提供了一条可扩展的补充路径. 仿真技术通过显式建模物理规律与环境动态, 能够在虚拟环境中生成大规模、高度可控的合成训练数据. 与真实数据采集相比, 仿真数据具备三方面优势: 在可扩展性上, 程序化生成与并行计算可以低成本产生海量数据, 突破硬件与人力限制; 在可控性上, 研究者可调整光照、物体属性与任务配置, 覆盖真实世界难以复现的长尾场景; 在数据完整性上, 仿真环境天然包含物体姿态、接触力等物理真值, 无需人工标注即可获得完整的多模态监督信号.

仿真数据的训练价值最终取决于其向真实场景迁移的有效性. 这要求仿真平台在视觉保真度、物理准确性、数据多样性与生成效率等多个维度上达到较高水平. 近年来, 神经渲染、高效物理引擎与智能化任务生成等技术的快速发展, 为构建高质量仿真环境奠定了坚实基础. 在此推动下, 面向 VLA 训练的仿真平台正朝着视觉高保真、多模态融合与任务智能化生成的技术体系演进.

视觉渲染是构建高保真仿真环境的基础. 传统光栅化方法效率较高, 但物理精度有限, 难以真实模拟复杂的光照与材质交互. 物理级光线追踪虽能显著提升渲染质量, 但计算开销较大, 在追求高保真度时往往需要牺牲生成效率. 神经辐射场 (neural radiance fields, NeRF)^[25] 作为神经渲染的代表性工作, 能够从稀疏图像重建较为真实的三维场景,

但其隐式体积表示与体渲染过程计算成本高昂, 难以满足实时交互与大规模数据生成的需求. 三维高斯泼溅 (3D Gaussian splatting, 3DGS)^[26] 的提出突破了这一瓶颈. 该方法采用显式可微分的高斯基元作为场景表示, 在保持高视觉质量的同时实现了实时渲染. 这一技术进步随后被机器人领域采纳, 为高通量、高保真的仿真数据生成提供了新的技术途径.

基于 3DGS 等先进渲染技术, 一系列仿真平台致力于构建高保真、可交互的数字孪生环境. RoboGSim^[27] 利用高斯重建技术从真实场景恢复几何与外观, 构建可交互的数字孪生, 形成了“Real-to-Sim-to-Real”的完整闭环. Physically Embodied Gaussians^[28] 将 3DGS 渲染与基于位置的动力学模型结合, 并通过“视觉-力”反馈实时修正物理状态, 在 30 Hz 下保持视觉与物理的一致性. Re³Sim^[29] 提出了一套高保真实景仿真重建流程, 旨在通过高质量重建与神经渲染来弥合仿真与真实世界之间的几何与视觉差异. 其研究表明, 仅依靠仿真数据训练的模型, 能够有效迁移到真实机器人平台并完成操作任务.

面对仿真环境与真实世界之间不可避免的域偏移问题, 增强数据的多样性成为提升模型泛化能力的关键策略. 域随机化方法通过程序化扰动光照、纹理、姿态及动力学参数, 有效拓宽数据分布. 近年来, 研究进一步向语义化、自动化的任务生成深化. MimicGen^[30] 以少量人类演示为基础, 通过几何空间变换与场景重组自动生成大量语义一致但配置多样的任务实例, 无需额外标注即可提升数据多样性. GenSim2^[31] 框架利用多模态大语言模型自动提出机器人操作任务、生成仿真场景代码、调用规划求解器并产生演示数据, 实现了从任务定义到数据生成的端到端自动化. RoboCasa^[32] 则进一步将生成过程扩展至环境层面, 借助生成式模型自动合成物体资产与室内布局, 构建出包含百余种房间配置和数千个可交互物体的复杂家居仿真平台, 从而在视觉与物理层面增强了任务场景的丰富性. DemoGen^[33] 聚焦于动作轨迹的泛化, 通过对原始演示轨迹进行运动路径重规划, 使其能够适应新的障碍布局与目标位置, 在复杂约束下提升策略的适应能力.

伴随数据多样性增强技术的成熟, 底层仿真平台的规模化生产能力成为决定其实用价值的关键因素. 近年来, 底层物理引擎与分布式架构的创新提升了数据生产效率. Genesis^[34] 通过高度优化的 GPU 并行, 实现了每秒千万帧量级的数据生成, 其系统集成生成式代理以自动构建交互场景. Isaac Lab^[35] 进一步将仿真、训练与数据管理深度集成, 可在单张 GPU 上并行运行上千个环境实例, 为大规模强

化学习 (reinforcement learning, RL) 与模仿学习提供了统一的高通量平台. ManiSkill3^[36] 设计了完全在 GPU 上运行的仿真与渲染管线, 在保持高质量物理仿真的前提下, 在复杂任务中实现了每秒超 3 万帧的高效生成, 为复杂操作场景的持续学习提供了高效数据源.

高保真、大规模的仿真数据为 VLA 模型提供了丰富的视觉-动作样本, 但复杂操作任务往往需要视觉、触觉、力觉等多模态信息的协同. 在接触密集型任务中, 机器人需要同时感知多维信号, 以在动态约束和物理交互下实现稳定控制与自适应操作. 为此, 现代仿真平台正构建完整的多模态感知生态系统. Isaac Sim 等系统已支持 RGB、深度图、语义分割、法线图等多通道同步输出, 可生成高密度结构化视觉数据. ManiSkill3^[36]、Sapien^[37] 等平台支持生成点云与体素网格, 强化了几何表征能力. 在物理模态方面, Tacel^[38] 专注于为视觉式触觉传感器提供高性能仿真, 能以数千个并行实例生成触觉图像, 为视觉-触觉跨模态学习提供了规模化数据基础. 上述技术进步使仿真数据的实用价值得到验证. Sim-and-Real Co-Training^[39] 表明, 仿真与真实数据的协同训练可带来显著的性能增益. GraspVLA^[40] 通过十亿帧仿真抓取数据与互联网语义的联合训练, 展现了显著的零样本泛化潜力. RoboGSim^[27] 基于高斯重建的合成数据, 在驱动真实机器人时取得了与真实数据相当甚至在未知视角与布局下更具优势的性能.

尽管如此, 物理仿真依赖显式的环境建模与参数配置, 在处理复杂开放域场景时面临建模成本高、迁移效果受限等问题. 互联网上海量的人类操作视频包含丰富的场景与任务知识, 且无需显式建模即可获取, 为突破上述局限提供了新的数据来源.

1.3 多源视频驱动的知识迁移

相较于需要主动交互的真实采集与仿真生成, 海量的互联网视频以极低成本提供了无与伦比的任务多样性与场景覆盖. 然而, 将视频知识迁移至机器人面临两大核心挑战: 一是动作空间的缺失与异构, 视频通常缺乏动作标注, 且人机在形态、视角与动作空间上存在根本差异; 二是环境动态的隐式性, 其中蕴含的物理规律与因果结构需被显式提取才能指导策略. 为应对这些挑战, 当前研究形成了两条互补的技术路径: 其一, 通过世界模型学习环境动态, 在潜在空间中构建可推演的交互模型; 其二, 通过端到端迁移, 直接建立从人类演示到机器人控制的映射. 前者重在理解“世界如何运作”, 后者重在

模仿“人类如何操作”, 共同构成了视频驱动的技能学习体系.

1.3.1 世界模型学习与应用

世界模型旨在以数据驱动的方式学习环境动态与因果结构, 能够在潜在空间中预测未来或生成轨迹, 从而构建出用于内部推演的“想象空间”^[41]. 这一技术最初为提升强化学习的样本效率而发展, 近年来被引入视频数据的学习场景. 由于世界模型无需人工显式设定环境物理参数, 而是从数据中自监督学习环境动力学, 因此适合处理仿真难以显式建模的复杂动态系统, 为 VLA 模型的数据生成与策略优化提供了有效途径.

利用世界模型从无标注视频中学习动作表示, 是缓解标注瓶颈的重要方向. UniVLA^[42] 在 DINO 特征空间中建立任务中心的潜在动作模型, 通过结合语言指令过滤任务无关的场景动态, 实现跨具身平台的动作表示迁移. 该方法将预训练计算量降至 OpenVLA 的 1/20、下游数据需求降至 1/10, 在 LIBERO 基准上超越 OpenVLA 18.5 个百分点. 另一类工作侧重于构建生成式世界模型, 以显式理解环境动态来辅助决策. Genie^[43] 在大规模互联网视频上无监督学习, 能从游戏画面中自动发现潜在动作并形成可交互的生成模型, 证明了从无标注视频学习可控环境模型的可行性. 后续研究进一步探索世界模型与动作生成的深度融合. WorldVLA^[44] 将二者统一于自回归 Transformer 框架, 使世界模型与动作模型相互增强; 3D-VLA^[45] 将其扩展至三维空间, 支持基于点云的预测与规划; Genie Envisioner^[46] 构建了统一的视频生成平台, 实现了策略学习、仿真与评估的集成; DreamVLA^[47] 则以更轻量化的方式, 将场景动态预测作为辅助任务引入训练, 使模型能隐式考虑动作后果. 此外, NVIDIA Cosmos^[48] 作为开源的世界基础模型平台, 可从有限真实轨迹进行光照变换、跨环境迁移等可控数据增强, 自动生成多样化训练样本.

除了与动作生成耦合, 世界模型的另一重要价值在于构建低成本的虚拟交互环境——智能体在这一安全、可重复的“想象空间”中进行试错与规划, 再将习得的策略转移至物理世界, 从而规避了直接真实交互的成本与风险. 一系列研究验证了该范式的可行性与优势. World-Env^[49] 从离线视频学习动态模型, 并引入视觉语言模型提供奖励信号, 仅基于五个演示就显著提升了 VLA 的泛化能力. Day-Dreamer^[50] 将这一思路扩展至在线学习, 通过交替进行少量真实环境交互与大量模型内想象训练实现高效策略学习, 使四足机器人在约 1 h 内学会行走,

展现了其在加速真实世界学习方面的潜力. Dreamer v4^[51] 则在该范式下实现了超长程规划, 在《我的世界》中自主完成了需数万个动作的复杂任务链, 证明了其对复杂长期推理的支持能力.

为了将人类视频中的知识迁移至形态各异的机器人, 研究进一步探索了跨本体表征. VLWM^[52] 将世界动态投射到语言空间, 利用大语言模型从互联网视频中提取结构化的轨迹表示, 以支持不同的规划模式. SWIM^[53] 则通过引入视觉可供性这一以物体为中心的表征, 弥合了人机在形态与运动方式上的差异, 使得仅需约 30 min 的真实机器人交互即可完成技能迁移.

1.3.2 人类操作模仿与迁移

与世界模型关注环境动态不同, 人类操作视频驱动的技能迁移旨在直接建立从人类演示到机器人控制的映射. 该路径通过解决跨本体、跨视角与跨动作空间的对齐问题, 力求实现人类技能的高保真迁移, 而无需构建环境中间表征.

早期研究侧重于从视频中学习通用的视觉操作表征. R3M^[54] 及 MVP^[55] 通过在大规模人类第一视角视频上进行预训练, 提取与操控相关的低维特征, 这些预训练表征随后被应用于行为克隆和强化学习任务, 显著提升了样本效率. MPI^[56] 引入交互预测机制, 将操作动态显式建模, LAPA^[57] 则通过 VQ-VAE^[58] 将视频帧间动态压缩为离散潜在动作 Token, 在无动作标注的条件下实现预训练并高效迁移. 这一系列工作奠定了从人类视频中学习通用感知与动作表征的基础. 随着表征学习方法的成熟, 用于机器人模仿学习的视频数据逐步从受控环境中的第一人称演示, 拓展至第三人称视角及互联网规模的开放环境视频. WHIRL^[59] 通过从真实世界人类操作视频中提取人物交互先验, 实现了跨视角的机器人模仿学习. VideoMimic^[60] 证明了手机随手拍摄的视频也可作为可扩展的技能学习数据源. RIGVid^[61] 则尝试利用视频生成模型构建虚拟演示, 通过 6D 姿态追踪提取显式的物体运动轨迹.

近期, 研究开始探索将人类视频经验直接融入 VLA 模型的端到端训练. 根据对操作知识的提取方式, 主要形成显式建模与隐式学习两类技术路线. 显式建模方法通过构建人类动作的结构化表示建立精确的跨本体映射. EgoVLA^[62] 基于 MANO^[63] 手部参数模型构建统一动作空间, 将人类视频中的手部动作编码为标准化的手腕姿态与关节参数, 再通过逆运动学与动作重定向技术实现跨本体对齐. Being-H0^[64] 则专注于灵巧操作, 在包含 1.65 亿运动-指令对的 UniHand 数据集上, 采用分组残余量

化对手指运动进行毫米级精度的建模与迁移. 隐式学习不依赖显式标注与对齐, 而是通过大规模视频生成预训练, 使模型内隐地掌握操作相关的视觉动力学. GR-2^[65] 模型通过在 3 800 万互联网视频上进行生成式预训练, 直接从开放域数据中学习视觉动力学与场景理解, 从而提升开放世界任务的泛化能力. RynnVLA-001^[66] 在 1 200 万第一人称操作视频上进行类似预训练, 并通过预测人体关键点轨迹来桥接视觉观测与动作空间, 再结合 ActionVAE 获得紧凑的动作表示. Gen2Act^[67] 则利用预训练视频生成模型合成人类操作演示作为视觉提示, 通过轨迹预测提取运动信息, 并训练一个轻量策略将其转化为机器人动作, 展示了生成式视频在引导模仿学习中的潜力.

综上所述, 视频驱动的知识迁移为 VLA 模型开辟了一条低成本获取大规模训练数据的新路径. 世界模型与人类操作迁移两类方案从不同角度应对了视频数据缺乏动作标注的核心挑战: 前者通过学习环境动态规律在潜在空间中构建可推演的交互模型, 后者通过跨本体对齐技术将人类操作经验直接映射为机器人控制信号. 两类方案各有侧重, 但也面临共同的挑战: 视频作为被动观察数据, 缺乏力学反馈、接触信息与物理约束, 这使得从视频学习的策略在接触密集型任务中的可靠性仍有待验证. 因此, 如何在保持视频数据规模优势的同时, 有效弥补其物理真实性的不足, 是该技术路线进一步发展需要解决的关键问题.

1.4 资源约束下的数据高效学习

当模型需要部署于资源受限的平台时, 数据质量的影响会变得尤为关键. 现有研究表明, 观测与执行之间的时序失配、推理延迟以及动作执行中的闭环反应性不足, 会显著影响动态操作任务中的策略稳定性与成功率. DynamicVLA 指出, 动态操作对快速感知、时序预判与连续控制提出了更高要求^[68]. RTC 与 A2C2 也进一步表明, 推理延迟会削弱动作分块策略的反应性, 并在执行边界处带来停顿或不稳定动作^[69-70]. 因此, 在资源受限场景下, 数据采集与策略训练需要更加重视时序一致性.

除了时序对齐, 示教轨迹本身的质量同样关键. Sakr 等^[71] 研究发现, 示教的一致性显著影响学习效果与泛化成功率, 并给出了路径长度、加加速度、曲率、能耗等一组可量化示教质量的指标. Belkhale 等^[72] 则指出, 状态多样性并非总是有益的, 影响数据质量的关键因素更应归结为动作发散度与转移多样性; 该文同时提到, 示教中的停顿或其他动作噪

声会增加学习难度. 这意味着, 对于资源受限的模型而言, 单纯增加数据规模并不一定带来更好的性能, 提升轨迹的一致性、平滑性以及动作分布与策略表征之间的匹配程度更为重要. 另一方面, 失败轨迹也并非毫无价值, 若经过质量评估与筛选, 保留其中的高质量片段, 失败数据同样可用于扩充训练集并提升下游性能^[73]. 此外, 采用课程学习方式组织数据, 也被证明能够提高最终成功率与样本效率^[74].

