达闼杯

机器人大模型与具身智能挑战赛

- 将大语言模型纳入机器人系统 -

2023 年 8 月 1 日

一、【大赛简介】

机器人应用是人工智能时代最具有挑战性的前沿科学技术难题之一，它汇集了人工智能和机器人核心技术，包括人工智能学的智能感知、认知和决策的各种算法能力，以及机器人学在传感器、控制器和执行器的高可靠、高精确的运动和控制能力。

预训练大模型 GPT 技术的突破，可以为机器人提供智慧的大脑。同时，具身智能机器人是一个具有物理实体、可与真实世界进行多模态交互，像人类一样感知和理解环境，并通过自主学习完成任务的智能体。二者的结合将使机器人做到“心灵手巧”。

达闼云端机器人国家新一代人工智能开放创新平台与中国计算机学会、AITISA 联盟、OpenI 启智、CCF 开源发展委员会、北京大学、复旦大学、北京航空航天大学、中山大学、北京邮电大学、北京智谱联合，共同举办达闼杯“机器人大模型与具身智能挑战赛”，旨在鼓励开发者能够利用大模型技术和具身智能技术，实现跨模态人机交互、并能自主完成各种复杂任务的机器人应用。

为此，我们提供一个比赛平台，在模拟环境中实现机器人（Agent）的各项任务。基于比赛平台，参赛者可以结合大规模视觉和语言预训练模型（LLMs， CLIP 等）算法来训练机器人完成复杂的任务。大赛鼓励参赛者采用各种人工智能技术和工具，包括但不限于大语言模型（LLMs）、多模态视觉理解模型、深度学习和强化学习，来提高机器人的具身智能能力。

“机器人具身智能挑战赛”作为 CCF 中国软件大会的赛事内容，向来自学术界、工业界和研究机构的团队开放。参赛者需要提交详细的技术报告，描述他们系统的设计、实现和评估。通过本次比赛，参赛者既可以学习和提升人工智能和机器人领域的知识与技术，还可以与顶级专家直接交流，共同探讨通用人工智能的未来发展。另外，大赛配备了免费的算力支撑，各路勇士们，还在犹豫什么，快快组队参赛吧！

二、【赛事亮点】

1、环境、智能体和任务

赛事选用日常生活中具有普遍性的咖啡厅场景，由经过训练的机器人自主完成咖啡厅服务员的日常工作。机器人工作的环境是复杂的、多种多样的、动态和开放的，且在环境中有可相互作用的、结构丰富的物体。智能体（Agent）是具备多种传感器和执行器的自主智能设备，可以需要完成复杂、多样和新颖的任务。智能体、环境和任务之间的相互作用是长期和持续的。

2、跨模态 AI 的具身智能

人工智能的技术发展正在从基于静态数据集（例如 ImageNet、COCO、VQA、LLM）的“Internet AI”转变为智能体 Agent 在现实环境中可以交互行动的“Embodied AI”（具身智能）。以服务机器人为代表的具身智能体，将汇聚跨模态的 AI 能力，突破机器人在开放环境、泛化场景、连续任务等复杂条件下的感知、认知、决策和执行技术，展示智能机器人在新一轮人工智能大模型变革浪潮中的巨大潜力。

3、预训练大模型 GPT

最近大语言模型（LLMs）的研究已经能够生成复杂的基于提示的文本。然而，如何将大语言模型（如 ChatGPT）的 AIGC 编码应用于机器人执行现实世界中的复杂任务，目前还是智能机器人面临的一个主要困难。通过大语言模型（LLMs）给出更多的人类语言输入后，形成机器人对任务的知识理解，具有推理能力并完成复杂的任务。

4、深度强化学习

深度强化学习（Deep Reinforcement Learning，DRL）是一种结合了深度学习和强化学习的方法，其主要特点在于通过神经网络（尤其是深度神经网络）来表示和优化策略或值函数。这使得 DRL 能够处理高维度、复杂的状态空间和动作空间，从而在许多领域取得了显著的成功，如游戏、机器人控制和自动驾驶等。DRL 具有自主学习和适应环境变化的能力，可以在没有人工干预的情况下，通过与环境的交互，自动地学习到完成任务的最优策略。

