基于多模态大模型的智能机器人平台研究与应用

Research and Application of An Intelligent Robot Platform Based on Multimodal Large Models

蒋　维[1]，朱　亮[1]，闵爱佳[1]，厉　智[1]，郭　庆[1]，刘梦雅[1]，谢　磊[2]（1.联通数字科技有限公司，北京 100032；2. 南京大学软件新技术国家重点实验室，南京 210023）

Jiang Wei[1], Zhu Liang[1], Min Aijia[1], Li Zhi[1], Guo Qing[1], Liu Mengya[1], Xie Lei[2] (1. China Unicom Digital Technology Co., Ltd., Beijing 100032, China; 2. State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China)

摘　要：

针对智能机器人系统存在的平台集中管控能力弱、数据闭环断裂和控制策略通用性差问题，提出了一种基于多模态大模型的智能机器人平台架构。智能机器人平台依托格物 AIoT 平台构建标准物模型，统一接入异构机器人实现对多种智能机器人的集中管控，打通“感知—训练—部署”数据闭环；引入 VLA（Vision Language Action）多模态大模型，通过“语言—视觉—动作”的端到端映射直接生成机器人控制指令，提高控制策略的灵活性与泛化性。

关键词：

具身智能；物联网平台；视觉语言动作大模型；智能机器人

doi：10.12045/j.issn.1007-3043.2025.07.003

文章编号：1007-3043（2025）07-0015-06

中图分类号：TP18

文献标识码：A

开放科学（资源服务）标识码（OSID）：

Abstract：

To address the challenges of weak centralized management, broken data loops, and poor generalization of control strategies in intelligent robotic systems, it proposes an intelligent robotic platform powered by the multimodal large model. The platform builds a standard object model based on the Gewu AIoT platform architecture, uniformly connects heterogeneous robots to achieve centralized management and control of various intelligent robots, and achieves the data closed loop of "perception – training – deployment". By integrating a Vision-Language-Action (VLA) multimodal large model, it can directly generate robot control instructions through end-to-end mapping of "language-vision-action", enhancing the flexibility and generalization of control strategies.

Keywords：

Embodied intelligence; Internet of things platform; Vision-language-action large model; Intelligent robot

引用格式：蒋维，朱亮，闵爱佳，等. 基于多模态大模型的智能机器人平台研究与应用[J]. 邮电设计技术，2025(7)：15-20.

1 概述

近年来，以大模型为代表的人工智能（Artificial Intelligence，AI）系统[1]在语言理解、视觉感知和推理生成等任务中取得了突破性进展，推动 AI 逐步从感知认知智能迈向具身行为智能的发展阶段。具身智能（Embodied Intelligence）[2-3] 作为新一代人工智能的重要方向，强调智能体通过身体与环境互动产生智能行为完成“感知—理解—推理—执行”的闭环交互，目前已成为机器人、智能制造与人机协作等领域的核心技术路径。在此背景下，智能机器人作为具身智能的重要载体，逐渐成为推动具身智能落地应用的关键力量，引发了学术界与产业界的高度关注。

我国高度重视人工智能，尤其是智能机器人方向的发展。2024 年，“人工智能+”首次被写入政府工作报告，标志着人工智能与实体经济融合的重要性进一步提升。2025 年的政府工作报告首次将“具身智能”纳入国家重点发展方向，提出要加快突破人形机器人、具身智能等关键技术，指出要大力培育具身智能、6G 等未来产业，持续推进“人工智能+”行动，大力发展

收稿日期：2025-06-04

人工智能手机和电脑、智能机器人等新一代智能终端及智能制造装备。

尽管人工智能在多个领域取得了长足进展,但现有智能机器人系统仍存在两大瓶颈。一方面,缺乏统一平台支撑的数据闭环和集中管控能力。现有机器人系统中,训练数据采集、模型训练与终端部署常处于割裂状态,模型难以快速迭代优化,需要专用的平台对异构智能机器人进行管理控制,限制了多机器人的协同部署。另一方面,缺乏通用、端到端的智能机器人控制模型,系统灵活性不足。传统机器人系统依赖规则系统或功能函数库进行控制逻辑拆解,难以适应复杂、动态的复杂环境,不同类型的机器人往往需要独立开发控制策略,导致控制方案高度定制化,通用性差。

