arXiv:2603.01813v1 [cs.RO] 2 Mar 2026

SSMG-Nav：借助语义骨架记忆图增强终身物体导航

Haochen Niu、Lantao Zhang、Xingwu Ji、Rendong Ying、Peilin Liu† 和 Fei Wen

摘要—**在陌生环境中根据人类指令导航至视野之外的目标，是服务机器人的核心能力。尽管相关研究已取得显著进展，大多数方法仍未充分利用可复用的持久记忆，限制了其在终身场景中的表现。此外，许多方法仅支持单模态输入，并采用短视的贪心策略，往往导致低效的往返机动（BFM）。为解决这些局限，我们提出 SSMG-Nav，这是一种基于**语义骨架记忆图**（SSMG）的物体导航框架。该框架将过去的观察整合为与空间对齐的持久记忆，并以拓扑关键点（例如交叉点、房间中心）为锚点。SSMG 将邻近实体聚类为子图，融合实体级与空间级语义，从而形成紧凑的候选目的地集合。为支持多模态目标（图像、物体和文本），我们集成了视觉语言模型（VLM）。对于每个子图，由记忆合成的多模态提示会引导 VLM 推断各目的地的目标信念。随后，长时域规划器权衡该信念与可通行性代价，生成使期望路径长度最小化的访问序列，从而减少回溯。在具有挑战性的终身基准测试集和标准 ObjectNav基准测试集上进行的大量实验表明，与强基线方法相比，我们的方法取得了更高的成功率和更高的路径效率，验证了 SSMG-Nav 的有效性。**

I. 引言

服务机器人需要根据人类指令，在此前未见过的环境中导航至当前视野之外的目标物体。传统方法通常依赖强化学习（RL）或模仿学习（IL） [1], [2]。尽管取得了显著进展，这些方法仍存在两项根本局限：（i）它们对数据和交互的需求量大，导致训练成本高昂；（ii）当部署场景偏离训练领域时，它们往往会在分布偏移下性能下降。

人类导航表明，应优先前往可能性较高的位置（例如，寻找毛绒玩具时前往卧室或货架），这凸显了常识空间语义推理的价值。同时，大规模预训练模型在常识与推理方面展现出了强大的泛化能力，为导航提供了一种新范式。近年来，涌现出许多基于视觉语言模型（VLM）或大语言模型（LLM）的零样本物体导航方法；这些方法利用广泛的先验知识，将环境语义线索与目标物体联系起来，展现出令人期待的优势 [3]–[6]。

本研究得到国家自然科学基金（编号：62276166）和科技创新2030—重大项目（编号：2022ZD0208700）的资助。

所有作者均来自中国上海市上海交通大学类脑应用技术中心（BATC），邮编200240（电子邮箱：{haochen niu,swagger,jixingwu,rdying,liupeilin,wenfei}@sjtu.edu.cn）。

然而，面向真实世界、以用户为中心的导航仍存在若干不足，尤其是在终身场景中，智能体需要利用历史经验不断改进 [7], [8]。首先，记忆机制的设计仍未得到充分探索：许多方法仅依据瞬时观测做出纯反应式决策；另一些方法构建价值地图 [4], [9] 或语义地图 [10]–[11]，但这些地图缺乏持久性和可复用性。近期关于图记忆的研究颇具前景，但其中大多数将“物体” [6], [12]或“航点” [13]建模为节点，通过邻接关系连接这些节点，并以文本形式构建记忆图，这限制了整体表达能力。其次，大多数方法针对单模态目标，无法处理多模态指令。第三，规划器通常采用贪心策略，促使智能体以短视的决策前往最可能的目标位置。这往往会导致低效的往返机动（BFM） [14]。

这些局限性促使我们构建一种能够维持持久记忆并支持多模态目标、从而实现高效终身任务执行的导航框架。其核心是语义骨架记忆图（SSMG），它提供可复用的空间锚定语义，以支持长时程规划。与现有以物体 [6]或航点 [13] 为节点的图表示不同，SSMG将物体聚合在环境骨架的拓扑关键点周围（例如走廊交叉口、房间中心）。由此得到一组候选目的地，每个候选目的地表示为一个子图，自然地结合了实体语义与空间语义。随后，我们引入视觉语言模型，并针对每个SSMG子图构建多模态提示，以引导视觉语言模型评估目标位于该子图中的信念。借助视觉语言模型的多模态常识与推理能力，我们可以处理以图像、物体或文本指定的目标，同时充分利用每个子图所编码的局部场景上下文。此外，我们设计了一种长时程规划器，联合考虑信念与可通行性代价，以计算使期望行驶距离最小化的访问序列。通过使用高效的近似求解器来处理这一优化问题，该规划器有效缓解了短视行为，并提升了导航效率。

主要贡献如下。

- 一种新颖的语义骨架记忆图（SSMG），将历史观测整合为持久的空间锚定记忆，统一实体语义与空间语义，以支持明智的决策。
- 一种SSMG驱动的长时程规划器，通过多模态提示引出基于视觉语言模型的环境推理，并采用一种高效近似来

