arXiv:2602.01930v1 [cs.RO] 2 Feb 2026

LIER示例：用于机器人探索的语言-图像嵌入 n

Felix Igelbrink[1*†]、Lennart Niecksch[2, 1†]、MarianRenz[2,1], Martin Günther[1], Martin Atzmueller[2,1]

[1]德国人工智能研究中心（DFKI），合作与自主系统研究部（CAS），德国奥斯纳布吕克。[2]奥斯纳布吕克大学，语义信息系统组，德国奥斯纳布吕克。

通讯作者。电子邮箱：felix.igelbrink@dfki.de；贡献作者：lennart.niecksch@dfki.de；marian.renz@dfki.de；martin.guenther@dfki.de；martin.atzmueller@uos.de；

†这些作者对本研究的贡献相同。

摘要

语义地图使机器人能够推理周围环境，从而完成在已知环境中导航、寻找特定物体以及探索未绘制地图区域等任务。传统的地图构建方法能够提供准确的几何表示，但往往受限于预先设计的符号词汇表。由于依赖固定的物体类别，这些方法难以处理设计时未定义的分布外知识。近年来，CLIP 等视觉语言基础模型取得了进展，使开放集地图构建成为可能：物体不再以固定标签表示，而是编码为高维嵌入。在 LIEREx 中，我们将这些视觉语言基础模型与成熟的三维语义场景图相结合，使自主智能体能够在部分未知环境中进行目标导向探索。

K关键词： 三维语义场景图、视觉语言模型、开放集语义地图构建、主动
P感知、机器人探索

1 引言

在部分或完全未知的环境中运行的自主移动机器人智能体，需要具备高度的场景理解能力，以确保运行有效且安全。这项基础能力主要依赖于利用传感器数据构建语义地图，将环境的几何信息与其中物体的详细信息（例如其类别和属性）结合起来 [1]。传统语义建图系统主要依赖僵化、封闭的符号词汇表，通常由一组固定的物体类别定义。这一限制降低了系统的灵活性和可扩展性，使其难以处理设计阶段未明确预定义的通用知识或概念。

随着视觉语言基础模型（VLFMs）的发展，一项重要的方法学进展应运而生，使开放词汇与多模态视觉数据得以有效结合。对比语言-图像预训练（CLIP）模型 [2] 是此类视觉语言基础模型的典型代表。这些模型在海量图文对数据集上进行训练，从而学习丰富的多模态特征，并将视觉概念与自然语言

本预印本尚未经过同行评审，也未在提交后进行任何改进或修正。本文的正式出版版本发表于《KI - Künstliche Intelligenz》，可在线查阅：https://doi.org/10.1007/s13218-026-00902-6

映射到联合特征空间中。因此，这种能力实现了开放集语义地图构建，其中物体不再通过固定标签表示，而是通过高维特征嵌入表示。

将这些视觉语言基础模型整合到语义建图架构中，可引入先进的查询能力，显著突破传统受限于固定词汇表的方法所存在的局限。这为目标导向探索和持续监视等新应用提供了契机，尤其适用于动态变化或仅部分探索过的环境。借助开放集语义查询，机器人智能体能够利用自然语言解读并搜索任意物体和抽象概念。

在 LIEREx[1] 项目（面向机器人探索的语言-图像嵌入）中，我们正研究如何将这些视觉语言基础模型的进展与混合机器学习方法相结合，以推动移动机器人学中的语义建图和自主探索。LIEREx 利用开放集地图表示的优势，同时结合当前流行的基于三维语义场景图（3DSSG）的语义方法所具备的空间推理能力。具体而言，该项目着重于动态生成探索策略，并借助基于神经网络的估计系统推导出最优观测位姿。这种方法有助于高效验证搜索和探索结果，使系统不再仅仅依赖语言查询本身。

这项工作与 ExPrIS 项目（将知识层面的预期作为先验，用于从传感器数据中解读物体）密切相关，双方共享基础设施，并关键性地采用 3DSSG 作为基础环境表示。

本报告的其余部分安排如下：第2 节介绍作为架构基础的开放集地图构建方法，随后在第3节介绍探索规划系统。第4 节介绍机器人演示系统。最后，我们在第5节讨论关键技术洞见并作出总结。

