arXiv:2303.03480v2 [cs.RO] 5 Nov 2023

具身智能体能找到你的“猫形杯子”吗？大语言模型引导的探索：用于零样本目标物体导航

Vishnu Sashank Dorbala[1],James F. Mullen Jr[1],和 Dinesh Manocha[1]补充材料（包括完整技术报告、代码和视频）请访问https://gamma.umd.edu/LGX/

摘要——**我们提出了 LGX（语言引导探索），这是一种用于**语言驱动的零样本物体目标导航（**L-ZSON**）**的新算法。在这一任务中，具身智能体需要在**此前未见过的**环境中，导航至**描述独特**的目标物体。我们利用大语言模型（LLM）的常识推理能力，依靠其为导航过程中的连续决策提供支持，从而完成这一任务。同时，我们使用预训练的视觉-语言定位模型进行通用目标物体检测。在 RoboTHOR 上，我们实现了当前最先进的零样本目标导航结果，成功率（SR）比当前基线 OWL-ViT CLIP on Wheels（OWL CoW）提高了 27% 以上。此外，我们研究了大语言模型在机器人导航中的应用，并分析了影响模型输出的各种提示策略。最后，我们通过**真实世界**实验展示了该方法的优势，实验结果表明，LGX 在检测视觉特征独特的物体并导航至其所在位置方面表现更优。**

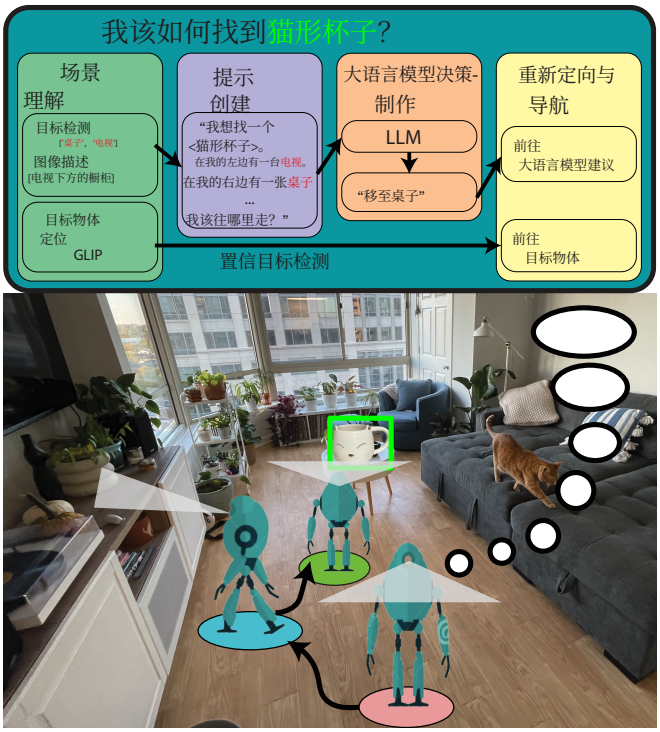

图 1：基于大语言模型的导航：我们的方法 LGX 旨在解决语言驱动的零样本目标导航（L-ZSON）问题。为了在未知环境中导航至并检测未见过、以任意方式描述的物体类别，我们首先提取环境中的视觉语义信息。随后利用这些信息为大语言模型（LLM）构造提示，大语言模型的输出则为我们提供物体子目标或笛卡尔方向，以引导具身智能体前往目标。同时，GLIP 会在环境中搜索目标物体，本例中的目标物体是“猫形杯子”。

I. 引言

人类在指代物体时不会拘泥于预设的类别标签，而是使用自由流畅的自然语言来描述。执行物体目标导航的机器人智能体必须能够理解这一看似无穷无尽、任意多样的物体集合，并在家庭环境中高效地导航至这些以自然语言描述的物体。例如，人类可能会让机器人智能体找到“猫形杯子”。在固定类别标签上训练的智能体可能会将这一请求理解为寻找一只“猫”或一个“杯子”，但人类实际指的是一个形状像猫的杯子。这类独特物体通常不属于 ImageNet 21k [1]和 OpenImages V4 [2]等大型图像数据集中常见的物体类别。此外，在家庭环境中部署的智能体可能需要在没有明确提供房屋地图或布局的情况下，导航至这些目标物体。

在我们的工作中，我们旨在通过解决*L-ZSON* 任务 [3]来应对这些问题。*L-ZSON* ，即语言驱动的零样本物体导航，要求智能体使用

以自然语言自由描述一个物体，并以“零样本”方式找到它，期间从未见过该环境或目标物体。传统的物体目标导航任务 [4]，[5]要求智能体在未见环境中，从预先确定的类别集合中找到一个物体。零样本物体导航（ZSON） [6], [7] 将此任务扩展为无需针对特定环境进行训练即可识别目标。*L-ZSON* 则进一步将 ZSON 扩展到自然语言自由描述。

用于物体导航任务的仿真环境，包括 RoboTHOR [8] 和 AI Habitat [9]，，仅包含使用

[1]作者来自美国马里兰大学帕克分校 vdorbala@umd.edu, mullenj@umd.edu, dmanocha@umd.edu

简单语言描述的常见日常家居物体（例如，杯子、桌子、床）。然而，人类在与智能体交流时往往使用不受约束的自然语言 [10]，，这可能导致智能体在理解上产生混淆 [11]。这一问题在常见物体导航模型的仿真到现实迁移[12] 中变得更加明显，因为这些模型无法理解人类的指令。在这项工作中，我们旨在通过开展针对独特物体指称（例如，橄榄色夹克）的真实世界实验来解决这一问题。

解决物体目标导航的常见方法基于全监督学习 [4]，[13]，，这对于预期能够检测任意描述的物体并在动态真实环境中稳定运行的智能体来说并不切实际。近期一些研究通过 ZSON 任务 [14],解决了在新地点的泛化问题，然而，通过 L-ZSON 任务 [15] 解决新物体泛化问题的研究更少，而且没有研究探讨包含大量不受约束语言的真实世界测试案例。

这些研究利用 CLIP [16] 和 GLIP [17] 等大规模预训练模型，在真实场景中执行零样本开放词汇目标检测。此类“基础模型” [18] 的下游迁移已在多种视觉与语言任务中展现出显著改进，例如图像描述 [19] 和问答 [20]。然而，将其迁移到机器人学并非易事，因为与视觉和语言任务不同，机器人任务通常涉及某种形式的经验性 决策：智能体会持续与环境交互。利用这些模型中蕴含的隐式知识来组合机器人动作，是一项独特的挑战。

主要贡献： 受上述挑战的启发，我们提出了一种新颖的方法：**LGX**，即语言引导探索，它利用大语言模型（LLM）和预训练视觉与语言模型中的隐式知识来应对 L-ZSON 任务。

包括 L-ZSON 在内的目标物体导航可分解为两个关键组成部分——序贯决策 和目标物体定位。前者指在探索过程中实时做出探索决策，后者则指从智能体感知中定位并接地 目标物体。

在这项工作中，我们利用大语言模型（LLM）和视觉-语言（VL）模型，解决阻碍这两个组成部分性能的泛化能力问题。由于大语言模型的表现依赖于所使用的提示[21]，我们通过上下文学习 [22]研究提示构造的影响，并逐一分析各种提示类型所产生的效果。此外，我们分析了视觉-语言模型在目标物体定位中的应用，并展示了采用独特物体指称能够提升性能。我们的贡献如下：

