核心挑战 s in 具身视觉语言规划（扩展摘要)*

Jonathan Francis[1,2†], **Nariaki Kitamura**[3], **Felix Labelle**[2], **XiaopengLu**[2], **Ingrid Navarro**[2], **Jean Oh**[2]

[1]博世人工智能中心[2]卡内基梅隆大学计算机科学学院[3]小松制作所{jmf1, ingridn, jeanoh}@cs.cmu.edu, Nariaki Kitamura@global.komatsu,{flabelle, xiaopen2}@alumni.cmu.edu

arXiv:2304.02738v1 [cs.RO] 5 Apr 2023

摘要

多模态机器学习和人工智能（AI）领域的最新进展催生了计算机视觉、自然语言处理和机器人学交叉领域的挑战性任务。尽管许多方法和先前的综述工作已经描述了这些维度中的一个或两个，但尚未有对所有三个维度的中心进行整体分析。此外，即使考虑了这些主题的组合，更多的重点也放在描述例如当前的架构方法上，而不是也 说明该领域的高级挑战和机遇。在本综述论文中，我们讨论具身视觉语言规划（EVLP）任务，这是一类突出的具身导航和操作问题，它们联合利用计算机视觉和自然语言在物理环境中进行交互。我们提出了一种分类法来统一这些任务，并对用于具身视觉与语言任务的当前和新的算法方法、评价指标、模拟器和数据集进行了深入分析和比较。最后，我们提出了我们认为新的EVLP工作应寻求解决的核心挑战，并倡导能够实现模型泛化并进一步推动现实世界部署的任务构建。

1 引言

随着人工智能（AI）和机器人学领域的最新进展，人们设想智能体能够在共享环境中与人类互动。这类智能体包括任何能够自主决策并采取行动的实体；人们期望它们能够理解这些环境中的语义概念，并利用通过传感器感知到的视觉、触觉、听觉或文本等信息来实现这一点[Wooldridge and Jennings, 1995; Castelfranchi, 1998]。为了开发具备这些感知和推理能力的智能体，具身人工智能（EAI）作为一个领域，已成为研究特定一组.....的热门方向。

围绕处于物理环境中的智能体的人工智能问题：近年来，要求智能体同时运用视觉和语言理解能力的任务，其相关论文和数据集数量显著增加 [Das *et al.*，2018；Gordon *et al.*，2018；Anderson *et al.*，2018；Krantz *etal.*，2020；Thomason *et al.*，2019；Nguyen and Daum´e III，2019；Majumdar *et al.*，2020；Li*et al.*， 2020]。在本文中，我们对这类问题的近期研究进行了综述，并将其称为具身视觉语言规划 （EVLP）任务。在本文中，我们旨在全面概述当前关于具身视觉语言规划问题的研究，讨论其主要挑战和未来方向。我们的主要贡献如下：(i) 我们对具身视觉语言规划领域进行了形式化定义，并提出了一套分类法，该分类法既统一了EVLP中一组相关任务，也为新任务的分类提供了基础；(ii) 我们综述了近期的EVLP 任务，比较了它们的任务属性，重点介绍了这些任务所采用的建模方法，并分析了用于评估这些任务中方法的数据集、模拟器和评价指标；最后，(iii) 我们确定了当前EVLP 系列研究面临的开放性挑战，重点鼓励实现未见过场景下的泛化，并将算法部署到现实世界中。关于更多细节，欢迎读者参阅我们的完整期刊论文 [Francis *et al.*，2022b]。