1.5 代表性数据集对比

在系统梳理了真实采集、仿真生成与视频迁移三类数据获取路径之后, 本节对当前代表性数据集进行横向对比, 以呈现不同数据策略在规模、模态与应用场景等维度上的具体差异.

如表 2 所示, 当前数据集生态呈现出明显的分层结构特征. 在规模维度上, 真实数据已实现从十万级到百万级的跃升, 混合数据通过整合策略达到百万级以上, 而人类视频数据则以时长计量, 反映

出不同数据类型在量化标准上的差异; 在模态配置上, 数据集根据任务特性呈现出差异化选择: 接触密集型任务倾向于集成功觉、触觉与事件相机等物理感知模态, 而通用操作任务则以视觉-语言-动作三元组为主; 在功能定位上, 跨平台整合型数据集 (如 Open X-Embodiment^[4]、ARIO^[77]) 支撑通用能力预训练, 专项基准型数据集 (如 LIBERO^[86]、RE-ASSEMBLE^[24]) 聚焦特定技能评测, 形成了“预训练-微调”的两阶段训练生态.

尽管如此, 现有数据集在覆盖度与质量上仍存在结构性不足. 在时间尺度上, 绝大多数数据集聚焦于分钟级短期任务, 缺乏对小时级及以上长程交互过程的系统记录; 在失败样本方面, 除 AGIBot World^[21] 等少数数据集外, 大部分仅包含成功轨迹, 限制了模型的错误恢复与鲁棒性学习能力; 在地理与场景多样性上, 数据采集高度集中于特定研究机构与结构化环境, 开放域、非结构化场景的代表性不足. 这些局限既反映了当前数据采集的技术与成本约束, 也指明了未来数据集建设需要突破的方向.

表 2 面向 VLA 模型训练的代表性数据集总结
Table 2 Summary of representative datasets for VLA model training

分类	数据集	来源	规模	模态	主要特点
真实机器人操作数据	RT-1 ^[14]	真实	130 K	V, L, A	13 台机械臂采集, 覆盖 700 余项任务
	BridgeData V2 ^[75]	真实	60 K	V, L, A, D	5 万条遥操作, 1 万条自动策略轨迹
	Open X-Emb. ^[4]	真实	1 M+	V, L, A	21 家机构, 22 种异构机器人, 统一数据格式
	DROID ^[5]	真实	76 K	V, L, A, D	564 个开放场景, 分布式遥操作采集
	RH20T ^[16]	真实	110 K	V, L, A, T, F, S, D	11 维多模态同步, 147 项接触密集型任务
	AGIBot World ^[21]	真实	1 M+	V, L, A, T	217 项任务, 含失败恢复轨迹与子任务标注
	RoboMIND ^[76]	真实	107 K	V, L, A, D	4 种平台, 479 项任务, 含失败轨迹
真实-仿真混合数据	ARIO ^[77]	混合	3 M+	V, L, A, T, S	258 种机器人, 32 万项任务, 统一数据格式
	MimicGen ^[30]	混合	50 K+	V, A	200 条演示自动扩展至 5 万条轨迹
	RoboCasa ^[32]	仿真	100 K	V, A	大规模厨房场景仿真框架, 基于人工示教生成轨迹
	NVIDIA Physical AI ^[78]	混合	320 K	V, L, A, D, P	基于 OpenUSD 的工业级仿真资产库
	RoboData ^[79]	混合	500 K	V, L, A, D, C	整合 9 个开源数据集, 统一评测接口
人类操作视频数据	Ego4D ^[80]	视频	3 670 h	V, L, S, E, I, D, P	74 个地点, 3 670 h 第一人称视频
	EPIC-K-100 ^[81]	视频	100 h	V, L, S	厨房场景, 97 类动作, 9 万个片段
	EgoExo4D ^[82]	视频	1 286 h	V, L, E, S	第一/第三人称同步, 8 个技能领域
	HOT3D ^[83]	视频	13.88 h	V, E, P	多视角 3D 手物交互标注
	TACO ^[84]	视频	2.5 K	V, D, P	2 500 条双手工具操作, 高精度 3D 标注
	HoloAssist ^[85]	视频	166 h	V, L, E, I, S	7 个模态同步, 人机协作场景
	UniHand ^[64]	视频	150 M	V, L, A	采用动捕、VR、RGB 采集, 涵盖超过 1.5 亿运动指令对
专项任务数据	LIBERO ^[86]	仿真	6.5 K	V, L, A	程序化生成机器人操作任务, 提供遥操作示教
	REASSEMBLE ^[24]	真实	4.5 K	V, L, A, F, S, C	涵盖事件相机与力觉采集, 接触密集装配任务
	RoboCerebra ^[23]	仿真	1 K	V, L, A, D	平均 2 972 步超长轨迹, 规划能力评测
	BRMData ^[87]	真实	约 500	V, A, D	双臂移动平台, 10 个家居场景任务

注: V = 视觉, L = 语言, A = 动作, D = 深度, T = 触觉, F = 力觉, S = 音频, E = 眼动, I = IMU, C = 事件相机, P = 3D 姿态/点云; 机器人数据多以轨迹/样本数计, 人类操作视频数据多以时长计, 部分数据集以序列数或样本数计.

2 高效模型设计与训练优化

第 1 节探讨了 VLA 模型的数据获取策略, 从真实采集、仿真生成到视频迁移的多源路径为模型训练提供了样本基础. 除数据瓶颈外, 计算效率是制约 VLA 模型实际部署的另一关键挑战. 当前主流 VLA 模型继承预训练 VLM 的架构, 参数量常达数十亿, 导致单次推理需要数十 GB 显存和数百毫秒延迟, 这与移动机器人平台严格的 SWaP 约束形成了根本性矛盾. VLA 模型的计算瓶颈是系统性的, 体现在多个方面. 在模型规模上, 数十亿参数的 VLM 骨干导致显存占用和计算量过高; 在序列处理上, Transformer 的自注意力机制在处理长序列时面临二次复杂度瓶颈, 高帧率视频流和密集图像块使计算开销急剧增长; 在计算分配上, 现有模型对所有输入采用统一的全量计算路径, 未能根据任务复杂度自适应调整; 在动作生成上, 主流的自回归离散解码在精度上受限于量化误差, 在效率上受制于串行延迟.

针对上述问题, 本节探讨如何在模型构建阶段, 通过架构创新, 从源头降低模型的理论复杂度与参数规模. 本节主要从架构设计与训练优化两个层面展开探讨. 在架构设计方面, 现有研究形成了以下

主要技术路径: 轻量化架构降低模型规模、高效序列建模突破复杂度瓶颈、分层架构实现功能解耦、条件计算实现资源自适应分配、动作生成创新解决精度与延迟问题. 在训练优化方面, 重点关注参数高效微调以降低训练代价, 并结合强化学习提升数据利用率. 如图 4 所示, 本节与第 3 节将分别系统阐述模型在设计阶段与部署阶段的核心优化方法与典型应用场景. 表 3 汇总了代表性 VLA 模型在参数规模、训练开销与推理性能上的关键指标, 可作为后续各节讨论的参照.

2.1 轻量化模型架构

大规模 VLA 模型的计算瓶颈首要源于其庞大的参数规模. 采用轻量级预训练 VLM 作为主干是降低开销的直接途径, 但如何在压缩规模的同时保持复杂具身任务所需的空間推理、物理理解及精确控制能力, 仍是核心挑战. 尽管 Phi-2^[96]、LLaVA-Phi^[97] 等工作已证明小模型在通用领域的潜力, 但将其应用于机器人控制仍需专门解决动作生成精度、实时性及物理交互鲁棒性等关键问题. 现有研究从三个角度探索轻量化 VLA 架构.

第一类工作主要聚焦于动作生成机制的优化与

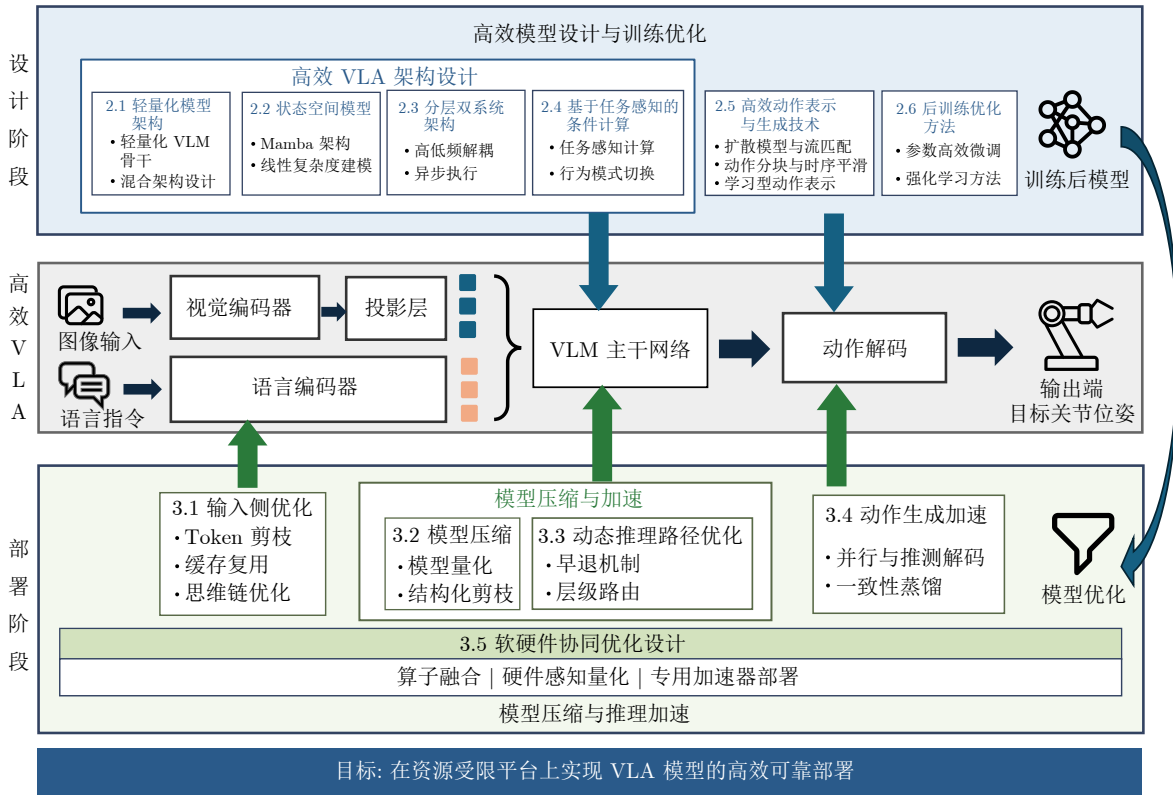


图 4 高效 VLA 模型的设计与部署优化技术框架

Fig. 4 Technical framework for efficient VLA model design and deployment optimization

表 3 主流 VLA 模型关键指标对比
Table 3 Comparison of key metrics for mainstream VLA models

模型	参数量	主要训练数据规模	训练开销		推理性能	
			训练耗时	步数/Epoch	频率 (Hz)	推理平台
RT-1 ^[14]	35.00 M	13 万条机器人轨迹	—	20 万步	3	TPU v4
RT-2 ^[88]	55.00 B	10 亿对图文与机器人数据	—	8 万步	1 ~ 3	TPU v5
Octo ^[89]	93.00 M	80 万条混合轨迹	0.3 万 TPU h	30 万步	30	RTX 4090
OpenVLA ^[90]	7.00 B	97 万条机器人轨迹	2.15 万 A100 h	27 轮次 ^d	6	RTX 4090
π_0 ^[91]	3.30 B	1 万小时以上机器人数据	—	70 万步	≤ 50	RTX 4090
GR00T N1 ^[92]	2.20 B	5.9 亿帧多源混合数据	5 万 H100 h	20 万步 ^a	10/120 ^b	L40
GraspVLA ^[40]	1.80 B	10 亿帧合成抓取数据	—	12 万步	5	L40s
RoboMamba ^[93]	3.20 B	31 万对图文与机器人数据	—	5 万步	9	A100
UniVLA ^[42]	8.50 B	62.2 万条多源视频	960 A100 h	2 万余步	10	RTX 4090
SmolVLA ^[19]	0.45 B	2.3 万条开源轨迹	3 万 GPU h	20 万余步	30	消费级平台
TinyVLA ^[94]	1.30 B	500 条微调轨迹	—	—	71 ^c	A6000
FLOWER ^[95]	0.95 B	25 万条混合轨迹	200 H100 h	36 万步	311	RTX 4090

注: a 代表预训练步数; b 代表 10 Hz (视觉语言模块)/120 Hz (动作策略模块); c 代表原文指出单次推理延迟 14 ms, 此处为换算得出的理论频率; d 代表原文明确指出训练轮次为 27 Epochs, 未披露具体训练步数; “—”表示数据未披露。

效率提升. TinyVLA^[94] 采用小规模 VLM 配合扩散解码器, 通过扩散策略解码器替代自回归机制, 在连续空间直接建模避免量化误差, 单次推理延迟达到 14 ms. DiffusionVLA^[98] 进一步探索了自回归推理与扩散动作的深度融合, 其 DiVLA-2B 模型通过推理注入模块将 VLM 的语义理解动态嵌入扩散策略头, 在 20 亿参数规模下达到 82 Hz 的推理速度. FLOWER^[95] 则从参数分配角度进行创新; 该工作发现 VLM 后层对文本任务存在过度特化现象, 基于这一发现, 通过剪枝语言模型后半部分并将节省的参数重新分配给 18 层的流匹配解码器, 在不足 1 B 参数下实现了 311 Hz 的高频推理.

第二类工作则针对模型进行系统性轻量化设计. SmolVLA^[19] 通过三层协同优化实现全栈轻量化: 感知端压缩视觉 Token 降低冗余, 融合端构建交错注意力机制平衡效率与表达力, 推理端实现感知与动作的异步解耦避免计算阻塞. 该方法以 4.5 亿参数达到与 7B 模型相当的性能, 证明了通过识别不同模块的计算特性并针对性优化能够实现整体效率的较大提升.

第三类工作通过轻量策略网络保持 VLM 完整性. VLA-Adapter^[99] 冻结预训练 VLM, 仅训练 5 亿参数量的策略网络, 在无需机器人数据预训练的条件下超越 OpenVLA^[90], 推理吞吐量达 200 Hz 以上, 为快速部署现有 VLM 提供了实用方案.

综上所述, 通过重构动作解码器、协同设计轻量架构及构建高效适配器, 轻量化 VLA 模型在参数规模、推理速度与任务性能间取得了有效平衡. 然而, 这些方法在压缩规模时, 可能难以完全保留

大模型中涌现的复杂推理能力. 此外, 其序列编码模块仍普遍依赖 Transformer, 在处理长视频、高帧率输入时面临二次复杂度的根本性瓶颈, 这引出了对更高效序列建模方法的需求.