1）我们提出了 LGX，这是一种用于解决 L-ZSON——一项语言引导的零样本目标物体导航任务——的新颖方法。我们的方法利用大规模视觉-语言（VL）模型对由非受限语言描述的物体进行定位，并利用大语言模型（LLMs）中内置的物体间语义关联。具体而言，我们研究了大语言模型在辅助完成零样本目标导航所需的序贯导航决策时所具备的隐式常识推理能力。2）我们的方法利用环境的视觉场景描述来构建提示供大语言模型使用，其输出驱动我们的导航方案。我们研究了多种类型的提示，并为如何成功利用这些提示进行机器人导航提供了见解。3）我们的方法在 RoboTHOR 上将最先进的零样本成功率（SR）和路径长度加权成功率（SPL）提高了 27%。4）最后，我们还将该方法迁移至一个真实世界机器人平台，并研究了这一场景中涉及的各种复杂性。据我们所知，我们的方法是首个在现实世界中评估 L-ZSON 方法性能的方法。

II. 相关工作

A. 语言引导机器人学

使用语言引导机器人是文献中一个热门的研究任务，相关工作涵盖了使用通用化的落地图 [23]进行机器人操作 [24]，以及执行语言引导导航 [25], [26]。Thomas 等人在 [27]中提出了一种方法，通过系统性的基于概率图的方法解析非受限自然语言。Jesse 等人较近期针对这一问题开展的工作 [28], [29]以及 Gao 等人的研究 [30]，探索了利用人机对话收集与完成任务相关信息的方法。解析非受限自然语言与我们的工作高度相关，我们的研究也受到了这些论文所提出技术的启发。

B. 语言驱动的零样本导航

近期研究尝试使用 CLIP [16] 执行零样本具身导航。CLIP 是一种经过大规模预训练的视觉-语言模型，能够进行零样本目标检测。Dorbala 等人在 [31] 中以零样本方式执行视觉与语言导航，而 Gadre 等人在 [15] 中使用它执行物体目标导航。这两项工作都基于未见环境这一假设。

Gadre 等人在 [3]中提出的 L-ZSON 针对使用不常见目标物体的零样本目标导航问题展开研究。他们使用 OWL-ViT（一种经过微调、用于目标物定位的视觉 Transformer）和基于前沿的探索（FBE） [32]，为该任务建立了基线。相比之下，我们的方法使用 GLIP [17], 这一经过预训练的视觉-语言模型进行零样本目标物接地。为了探索环境，我们引入 GPT-3 [22], 这一大语言模型来做出导航决策。

C. 语言引导的场景操作

语言引导的场景操作是一项热门任务，要求智能体处理自然语言指令，以操控环境中的物体。[33]–[35] 的作者提出了多种基于语义分割的方法，以语言提示为输入执行场景操作和物体放置。此外，也有许多经典的语言引导操作方法 [24],[36]–[39] ，它们利用不同形式的接地图来完成这项任务。Raman 等人在 [40]中解决了一项独特的操作任务：合成机器人可用来向用户解释其失败原因的语言。

近来，若干研究提出了利用预训练嵌入进行语言条件操作任务的新方法 [41]–[44]。与这些操作任务不同，我们的工作聚焦于语言引导探索（LGX）。

D. 用于高级任务规划的大语言模型

近期研究已证明，大语言模型可作为一种工具，为智能体制定执行计划，将高级指令（例如“给我做早餐”）拆解为供智能体执行的低级任务（例如“做鸡蛋”和“拿果汁”） [45]–[47]。请注意，这些方法利用大语言模型为机器人生成和规划动作，但并未明确将其用于未知环境中的探索。这些方法假定环境信息已知，并使用大语言模型将复杂任务拆解为机器人可执行的任务。

Huang 等人在 [45] 中探讨如何利用大语言模型将复杂任务简化为“可执行子任务”，供机器人执行。他们的任务涉及利用预训练大语言模型在规划层级上，确定如何在可交互环境中

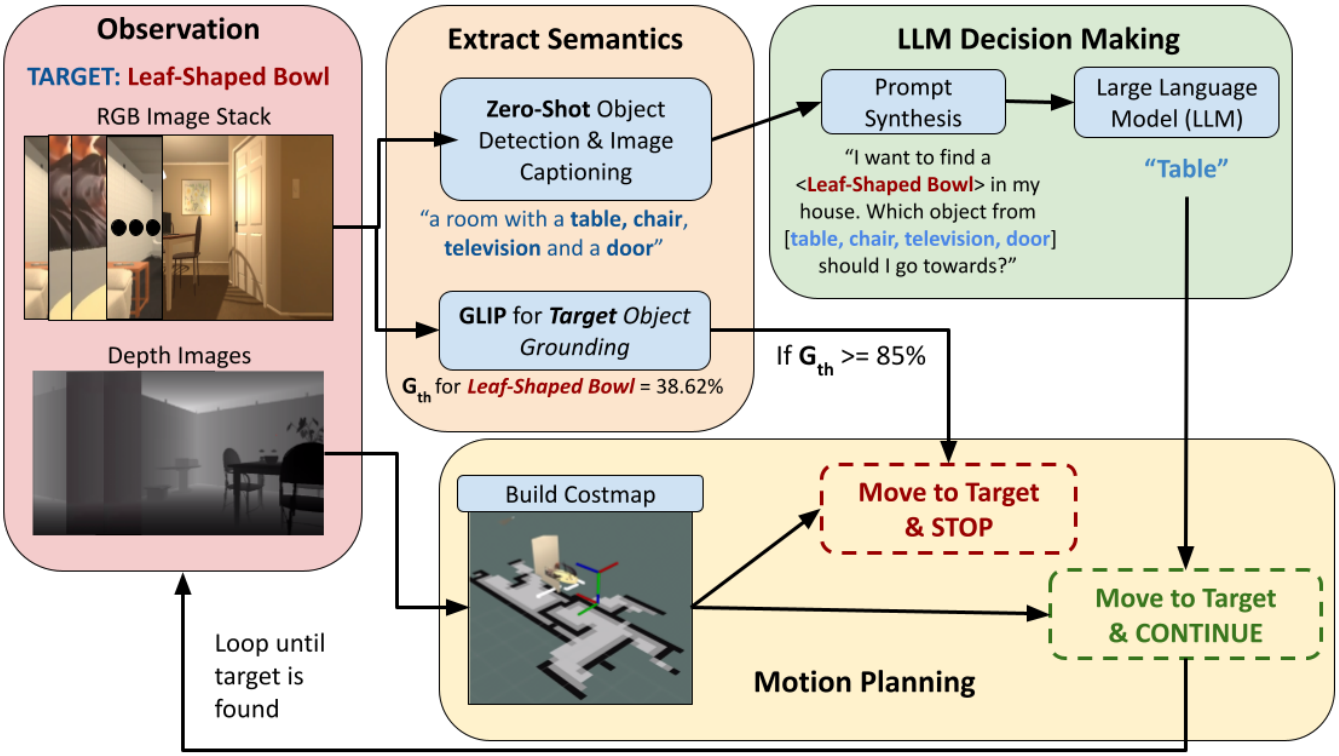

图 2： 我们方法的概览。首先，我们通过进行 360度旋转，从环境中收集观测数据，以获取智能体周围的深度图像和 RGB图像。RGB图像为我们提供智能体视野中物体的语义信息，而深度图像则使我们能够创建代价地图。随后，我们利用提取出的物体标签为大语言模型合成提示。最后，大语言模型根据物体列表生成输出，驱动导航方案，告知智能体应朝哪个方向前进。与此同时，我们尝试借助 GLIP 在场景中定位目标物体。找到目标后，我们便退出决策循环并直接导航至目标物体。

与已知的 VirtualHome 环境进行交互，且不利用交互结果来探索环境。

SayCan 在 [46] 中以类似的方式利用大语言模型推断智能体可以执行的动作，但还使用可供性函数，将生成的动作与环境联系起来。他们假设地图中的物体位置已知（见该论文第 D.2 节），只解决“规划问题”，并预先提供了拾取、前往和放置等动作。他们没有考虑像 LGX 那样在未见环境中进行导航的零样本探索情形。