1.1 问题定义

我们讨论一大类问题，这些问题与具身智能体在物理环境中进行规划决策的能力有关。形式化地，令 S 和 A 分别表示状态和动作的集合； V 和 L 表示智能体可用的视觉和语言输入集合。规划问题由元组 $\Phi = \{S, A, s_{ini}, s_{goal}\}$ 定义，其中 s_{ini}、 $s_{goal} \in S$ 分别表示初始状态和目标状态。规划问题 Φ 的一个解 $\psi \in \Psi_\Phi$是在每个状态中采取的动作序列，从初始状态开始以到达目标状态，$\psi = [s_{ini}, a_0, ..., s_t, a_t, ..., a_T, s_{goal}]$，其中 $t \in T$ 是回合长度 T 中的有限时间步， Ψ_Φ 是 Φ的可能解集合。给定一个特定的 EVLP 问题 Φ，时间步 t 时的状态$s_t \in S$ 可以根据截至当前时间步的视觉和语言输入来定义，即， s_t ={(v_0, l_0)、(v_1, l_1).....、(v_t, l_t).....、(v_T, l_T)}，其中 $v_t \in V$ 和 $l_t \in L$。智能体的目标是最小化一个可接受解 $\psi \in \Psi_\Phi$ 与其预-¯测解 ψ 之间的差异。这一定义广泛概括了以下核心：

*完整版发表于JAIR [Francis等人， 2022]。

†联系作者。

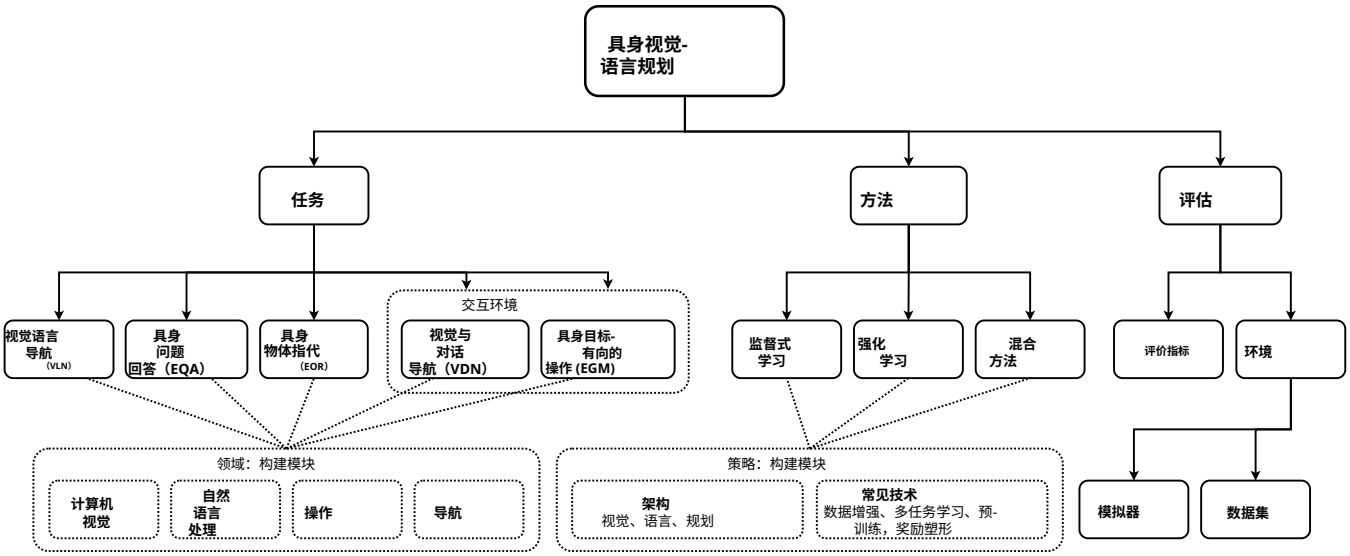


图1：.....的分类法具身视觉语言规划。

具身视觉语言规划问题。需要针对特定任务定制定义，其中添加了额外的约束或假设，以专注于这一通用问题的特定子领域。

1.2 分类法

我们提出了一种EVLP研究的分类法，如图1所示，本文的其余部分围绕该分类法组织。该分类法将领域细分为三个分支：任务、方法和评估方法。任务分支提出了一个框架，用于对现有任务进行分类，并作为区分新任务的基础。方法分支涉及学习范式、用于不同任务的常见架构，以及用于提高性能的常见技巧。分类法最右侧的分支讨论任务评估方法论，该方法论细分为两部分：评价指标和环境。评价指标小节引用了贯穿具身视觉与语言任务的许多常见评价指标，而环境小节介绍了当前使用的不同模拟器和数据集。

