基于大语言模型的机械臂系统

2021251124　古翱翔

1　背景

1.1 背景介绍

近年来,以 ChatGPT[①]为代表的**大语言模型(LLMs)**[②]在自然语言处理领域取得了飞速进展,在各类语言任务中表现卓越[1,2]。同时随着**多模态模型**(如 CLIP、DALL-E、SORA 等)的兴起,研究人员开始探索将视觉等不同模态信息相结合,以更好地理解和解释复杂的真实世界。这也为将 LLMs 应用于机器人领域奠定了基础[3,4]。

传统的机械臂控制系统普遍依赖预编程规则或视觉识别技术,在处理复杂指令和动态环境时存在明显局限[3]。而集成 LLMs 有望赋予机械臂**诸多增强能力**,如:

1)自然语言理解 机械臂能够解析复杂语言指令,通过简单语言控制实现复杂任务[3]。

2)任务规划 对高级任务进行分解规划,提高机械臂在多步骤任务中的灵活性和效率[3]。

[①]**GPT:**Generative Pre-trained Transformer

[②]**大模型:**指拥有巨大参数量的机器学习模型

3)多模态感知与推理 结合视觉、语言等信息,机械臂对环境的理解和推理能力将大幅提升,适应性和鲁棒性也随之增强

4)人机交互 自然语言界面和对话系统将极大改善人机交互体验,实现高效协作和沟通[5]。

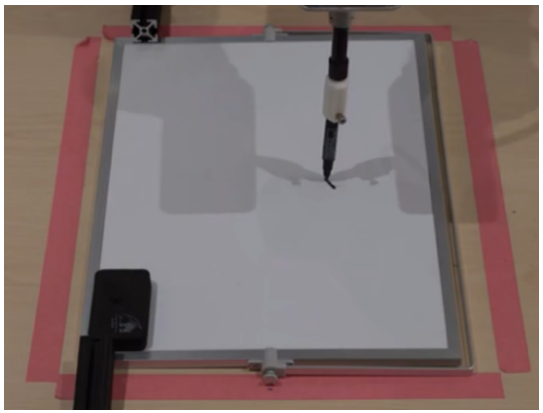
将这一领域又被称作 **具身智能**[6]，是人工智能下一个风口。

1.2 研究现状

目前,多个团队已开展相关研究,取得了初步进展。主要来自各大高校和研究机构，包括清华大学、北京大学、麻省理工学院（MIT）、斯坦福、谷歌、字节跳动等。我们根据使用方法的不同，可以分为以下几类

1.2.1 LLM+CODE

使用 LLM 来完成复杂任务的分解和规划，调用封装好的机器人函数来完成分解后的动作。这部分工作主要有 **Code as Policies** (谷歌，2023)[7] ， **SayPlan**(2023，使用 3D 场景图)[8] 、**ReKep**（2024, 斯坦福李飞飞组, 关系关键点约束）[9]、**ManipVQA**(2024, 北京大学黄思源组, 模型微调)[10] 等 、**Instruct2Act**(2023，北京大学黄思源组，SAM+CLP)[11] , **VoxPoser**(斯坦福李飞飞组，2023)[12]