在长期访问规划中平衡信念与遍历成本，从而提高效率。

- 一个完整的终身多模态物体导航框架 SSMG-Nav，在公开基准测试中取得了最先进的性能。

II. 相关工作

A. 物体导航

现有的物体导航方法大致可分为基于学习的方法和零样本方法。

1）基于学习的物体导航：基于学习的方法发展迅速，其中强化学习（RL）和模仿学习（IL）发挥着核心作用。Habitat-Web[2]基于大规模人类示范，表明模仿学习能够内化物体搜索中隐含且往往复杂的行为。在模块化方法方面，PONI [9] 使用势函数网络来决定“该去哪里寻找？”，并结合解析局部策略进行动作生成；SGM [15] 则通过自监督，在此基础上预测智能体可见范围之外的地图。ZSON [16] 通过使用 DD-PPO 训练策略，并利用 CLIP [17]进行图像目标导航，将零样本泛化能力融入其中。最近，MVVT[1] 利用多视角输入，将全局图像线索与局部空间上下文相结合，进一步提升了策略学习效果。尽管取得了这些进展，基于学习的方法通常需要大规模数据，并且在分布偏移下往往会出现性能下降。

2）零样本物体导航：近年来，随着大型模型的进步，零样本物体导航涌现出大量新方法。一些研究利用视觉语言模型（VLM），例如 CLIP [17],，将观测与目标物体进行对齐，并结合先进的探索策略或基于学习的策略来实现开放集物体导航 [3]。然而，这些方法缺乏对环境语义和记忆的分析，导致探索效率较低。

PIVOT [18]和 VLMNav [19] 将导航视为视频问答，通过向视觉语言模型提问来选择移动方向。在这一思路的基础上，DyNaVLM [20] 为系统加入了图记忆架构，将物体表示为节点，将关系邻接表示为边。

PONI [9], 、VLFM [4] 和 WMNav [21]等方法会构建价值地图，对地图上的前沿点进行评分，从而引导智能体优先探索目标相关性较高的区域。然而，这类价值地图通常难以重复利用，因为它们与当前任务紧密耦合。

此外，还有一些方法会构建语义地图。ESC [10]、L3MVN [22],和 OpenFMNav [11]等方法将物体类别、位置和关系整合到语义地图中，并将自我中心视图中的候选航点呈现为提示；随后由大语言模型选择下一个航点。在此范式的基础上，InstructNav [5] 利用动态导航链（DCoN）和时空链 [23]，以自然语言明确组织空间推理。

近期的一些研究聚焦于能够高效整合不同类型信息的记忆/地图图表示。SG-Nav [12] 构建了分层场景图 [24]，并提出一种思维链提示方法，使大语言模型能够推理目标位置。类似地，SayNav [25]采用自下而上的方式，以物体作为节点构建物体-房间-场景拓扑地图。MCGPT [13]在移动过程中采样视点，将其作为拓扑节点，并利用离线构建的思维链数据库，检索最适合在线分析的提示模板。

B. 终身多模态物体导航

近年来，一些研究者开始关注更贴近真实场景的物体导航任务。[6] 指出，现有零样本方法在针对特定任务构建基于大语言模型的推理框架方面存在局限，无法泛化到不同类型的目标。因此，他们提出了一种以物体为节点的图结构，用于统一的多模态目标表征，并基于大语言模型推理通过多阶段子图匹配验证目标，但未考虑终身任务执行。SayNav [25] 研究了多物体导航任务，但只能一次性呈现所有目标，且仅支持单一模态。OpenIN [26] 侧重于终身任务执行，但同样局限于单一模态。相比之下，GOAT [7] 将终身多模态物体导航形式化为以物体类别、图像或语言描述指定顺序目标的任务，而 GOAT-Bench [8]则提供了近期推出的标准化基准测试，包含统一协议和公开基线方法，以便进行公平比较。

III. 方法

A. 任务定义与系统概述

终身多模态物体导航任务将智能体置于一个此前未见过的场景中，并要求其依次导航至多个目标物体。每个目标物体由物体类别（例如“毛巾”）、文本描述（例如“浴室中悬挂在挂钩上的浴巾，位于淋浴间和淋浴凳的相对位置”）或唯一图像指定。每个目标定义一个子任务，当时间预算耗尽或智能体发出“停止”动作后，该子任务结束；随后智能体切换到下一个子任务，直至全部完成。在每次切换时，如果智能体与目标物体之间的距离小于距离阈值，则该子任务成功；否则失败。