2 当前方法

我们简要概述 EVLP 任务系列中使用的任务、方法、学习范式、数据集、模拟器和评价指标；有关更多参考文献、特定任务的问题定义、架构描述与训练目标、数据集和模拟器的比较与统计数据，以及评价指标公式，请参阅 [Francis 等，2022b]。

2.1 任务、方法与学习范式

EVLP 任务。 目前已有许多 EVLP 任务，每项任务都侧重于智能体面临的不同技术挑战和推理要求。任务的差异体现在动作空间（可能执行的动作类型和数量）、所需的推理模式（例如指令遵循，或探索与信息收集），以及任务是否要求与另一智能体交互等方面。视觉与语言导航和视觉与对话导航要求智能体依据自然语言指令，在环境中导航至目标位置；后者还会为智能体提供中间监督和澄清信息。在具身问答任务中，智能体首先会收到一个以语言形式提出的问题，然后必须对环境进行引导式探索，以收集足够的周边信息并生成答案。在具身物体指代任务中，智能体需要导航至给定指令中提及的物体，并在抵达其所在位置后识别（或选择）该物体。具身目标导向操作任务将基于操作的环境交互与前述任务的要求结合起来，例如导航与路径规划、状态跟踪、指令遵循、指令分解和物体选择。

方法与学习范式. 追求解决具身视觉与语言任务的技术方案必须对各个方面进行建模。首先，视觉建模通常涉及构建智能体环境的无损且具有预测性的状态表示；因为自我中心视觉观测会随着智能体在空间中导航或操作物体而变化，智能体还必须包含时序建模机制，以便表示其环境中随时间观察到的状态变化并监控任务执行的进度。接下来，具身视觉与语言任务中的语言建模通常需要使用提供的指令或问题提示来生成智能体目标的丰富描述；因为语言在实践中可能存在歧义，在获得无偏表示方面仍存在挑战。接下来，智能体必须能够将其任务进度与目标表示进行比较，这通常需要用于多模态表示、对齐和融合的复杂策略。最后，为了与其环境交互，智能体必须包含动作生成和规划的机制；受机器人学中经典方法的启发，许多此类方法源于建图和探索策略、搜索和拓扑规划的早期工作，

以及分层任务分解。为了使智能体倾向于期望的任务导向行为，各种方法利用了不同的学习范式（例如，半监督/自监督/全监督学习、强化学习等）和策略（例如，预训练、数据增强、多任务学习、奖励塑造、循环一致性等）。

2.2 数据集、模拟器和评价指标

数据集。 EVLP 数据集在三个主要维度上有所不同：视觉观察、自然语言输入和专家演示。视觉观察通常由 RGB 图像组成，常与深度数据或语义掩码配对。这些观察可以表示室内和室外环境，来自照片级真实或基于合成的设置。相比之下，语言在提示类型上有所不同。语言提示可能以问题、逐步指令或需要通过对话或描述进行某种澄清的模糊指令的形式出现。语言还可以在语言序列的复杂性和词汇范围方面有所不同。最后，导航轨迹在动作空间的粒度（或离散化）以及所提供动作序列与其他两个维度的隐式对齐等方面有所不同。

模拟器。 EAI 研究的早期模拟平台通常利用简单的视频游戏环境来创建和训练神经控制器。在许多这些平台上很快就达到了人类表现，因为简化环境通常缺乏现实世界设置的多样性和复杂性。近期工作通过使用照片级真实感以及智能体能够修改环境中物体状态的交互式上下文，解决了这种缺乏真实感的问题。为此，人们还对开发专注于模拟到现实迁移和评估的框架感兴趣，从而研究真实设置与模拟设置之间的差异。最后，其他平台还专注于鼓励工作的可重复性、设计的灵活性和基准测试。