[1] https://www.dfki.de/en/web/research/ 项目与出版物/项目/lierex

2 开放集语义建图

LIEREx 和 ExPrIS 都致力于构建一个共同的统一语义框架。ExPrIS 专注于推导结构化表示，以便与现有的结构化知识进行集成；而 LIEREx 则通过引入现代视觉语言模型，扩展并泛化了这一目标。这使系统能够利用语义地图中的语言概念所蕴含的常识知识。具体而言，LIEREx 旨在增强物体检索能力：它不仅提供与查询对应的位置或几何信息，还会直接生成适合的观测位姿，使机器人更有可能在这些位置观测到所查询的物体。随后，机器人智能体可以利用这些位姿来搜索所请求的物体或位置。LIEREx 的基本流程如图1所示。两个项目都以 3DSSG 作为基础表示，并被设计为可互操作，将符号知识（ExPrIS）与多模态开放集感知（LIEREx）相结合。

我们的 3DSSG 表示结构受到机器人学领域其他主流方法的启发 [3, 4]。它以异构图的形式实现，并组织为包含多个层次的动态层级结构。这些层次表示不同级别的语义概念，范围从低级概念（如单个物体）到高级概念（如房间）。LIEREx 的一项核心贡献是扩展这一 3DSSG 结构，以纳入从视觉语言基础模型（VLFM）推断出的视觉语言特征。

这些模型，尤其是开创性的 CLIP 模型 [2]，，能够学习丰富的多模态特征空间，将自然语言与视觉概念相结合。最初的 CLIP 模型在零样本图像分类任务中已展现出卓越性能，甚至媲美在基准数据集上训练的监督模型 [5, 6]。再加上模型相对简单，且易于集成到下游任务中，视觉语言基础模型在许多计算机视觉任务中的受欢迎程度日益提升。最近，大语言模型（LLMs）也成为这些应用中一种前景广阔的辅助工具 [7]。借助视觉语言基础模型的集成，新一代大型多模态模型（LMMs）也能够直接处理来自图像或视频的视觉输入 [8]。

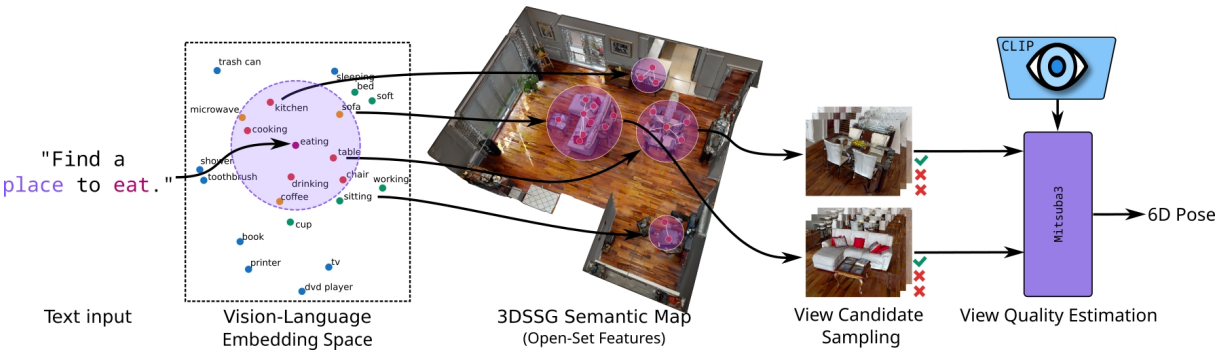


图 1：LIEREx 流程概览。文本查询会从 3DSSG 中检索最匹配的实例，同时考虑图所提供的场景上下文，以及它们在视觉语言模型（VLM）嵌入空间中的接近程度。随后，从 3DSSG 中采样候选视点位姿，并由估计模块根据它们与查询的语义相似度进行评分。

