



汇报

AIR-VLA: Vision-Language-Action Systems for Aerial Manipulation

哈尔滨工业大学

汇报人：李明谦



目录

CONTENT

01

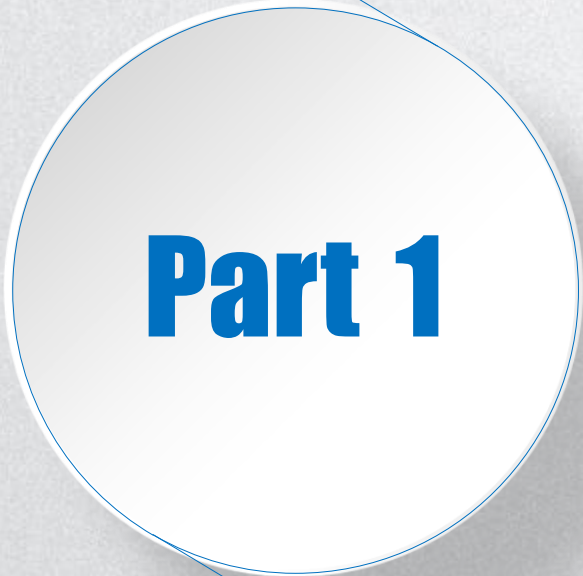
VLA和VLM

02

前沿模型

03

论文介绍



Part 1

VLA和VLM

VLM 的演进可以分为三个主要阶段：

1. 对齐阶段（2021年左右）：以 CLIP 为代表。通过对比学习将图像特征和文本特征映射到同一个空间，实现了“图文匹配”。

2. 融合阶段（2022-2023年）：以 Flamingo（Google DeepMind 提出的模型）、BLIP-2（由 Salesforce 提出）为代表。引入了 Q-Former 等结构，将视觉特征转化为 LLM（大语言模型）能理解的 Token。

3. 原生多模态与大模型阶段（2023年至今）：以 LLaVA、GPT-4o、Qwen-VL 为代表。将视觉编码器与强大的 LLM 直接拼接，通过指令微调（Instruction Tuning）让模型具备了极强的理解、推理和对话能力。

CLIP (Contrastive Language-Image Pre-training): 视觉-语言对齐的开山之作。它通过对比学习，让模型知道“狗”这个词和“狗的图片”在特征空间上应该是接近的，从而实现了跨模态的理解。

Q-Former: 由 BLIP-2 提出的一种特殊的 Transformer 结构。它的作用是像“过滤器”一样，从成千上万个视觉特征中提取出对理解指令最有帮助的核心信息，减轻 LLM 的计算压力。

VLA 是 VLM 向机器人控制领域的延伸：

- 1.早期探索： 如 SayCan，利用 LLM 拆解任务，再调用底层的低级控制策略。
- 2.端到端突破（RT-1）： 谷歌推出的 RT-1 证明了 Transformer 可以像处理文本一样处理机器人的传感器数据和动作指令。
- 3.视觉-语言-动作的真正融合（RT-2）： 2023年，谷歌提出了 RT-2。这是第一个真正的 VLA 模型，它直接在视觉语言模型（如 PaLM-E）上进行微调，让模型直接输出机器人控制 Token。
- 4.开源与通用化（2024年至今）： 以 OpenVLA、Octo 为代表，研究重点转向如何在多机型、多任务的大规模机器人数据集上训练通用的“机器人大脑”。

VLM和VLA区别和联系

VLM=视觉编码器+适配器+大预言模型

VLA=动作令牌化+统一预测+闭环控制

特性	VLM (视觉语言模型)	VLA (视觉-语言-动作模型)
主要目标	理解图像内容并生成文本	根据图像和指令控制物理硬件
输出内容	文本、描述、推理过程	机器人动作序列（如：前进、抓取）
数据来源	互联网图文对、视频、书籍	机器人轨迹数据、人类演示数据 (Teleoperation)
具身性	“旁观者”，不与物理世界交互	“执行者”，直接干预物理世界
典型模型	LLaVA, GPT-4o, Claude 3.5	RT-2, OpenVLA, Octo