评价指标。 EVLP研究中的常用评价指标可以分为多个类别——每一类都衡量智能体性能的不同方面，例如距离（量化智能体穿越空间的方式）、成功度（描述智能体完成整体任务的程度）、路径-路径相似度（评估智能体轨迹与真实轨迹的相似程度）、基于指令的评价指标（衡量自然语言指令与智能体轨迹之间的对齐程度），以及以对象为中心的评价指标（评估对象选择、识别或操纵的有效性）。我们在图 2 中展示了前三类。

3 核心挑战

3.1 EVLP研究的新方向

我们着重指出三个有前景的方向，以推动人机交互更加普及，并提升智能体的泛化能力。首先，我们主张改善社会互动：我们认为，从静态指令转向主动对话，将有助于催生新的协作与辅助能力。其次，为了使智能体能够应对现实世界部署中的更多复杂情况，我们主张在 EVLP 研究中引入动态环境，鼓励智能体采用能够适应环境不确定性和非平稳性的推理策略。我们探讨了跨任务机器人学习的愿景：智能体可先从相关的模态中心任务中积累经验，再部署到任务重叠度较高的共享多模态设置中。最后，我们还将指出迁移学习中交互式物体感知的新方向；智能体必须与环境进行物理交互，才能学习新概念[Tatiya等人，2023b；Tatiya 等人，2023a]。

3.2 领域知识的使用

我们进一步鼓励开发以严谨方式利用领域知识的方法，以指导模型的学习和迁移；尽管这一理念近期在其他领域重新兴起 [Francis, 2022; Park 等, 2020; Francis 等, 2022a;Herman 等 , 2021;Andreas 等 , 2016; Francis等，2019], 但我们注意到，EVLP 领域中此类研究寥寥无几。领域知识有多种形式，例如图模型、逻辑规则、约束、预训练、知识图谱等；虽然领域知识有望提升智能体的样本效率、可解释性、安全性和泛化性，但如何在任意学习问题中有效地表达和利用领域知识仍是一项挑战。预训练 和常识知识尤其是两种有望赋予智能体上述特性的体现。

预训练任务经过精心设计，并与广受欢迎的高容量模型相结合，在图像分类等领域开展自监督 [He 等人，以及自然语言处理领域 [Devlin 等人，2019；Yang等人，2019b；Ma 等人， 2021], 旨在最大限度地提升迁移或微调方法的泛化性。尽管在特定多模态问题方面已取得一些进展 [Majumdar 等人，2020；Hao 等人，2020；Lu 等人， 2019], 但要开发能够涵盖更广泛 EVLP 任务系列的通用预训练策略，仍面临诸多挑战。

在自然语言处理领域，常识知识的获取与注入模型仍是一个活跃的研究方向 [Talmor 等人，2019；Ma 等人，2019；Ma等人，2021；Li 等人， 2021], 一些研究提出直接利用结构化常识知识库为观测提供依据，从而提升模型在相关任务上的下游表现。然而，在 EVLP 任务背景下，常识知识的应用仍鲜有探索。EVLP 任务的最终目标是开发能够在真实环境中解决现实问题的智能体，因此，可以考虑为模型提供关于世界的结构化外部知识 [Yang 等人，2019a；Tatiya等人， 2022]。