Huang、Xia 和 Xiao 等人在 [47]，中也遵循类似的规划目标，并将大语言模型与人类反馈结合起来完成指令。

相比之下，我们的工作将大语言模型用作先验，并非用于规划，而是用于探索未见环境，以预测机器人需要前往的方向，且无需地图。

E. 语言引导导航中的大语言模型

大语言模型在机器人学中的应用最近引起了广泛关注。近期的一些研究 [45]–[47] 也专门利用大语言模型的规划能力。请注意，我们与这些研究不同：它们将大语言模型用于规划目标，而我们则将大语言模型用于环境探索。

已有少数研究探讨了将生成模型用于导航，具体而言，包括 LM-Nav [48] 和 VLMaps [49]。这两项研究都致力于解决视觉与语言导航（VLN）问题，其输入是对通往目标的路径进行描述的不受限制的语言。后者使用 GPT-3 [22] 获取“地标”或子目标，而前者则侧重于利用大语言模型进行“代码编写” [50]。相比之下，我们关注的是将视觉场景语义转化为大语言模型的输入提示，以获取以动作形式呈现的导航指引。我们直接将大语言模型的输出纳入序贯决策流程，从而驱动智能体的导航方案。

III. 使用语言引导探索（LGX）求解 L-ZSON

A. 方法概述

我们在图 2 中概述了我们的方法。我们的方法使用大语言模型（LLM）来预测智能体需要导航前往的位置。为此，我们首先从场景中提取上下文线索，形式可以是物体标签或场景描述。随后，我们使用其中任一线索来设计提示，询问 LLM 智能体应如何继续探索环境。LLM 利用其关于环境中物体关系的常识知识，为智能体提供移动方向或目标物体。

同时，我们使用视觉-语言模型 GLIP[17]来获取目标对象定位分数，该分数反映所描述的目标物体位于场景中的置信度。一旦置信度达到阈值 G_{th}，智能体便会假定目标物体位于其自我中心视角中。一个

若在原地旋转过程中，目标物体出现在智能体的视野中，则判定该回合成功。

B. 场景理解

在每次运行中，智能体都会观察环境，采集 RGB 和深度图像以供检查。每次观察时，我们让机器人原地旋转 360 度，并以设定的分辨率r拍摄图像。这样我们便获得了 $360/r$ 张 RGB图像I_r和深度图像I_d。I_d 用于构建环境的二维代价地图。随后，将I_r 中的每张图像输入目标检测模型或图像描述模型。这两种模型会给出不同的结果，我们将在实验部分对此进行讨论。目标检测方面，我们使用 YOLO [51], ，它包含常见的家居类别；而 BLIP [52] 则为我们生成图像描述。BLIP 会生成描述环境的图像描述C，YOLO 则会列出智能体周围可能前往的物体O。我们选择C 或O 作为提供给大语言模型的提示的一部分。

每一步都原地旋转，可以让智能体充分观察周围环境，为大语言模型提供足够的信息，使其能够根据智能体当前在环境中的位置，做出充分知情的导航决策。若不进行原地旋转，智能体就会朝着已看到的物体穿过未知区域前进，即使这些物体都与目标物体o_g无关。例如，如果o_g 是“蓝色枕头”，但智能体初始化时面朝厨房，并看到了“微波炉”“杯子”和“桌子”等物体，那么机器人就会继续探索这些物体附近，因为它不知道正后方是“床”或“沙发”，而枕头可能就在这些地方。

同时，在进行整圈旋转的过程中，智能体使用I_d 构建环境的代价地图。我们使用 RTABMAP [53]，结合视觉对应关系和标准 ROS 二维代价地图软件包 [54], 提供的深度信息来计算代价地图。大语言模型通过提供物体或方向（取决于输入中传入的是C 还是O ）做出导航决策后，我们会相应地调整智能体的朝向，并在代价地图中沿智能体的自我中心视野随机选择一个点。代价地图帮助我们在探索环境时避开障碍物。

我们使用 GLIP 进行目标物体定位。在智能体每次旋转时，都会将采集到的 RGB图像连同作为提示的目标物体一起输入 GLIP。当 GLIP 的定位准确率超过阈值G_{th}时，我们便认为目标物体位于智能体的视野中，并触发停止信号。否则，智能体会继续探索，直至原地旋转n_r 次。如果真实目标物体位于视野中，则判定该回合成功。

G_{th} 和n_r 是通过消融实验经验性选定的超参数。为选择G_{th}，我们在某个环境中对多个阈值进行消融实验，并选取成功率最高的阈值（参见表 I）。n_r 则根据环境的大小选定。

C. 使用大语言模型进行智能探索

我们利用提取出的语义，为大语言模型 GPT-3 设计提示。我们探索了两种情形：

1)**YOLO $\rightarrow$ 大语言模型**：在此情况下，我们利用 YOLO 检测到的智能体周围的物体列表O 来合成提示。为提升目标检测效果，我们在此将旋转分辨率r 设得更低。

2) **BLIP $\rightarrow$ 大语言模型**：在此，我们利用上一步生成的图像描述C来创建提示。这里将旋转分辨率r 设为 90，表示智能体在探索时可采取的四个方向（左、右、前、后）之一。

采用 YOLO + 大语言模型方法时，大语言模型的输出会从O 中给出一个智能体应前往的物体。随后，智能体会重新调整朝向，面向该物体。采用 BLIP +大语言模型时，输出会给出一个方向，智能体据此重新调整朝向。当大语言模型的输出不符合预期时，智能体会随机选择一个方向。

D. 目标检测与运动规划

在完成 360° 原地旋转的同时，我们还运行 GLIP，并将目标物体 o_g 作为其目标标签。如果 GLIP 以足够高的置信度G_{th}找到目标物体o_g，我们便在目标位于智能体前方时终止探索循环。随后，我们将模拟器作为真实值，以判断目标是否确实位于智能体前方，从而确定是否成功。如果 GLIP 未找到目标，我们便继续基于大语言模型进行探索。

每一步中，我们首先根据大语言模型的决策调整智能体的朝向。然后，我们使用构建的代价地图（参见 III-B 节），沿智能体当前朝向选取一个距离e_d 固定的探索点。接着，我们使用标准的 ROS *move base*软件包来避开障碍物。

IV. 分析我们的方法

A. 使用 *GLIP* 进行零样本检测

开放词汇（OV）定位模型在物体级识别任务中展现出强大的零样本性能，并证明能够很好地泛化到众多数据源。在我们的方法中，我们使用 GLIP 对感兴趣的目标物体进行定位。它会输出一个边界框，使我们能够在智能体的视场中定位用于导航的方向。GLIP 的输出可定义为

$$\{o_{t,i}, b_{t,i}\} = GLIP(I_t, P_o) \tag{1}$$

其中$o_{t,\ i}$ 和$b_{t,\ i}$ 分别表示目标检测结果和边界框。I_t 和P_o 分别表示输入图像和输入提示词，用于定义感兴趣的物体。

我们选择 GLIP 作为定位模型。在与其他先进的开放词汇检测模型进行消融实验后，我们认为其表现更优，这些模型包括 OWL-ViT[55] 、以物体为中心的开放词汇检测 [56] 和 Detectron2 [57]。图3展示了 GLIP 对 L-ZSON 的应用价值。它不仅能识别通过自然语言定义的物体“猫形杯子”，还能将其与相关物体区分开来。基于这一特性，在原地旋转过程中运行 GLIP，使我们能够可靠地检测目标物体o_g，无论对它们的描述方式如何。

B. 检视用于探索的大语言模型提示词

GPT-3 的结果会受到所提供提示词的极大影响。由于我们直接利用它的常识推理能力（猫形杯子更可能在桌子附近，而非床边）来进行导航，因此我们必须考虑各种提示策略，这一点很重要