本文设计了基于多模态大模型的智能机器人平台架构以解决上述问题。一方面依托格物 AIoT 平台[4]在海量设备接入、物模型建模[5]与数据统一管理等方面的技术能力,构建了面向异构智能机器人的统一管控平台,使异构机器人的协同部署成为可能,打通了数据采集、模型训练与终端部署的链路,有效提升了模型迭代效率;另一方面引入 VLA 多模态大模型进行智能机器人控制,该模型融合语言理解、视觉感知与动作控制,实现了自然语言指令到具体动作的端到端控制。基于多模态大模型的智能机器人平台架构提升了模型迭代效率,还有效增强了机器人控制策略的泛化性和跨设备适应能力,为多机器人协同控制的落地提供了支撑。

2 智能机器人发展现状

智能机器人(Intelligent Robot)是指具有感知、认知、决策和执行能力,能够自主或半自主地完成复杂任务,并能与环境进行有效交互的机器人系统。近年来,智能机器人已广泛应用于工业、服务、医疗等多个场景,展现出强大的环境感知、自主决策与任务执行能力。随着感知技术、边缘计算与人工智能的发展,机器人从“自动执行”逐步向“自主协作”演进,智能化水平不断提升。然而,要真正实现面向大规模实际场景的广域部署与持续优化,仍依赖于强有力的基础支撑平台与高效灵活的控制方法。

随着智能机器人技术的快速发展,多机器人的集成管理成为智能机器人大规模落地的关键。早期机器人多依赖专用系统进行单设备管理而难以支持多设备统一调度。随着物联网和人工智能技术的进步,人们开始尝试通过数字孪生、物模型等技术实现设备描述和数据接口的统一,以提高系统的集成与智能化水平。但目前不同机器人系统大多仍依赖独立平台,导致跨设备协同与管理受限。此外,仍存在训练数据采集、模型训练与部署互相独立,模型迭代缓慢的问题。

在控制方法层面,早期的智能机器人多采用模块化架构进行控制,各个定制的子模块之间通过接口离散连接。虽然这种方式具有逻辑可解释的优点,但是系统高度定制化不易泛化,且模块间互相依赖容易产生误差。近年来,随着大语言模型(Large Language Model,LLM)和视觉基础模型的发展,研究者们尝试将大语言模型引入智能机器人系统,通过自然语言对任务进行表达,由视觉模型编码分析环境状态,再由大语言模型理解任务并匹配预定义函数进行调用。例如,SayCan[6]、Code-as-Policies[7] 等方法就采用了“语言理解+行为函数库”的执行策略。这类方法虽然提升了系统的泛化能力,但由于控制链路中仍依赖预定义函数库,大语言模型仅作为决策器,没能实现从感知到控制指令的真正端到端控制系统。

格物 AIoT 平台和 VLA 大模型的出现为实现智能机器人的集中管理和端到端灵活控制提供了契机。格物 AIoT 平台[4]为设备提供安全可靠管理能力,是一个可以快速连接设备和企业系统的物联网平台。向下接入分散的物联网传感器,汇集传感数据,具备多种传感器兼容能力;向上面向应用层服务提供应用开发的基础性平台和面向底层网络的统一数据接口。格物平台提供了如设备管理、规则引擎、物网管理等功能,具有高可靠性、高并发、低时延、数据处理功能强等优势,因此不仅可以为上层应用提供丰富的开发工具和服务,还能够支持大规模设备的统一管理与协同调度,有效缓解设备管理分散的问题,可广泛应用于工业、运输、智慧城市等领域。VLA 多模态大模型通过构建统一的语义空间将自然语言指令与视觉场景信息进行跨模态对齐,可以根据视觉、自然语言指令直接推理生成结构化动作指令,提升了机器人控制系统的灵活性和泛化能力。VLA 多模态大模型的端到端控制能力结合格物 AIoT 平台的强大设备管理与数据服务能力为智能机器人控制和集中管理方法的升级提供了新的契机,有望实现从感知数据采集、模型训练到终端部署的闭环协同,推动智能机器人系统

