毕设具身智能综述

1 具身智能

[1] 是一篇综述文献。

我们概述了预训练基础模型在机器人领域的应用。传统的机器人领域深度学习模型通常使用针对特定任务的小规模数据集进行训练，这限制了它们在不同应用场景中的适应性。相比之下，预训练于大规模互联网数据的基础模型展现出更强的泛化能力，并且在某些情况下能展现新兴的能力以找到不在训练数据中存在的零样本解决方案。这些基础模型可能有助于提升机器人自主技术栈的各个组件，从感知、决策到控制。例如，大型语言模型能够生成代码或提供常识推理，而视觉-语言模型则支持开放式词汇的视觉识别。然而，仍然存在许多挑战，特别是缺乏相关的训练数据、安全保证和不确定性量化，以及实时执行的问题。在本文中，我们研究了已使用或构建基础模型解决机器人问题的相关论文，并探讨这些基础模型如何提升机器人在感知、决策及控制领域的性能。同时，我们分析阻碍基础模型在机器人自主性中应用的挑战，并提供了未来发展方向的可能性途径。与本文相关的 GitHub 项目可以在以下链接找到：1 https://github.com/your-repo-name

关键词——机器人技术，大型语言模型（LLMs），视觉-语言模型（VLM），大规模预训练模型，基础模型

基金会模型是在大量互联网数据上进行预训练的，可以针对广泛的数据下游任务进行微调。基金会模型在视觉和语言处理方面取得了显著突破；其中包括 BERT [1]、GPT-3 [2]、GPT-4 [3]、CLIP [4]、DALL-E [5] 和 PaLM-E [6]。基金会模型有可能为机器人学领域开辟新的可能性，如自主驾驶、家用机器人、工业机器人、辅助机器人、医疗机器人、场域机器人以及多机器人系统。预训练的大语言模型（PLMs）、大视觉-语言模型（VLMs）、大音频-语言模型（ALMs）和大视觉导航模型（VNMs）可以用于改进各种机器人任务。将基金会模型整合到机器人学中是一个快速发展的领域，机器人学社区最近才开始探索如何在感知、预测、规划和控制等方面利用这些大型模型来推动机器人技术的发展。

在基础模型出现之前，传统的机器人学深度学习模型通常是在为特定任务收集的有限数据集上进行训练 [7]。相比之下，基础模型是基于广泛且多样化的数据进行预训练的，在其他领域（如自然语言处理、计算机视觉和医疗保健 [8]）中已经证明能够显著增强适应性、泛化能力和整体性能。最终，基础模型有可能在机器人学中同样带来这些好处。从基础模型转移知识可能比针对特定任务的训练能减少训练时间和计算资源。

特别是在机器人技术领域，多模态预训练模型能够将来自各种传感器的异构多模态数据融合并对齐，形成适合机器人理解和推理所需的紧凑同质表示[9]。这些学习到的表示形式有可能应用于自主系统堆栈中的任何部分，包括感知、决策和控制。此外，预训

练模型提供了零样本能力，即 AI 系统能够在没有特定任务示例或专门训练数据的情况下执行任务的能力。这将使机器人在遇到新情况时能够泛化其学到的知识，增强其在非结构化环境中的适应性和灵活性。将预训练模型集成到机器人系统中可能会使系统的上下文感知更加敏锐，加强了机器人对环境的感知和交互能力。例如，在感知领域，大型视觉-语言模型（VLM）已被发现通过学习视觉与文本数据之间的关联提供跨模态理解，从而辅助零样本图像分类[10]、零样本目标检测[10]以及 3D 分类[11]等任务。在三维语义接地方面（即将视觉-语言模型的上下文理解与真实世界的三维环境对齐），通过将词语与三维环境中特定的对象、位置或动作关联起来，可以增强机器人的空间感知能力。在决策或规划领域，大模型和 VLM 被发现能帮助机器人进行高阶任务的规范描述[13]，通过利用语言线索来执行复杂操作、导航和交互等任务。例如，在模仿学习[14]和强化学习[15]中，预训练模型似乎提供了提高数据效率并增强上下文理解的可能性。特别是，基于语言的奖励能够引导强化学习代理，提供塑形奖励[16]。此外，研究人员已经使用语言模型为策略学习技术提供反馈[17]。一些研究工作表明视觉-语言模型（VLM）的视觉问答能力可以在机器人场景中得到利用；例如，研究人员已通过 VLM 回答与视觉内容相关的提问来帮助机器人完成任务[18]。另外，研究人员还表示用 VLM 用于数据标注，生成描述性标签以对视觉内容进行注释[19]。