2.2 状态空间模型

状态空间模型 (state space models, SSM) 通过隐状态的递归更新建模序列依赖, 实现线性时间复杂度. Mamba 架构^[100] 引入选择性机制, 通过动态调整状态转移矩阵, 使模型能够根据输入内容自适应地保留或遗忘历史信息. 该特性在处理高帧率视频、密集图像块以及长历史动作序列等场景时, 相比 Transformer 的二次复杂度提供了显著的效率优势. 近年来, SSM 在多模态领域取得了一系列进展. VL-Mamba^[101] 将二维图像转换为一维序列进行建模, Video-Mamba^[102] 扩展至视频级长上下文处理, Mamba-Adapter^[103] 通过轻量级集成实现上下文扩展. 这些工作验证了 SSM 处理多模态输入的可行性.

基于上述进展, RoboMamba^[93] 将 Mamba 架构引入 VLA 建模. 该方法通过轻量级投影将视觉特征映射至语言空间, 并采用 Mamba 单元替代自注意力机制. 其具体架构与训练流程如图 5 所示. 为提升适配效率, 仅在解码端集成 3.7 M 参数的策略头, 可通过单 GPU 快速微调. RoboMamba-2.8B 的控制频率达到 9 Hz, 较 OpenVLA^[90] 实现了显著提升. 在长序列处理中, 线性复杂度带来的显存优势尤为明显.

凭借线性复杂度的序列建模特性, SSM 在处理长序列时展现出显著的效率优势. 然而, 其选择性

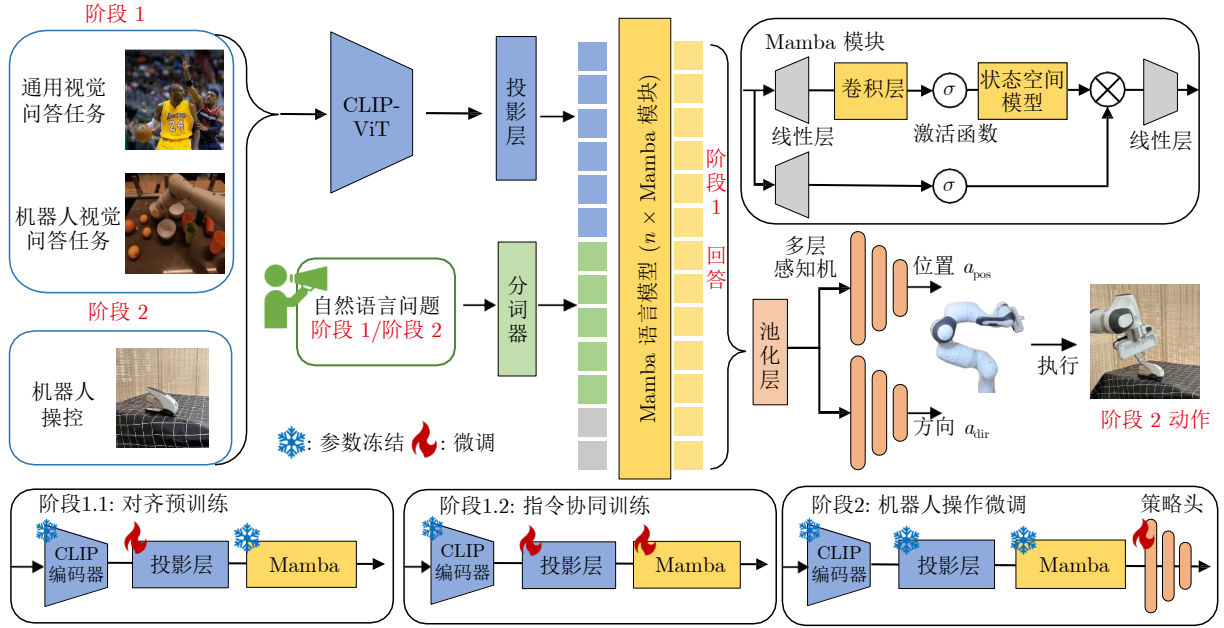


图5 RoboMamba 模型架构与训练流程示意图

Fig.5 Schematic diagram of the RoboMamba model architecture and training pipeline

机制能否有效捕获具身场景中复杂的空间关系与物理因果联系, 目前仍是一个开放的科学问题. 与 Transformer 通过全局注意力显式建模任意位置依赖关系不同, SSM 依赖有限维度的隐状态进行递归压缩以保存历史信息. 对于需要精确推理三维空间几何关系的操作任务, 远距离的空间依赖可能在递归压缩过程中被逐步稀释, 进而导致建模精度下降. 现有研究表明, 在需要精确上下文内检索与长程精确回溯的任务上, Mamba 相较 Transformer 仍存在一定差距^[104-105], 这一局限在高精度具身操作场景中尤需关注. 此外, SSM 依赖隐式的记忆保持机制来处理长时依赖^[100], 虽在效率方面有所增益, 但在精确性方面也引入了不确定性. 上述潜在的精度损失有待更系统的消融实验加以量化验证. 当前研究仍处于初步探索阶段, SSM 在具身智能领域的适用边界尚需更深入的理论分析与实证支撑.

2.3 分层双系统架构

当前多数机器人控制模型仍采用端到端的统一架构, 所有组件以相同的推理频率执行计算. 对于 VLA 模型, 其高层语义理解通常计算复杂且更新缓慢, 而底层动作生成则计算相对简单但需要高频输出. 端到端统一架构难以同时满足两者要求, 制约了系统的整体性能. 作为一种有效的解决思路, 分层双系统架构将 VLA 模型解耦为高层理解模块和底层执行模块, 并通过异步计算机制实现不同模块的协同工作, 以期解决大模型的实时控制难题. 但

模块间如何高效通信仍存在挑战, 研究者对此提出了多种解决方案.

HiRT^[106] 与 LCB^[107] 从异步解耦的基本思想出发, 探索了模块间通信的实现方法. HiRT 采用低频更新 (2 Hz) 的理解模块提取语义特征并缓存, 轻量级执行模块则以较高频率 (20 Hz) 读取缓存并生成动作, 有效避免了大模型推理延迟对实时控制的限制, 在 Metaworld 任务中将成功率从 48% 提升至 75%. LCB 则引入可学习的连续潜在编码作为模块间通信媒介, 替代传统的语言指令接口, 在保持语义表达能力的同时将推理时间缩短约 25%. RoboDual^[108] 进一步探索了异构策略之间的协同机制, 将通用视觉-语言-动作策略 (3.9 Hz) 与专用扩散策略 (15 Hz) 通过异步结构相结合, 在参数数量为 2 000 万的设定下实现了控制频率提升 3.8 倍. 其模块间采用直接输出传递方式, 与 HiRT 的特征缓存机制形成对比. OpenHelix^[109] 通过系统实验验证了异步架构的稳定性, 表明高层策略更新频率在 0.5 ~ 2.0 Hz 范围内变化时, 系统性能波动小于 5%, 在内存限制为 500 MB 的条件下任务成功率仍可达到 70% ~ 80%.

这些研究通过功能解耦与异步执行的显式结构设计, 获得了显著的效率提升, 在保留大模型语义理解能力的同时满足了机器人高频控制需求. 然而, 对于需要高频语义更新的动态任务, 固定的模块划分可能限制系统的适应性. 此外, 模块间接口的设计与通信机制的鲁棒性仍是影响整体性能的关键因

素:一方面,如何设计兼顾表达能力与传输效率的中间表示有待深入探索;另一方面,当高层模块出现推理延迟或失效时,底层执行模块的容错与降级策略尚未得到充分研究.未来研究需在模块自适应划分、接口标准化以及异常处理机制等方向取得突破,以构建更加灵活可靠的分层控制架构.

2.4 基于任务感知的条件计算

前述方法在架构确定后,对所有输入仍采用固定的计算路径.研究者发现,机器人操作任务的计算需求存在明显差异.简单的移动任务可能只需浅层特征就能完成,而精细的操作则需要深层网络的充分计算.标准 VLA 模型对所有输入执行完整前向传播,未能根据任务复杂度自适应调整.基于任务感知的条件计算使模型能够根据实时需求调整计算深度,避免不必要的开销.需要指出的是,本节探讨的条件计算与第 3.3 节的动态推理路径优化存在本质区别.条件计算是架构层面的原生设计:在训练阶段引入多路径执行结构,使模型能够根据任务状态自适应地选择计算路径,这种多模式行为直接编码在模型参数中.动态推理路径优化则是部署层面的后处理加速策略:针对已训练完成的模型,通过早退、层跳跃等外部机制在推理时动态截断计算图.前者改变模型的功能结构,后者仅调整推理时的执行深度.两者可以组合使用,但设计目标与作用阶段不同.

当前研究从行为模式切换、计算路径优化及能力集成等角度探索条件计算. SwitchVLA^[110] 通过联合编码视觉、语言及接触状态,构建任务感知的多模态表示,并基于此表示将控制策略划分为多个行为模式,根据任务阶段自适应选择执行分支.在 LIBERO^[86] 仿真环境中,该方法在任务切换场景下将成功率从基线的 8.0% ~ 10.0% 提升至 50.0% ~ 68.0%,真实机器人实验中延迟切换成功率达 94.4%. SP-VLA^[111] 结合模式切换与输入优化两类机制,设计“深思”与“直觉”双推理模式,并提出空间-语义双感知的 Token 剪枝策略.该方法在 LIBERO^[86] 基准上实现 1.5 倍推理加速且任务成功率无损,推理频率从 4.0 Hz 提升至 8.0 ~ 12.5 Hz. OneTwoVLA^[112] 则在单一模型中集成 System-1 (执行) 与 System-2 (推理) 能力,通过可学习的模式切换逻辑实现动态协调,在长视野任务规划中成功率达 87.0%,在开放世界视觉定位任务中成功率从基线的 3.0% 提升至 73.0%.

这些工作为 VLA 模型在算力受限环境下的实时交互开辟了新路径,验证了任务感知机制对计算

效率与泛化性的双重优势.条件计算架构通过在 VLA 内部显式引入“计算资源调度”机制,使模型在执行任务决策的同时,对自身计算开销进行实时管理,在推理效率和任务性能之间取得较优平衡.这一额外决策层也增加了系统的设计与调试复杂度:当控制失败时,难以区分问题源于基础策略能力不足还是资源调度策略不当,调试与优化需要在任务策略与计算策略两个层面同时进行.现有工作多在已知任务分布和相对受控的环境中学习或预设切换规则,其鲁棒性和泛化性在环境分布偏移及全新任务下仍缺乏系统验证,如何在保持计算效率优势的同时确保条件计算机制在开放场景下可靠运行,是后续研究亟待解决的核心问题.

2.5 高效动作表示与生成技术

动作解码模块负责将 VLM 主干的输出映射为连续高维控制信号,是连接感知与执行的关键环节.早期 VLA 模型采用自回归解码,将连续动作离散化后逐 Token 生成.这种方法虽然继承了语言模型的训练框架,但存在两方面限制:离散化引入的量化误差损害控制精度,串行生成导致的累积延迟难以支撑实时控制.为此,研究者从生成范式、解码架构及表示方法三个维度展开探索.

2.5.1 扩散模型与流匹配

生成式方法通过在连续空间直接建模动作分布,消除了离散化误差并实现并行生成.

扩散模型^[113]通过迭代去噪逐步还原数据分布.尽管存在 DDIM 等确定性采样方法,但其采样轨迹通常呈现复杂的曲线特征,往往需要数十次迭代才能保证生成质量,这成为高频闭环控制的瓶颈. TinyVLA^[94] 首次将扩散策略作为解码头融入 VLA 框架,相比自回归方案推理延迟降低 20 倍. DiffusionVLA^[98] 进一步探索了语义与动作的深度融合,通过推理注入模块将 VLM 的思维链以特征调制的方式引导扩散去噪过程.实验表明, DiVLA-7B 在保持 42 Hz 高频推理的同时,零样本泛化成功率较 OpenVLA^[90] 提升逾一倍.

流匹配^[114]则通过学习从噪声到数据的确定性速度场 (velocity field) 构建生成路径.以 π_0 ^[91] 采用的 Rectified Flow 为例,其训练目标明确导向平直的生成轨迹 (straight path).由于传输路径曲率较低,模型可用极少的积分步数完成推理,显著提升了效率. π_0 -Fast 进一步采用一致性蒸馏技术,将多步积分压缩为单步推理,在 Jetson 嵌入式平台上实现了高达 50 Hz 的推理频率.

如图 6 所示,自回归、扩散与流匹配代表了

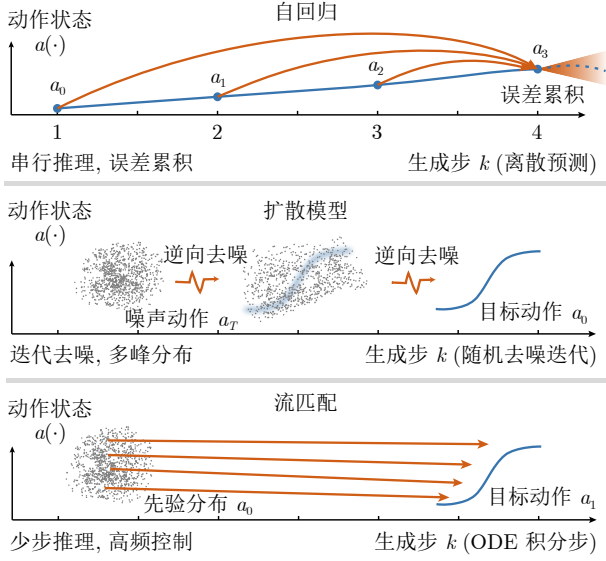


图6 三种主流 VLA 动作生成范式对比

Fig.6 Comparison of three mainstream VLA action generation paradigms

种不同的动作生成范式. 自回归通过逐 Token 串行解码实现动作生成, 虽然训练框架成熟但推理延迟较高, 主要适用于对实时性要求不高的场景. 扩散模型通过多轮去噪迭代, 逐步细化动作输出. 虽可处理多峰分布但推理步数较多. 流匹配则通过最小化概率传输路径的曲率, 实现了少步数推理, 在效率上具有明显优势, 但对训练数据的一致性要求更高. 在实际应用中, 动作方法选择需权衡任务特性与资源约束: 对于需要探索多种可能策略的复杂操作, 扩散模型的充分采样能力更为适合; 对于轨迹明确且需要高频响应的任务, 流匹配的效率优势更加突出; 自回归方法在保持 VLM 原生能力、简化训练流程方面具有优势.

2.5.2 动作分块与时序平滑

自回归解码的串行瓶颈严重制约了 VLA 模型的实时性能. 动作分块策略通过一次性预测未来多个时间步的动作序列来缓解这一问题. ACT^[17] 系统地提出了这一思路, 通过降低推理调用频率提升系统效率, 同时在块内进行时序建模以提升动作平滑性. 动作分块方法生成的序列在块边界处存在不连续问题. RTC^[69] 提出异步分块生成机制, 在执行当前块的同时并行预测下一块, 通过修复过程保持跨分块平滑衔接, 无需重新训练即可作为推理插件集成至任意模型. 与此互补, A2C2^[70] 聚焦块内时间感知校正, 通过轻量校正模块在每个控制周期对动作进行实时调整, 与 RTC 构成完整的时序稳定化方案. 这些工作表明, 通过精心设计的修正机制, 动作分块方法能够在效率与鲁棒性间取得有效平衡.