3.3 智能体训练目标

针对智能体执行给定任务选择适当的训练目标（一个或多个），一直是机器学习和人工智能领域的长期问题；这种选择取决于可用训练信号的性质（例如，

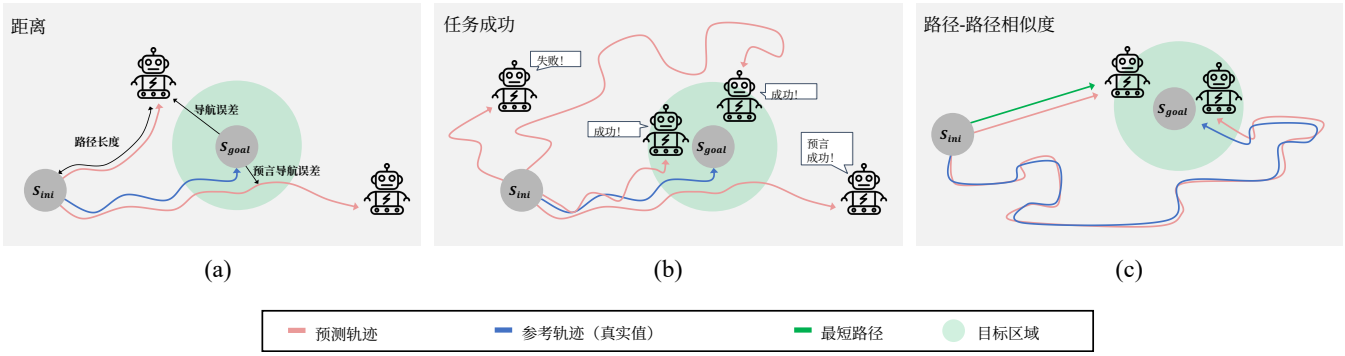


图 2：在具身视觉语言规划任务中使用的典型评价指标所衡量的现象 .

奖励或成本函数、监督水平、环境可观测性）以及外部知识（例如，辅助目标、常识、约束）被认为对有效偏置智能体行为必要的程度。对于具身视觉与语言任务，训练目标的选择因智能体操作环境的复杂性质而更具挑战性。这通常需要由多种偏置策略组成的框架。鉴于优化泛化性和可解释性的根本动机，应明确处理寻找最有效地整合各种相关来源信息并将智能体的归纳偏置泛化到新环境的学习范式；实际上，训练范式应包括鼓励我们希望赋予的属性的明确机制。

3.4 仿真到现实的差距

数据集和模拟环境是 EVLP 研究背后的主要驱动力，因为模型效能的衡量依赖于强大的测试场景和适当的评价指标。当前的 EVLP 任务以一组目标和评价指标的形式，构建于现有模拟器或数据集之上。在本节中，我们敦促社区考虑并优先将 EVLP 智能体部署到现实世界环境中。具体而言，我们认为，各种 EVLP 任务和评价指标可以基于三个维度加以改进：模拟器真实度、数据集真实度和模型泛化能力测试。

基于模拟的训练与执行在对序列学习问题进行建模时尤为有吸引力，因为离线数据集无法实现与环境之间这种递归式交互。然而，科学家和从业者在促使模型形成适用于现实应用场景的预期行为方面，能力仍然有限。由于这种差异降低了模型在现实世界部署中的即时可行性，我们强调，计算机视觉和机器人学领域应更加关注模拟到现实的迁移、对未见情形的泛化，以及对分布外环境的鲁棒性。我们鼓励制定评价指标，以评估智能体的中间行为和任务效率，而非仅仅表明其是否完成了域内任务。

基于数据集的训练可以非常有效，例如，当为模型提供关于智能体行为的强先验时。Re-

然而，仅依赖数据集进行训练可能会带来重大问题，尤其是因果混淆、对环境转移动态的有限观察，以及由于类别不平衡导致的不现实先验。此外，基于数据集定义的任务评价指标在真正评估智能体是否准备好部署到现实世界方面的能力可能有所不同。尽管存在这些挑战，精心构建的数据集可以在鼓励模型学习专业技能方面发挥重要作用；使智能体能够在多个环境中训练的数据集可以导致更可泛化的行为。

我们还将讨论在现实世界部署中评估 EVLP 智能体所面临的问题，并提出新的评估方法以应对这些挑战。我们建议，当前的评估范式应明确测试智能体跨领域和跨任务的泛化性，因为模型需要能够迁移到未见环境，并能抵御模型和环境的不确定性。我们还指出，需要建立包含不同（子）领域划分的测试平台，因为当前的数据划分可能无法代表预期现实场景中的变化。总而言之，我们旨在强调全面评估范式对于 EVLP 智能体的重要性，并提出新的方法来评估其在现实世界部署中的泛化性和鲁棒性。