在本综述中，我们研究了目前基础模型在机器人学中的应用文献。我们探讨了当前的方法和应用场景，提出了现有挑战，并指出了未来研究方向以解决这些挑战，同时指出了将基础模型集成到机器人自主性中可能暴露的风险。与我们的研究同期，另一篇关于基础模型在机器人学方面的综述也在 arXiv [20]上发表。相比那篇文章，我们的研究更侧重于未来的挑战和机遇，特别强调了安全性及风险问题，并更加关注现有文献在应用、算法和架构方面进行了比较。相比之下，一些现有的综述仅集中在特定的上下文指令上，例如提示[Prompt] [21]、视觉转换器[Vision Transformers] [22]或决策制定[Decision-Making] [13, 23]等方面；而我们则从更广泛的角度出发，将基础模型的研究线程联系起来，强调其与机器人学的相关性和应用。虽然我们的研究范围比论文[24]要窄得多，后者探讨了基础模型在多个学科中的广泛应用（其中也包括机器人学）。我们希望这篇综述能为研究人员提供清晰的近期进展和现有不足领域的指引，并指明未来研究的方向及其机遇与挑战。最终目标是为机器人学家提供一份参考文献，让他们了解这一令人兴奋的新领域。

在该综述的研究范围中，我们限定了以下几类论文：

- 1) 背景论文：虽然没有明确链接到机器人学，但理解基础模型必不可少的论文将在综述文稿的背景部分（第二部分）讨论。
- 2) 机器人学论文：将基础模型集成于插件式机器系统、对基础模型进行适应或微调以用于机器人系统，或是构建新的针对特定领域的机器人基础模型的论文。

- 3) 靠近机器人学的研究论文：提出应用于邻近领域（如计算机视觉、具身智能）的方法和技术，并明确指出了未来可能应用于机器人学的路径。

本文结构如下: 第二部分介绍了基础模型, 包括大型语言模型（LLM）、视觉变换器、多模态视觉语言模型、具身多模态语言模型以及视觉生成模型。此外, 在这一部分的最后, 我们将讨论训练这些基础模型的不同方法。第三部分探讨了基础模型在机器人决策任务中的集成应用。首先，我们讨论使用语言条件化模仿学习和语言辅助强化学习来学习机器人的策略。接着，我们探讨如何利用基础模型设计一种可用于规划的语言条件化价值函数。随后，本文将介绍如何使用基础模型进行任务规划相关的具身体现与代码生成。第四部分研究了各种具有潜在提升空间的机器人感知任务，这些任务包括语义分割、三维场景表示、零样本 3D 分类、可操作性预测以及动力学预测。第五部分讨论了一系列关于具身人工智能代理、通用人工智能代理以及用于具身人工智能研究开发的模拟器和基准测试的文章。第六部分总结了基础模型在机器人系统中的应用挑战，并提出了未来研究的方向。最后，在第七部分，我们将提出结论性的意见。