向更高效、统一和智能化方向发展。

3 基于多模态大模型的智能机器人平台架构

本文研究并实现了一个基于多模态大模型的智能机器人平台架构。该平台架构创新性地利用格物AIoT平台的标准机器人物模型统一了异构机器人的数据接口，从而实现对异构智能机器人进行集中管控，实现了跨具身[8]和可泛化；平台实现了多源感知数据的自动对齐与标记，并且通过格物AIoT平台构建了完整的“感知—训练—部署”闭环系统，解决了数据闭环断裂导致的迭代延迟问题；通过设计部署VLA多模态大模型实现了端到端的机器人控制避免模块化依赖，从而使得智能机器人控制方法具有灵活性。多模态大模型和格物AIoT平台的结合使得智能机器人系统具有强泛化性，解决了传统定制化平台方案通用性差、协同不足的问题。

3.1 基于格物AIoT平台的数据闭环和集中管控

如图1所示，本文基于格物AIoT平台构建了一套面向智能机器人的数据闭环与集中管控体系。构建了集数据采集、模型训练与统一部署于一体的闭环系统，并基于平台内构建的标准化机器人物模型，实现对多类型异构机器人的统一描述与集中控制，为大规模、跨具身的智能机器人系统落地提供平台支撑。

在集中管控方面，本文利用了格物AIoT平台的机器人标准物模型机制，对异构机器人进行统一建模与调度管理。物模型是设备的标准数字化映射，抽象定义了机器人设备的状态属性（如关节角度、姿态信息等）、行为能力（如移动、抓取、旋转等操作）和事件机制（如任务完成、故障报警等），通过语义化建模将不同类型机器人统一映射至平台通用的通信接口之上。借助这一机制，平台能够实现对具身形式各异的智能机器人的统一接入、集中调度与状态管理（如图1中跨机器人标准物理型数据上报所示），有效解决异构设备接口不兼容、运维割裂等问题，推动多机器人协同控制在复杂应用场景中的可行性与可扩展性。

在数据闭环方面，本文依托格物AIoT平台构建了可以支撑大模型持续迭代的“感知—训练—部署”闭环流程。该流程始于用户通过平台操控智能机器人本体执行标准化任务。平台通过格物OS的端侧智能分析能力，自动采集并上报任务执行过程中的多模态感知数据（图像、音频、位姿等），这些原始数据首先进入平台的设备管理平台（Device Management Platform，DMP），经历协议解析、数据预处理、物理型数据解析等环节，并完成数据的时序对齐处理，确保不同传感器数据流（如图像、音频、位姿）拥有统一精确的时间戳。AIoT模型应用服务平台内置了自动化标注机制，当机器人执行特定控制指令时，平台同步记录该指令的语义信息，并自动绑定为同时产生的多模态感知数据的语义标签，形成高质量的指令—数据对。这些经过自动标注和精准对齐的高质量数据为训练大模型奠定了坚实基础。

训练完成的模型随后进入支撑管理平台（Backbone Management Platform，BMP）。BMP负责业务部署与优化，包括端边云协同、模型轻量化、模型推理适配转换以及仿真与测试等，确保模型能高效运行在资源受限的终端设备上。最终，平台提供模型快速下发功