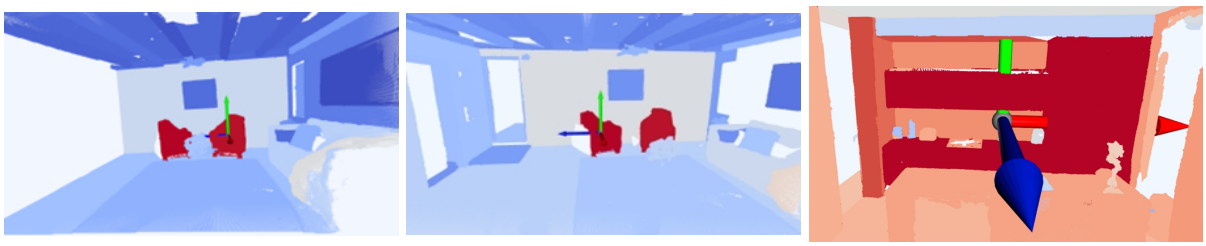
然而，CLIP 及类似模型（例如 SigLIP [9]，BLIP [10]，和 BLIPv2 [11] ）只能为整张图像生成特征向量。因此，它们无法直接用于语义建图所需的语义分割和实例分割任务，因为这些任务需要精细的特征来定位图像中检测到的物体。基于像素和区域的方法 [12–15] 通过对每个像素或较大区域执行二维开放词汇语义分割，弥补了这一不足。这些方法以基线视觉语言基础模型和近期的实例无关分割模型 [16]为基础，学习为像素或区域分别生成特征向量，但其运行时间显著长于基于图像的视觉语言基础模型和传统模型 [17]。

与传统分割模型不同，这些方法不会生成与特定标签关联的分割掩码，除非在运行时提出请求。相反，它们会生成高维特征图，将语义信息直接编码到视觉表示中。这使映射过程更加灵活：视觉特征保留在潜在空间中，并可灵活检索。

在 LIEREx 中，我们扩展了混合式 3DSSG 结构，加入额外的 CLIP特征向量，以支持开放集语义查询，并利用图结构提供的空间推理能力进行语言驱动的推理。我们没有依赖将 CLIP特征投影到稠密像素或 3D 点上的蒸馏模型，而是采用与其他流行方法一致的两步法来推断特征向量：

首先，从 [18–21]中使用类别无关的分割模型 [16, 22, 23]对传入的 RGB-D帧分割出潜在物体候选项 [20]。随后，利用获得的掩码推断各个 CLIP特征向量，并将其系统地整合到相应的 3DSSG 节点中，优先选择感兴趣物体清晰可见的视角 [20]。最终的 3DSSG 可通过任意文本进行查询：将查询特征向量与各节点进行余弦相似度比较，并返回所有匹配的物体（见图2）。

此外，我们将环境的空间结构与这些开放集语义能力相结合。不同于现有方法通常将场景图视为扁平的物体集合或严格的层级结构，我们的方法旨在自动推断室内环境固有的层级关系。通过将视觉-语言模型中的多尺度特征图与 3DSSG 结构对齐 [18, 21], ，我们确保语义嵌入不仅以视觉外观为基础，也融入其空间上下文；同时，低层级物体能够动态聚类为更高层级的物体。这种层级架构超越了静态分类，能够支持针对抽象查询的复杂推理，而这类查询通常超出标准视觉语言基础模型的能力范围。



(a) (b) (c)

图 2：VL地图中的示例查询。（a）和（b）展示了文本查询椅子的前两个结果，其中包含周围环境的裁剪区域。（c）展示了对高阶概念小厨房的查询，该概念由多个物体实例组成。

3 探索规划系统

在开放词汇探索领域，近期方法主要侧重于利用视觉语言基础模型或视觉语言模型提供的语义引导，增强经典的基于前沿的策略。这些方法通常保留探索已知空间与未知空间边界的基本逻辑，但会根据语义相关性对这些前沿进行加权。

在 CoW [24]，中，作者通过整合基于CLIP的相关性地图来增强标准的前沿选择，从而优先考虑感兴趣区域。类似地，GOAT [25] 采用基于图像的物体实例记忆，并结合全局二维语义地图，朝目标物体位置导航。为进一步引导这一过程，Ren等人 [26] 等方法将当前RGB帧投影到三维体素地图中，并利用视觉语言模型根据识别出的自由空间对潜在探索方向进行评分。

这些工作的多数都存在一个显著局限：依赖二维地图表示或二维投影进行规划。Laina等人 [19]，的工作是一个显著例外，它直接在三维空间中运行，利用基于CLIP的余弦相似度调节前沿附近的采样。然而，为了估算候选位姿的信息增益，该方法依赖于对TSDF子地图进行光线投射——这一过程计算开销大，且难以在资源受限平台上扩展。