5、虚实训练与智能迁移

采用高精度、高逼真度的数字孪生与物理仿真引擎，构建真实世界场景的数字孪生语义场景和各种常用 3D 物品模型。具有 38 个自由度的人形机器人数字孪生体在仿真场景中进行算法、技能和任务训练。通过 HARIX 海睿云端大脑完成由虚拟环境到物理场景的无缝智能迁移，赋能物理实体机器人完成在现实世界的工作。

6、开源开放的生态系统

支撑大赛的 HARIX 海睿云端大脑是一个开源开放的生态平台，开发者通过 HARIX RDK（Robot Development Kit）进行机器人应用开发，包括物理世界场景数据模型、机器人数字孪生模型、人工智能算法、行业应用程序等，形成资源共建和共享的 UGC 社区生态。同时，HARIX OS 海睿生态平台实现虚拟线上资源与实体线下应用的商业互通，开发者贡献的机器人技能可以通过 RSS（Robot Skill Store）实现收费服务，从而促进云端机器人商业生态的良性发展。

三、【赛题考核要点】

1、总体任务：

在仿真环境中，参赛者通过大模型训练机器人在咖啡厅场景成为合格的咖啡厅服务员。这项比赛的考核要点是将大语言模型（LLMs）整合到机器人系统中，开发能够理解自然语言并以友好和有效的方式与人类互动，并能在咖啡店仿真场景中自主完成各种服务任务的智能机器人。

选手可以按照一般常识性理解，进行机器人任务设定和训练。在仿真场景中，机器人可以与可交互的物品和 NPC 进行互动操作的训练，比如：咖啡店服务员与顾客（NPC）互动、接受和执行订单以及回答有关菜单的问题、导航、操作咖啡机、清理桌子/地面、开空调/开灯、递送咖啡/饮料/食物等。

2、考核要点：

- 主动探索和记忆（Active Exploration and Memorization）：机器人在环境中通过主动探索获得各种环境信息，实现对位置环境的感知，形成以环境感知信息以及运动轨迹等历史信息维护一个机器人自身的记忆库。

- 场景多轮对话（Grouded Question Answering）：多轮对话要求机器人智能体具有与人进行流畅的交流能力，具身对话是机器人利用视觉等传感器获得的场景信息基础上，完成于场景相关的对话。

- 视觉语言导航（Vision Language Navigation）：导航是构建智能机器人的一个基本要素。在现实场景中，一个机器人要在不同的场景下承担多种复杂的导航任务。我们的模拟器支持多任务的现实世界导航和物体互动。对于这个任务中的导航，尽管有传统的 ObjectNav 和 PointNav，你可以利用我们的环境完成简单到复杂的视觉语言导航，并有不同难度的指示，以及交流导航，机器人智能体可以在导航中寻求帮助。

- 视觉语言操作（Vision Language Manipulation）：抓取是指机器人使用机械臂抓取物体并将其从一个原始位置移动到目标位置的动作。尽管机器人学习算法在现有的挑战上取得了很大的突破和改进，但仍有许多问题亟待解决。这项任务要求机器人按照视觉和语言的场景描述来抓取一个物体。虽然 Saycan 和 RT-1 在以前的研究中被用来实现使用 Deep-RL 算法的抓取，但这项任务更侧重于在现实环境中抓取薄、大、平、软的物体，避免碰撞，以及多任务抓取。参赛者需要根据大语言模型提供的指令，解决在不同场景下抓取不同物体的问题。具体抓取物品的技能需要参赛者基于提供的环境和工具接口，通过强化学习等方式进行训练。