视频 1.1: Code as Policies example

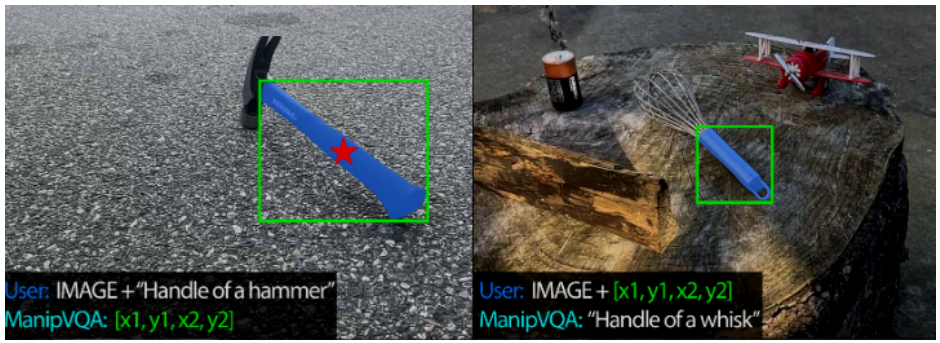


图 1.1: ManipVQA 模型统一的 VQA 格式生成预测

1.2.2 端到端

端到端最早的代表工作有 ACT(2023，斯坦福大学 Mobile Aloha)[13,14] 和 Diffusion[3] Policy (2023,MIT,哥伦比亚大学)[15]。这部分工作没有做过多假设，理论上什么都可以做，但是由于它模型比较小，训练数据少，所以实际上只能用于几个特定的任务，泛化性差。

[3]Diffusion 是一种生成方法，如今图像生成领域的成就基本都是基于 Diffusion 方法，比如 Stable Diffusion 和 Midjourney

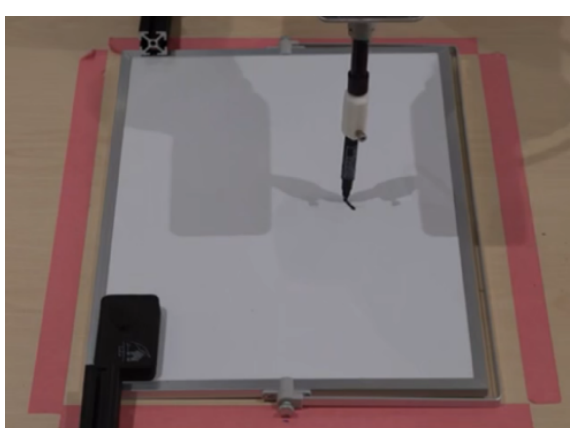
一种可能的解决的方案是使用模仿学习/强化学习，将模型训练的大一些，这部分工作主要有

Transformer+MSE 直接回归 GR-1(2023,字节跳动)[16],GR-2(2024,字节跳动)[17]

Transformer+Diffusion Head Octo(2023,伯克利、斯坦福)[18] HPT(2024,MIT 何恺明组)[19]

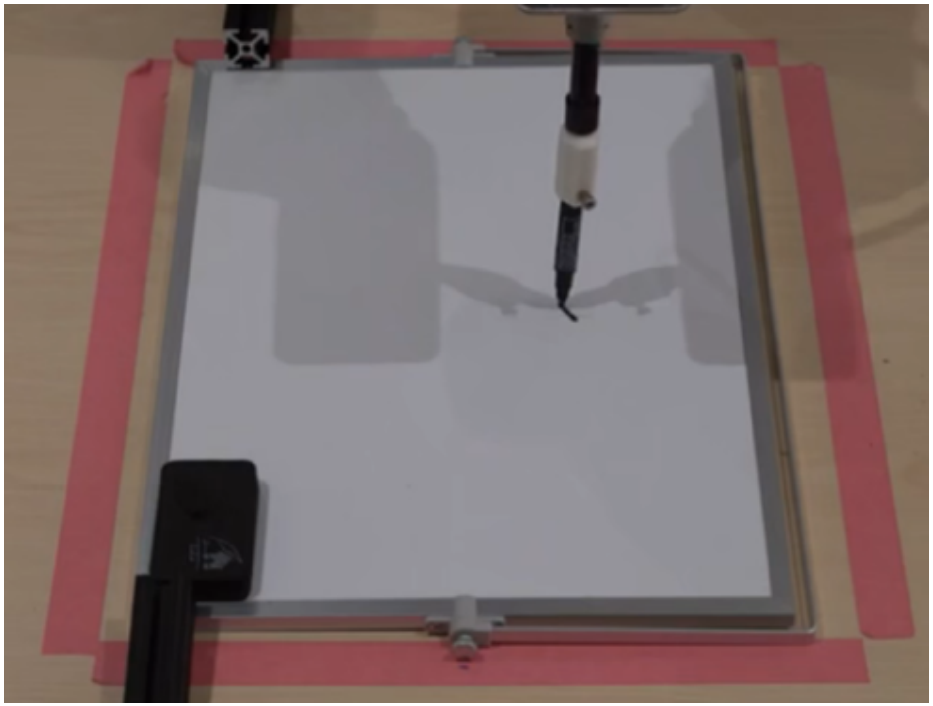
Transformer+Discretized Token RT-1,2(2022-2023,谷歌)[7,20] ,OpenVLA(2024,斯坦福、伯克利)[21]