VLM和VLA区别和联系

传统框架局限：感知、规划、控制模块分离，在空中动态环境中极易产生级联误差。

VLM 的崛起：提供了强大的场景理解和逻辑推理，但缺乏对物理动作的直接输出能力。

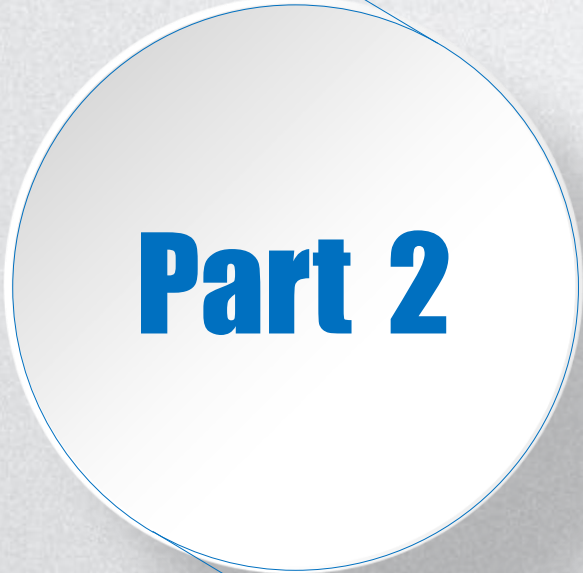
VLA 范式：将多模态输入（视觉+语言+状态）直接映射为底层连续动作，实现“所见即所动”。

VLA 是 VLM 的“进阶形态”。

VLM 是大脑：它负责识别出“桌子上有一个红色的苹果”以及“红苹果通常比绿苹果更甜”。

VLA 是大脑 + 小脑 + 神经：它不仅知道那是红苹果，还知道为了抓起这个苹果，自己的机械臂电机需要转动多少度，并在抓取过程中根据视觉反馈实时调整力度。

目前的趋势是利用 VLM 在互联网海量数据中习得的通用常识，通过动作微调（Action Fine-tuning），将其迁移到具体的机器人硬件上，从而解决机器人领域“数据匮乏”的难题。



Part 2

前沿模型

前沿模型

$\pi 0$ ：首个通用机器人“大脑”

是 PI（Physical Intelligence）公司在 2024 年底推出的初代通用模型，旨在用一个大模型控制任何机器人完成任何任务。

核心原理

基于 Flow Matching (流匹配)：这是 $\pi 0$ 最大的技术特色。传统的模型（如 OpenVLA）通常使用 Diffusion (扩散模型) 来生成动作，但扩散模型计算量大、推理慢。 $\pi 0$ 采用流匹配算法，生成的动作更平滑、频率更高（通常能达到 50Hz 以上），非常适合精细的操纵任务（如折衣服、拿取细小物体）。