3、赛道和赛题：

任务		大模型+具身智能	考核点
规定任务	1、环境主动探索和记忆 AEM（Active Exploration and Memorization）	输出探索结果（语义地图）对环境重点信息记忆。 生成环境的语义拓扑地图，和不少于10个环境物品的识别和位置记忆，可以是图片或者文字或者格式化数据。	1、在最短时间内完成全部环境的自主探索 2、语义地图完整度、精准度（知道物品空间位置对应是否准确） 3、关键信息的记忆完整度
	2、视觉语言导航 VLN（Vision Language Navigation）	识别顾客（NPC）靠近、打招呼、对话、领位导航到适合人数的空闲餐桌 开始条件：监测到顾客靠近 结束条件：完成领位，语音：“请问您想喝点什么？”，并等待下一步指令	1、与顾客打招呼 2、识别空闲座位，以及分配正确人数的餐桌 3、正确指引顾客到指定位置。
	3、具身多轮对话 GQA（Grouded Question Answering）	1、点餐（order）的对话，咖啡厅服务员可以为客人（NPC）完成点餐基本对话 2、场景对话（GQA）结合场景：询问卫生间、附近娱乐场所（数据来源学生根据常识自主定义） 开始条件：顾客NPC发出点餐指令 结束条件：顾客NPC发出指令，表示不再需要服务	1、大语言模型接入和使用 2、大语言模型对话能力考核
	4、视觉语言操作 VLM（Vision Language Manipulation）	机器人根据指令人的指令调节空调，自主探索环境导航到目标点，通过手臂的运动规划能力操作空调，比如开关按钮、调温按钮、显示面板	1、导航任务完成的成功率、精度、效率 2、操作任务完成的成功率、视觉反馈的运动控制精度、动作合理拟人化 实现方法：可选择接入大模型，或自行训练机器人运动规划
开放任务	1、人提出请求，机器人完成任务	1、做咖啡（固定动画）：接收到做咖啡指令、走到咖啡机、拿杯子、操作咖啡机、取杯子、送到客人桌子上 2、倒水 3、夹点心 具体描述：设计一套点单规则（如菜单包含咖啡、水、点心等），按照规则拟造随机的订单。在收到订单后，通过大模型让机器人输出合理的备餐规划，并尝试在模拟环境中按照这个规划实现任务	1、对语义的意图理解（认知），形成任务目标 2、思维链 long horizon 3、与机器人技能结合完成复杂任务
	2、在特定环境下，机器人发现目标，可自主完成任务	1、打扫地面：地面有垃圾，机器人主动扫地、清理地面垃圾 2、收拾桌子：桌子上的污渍，机器人主动擦桌子 3、摆椅子：椅子不正，机器人主动摆正椅子 4、开灯：室内光线暗，机器人主动打开房屋的灯	1、对环境的自主认知并形成任务目标 2、思维链 long horizon

四、【赛事仿真环境】

为了更好地完成这项挑战，我们给出了一个还原真实场景效果的仿真环境，参赛者可以用它来训练机器人并在这个仿真环境中部署代码。在仿真环境中考虑了基于物理（物体状态和关系）和社会（人的行动和目标）信息的具有模糊性的人类指令的对应，可以对物理和社会环境中的指令理解和跟随进行整体评价。

我们导入了 Ginger 人形机器人智能体 Agent，具有丰富的环境交互界面，支持机器人关节控制和底盘速度控制。为了方便图像采集，机器人智能体提供了多种传感器模拟（RGBD 相机、分割相机、激光雷达、IMU、里程表等），并支持模拟环境中的并行训练。目前比较流行的同类研究，如 AI2-THOR[1]、Habitat-Sim[2]和 iGibson[3]都无法做到我们模拟的内容和我们提供的功能。据我们所知，将 ChatGPT 与机器人技术相结合的是微软的 ChatGPT for Robotics[4]等研究机构在开展类似的工作，他们提供的机器人并不具备 Ginger 人形机器人的硬件能力。在这次挑战中，我们参赛者将把最新的大语言模型提供的指令与机器人的语音交互、视觉、导航和操作结合起来，激发出更多的创造力。