图3： 在给定图像上输入字符串“猫形杯子 . 猫 . 杯子”时，GLIP 输出的示例。GLIP 能够成功定位独特的物体，例如“猫形杯子”，并将其与相关物体区分开来，例如猫或杯子。

对于我们而言，如果大语言模型未能提供有效回复，智能体就会朝随机方向移动。我们比较了七种不同的大语言模型提示词，它们都是对以下模板的变体：

> “你正在控制一台家用机器人。这台机器人想要在我家里找到一个 o_g 。从 $\{O\}$ 机器人应该 g应该前往哪里？请用一个物体列表中的物体。”

我们首先探究不同视角如何影响大语言模型反馈。以下是这些提示：

- **Robot-Prompt**: “你正在控制一台家用机器人。这台机器人想在家里找到一只 og。它应该前往 $\{O\}$ 中的哪个物体？请从物体列表中回复一个物体。”
- **I-Prompt**: “我想在家里找到一只 og。我应该前往 $\{O\}$ 中的哪个物体？请用一个词作答。”
- **第三人称提示**：“房子里有一只 og。 $\{O\}$ 中的哪个物体最有可能离 og 最近？请从物体列表中回复一个物体。”

其次，我们改变提供给提示的信息顺序。

- $\{O\}$-**优先提示**：“你正在控制一台家用机器人。你必须从 $\{O\}$ 中选择一个物体，让机器人朝它前进，以便尝试在我家中找到 og。请从物体列表中回复一个物体。”
- **获取最近物体提示**：“你正在控制一台家用机器人。机器人想在我家中找到一个 og。 $\{O\}$ 中的哪个物体可能离 og 最近？请从物体列表中回复一个物体。”
- **“一个词”优先提示**：“请用一个词回复。你正在控制一台家用机器人。机器人想在我家中找到一个 og。 $\{O\}$ 中的哪个物体应该是机器人前往的目标？”

最后，我们使用场景的自然语言描述来创建提示。

- **BLIP-Prompt**：”我想在家里找到一只 og。你前方有 $<caption>$。右方有 $<caption>$。后方有 $<caption>$。左方有 $<caption>$。我应该朝前、右、后还是左走？请用一个词回答。”

五、实验与结果

A. 实验设置

模拟设置。我们使用 RoboTHOR [8]验证集作为实验的模拟环境。它包含 1800 个验证回合，涵盖 15 个验证环境。共有 12 种不同的目标物体类别。每个探索回合都会执行前文第 III-B 和 III-C 节所述的场景理解流程。对于每个回合，我们运行 LGX 进行 n_r 轮探索，或直到检测到置信度高于 G_{th} 阈值的目标物体。这些限制构成了停止条件。鉴于 RoboTHOR 的环境较小，目标物体通常位于距离出生点 10 米以内，因此将n_r 设为 5。根据表 I 中所述的消融实验，RoboTHOR 的G_{th} 设为 0.85。第 III-D 节所述的探索距离e_d 设为 5 米，以确保机器人移动后 RoboTHOR 的场景发生显著变化。

G_{th}	成功率（%）	SPL（%）
0.6	13.8	7.2
0.75	20.3	10.8
0.8	32.5	18.7
0.85	**35.0**	**21.9**
0.95	18.0	11.3

表 I： 在 RoboThor 验证集上对G_{th} 进行消融实验：G_{th} 根据消融实验的实证结果设置阈值。取值较低会产生大量假阳性，导致性能不佳。相反，取值较高则会排除许多可能成功的情况。

提示选择设置。 我们在表现最佳和最差的 500 个回合子集中，分别运行上一节介绍的 7 个提示。

指标。 我们报告并比较成功率（SR）以及按路径长度倒数加权的成功率（SPL） [58]。SR 和 SPL 是 Habitat 和 RoboTHOR 挑战赛中使用的主要指标。在提示消融实验中，我们定义了一项新指标——**提示成功率（PSR）**：

$$PSR = \frac{p_{suc}}{p_{total}} \tag{2}$$

其中，p_{suc} 表示大语言模型选择有效响应的次数，p_{total} 表示智能体向大语言模型发送提示的总次数。根据所用的语义提取方案，当大语言模型选择智能体检测到的物体或用于导航的方向时，其响应即为有效响应。

真实世界设置。 我们使用 TurtleBot 2 开展实验，以验证 LGX 在现实世界中的两个方面：i）**大语言模型的探索能力**；ii）**基于 GLIP 的开放词汇定位**。

为验证 i)，我们考察一种**两阶段方法**，其中智能体需要从一个房间出发，穿过一条“走廊”，到达包含目标物体的房间。在**第一阶段**，智能体在生成所在的房间内原地旋转，收集周围物体的信息。由于这个房间里没有目标物体，预期大语言模型的输出为“走廊”。在**第二阶段**，智能体位于走廊中，需要根据各个房间内常见物体的集合，选择正确的房间前往（参见表 II）。此时，预期大语言模型的输出会指向最有可能包含目标物体的房间，这一判断基于常识知识（“遥控器”通常在“沙发”附近）。图 4 对此进行了详细说明。

为验证 ii)，我们检验 GLIP 对独特家居物体类别的分类准确率。为此，我们首先定义并选取独特的目标物体，以及属于不同家庭房间的常见物体。这些物体

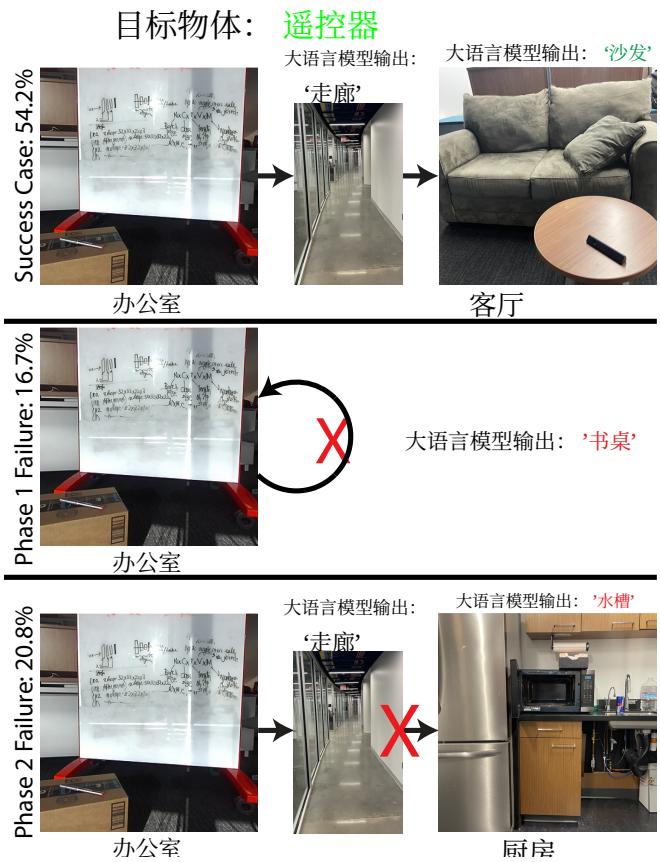

图 4： 为验证大语言模型的探索能力，我们定义了一个两阶段过程。目标物体位于另一个房间中，要求智能体从当前房间导航至‘走廊’。LGX 中的大语言模型将当前房间内的物体以及走廊作为输入传递给大语言模型。未到达‘走廊’则视为第一阶段失败。在第二阶段中，可以看到四个可能的房间，智能体必须导航至包含目标物体的房间。我们将表 II 所示的每个房间的一组常见物体作为输入传递给 LGX 中的大语言模型。未选择正确的房间则视为第二阶段失败。