2.5.3 学习型动作表示

传统离散化采用均匀划分, 在高频控制场景下因时序冗余导致模型难以学习有效的运动模式. 学习型分词器通过从数据中提取动作基元, 实现更高效的编码. VQ-VLA^[115] 采用矢量量化自编码器^[58], 通过多阶段残差编码将高维轨迹映射为离散码本索引. 该方法需要在大规模数据集上预训练以学习通用动作基元, 初期训练成本较高, 但所学码本可跨任务复用, 显著提升了后续任务的适配效率. FAST^[116] 则结合分析型与学习型压缩. 该方法首先通过离散余弦变换将动作序列投影至频域, 利用信号的频域稀疏性实现初步压缩; 其次采用字节对编码压缩稀疏系数矩阵中的冗余模式. 这种混合策略使分词器构建时间缩短至数分钟, 同时保持了与端到端学习方法相当的表征能力, 训练效率提升 5 倍.

上述动作表示方法均针对自回归 VLA 架构设计, 旨在将连续动作压缩为离散 Token 以适配语言模型的预测机制. VQ-VLA^[115] 通过引入可学习的神经分词器, 以较高的训练成本换取更优的表示能力. 而 FAST^[116] 则通过轻量的信号处理方法优先保障训练效率, 其性能表现需在具体任务的精度与效率需求间权衡.

2.6 后训练优化方法

第 2.1 ~ 2.5 节从架构设计层面探讨了如何降低 VLA 的固有复杂度, 但训练所需的巨大计算成本, 构成了其广泛应用的另一个核心障碍. 本节将重点讨论训练策略本身, 即如何在给定的架构上, 通过创新的训练范式来降低训练门槛、提升数据效率. 具体而言, 参数高效微调通过模仿专家演示实现基础能力学习, 强化学习方法则通过环境反馈实现策略对齐与优化.

2.6.1 参数高效微调

VLA 模型的后训练过程通常首先依赖监督微调 (supervised fine-tuning, SFT), 通过模仿专家演示掌握任务先验. 然而, 对于数十亿参数的 VLM, 全参数微调 (full fine-tuning, FFT) 在计算和存储上代价高昂. 参数高效微调 (parameter-efficient fine-tuning, PEFT) 通过冻结 VLM 主干大部分参数, 仅更新少量可训练模块, 成为 VLA 训练的关键技术. 低秩适配 (low-rank adaptation, LoRA)^[117] 是当前应用最广泛的 PEFT 方案之一, 其基本思想是预训练模型微调时的权重变化矩阵具有低秩结构. 通过在原权重旁并联低秩矩阵, 可训练参数量降低数百倍, 且推理时可将低秩增量合并回原权重实现零额外开销. OpenVLA^[90] 通过系统实验验证

了 LoRA 的有效性, 在保持性能的同时将训练成本降低一个数量级. 第 2.1 节提到的 TinyVLA^[94] 及 VLA-Adapter^[99] 等方法亦采用该策略, 实现了算力受限场景下的高效适配.

2.6.2 强化学习方法

强化学习通过引入环境交互与奖励信号, 为 VLA 模型提供了超越模仿学习的策略优化路径. RL 方法以试错探索和价值估计为驱动, 能够同时优化模型对环境、本体和任务的综合理解, 从而在数据质量不均或稀缺的场景下展现优势. 将 VLA 模型的训练从模仿学习扩展至强化学习, 常面临探索导致遗忘与离线价值高估两大稳定性挑战. 一方面, 在线探索可能导致策略偏离预训练学到的视觉-语言理解能力, 造成“灾难性遗忘”; 另一方面, 在利用离线数据进行策略评估时, 模型难以准确估计训练数据分布之外动作的真实价值, 并倾向于产生不切实际的过高估值, 这可能导致策略“漂移”至未知且低效的动作空间.

为解决上述挑战, 现有研究主要从离线优化与稳定化、自动化奖励建模以及在线适应三个层面展开探索.

离线强化学习. 为充分利用现有演示数据并保持训练稳定, 研究者首先探索了离线强化学习与混合训练机制. CO-RFT^[118] 提出分块离线强化学习框架, 将时序差分学习扩展至动作序列级别, 并通过 Transformer-Critic 评估长期价值, 仅需少量演示即可大幅提升成功率. ReinboT^[119] 则聚焦于离线数据的质量异构性, 采用回报条件化 (reward-conditioned) 策略, 使模型能够在混合质量数据集上自动识别并利用高质量轨迹. 为进一步增强稳定性, ConRFT^[120] 在离线 RL 阶段引入人在回路 (human-in-the-loop) 作为安全保障, 并通过保留行为克隆损失来约束策略, 防止其偏离先验知识.

自动化奖励建模. 强化学习的效果取决于奖励设计. 依赖人工精心设计的奖励函数通常需要大量工程调试, 且稀疏的奖励信号也会导致模型探索效率低下. 部分工作探索了自动化奖励建模, 以提升策略学习效率、降低人工依赖. 一类方法是从数据中学习结构化奖励. ManipLVM-R1^[121] 提出基于可验证奖励的训练框架, 利用空间可操作区域预测与轨迹规划的结构化奖励替代人工监督. RFTF^[122] 则通过训练时序价值模型, 从无需动作标签的专家演示中学习状态价值的单调性规律, 并将其转化为稠密奖励信号. 另一类方法则利用 VLM 自身的理解能力来自动化生成奖励. IKER^[123] 利用 VLM 生成程序化的奖励函数, 为多步骤操作提供精确反馈,

取代了繁琐的人工奖励设计. 此外, 近期研究探索基于偏好对齐的隐式奖励建模. GRAPE^[124] 提出轨迹级偏好优化 (trajectory-wise preference optimization, TPO) 框架, 通过对比轨迹优劣直接更新策略, 并设计引导式成本偏好生成 (guided-cost preference generation, GCPG) 流程. 利用 VLM 自动生成定制化约束, 实现对特定操控目标 (如安全性、效率) 的灵活对齐.

在线强化学习与适应. 在线强化学习则聚焦于如何在保持先验知识的同时进行高效探索. iRe-VLA^[125] 提出了一种交替优化框架: 在 RL 阶段仅优化动作策略头, 通过冻结 VLM 主干网络以保持模型的多模态理解能力; 在 SFT 阶段, 将强化学习生成的高质量轨迹与原始演示数据联合训练, 实现知识融合. PA-RL^[126] 则利用离线 RL 预训练一个价值函数 (Q-function), 在在线交互时, 该函数用于评估基础策略采样的动作, 从而在不破坏先验知识的前提下选择更优动作. SimpleVLA-RL^[127] 采用了更为简化的训练策略, 仅使用二元奖励信号, 在单条演示轨迹初始化的条件下, 通过少量环境交互将成功率从 17.3% 大幅提升至 91.7%, 验证了 RL 在策略创新方面的潜力.

综上所述, 强化学习方法为 VLA 模型提供了从静态模仿向动态优化演进的技术路径. 现有研究通过离线学习、自动化奖励及约束机制等手段, 初步实现了探索效率与知识保持的平衡. 如何在保持预训练先验知识的同时充分释放强化学习的探索潜力, 仍是该领域有待突破的核心问题.

3 模型部署优化

第 2 节探讨了 VLA 模型在架构设计与训练阶段的效率优化策略, 这些方法旨在从源头降低模型的固有复杂度. 然而, 在实际部署中, 常需将已有的、参数庞大的预训练模型适配到算力、内存和功耗严格受限的机器人平台. 因此, 本节聚焦于模型的后处理与部署阶段, 系统探讨如何通过模型压缩、推理加速等部署优化技术, 推动 VLA 模型在边缘侧的实际应用.

为明确优化工作的边界与约束条件, 表 4 列举了机器人领域主流边缘计算平台的典型硬件规格参数.

以典型的 7B 参数 VLA 模型为例, FP16 精度下仅静态权重即需约 14 GB 存储空间; 若进一步计入运行时激活值、临时工作区及 KV Cache, 整体内存占用将进一步上升. 对于 Jetson Orin NX 这类配备 16 GB 统一内存的平台, 在不采用量化、裁剪

表 4 典型边缘计算平台硬件规格对比
Table 4 Comparison of hardware specifications of typical edge computing platforms

平台	算力	内存	功耗
NVIDIA Jetson AGX Orin	275 TOPS	64 GB	15 ~ 60 W
NVIDIA Jetson Orin NX	100 TOPS	16 GB	10 ~ 25 W
Ascend 310B	20 TOPS	24 GB	8 W
移动端 NPU	约 35 TOPS	共享内存	5 ~ 15 W
Raspberry Pi 5	无独立 NPU	8 GB	约 5 W

或分层卸载等缩减手段的情况下, 直接部署 7B 量级模型的空间裕度十分有限. 将权重量化至 INT4 后, 权重存储可压缩至约 3.5 GB, 在一定程度上缓解了内存压力, 但实际部署可行性仍需综合考虑量化元数据开销、运行时缓冲区占用、后端算子对低精度的支持程度, 以及控制频率与推理精度之间的折中. 同时, 低功耗边缘推理芯片在 INT8 算力、片上存储容量及动态计算支持方面的资源通常也相对有限, 使得大规模 VLA 模型在边缘硬件上的部署面临较严格的资源约束.

综合来看, VLA 模型在边缘侧部署主要受三方面因素制约: 一是内存容量紧张, 大参数量与 KV Cache 的组合易超出嵌入式设备的内存预算; 二是推理延迟较高, 单次前向传播耗时多在数百毫秒量级, 与实时控制所要求的毫秒级响应存在差距; 三是运行能耗较大, 持续高负载状态对移动平台的续航能力构成明显限制.

针对上述部署效率瓶颈, 本节在既定模型架构下, 沿数据处理流程系统梳理部署阶段的关键优化技术. 主要技术路径包括: 通过输入侧优化减少源头计算冗余; 采用量化、剪枝等压缩方法降低存储与计算开销; 基于输入复杂度的动态推理机制自适应分配计算资源; 加速自回归解码过程以缓解动作生成的序列化瓶颈; 通过算法与硬件的协同设计实现更深层次的适配. 下文将围绕上述方法在 VLA 场景中的适用特点、实现机制与面临的技术挑战展开具体分析.

3.1 输入侧优化

输入侧优化从数据源头减少推理计算量, 通过识别并压缩图像中的冗余信息降低计算负担. 尽管通用视觉语言模型已有成熟的 Token 剪枝方法, 但这些方法在 VLA 任务中往往效果不佳. 机器人操作场景具有视野静态背景多、关键动态区域小的特点, 通用方法难以准确识别任务相关特征, 盲目压缩易导致控制性能下降. 因此, VLA 的输入优化需着重从利用时空冗余与感知任务相关性两个维度进行.

如图 7 所示, 视觉 Token 优化主要从时空冗余利用和任务相关性两个维度展开. 在时空冗余利用方面, 主要思路是复用连续帧间的相似信息. VLA-Cache^[128] 利用连续帧间的高度相似性设计 Token 复用机制, 通过余弦相似度检测静止图像块, 但其在快速动态场景中的适应性仍有待验证. FlashVLA^[129] 从时空双维度分析推理冗余, 空间上通过奇异值分解定义 Token 信息贡献分数, 时间上设计双条件触发器判断动作复用时机, 然而信息贡献分数的计算过程本身可能引入额外开销. SpecPrune-VLA^[130] 提出两级剪枝框架, 第一级结合历史观测识别长期稳定区域, 第二级在不同网络层执行差异化筛选, 但剪枝阈值的设定仍依赖经验参数. 在任务相关性方面, 核心是依据当前指令与场景识别关键区域. 多数方法借助跨模态注意力分数来保留任务相关 Token. LightVLA^[131] 更进一步通过插入可微分的评分头进行端到端训练, 以自适应学习剪枝策略, 但这引入了额外的训练成本. 除了视觉 Token, 思维链文本也存在优化空间. Fast ECoT^[132] 观察到高层任务规划比视觉观测变化更慢的特性, 通过缓存并复用思维链文本, 仅在规划需要变更时才重新生成, 但其缓存一致性的保证机制仍需进一步完善. ECoT-Lite^[133] 则探索训练与推理解耦策略, 在训练时利用思维链数据改善模型表征, 推理时完全跳过思维链生成, 这一发现对思维链在 VLA 中的实际价值提出了新的思考.

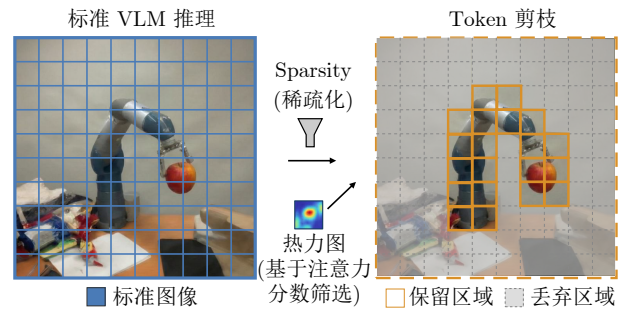


图 7 基于注意力热力图的视觉 Token 剪枝与稀疏化推理
Fig.7 Visual Token pruning and sparse inference based on attention heatmap

综上所述, 输入侧优化通过挖掘数据冗余有效降低了计算负载. 当前研究仍面临动态场景下优化稳定性不足与关键特征识别精度有限等挑战. 此外, 免训练方法与基于训练的方法在部署便捷性与优化效果之间存在固有权衡, 需根据实际场景的资源约束与性能要求进行选择. 输入侧优化从数据源头削减了冗余计算, 但 VLA 模型主干网络的参数量和计算复杂度仍是实际部署的核心瓶颈. 因此, 第 3.2

节将视角转向模型本身,探讨量化与剪枝等压缩技术如何在保持模型能力的前提下降低资源占用。

3.2 模型压缩

模型压缩技术面向已训练模型,通过对参数或结构进行后处理,在推理阶段实现持久的资源节约。量化与剪枝分别从数值精度与结构冗余两方面入手,为 VLA 模型的边缘部署提供了重要技术路径。

量化通过降低权重与激活值的数值精度来减少存储与计算开销。根据其训练流程的结合方式,主要分为两类:训练后量化 (post-training quantization, PTQ) 直接对已训练模型进行转换,实施简便,但在极低位宽下性能下降明显;量化感知训练 (quantization-aware training, QAT) 将量化过程融入训练,使模型主动适应低精度表示,因而在极低位宽下仍能保持较高性能。

大语言模型的量化技术为 VLA 模型提供了重要基础。GPTQ^[134] 利用二阶信息优化量化误差,AWQ^[135] 基于激活感知保护关键权重,SmoothQuant^[136] 通过权重-激活平滑化缓解激活量化难题。然而,这些方法主要针对语言模型设计,未充分考虑 VLA 任务中量化误差会通过环境反馈在连续决策环中累积。

因此,研究者开始探索适应 VLA 特性的量化方法。QAIL^[137] 提出量化感知模仿学习,通过对齐量化模型与全精度模型的动作分布缓解误差累积,在 4 位权重量化下保持任务成功率,并在边缘 GPU 上实现约 2.5 倍加速。值得注意的是,OpenVLA^[90] 团队发现,在多数 GPU 上将模型量化为 W8A8 格式后,因量化与反量化操作引入的额外开销,推理速度反而可能下降,导致控制频率降低。这一现象表明,量化策略的实际收益高度依赖底层硬件的算子优化,算法与硬件协同设计至关重要。