1、仿真场景和物体

对于模拟器，我们的模拟器中使用高分辨率的模型和精确的材料和纹理，高度详细的物理引擎可以准确地模拟了物体的动力学和机械行为，使模拟器环境中物体的外观和物理特性与真实世界非常接近。此外，我们在模拟器中，包括机器人控制系统和物体的互动，以满足强化学习的需要。

模拟器中的各种场景和物体为机器人的导航和操作训练提供了强有力的支持，例如，在咖啡馆场景中，总共有 13 类近 800 个物体。此外，还有 73 个类别的物体可以被添加（尽管这些物体可能与咖啡馆没有什么关系）。通过使用模拟环境，产生了大量不同的数据集。我们分别构建了 5k 个用于导航和操纵的人类演示，涵盖了各种情况，ChatGPT 被用来生成 10k 个人类指令。

1.1 咖啡厅场景：

1.2 仿真环境可交互的物体

- 物品：咖啡机、咖啡杯、茶杯、一次性纸杯、汤匙、托盘、毛巾、抹布、喷壶、扫帚、报纸、词典、饭盒、碗盘、保温杯、茶叶罐、纸巾包、维达卫生纸、鼠标、老花镜、牙膏、网球、小熊玩具、小狗玩具、魔方、计算器、胶棒、笔筒、名片盒、订书机、透明胶带、水壶等

- 食物与饮料：蛋糕、百事薯片、水果：苹果、香蕉、橘子、西瓜、山竹、榴莲、西红柿、大枣、核桃、土豆、山药、蒜头、烧饼、面包、火腿肠、冰红茶、纯甄酸奶、AD 钙奶、农夫山泉、冰红茶、蒙

牛真果粒、伊利果粒多、伊利优酸乳、绿箭口香糖、益达口香糖、夏进甜牛奶、安慕希、贝纳颂咖啡、雀巢咖啡、果亿特椰汁水、椰树牌椰汁、舒耐喷雾、美汁源果粒橙、橙汁等

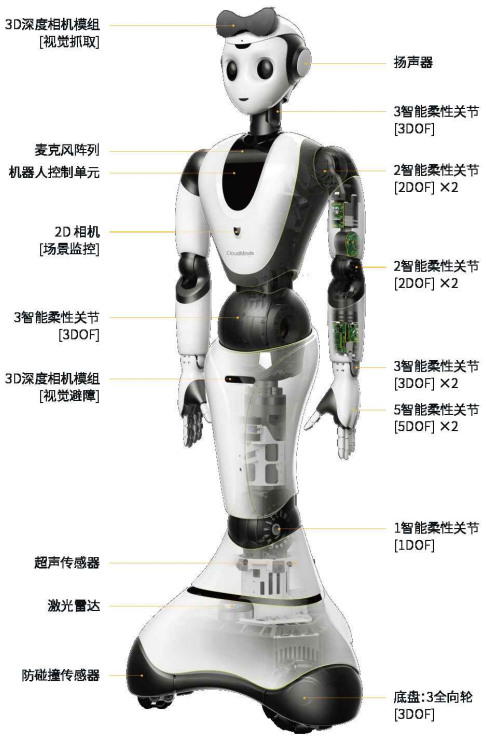
- 设施与设备：空调、冰箱、桌子、椅子、沙发、门、窗、窗帘、灯、吸尘器、抽屉、橱柜、听诊器、水壶、水杯、电梯等