见表 II。成功案例定义为 GLIP 成功检测到目标物体。

Room	目标物体	常见物体
厨房	红牛罐、甜菊糖包	水槽，冰箱
客厅	遥控器，茶几	沙发，电视
卧室	半身像，橄榄色夹克	床、毯子
办公室	银色钢笔、白板	书桌、电脑

表 II：用于验证大语言模型探索的物体设置。我们定义了四种家庭房间，其中摆放着可能在这些房间中找到的常见物体和目标物体。常见物体是普通的家居用品，而目标物体则使用自由形式语言进行独特描述。

此实验设置通过真实世界的两个主要复杂因素来验证我们的系统：自由形式自然语言和部分可观测环境。对于自由形式自然语言问题，我们为每个目标物体提供独特描述，而不是使用常见的视觉类别标签。对于部分可观测环境这一因素，我们开展了两阶段探索实验，每个房间中只有少量物体可见，以此模拟真实世界的住宅。

模型	RoboTHOR 成功率 (%) ↑	RoboTHOR SPL (%) ↑
CoW (FBE +CLIP) [3]	15.2	9.7
OWL-CoW（FBE + OWL-VIT） [3]	27.5	17.2
轮上 GLIP (FBE + GLIP)	33.2	20.3
LGX（BLIP → 大语言模型 + GLIP）	28.46	13.5
LGX（YOLO → 大语言模型 + GLIP）	**35.0**	**21.9**

表 III： RoboTHOR 结果：请注意，与当前最先进方法 CoW 和 OWL 相比，我们的两种方法在成功率（SR）和 SPL 方面均有所提升。

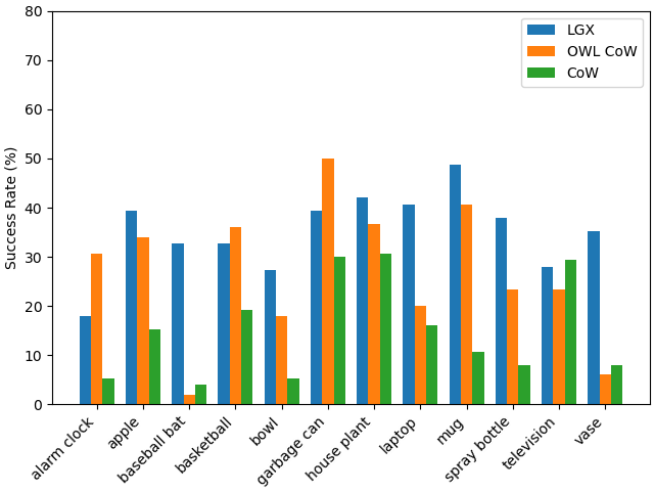

图 5： LGX 与 RoboTHOR 上的 OWL CoW 和原始 CoW 的类别细分对比。LGX 在定位棒球棒、碗、笔记本电脑、喷雾瓶和花瓶类别方面取得了显著提升。在电视和垃圾桶等较大类别上也表现出相近的性能。

B. 基线与消融实验

我们将自己的方法 LGX 与两种最先进的方法以及一种消融方法进行比较：

轮上 CLIP（CoW）。 [3] 使用 Grad-CAM——一种基于梯度的可视化技术——结合 CLIP [16] 在自我中心视角中定位目标物体。CoW 采用基于 Frontier 的探索技术进行零样本目标导航。

OWL CoW。 [3] 使用 OWL-ViT 变换器代替 CLIP 模型来进行目标物体定位。随后，该检测器在 CoW 方法中取代 CLIP。

轮上 GLIP（GoW）。 [3] 使用 OWL-ViT 变换器进行视觉对象定位，而我们将其替换为基于 GLIP 的定位系统。这也构成了对我们方法的消融，即去除基于大语言模型的探索机制。

随机方法与 GLIP。作为对 LGX 进行真实世界分析时的基线，我们还选择了一个随机方向选择器进行探索。智能体随机做出决策，复现一种“无先验信息”探索方法的行为。

C. 与仿真基线方法的比较

我们在表 III 中将我们的方法与为 L-ZSON 任务配置的其他模型进行了性能比较。我们的方法显著优于 OWL CoW 和原始 CoW，在成功率（SR）和 SPL 两项指标上均有所提升。LGX 的成功率也高于 GoW。这可以归因于我们在使用 GLIP 进行目标定位的基础上，改进了大语言模型驱动的探索方案。

模型	RoboTHOR 成功率 (%) ↑	RoboTHOR PSR (%) ↑
BLIP-Prompt	29.2	**100**
I-Prompt	33.3	87.7
Robot-Prompt	**33.8**	71.1
第三人称	31.3	99.4
{O}第一	**33.8**	95.3
获取最近物体	32.1	95.7
”一个词” 第一	28.1	52.3

表 IV： RoboTHOR 上七种不同提示在三个变化轴上的比较。基于物体的提示（中间和底部）优于基于自然语言的提示。

图 5 中，我们直接比较了 CoW 和 OWL 在 RoboTHOR 中对不同目标物体的表现。我们的方法在“碗”和“花瓶”等较小物体上的表现优于这两个基线方法；在“电视”等较大物体上的表现则与 OWL 相近。这些结果表明，CoW 的性能存在不足，这很可能是由于 CLIP 无法有效地在图像中定位目标物体。

D. 提示调优策略的影响

如表 IV 所示，尽管 BLIP 生成的基于自然语言的提示具有完美的 PSR，但其表现仍不如基于物体的提示。我们认为，这是因为采用基于 BLIP 的提示方案时，动作空间有限。基于物体的提示为大语言模型提供了许多不同的路径选择，而基于 BLIP 的提示则按定义与四个基本方向相关联，可能会将智能体引向死胡同。此外，我们注意到，在某些回合中，大语言模型会使智能体陷入循环，连续选择相反的方向。

在我们提示调优的第二个维度上，即大语言模型相对于机器人的视角，任务成功率没有显著差异。尽管不同视角对应的 PSR 差异很大，结果仍是如此。机器人视角的提示取得了最高的成功率，但在所研究的各视角中，其 PSR 也最低。值得注意的是，使用机器人视角的提示时，相较于其他提示中更常见的空响应，大语言模型更频繁地回复“no”或“nothing”。

在对提示结构所做的各种改动中，物体集优先提示和获取最近物体提示的成功率均无显著差异。然而，“一个词”优先提示（即将“用一个词回复”这一短语置于提示其余部分之前）表现出的成功率和 PSR 显著较低。我们认为，这是因为当这条指令被放在提示其余部分之前时，大语言模型不再遵循它。获取最近物体提示的高 PSR 表明，选择可能距离最近的物体或许

这对大语言模型而言可能是一个更简单的问题。同样，物体集优先提示的 PSR 较高，表明将物体集放在提示开头时，大语言模型能够更好地参考物体集。

我们认为，尽管 PSR 发生了较大变化，任务成功率却几乎没有差异，这进一步表明，RoboTHOR 为这类依赖上下文的场景智能探索提供了有偏差的依据。

E. 与真实世界基线方法的比较

在我们的真实世界实验中，我们采用了前文实验设置中所述的两阶段方法。如表 V

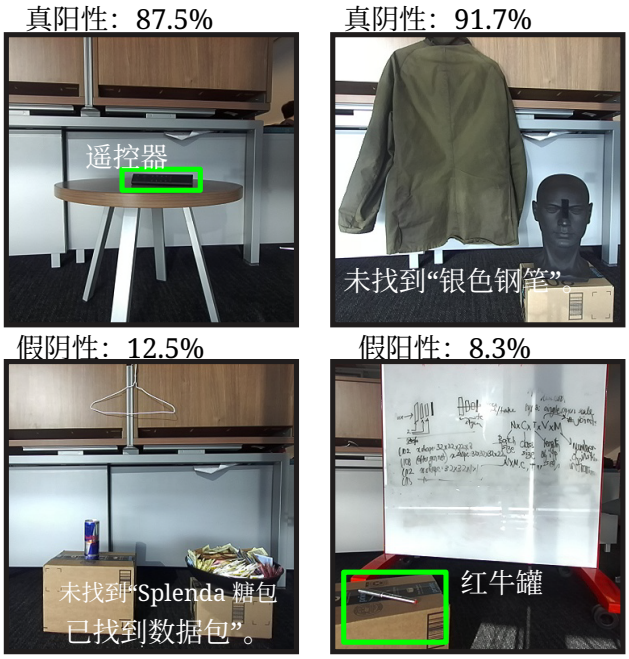

图 6： 我们在现实世界实验中，GLIP 成功和失败案例的示例。当场景中存在目标物体时，GLIP 有 87.5% 的概率准确检测到它。相反，当场景中不存在目标物体时，GLIP 有 8.3% 的概率误检。