结构化剪枝通过移除完整的结构单元 (如注意力头、层或通道) 降低计算量,其优势在于可直接在标准硬件上实现加速。大语言模型的剪枝方法如 LLM-Pruner^[138]、SliceGPT^[139] 和 Sheared LLaMA^[140] 等在语言任务上表现良好,但直接应用于 VLA 时往往导致明显的性能退化。GLUESTICK^[141] 的谱分析显示,VLA 权重能量分布较语言模型更为均匀,缺乏明显冗余结构,这一特性导致传统基于幅值或梯度的重要性度量在 VLA 中效果有限。为此,该方法提出剪枝后恢复策略,通过对剪枝前后权重差值进行奇异值分解,构建低秩校正项以补偿性能损失。该方法无需额外训练即可部分恢复性能,但其有效性依赖于误差估计的准确性。EfficientVLA^[142] 采用

动态层跳跃策略,通过相邻层输出的余弦相似度判断冗余性,在运行时实现自适应剪枝。该方法的优势在于避免不可逆的永久性损失,但动态决策本身也可能引入额外开销。

协同压缩策略通过技术组合实现更深度的优化。RLRC^[143] 采用三阶段流程,依次进行结构化剪枝、监督微调和强化学习优化,最终实现 8 倍内存压缩。SQAP-VLA^[144] 则从协同设计出发,通过 Hadamard 变换改善激活分布,并设计量化感知的剪枝准则,在 W4A4 量化下实现 1.93 倍加速且成功率提升 4.5%。此外,BitVLA^[145] 将蒸馏机制融入量化训练,利用全精度模型指导三元量化模型学习,将模型大小降至约 1.4 GB。尽管如此,独立的知识蒸馏方案在 VLA 领域仍较为少见。这些成果表明,精心设计的多技术协同流程对提升压缩效果具有重要作用。

总体而言,VLA 模型压缩受制于量化误差累积、权重分布均匀等多重因素。当前研究虽在特定场景下取得进展,但通用化的即插即用方案仍有所欠缺。未来工作需要建立更完善的理论基础,探索压缩方法与模型内在特性的关联。上述方法均在推理前确定了模型的精简形式,属于静态优化范畴。然而,不同输入的处理难度存在显著差异:简单抓取任务与复杂装配任务对计算资源的需求截然不同。若对所有输入采用相同的计算深度,则会造成资源浪费或性能不足。因此,研究者开始探索在推理过程中动态调整计算路径的方法。

3.3 动态推理路径优化

动态推理路径优化通过在推理过程中依据输入复杂度自适应地调整计算深度,在效率与精度之间取得平衡。与第 2.4 节所述在训练阶段引入任务感知机制的做法不同,本节方法属于面向已训练模型的部署层面加速技术,其核心在于通过早退、层跳跃及层级路由等外部控制手段减少冗余计算,而无需改动模型原有的参数与功能特性。这类方法可灵活应用于 OpenVLA 等主流 VLA 模型,部分方法无需额外训练即可直接部署。

层跳跃策略通过识别冗余层来降低计算深度。如第 3.2 节所述,EfficientVLA^[142] 采用基于相邻层输出余弦相似度的免训练判断机制,可与其他优化技术协同工作,整体加速效果达到 1.93 倍。早退机制在网络不同深度设置多个出口,根据输出置信度动态选择退出点。DeeR-VLA^[146] 在多模态大语言模型中引入多层中间出口,基于动作一致性判定当前任务复杂度,简单场景下在浅层网络即可输出动作,

复杂场景则启用全部模型深度. 该方法配合专门的训练策略与辅助损失, 在 CALVIN^[147] 上将计算量降低 5 ~ 6 倍, 显存占用相应减少. 该方法的优势在于能够根据计算预算自动调整退出策略, 为资源受限平台提供了灵活的部署方案.

MoLe-VLA^[148] 提出了一种适配于现有 VLA 模型 (如 OpenVLA^[90]、CogAct^[149]) 的动态层级路由框架. 该框架通过训练空间-时间感知路由器, 联合建模视觉输入的空间结构与语言输入的时序依赖, 按需激活与当前任务状态最相关的 Transformer 层. 为弥补层跳跃可能带来的语义损失, 研究者引入认知自蒸馏机制, 通过认知 Token 和蒸馏损失确保稀疏化推理保持必要的语义表征能力. 实验表明, 该方法在降低 5.6 倍计算成本的同时, 任务成功率反而有所提升.

上述方法在不同粒度上实现了计算的自适应分配. 层跳跃策略基于简单的统计分析实现免训练部署, 早退机制通过置信度评估在深度维度进行动态调整, 层级路由则在每一层都进行独立的执行决策. 这些方法的核心挑战在于如何准确判定计算的必要性, 避免过早退出或错误跳跃关键层导致性能下降. 当前研究多在特定基准任务上验证效果, 其在开放场景下的泛化能力与鲁棒性仍需深入探索. 此外, 动态决策本身引入的额外计算开销与节省的计算量之间的平衡, 也是实际部署中需要仔细权衡的因素. 动态推理路径优化主要作用于模型的主干网络, 有效降低了编码与推理阶段的计算开销. 然而, 对于采用自回归范式的 VLA 模型, 动作解码阶段的串行依赖构成了实时控制的主要瓶颈之一: 每个动作 Token 的生成都必须等待前序 Token 完成, 导致解码延迟随动作维度线性增长. 因此, 第 3.4 节将聚焦于动作生成环节的加速策略.

3.4 动作生成加速

主流 VLA 模型普遍采用自回归方式生成动作序列, 这种设计虽能复用语言模型的生成框架, 但其固有的串行依赖关系导致推理延迟随动作维度线性增长, 难以满足机器人实时控制的频率要求.

针对此瓶颈, 现有研究主要从并行化解码及推测加速两个角度展开探索. PD-VLA^[150] 将自回归过程重新表述为非线性方程系统, 通过 Jacobi 不动点迭代求解, 将因果注意力替换为双向注意力, 使模型在一次前向传播中同时更新所有动作 Token, 实现约 2.5 倍加速且无需训练. CEED-VLA^[151] 采用一致性蒸馏技术, 训练模型直接从任意中间状态预测最终收敛结果, 从而减少迭代次数, 在 CALVIN^[147]

及 LIBERO^[86] 上实现 4 倍以上加速, 但其训练过程相对复杂. OpenVLA-OFT^[152] 保持视觉语言主干不变, 仅训练一个新的动作回归头, 采用 L_1 损失直接预测连续动作值, 将多次串行推理合并为单次前向传播. 该方法在 A100 GPU 上实现了约 25 倍的吞吐量提升.

推测解码方法则借鉴了大语言模型的加速思路, 通过轻量草稿模型生成候选序列, 再由验证模型并行校验. Spec-VLA^[153] 将此方法引入 VLA 模型, 但发现直接应用效果有限, 这主要是因为 VLA 的动作预测任务比语言建模更为复杂和精确, 导致草稿模型的接受率极低. 为此, 该方法提出了宽松接受机制, 允许草稿模型的预测在一定误差范围内被接受, 从而将加速比提升至 1.2 ~ 1.4 倍. 其加速比远低于在语言模型上的表现, 这反映出动作预测对精度与稳定性的要求更高, 导致草稿模型难以生成高质量候选序列.

动作生成加速方法的研究需要在训练成本与加速效果之间进行权衡. 无需训练的并行化方法为快速部署提供了便利, 但其性能增益通常较为有限. 相比之下, 基于训练的方法能够实现更优越的加速比, 但需要较高的再训练成本或对模型架构进行改造. 需要注意的是, 上述技术路径目前主要适用于自回归解码范式, 无法直接迁移至采用扩散模型或流匹配等生成方式的 VLA 模型.

在计算资源极度受限的场景下, 三类动作生成方法在实时性与部署代价上存在明显差异, 需依据任务特性与硬件条件进行选择. 流匹配方法 (如 π_0 ^[91]) 通过对连续动作分布进行建模, 并以相对较少的积分步生成动作序列, 可以达到较高的控制频率与较好的实时性. 扩散模型在机器人控制中具有较强的多模态动作分布建模能力, 但其多步迭代去噪过程通常带来较高的推理开销. 自回归方法则受串行生成依赖限制, 在动作维度较高或动作分块长度增大时容易出现推理瓶颈. 不过, 借助动作分块与并行解码机制 (如 PD-VLA^[150]), 其执行频率可得到显著提升, 同时这类方法通常更容易复用现有 VLM 的训练与解码框架. 总体而言, 流匹配方法在低采样步数控制方面具有优势, 扩散模型在复杂多模态动作建模方面更具竞争力, 而自回归方法则在架构一致性与工程复用性方面更具优势.

3.5 软硬件协同优化设计

在资源受限的机器人平台上, 算法层面的模型压缩与推理加速并不能直接带来端到端性能收益, 其实际效果高度依赖底层硬件架构与系统执行特性

的支持情况. 以 OpenVLA 的量化实验为例, 研究发现, 在缺乏高效低精度算子支持的硬件平台上, 低比特量化不仅难以带来预期加速, 甚至可能降低整体控制频率^[90]. 这一现象表明, VLA 模型的部署瓶颈并非单一算法问题, 而是由模型结构、系统执行方式与硬件特性共同决定的系统性约束.

大语言模型领域已形成一系列成熟的系统级优化技术, 如通过 I/O 感知算法优化片上数据复用的 FlashAttention^[154], 实现显存高效管理的 PagedAttention^[155], 以及提升硬件利用率的连续批处理技术^[156]. 这些技术主要针对云端、高并发、批处理的文本推理场景, 与 VLA 模型在机器人等边缘控制场景中的计算特征存在显著差异. 具体来说, VLA 的高效部署面临两大挑战. 首先是单样本流式推理中的存储瓶颈问题. 现有大语言模型优化技术主要通过扩大批处理规模来提升算术强度, 从而掩盖内存访问延迟. 但 VLA 在实时推理中通常需以单样本、流式方式处理视频数据. 在此模式下, 硬件计算单元利用率低, 且每一次自回归推理步都需要从内存中加载全部模型参数与动态增长的 KV 缓存, 导致内存访问开销成为系统延迟的主要问题. 此外, VLA 模型处理的视觉上下文具有时序连续性, 一般采用滑动窗口方式进行更新, 这与大语言模型相对静态的文本上下文不同. 因此, 直接采用为静态上下文设计的缓存管理策略可能因内存数据频繁调度而导致性能抖动. 其次, 动态推理存在时序不确定性. 控制系统对推理延迟的分布有严格要求. 第 3.3 节所述的早退机制、混合专家模型中的动态路由等条件计算技术, 虽可降低平均计算成本, 但其数据依赖的执行路径会导致推理时间产生难以预测的波动. 这种延迟不确定性可能破坏控制回路的稳定性. 同时, 动态稀疏化计算可能破坏内存访问的局部性, 在硬件缺乏相应稀疏计算支持时, 甚至可能增加实际延迟.

针对上述差异, 尽管目前直接针对 VLA 的软硬件协同研究尚处于起步阶段, 但基于边缘计算与实时控制系统的固有特性, 工程实践中存在若干极具潜力的优化路径:

计算与控制解耦的分层架构. 为缓解高带宽带来的实时性压力, 一种可行的策略是采用异构频率执行机制. 将计算密集的视觉编码器与轻量级的策略解码头解耦, 前者以较低频率运行并维持状态估计, 后者则以高频匹配控制回路. 这种设计在系统层面规避了每个控制周期全量加载视觉主干的带宽开销^[157-158]. 同时, 利用算子融合与运行时图优化 (如 CUDA graph) 技术, 可有效减少单样本推理中

的内核启动延迟与 CPU 调度开销^[159-160].

面向预算约束下的确定性推理. 为抑制动态推理引发的抖动, 系统设计往往需引入严格的资源预算机制. 可通过在编译期或运行时锁定最大 Token 生成数或强制截断层数, 确保推理延迟的上界可控^[161-163]. 在架构层面, 部分工作尝试构建分层控制体系: 由时序确定的传统控制器保障底层实时性, 而 VLA 模型作为高层规划器容忍偶发的推理延迟, 从而在非实时操作系统上实现可靠控制^[158].

硬件感知的算法与编译协同优化. 算法设计必须与目标平台的硬件特性深度耦合. 量化方案的选择需优先考量硬件原生指令集的支持情况. 对于非结构化稀疏模型, 若缺乏硬件稀疏计算单元支持, 其性能收益往往被不规则内存访问抵消^[164]. 因此, 基于硬件特性的结构化剪枝与编译器的自动张量优化成为边缘端实现实际加速的关键路径^[159-160].

在编译与部署方面, 不同工具链对低比特量化与异构执行的支持重点有所差异. TensorRT 对 INT8 推理的支持已较为成熟, 并能够处理动态形状、条件分支及循环等控制流结构; 在稀疏计算方面, 其硬件加速主要依赖 Ampere 及以上架构所支持的 2:4 结构化稀疏模式, 因此对任意非结构化稀疏的通用加速能力仍相对有限. 相比之下, TVM 更侧重跨平台代码生成与面向异构目标的性能调优, 而 MLIR 则倾向于提供量化、形状推断以及方言转换等编译基础设施支撑. 尽管二者在机制上均具备对低比特推理与动态形状处理的一定支持, 但在面向 VLA 这类兼具多模态输入、混合精度需求以及复杂控制流特征的任务时, 通常仍需结合具体后端进行针对性适配.

在硬件实现方面, NPU 普遍更适于执行规则化、静态的量化算子, 而对于早退机制、专家路由等动态性较强的推理路径, 实际部署中往往仍需借助图分割或异构协同方案加以应对. FPGA 因其可重构特性与较好的时延确定性, 在时序敏感场景中具有一定应用潜力, 但受限于片上存储容量与带宽资源, 在执行大规模矩阵计算时通常难以显现优势. 总体而言, 面向 VLA 模型的软硬件协同优化仍处于持续演进的阶段, 如何将模型压缩、编译映射与硬件执行路径有效贯通, 并将理论计算量的减少稳定地转化为实际时延、运行频率与能效方面的收益, 仍是一个值得深入研究的课题.