1.1 基础模型

基础模型具有数十亿个参数，并且是在大规模互联网数据集上预训练的。如此规模和复杂性的模型训练涉及重大成本。获取、处理和管理数据可能也代价高昂。训练过程需要大量计算资源, 需要特定的硬件如 GPU 或 TPU, 还需要支撑模型训练所需的软件和基础设施，这同样需要资金支持。此外，训练基础模型耗时巨大，这也会进一步增加成本。因此这些模型通常作为即插即用模块使用（指的是无需进行大量定制化即可将基础模型集成到各种应用中）。表 I 提供了常用基础模型的详细信息。在本节其余部分，我们将介绍大规模语言模型（LLM）、视觉变换器、联合视觉语言模型（VLMs）、具身多模态语言模型以及视觉生成模型。在本节的最后一部分，我们将介绍用于训练基础模型的不同训练方法。

1.1.1 A. 术语和数学预备知识：

首先，在本节中我们介绍基础模型领域的常用术语，并描述不同类型的基模型的基本数学细节和训练实践。

分词 给定一串字符序列，分词是将该序列分割成更小单位（称为标记）的过程。根据分词策略，标记可以为单个字符、词语片段、完整单词或句子片段。这些标记通常表示为维度与词汇量大小相等的 1-热向量，并通过学习嵌入矩阵映射到低维实数空间中。LLM 以这些嵌入向量序列作为原始输入，生成同样形式的嵌入向量序列作为原始输出。然后将这些输出向量映射回标记并最终形成文本。例如，GPT-3 的词汇表包含 50,257 个不同的标记，嵌入维度为 12,288。

生成词元（从低维度的实值嵌入向量转换为高维度的一 hot 向量）具有一定的随机性，从而对词汇表中每个可能的词元赋予了权重。这些权重通常被大型语言模型 (LLM) 用作各词元的概率分布，以此在文本生成过程中引入随机性。例如，在 GPT-3 中，温度参数介于始终选择最高权重的词元（温度为 0）和基于权重建议的概率分布来抽取词元（温度为 1）之间。这种随机性仅存在于词元解码过程中，并非出现在 LLM 本身中。从作者所知，这是 GPT 家族模型中唯一的随机源。

生成模型是一种能够从概率分布中采样，从而创建类似于训练数据的数据实例的模型。例如，面部生成模型可以生产出难以与用于训练该模型的真实图像集区分的图像。这些模型可以被训练为条件性的，这意味着它们可以根据广泛变化的条件信息产生样本。例如，性别条件性面孔生成器可以在给定期望性别的情况下，生成女性或男性的面部图像。

鉴别模型：鉴别模型用于回归或分类任务。与生成模型不同，鉴别模型旨在区分不同的类或类别，在输入空间中学习类之间的边界。生成模型学习从数据分布中采样，而鉴别模型则学习评估给定输入特征的输出标签的概率分布，或者（取决于模型是如何训练的）学习评估输出概率分布的一些统计量，例如给定输入时的期望输出。

transformer 架构：大多数基础模型都是基于 transformer 架构构建的，这种架构对基础模型和大型语言模型的发展起到了关键作用。以下讨论综合自文献[76]以及在线博客、未发表报告和维基百科[77-79]。transformer 同时操作一组嵌入令牌向量（x1, . . . , xN），被称为上下文窗口。transformer 架构的关键创新之处在于最初的里程碑工作[76]提出的多头自注意力机制。在该架构中，每个注意力头计算一个表示与上下文窗口其他令牌 xin 相关性强度的重要性权重的向量。每个注意力头通过在计算重要性权重时使用的不同投影矩阵来数学编码不同的相似性概念。每个头部可以在所有令牌和所有头部之间并行地进行训练（反向传播过程）和评估（正向传播过程），相比于基于 RNN 或 LSTM 的先前模型，这使得训练和推理速度更快。