ELSE RDT(2024 10.10，清华)[22]，π_0 (2024 10.24 Physical Intelligence(π))[23]

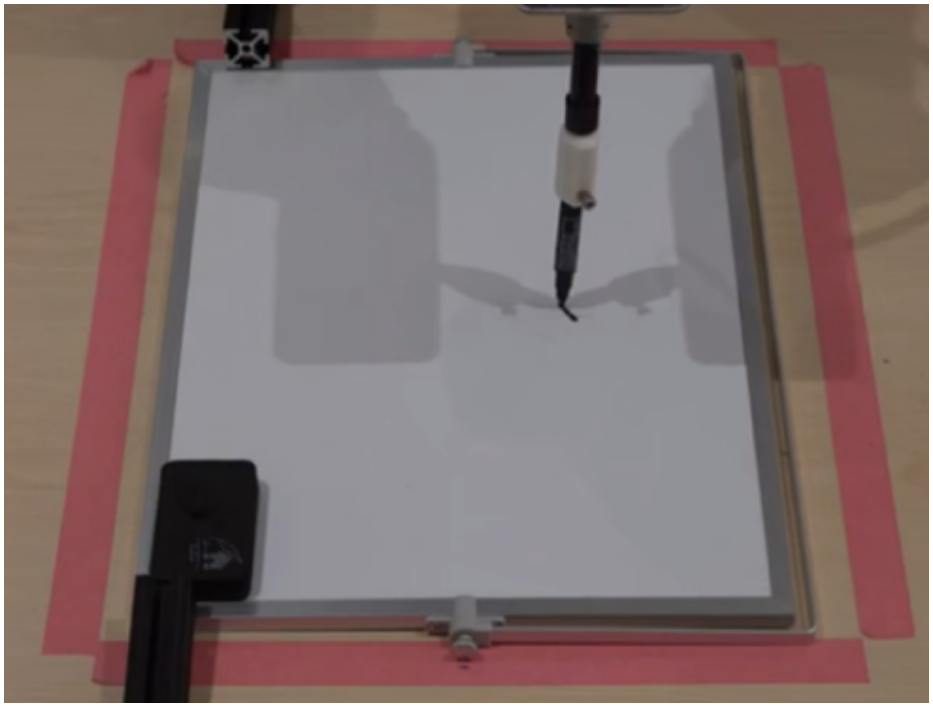


视频 1.2：π_0 example

视频 1.3：RDT example



视频 1.4：Diffusion Policy 将积木推到 T 型框内



视频 1.5：Diffusion Policy 涂抹酱

2　方案拟定

2.1 系统框架

我们将调研、综述并分析相关技术发展，最终开发一个基于大语言模型的仿真系统，使机械臂能够在语音指令的控制下自主移动，完成各种指令。

本课题应完成的工作主要包括**仿真、软件和交互**三大模块的开发。其中

1. 仿真模块搭建一个支持机械臂的仿真平台。
2. 软件模块利用大语言模型和视觉模型处理自然语言指令和 RGBD 图像，生成合理的操作步骤。
3. 交互模块开发语音指令解析与反馈系统，实时检测和解析用户语音，将其转化为自然语言指令并反馈或进一步询问细节。