4 未来展望

VLA 模型在仿真与受控环境中已展现出强大潜力, 但其向真实世界部署的进程仍面临性能、鲁

棒性与可信度的系统性挑战。为推动 VLA 从实验室走向规模化应用, 如图 8 所示, 未来研究应在数据、架构、协同优化与评估四个关键方向深化探索。

4.1 数据价值驱动的高效扩展

为使 VLA 模型走向实际应用, 需首先建立统一的理论框架, 以系统分析模型规模、数据规模与最终性能间的约束关系。需摒弃单纯扩张参数规模的范式, 转而优先探究资源受限条件下的模型规模定律 (scaling laws), 从而明确计算预算、数据质量与模型能力之间的定量联系。

在此基础上, 应构建可解释、可量化的数据价值评估体系, 从根本上推动研究范式从“数据规模优先”向“数据价值优先”转变。

为实现上述转变, 未来工作还需面向实际部署场景, 构建“感知-决策-数据”深度耦合的闭环优化架构。一方面, 发展模型驱动的主动数据构建机制, 针对边界案例、关键决策点及失败场景进行精准采样; 另一方面, 将数据价值评估结果实时反馈至模型训练与架构设计环节, 从而在资源受限条件下, 持续提升模型的学习效率与泛化稳健性。

4.2 脑机制启发的高效计算架构

当前人工智能模型普遍以 Transformer 架构为主导, 这一“算力驱动”的范式带来了巨大的能耗与成本压力。这一瓶颈在尺寸、重量与功耗 (SWaP) 受限的具身智能平台上尤为突出。因此, VLA 模型的未来演进, 亟须在模型架构层面实现根本性创新,

推动发展模式从“算力驱动”向“效率优先”转型。

未来研究应深入借鉴脑科学与认知科学, 探究人脑在极低功耗约束下实现高效信息处理的神经计算原理。例如, 脉冲神经网络 (spiking neural networks, SNN) 以事件驱动的稀疏脉冲编码替代传统密集浮点激活, 在理论上可将推理能耗降低 1 ~ 2 个数量级, 其异步计算特性与机器人传感器的稀疏时序事件高度契合; 预测编码框架将大脑建模为持续生成预测并校正预测误差的层级系统, 模型仅需传递预测误差而非原始感知输入, 从而大幅减少前向计算开销, 基于此框架的主动推理方法已在机器人控制中展现出对不确定环境的高效适应能力。此外, 借鉴人脑稀疏编码、模块化分工等机制, 设计新型记忆架构以克服“灾难性遗忘”, 亦是提升 VLA 持续学习能力的重要方向。这些脑启发方法有望从架构源头降低模型的计算与存储开销, 朝着绿色 AI 与可持续计算的目标迈进。

4.3 面向部署的端到端协同优化

当前 VLA 模型的实验室性能与实际部署效果存在显著差距, 其根源在于传统的孤立优化模式难以应对边缘设备的严苛约束。因此, 亟须推动发展范式从后处理的局部优化, 转向全链路协同设计, 将效率约束贯穿模型开发的全生命周期。

未来的研究需在不同阶段实现深度协同: 在训练阶段引入面向边缘设备原生精度的量化感知训练, 并从感知源头对高冗余输入进行压缩; 在模型架构层面发展动态推理路径与结构化剪枝相结合的



图 8 资源受限场景下 VLA 模型的未来研究展望

Fig.8 Future research prospects for VLA models under resource-constrained settings

自适应计算机制;在部署阶段通过算子融合、内存访问优化等技术,与特定硬件架构深度适配。

上述跨层协同的核心,在于打破算法与硬件之间的传统壁垒,实现从数据、训练、架构到推理的端到端优化。唯有建立这种全链路协同的优化范式,VLA 模型方能在严格的 SWaP 约束下实现高效可靠的部署,从而真正从实验室演示走向实际应用。

4.4 标准化的多维评测基准构建

VLA 作为面向机器人控制的模型,其最终价值必然体现在真实世界的部署之中,然而当前的评测体系难以有效支撑这一目标。在推理效率评测方面,存在较为严重的标准缺失问题:不同工作在异构硬件平台上报告推理频率,导致数据之间缺乏横向可比性;延迟测试的口径不统一,部分工作仅报告单次前向传播的核心推理时间,另一些则包含了预处理、后处理及通信开销;吞吐量与控制频率等概念存在混用现象;能耗指标几乎从未被系统地报告。这种评测上的混乱状态,严重阻碍了研究成果的客观比较。

一方面,由于缺乏面向部署的统一规范,不同硬件与测量方法下报告的性能难以公平比较;另一方面,评测多局限于任务成功率等单一指标,严重忽视了在资源受限平台上的实际部署可行性(如延迟、能耗),导致实验室中的优异表现与工程应用价值严重脱节。

因此,未来亟须构建一个标准化的、多维度的评测基准体系。首先,应建立“透明化报告”规范,倡导研究者在发布成果时,报告模型在代表性边缘平台上的关键性能指标,包括任务平均推理延迟、平均推理能耗、平均显存占用等,并详细说明测试环境配置。在此基础上,评测基准还应从单一维度转向“准确、高效、鲁棒、可信”的多维框架。在鲁棒性方面,应建立覆盖非理想条件的压力测试集,不仅包含传感器噪声、运动模糊和计算资源抖动,还应系统地评估模型在长时程任务中的稳定性以及失败后的恢复能力。在可信性方面,则需通过场景化测试,量化模型的“幻觉”风险、决策可解释性及安全伦理边界。

5 结束语

本文系统剖析了资源受限条件下 VLA 模型的核心挑战与前沿进展。面对数据稀缺与计算受限的双重瓶颈,研究从数据获取、高效模型设计、训练范式及部署优化四个层面,构建了覆盖“数据-模型-训练-部署”全周期的系统性技术框架与解决方案。

展望未来,VLA 模型的发展将在数据效率、计算架构、协同优化与评测标准等方向持续突破。构建高效、可靠、可信的 VLA 模型,是推动机器人从受控实验室走向开放世界的核心技术路径。期望本文的分析框架与技术梳理能够为相关研究提供参考,最终推动 VLA 模型在资源受限平台上的规模化应用。

参考文献

- Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning. Virtual Event: PMLR, 2021. 8748–8763
- Li J N, Li D X, Savarese S, Hoi S. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proceedings of the 40th International Conference on Machine Learning, Honolulu, USA: PMLR, 2023. Article No. 814
- Liu H T, Li C Y, Wu Q Y, Lee Y J. Visual instruction tuning. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 1516
- O'Neill A, Rehman A, Maddukuri A, Gupta A, Padalkar A, Lee A, et al. Open X-Embodiment: Robotic learning datasets and RT-X models: Open X-Embodiment collaboration. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Yokohama, Japan: IEEE, 2024. 6892–6903
- Khazatsky A, Pertsch K, Nair S, Balakrishna A, Dasari S, Karamcheti S, et al. DROID: A large-scale in-the-wild robot manipulation dataset. In: Proceedings of the 20th International Conference on Robotics: Science and Systems. Delft, The Netherlands: Robotics: Science and Systems Foundation, 2024.
- Oquab M, Darcet T, Moutakanni T, Vo H V, Szafraniec M, Khalidov V, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.
- Zhai X H, Mustafa B, Kolesnikov A, Beyer L. Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 11941–11952
- Zhong Y F, Bai F S, Cai S F, Huang X C, Chen Z, Zhang X W, et al. A survey on vision-language-action models: An action tokenization perspective. *arXiv preprint arXiv: 2507.01925*, 2025.
- Zhang Hui, Liang Shu-Tong, Li Ming-Xuan, Tian Yong-Lin, Ge Jing-Wei, Yu Hui, et al. Vision-language-action models: From the early foundations to the state-of-the-art. *Acta Automatica Sinica*, 2025, **51**(9): 1922–1950
(张慧, 梁殊彤, 李明轩, 田永林, 葛经纬, 于慧, 等. 视觉-语言-动作模型综述: 从前史到前沿. *自动化学报*, 2025, **51**(9): 1922–1950)
- Wang Wen-Sheng, Tan Ning, Huang Kai, Zhang Yu-Nong, Zheng Wei-Shi, Sun Fu-Chun. Embodied intelligence systems based on large models: A survey. *Acta Automatica Sinica*, 2025, **51**(1): 1–19
(王文晟, 谭宁, 黄凯, 张雨浓, 郑伟诗, 孙富春. 基于大模型的具身智能系统综述. *自动化学报*, 2025, **51**(1): 1–19)
- Guan W L, Hu Q H, Li A S, Cheng J. Efficient vision-language-action models for embodied manipulation: A systematic survey. *arXiv preprint arXiv: 2510.17111*, 2025.
- Yu Z S, Wang B, Zeng P P, Zhang H N, Zhang J, Wang Z, et al. A survey on efficient vision-language-action models. *arXiv*

- preprint arXiv: 2510.24795, 2025.
- 13 Ma Y E, Song Z X, Zhuang Y Z, Hao J Y, King I. A survey on vision-language-action models for embodied AI. *IEEE Transactions on Neural Networks and Learning Systems*, 2026, **37**(7): 3031–3051
 - 14 Brohan A, Brown N, Carbajal J, Chebotar Y, Dabis J, Finn C, et al. RT-1: Robotics Transformer for real-world control at scale. In: Proceedings of the 19th International Conference on Robotics: Science and Systems. Daegu, Republic of Korea: Robotics: Science and Systems Foundation, 2023.
 - 15 Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc., 2017. 6000–6010
 - 16 Fang H S, Fang H J, Tang Z Y, Liu J R, Wang C X, Wang J B, et al. RH20T: A comprehensive robotic dataset for learning diverse skills in one-shot. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Yokohama, Japan: IEEE, 2024. 653–660
 - 17 Zhao T Z, Kumar V, Levine S, Finn C. Learning fine-grained bimanual manipulation with low-cost hardware. In: Proceedings of the 19th International Conference on Robotics: Science and Systems. Daegu, Republic of Korea: Robotics: Science and Systems Foundation, 2023.
 - 18 Chi C, Xu Z J, Pan C E, Cousineau E, Burchfiel B, Feng S Y, et al. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. In: Proceedings of the 20th International Conference on Robotics: Science and Systems. Delft, The Netherlands: Robotics: Science and Systems Foundation, 2024.
 - 19 Shukor M, Aubakirova D, Capuano F, Kooijmans P, Palma S, Zouitine A, et al. SmolVLA: A vision-language-action model for affordable and efficient robotics. arXiv preprint arXiv: 2506.01844, 2025.
 - 20 Jiang T, Yuan T Y, Liu Y C, Lu C H, Cui J N, Liu X, et al. Galaxea open-world dataset and G0 dual-system VLA model. arXiv preprint arXiv: 2509.00576, 2025.
 - 21 Bu Q W, Cai J S, Chen L, Cui X Q, Ding Y, Feng S Y, et al. AgiBot World Colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Hangzhou, China: IEEE, 2025. 3549–3556
 - 22 Jiang S, Li H N, Ren R C, Zhou Y M, Wang Z P, He B. Kaiwu: A multimodal manipulation dataset and framework for robot learning and human-robot interaction. *IEEE Robotics and Automation Letters*, 2025, **10**(11): 11482–11489
 - 23 Han S H, Qiu B X, Liao Y, Huang S Y, Gao C, Yan S C, et al. RoboCerebra: A large-scale benchmark for long-horizon robotic manipulation evaluation. In: Proceedings of the 39th International Conference on Neural Information Processing Systems. San Diego, USA: Curran Associates Inc., 2025.
 - 24 Sliwowski D, Jadav S, Stanovcic S, Orbik J, Heidersberger J, Lee D. Demonstrating REASSEMBLE: A multimodal dataset for contact-rich robotic assembly and disassembly. In: Proceedings of the 21st International Conference on Robotics: Science and Systems. Los Angeles, USA: Robotics: Science and Systems Foundation, 2025.
 - 25 Mildenhall B, Srinivasan P P, Tancik M, Barron J T, Ramamoorthi R, Ng R. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 2022, **65**(1): 99–106
 - 26 Kerbl B, Kopanas G, Leimkühler T, Drettakis G. 3D Gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 2023, **42**(4): Article No. 139
 - 27 Li X H, Li J L, Zhang Z H, Zhang R, Jia F, Wang T C, et al. RoboGSim: A Real2Sim2Real robotic Gaussian splatting simulator. arXiv preprint arXiv: 2411.11839, 2024.
 - 28 Abou-Chakra J, Rana K, Dayoub F, Stünderhauf N. Physically embodied Gaussian splatting: A visually learnt and physically grounded 3D representation for robotics. In: Proceedings of the 8th Conference on Robot Learning. Munich, Germany: PMLR, 2025. 513–530
 - 29 Han X S, Liu M H, Chen Y L, Yu J Q, Lyu X Y, Tian Y, et al. Re3Sim: Generating high-fidelity simulation data via 3D-photorealistic real-to-sim for robotic manipulation. arXiv preprint arXiv: 2502.08645, 2025.
 - 30 Mandelkar A, Nasiriany S, Wen B W, Akinola I, Narang Y, Fan L X, et al. MimicGen: A data generation system for scalable robot learning using human demonstrations. In: Proceedings of the 7th Conference on Robot Learning. Atlanta, USA: PMLR, 2023. 1820–1864
 - 31 Hua P, Liu M H, Macaluso A, Lin Y F, Zhang W N, Xu H Z, et al. GenSim2: Scaling robot data generation with multi-modal and reasoning LLMs. In: Proceedings of the 8th Conference on Robot Learning. Munich, Germany: PMLR, 2024. 5030–5066
 - 32 Nasiriany S, Maddukuri A, Zhang L C, Parikh A, Lo A, Joshi A, et al. RoboCasa: Large-scale simulation of household tasks for generalist robots. In: Proceedings of the 20th International Conference on Robotics: Science and Systems. Delft, The Netherlands: Robotics: Science and Systems Foundation, 2024.
 - 33 Xue Z R, Deng S Y, Chen Z Y, Wang Y X, Yuan Z C, Xu H Z. DemoGen: Synthetic demonstration generation for data-efficient visuomotor policy learning. In: Proceedings of the 21st International Conference on Robotics: Science and Systems. Los Angeles, USA: Robotics: Science and Systems Foundation, 2025.
 - 34 Johnnynunez. Genesis: A generative and universal physics engine for robotics and beyond [Online], available: <https://forums.developer.nvidia.com/t/genesis-a-generative-and-universal-physics-engine-for-robotics-and-beyond/323946>, February 27, 2025
 - 35 Mittal M, Yu C, Yu Q X, Liu J Z, Rudin N, Hoeller D, et al. Orbit: A unified simulation framework for interactive robot learning environments. *IEEE Robotics and Automation Letters*, 2023, **8**(6): 3740–3747
 - 36 Tao S, Xiang F B, Shukla A, Qin Y Z, Hinrichsen X, Yuan X D, et al. ManiSkill3: GPU parallelized robot simulation and rendering for generalizable embodied AI. In: Proceedings of the 7th Robot Learning Workshop: Towards Robots with Human-Level Abilities. Virtual: ICLR, 2025.
 - 37 Xiang F B, Qin Y Z, Mo K C, Xia Y K, Zhu H, Liu F C, et al. SAPIEN: A simulated part-based interactive environment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 11094–11104
 - 38 Li Y Y, Du W X, Yu C, Li P H, Zhao Z H, Liu T Y, et al. Taccel: Scaling up vision-based tactile robotics via high-performance GPU simulation. arXiv preprint arXiv: 2504.12908, 2025.
 - 39 Maddukuri A, Jiang Z Y, Chen L Y, Nasiriany S, Xie Y Q, Fang Y, et al. Sim-and-real co-training: A simple recipe for vision-based robotic manipulation. In: Proceedings of the 21st International Conference on Robotics: Science and Systems. Los Angeles, USA: Robotics: Science and Systems Foundation, 2025.
 - 40 Deng S L, Yan M, Wei S L, Ma H X, Yang Y X, Chen J Y, et al. GraspVLA: A grasping foundation model pre-trained on billion-scale synthetic action data. arXiv preprint arXiv: 2505.03233, 2025.
 - 41 Ha D, Schmidhuber J. Recurrent world models facilitate policy evolution. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Montréal, Canada: Curran Associates Inc., 2018. 2455–2467