方法（导航决策 +定位）成功率（%）↑	
随机 + GLIP	6.9
GLIP-on-Wheels（GOW）	27.8
LGX（YOLO → 大语言模型 +GLIP）	**54.2**

表 V： 在现实世界的实验中，我们的模型显著优于随机基线和 GLIP-on-Wheels 基线。尽管这三种方法都使用 GLIP 进行目标物体检测，但两种基线方法都没有在任务的探索阶段整合场景上下文。

所示，LGX 显著优于选定的基线方法，成功率比 GoW 提高了 26.4%，比随机方法与 GLIP 提高了 47.3%。因此，GoW 有 33% 的时间能够导航到正确的房间，而随机方法与 GLIP 探索起始场景中物体的频率，与探索走廊的频率相同。所有成功率也受到 GLIP 失败案例的影响，具体而言，在尝试检测“甜菊糖包”时出现假阴性，以及将“红牛罐”和“甜菊糖包”误检为目标的假阳性（见图 6）。

在现实世界实验中，我们的方法中的大语言模型行为可分为三种可能情况，如图 4 所示。成功情况占 54.2%，此时机器人智能体成功从起始房间进入走廊，再进入包含目标的房间，随后使用 GLIP 检测到目标。

在一个第一阶段的失败案例中，智能体没有进入走廊，因为大语言模型认为起始场景中的某个物体很可能会引向目标。我们注意到的一个例子是，当目标物体是‘红牛罐’时，如果起始场景是办公室，大语言模型会输出‘书桌’。尽管目标罐子实际上在厨房里，但它有可能放在“书桌”上，这解释了大语言模型为何会给出这一输出。在第二阶段的失败案例中，智能体

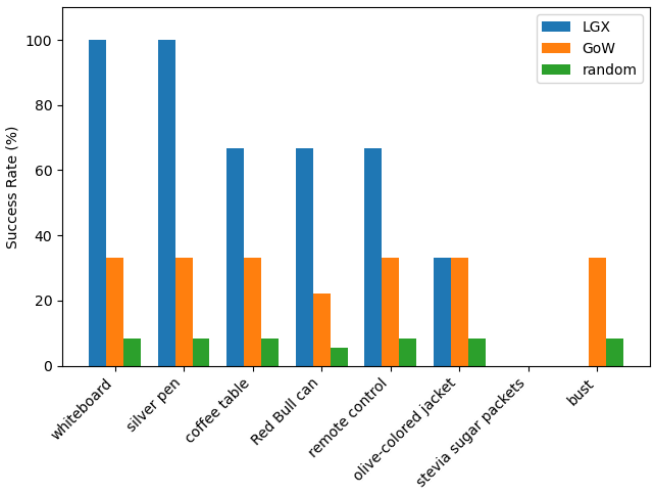

图 7： 在我们的真实世界研究中，对比我们的方法与基线方法在各个物体上的成功率。我们的方法在大多数目标上均优于基线方法，但值得注意的是，它未能定位“半身像”物体。由于 GLIP 未能检测到“甜菊糖包”，所有方法均无法定位该物体。

进入了一个不包含目标物体的房间。使用我们的方法时，这种情况占 20.8%。这类情况与大语言模型未能妥善关联目标物体与目标房间中的其他物体有关。一个例子是“橄榄色夹克”，大语言模型通常认为它会出现在“书桌”附近。图 7 展示了系统在各个目标物体上的性能细分。我们的方法未能定位“半身像”，误以为它与“书桌”有关。然而，基线方法相对成功，这表明 GLIP 在进入正确房间后成功检测到了“半身像”。

更多细节和消融实验将在讨论部分介绍。

VI. 局限性、结论与未来工作

在本工作中，我们提出了一种用于基于语言的零样本物体目标导航的新型算法。我们的方法利用大语言模型（LLM）的能力进行导航决策，并利用开放词汇的基础模型检测以自然语言描述的物体。我们在 RoboTHOR 基准上展示了当前最先进的结果，研究了驱动探索的大语言模型提示词的结构与措辞，并通过真实世界实验验证了我们的方法。

我们的方法仍存在一些失败情况，尤其是在大语言模型错误定位目标物体时。未来工作应通过过滤检测到的物体列表或提供已访问物体的历史记录，探索改变输入大语言模型的上下文。同样，研究哪些物体会产生异常显著的影响也将十分有益。未来工作还应致力于改进 SR 和 SPL 指标，使其能够为零样本导航任务提供更有意义的信息。

VII. 讨论

本节中，我们展示了一些补充实验和分析，同时也回答了有关 LGX 在现实世界中适用性的一些相关问题。

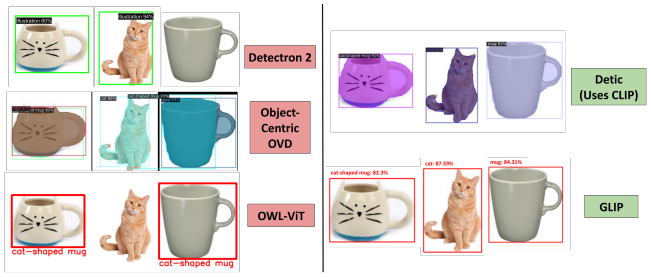

图 8：歧义性消融实验：我们针对各种开放词汇模型进行消融实验，以评估使用自然语言描述物体时存在的歧义性。Detectron2 [57]，OCOVD [56] 和 OWL-ViT[55] （用于 OWL-CoW）会产生假阳性，有时还无法检测到物体。另一方面，Detic [59] 和 GLIP [17] 在此类开放词汇定位任务中表现良好。我们决定使用 GLIP，因为它易于实现；在我们的框架中使用 Detic 的实验将作为未来工作的一部分。

A. 现实世界挑战

自由形式自然语言描述和部分可观测环境使物体导航成为现实世界中的一项复杂任务。在零样本设置下，性能通常取决于两个主要因素；**探索** 和 **视觉对象定位**，二者在现实世界中的实现都面临挑战。具体如下

- **视觉对象定位**：仿真环境中的目标物接地比现实世界中的目标物接地受到更多限制。仿真器具有数量固定的可用类别，而有时这些物体的特征并不可用于根据其自然语言描述进行定位。现实世界中的用户通常会使用自由形式语言，因此智能体必须相应地进行适应。我们尝试通过使用开放词汇定位模型 GLIP 来克服这一挑战。GLIP 根据提示中使用的语言定位目标，而不是寻找所有视觉对应关系，从而能够检测以自由形式自然语言描述的物体。我们在现实世界实验的第二部分中探究了 GLIP 的性能，相关内容见图 **5**和**6**。