2.2 交互模块

交互模块的核心功能包括自动语音识别（ASR）、大语言模型（LLM）和文本转语音（TTS）[24]。这部分我们也将综述现有的技术，着重关注于其实时性和准确性。

ASR Whisper（OpenAI）[25]、DeepSpeech[26] 等。

TTS ChatTTS 等。

音频 2.1：使用 Chattts 生成的语音

图 2.1：生成语音代码示例

2.3 仿真模块

而 Sim-to-Real 的研究更是方兴未艾。我们计划初步使用 Webots 与 ROS 搭建仿真环境，以 UR5 机械臂及其夹爪为主体，构建用于模拟机械臂运动与操作的仿真平台。

与此同时，我们希望收集并整理 Sim-to-Real 领域的高质量综述文章，深入了解其最新进展与挑战。同时，将尝试文中提到的开源仿真环境和方法，为后续工作提供实践支持与改进方向。

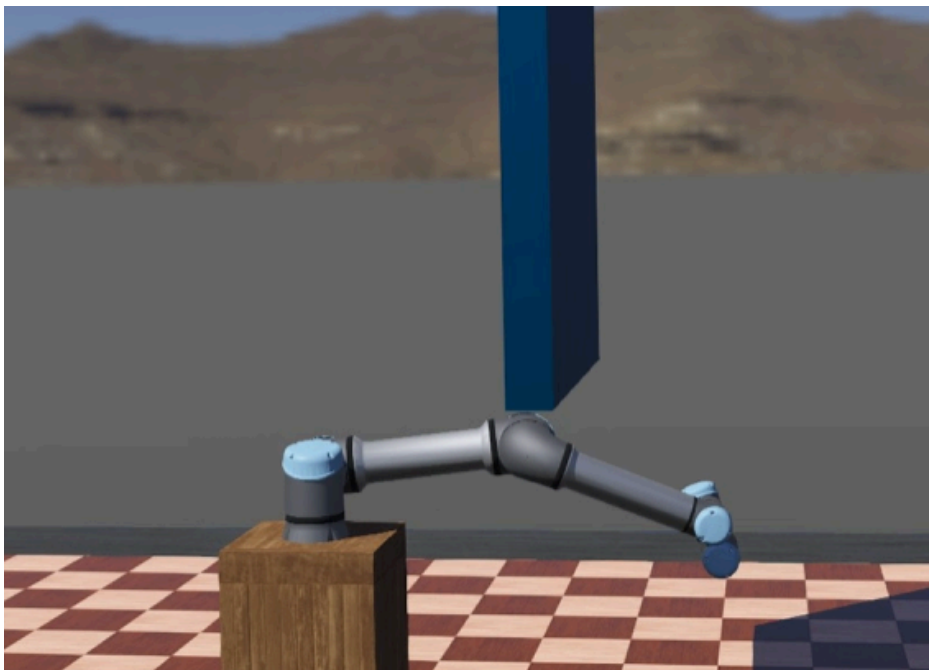


图 2.2：Webots UR5 机械臂仿真环境测试

2.4 软件模块

在本部分工作中，我们计划对一些端到端的大模型（如 RDT、GT-2、Octo 等）进行微调，并将其部署在仿真环境中进行效果测试。

同时，我们将复现基于 LLM 与 API 的相关代码，例如 Code as Policies 和 Instruct2Act 等，进一步探索这些方法在我们的场景中的应用。最终，我们将基于这些实验结果，确定软件部分的系统框架，并进行优化。

3　进度安排

3.1 进度安排

时间	内容安排 1	内容安排 2
第 1-2 周	查阅文献，明确研究目标与技术方案	搭建开发环境
第 3-4 周	完成仿真环境和机械臂模型的搭建与测试	实现开源项目微调及环境验证
第 5-6 周	开发语音交互模块并初步整合	搭建并测试软件模块框架
第 7-8 周	完善软件模块并整合全系统	运行全流程测试并优化关键问题
第 9-10 周	全面优化系统和代码，生成开发文档	测试多种指令场景并记录实验结果
第 11-12 周	撰写实验报告并对比现有算法	根据进度决定是否进行实物实验
第 13-14 周	整理实验数据并撰写论文	准备答辩材料并进行预演