2 具体论文 1

[2] 本研究探讨了使用 OpenAI 的 ChatGPT [1] 于机器人应用中的可行性。我们提出了一种结合提示工程技术设计原则和高级功能库创建的方法，使 ChatGPT 能够适应不同的机器人任务、模拟器以及形态因素。我们的评估集中在不同提示工程技术及对话策略对执行各种类型机器人任务效果的评估上。此外，我们还探讨了 ChatGPT 在使用自然语言对话进行自由格式交流、解析 XML 标签、合成代码方面的能力，同时还包括特定任务提示函数和闭环推理的应用。研究涵盖了从基本逻辑、几何和数学推断到复杂领域如空中导航、操作以及实体代理等一系列机器人任务。结果显示，ChatGPT 能够有效解决多种此类任务，并允许用户主要通过自然语言指令与之交互。此外，我们还介绍了一种开

源的研究工具——PromptCraft，该工具包含一个协作上传和投票选定的优秀提示方案平台，以及集成了 ChatGPT 的样本机器人模拟器，这使得使用 ChatGPT 进行机器人应用更加便捷。

图 1：当前的机器人流程需要一个专业的工程师在环中编写代码以改进过程。我们的目标是利用 ChatGPT 让非技术用户也能参与到环中，通过高级语言命令与语言模型交互，并能够无缝部署各种平台和任务。

https://github.com/microsoft/promptcraft-robotics?tab=readme-ov-file 上述链接是一个社区，供人们测试和分享机器人领域内大型语言模型（LLM）的有趣提示示例。我们还提供了一个带有 ChatGPT 集成的示例机器人模拟器（基于 Microsoft AirSim 构建），供用户开始使用。

模型如 GPT-3、LaMDA 和 Codex 在机器人学领域中也展现出零样本应用的潜力，特别是在高层代理规划[12, 13]或代码生成方面[14, 15]。这些早期演示激发了我们探索 ChatGPT 作为机器人领域更加多功能工具的可能性，因为它结合了自然语言和代码生成模型的优点，并具备对话灵活性。

在本文中，我们的目标是展示 ChatGPT 在机器人应用方面的潜力。我们提出的核心概念在于创建一个高层函数库，来解锁利用 ChatGPT 解决机器人应用程序的能力。由于机器人学是一个涵盖多种平台、场景和工具的多样化领域，存在大量的库和 API。与其要求大型语言模型（LLMs）输出特定于某平台或库的代码，这可能需要大量微调，我们选择创建一个简单的高层函数库供 ChatGPT 使用，并在后台将其与所选平台的实际 API 进行关联。这样可以允许 ChatGPT 从自然对话中解析用户意图，并将其转换为高层次函数调用的逻辑链条。此外，我们还详细介绍了若干提示工程技术指南，以帮助 ChatGPT 解决机器人任务。

我们的研究显示，ChatGPT 能够以零样本的方式解决各种与机器人相关的任务，并且能够适应多种形态因素，同时通过对话实现闭环推理。此外，我们还旨在展示当前模型的局限性，并提出克服这些局限性的想法。以下是我们的主要贡献：

- 我们展示了一个将 ChatGPT 应用于机器人任务的工作流程。该工作流程包括多种形式的提示技术，如自然语言对话、代码提示、XML 标签以及闭环推理。我们还说明了用户如何利用高水平的功能库来快速解析人类意图并生成解决问题的代码；
- 我们通过实验评估了 ChatGPT 在执行各种机器人任务方面的能力与局限性。我们在处理数学运算、逻辑推理和几何学运算时展示了模型的能力，并探索了涉及实体智能体、飞行导航及操作更复杂场景的可能性。我们包括了仿真以及来自 ChatGPT 计划的真实世界实验；

- 我们引入了一个协作性的开源平台 PromptCraft，研究人员可以在其中共同提供在机器人背景下的大语言模型 (LLM) 工作时正反面提示策略的示例。提示工程多是一个经验性科学领域，我们希望为研究人员提供一个简单接口，让他们能够以社区的形式贡献知识。随着时间的推移，我们将提供不同的测试环境，让用户可以检验他们的提示，并欢迎新的贡献；
- 我们发布了一个基于 Microsoft AirSim 的仿真工具，并集成了 ChatGPT。该 AirSim-ChatGPT 仿真包含无人机导航的示例环境，旨在作为研究人员探索如何利用 ChatGPT 支持机器人应用场景的起点。通过本项工作，我们希望能够开启融合大语言模型和机器人学的新机遇与研究方向。我们认为我们的研究成果将激励并指导这个引人入胜领域的进一步研究，铺平开发能够自然、直观地与人类交互的新颖创新性机器人系统的道路。欲了解更多信息，请参阅项目网页上的详细实验视频。