- **探索**：在真实环境中探索时，智能体必然身处动态变化的部分可观测环境中，整个地图及相关目标物体可能持续变化。诸如基于前沿的探索（FBE）这类无信息的探索方法可能效率低下，因为静态环境这一假设并不成立，导致智能体重复探索已经到访过的区域。我们利用大语言模型进行有信息的决策，依靠常识知识探索环境，以应对这一问题。与依赖在自由空间中选择前沿的 FBE 不同，我们基于大语言模型的探索方法不依赖地图，而是根据智能体周围的环境，对其应采取的方向作出有依据的选择。在现实世界基准测试中，我们通过第**V.A** 节和图**4**所介绍的两阶段方法，验证大语言模型的决策能力。

B. 为什么选择 *GLIP*?

我们的方法使用 GLIP [60]，，它是一种开放词汇目标检测器。它不会从开放集中提议所有物体，而只检测文本提示中提及的物体

。作为一种开放词汇目标检测器，GLIP 能够处理含糊的文本描述，并在环境中定位这些描述所指的内容。LGX 的物体导航性能确实对 GLIP 的性能十分敏感。论文中的图3展示了 GLIP 在应对对不具备开放词汇能力的模型而言模棱两可的场景时所具备的卓越能力。当用户提到“猫形杯子”时，GLIP 能够从包含猫 和杯子的图像组中成功识别出这一独特物体。我们在下方图8中展示了其他最先进的零样本开放词汇目标检测器的性能消融实验—— *Detectron2*[57]，*OCOVD* [56]，*OWL-ViT* [55]，*Detic* [59]和*GLIP*[60] 。图8定性地展示了积极的推断结果：GLIP 能够成功应对使用自由形式语言描述的模糊情形，进一步印证了我们在 LGX 中采用开放词汇定位的合理性。

参考文献

[1] J. Deng、W. Dong、R. Socher、L.-J. Li、K. Li 和 L. Fei-Fei，“ImageNet：大规模层次化图像数据库”，发表于 *2009*年*IEEE*计算机视觉与模式识别会议，2009年6月，第248–255页。[2] A. Kuznetsova、H. Rom、N. Alldrin、J. Uijlings、I. Krasin、J. Pont-Tuset、S. Kamali、S. Popov、M. Malloci、A. Kolesnikov 等，“Open Images数据集v4：大规模统一图像分类、目标检测与视觉关系检测”，发表于《国际计算机视觉杂志》，第128卷第7期，第1956–1981页，2020年。[3] S. Y. Gadre、M. Wortsman、G. Ilharco、L. Schmidt 和 S. Song，“CoWs on Pasture：语言驱动的零样本目标导航的基线与基准”，2022年12月。[4] S. K. Ramakrishnan、D. S. Chaplot、Z. Al-Halah、J. Malik 和 K. Grauman，“Poni：通过无交互学习实现物体目标导航的势函数”，发表于 *IEEE/CVF*计算机视觉与模式识别会议论文集，2022年，第18 890–18 900页。[5] D. Batra、A. Gokaslan、A. Kembhavi、O. Maksymets、R. Mottaghi、M. Savva、A. Toshev 和 E. Wijmans，“重新审视ObjectNav：关于具身智能体导航至物体的评估”，发表于 *arXiv*预印本 *arXiv:2006.13171*，2020年。[6] D. S. Chaplot、D. P. Gandhi、A. Gupta 和 R. R. Salakhutdinov，“使用目标导向语义探索进行物体目标导航”，发表于 神经信息处理系统进展，第33卷，第4247–4258页，2020年。[7] S. Y. Min、Y.-H. H. Tsai、W. Ding、A. Farhadi、R. Salakhutdinov、Y. Bisk 和 J. Zhang，“通过端到端自监督实现物体目标导航”，发表于 *arXiv*预印本 *arXiv:2212.05923*，2022年。[8] M. Deitke、W. Han、A. Herrasti、A. Kembhavi、E. Kolve、R. Mottaghi、J. Salvador、D. Schwenk、E. VanderBilt、M. Wallingford、L. Weihs、M. Yatskar 和 A. Farhadi，“RoboTHOR：开放的仿真到现实具身人工智能平台”，2020年4月。[9] M. Savva、A. Kadian、O. Maksymets、Y. Zhao、E. Wijmans、B. Jain、J. Straub、J. Liu、V. Koltun、J. Malik、D. Parikh 和 D. Batra，“Habitat：具身人工智能研究平台”，2019年11月。[10] J. Hatori、Y. Kikuchi、S. Kobayashi、K. Takahashi、Y. Tsuboi、Y. Unno、W. Ko 和 J. Tan，“使用不受约束的口语指令交互式拾取真实世界物体”，发表于 *2018 IEEE*国际机器人与自动化会议（*ICRA*）。IEEE，2018年，第3774–3781页。[11] D. Whitney、E. Rosen、J. MacGlashan、L. L. Wong 和 S. Tellex，“通过社会反馈减少物体取送交互中的错误”，发表于 *2017 IEEE*国际机器人与自动化会议（*ICRA*）。IEEE，2017年，第1006–1013页。[12] P. Anderson、A. Shrivastava、J. Truong、A. Majumdar、D. Parikh、D. Batra 和 S. Lee，“视觉与语言导航的仿真到现实迁移”，发表于 机器人学习会议。PMLR，2021年，第671–681页。[13] S. Chen、P.-L. Guhur、M. Tapaswi、C. Schmid 和 I. Laptev，“全局思考，局部行动：用于视觉与语言导航的双尺度图变换器”，发表于 *IEEE/CVF*计算机视觉与模式识别会议论文集，2022年，第16 537–16 547页。[14] A. Majumdar、G. Aggarwal、B. Devnani、J. Hoffman 和 D. Batra，“Zson：使用多模态目标嵌入的零样本物体目标导航”，发表于 *arXiv*预印本 *arXiv:2206.12403*，2022年。[15] S. Y. Gadre、M. Wortsman、G. Ilharco、L. Schmidt 和 S. Song，“Clip on wheels：将零样本目标导航视为目标定位与探索”，发表于 *arXiv*预印本 *arXiv:2203.10421*，2022年。

[16] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark等，《从自然语言监督中学习可迁移视觉模型》，载于《国际机器学习会议》。PMLR，2021年，第8748–8763页。[17] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang, K.-W. Chang和J. Gao，《基于落地的语言图像预训练》，2022年6月。[18] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill等，《基础模型的机遇与风险》，arXiv预印本 arXiv:2108.07258，2021年。[19] R. Mokady, A. Hertz和A. H. Bermano，《Clipcap：用于图像描述的CLIP前缀》，arXiv预印本 arXiv:2111.09734，2021年。[20] R. Dale，《GPT-3：它有什么用？》，《自然语言工程》，第27卷，第1期，第113–118页，2021年。[21] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan和J. Ba，《大语言模型是人类水平的提示工程师》，arXiv预印本 arXiv:2211.01910，2022年。[22] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell等，《语言模型是少样本学习者》，《神经信息处理系统进展》，第33卷，第1877–1901页，2020年。[23] S. Tellex, T. Kollar, S. Dickerson, M. Walter, A. Banerjee, S. Teller和N. Roy，《理解用于机器人导航和移动操作的自然语言命令》，《AAAI人工智能会议论文集》，第25卷，第1期，第1507–1514页，2011年8月。 [在线]。可访问：https://ojs.aaai.org/index.php/AAAI/article/view/7979