4 结论

在这篇扩展摘要中，我们提出了一个关于具身视觉语言规划（EVLP）领域的分类法，重点涵盖任务、建模方法、学习范式和评估设置；我们还根据解决各项任务所需的技能，构建了一个用于讨论现有及未来任务的框架。我们简要介绍了各种学习范式、训练策略和常用的优化目标，并探讨了不同形式的评估，例如使用专家演示数据集、模拟器和各种任务评价指标。最后，我们聚焦于该领域目前正在应对的挑战，以及仍有待解决的问题。具体而言，我们讨论了可能阻碍现实世界部署的问题，例如泛化能力不足、鲁棒性欠缺、模拟器的真实性不足以及交互不够丰富；我们指出，这些问题是未来最有前景、最值得投入的研究方向。

致谢

作者感谢 Alessandro Oltramari、Yonatan Bisk、Eric Nyberg 和 Louis-Philippe Morency 就原期刊论文展开富有见地的讨论；作者还感谢 GyanTatiya、Jimin Sun、Prasoon Varshney、Sahiti Yerramilli 和 Jayant Tamarapalli 在 EVLP 方面开展新的／近期的合作。

参考文献

[Anderson*et al.*，2018] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko S¨ underhauf, Ian D. Reid, Stephen Gould, and Anton van den Hengel. 视觉与语言导航：在真实环境中解释视觉接地的导航指令。In*2018*年*IEEE*计算机视觉与模式识别会议，*CVPR 2018*，美国犹他州盐湖城，*2018*年*6*月*18-22*日，第3674–3683页。IEEE计算机学会，2018。

[Andreas*et al.*，2016] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 神经模块网络。In *2016*年*IEEE*计算机视觉与模式识别会议，*CVPR 2016*，美国内华达州拉斯维加斯，*2016*年*6*月*27-30*日，第39–48页。IEEE计算机学会，2016。

[Castelfranchi，1998] CristianoCastelfranchi。为人工智能智能体建模社会行动。人工智能，103(1- 2):157–182，1998。

[Das 等， 2018] Abhishek Das、Samyak Datta、Georgia Gkioxari、StefanLee、Devi Parikh 和 DhruvBatra。具身问答。载于*2018IEEE*计算机视觉与模式识别会议，*CVPR 2018*，美国犹他州盐湖城，*2018*年*6*月*18*日至*22*日，第1–10页。IEEE计算机学会，2018年。

[德夫林等， 2019] Jacob Devlin、Ming-Wei Chang、Ken- ton Lee 和Kristina Toutanova。BERT：用于语言理解的深度双向变换器预训练。在《*2019*年北美计算语言学协会分会会议论文集：人类语言技术，第 *1* 卷（长篇与短篇论文）》，第 4171–4186 页，美国明尼阿波利斯，2019 年。计算语言学协会。

[Francis 等人， 2019] Jonathan Francis、Matias Quintana、Nadine von Frankenberg、Sirajum Munir 和 Mario Berges。Occutherm：利用体形信息推断居住者的热舒适度。载于第六届国际节能建筑、城市与交通系统会议论文集，BuildSys '19，美国纽约州纽约市，2019 年。ACM。

[Francis 等，2022a] Jonathan Francis、Bingqing Chen、Siddha Ganju、Sidharth Kathpal、Jyotish Poonganam、Ayush Shivani、Vrushank Vyas、Sahika Genc、Ivan Zhukov、Max Kumskoy、Jean Oh、Eric Nyberg 和 Sylvia L. Herbert。《Learn-to-race 挑战赛 2022：自主赛车中安全学习与跨领域泛化的基准测试》。发表于第 *1* 届 *ICML* 自动驾驶安全学习研讨会，2022 年。

[Francis 等，2022b] Jonathan Francis、Nariaki Kitamura、Felix Labelle、Xiaopeng Lu、Ingrid Navarro 和 Jean Oh。具身视觉语言规划的核心挑战。《人工智能研究杂志》，74：459–515，2022 年。[Francis， 2022] Jonathan Francis。《面向多视图上下文理解的知识增强表示学习》。博士论文，卡内基梅隆大学，2022 年。