如图 1 所示，我们提出的 SSMG-Nav 结构如下。在每个步骤中，智能体从模拟器获取 RGB、深度和位姿。利用这些传感器输入，我们增量式地更新几何信息（占据地图和骨架拓扑）与语义信息（价值地图和对象地图），从而随着时间推移构建 SSMG（第 III-B 节）。随后，导航策略在状态之间进行转换（第 III-D 节）。以当前状态为条件，我们从 SSMG子图生成多模态提示，以引导视觉语言模型进行信念估计（第 III-C.1 节）。最后，长时域规划器联合优化遍历成本和信念，生成长期访问路线，并执行其第一个航点（第 III-C.2 节）。

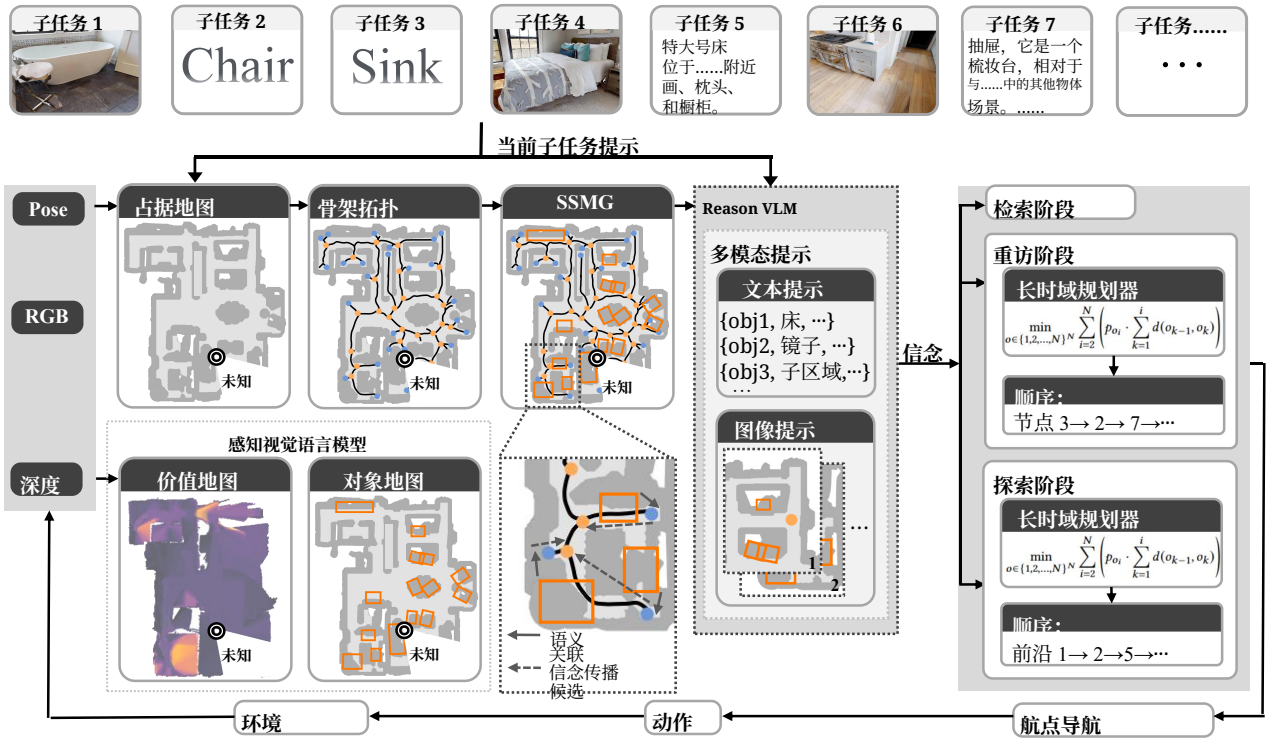


图 1：所提出的 SSMG-Nav 概述。

B. 语义骨架记忆图构建

1) 视觉感知： 在每个时间步 t，智能体获取传感器观测 $obs_t = \{I_t, D_t, P_t\}$，包括 RGB I_t、深度 D_t 和位姿 P_t。

我们首先将深度 D_t 转换为三维点云，并将其投影到鸟瞰视图（BEV）平面上。通过从传感器原点进行射线投射，我们构建并增量更新二维占据栅格 M_{occ}。每个栅格单元都会被标记为空闲、占据或未知，以跟踪探索进度。

为了分析观测中的语义信息，我们使用 BLIP-2 [27] 评估当前观测与目标之间的相关性。遵循 [4],，我们采用文本模板“前方似乎有一个 <目标物体> 。”，对当前 RGB 输入 I_t 执行文本图像匹配。与 M_{occ} 类似，我们将相关性分数投影到鸟瞰图平面上，得到价值地图M_{value}，它编码了已探索区域内场景上下文与目标物体之间的关联。