在 LIEREx 中，我们通过摒弃对候选视角进行纯几何评估来解决这些局限。我们不再执行成本高昂的在线光线投射来评估每个潜在观测位姿，而是提出了一种学习式视角质量估计系统。通过训练模型，根据部分建图的物体及其上下文，预测视角相对于语义查询的质量，我们可以直接利用 3DSSG 结构，高效预测候选位姿的质量分数。这使智能体能够以显著降低的计算开销验证搜索结果，并探索可能存在目标物体的位置。我们将这种针对性的视角估计与语义引导的基于前沿的探索策略相结合：系统优先采用学习得到的视角提议，以定位已知或假设存在的语义目标；同时，在探索环境中未知区域时，则改用加权前沿探索。

3.1 视角质量估计

为了实现物体目标导航，必须在地图的坐标系中提供观测位姿和轨迹。传统上，这通常通过启发式采样（例如 CAOS [27] 的 FLAP）或手工设计的代价函数（例如 OK-Robot [28]）来完成。

对所有位姿进行光线投射和渲染的计算成本很高（例如参见 Linok等人 [21]），因此在资源受限的系统上往往难以实现。数据驱动的方法效率可能高得多，但需要适当的训练数据。

据我们所知，目前没有可用于杂乱场景中视图位姿估计任务的公开数据集。一个见解是，真实标注的最佳观测位姿往往很难定义。

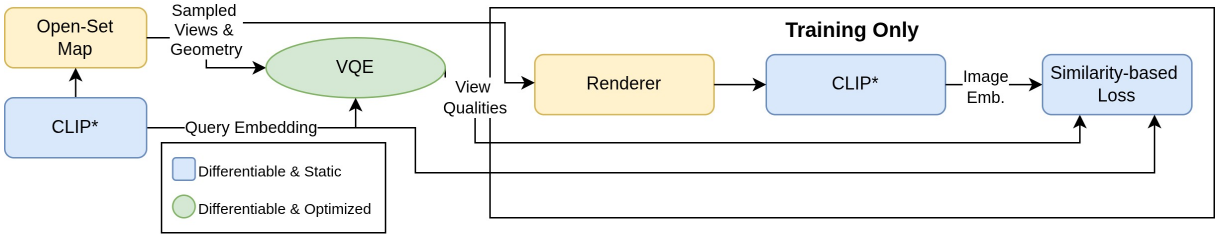


图 3：视角质量估计（VQE）模块的自监督训练流程。该模型通过将渲染视图的 CLIP嵌入与查询嵌入进行比较，学习预测质量分数。

在LIEREx中，我们转而提出一种基于视图质量评估的半监督方法。受先前抓取相关工作的启发，[29, 30]，我们提出一种方法，用于估计潜在目标周围均匀采样位姿的视图质量，而不是直接回归位姿。该方法并不关注单个实例的可见性，而是在训练期间渲染位姿，并将渲染结果所推断的CLIP嵌入与查询的嵌入进行比较（见图3）。其核心思想是，模型应学会偏好那些能在CLIP特征空间中清晰区分所查询物体或概念的观测位姿。

3.2 数据生成

为支持视角质量估计模型所需的大规模训练和评估，我们采用 Habitat模拟器 [31]。结合 Matterport3D [32] 和 Matterport-Habitat (HM3D) [33]数据集，该模拟器提供了大量多样且逼真的室内环境，这些环境基于高分辨率三维摄影测量构建。

与Gazebo等传统机器人仿真环境不同，Habitat经过优化，可在无需模拟完整机器人动力学开销的情况下高效测试感知与规划算法。该设置支持快速迭代，并为我们视点位姿模型的自监督训练提供大规模数据生成。我们利用HM3D数据集，并通过视觉语言建图流程，基于真实轨迹和语义标注生成语义地图。

3.3 训练

训练期间，从地图集合和预期词汇表中随机采样查询-地图对，并基于CLIP嵌入的余弦相似度提取top-k区域（见图2）。随后，使用数据集中的真实网格渲染所评估位姿处的视图。渲染结果的CLIP特征使用与VL建图流程中相同的冻结CLIP模型计算。最后，计算渲染结果与查询之间的一种余弦损失变体，并反向传播到模型。图3 给出了所提出数据流的高层概览。