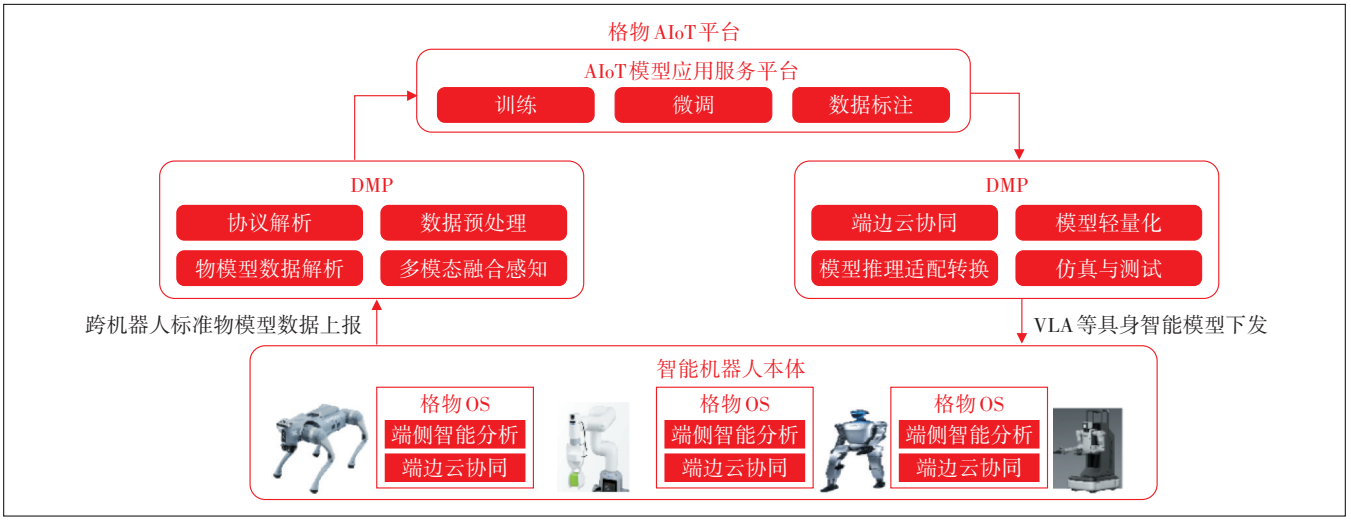

图1　基于格物AIoT平台的集中管控与数据闭环

能，将优化后的模型快速部署到智能机器人本体的格物OS环境中，实现端侧智能分析能力的即时更新与闭环验证。新部署的模型在机器人上运行，又会持续产生新的感知数据，从而开启下一轮迭代。这一从真实场景数据采集、DMP处理与数据标注、平台训练/微调、BMP优化到模型部署的全流程自动化闭环，显著缩短了模型迭代周期，成为驱动机器人智能持续进化的核心引擎。

综上所述，格物AIoT平台通过其标准化物模型实现了对异构机器人的集中统一管控，并通过构建数据闭环为智能机器人的大规模应用与持续智能化升级提供了强大的平台级支撑。

3.2 VLA多模态大模型的构建

考虑到VLA多模态大模型在跨模态语义对齐与任务泛化方面的潜力，本文使用VLA多模态大模型来克服传统智能机器人控制方法在泛化能力与交互灵活性方面的局限，构建面向实际场景的端到端智能机器人控制系统，推动智能机器人系统从模块化向端到端范式演进以提高控制系统的泛化性。该系统将自然语言指令与场景图像联合输入至VLA多模态大模型，直接输出结构化的动作指令来驱动智能机器人完成任务动作。相较于传统的多阶段处理流程，该方法简化了系统结构，降低了感知误差传播带来的执行偏差，同时大幅提升了对开放环境和复杂指令的理解与响应能力，使得控制方法泛化性强、可跨具身。

具体来说，本文借鉴OpenVLA[9]和OpenVLA-OFT[10]的思路实现了由多模态大模型驱动的智能机器人端到端控制系统。通过构建统一的多模态嵌入空间，并采用对比学习策略强化匹配样本之间的相似性，实现了语言、图像和动作等异构模态信息的高效对齐，解决了多模态语义对齐与联合建模的复杂性的挑战。为应对多模态推理结果向底层控制命令的映射能力不足的挑战，本文通过端到端微调机制，在控制链路上对大模型进行适配训练，使其能够直接生成与硬件平台高度匹配的控制参数，从而有效避免模块化架构中因阶段性误差积累带来的性能退化问题，提升了控制的精度与系统整体稳定性。