有效利用 ChatGPT 进行机器人应用面临几个挑战，包括提供完整且准确的问题描述、识别合适的函数调用和 API 集合，以及通过特殊参数引导答案结构。为了充分利用 ChatGPT 在机器人领域的应用，我们构建了一个包含以下步骤的流程：

1. 首先，定义一个高层机器人功能库。该库可以针对特定的形式因子或应用场景，并应映射到实际的机器人平台实施上，同时命名需足够描述性以便 ChatGPT 能够理解；
2. 然后，为 ChatGPT 构建一个提示信息，该提示不仅描述目标，还指明允许使用的高层函数从库中选择。提示还可以包含约束条件或 ChatGPT 如何结构化其响应的信息；
3. 用户保持在环中，通过直接分析或模拟评估 ChatGPT 生成的代码输出，并向 ChatGPT 提供关于代码质量与安全性的反馈；
4. 在对 ChatGPT 生成的实现进行迭代后，最终代码可以部署到机器人上。我们将在图 2 中以家庭清洁机器人的示例直观展示此流程。

2.1 机器人 API 库的构建与描述

作为一项成熟的领域，机器人学已经存在了多种库，无论是黑盒还是开源的形式，均能够用于感知和动作的基本功能（例如目标检测与分割、建图、路径规划、控制、抓取等）。如果在指令中正确规范这些预定义的功能，LLM 便可以利用这些函数进行机器人推理和执行。一个重要的提示设计要求是所有 API 名称都必须描述其整体功能行为。清晰的命名对于允许 LLM 推理解析不同 API 的功能连接并产生所需结果至关重要。因此，我们可以定义高级函数，作为实际库实现的具体封装。例如，名为 detect_object(object_name)的函数内部可能与 OpenCV 函数或计算机视觉模型关联，而类似 move_to(x, y, z)这样的函数则可能内部调用路径规划和避障流水线以及适合无人机的低级电机指令。在提示中列出如此一组高级函数对于 ChatGPT 来说至关重要，使其能够创建逻辑的行为原语序列，并在不同场景和平台中进行泛化。

根据具体应用场景，我们建议解释 API 的功能，必要时将其分解为具有清晰输入输出的子组件，类似代码文档的方式。图 3 展示了用于家庭烹饪机器人场景的良好 API 提示策略示例，该策略使 ChatGPT 能够根据不同功能推断任务的顺序与内容，而实际由机器人执行。相比之下，请参阅附录 A.1 中的例子，在没有 API 指导的情况下，ChatGPT 会生成无法边界约束的文字响应；或者参见附录 A.2 中的例子，由于 API 定义不充分，导致 ChatGPT 对功能调用参数进行错误设想。

我们注意到，与经典的符号人工智能脆弱的结构不同，后者要求物体和函数之间严格预定义的关系。而 LLM 能够根据特定问题的需求来界定新的功能和概念。这一能力赋予了 LLM 在处理机器人应用时的灵活性与鲁棒性。图 4 展示了当 ChatGPT 需要解决特定问题时，如何创造新的高级概念甚至底层代码的情况。用户可以利用这种能力作为设计策略，并在 LLM 协助下迭代定义新的 API，直到当前 API 不足以解决问题为止。