大规模预训练：它基于 PaliGemma（谷歌的轻量化 VLM）进行初始化，这让它天生具备了强大的视觉常识。

VLA 架构：它将图像和指令作为输入，直接输出机械臂的关节角度或末端位姿。

前沿模型

π 0.5：面向“开放世界”的进化

是在 2025 年发布的升级版，其核心关键词是 Open-World Generalization（开放世界泛化性）。

核心原理与改进

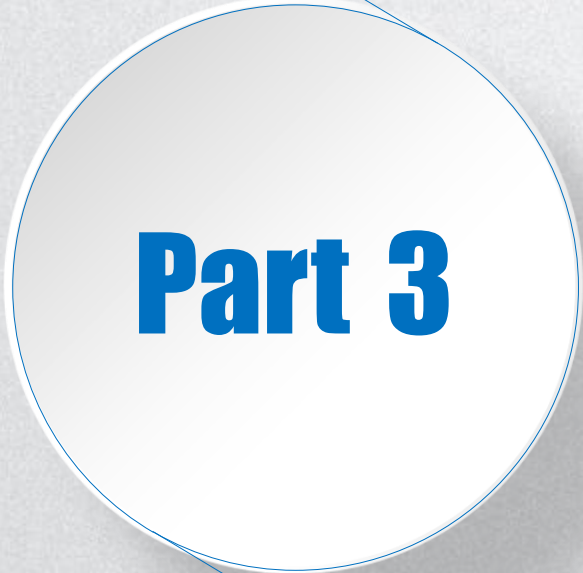
知识隔离 (Knowledge Insulation): 研究者发现，在给 VLM 喂入大量具体的机器人动作数据时，模型原有的“互联网常识”会被冲淡（这种现象叫“灾难性遗忘”）。 π 0.5引入了知识隔离技术，确保模型在学习如何转动 Z1 机械臂的同时，不会忘记“苹果是圆的、红色的”这种常识。

分层架构 (System 1 & System 2)：

高层（语义规划）：像人类的“系统 2”，负责理解复杂的逻辑，比如“清理厨房”。

底层（动作执行）：像人类的“系统 1”，负责低延迟的避障和抓取。

异构任务协同训练： π 0.5在训练时不仅学习动作数据，还同时学习物体的语义检测、子任务分解等。这让它即使在从未见过的家政环境（如陌生的厨房）中，也能靠常识推断出该怎么做。



Part 3



论文介绍

研究背景

视觉-语言-动作（Vision-Language-Action, VLA）模型在地面具身智能领域已取得显著进展，但其在空中操作系统（Aerial Manipulation Systems, AMS）中的应用仍处于起步阶段。AMS 独特的物理特性，包括浮动基座动力学（floating-base dynamics）、无人机与操作手之间的强耦合，以及操作任务的多步长周期（long-horizon）属性，对现有专为二维平面或静态基座设计的 VLA 范式提出了严峻挑战。

为填补这一空白，本研究提出了 AIR-VLA，这是首个为空中操作量身定制的 VLA 基准测试。我们构建了基于物理引擎的仿真环境，并发布了一个包含 3000 条高质量手动遥操作演示记录的多模态数据集。该基准涵盖了基础操作、对象与空间理解、语义推理及长周期规划四大维度。通过系统评估主流 VLA 和 VLM 模型，实验验证了将 VLA 范式迁移至空中系统的可行性。同时，利用为空中任务设计的专用多维指标，揭示了当前模型在无人机机动性、精细控制及高级规划方面的能力边界。AIR-VLA 为通用空中机器人研究提供了标准化的测试台与数据基础。

平台框架

包含四个维度：任务设计分类、评估框架、仿真环境、数据集构建

任务设计分类 (Task Design Taxonomy): 任务平均长度约 475 个时间步，包含四个套件：

基础操作 (Base Manipulation): 评估 3D 悬停、低级运动控制及无人机-机械臂协作。

对象与空间 (Object & Spatial): 测试对物体颜色、材质及三维空间关系的精细理解。

语义理解 (Semantic Understanding): 测试对非结构化、隐性意图指令的遵循能力。

长周期任务 (Long-Horizon): 评估在广域机动与局部操作间切换的多步规划能力。

3.2 评估框架 (Evaluation Framework):

VLA 层: 在线闭环仿真评估，关注空间导航、末端执行器精度及动态安全性。

VLM 层: 针对高级任务规划，评估子任务分解、三维感知及技能选择。

3.3 仿真环境 (Simulation Environment): 基于 NVIDIA Isaac Sim 和 PhysX 5，利用光线追踪确保物理真实性。机器人系统由四旋翼无人机与 7-DoF Franka Panda 机械臂组成，构成 12-DoF 的高维控制问题。