[24] J. S. Park, B. Jia, M. Bansal和D. Manocha，《使用动态约束映射，根据基于属性的自然语言指令高效生成运动计划》，载于《2019年国际机器人与自动化会议》（ICRA）。IEEE，2019年，第6964–6971页。[25] Z. Hu, J. Pan, T. Fan, R. Yang和D. Manocha，《复杂场景中遵循人类指令的安全导航》，《IEEE机器人与自动化快报》，第4卷，第2期，第753–760页，2019年。[26] V. S. Dorbala, A. Srinivasan和A. Bera，《机器人能信任你吗？一种基于深度强化学习的方法，实现信任驱动的人类引导导航》，载于《2021年IEEE国际机器人与自动化会议》（ICRA）。IEEE，2021年，第3538–3545页。[27] T. Kollar, S. Tellex, D. Roy和N. Roy，《迈向对自然语言方向指令的理解》，载于《2010年第五届ACM/IEEE人机交互国际会议》（HRI）。IEEE，2010年，第259–266页。[28] J. Thomason, A. Padmakumar, J. Sinapov, N. Walker, Y. Jiang, H. Yedidsion, J. Hart, P. Stone和R. J. Mooney，《通过人机对话改进基于语境的自然语言理解》，载于《2019年国际机器人与自动化会议》（ICRA）。IEEE，2019年，第6934–6941页。[29] J. Thomason, M. Murray, M. Cakmak和L. Zettlemoyer，《视觉与对话导航》，载于《机器人学习会议》。PMLR，2020年，第394–406页。[30] X. Gao, Q. Gao, R. Gong, K. Lin, G. Thattai和G. S. Sukhatme，《DialFRED：用于具身指令跟随的对话驱动智能体》，《IEEE机器人与自动化快报》，第7卷，第4期，第10 049–10 056页，2022年。[31] V. S. Dorbala, G. A. Sigurdsson, J. Thomason, R. Piramuthu 和 G. S. Sukhatme，《CLIP-Nav：利用CLIP实现零样本视觉与语言导航》，载于CoRL 2022语言与机器人研讨会，2022年。[32] B. Yamauchi，《一种用于自主探索的基于边界的探索方法》，载于《1997年IEEE机器人与自动化计算智能国际研讨会论文集：迈向机器人与自动化的新计算原理》（CIRA’97）。IEEE，1997年，第146–151页。[33] O. Mees, A. Emek, J. Vertens和W. Burgard，《通过虚构场景表征学习关系指令对应的物体摆放》，载于《2020年IEEE国际机器人与自动化会议》（ICRA）。IEEE，2020年，第94–100页。[34] O. Mees和W. Burgard，《通过语言落地组合抓取与放置任务》，载于《实验机器人学：第17届国际研讨会》。施普林格，2021年，第491–501页。[35] R. Kartmann, D. Liu和T. Asfour，《基于三维空间物体关系和语言指令的语义场景操控》，载于《2020年IEEE-RAS第二十届国际人形机器人会议》（Humanoids）。IEEE，2021年，第306–313页。[36] A. Boteanu, T. Howard, J. Arkin和H. Kress-Gazit，《复杂自然语言指令的可验证语义落地与执行模型》，载于《2016年IEEE/RSJ智能机器人与系统国际会议》（IROS）。IEEE，2016年，第2649–2654页。[37] T. M. Howard, S. Tellex和N. Roy，《面向移动操作机器人的自然语言规划器接口》，载于《2014年IEEE国际机器人与自动化会议》（ICRA）。IEEE，2014年，第6652–6659页。

[38] A. Boteanu, J. Arkin, S. Patki, T. Howard和H. Kress-Gazit, 《通过基于落地的语言交互修复机器人发起的规格》, arXiv预印本 arXiv:1710.01417, 2017年。 [39] H. Kress-Gazit, M. Lahijanian和V. Raman, 《面向机器人的综合：机器人行为的保证与反馈》, 《控制、机器人与自主系统年度评论》, 第1卷, 236, 2018年, 第211页。 – . [40] V. Raman, C. Lignos, C. Finucane, K. C. Lee, M. P. Marcus和H. Kress-Gazit, 《抱歉, 戴夫, 恐怕我做不到：用自然语言解释无法实现的机器人任务》, 载于《机器人学：科学与系统》, 第2卷, 第1期。Citeseer, 2013年, 第2–1页。 [41] Z. Zhao, W. S. Lee和D. Hsu, 《用于物体放置的自然语言指令的可微解析与视觉落地》, 载于《2023年IEEE国际机器人与自动化会议》（ICRA）。IEEE, 2023年, 第11 546–11 553页。 [42] M. Shridhar, L. Manuelli和D. Fox, 《Cliport：用于机器人操作的“做什么”和“在哪里”通路》, 载于《机器人学习会议》。PMLR, 2022年, 第894–906页。 [43] S. Stepputtis, J. Campbell, M. Phielipp, S. Lee, C. Baral和H. Ben Amor, 《用于机器人操作任务的语言条件模仿学习》, 《神经信息处理系统进展》, 第33卷, 第13 139–13 150页, 2020年。 [44] C. Lynch和P. Sermanet, 《基于非结构化数据的语言条件模仿学习》, arXiv预印本 arXiv:2005.07648, 2020年。 [45] W. Huang, P. Abbeel, D. Pathak和I. Mordatch, 《大语言模型作为零样本规划器：为具身智能体提取可执行知识》, 载于《国际机器学习会议》。PMLR, 2022年, 第9118–9147页。 [46] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman等, 《照我能做的做, 而不是照我说的做：将语言落地于机器人的可供性》, arXiv预印本 arXiv:2204.01691, 2022年。 [47] W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar等, 《内在独白：通过使用语言模型规划实现具身推理》, arXiv预印本 arXiv:2207.05608, 2022年。 [48] D. Shah, B. Osinski, B. Ichter和S. Levine, 《LM-Nav：利用大型预训练语言、视觉与动作模型实现机器人导航》, 2022年7月。 [49] C. Huang, O. Mees, A. Zeng和W. Burgard, 《用于机器人导航的视觉语言地图》, 2022年10月。 [50] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason和A. Garg, 《Progprompt：利用大语言模型生成情境化机器人任务计划》, arXiv预印本 arXiv:2209.11302, 2022 年。 [51] J. Redmon, S. Divvala, R. Girshick 和 A. Farhadi, 《只看一次：统一的实时目标检测》, 载于《IEEE计算机视觉与模式识别会议论文集》, 2016年, 第779–788页。 [52] J. Li, D. Li, C. Xiong和S. Hoi, 《BLIP：用于统一视觉语言理解与生成的自举式语言图像预训练》, 载于《国际机器学习会议》。PMLR, 2022年, 第12 888–12 900页。 [53] M. Labb´e和F. Michaud, 《RTAB-Map：用于大规模、长期在线运行的开源激光雷达与视觉同步定位和建图库》, 《野外机器人学杂志》, 第36卷, 第2期, 第416–446页, 2019年。 [54] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, A. Y. Ng等, 《ROS：一个开源机器人操作系统》, 载于ICRA开源软件研讨会, 第3卷, 第3.2期。日本神户, 2009年, 第5页。 [55] M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen等, 《简单的开放词汇目标检测》, 载于《欧洲计算机视觉会议》。施普林格, 2022年, 第728–755页。 [56] H. Bangalath, M. Maaz, M. U. Khattak, S. H. Khan和F. Shahbaz Khan, 《弥合目标级与图像级表征之间的差距，实现开放词汇检测》, 《神经信息处理系统进展》, 第35卷, 第33 781–33 794页, 2022年。 [57] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo和R. Girshick, 《Detectron2》, https://github.com/facebookresearch/detectron2, 2019年。 [58] P. Anderson, A. Chang, D. S. Chaplot, A. Dosovitskiy, S. Gupta, V. Koltun, J. Kosecka, J. Malik, R. Mottaghi, M. Savva等, 《具身导航智能体的评估》, arXiv预印本 arXiv:1807.06757, 2018年。 [59] X. Zhou, R. Girdhar, A. Joulin, P. Kr¨ahenb¨uhl和I. Misra, 《利用图像级监督检测两万类》, 载于ECCV, 2022年。 [60] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Z hong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang, K.-W. Chang和J. Gao, 《基于落地的语言图像预训练》, 2022年6月。