通过在指令中提供清晰且简洁的任务描述及其背景，ChatGPT 可以生成更准确的响应。好的上下文描述应包含以下内容，除了机器人 API：

- 约束和要求：明确与任务相关的约束或要求。如果任务涉及搬运物体，则需指定被搬运物的重量、尺寸及形状。
- 环境：描述执行机器人任务时所处的环境。例如，如果任务是导航迷宫，则需描述迷宫的大小及形状，并说明需要避开的障碍物和危险。
- 当前状态：描述当前机器人的系统状态。例如，若任务是拾取物体，则需描述机器人及目标物当前的位置与姿态。
- 目标和任务：明确任务的目标与目的。如果任务是拼装拼图，则需指定需要拼接的块数以及预计完成时间。
- 解决方案示例：通过展示如何解决类似的任务，帮助指导大模型的解决方案策略。例如，如果任务涉及与用户的交互，则可描述机器人何时及如何向用户提供输入请求（参见图 5）。值得注意的是，预热也可能引入偏差，因此应提供多样化的示例并避免使用过度规定性的语言。

即便是一个精心设计的提示，也可能无法包含解决问题所需的所有必要信息，或者在某些情况下，ChatGPT 可能无法以零样本的方式生成正确的响应。在这种情况下，我们发现用户可以采取的一个简单而有效的策略是在聊天格式中向 ChatGPT 发送额外指令，描述问题，并要求它自我修正。之前依赖 GPT-3 或 Codex 模型的方法[15, 14]需要用户重新设计输入提示并从头开始生成新的输出。然而，ChatGPT 的对话能力意外地成为行为矫正的一种非常有效的途径。第 3.2 节和补充视频中展示了用户与 ChatGPT 之间交互行为的例子。

2.3. 特殊提示用于引导答案结构 不同的提示方法可以用来迫使模型的输出遵循某种特定模式。例如，用户可能希望自动解析 ChatGPT 的输出，以便在其他脚本中实时执行。如图 3 所示，一种简单的方法是直接要求 ChatGPT 生成特定语言（如 Python、C++）的代码。通常，文本部分会跟着一个代码块出现。通过请求模型使用 XML 标签来帮助我们自动解析输出，可以产生更结构化的回答，如图 5 所示。 在其他情况下，用户可能希望迫使模型按照列表形式输出答案，而不是以代码或自由格式文本的形式。附录 A.3 展示了这样一种方法的示例，其中用户的提示的最后一行指定了模型的输出格式。

我们要求 ChatGPT 编写一套具备避障功能的目标达成算法，用于配备前方距离传感器的无人机。ChatGPT 构建了避障算法的主要组成部分，但在某些关于无人机姿态处理的步骤上有所遗漏，需要人类提供反馈。尽管这些反馈仅是高层次的文字描述，ChatGPT nonetheless 依据适当的地方修改了代码，从而改进了解决方案。

假设我给你提供了一些 XYZ 坐标作为目标。我希望你控制无人机，让它朝着目标移动的同时避开前方的障碍物。无人机不应直接飞向目标，而应该逐小步前进，并在每一步重新评估前方的障碍情况。无人机配备了一个距离传感器，可以返回无人机前方最近障碍物的距离，可通过调用 get_distance()访问。

确保只有在无人机前方至少有 10 米的距离时才进行飞行。如果不是这样，则应旋转无人机直至前方有足够的 10 米空隙，然后沿当前面向的方向再移动一小步。在每一步过后，请再次面向目标。听清楚了吗?

2.1 具体论文 2

[3] 机器人使用语言需要将语言与物理世界联系起来（或说是情境化），以便能够参照外部现实并将词汇、感知和行动建立起桥梁 [4]。经典的方法通过词法分析提取语义表示来指导政策，但它们往往难以应对未见过的指令 [5]–[7]。近年来的方法则实现了从语言到行动的端到端建模 [8]–[10]，但这需要大量的训练数据，而这些数据在实际机器人上获取可能会十分昂贵。与此同时，自然语言处理技术的进步表明，预训练于大规模互联网数据的大规模语言模型（LLMs）具备现成的能力 [11]–[13]，可以在无需额外微调的情况下应用于使用语言的机器人中，比如从自然语言指令规划一系列步骤 [16]–[18]。这些步骤可以基于固定技能集中的价值函数来建立在真实机器人可操作性之上，即通过行为克隆或强化学习预先训练的策略 [19]–[21]。