4 室内演示平台

由于我们的重点是开发时空语义环境表示，而不是构建完整的同步定位与建图系统，因此我们在TIAGo[2] 平台 [34]上集成了MICP-L方法。这使我们能够利用高分辨率的预先录制的地图（例如来自地面激光雷达或CAD模型的地图），在无需外部跟踪系统或运行完整同步定位与建图系统的情况下，获得类似真实值的位姿估计。这使我们能够将我们的方法从仿真环境受控地迁移到真实世界环境中，并能够针对含噪传感器数据和感知数据对系统进行评估。该机器人配备了用于定位的Ouster OS0激光雷达和Femto Bolt飞行时间RGB-D相机。我们通过使用ICP配准将相机点云与三维激光雷达参考坐标系对齐，完成了外参标定。

[2]https://pal-robotics.com/es/robot/tiago/

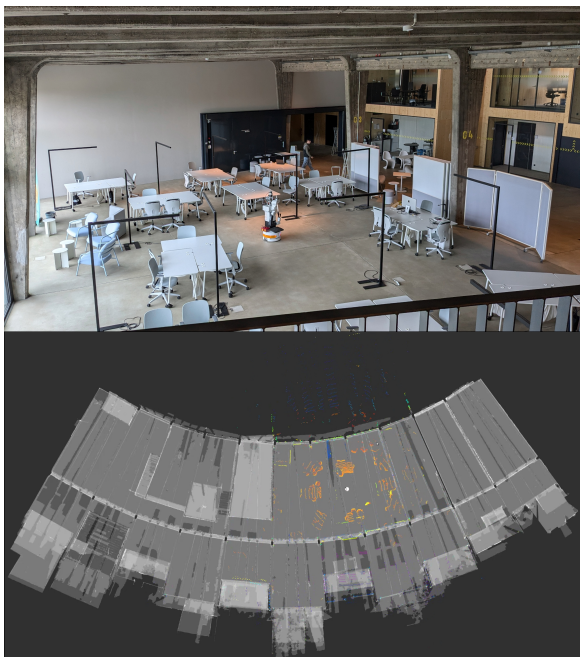


图 4：TIAGo机器人在大型环境中的定位。下图展示了该建筑预先记录的多边形地图（不含家具），并叠加了配准过程中未与地图几何信息关联的激光雷达点。

此外，将测量结果与已知的地图几何相关联，可以有效过滤常见的全景类别（例如墙壁、天花板）；这些类别不一定需要在语义地图中作为独立实例进行跟踪。这样可以从地图中删去不必要的实例并降低不确定性，从而提高整体效率（见图4）。

该机器人平台将在未来工作中用于评估完整的建图与探索流水线，包括 3DSSG 构建、视角质量估计、语义查询和空间推理。

5 讨论与展望

LIEREx 的开发带来了若干重要见解，涉及将视觉语言基础模型集成到三维探索中。

我们发现，生成良好的 观测位姿很少仅取决于几何信息。简单的几何启发式（例如视角或到质心的距离）在杂乱场景中会失效，因为遮挡和语义朝向（例如物体的正面）至关重要。因此，仅凭二维地图表示是不够的；要有效评估可见性和语义相关性，密集的三维信息必不可少。

此外，我们的初步实验表明，直接回归最优视点位姿（例如通过可微渲染）并不实用。优化空间的非凸特性常常导致系统收敛于局部极小值，此时视点在几何上有效，但在语义上毫无意义。因此，我们转而采用视角质量估计 方法，由系统对采样候选视点的质量分数进行回归，而不是直接优化位姿参数。

最后，我们指出基于实例的地图表示与统一地图表示之间存在显著张力。尽管我们的 3DSSG 有助于复杂的符号推理和开放集查询，但在传感器噪声下为各个实例保持几何一致性，要比在统一的基于体素的表示（例如截断符号距离函数）中困难得多。虽然统一地图在光线投射和可见性检查方面效率更高，但它们往往缺乏图结构所提供的高层语义操作所需的灵活性。

在未来工作中，我们旨在通过将学习到的视角质量估计集成到混合式、语义引导的基于前沿的探索规划器中，来应对这些权衡。该完整流程将在 TIAGo 演示平台上进行评估，以验证开放集 3DSSG 推理在真实场景中的有效性。