- 42 Bu Q W, Yang Y T, Cai J S, Gao S Y, Ren G H, Yao M Q, et al. UniVLA: Learning to act anywhere with task-centric latent actions. In: Proceedings of the 21st International Conference on Robotics: Science and Systems. Los Angeles, USA: Robotics: Science and Systems Foundation, 2025.
- 43 Bruce J, Dennis M D, Edwards A, Parker-Holder J, Shi Y G, Hughes E, et al. Genie: Generative interactive environments. In: Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria: PMLR, 2024. 4603–4623
- 44 Cen J, Yu C H, Yuan H J, Jiang Y M, Huang S T, Guo J Y, et al. WorldVLA: Towards autoregressive action world model. arXiv preprint arXiv: 2506.21539, 2025.
- 45 Zhen H Y, Qiu X W, Chen P H, Yang J C, Yan X, Du Y L, et al. 3D-VLA: A 3D vision-language-action generative world model. In: Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria: PMLR, 2024. 61229–61245
- 46 Liao Y, Zhou P F, Huang S Y, Yang D L, Chen S C, Jiang Y X, et al. Genie Envisioner: A unified world foundation platform for robotic manipulation. arXiv preprint arXiv: 2508.05635, 2025.
- 47 Zhang W Y, Liu H S, Qi Z K, Wang Y N, Yu X Q, Zhang J Z, et al. DreamVLA: A vision-language-action model dreamed with comprehensive world knowledge. arXiv preprint arXiv: 2507.04447, 2025.
- 48 NVIDIA. Cosmos world foundation model platform for physical AI. arXiv preprint arXiv: 2501.03575, 2025.
- 49 Xiao J J, Yang Y D, Chang X Y, Chen R H, Xiong F, Xu M, et al. World-Env: Leveraging world model as a virtual environment for VLA post-training. arXiv preprint arXiv: 2509.24948, 2025.
- 50 Wu P, Escontrela A, Hafner D, Goldberg K, Abbeel P. Day-Dreamer: World models for physical robot learning. In: Proceedings of the 6th Conference on Robot Learning. Auckland, New Zealand: PMLR, 2022. 2226–2240
- 51 Hafner D, Yan W, Lillicrap T. Training agents inside of scalable world models. arXiv preprint arXiv: 2509.24527, 2025.
- 52 Chen D L, Moutakanni T, Chung W, Bang Y J, Ji Z W, Bolourchi A, et al. Planning with reasoning using vision language world model. arXiv preprint arXiv: 2509.02722, 2025.
- 53 Mendonca R, Bahl S, Pathak D. Structured world models from human videos. In: Proceedings of the 19th International Conference on Robotics: Science and Systems. Daegu, Republic of Korea: Robotics: Science and Systems Foundation, 2023.
- 54 Nair S, Rajeswaran A, Kumar V, Finn C, Gupta A. R3M: A universal visual representation for robot manipulation. In: Proceedings of the 6th Conference on Robot Learning. Auckland, New Zealand: PMLR, 2022. 892–909
- 55 Radosavovic I, Xiao T T, James S, Abbeel P, Malik J, Darrell T. Real-world robot learning with masked visual pre-training. In: Proceedings of the 6th Conference on Robot Learning. Auckland, New Zealand: PMLR, 2022. 416–426
- 56 Zeng J, Bu Q W, Wang B J, Xia W K, Chen L, Dong H, et al. Learning manipulation by predicting interaction. In: Proceedings of the 20th International Conference on Robotics: Science and Systems. Delft, The Netherlands: Robotics: Science and Systems Foundation, 2024.
- 57 Ko P C, Mao J Y, Du Y L, Sun S H, Tenenbaum J B. Learning to act from actionless videos through dense correspondences. In: Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: OpenReview.net, 2024.
- 58 van den Oord A, Vinyals O, Kavukcuoglu K. Neural discrete representation learning. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc., 2017. 6309–6318
- 59 Bahl S, Gupta A, Pathak D. Human-to-robot imitation in the wild. In: Proceedings of the 18th International Conference on Robotics: Science and Systems. New York, USA: Robotics: Science and Systems Foundation, 2022.
- 60 Allshire A, Choi H, Zhang J Y, McAllister D, Zhang A, Kim C M, et al. Visual imitation enables contextual humanoid control. arXiv preprint arXiv: 2505.03729, 2025.
- 61 Patel S, Mohan S, Mai H L, Jain U, Lazebnik S, Li Y Z. Robotic manipulation by imitating generated videos without physical demonstrations. arXiv preprint arXiv: 2507.00990, 2025.
- 62 Yang R H, Yu Q X, Wu Y C, Yan R, Li B R, Cheng A C, et al. EgoVLA: Learning vision-language-action models from ego-centric human videos. arXiv preprint arXiv: 2507.12440, 2025.
- 63 Romero J, Tzionas D, Black M J. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (TOG)*, 2017, **36**(6): Article No. 245
- 64 Luo H, Feng Y C, Zhang W P, Zheng S P, Wang Y, Yuan H Q, et al. Being-H0: Vision-language-action pretraining from large-scale human videos. arXiv preprint arXiv: 2507.15597, 2025.
- 65 Cheang C L, Chen G Z, Jing Y, Kong T, Li H, Li Y F, et al. GR-2: A generative video-language-action model with web-scale knowledge for robot manipulation. arXiv preprint arXiv: 2410.06158, 2024.
- 66 Jiang Y M, Huang S T, Xue S K, Zhao Y X, Cen J, Leng S C, et al. RynnVLA-001: Using human demonstrations to improve robot manipulation. arXiv preprint arXiv: 2509.15212, 2025.
- 67 Bharadhwaj H, Dwibedi D, Gupta A, Tulsiani S, Doersch C, Xiao T, et al. Gen2Act: Human video generation in novel scenarios enables generalizable robot manipulation. In: Proceedings of the 9th Conference on Robot Learning. Seoul, Republic of Korea: PMLR, 2025. 3936–3951
- 68 Xie H Z, Wen B C, Zheng J R, Chen Z X, Hong F Z, Diao H W, et al. DynamicVLA: A vision-language-action model for dynamic object manipulation. arXiv preprint arXiv: 2601.22153, 2026.
- 69 Black K, Galliker M Y, Levine S. Real-time execution of action chunking flow policies. In: Proceedings of the 39th International Conference on Neural Information Processing Systems. San Diego, USA: Curran Associates Inc., 2025.
- 70 Sendai K, Alvarez M, Matsushima T, Matsuo Y, Iwasawa Y. Leave no observation behind: Real-time correction for VLA action chunks. arXiv preprint arXiv: 2509.23224, 2025.
- 71 Sakr M, Zhang J Y, van der Loos H F M, Kulić D, Croft E. Consistency matters: Defining demonstration data quality metrics in robot learning from demonstration. *ACM Transactions on Human-Robot Interaction*, 2026, **15**(2): Article No. 34
- 72 Belkhale S, Cui Y C, Sadigh D. Data quality in imitation learning. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 3524
- 73 Wu K, Liu N, Zhao Z, Qiu D, Li J M, Che Z P, et al. Learning from imperfect demonstrations with self-supervision for robotic manipulation. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Atlanta, USA: IEEE, 2025. 16899–16906
- 74 Hou M H, Hindriks K V, Eiben A E, Baraka K. Active robot curriculum learning from online human demonstrations. In: Proceedings of the 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI). Melbourne, Australia: IEEE, 2025. 810–818
- 75 Walke H R, Black K, Zhao T Z, Vuong Q, Zheng C Y, Hansen-Estruch P, et al. BridgeData V2: A dataset for robot learning at scale. In: Proceedings of the 7th Conference on Robot Learning. Atlanta, USA: PMLR, 2023. 1723–1736
- 76 Wu K, Hou C K, Liu J M, Che Z P, Ju X Z, Yang Z Q, et al. RoboMIND: Benchmark on multi-embodiment intelligence normative data for robot manipulation. In: Proceedings of the 21st International Conference on Robotics: Science and Sys-

- tems. Los Angeles, USA: Robotics: Science and Systems Foundation, 2025.
- 77 Wang Z Q, Zheng H, Nie Y S, Xu W J, Wang Q W, Ye H, et al. All Robots In One: A new standard and unified dataset for versatile, general-purpose embodied agents. *arXiv preprint arXiv: 2408.10899*, 2024.
 - 78 NVIDIA Research. Physical AI [Online], available: <https://research.nvidia.com/research-area/physical-ai>, May 13, 2025
 - 79 Yan F, Liu F F, Huang Y Y, Guan Z C, Zheng L M, Zhong Y F, et al. RoboTron-Mani: All-in-one multimodal large model for robotic manipulation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Honolulu, USA: IEEE, 2025. 13707–13718
 - 80 Grauman K, Westbury A, Byrne E, Chavis Z, Furnari A, Girdhar R, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In: *Proceedings of the 25th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, USA: IEEE, 2022. 18973–18990
 - 81 Damen D, Doughty H, Farinella G M, Furnari A, Kazakos E, Ma J, et al. Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100. *International Journal of Computer Vision*, 2022, **130**(1): 33–55
 - 82 Grauman K, Westbury A, Torresani L, Kitani K, Malik J, Afouras T, et al. Ego-Exo4D: Understanding skilled human activity from first- and third-person perspectives. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, USA: IEEE, 2024. 19383–19400
 - 83 Banerjee P, Shkodrani S, Moulon P, Hampali S, Han S C, Zhang F, et al. HOT3D: Hand and object tracking in 3D from egocentric multi-view videos. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, USA: IEEE, 2025. 7061–7071
 - 84 Liu Y, Yang H L, Si X, Liu L, Li Z P, Zhang Y X, et al. TACO: Benchmarking generalizable bimanual tool-action-object understanding. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, USA: IEEE, 2024. 21740–21751
 - 85 Wang X, Kwon T, Rad M, Pan B W, Chakraborty I, Andrist S, et al. HoloAssist: An egocentric human interaction dataset for interactive AI assistants in the real world. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Paris, France: IEEE, 2023. 20213–20224
 - 86 Liu B, Zhu Y F, Gao C K, Feng Y H, Liu Q, Zhu Y K, et al. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. New Orleans, USA: Curran Associates Inc., 2023. Article No. 1939
 - 87 Zhang T L, Li D J, Li Y H, Zeng Z C, Zhao L, Sun L, et al. Empowering embodied manipulation: A bimanual-mobile robot manipulation dataset for household tasks. *arXiv preprint arXiv: 2405.18860*, 2024.
 - 88 Zitkovich B, Yu T H, Xu S C, Xu P, Xiao T, Xia F, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In: *Proceedings of the 7th Conference on Robot Learning*. Atlanta, USA: PMLR, 2023. 2165–2183
 - 89 Ghosh D, Walke H, Pertsch K, Black K, Mees O, Dasari S, et al. Octo: An open-source generalist robot policy. In: *Proceedings of the 20th International Conference on Robotics: Science and Systems*. Delft, The Netherlands: Robotics: Science and Systems Foundation, 2024.
 - 90 Kim M J, Pertsch K, Karamcheti S, Xiao T, Balakrishna A, Nair S, et al. OpenVLA: An open-source vision-language-action model. In: *Proceedings of the 8th Conference on Robot Learning*. Munich, Germany: PMLR, 2024. 2679–2713
 - 91 Black K, Brown N, Driess D, Esmail A, Equi M R, Finn C, et al. π_0 : A vision-language-action flow model for general robot control. In: *Proceedings of the 21st International Conference on Robotics: Science and Systems*. Los Angeles, USA: Robotics: Science and Systems Foundation, 2025.
 - 92 NVIDIA. GR00T N1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv: 2503.14734*, 2025.
 - 93 Liu J M, Liu M Z, Wang Z Y, An P J, Li X Q, Zhou K C, et al. RoboMamba: Efficient vision-language-action model for robotic reasoning and manipulation. In: *Proceedings of the 38th International Conference on Neural Information Processing Systems*. Vancouver, Canada: Curran Associates Inc., 2024. Article No. 1266
 - 94 Wen J J, Zhu Y C, Li J M, Zhu M J, Tang Z B, Wu K, et al. TinyVLA: Toward fast, data-efficient vision-language-action models for robotic manipulation. *IEEE Robotics and Automation Letters*, 2025, **10**(4): 3988–3995
 - 95 Reuss M, Zhou H Y, Rühle M, Yağmurlu Ö E, Otto F, Lioutikov R. FLOWER: Democratizing generalist robot policies with efficient vision-language-flow models. In: *Proceedings of the 9th Conference on Robot Learning*. Seoul, Republic of Korea: PMLR, 2025. 3736–3761
 - 96 Microsoft Research. Phi-2: The surprising power of small language models [Online], available: <https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/>, December 12, 2023
 - 97 Zhu Y C, Zhu M J, Liu N, Xu Z Y, Peng Y X. LLaVA-Phi: Efficient multi-modal assistant with small language model. In: *Proceedings of the 1st International Workshop on Efficient Multimedia Computing under Limited*. Melbourne, Australia: ACM, 2024. 18–22
 - 98 Liang Z X, Li Y Z, Yang T S, Wu C Y, Mao S T, Pei L A, et al. Discrete diffusion VLA: Bringing discrete diffusion to action decoding in vision-language-action policies. *arXiv preprint arXiv: 2508.20072*, 2025.
 - 99 Wang Y H, Ding P X, Li L X, Cui C, Ge Z R, Tong X Y, et al. VLA-Adapter: An effective paradigm for tiny-scale vision-language-action model. In: *Proceedings of the 40th AAAI Conference on Artificial Intelligence*. Singapore: AAAI Press, 2026. 18638–18646
 - 100 Gu A, Dao T. Mamba: Linear-time sequence modeling with selective state spaces. In: *Proceedings of the 1st Conference on Language Modeling*. Virtual Event: OpenReview.net, 2024.
 - 101 Qiao Y Y, Yu Z, Zhao Z J, Chen S H, Sun M Z, Guo L T, et al. VL-Mamba: Exploring state space models for multimodal learning. In: *Proceedings of the 4th Neural Information Processing Systems Efficient Natural Language and Speech Processing Workshop*. Vancouver, Canada: PMLR, 2024. 102–113
 - 102 Li K C, Li X H, Wang Y, He Y N, Wang Y L, Wang L M, et al. VideoMamba: State space model for efficient video understanding. In: *Proceedings of the 18th European Conference on Computer Vision*. Milan, Italy: Springer, 2024. 237–255
 - 103 Xie F, Nie J H, Tan Y J, Zhang W K, Zhao H S. Mamba-Adaptor: State space model adaptor for visual recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, USA: IEEE, 2025. 20124–20134
 - 104 Park J, Park J, Xiong Z Y, Lee N, Cho J, Oymak S, et al. Can Mamba learn how to learn? A comparative study on in-context learning tasks. In: *Proceedings of the 41st International Conference on Machine Learning*. Vienna, Austria: PMLR, 2024. Article No. 1612
 - 105 Ye Z F, Xia K J, Fu Y G, Dong X, Hong J, Yuan X C, et al. LongMamba: Enhancing Mamba's long-context capabilities via training-free receptive field enlargement. In: *Proceedings of the 13th International Conference on Learning Representations*. Singapore: OpenReview.net, 2025.
 - 106 Zhang J K, Guo Y J, Chen X Y, Wang Y J, Hu Y C, Shi C M, et al. HiRT: Enhancing robotic control with hierarchical robot Transformers. In: *Proceedings of the 8th Conference on Robot Learning*. Munich, Germany: PMLR, 2024. 933–946