1.3 人（NPC）

咖啡厅客人：男客人 1Mark、男客人 2Altman、女客人 1Amy、女客人 2Cathy

2、具身智能机器人

2.1 人形机器人 Ginger 硬件能力



2.2 灵巧手硬件能力

自由度（DOF）：7DOF

触觉传感器：电流反馈

手臂负载：5kg

功能：抓取、开关门、开关灯、搬桌椅、推手推车、使用吸尘器、使用咖啡机、操作空调控制面板等

五、【赛事开发环境】

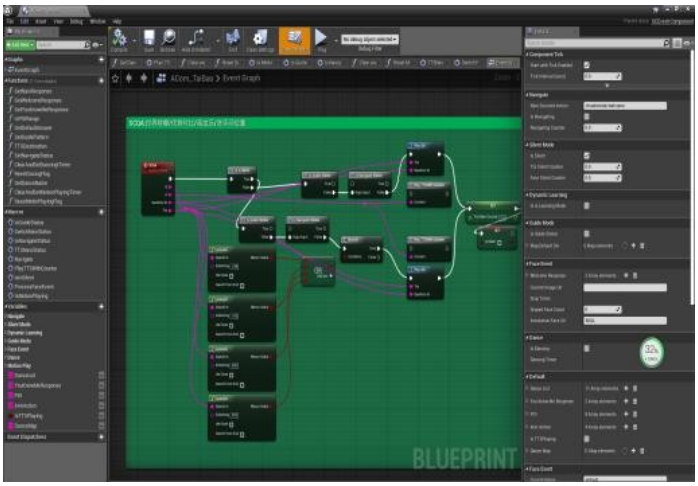
1、操控机器人及开发文档由达闼机器人提供技术支持，参考网址：https://harix.cloudminds.com/#/index/community/doc。赛事答疑请使用大赛官方技术支持论坛：

https://bbs.csdn.net/forums/harix

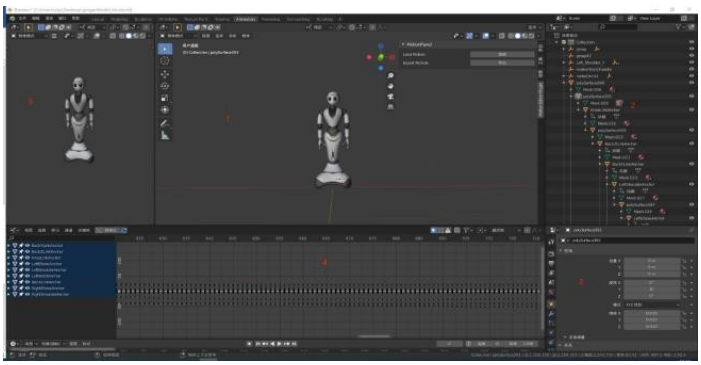
2、达闼海睿 RDK 开发者门户支持，参考网址：https://harix.cloudminds.com/，参赛选手开发过程中，需要登录 RDK 开发者门户（账号在报名时联系工作人员分配），进行智能语音、地图、算法等相关配置。