尽管前景广阔，但这一抽象化让大语言模型无法直接调控感知-行动反馈循环，从而难以以如下方式情境化语言：（i）推广具有共同感知和动作反馈模式的行为，例如从“把苹果放在橙子上”到“在看到橙子时把苹果放下”；（ii）表达控制中的常识先验，例如“加快速度”或“用力推”；（iii）理解空间关系，“稍微将苹果向左移动”。因此，引入每项新技能

及其对应情境化方式都需要更多数据和重新训练——换句话说，数据负担仍然存在，只是转移到了技能获取阶段。

2.2 开放词汇导航

Improving Vision-and-Language Navigation with Image-Text Pairs from the Web[4] 开放词汇导航：开放词汇导航解决了机器人在未见过的环境中的导航问题。其核心在于机器人能够理解和响应语言提示、指令或语义信息，而不受预定义数据集限制。本节探讨了将 LLMs 和 VLMs 相结合，以插拔式方式集成到机器人导航任务中的相关研究。此外，还讨论了通过构建专门针对机器人导航任务的基础模型来采取不同方法的研究。

在 VLN-BERT [135]中，作者提出了一种基于视觉语言变换器的模型，该模型利用多模态视觉和语言表示来使用网络数据进行视觉导航。模型旨在评估指令（例如，“在棕色沙发上停下”）与由代理拍摄的一系列全景 RGB 图像之间的兼容性。

摘要。遵循类似于“下楼走，在棕色沙发前停下”的导航指令，需要具身 AI 代理将通过语言引用的场景元素（例如，“楼梯”）与环境中的视觉内容（对应于“楼梯”的像素）进行关联。我们提出如下问题：是否可以利用庞大的未具身网络爬取的视觉-语言数据集（如 Conceptual Captions [24]）来学习视觉语义关系（“楼梯”长什么样子?），从而改善相对数据匮乏的具身感知任务（视觉-语言导航）的表现? 具体来说，我们开发了 VLN-BERT 模型，这是一种基于变压器的语言-视觉变换器，用于评估指令（如“…在棕色沙发前停下”）与代理拍摄的一系列全景 RGB 图像之间的兼容性。实验结果表明，在使用网络图片文本对预训练 VLN-BERT 之前，并在其上进行具身路径指令数据的微调，显著提升了 VLN 的表现——在完全可观测的情况下，成功率相比之前的最佳方法提高了 4 个百分点。我们对预训练课程的设计进行了剥离分析，结果显示每个阶段都具有重要作用，并且它们的结合使用产生了积极的协同效应。

考虑图 1 中的导航指令：“Walk through the bedroom and out of the door into the hallway. Walk down the hall along the banister rail through the open door. Continue into the bedroom with a round mirror on the wall and butterfly sculpture.” 在视觉与语言导航（VLN）[4]任务中，智能体必须解释此类指令以在逼真环境中导航。在这个例子中，智能体需要从卧室走出去进入走廊，沿着扶手前行穿过一个敞开的门，然后继续走进有一个圆形镜子和蝴蝶雕塑的卧室。但如果智能体从未见过蝴蝶会怎样呢?

LM-Nav [136]系统考虑了视觉导航任务。该系统利用预训练的图像和语言模型，提供了一个基于文本的接口以进行视觉导航。LM-Nav 通过自然语言指示，在现实世界的户外环境中展示了视觉导航能力。LM-Nav 利用了大规模语言模型（LARGE LANGUAGE MODEL, GPT-3 [2]），视觉语言模型（VISUAL LANGUAGE MODEL, CLIP [4]）以及视觉导航模型（VNM）。首先，VNM 通过估计图像间的距离构建环境的拓扑图。随后，LLM