虽然价值地图捕捉了特定目标的相关性，但终身导航需要对观测进行与目标无关的语义表征。因此，我们再次使用 BLIP-2 为当前图像 I_t生成图像描述，记为 cap_t。为获得结构化语义，我们对 I_t 应用开放集检测器 Grounding-DINO [28]，并生成一组检测结果

$$\text{Det}_t = \{(\mathbf{b}_i, s_i, c_i) \mid i = 1, \ldots, N\}, \tag{1}$$

其中$\mathbf{b}_i$ 为边界框， s_i 为置信度分数， c_i 为预测类别， N 为检测到的物体数量。对于 Det_t 中的每个检测结果，我们应用 MobileSAM [29] 在相应的边界框内获取实例掩码。通过将掩码与深度融合，我们将像素反投影到三维空间，并构建对象级点云地图M_{obj}。对于每个物体 $j \in \mathcal{O}$，我们存储：

$$\mathcal{O}_j := \{\mathbf{PC}_j, c_j, \mathcal{V}_j\}, \quad \mathcal{V}_j := \{(\mathbf{I}_j^k, \mathbf{P}_j^k, \text{cap}_j^k, s_j^k)\}_{k=1}^{K}. \tag{2}$$

其中，$\mathbf{PC}j$ 表示点云， c_j 表示类别， $\mathcal{V}_j$ 收集最多 K 个视图：RGB $\mathbf{I}_j^k$、位姿$\mathbf{P}_j^k$、图像描述 cap_j^k和分数 s_j^k。我们在实验中设置 $K = 3$。

随着时间推移， M_{occ} 和 M_{value} 通过覆盖率进行增量更新。 M_{obj} 的更新受语义类别一致性以及点云三维边界框之间对齐情况的约束。对于进行融合的物体，根据检测置信度替换历史条目 $\mathcal{V}_j$，从而保留置信度较高的观测结果，这些结果通常对应更优的视点。

2） 骨架拓扑提取：对象簇图[12] 可能会失去几何保真度，而基于轨迹的节点生成 [30] 可能产生缺乏代表性的节点。因此，我们旨在提取简洁的几何骨干，并在之后为其补充语义信息。

按照 [14], ，我们通过形态学细化近似 M_{occ} 的骨架，以获得单像素宽的中轴。每个骨架像素都被分类为端点 $\mathcal{N}_{\text{e}}$（度为 1）、连接点 $\mathcal{N}_{\text{c}}$（度为 2）或交叉点 $\mathcal{N}_{\text{j}}$（度为 ≥ 3）。随后，我们剪除连接点，仅保留 $\mathcal{N}_{\text{e}}$ 和 $\mathcal{N}_{\text{j}}$（特殊节点），并在它们之间添加边。由此得到图$\mathcal{G} = (\mathcal{N}, \mathcal{E})$，其中连接点作为边属性存储，以支持快速路径规划。

3）语义关联：我们利用来自 M_{obj} 的物体信息扩充骨架拓扑 $\mathcal{G}=(\mathcal{N},\ \mathcal{E})$，从而构建完整的 SSMG：对于每个物体 $j\in\mathcal{O}$，我们选取其最高置信度视点 $k_j^\star:=\arg\max_k s_j^k$，并将参考位姿设置为$\mathbf{P}_j^\star:=\mathbf{P}_j^{k_j^\star}$；通过 $\pi(\cdot)$ 将位姿投影到鸟瞰图后，我们将物体关联到其最近节点：

$$n_j^\star:=\arg\min_{n\in\mathcal{N}}\left\|\mathbf{x}_n-\pi(\mathbf{P}_j^\star)\right\|,\mathcal{O}(n_j^\star)\leftarrow\mathcal{O}(n_j^\star)\cup\{j\}.\ (3)$$

为提高覆盖率并保持简洁，每个节点n 都根据当前关联物体中距离最远的物体定义一个数据驱动半径，

$$r_n:=\begin{cases}\max\limits_{j\in\mathcal{O}(n)}\left\|\mathbf{x}_n-\pi(\mathbf{P}_j^\star)\right\|, & \mathcal{O}(n)\neq\emptyset,\\ 0, & \text{otherwise},\end{cases}\quad(4)$$

并纳入该半径范围内的所有物体，

$$\mathcal{O}(n)\leftarrow\mathcal{O}(n)\cup\left\{j\in\mathcal{O}\ \middle|\ \left\|\mathbf{x}_n-\pi(\mathbf{P}_j^\star)\right\|\le r_n\right\},\quad(5)$$

这自然允许物体归属于多个节点（即 $j\in\mathcal{O}(n_1)\cap\mathcal{O}(n_2)$）。为支持下游信念传播，我们通过将每个端点连接到骨架上与其直接相连的交叉节点，扩展边集（即骨架图上的一跳连通性）：

$$\mathcal{E}\leftarrow\mathcal{E}\cup\left\{(n_e,n_j)\in\mathcal{N}_{\text{e}}\times\mathcal{N}_{\text{j}}\ \middle|\ \text{skeleton_adj}(n_e,n_j)\right\},\ (6)$$