VLA多模态大模型架构如图2所示。

首先，大模型整体结构由分词器、视觉编码器、多层感知机映射器、语言模型主干和动作解码器5个部分构成。其中，分词器将原始语言指令切分并映射为离散的token序列e_{text}，作为大模型的输入以进行后续处理；视觉编码器采用SigLIP[11]和DinoV2[12]2种预训练视觉模型提升模型的空间推理能力，分别提取当前环境观测图像x_{image}的特征后进行拼接得到视觉表示e_{image}；多层感知机映射器将融合后的视觉特征投射至语言模型的嵌入空间，使多模态信息在同一语义空间中对齐，便于大语言模型理解图像内容；大语言模型主干f_{fuse}采用Llama 2接收图像与语言的编码表示并生成多模态任务表示e_{task}：

$$e_{\text{task}} = f_{\text{fuse}}(e_{\text{text}}, e_{\text{image}}) \tag{1}$$

该表示向量作为动作解码器的输入，最终预测出一系列智能机器人控制指令：

$$\hat{a} = \{a_1, a_2, \cdots, a_T\}, \quad a_t \in \mathbb{R}^d \tag{2}$$

其次，VLA多模态大模型训练数据采用人工演示的方式采集。每条样本包含语言指令、图像观测及对应的专家动作轨迹，形式为：$(x_{\text{text}},\ x_{\text{image}},\ a)$。

在训练阶段，机器人控制任务被转化为语言建模任务和离散化机器人动作并映射到语言模型的token空间。模型基于图像和语言指令进行微调，使用标准的下一个token预测目标，最小化预测动作token的交叉熵损失，从而学习如何在给定视觉和语言输入的条件下生成机器人动作指令。

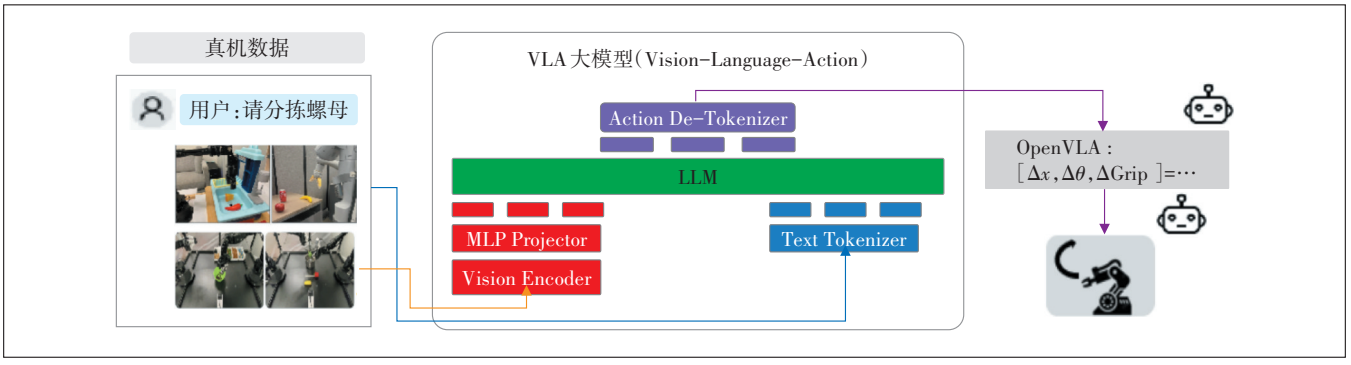

图2　VLA多模态大模型架构

最后，在模型部署阶段，得益于语言模型结构的高度通用性，VLA多模态大模型可实现跨任务、跨平台的迁移，不依赖于平台级重训练。同时，系统支持低秩适配（LoRA）等轻量化微调方式，仅需更新极少量参数即可快速适应新任务或新硬件环境，具备良好的可扩展性与应用灵活性。

3.3 多模态大模型驱动的智能机器人控制总体方案

在前述VLA多模态大模型基础上，本节构建了以VLA多模态大模型为核心的智能机器人控制总体方案，实现了从用户语音指令接收、VLA多模态大模型理解并推理、智能机器人控制执行、再到数字孪生仿真的全流程闭环控制。方案整体流程如图3所示。

Step1：多模态输入与感知。主要包括语音输入（采集用户语音指令〔如“将圆柱体积木移动到中国联通标志处”〕）、语音识别（将语音转写为文本指令为大模型提供语言输入）和环境图像采集（采集场景图像作为视觉输入）。

Step2：VLA多模态大模型处理。将语言指令和图像输入VLA多模态大模型，模型通过融合视觉和语言信息，理解任务意图和环境状态并在统一的语义空间中进行推理，端到端生成智能机器人控制指令序列。