将自然语言指令转换为一系列中间语言地标。VLM 利用联合概率分布将视觉观察与地标描述关联起来。利用 VLM 的概率分布、LLM 指令和 VNM 的图形连接性，使用搜索算法规划最优路径。最后，通过 VNM 的目标导向策略执行该计划。

在开放词汇导航领域，另一条研究线是对象导航任务。在这个任务中，机器人必须能够找到由人类描述的对象并导航至该对象。导航任务被分解为探索阶段（当语言目标未被检测到时）和利用阶段（当目标被检测到后，机器人开始朝向目标导航）。随着机器人在环境中移动，它会使用 RGB-D 观测值和姿态估计创建一个自上而下的环境地图。在文献[143]中，研究人员引入了一种零样本对象导航设置，该设置利用了诸如 CLIP [4] 等开放词汇分类器来计算图像与用户指定描述之间的余弦相似度。

3 任务书

基于大模型的零样本机械臂视觉导航

我们将通过 OpenAI Whisper 实现语音识别，https://github.com/ahmetoner/whisper-asr-webservice 参考此处直接使用 docker 进行部署。 https://zhuanlan.zhihu.com/p/678406937, 并且通过 Bark 来实现文字到语音实现机器人和用户的交互。

前端使用图像和文字提示，同时使用 YOLO 提供目标检测和基础的文字提示，同时调用多模态模型分析图象和文字，同时使用 GPT-3 进行对话，同时使用 BERT 进行文本分类, 同时使用 transformer 进行文本生成, 同时使用 CLIP 进行图像分类, 同时使用 VNM 进行视觉导航，同时使用 LM-Nav 进行视觉导航，同时使用 VLN-BERT 进行视觉导航

同时编写预先的函数以供大模型调用，使用类似专家的模式，将 AI 分为不同的专家，再执行导航任务时候，调用不同的专家，以实现更好的效果。

在仿真平台上搭建一套测试系统，包括环境，机械臂，测试不同的场景下的执行效率。

最后如有可能，在实际机械臂上进行部署和测试。

参考[1–8]

参考文献

[1] FIROOZI R, TUCKER J, TIAN S, u. a. Foundation Models in Robotics: Applications, Challenges, and the Future[EB/OL](2023). https://arxiv.org/abs/2312.07843

[2] VEMPRALA S H, BONATTI R, BUCKER A F C, u. a. ChatGPT for Robotics: Design Principles and Model Abilities[J/OL]IEEE Access, 2023, 12: 55682-55696. https://api.semanticscholar.org/CorpusID:259141622

[3] LIANG J, HUANG W, XIA F, u. a. Code as Policies: Language Model Programs for Embodied Control[EB/OL](2023). https://arxiv.org/abs/2209.07753

[4] MAJUMDAR A, SHRIVASTAVA A, LEE S, u. a. Improving Vision-and-Language Navigation with Image-Text Pairs from the Web[EB/OL](2020). https://arxiv.org/abs/2004.14973

[5] HUANG W, WANG C, ZHANG R, u. a. VoxPoser: Composable 3D Value Maps for Robotic Manipulation with Language Models[EB/OL](2023). https://arxiv.org/abs/2307.05973

[6] ZHAO H, PAN F, PING H, u. a. Agent as Cerebrum, Controller as Cerebellum: Implementing an Embodied LMM-based Agent on Drones[EB/OL](2023). https://arxiv.org/abs/2311.15033

[7] LI X, ZHANG M, GENG Y, u. a. ManipLLM: Embodied Multimodal Large Language Model for Object-Centric Robotic Manipulation[EB/OL](2023). https://arxiv.org/abs/2312.16217

[8] 王文晟, 谭宁, 黄凯, u. a. 基于大模型的具身智能系统综述[J/OL]自动化学报, 2024, 51(AAS-CN-2024-0542): 1. http://www.aas.net.cn/article/doi/10.16383/j.aas.c240542. DOI:10.16383/j.aas.c240542