[Gordon 等人， 2018] Daniel Gordon、Aniruddha Kembhavi、Mohammad Rastegari、Joseph Redmon、Dieter Fox 和 Ali Farhadi。IQA：交互式环境中的视觉问答。载于*2018* 年 *IEEE* 计算机视觉与模式识别会议，*CVPR 2018*，美国犹他州盐湖城，*2018* 年 *6* 月 *18—22* 日，第 4089–4098 页。IEEE计算机学会，2018 年。

[Hao 等人， 2020] Weituo Hao、Chunyuan Li、Xiujun Li、Lawrence Carin 和 Jianfeng Gao。通过预训练学习通用视觉与语言导航智能体。载于*2020* 年 *IEEE/CVF* 计算机视觉与模式识别会议，*CVPR 2020*，美国华盛顿州西雅图，*2020* 年 *6* 月 *13—19* 日，第 13134–13143 页。IEEE，2020 年。

[He等人， 2016] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。用于图像识别的深度残差学习。载于*2016*年*IEEE*计算机视觉与模式识别会议（*CVPR 2016*），美国内华达州拉斯维加斯，*2016*年*6*月*27*日至*30*日，第770–778页。IEEE计算机学会，2016年。

[Herman等人， 2021] James Herman、Jonathan Francis、Siddha Ganju、Bingqing Chen、Anirudh Koul、Abhinav Gupta、Alexey Skabelkin、Ivan Zhukov、Max Kumskoy 和 Eric Nyberg。Learn-to-race：用于自主赛车的多模态控制环境。载于*IEEE/CVF*国际计算机视觉会议论文集，第9793–9802页，2021年。

[Krantz等人， 2020] Jacob Krantz、Erik Wijmans、Arjun Majumdar、Dhruv Batra 和 Stefan Lee。超越导航图：连续环境中的视觉与语言导航。*CoRR*，abs/2004.02857，2020年。

[Li 等， 2020] Dongxu Li、Cristian Rodriguez Opazo、Xin Yu 和 Hongdong Li。基于视频的词级深度手语识别：一个新的大规模数据集与方法比较。载于 *IEEE* 冬季计算机视觉应用会议，*WACV 2020*，美国科罗拉多州斯诺马斯村，*2020* 年 *3* 月 *1–5* 日，第 1448–1458 页。IEEE，2020 年。

[Li 等人, 2021] Yikang Li, Pulkit Goel, Varsha Kuppur Rajendra, Har Simrat Singh, Jonathan Francis, Kaixin Ma, Eric Nyberg, 和 Alessandro Oltramari. 通过常识知识提取与注入的词汇约束文本生成. 在 第*35*届*AAAI*人工智能会议上的常识知识图谱 *(CSKGs@AAAI-21)*, 2021.

[Lu等人，2019] Jiasen Lu、Dhruv Batra、Devi Parikh 和 Stefan Lee。Vilbert：为视觉与语言任务预训练任务无关的视觉语言表征。载于 Hanna M. Wallach、Hugo Larochelle、Alina Beygelzi mer,