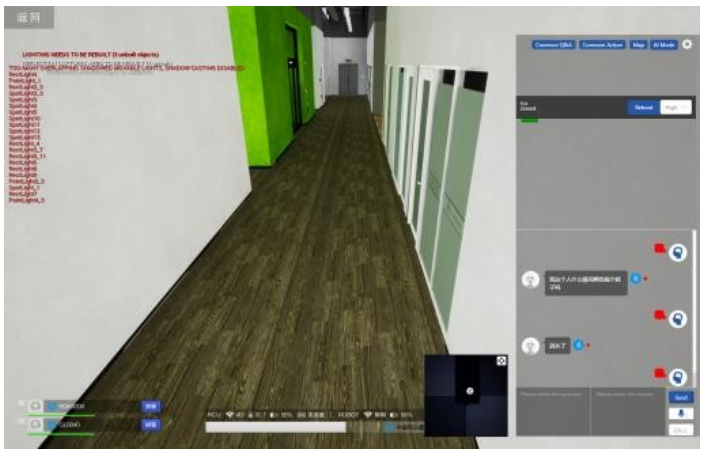
3、大赛开发环境

PC 端 RDK 开发环境由大赛组委会赛事官网发布赛题信息，提供赛题答疑。

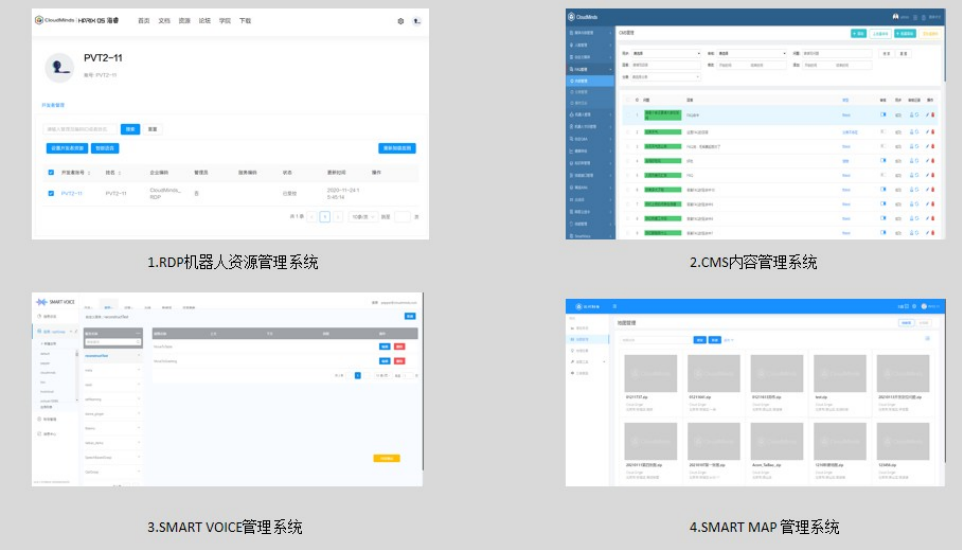


RDK 地图编辑器 RDK 蓝图编辑器





RDK 动作编辑器 RDK 人工坐席辅助



云端 RDK 开发环境

六、【赛事组织】

1、标题：【机器人大模型与具身智能挑战赛：将大语言模型纳入机器人系统】

2、组委会：中国计算机学会、达闼机器人、AITISA 联盟、OpenI 启智、北京大学、CCF 开源发展委员会、复旦大学、北京航空航天大学、中山大学、北京邮电大学、北京智谱

3、规则和评估：参赛队必须将 LLMs 纳入其机器人系统，以促进自然语言的理解和互动。比赛将根据以下标准来评估机器人系统的性能：

- 任务完成的准确性和效率：

参赛者将被评估任务完成的准确性和效率，包括物体操作、导航到准确位置的精度，执行推理速度，订单执行和人机互动。机器人必须准确、高效地完成任务，才能获得分数。

- 人与机器人的互动：

参赛者将被评估其机器人与顾客和工作人员互动的自然度和友好度。机器人必须以自然和友好的方式进行交流，以获得积分。

- 时间限制：

参赛者将有规定的时间来完成任务，在规定时间内得分最高的团队将被宣布为获胜者。

4、比赛的程序：

机器人具身智能挑战赛将在 CCF 中国软件大会会议期间举行。比赛将由多个任务组成，组委会评委将根据赛题要求测试机器人执行任务的能力。安排如下：

- 规定任务：测试Agent完成规定任务的能力

- 开放任务：测试Agent完成开放任务的能力

5、可能的参与者名单：

机器人具身智能挑战赛对来自全球学术界、工业界和研究机构的团队开放，包括：

- 在机器人和人工智能方面有专长的大学和研究机构

- 专门从事机器人和人工智能应用的初创企业和公司

- 独立的机器人开发者和业余爱好者

6、参赛作品提交：

除挑战赛统一要求文档外，提交内容还包括方案设计文档、RDK 技能开发蓝图源文件，以及录制的运行结果视频。凡提交的参赛作品，赛事组织方拥有永久的使用权，如经过审核后发布到海睿机器人技能应用商店（Robot Skill Store），获得的收益将按一定比例和平台分成（具体细节以最终官方公布为准）。