Step3：智能机器人控制执行。智能机器人执行模型生成的控制指令。

Step4：数字孪生仿真。在智能机器人执行任务过程中，内置传感器持续采集关键运行状态（如末端位置、关节角度、负载等），并将数据实时回传至格物AIoT平台。格物AIoT平台根据实时数据驱动虚拟智能机器人模型同步运行，实现操作状态的三维仿真可视化。

通过将自然语言理解、视觉感知与动作指令生成整合到VLA多模态大模型中，本方案实现了对复杂任务的端到端理解与控制，进一步提升了系统在工业场景中的实用性和智能水平。

4 智能机器人平台构建与应用效果

本章将介绍智能机器人平台的集中管控能力，并以智能机械臂为例，介绍基于VLA多模态大模型的智能机器人控制的完整操作流程与效果。平台通过语义化建模将不同类型机器人统一映射至平台通用的通信接口之上，实现了对双足人形交互机器人、轮式酒店服务机器人、臂式分拣机器人等的集中管控。对于所管控的机器人，平台可以实时地采集传感器数据并进行状态监控，当智能机器人状态出现异常时及时发出告警。基于VLA多模态大模型的智能机器人控制流程如图4所示，系统初始化完毕后，进入用户指令待命状态。当用户发出自然语言操作指令：“请将圆形木块移动到中国联通Logo的上方”时，系统获取语言指令和环境信息，并使用VLA多模态大模型进行任务理解和指令推理，机械臂接收到指令后执行相应的移动与抓取操作，同时将执行过程中的传感器数据实时同步至格物DMP平台。数字孪生系统则基于同步数据，动态渲染出虚拟现实可视化仿真界面。最终，圆形木块被准确移动至中国联通Logo的上方。系统支持对任意属性目标物的识别与操作指令响应，例如当用户下达“请将红色木块移动到蓝色Logo上方”的指令时，机械臂能够准确理解语义内容，完成对红色