Florence d’Alch´e-Buc、Emily B. Fox 和 Roman Gar- nett 编，《神经信息处理系统进展 *32*：*2019* 年神经信息处理系统年会（*NeurIPS 2019*），*2019* 年 *12* 月 *8* 日至 *14* 日，加拿大不列颠哥伦比亚省温哥华》，第 13–23 页，2019 年。[Ma等人，2019] Kaixin Ma、Jonathan Francis、Quanyang Lu、Eric Nyberg 和 Alessandro Oltramari。面向常识问答的可泛化神经符号系统。载于自然语言处理常识推理第一届研讨会论文集，第 22–32 页，中国香港，2019 年。计算语言学协会。[Ma等人，2021] Kaixin Ma、Filip Ilievski、Jonathan Fran- cis、Yonatan Bisk、Eric Nyberg 和 Alessandro Oltra mari。面向常识问答零样本评估的知识驱动数据构建。载于第 *35* 届 *AAAI* 人工智能会议论文集（*AAAI-21*），2021 年。[Majumdar等人，2020] Arjun Majumdar、Ayush Shrivas- tava、Stefan Lee、Peter Anderson、Devi Parikh 和 Dhruv Batra。利用来自网络的图像—文本对改进视觉与语言导航。*CoRR*，abs/2004.14973，2020 年。[Nguyen 和 Daum´e III，2019] Khanh Nguyen 和 Hal Daum´e III。救命，安娜！通过鼓励回顾性好奇心的模仿学习，在自然多模态辅助下进行视觉导航。载于*2019* 年自然语言处理经验方法会议暨第九届自然语言处理国际联合会议论文集（*EMNLP-IJCNLP*），第 684–695 页，中国香港，2019 年。计算语言学协会。[Park等人，2020] Seong Hyeon Park、Gyubok Lee、Manoj Bhat、Jimin Seo、Minseok Kang、Jonathan Francis、Ash- win R Jadhav、Paul Pu Liang 和 Louis-Philippe Morency。通过多模态上下文理解进行多样且合理的轨迹预测。载于欧洲计算机视觉会议（*ECCV*），2020 年。[Talmor等人，2019] Alon Talmor、Jonathan Herzig、Nicholas Lourie 和 Jonathan Berant。CommonsenseQA：一项针对常识知识的问答挑战。载于*2019* 年北美计算语言学协会分会年会：人类语言技术论文集，第 *1* 卷（长篇与短篇论文），第 4149–4158 页，美国明尼苏达州明尼阿波利斯，2019 年。计算语言学协会。[Tatiya等人，2022] Gyan Tatiya、Jonathan Francis、Luca Bondi、Ingrid Navarro、Eric Nyberg、Jivko Sinapov 和 Jean Oh。用于语义音视频具身导航的知识驱动场景先验。*arXiv* 预印本 *arXiv:2212.11345*，2022 年。[Tatiya等人，2023a] Gyan Tatiya、Jonathan Francis 和 Jivko Sinapov。用于落地物体识别的跨工具与跨行为感知知识迁移。*arXiv* 预印本 *arXiv:2303.04023*，2023 年。

[Tatiya等，2023b] Gyan Tatiya、Jonathan Francis 和 Jivko Sinapov。跨异构机器人形态迁移非视觉物体属性的隐式知识。载于国际机器人与自动化会议（*ICRA*）论文集，2023 年。[Thomason 等，2019] Jesse Thomason、Michael Murray、Maya Cakmak 和 Luke Zettlemoyer。视觉与对话导航。*CoRR*，abs/1907.04957，2019 年。[Wooldridge 和 Jennings，1995] Michael Wooldridge 和 Nicholas R Jennings。智能体：理论与实践。知识工程评论，10(2):115–152，1995 年。[Yang 等，2019a] Wei Yang、Xiaolong Wang、Ali Farhadi、Abhinav Gupta 和 Roozbeh Mottaghi。使用场景先验的视觉语义导航。载于第七届国际学习表征会议，*ICLR 2019*，*2019* 年 *5* 月 *6—9* 日，美国路易斯安那州新奥尔良。OpenReview.net，2019 年。[Yang 等，2019b] Zhilin Yang、Zihang Dai、Yiming Yang、Jaime G. Carbonell、Ruslan Salakhutdinov 和 Quoc V. Le。XLNet：用于语言理解的广义自回归预训练。载于 Hanna M. Wallach、Hugo Larochelle、Alina Beygelzimer、Florence d’Alch´e-Buc、Emily B. Fox 和 Roman Garnett 编，《style id='28'>神经信息处理系统进展 32：2019 年神经信息处理系统年会，NeurIPS 2019，2019 年 12 月 8—14 日，加拿大不列颠哥伦比亚省温哥华》，第 5754–5764 页，2019 年。