其中，当且仅当 n_e 和 n_j 是骨架拓扑上的一跳邻居时，skeleton adj(n_e,n_j) 为真。

*C. SSMG*驱动的长时程规划器

零样本物体导航日益借助视觉语言模型/大语言模型，将感知转化为规划先验。然而，当前主流系统在每一步只选择最有希望的前沿 [6]，[10]，[12]，，而不会在长时程范围内对多个候选项进行排序，这常常导致 BFMs 和低效轨迹——自主探索领域也报告了这一现象[14]。此外，空间理解 [31]和长时程规划 [32]，[33] 方面的局限，使人们不禁怀疑：以某种形式将大规模全局地图和丰富的语义信息直接输入大语言模型，能否可靠地支持长时程规划。

因此，我们提出一种 SSMG 驱动的长时程规划器（LHP），它构建基于 SSMG 的多模态提示，以对 SSMG 的节点进行推理，随后采用近似算法进行长期路径规划。

1）基于视觉语言模型的信念评估：我们针对每个特殊节点进行信念估计。首先，根据端点构建多模态提示$\mathcal{N}_{\text{e}}$。每个节点都会存储附近的物体信息；对于每个物体 $j\in\mathcal{O}(n_i)$，我们会构建一个文本提示 {物体标识$_j$，物体类别$_j$，物体描述$_j$}。随后，我们将一个子图提取为图像提示：以节点n_i为中心，在鸟瞰平面上标注每个物体的点云质心及其物体标识$_j$。将文本提示与视觉提示拼接后，即可生成背景知识提示。最后，思维链推理会引导视觉语言模型（Qwen-VL-Plus [34]）逐步分析节点 n_i，并生成格式化的信念估计值，信念$_i$。

评估完所有端点 $\mathcal{N}_{\text{e}}$ 后，我们将端点线索自底向上传播至 $\mathcal{N}_{\text{j}}$。这使候选节点集中于交叉点，从而缩小规划搜索空间。对于每个交叉点，我们将其关联端点的信息补充到背景知识提示中，并再次查询视觉语言模型，以输出交叉点信念。

2）长时程规划：我们的长时程规划器将所有 $\mathcal{N}_{\text{j}}$ 视为候选节点，并计算一条依次访问这些节点的访问路径。智能体沿该路径前进，直到找到目标物体。给定 N 个候选节点 $\{n_1,\dots,n_N\}$、目标概率 $\{p_1,\dots,p_N\}$ 和成对距离 $d(i,\ j)$，我们寻求一种顺序 $\{n_{o_1},\dots,n_{o_N}\}$，使搜索目标物体时的期望行程最小：

$$\min_{o\in\{1,2,\dots,N\}^N}\sum_{i=2}^{N}\left(p_{o_i}\cdot\sum_{k=2}^{i}d(o_{k-1},o_k)\right),\quad(7)$$

满足以下条件：

$$\sum_{i=1}^{N}p_i=1,\quad p_i\ge0,\qquad|\{o_1,o_2,\dots,o_N\}|=N.\quad(8)$$

该问题是 NP 难问题。受 2-opt局部搜索启发式 [35], 的启发，我们提出一种近似求解方法。我们首先将第 III-C.1 节中的估计信念$_i$ 通过温度为 τ 的 softmax 转换为概率 p_i：

$$p_i=\frac{\exp\left(belief_i/\tau\right)}{\sum_{j=1}^{N}\exp\left(belief_j/\tau\right)},\quad\tau>0.\quad(9)$$

我们首先使用贪心启发式生成一条初始路径，该启发式始终访问最近的未访问节点。为简化起见，令 n_1 表示 n_{o_1}，并将初始路径记为 $\pi_0=\{n_1,n_2,\dots,n_N\}$。其期望行驶成本 $E(\pi_0)$ 根据式（7）计算。接下来，选取一段 $\{n_p,n_{p+1},\dots,n_q\}$，将其反转并替换原始片段，得到 $\pi_{pq}=\{n_1,\dots,n_{p-1},n_q,n_{q-1},\dots,n_{p+1},n_p,\dots,n_N\}$。计算 $E(\pi_{pq})$；如果其低于当前期望值，则接受 π_{pq}。遍历所有 $(p,\ q)$ 对，并重复上述过程，直至达到最大迭代次数或无法进一步改进。

D. 导航策略

基于上述内容，我们制定了一种包含四个阶段的策略：检索、重访、探索和航点导航。

1）检索阶段：切换到新的子任务时，我们首先在 SSMG 中查询指定目标。如果找到同类别的实例，我们会按照第 III-C.1 节构造提示。如果确认某个实例就是目标，则继续进行航点导航。如果当前 SSMG 中不存在该目标，我们会评估 M_{occ}上的探索状态。如果仍有有效的前沿，则进入探索阶段，搜索未探索区域。如果探索已完成，我们会考虑可能存在的漏检（例如因视点所致），并转入重访阶段。