致谢。 本研究工作部分得到 LIEREx 项目的支持，该项目由德国联邦研究、技术与航天部（BMFTR）资助，资助编号为 16IW24004。DFKI 下萨克森（DFKI NI）由下萨克森州科学与文化部以及大众基金会提供支持。

声明

作者声明，不存在与本文内容相关的利益冲突。

参考文献

[1]Nüchter A, Hertzberg J. 面向移动机器人的语义地图。《机器人与自主系统》。2008年11月;56(11):915–926. https://doi.org/10.1016/j.robot.2008.08.001。

[2] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S 等. 从自然语言监督中学习可迁移的视觉模型. 载于：国际机器学习会议（ICML）. PMLR；2021年. 第 8748–8763 页。

[3] Rosinol A, Gupta A, Abate M, Shi J, Carlone L. 三维动态场景图：利用地点、物体和人类实现可操作的空间感知。《机器人学：科学与系统》（RSS）。2020年；。[4] Hughes N, Chang Y, Carlone L. Hydra：用于三维场景图构建与优化的实时空间感知系统。载于：《机器人学：科学与系统》（RSS）；2022年。第50篇文章。[5] Liu S, Zeng Z, Ren T, Li F, Zhang H, Yang J, 等。Grounding DINO：将 DINO 与接地预训练相结合，实现开放集目标检测。载于：欧洲计算机视觉会议（ECCV）。Springer；2024年。第38–55页。[6] Ren T, Jiang Q, Liu S, Zeng Z, Liu W, Gao H, 等。Grounding DINO 1.5：推进开放集目标检测的“边缘”。arXiv 预印本 arXiv:240510300。2024年；https://doi.org/10.48550/ARXIV.2405.10300。2405.10300。[7]Zeng F, Gan W, Wang Y, Liu N, Yu PS. 面向机器人学的大语言模型：综述。arXiv 预印本 arXiv:231107226。2023年；https://doi.org/10.48550/ARXIV.2311.07226。2311.07226。[8] Huang D, Yan C, Li Q, Peng X. 从大语言模型到大型多模态模型：文献综述。《应用科学》。2024年；14(12):5068。

[9]Zhai X, Mustafa B, Kolesnikov A, Beyer L. 语言图像预训练的 Sigmoid 损失。载于：IEEE/CVF 国际计算机视觉会议（ICCV）论文集；2023年。第11975–11986页。[10] Li J, Li D, Xiong C, Hoi S. BLIP：用于统一视觉-语言理解与生成的语言-图像预训练引导。载于：国际机器学习会议（ICML）。PMLR；2022年。第12888–12900页。[11]Li J, Li D, Savarese S, Hoi S. BLIP-2：利用冻结图像编码器和大型语言模型进行语言-图像预训练引导。载于：国际机器学习会议（ICML）。PMLR；2023年。第19730–19742页。[12] Zhou C, Loy CC, Dai B. 从 CLIP 中提取免费的密集标签。载于：欧洲计算机视觉会议（ECCV）。Springer；2022年。第696–712页。[13] Ghiasi G, Gu X, Cui Y, Lin TY. 利用图像级标签扩展开放词汇图像分割。载于：欧洲计算机视觉会议（ECCV）。Springer；2022年。第540–557页。[14]Ding Z, Wang J, Tu Z. 使用 MaskCLIP 实现开放词汇通用图像分割。载于：国际机器学习会议（ICML）。第202卷。PMLR；2023年。第8090–8102页。[15] Lüddecke T, Ecker AS. 使用文本和图像提示进行图像分割。载于：IEEE/CVF 计算机视觉与模式识别会议（CVPR）。IEEE；2022年。第7076–7086页。[16]Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L 等。分割一切。载于：IEEE/CVF 国际计算机视觉会议（ICCV）论文集；2023年。第4015–4026页。

[17] Yamazaki K, Hanyu T, Vo K,Pham T,Tran M, Doretto G 等。Open-Fusion：实时开放词汇三维映射与可查询场景表示。载于：2024年IEEE国际机器人与自动化会议（ICRA）。IEEE；2024年。第9411–9417页。