7、考核与奖励：组委会可以设定比赛任务，考察选手提交作品的任务完成能力，并依照评分规则进行评判。根据评委评分，大赛设定一、二、三等奖和专项奖若干。评分规则：

总分排行与单项排行两种名次

1）总分排行规则：场景下有多个任务，任务得分为完成这些任务的分数相加；每位选手按比赛要求提交设计文档、录制视频、蓝图源文件等内容，组委会进行评审打分。

2）单项排行规则：场景下有多个任务，每个任务单独进行分数排行；每位选手按比赛要求提交设计文档、录制视频、蓝图源文件等内容，组委会进行评审打分。

3）决赛：赛事评委按照大赛组委会的相关规定和要求，对选手的作品进行计分、排名。当分数相同时，文档质量等将成为附加排序依据。约前 30% 的队伍进入答辩环节，参加答辩的队伍最终排名依据测试排名（权重 65%）与答辩排名（权重 35%）的综合排名。

4）如果因题目难度过大或版本等不可抗因素，导致无团队完成任意一个任务，组委会重新评审后，设定新的打分标准，按新的标准重新打分。如违背大赛规定规范，可能会被取消参赛资，具体内容由大赛组委会规定。

8、反抄袭、反作弊：

（1）禁止多账户报名，否则取消所有成绩及比赛资格

（2）禁止利用任何规则漏洞或技术漏洞等不良途径提高成绩排名，禁止任何不诚信行为，一经发现取消所有成绩及比赛资格

9、在完成赛事任务过程中，对于由版本等不可抗因素所触发的问题，组委会将本着公平、公正、透明的原则积极解决，最终解释权归组委会所有。

10、重要日期

作品提交截止日期:2023年11月20日

公布比赛结果和颁奖日期:2023年12月03日

11、作品提交方式

作品提交链接网址：

https://easychair.org/conferences/?conf=chinasoft2023

（提交作品详细步骤见使用说明—参赛作者 https://kdocs.cn/l/cuSwixcIWMR8）

12、交流方式

达闼机器人的赛事网址：

https://harix.cloudminds.com/#/index/community/race

海睿 OS 技术社区网址：

https://bbs.csdn.net/forums/harix

加入“达闼杯机器人大模型与具身智能挑战赛服务群”，扫描以下微信二维码会有专人邀请入微信群：

七、【其他】

1、关于 AI 工具应用：

许多研究者在多模态机器人模型方面取得了优异的成果。在这个挑战中，参赛选手可以参考我们提供的基本模型作为你工作的起点（仅作建议，不限制采用任何大模型），或者可以依照自己的想法来解决问题：

Gato[5]：它是一个作为多模态、多任务、多体现通用策略的机器人。相同的网络使用相同的权重可以在 Atari 游戏、图像标题、聊天、使用真实机械臂堆叠积木等方面发挥作用，并根据其上下文决定是否输出文本、关节扭矩、按钮按下或其他标记。链接：https://arxiv.org/abs/2205.06175。

RT-1[6]：它建立在一个 transformer 架构上，以机器人摄像头的短历史图像和以自然语言表达的任务描述作为输入，直接输出标记化的行动。RT-1 的架构类似于当代的仅解码序列模型，针对标准的分类交叉熵目标进行训练，并进行因果屏蔽。它的关键特点包括图像标记化、行动标记化和标记压缩。链接：https://arxiv.org/abs/2212.06817。

Saycan：它是一种机器人控制方法，使用大型语言模型（LLM）规划机器人动作序列，以实现用户指定的目标。SayCan 使用提示工程将用户的输入转换为对话，询问机器人如何完成将海绵带给用户的步骤。机器人的每个技能都有一个文本描述，LLM 可以使用它来计算其完成步骤的概率，以及一个价值函数，该函数指示技能在给定世界当前状态的情况下成功的可能性。链接：https://say-can.github.io/。