表I：终身多模态物体导航对比。

方法	零样本	GOAT已见验证集			GOAT未见验证集		
		s-SR（↑）	e-SR(↑)	SPL(↑)	s-SR (↑)	e-SR(↑)	SPL(↑)
SenseAct-NN Skill Chain [8]	×	0.292	–	0.128	0.295	–	0.113
SenseAct-NN 单体式 [8]	×	0.168	–	0.094	0.123	–	0.068
SenseAct-NN 单体式†	×	0.159	0.003	0.104	0.126	0.008	0.088
CoW [3]	✓	0.148	–	0.087	0.161	–	0.104
Modular GOAT [8]	✓	0.263	–	0.175	0.249	–	0.172
PIVOT [18]	✓	–	–	–	0.102	–	0.055
VLMNav [19]	✓	–	–	–	0.201	–	0.096
DyNaVLM [20]	✓	–	–	–	0.255	–	0.102
VLFM† [4]	✓	0.308	0.025	0.129	0.406	0.028	0.159
SSMG-Nav（本文）	✓	**0.412**	**0.056**	**0.318**	**0.465**	**0.086**	**0.341**

2）重访阶段：在此阶段，我们按照第 III-C 节计算访问序列 π，并依次访问各节点。在每个交叉点，智能体会进行 360° 旋转，以获取全景观测。如果匹配到目标，则转入航点导航；否则，我们会检查是否有相邻的末端节点，若有则访问它们，并重复进行 360°扫描。如果仍未匹配到目标，则前往 π中的下一个交叉点并重复上述过程。

3）探索阶段：在此阶段，我们同样进行长视野规划，以提高探索效率。具体而言，我们将已知区域与未知区域之间的边界提取为前沿。随后对这些前沿进行聚类，得到若干前沿点 $\{f_1, f_2, \ldots, f_n\}$。然后，我们采用第III-C.2节所述方法求解路径。考虑到探索过程中SSMG信息不完整以及未知区域的不确定性，我们简单地使用价值地图 M_{value} 作为每个前沿点的信念，类似于 [4]中的方法。在计算出访问路径后，我们导航至第一个前沿；随着运动过程中环境的更新，我们在每一步重新规划，并通过航点导航前往新的目标。

4）航点导航阶段：在导航阶段，我们基于目标点执行A*路径规划，并在特定距离处生成前瞻点。随后，我们采用点导航方案，将在HM3D数据集 [36]上使用可变经验展开（VER）强化学习算法训练得到的模型作为局部规划器，如 [4]所述。该规划器以深度以及目标点的相对位置和朝向作为输入，输出相应的动作。

IV. 实验结果

A. 实验设置

1) 基准测试： 我们在基于 Habitat 模拟器的四个高分辨率基准测试 [37] 上进行评估，涵盖两项任务。首先，我们考虑 GOAT-Bench [8] 引入的、具有挑战性的终身多模态物体导航任务。我们使用 *GOAT*已见验证集 和 *GOAT*未见验证集 这两个划分；每个划分包含 36 个室内场景，每个场景包含 10 个回合，每个回合由 5–10 个子任务组成。每个子任务

表II：物体导航对比。

方法	零-shot	HM3D		MP3D	
		成功率（↑）	SPL(↑)	成功率 (↑)	SPL(↑)
PONI [9]	×	–	–	0.318	0.121
Habitat-Web [2]	×	0.415	0.160	0.316	0.085
ZSON [16]	×	0.255	0.126	0.153	0.048
SGM [15]	×	**0.602**	0.308	0.377	0.147
ESC [10]	✓	0.392	0.223	0.287	0.142
L3MVN [22]	✓	0.504	0.231	0.349	0.145
OpenFMNav [11]	✓	0.549	0.244	0.372	0.157
SG-Nav [12]	✓	0.540	0.249	0.402	0.160
VLFM [4]	✓	0.525	0.304	0.364	0.175
VLFM†	✓	0.523	0.303	0.339	0.165
InstructNav [5]	✓	0.510	0.187	0.420	0.161
UniGoal [6]	✓	0.545	0.251	0.410	0.164
WMNav [21]	✓	0.581	0.312	**0.454**	0.172
Ours	✓	0.595	**0.337**	0.443	**0.191**

通过以下三种模态之一指定目标物体：物体类别、参考图像或自然语言描述。

为进一步评估我们方法的普适性，我们还使用两个标准数据集*HM3D* [36] 和*MP3D* [38]对经典物体导航任务进行评估。第一个划分包含 20 个场景中的 2,000 个回合，涵盖 6 个目标类别；第二个划分包含 11 个场景中的 2,195 个回合，涵盖 21 个目标类别。该任务可视为终身多模态设置的一种特殊情况——仅限于物体类别这一种模态，且每个回合恰好包含一个子任务。

2）指标：在物体导航任务中，成功率（*SR*）和按路径长度倒数加权的成功率（*SPL*）是常用的评估指标。

鉴于终身多模态物体导航的特殊性，我们将总体成功率细化为两项粒度更细的指标：子任务成功率（*s-SR*）和回合成功率（*e-SR*）。具体而言，s-SR 的计算方式是：先衡量每个回合中各子任务的成功情况，再对所有回合的子任务级成功率取平均。只有当某个回合中的所有子任务均已完成时，我们才将该回合视为成功