参考文献

1. Vemprala, S., Bonatti, R., Bucker, A. & Kapoor, A. ChatGPT for Robotics: Design Principles and Model Abilities. (2023)
2. Achiam, J. *u. a.* Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023)
3. Li, J. *u. a.* MMRo: Are Multimodal LLMs Eligible as the Brain for In-Home Robotics?. (2024)
4. Firoozi, R. *u. a.* Foundation Models in Robotics: Applications, Challenges, and the Future. (2023)
5. Lykov, A. & Tsetserukou, D. LLM-BRAIn: AI-driven Fast Generation of Robot Behaviour Tree Based on Large Language Model. (2023)
6. 中国信息通信研究院、北京人形机器人创. 具身智能发展报告（*2024* 年）. (2024)
7. Liang, J. *u. a.* Code as Policies: Language Model Programs for Embodied Control. (2023)
8. Rana, K., Haviland, J., Garg, S., Abou-Chakra, J. & Reid, I. SayPlan: Grounding Large Language Models Using 3D Scene Graphs for Scalable Robot Task Planning.
9. Huang, W., Wang, C., Li, Y., Zhang, R. & Fei-Fei, L. ReKep: Spatio-Temporal Reasoning of Relational Keypoint Constraints for Robotic Manipulation. (2024) doi:10.48550/arXiv.2409.01652
10. Huang, S. *u. a.* ManipVQA: Injecting Robotic Affordance and Physically Grounded Information into Multi-Modal Large Language Models. (2024) doi:10.48550/arXiv.2403.11289

11. Huang, S. *u. a.* Instruct2Act: Mapping Multi-modality Instructions to Robotic Actions with Large Language Model. (2023) doi:10.48550/arXiv.2305.11176
12. Huang, W. *u. a.* VoxPoser: Composable 3D Value Maps for Robotic Manipulation with Language Models. (2023)
13. Fu, Z., Zhao, T. Z. & Finn, C. Mobile ALOHA: Learning Bimanual Mobile Manipulation with Low-Cost Whole-Body Teleoperation. (2024) doi:10.48550/arXiv.2401.02117
14. Zhao, T. Z., Kumar, V., Levine, S. & Finn, C. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware. (2023) doi:10.48550/arXiv.2304.13705
15. Chi, C. *u. a.* Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. (2024) doi:10.48550/arXiv.2303.04137
16. Wu, H. *u. a.* Unleashing Large-Scale Video Generative Pre-training for Visual Robot Manipulation. (2023) doi:10.48550/arXiv.2312.13139
17. Cheang, C.-L. *u. a.* GR-2: A Generative Video-Language-Action Model with Web-Scale Knowledge for Robot Manipulation. (2024) doi:10.48550/arXiv.2410.06158
18. Team, O. M. *u. a.* Octo: An Open-Source Generalist Robot Policy. (2024)
19. Wang, L., Chen, X., Zhao, J. & He, K. Scaling Proprioceptive-Visual Learning with Heterogeneous Pre-trained Transformers. (2024) doi:10.48550/arXiv.2409.20537
20. Brohan, A. *u. a.* RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. (2023) doi:10.48550/arXiv.2307.15818

21. Kim, M. J. *u. a.* OpenVLA: An Open-Source Vision-Language-Action Model. (2024) doi:10.48550/arXiv.2406.09246
22. Liu, S. *u. a.* RDT-1B: A Diffusion Foundation Model for Bimanual Manipulation. (2024)
23. Black, K. *u. a.* $Pi0$: A Vision-Language-Action Flow Model for General Robot Control.
24. Cui, W. *u. a.* Recent Advances in Speech Language Models: A Survey. (2024) doi:10.48550/arXiv.2410.03751
25. Radford, A. *u. a.* Robust Speech Recognition via Large-Scale Weak Supervision. (2022) doi:10.48550/arXiv.2212.04356
26. Hannun, A. *u. a.* Deep Speech: Scaling up End-to-End Speech Recognition. (2014) doi:10.48550/arXiv.1412.5567

感谢聆听，请老师批评指正！