[18]Maggio D, Chang Y, Hughes N, Trang M, Griffith D, Dougherty C, 等. Clio：实时任务驱动的开放集3D场景图。IEEE 机器人自动化快报。2024年；9(10)：8921–8928。https://doi.org/10.1109/LRA.2024.3451395。

[19] Laina SB, Boche S, Papatheodorou S, Schae- fer S, Jung J, Leutenegger S. FindAnything：面向任意环境中机器人探索的开放词汇、以物体为中心的建图。arXiv 预印本 arXiv:250408603。2025年；。

[20]Kassab C, Mattamala M, Morin S, Büchner M, Valada A, Paull L, 等. 基本要素：设计简单有效的开放词汇场景图。arXiv 预印本 arXiv:241201539。2024年；。

[21] Linok S, Zemskova T, Ladanova S, Titkov R, Yudin D, Monastyrny M 等. 超越基本查询：基于三维场景图的开放词汇对象定位. 载于：IEEE机器人与自动化国际会议（ICRA）. IEEE；2025年. 第13582–13589页.

[22] Zhao X, Ding W, An Y, Du Y, Yu T, Li M 等. Fast Segment Anything. arXiv预印本 arXiv:230612156. 2023年；.

[23]Zhang C, Han D, Qiao Y, Kim JU, Bae SH, Lee S 等. Faster Segment Anything：面向移动应用的轻量级SAM. arXiv预印本 arXiv:230614289. 2023年；.

[24] Gadre SY、Wortsman M、Ilharco G、Schmidt L、Song S。《牧场上的奶牛：语言驱动的零样本物体导航的基线与基准》。载于：IEEE/CVF计算机视觉与模式识别会议论文集（CVPR）；2023年。第23171–23181页。

[25] Chang M、Gervet T、Khanna M、Yenamandra S、Shah D、Min SY 等。《GOAT：前往任意事物》。载于：机器人：科学与系统会议论文集（RSS）。荷兰代尔夫特；2024年。文章编号73。

[26] Ren AZ, Clark J, Dixit A, Itkina M, Majumdar A, Sadigh D. 探索直至确信：面向具身问答的高效探索. 载于：机器人学：科学与系统会议论文集（RSS）. 荷兰代尔夫特；2024年. 文章89.

[27] Gedicke T, Günther M, Hertzberg J. FLAP for CAOS：面向杂乱环境感知物体搜索的前瞻性主动感知。《IFAC-PapersOnLine》。2016年；49(15)：114–119。

[28] Liu P, Orru Y, Vakil J, Paxton C, Shafiullah NMM, Pinto L. 展示 OK-Robot：在机器人学中集成开放知识模型，真正重要的是什么。载于：《机器人学：科学与系统》（RSS）会议论文集。代尔夫特，荷兰；2024年。第 91 篇文章。

[29] Wang C, Fang HS, Gou M, Fang H, Gao J, Lu C. 杂乱环境中的可抓取性发现：实现快速准确的抓取检测。载于：IEEE/CVF 国际计算机视觉会议（ICCV）论文集；2021年。第 15964–15973 页。

[30] Fang HS, Wang C, Fang H, Gou M, Liu J, YanH 等. AnyGrasp：时空域中稳健且高效的抓取感知。《IEEE机器人学汇刊》。2023年；39(5)：3929–3945。

[31] Savva M, Kadian A, Maksymets O, Zhao Y, Wijmans E, Jain B 等。Habitat：具身人工智能研究平台。载于：IEEE/CVF 国际计算机视觉会议（ICCV）论文集；2019年。第 9339–9347 页。

[32] Chang A、Dai A、Funkhouser T、Halber M、Niessner M、Savva M 等。Matterport3D：从室内环境中的 RGB-D数据中学习。arXiv 预印本 arXiv:170906158。2017年。

[33]Ramakrishnan SK、Gokaslan A、Wijmans E、Maksymets O、Clegg A、Turner J 等。Habitat-Matterport 3D 数据集（HM3D）：用于具身人工智能的 1000 个大规模 3D 环境。arXiv预印本 arXiv:210908238。2021年。

[34] Mock A, Wiemann T, Pütz S, Hertzberg J. MICP-L：利用硬件加速光线投射进行机器人定位的基于网格的 ICP。载于：2024年 IEEE/RSJ 智能机器人与系统国际会议（IROS）。IEEE；2024年。第 10664–10671 页。