；相应地，e-SR 表示满足这一标准的回合所占比例，即全部完成的回合所占的比例。

3）基线方法：针对终身多模态物体导航，我们将 SSMG-Nav 与两种强化学习方法 SenseAct- NN Skill Chain 和 Monolithic [8],以及六种零样本方法进行比较。其中，CoW [3], PIVOT [18], 和 VLMNav [19]是不使用记忆（无状态）的方法。Modular GOAT [8]构建显式语义地图，而 DyNaVLM [20]采用基于图的记忆架构。为建立强有力的价值地图基线，我们还将 VLFM [4] 适配到该基准测试，并仔细调整超参数，以确保比较公平。

针对物体目标导航任务，我们评估了四种基于学习的方法和八种零样本方法。其中，ESC [10],L3MVN [22],和 OpenFMNav [11]基于语义地图；VLFM [4], InstructNav [5], 和 WMNav[21] 基于价值地图；SG-Nav [12] 和 UniGoal [6]则基于场景图。

4) 配置详情：按照 [8]，Habitat中的智能体被建模为半径0.17 m、高度1.44 m的圆柱体。智能体安装了一台高度为1.31 m的RGB相机，分辨率为 640 × 480 ，水平视野为 79°。此外，还安装了一台具有相同内参和外参的深度相机，其有效范围为0.5–5.0 m。动作空间为 {前进、左转、右转、停止}；每次前进移动0.25 m，每次旋转 30°。当智能体发出停止动作且距离任一候选目标视点不超过0.5 m时，任务即告成功。每项任务最多允许500个时间步。

B. 基准测试结果

1) 终身多模态物体导航任务上的性能：定量结果如表I所示，其中粗体和下划线分别表示最佳和次佳结果。带有† 标记的方法（SenseAct- NN Monolithic和 VLFM）是我们重新实现的方法，并补充了e-SR指标，以增强分析的全面性。从结果可以看出，VLFM取得了较高的成功率，体现了其探索能力；然而，由于缺乏持久记忆，其SPL仍然较低。相比之下，Modular GOAT得益于其记忆模块，SPL显著高于VLFM，尽管成功率较低。我们的方法在所有指标上均取得最佳表现，明显优于基线方法，尤其在SPL方面提升显著。这些结果表明，SSMG能够有效抽象并保留历史观测，使智能体在终身执行过程中利用过去的信息，并借助大语言模型的通用知识和推理能力进行长时域决策，从而兼顾高成功率与高效率。另一方面，所有方法的整体e-SR仍相对较低，这凸显了终身多模态物体导航仍极具挑战性，值得进一步研究。

我们还对多模态任务的性能进行了统计分析，如图 2 所示。

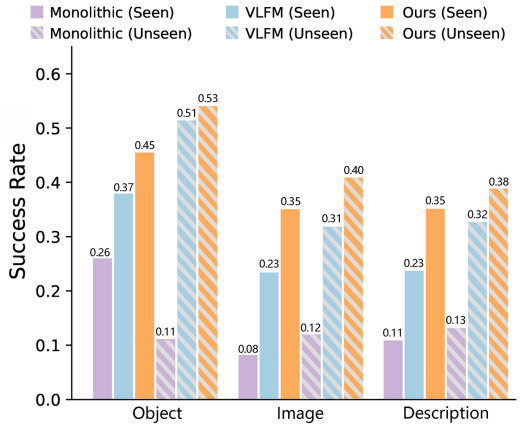


图2：不同模态上的性能。

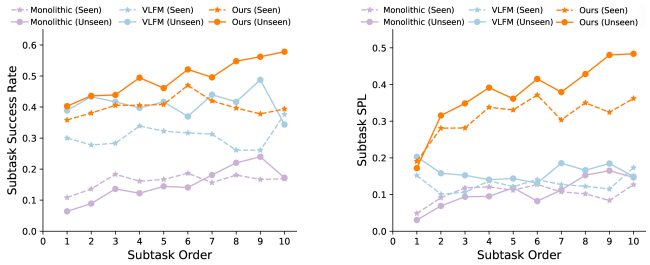
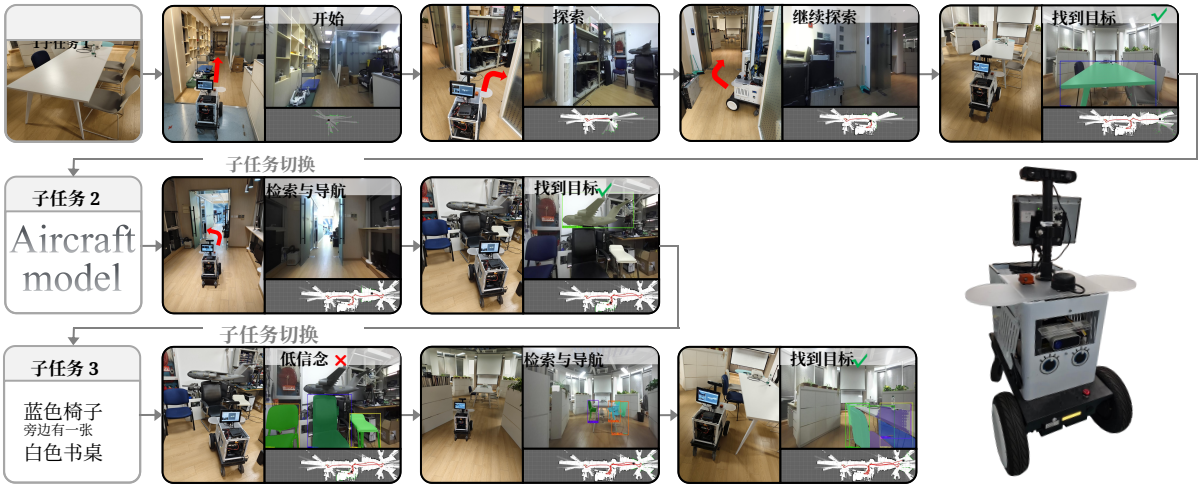