3.4 数据集构建 (Dataset Construction):

数据采集: 采用手动遥操作，记录了涵盖多种协作方案的专家演示。

感知系统: 配备前下视 RGB-D、腕部 RGB-D 及第三方视角相机。

环境多样性: 包含住宅 (Home)、工厂 (Factory) 及户外 (Outdoor) 场景，并模拟了多样的光照条件。



各模型在 AIR-VLA 四类任务套件下的详细性能评估 (归一化分数)

模型	基础操作 (Pos/Arm/Safe/Task/Tot)	对象与空间 (Pos/Arm/Safe/Task/Tot)	语义理解 (Pos/Arm/Safe/Task/Tot)	长周期任务 (Pos/Arm/Safe/Task/Tot)	总平均 (Tot)
ACT	0.0/20.0/100/0.0/ 15.0	0.0/16.0/100/0.0/ 14.0	0.0/16.0/100/0.0/ 14.0	0.0/10.0/100/0.0/ 12.5	13.9
Diffusion Policy	0.0/20.0/100/0.0/ 15.0	0.0/16.0/100/0.0/ 14.0	0.0/16.0/100/0.0/ 14.0	0.0/10.0/100/0.0/ 12.5	13.9
π 0-FAST	27.5/23.9/98.1/0.0/ 22.7	5.1/18.9/98.3/0.0/ 15.8	6.5/16.9/99.3/0.0/ 15.8	41.0/32.6/99.4/0.0/ 28.3	18.8
π 0	56.6/48.0/87.5/8.0/ 38.1	58.1/38.2/75.4/2.0/ 32.4	55.7/36.1/85.1/2.0/ 32.3	65.1/38.5/85.9/2.0/ 32.3	34.5
π 0.5	78.8/49.9/87.2/11.0/ 45.3	70.9/44.1/81.2/13.5/ 42.3	65.4/44.7/82.7/11.0/ 40.2	62.4/40.8/84.8/7.0/ 37.1	42.0

注：加权总分 (Tot) 计算权重为 $w_{pos} = 0.25, w_{arm} = 0.25, w_{safe} = 0.10, w_{task} = 0.40$ 。缩写解释：Pos (位置精度), Arm (机械臂效能), Safe (安全性), Task (任务进度)。

π 0.5 模型在不同干扰条件下的鲁棒性评估

测试条件	<i>Spos</i>	<i>Sarm</i>	<i>Ssafe</i>	<i>Stask</i>	<i>Stotal</i>
标准环境 (无干扰)	71.3	45.9	84.7	10.6	42.0
无人机动 态干扰	68.2 (↓3.1)	45.6 (↓0.3)	83.2 (↓1.5)	10.5 (↓0.1)	41.0 (↓1.0)
无固定第 三方视角	显著下降	显著下降	显著下降	显著下降	显著下降

Spos: 基座定位准确度
Sarm: 机械臂操作效能
Ssafe: 环境安全性
Stask: 任务进度
Stotal: 加权总分

实验与结果分析

主要结果分析：

模型表现： $\pi 0.5$ 与 $\pi 0$ 显著优于 ACT。但在空中场景中，模型常遭遇“空间定位失败”(Spatial Grounding Failure)：虽然识别出物体，却无法解析相对位置。

协作瓶颈： 模型在粗粒度导航上表现尚可，但在机械臂精细操作上遇到技术瓶颈。

安全性： 由于浮动基座特性，碰撞引发的系统扰动远超地面机器人，安全约束仍是核心挑战。

AIR-VLA 填补了空中操作 VLA 研究的空白。实验表明，尽管大规模预训练提供了强大的操作先验，但在应对 3D 运动耦合、精细协作及长周期语义规划方面，现有的 VLA 模型仍有巨大提升空间。本项目提供的标准化基准和 3000 条高质量数据集将为未来的具身智能研究提供有力支持。

请各位师兄批评指正！