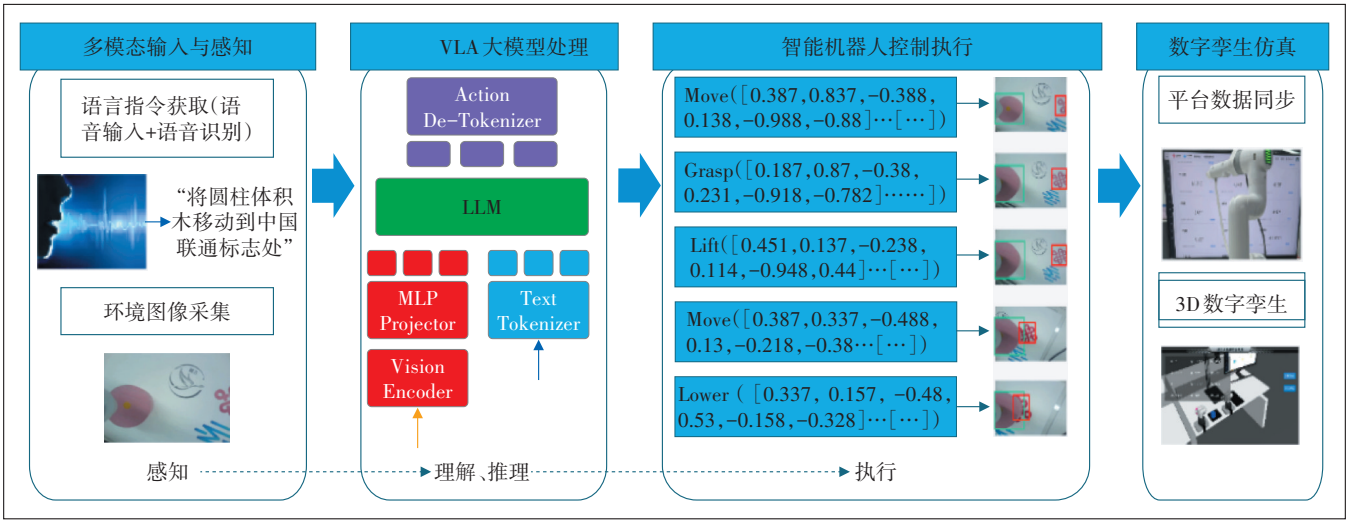

图3　VLA多模态大模型驱动的智能机器人控制总体方案设计

图4　智能机器人控制系统操作效果

木块的抓取与搬运，并将其放置于蓝色联通数科Logo上方。以上实验结果表明，本文提出的智能机器人平台可以实现对异构智能机器人的集中管控和状态监测，具有良好的泛化性，由VLA多模态大模型驱动的智能机器人控制方法具备端到端任务指令生成能力，能够根据用户语言灵活识别目标属性并完成复杂操作任务，展现出良好的通用性与智能性。

5 总结

本文针对智能机器人系统中存在的数据闭环断裂、管控分散和控制通用性差等问题，提出了一种基于多模态大模型的端到端智能机器人平台架构。智能机器人平台依托格物AIoT平台构建标准物模型，实现了异构机器人的统一接入与集中管理，打通“感知—训练—部署”全流程数据闭环；平台通过引入VLA视觉—语言—动作多模态大模型，支持智能机器人的语言与视觉输入到控制指令的端到端映射，提升了控制策略的泛化性与灵活性。实验结果表明，该平台能够实现多机器人协同控制与状态监测，具备良好的跨设备适应能力，所提出的VLA大模型能够精准执行复杂自然语言指令，具有良好的泛化性，本研究为智能机器人的规模化部署与灵活控制提供了有力支撑。

参考文献：

[1] 施志雄，段该甲，马龙轩，等. 基于大语言模型命名实体识别的AI智能问答优化[J]. 邮电设计技术，2025(3)：80-84.
[2] 王文晟，谭宁，黄凯，等. 基于大模型的具身智能系统综述[J]. 自动化学报，2025，51(1)：1-19.
[3] 张伟男，刘挺. 具身智能的研究与应用[J]. 智能系统学报，2025，20(1)：255-262.
[4] 中国联通. 格物AIoT平台[EB/OL]. [2025-05-12]. https://dmp.cuiot.cn//.
[5] 蒋维，王延红，朱亮，等. 基于大模型的工业互联网物模型智能生成技术研究[J]. 邮电设计技术，2024(11)：19-24.
[6] AHN M，BROHAN A，BROWN N，et al. Do as i can，not as i say：grounding language in robotic affordances[EB/OL]. [2025-02-09]. https://arxiv.org/abs/2204.01691.
[7] LIANG J，HUANG W L，XIA F，et al. Code as policies：language model programs for embodied control[C]//2023 IEEE International Conference on Robotics and Automation (ICRA). London：IEEE，2023：9493-9500.
[8] DOSHI R，WALKE H，MEES O，et al. Scaling cross-embodied learning：one policy for manipulation，navigation，locomotion and aviation[EB/OL]. [2025-02-09]. https://arxiv.org/abs/2408.11812.
[9] KIM M J，PERTSCH K，KARAMCHETI S，et al. Openvla：an open-source vision-language-action model [EB/OL]. [2025-02-09]. https://arxiv.org/abs/2406.09246.
[10] KIM M J，FINN C，LIANG P. Fine-tuning vision-language-action models：optimizing speed and success [EB/OL]. [2025-02-09]. https://arxiv.org/abs/2502.19645.
[11] ZHAI X H，MUSTAFA B，KOLESNIKOV A，et al. Sigmoid loss for language image pre-training[C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris：IEEE，2023：11941-11952.
[12] OQUAB M，DARCET T，MOUTAKANNI T，et al. Dinov2：learning robust visual features without supervision[EB/OL]. [2025-02-09]. https://arxiv.org/abs/2304.07193.

作者简介：
蒋维，高级工程师，博士，主要从事人工智能和工业互联网平台研发工作；朱亮，工程师，博士，主要从事人工智能研发工作；闵爱佳，高级工程师，硕士，长期从事物联网终端产品研发工作；厉智，硕士，主要从事人工智能研发工作；郭庆，博士，主要从事移动计算、人工智能相关研究工作；刘梦雅，硕士，主要从事人工智能研发工作；谢磊，教授，博士生导师，博士，主要从事物联网、移动计算研究工作。