- 107 Shentu Y D, Wu P, Rajeswaran A, Abbeel P. From LLMs to actions: Latent codes as bridges in hierarchical robot control. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Abu Dhabi, United Arab Emirates: IEEE, 2024. 8539–8546
- 108 Bu Q W, Li H Y, Chen L, Cai J S, Zeng J, Cui H M, et al. Towards synergistic, generalized, and efficient dual-system for robotic manipulation. In: Proceedings of the 13th International Conference on Learning Representations. Singapore: OpenReview.net, 2025.
- 109 Cui C, Ding P X, Song W X, Bai S H, Tong X Y, Ge Z R, et al. OpenHelix: A short survey, empirical analysis, and open-source dual-system VLA model for robotic manipulation. arXiv preprint arXiv: 2505.03912, 2025.
- 110 Li M, Zhao Z, Che Z P, Liao F, Wu K, Xu Z Y, et al. SwitchVLA: Execution-aware task switching for vision-language-action models. arXiv preprint arXiv: 2506.03574, 2025.
- 111 Li Y, Meng Y, Sun Z W, Ji K Y, Tang C, Fan J J, et al. SP-VLA: A joint model scheduling and token pruning approach for VLA model acceleration. arXiv preprint arXiv: 2506.12723, 2025.
- 112 Lin F Q, Nai R Q, Hu Y D, You J C, Zhao J M, Gao Y. OneTwoVLA: A unified vision-language-action model with adaptive reasoning. arXiv preprint arXiv: 2505.11917, 2025.
- 113 Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2020. Article No. 574
- 114 Lipman Y, Chen R T Q, Ben-Hamu H, Nickel M, Le M. Flow matching for generative modeling. In: Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: OpenReview.net, 2023.
- 115 Wang Y T, Zhu H Y, Liu M Y, Yang J G, Fang H S, He T. VQ-VLA: Improving vision-language-action models via scaling vector-quantized action tokenizers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Honolulu, USA: IEEE, 2025. 11089–11099
- 116 Pertsch K, Stachowicz K, Ichter B, Driess D, Nair S, Vuong Q, et al. FAST: Efficient action tokenization for vision-language-action models. In: Proceedings of the 21st International Conference on Robotics: Science and Systems. Los Angeles, USA: Robotics: Science and Systems Foundation, 2025.
- 117 Hu E J, Shen Y L, Wallis P, Allen-Zhu Z, Li Y Z, Wang S A, et al. LoRA: Low-rank adaptation of large language models. In: Proceedings of the 10th International Conference on Learning Representations. Virtual Event: OpenReview.net, 2022.
- 118 Huang D C, Fang Z R, Zhang T L, Li Y H, Zhao L, Xia C H. CO-RFT: Efficient fine-tuning of vision-language-action models through chunked offline reinforcement learning. arXiv preprint arXiv: 2508.02219, 2025.
- 119 Zhang H Y, Zhuang Z F, Zhao H, Ding P X, Lu H C, Wang D L. ReinboT: Amplifying robot visual-language manipulation with reinforcement learning. In: Proceedings of the 42nd International Conference on Machine Learning. Vancouver, Canada: PMLR, 2025. 77254–77271
- 120 Chen Y H, Tian S, Liu S G, Zhou Y T, Li H R, Zhao D B. ConRFT: A reinforced fine-tuning method for VLA models via consistency policy. In: Proceedings of the 21st International Conference on Robotics: Science and Systems. Los Angeles, USA: Robotics: Science and Systems Foundation, 2025.
- 121 Song Z R, Ouyang G X, Li M Z, Ji Y H, Wang C X, Xu Z X, et al. ManipLVM-R1: Reinforcement learning for reasoning in embodied manipulation with large vision-language models. In: Proceedings of the 40th AAAI Conference on Artificial Intelligence. Singapore: AAAI Press, 2026. 18558–18566
- 122 Shu J Y, Lin Z W, Wang Y T. RFTF: Reinforcement fine-tuning for embodied agents with temporal feedback. arXiv preprint arXiv: 2505.19767, 2025.
- 123 Patel S, Yin X C, Huang W L, Garg S, Nayyeri H, Fei-Fei L, et al. A real-to-sim-to-real approach to robotic manipulation with VLM-generated iterative keypoint rewards. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Atlanta, USA: IEEE, 2025. 8258–8266
- 124 Zhang Z J, Zheng K Y, Chen Z R, Jang J, Li Y, Han S W, et al. GRAPE: Generalizing robot policy via preference alignment. In: Proceedings of the 13th International Conference on Learning Representations. Singapore: OpenReview.net, 2025.
- 125 Guo Y J, Zhang J K, Chen X Y, Ji X, Wang Y J, Hu Y C, et al. Improving vision-language-action model with online reinforcement learning. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Atlanta, USA: IEEE, 2025. 15665–15672
- 126 Mark M S, Gao T, Sampaio G G, Srirama M K, Sharma A, Finn C, et al. Policy-agnostic RL: Offline RL and online RL fine-tuning of any class and backbone. In: Proceedings of the 13th International Conference on Learning Representations. Singapore: OpenReview.net, 2025.
- 127 Li H Z, Zuo Y X, Yu J L, Zhang Y H, Yang Z H, Zhang K Y, et al. SimpleVLA-RL: Scaling VLA training via reinforcement learning. arXiv preprint arXiv: 2509.09674, 2025.
- 128 Xu S Y, Wang Y K, Xia C H, Zhu D H, Huang T, Xu C. VLA-Cache: Efficient vision-language-action manipulation via adaptive token caching. In: Proceedings of the 39th International Conference on Neural Information Processing Systems. San Diego, USA: Curran Associates Inc., 2025.
- 129 Tan X D, Yang Y X, Ye P, Zheng J L, Bai B Z, Wang X Y, et al. Think twice, act once: Token-aware compression and action reuse for efficient inference in vision-language-action models. arXiv preprint arXiv: 2505.21200, 2025.
- 130 Wang H Z, Xu J M, Xiang Y S, Pan J Y, Zhou Y K, Li Y L, et al. SpecPrune-VLA: Accelerating vision-language-action models via action-aware self-speculative pruning. arXiv preprint arXiv: 2509.05614, 2025.
- 131 Jiang T T, Jiang X F, Ma Y, Wen X, Li B L, Zhan K, et al. The better you learn, the smarter you prune: Towards efficient vision-language-action models via differentiable token pruning. arXiv preprint arXiv: 2509.12594, 2025.
- 132 Duan Z K, Zhang Y, Geng S K, Liu G W, Boedecker J, Lu C X X. Fast ECoT: Efficient embodied chain-of-thought via thoughts reuse. arXiv preprint arXiv: 2506.07639, 2025.
- 133 Chen W, Belkhal S, Mirchandani S, Pertsch K, Driess D, Mees O, et al. Training strategies for efficient embodied reasoning. In: Proceedings of the 9th Annual Conference on Robot Learning. Seoul, Republic of Korea: PMLR, 2025. 365–391
- 134 Frantar E, Ashkboos S, Hoefler T, Alistarh D. OPTQ: Accurate quantization for generative pre-trained Transformers. In: Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: OpenReview.net, 2023.
- 135 Lin J, Tang J M, Tang H T, Yang S, Chen W M, Wang W C, et al. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. In: Proceedings of the 7th Annual Conference on Machine Learning and Systems. Santa Clara, USA: mlsys.org, 2024. 87–100
- 136 Xiao G X, Lin J, Seznec M, Wu H, Demouth J, Han S. SmoothQuant: Accurate and efficient post-training quantization for large language models. In: Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: PMLR, 2023. Article No. 1585
- 137 Park S, Kim H, Jeon W, Yang J, Jeon B, Oh Y, et al. Quantization-aware imitation-learning for resource-efficient robotic control. arXiv preprint arXiv: 2412.01034, 2024.
- 138 Ma X Y, Fang G F, Wang X C. LLM-Pruner: On the structural pruning of large language models. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 950

- 139 Ashkboos S, Croci M L, Nascimento M G D, Hoefler T, Hensman J. SliceGPT: Compress large language models by deleting rows and columns. In: Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: OpenReview.net, 2024.
- 140 Xia M Z, Gao T Y, Zeng Z Y, Chen D Q. Sheared LLaMA: Accelerating language model pre-training via structured pruning. In: Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: OpenReview.net, 2024.
- 141 Jabbour J, Kim D K, Smith M, Patrikar J, Ghosal R, Wang Y H, et al. Don't run with scissors: Pruning breaks VLA models but they can be recovered. arXiv preprint arXiv: 2510.08464, 2025.
- 142 Yang Y T, Wang Y H, Wen Z C, Luo Z W, Zou C, Zhang Z P, et al. EfficientVLA: Training-free acceleration and compression for vision-language-action models. arXiv preprint arXiv: 2506.10100, 2025.
- 143 Chen Y X, Han Y X, Huang Y Z, Li X. RLRC: Reinforcement learning-based recovery for compressed vision-language-action models. *IEEE Robotics and Automation Letters*, 2026, 11(7): 8864–8871
- 144 Fang H Y, Liu Y J, Du Y, Du L, Yang H R. SQAP-VLA: A synergistic quantization-aware pruning framework for high-performance vision-language-action models. arXiv preprint arXiv: 2509.09090, 2025.
- 145 Wang H Y, Xiong C Y, Wang R P, Chen X L. BitVLA: 1-bit vision-language-action models for robotics manipulation. arXiv preprint arXiv: 2506.07530, 2025.
- 146 Yue Y, Wang Y L, Kang B Y, Han Y Z, Wang S Z, Song S J, et al. DeeR-VLA: Dynamic inference of multimodal large language models for efficient robot execution. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2024. Article No. 1803
- 147 Mees O, Hermann L, Rosete-Beas E, Burgard W B. CALVIN: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 2022, 7(3): 7327–7334
- 148 Zhang R Y, Dong M H, Zhang Y, Heng L, Chi X W, Dai G L, et al. MoLe-VLA: Dynamic layer-skipping vision language action model via mixture-of-layers for efficient robot manipulation. In: Proceedings of the 40th AAAI Conference on Artificial Intelligence. Singapore: AAAI Press, 2026. 18764–18772
- 149 Li Q X, Liang Y B, Wang Z Y, Luo L, Chen X, Liao M Z, et al. CogACT: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. arXiv preprint arXiv: 2411.19650, 2024.
- 150 Song W X, Chen J Y, Ding P X, Zhao H, Zhao W, Zhong Z D, et al. PD-VLA: Accelerating vision-language-action model integrated with action chunking via parallel decoding. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Hangzhou, China: IEEE, 2025. 13162–13169
- 151 Song W X, Chen J Y, Ding P X, Huang Y X, Zhao H, Wang D L, et al. CEED-VLA: Consistency vision-language-action model with early-exit decoding. arXiv preprint arXiv: 2506.13725, 2025.
- 152 Kim M J, Finn C, Liang P. Fine-tuning vision-language-action models: Optimizing speed and success. In: Proceedings of the 21st International Conference on Robotics: Science and Systems. Los Angeles, USA: Robotics: Science and Systems Foundation, 2025.
- 153 Wang S S, Yu R C, Yuan Z H, Yu C, Gao F, Wang Y, et al. Spec-VLA: Speculative decoding for vision-language-action models with relaxed acceptance. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Suzhou, China: Association for Computational Linguistics, 2025. 26928–26940
- 154 Dao T, Fu D Y, Ermon S, Rudra A, Ré C. FLASHATTENTION: Fast and memory-efficient exact attention with IO-awareness. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. Article No. 1189
- 155 Kwon W, Li Z H, Zhuang S Y, Sheng Y, Zheng L M, Yu C H, et al. Efficient memory management for large language model serving with PagedAttention. In: Proceedings of the 29th Symposium on Operating Systems Principles. Koblenz, Germany: ACM, 2023. 611–626
- 156 Yu G I, Jeong J S, Kim G W, Kim S, Chun B G. Orca: A distributed serving system for Transformer-based generative models. In: Proceedings of the 16th USENIX Symposium on Operating Systems Design and Implementation. Carlsbad, USA: USENIX Association, 2022. 521–538
- 157 Xue H, Ren J J, Chen W D, Zhang G, Yuan F, Gu G Y, et al. Reactive diffusion policy: Slow-fast visual-tactile policy learning for contact-rich manipulation. In: Proceedings of the 21st International Conference on Robotics: Science and Systems. Los Angeles, USA: Robotics: Science and Systems Foundation, 2025.
- 158 Ma Y C, Zhou Y Z, Yang Y H, Wang T C, Fan H Q. Running VLAs at real-time speed. arXiv preprint arXiv: 2510.26742, 2025.
- 159 Xia C W, Zhao J C, Sun Q Q, Wang Z, Wen Y, Yu T, et al. Optimizing deep learning inference via global analysis and tensor expressions. In: Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems. La Jolla, USA: ACM, 2024. 286–301
- 160 Shao J R, Zhou X Y, Feng S Y, Hou B H, Lai R H, Jin H Y, et al. Tensor program optimization with probabilistic programs. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. Article No. 2593
- 161 Yang Y Q, Jiao L, Xu Y D. A queueing theoretic perspective on low-latency LLM inference with variable token length. In: Proceedings of the 22nd International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks. Seoul, Republic of Korea: IEEE, 2024. 273–280
- 162 Zhen R R, Li J T, Ji Y X, Yang Z L, Liu T, Xia Q R, et al. Taming the titans: A survey of efficient LLM inference serving. In: Proceedings of the 18th International Natural Language Generation Conference. Hanoi, Vietnam: Association for Computational Linguistics, 2025. 522–541
- 163 Xin J, Tang R, Lee J, Yu Y L, Lin J. DeeBERT: Dynamic early exiting for accelerating BERT inference. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Virtual Event: Association for Computational Linguistics, 2020. 2246–2251
- 164 Liu J, Dai G H, Xia H, Guo L D, Shi X S, Xu J M, et al. TSTC: Two-level sparsity tensor core enabling both algorithm flexibility and hardware efficiency. In: Proceedings of the IEEE/ACM International Conference on Computer Aided Design (ICCAD). San Francisco, USA: IEEE, 2023. 1–9



周毅 河南大学人工智能学院教授。2011年获得同济大学博士学位。主要研究方向为认知计算, 具身智能和智能网联汽车。

E-mail: zhouyi@henu.edu.cn

(ZHOU Yi) Professor at the School of Artificial Intelligence, Henan University. He received his Ph.D. degree from Tongji University in 2011. His research interests include cogni-

ive computing, embodied intelligence, and intelligent connected vehicles.)



孙煜坤 河南大学人工智能学院硕士研究生. 主要研究方向为具身智能与多智能体系统协同.
E-mail: sunyk@henu.edu.cn
(**SUN Yu-Kun** Master student at the School of Artificial Intelligence, Henan University. His research interests include embodied intelligence and multi-agent system coordination.)



石磊 河南大学人工智能学院教授. 2020 年获得电子科技大学博士学位. 主要研究方向为无人系统协同定位与控制, 社交网络观点动力学. 本文通信作者.
E-mail: shilei910918@126.com
(**SHI Lei** Professor at the School of Artificial Intelligence, Henan University. He received his Ph.D. degree from University of Electronic Science and Technology of China in 2020. His research interests include cooperative localization and control of unmanned systems, and opinion dynamics in social networks. Corresponding author of this paper.)



程玉华 电子科技大学自动化工程学院教授. 2007 年获得电子科技大学博士学位. 主要研究方向为电子测试技术, 精密无损检测, 故障诊断与健康健康管理.
E-mail: chengyuhua_auto@uestc.edu.cn
(**CHENG Yu-Hua** Professor at the School of Automation Engineering, University of Electronic Science and Technology of China. He received his Ph.D. degree from University of Electronic Science and Technology of China in 2007. His research interests include electronic testing technology, precision non-destructive testing, fault diagnosis, and health management.)



秦家虎 中国科学技术大学自动化系教授. 2013 年获得澳大利亚国立大学博士学位. 主要研究方向为多智能体系统分布式决策与控制, 信息物理系统安全与控制.
E-mail: [jqhin@ustc.edu.cn](mailto:jhqin@ustc.edu.cn)
(**QIN Jia-Hu** Professor at the Department of Automation, University of Science and Technology of China. He received his Ph.D. degree from Australian National University in 2013. His research interests include distributed decision and control of multi-agent systems, and security and control of cyber-physical systems.)