图3：随时间变化的性能。

结果表明，我们的方法在所有模态下均表现出优势。值得注意的是，使用物体模态的任务通常不如使用图像和描述模态的任务，这可能是因为实例级物体导航需要对目标物体进行更深入的理解和分析。

为探究不同方法在终身任务上的特性，我们以子任务索引为自变量计算性能统计量；结果如图3所示。我们首先观察到，我们的方法在几乎所有子任务顺序上均持续优于基线方法，从而强化了前文图表所得出的结论。此外，相对于各自的初始状态，我们的方法与Monolithic基线均随子任务推进（即随时间）取得性能提升，其中我们的方法在*GOAT*未见验证集 划分上的SPL提升尤为显著。如第IV-A.3节所述，基于强化学习的Monolithic方法隐式编码地图信息，而我们的方法则将历史观测显式抽象为SSMG以辅助决策；因此，性能随时间提升是预期且一致的发现。相比之下，VLFM凭借视觉语言模型的通用知识在整体上优于Monolithic，但并未表现出随时间的明显提升，这进一步凸显了记忆在终身任务中的重要性，无论其以显式还是隐式方式实现。

2）物体导航任务上的表现：表II中的定量结果表明，我们的方法在 SPL 上始终优于强大的基线方法，同时取得了相当的成功率



Fi图 4：SSMG-Nav 在包含三个子任务的办公室环境中的真实世界演示。

表III：消融研究。

方法		s-SR	e-SR	SPL
Map	规划器			
语义地图	贪心	0.389	0.056	0.278
价值地图	贪心	0.442	**0.083**	0.258
SSMG	贪心	0.453	0.056	0.328
SSMG	无重复访问的长时程规划器	0.479	**0.083**	0.361
SSMG	LHP	**0.502**	**0.083**	**0.376**

，而在 GOAT-Bench 上则并非如此。我们认为，该基准测试不具备终身学习特性，因此持久记忆带来的边际收益较小。我们进一步认为，我们的 SPL 优势主要归功于所提出的长时域规划器。值得注意的是，表现最佳的零样本方法 WMNav 在每一步都会接收由六个摄像头组成的全景观测，而我们的方法仅依赖单个摄像头。

C. 消融研究

在消融研究中，我们评估地图表示和规划器策略对性能的影响。表III报告了在下采样的*GOAT*未见验证集子集上的结果。“语义地图”仅记录物体语义。“价值地图”完全依赖根据当前目标生成的价值地图（即没有可重复使用的长期记忆）。“贪心”会选择价值最高或距离最近的区域。尽管“语义地图”的 s-SR 和 e-SR 低于“价值地图”，但它通过利用记忆获得了更高的 SPL，这凸显了持久记忆对于终身任务的重要性。对于我们的方法，将“SSMG”与“贪心”相结合，效果显著优于其他地图格式，证明了 SSMG 作为记忆表示的优势。此外，采用“无重复访问的长时程规划器”后性能有所提升，采用完整的“长时程规划器”后性能进一步提升，这表明其收益逐步增加，也验证了我们的模块设计。我们方法的完整版本取得了总体最佳结果。

D. 真实世界部署

我们的方法部署在轮式机器人平台上。ZED立体相机提供 RGB 和深度数据，二维激光雷达则通过 Cartographer 提供定位。我们采用 [39]作为局部规划器，以导航至航点。实验在办公室环境中进行，包含三个子任务：（i）指定为“白色书桌”的图像模态目标；（ii）指定为飞机模型的物体模态目标；以及（iii）描述为“白色书桌旁的一把蓝色椅子”的文本模态目标。图4展示了实验过程概览。这表明，我们的方法能够在物理平台上执行终身多模态物体导航。

V. 结论

本文提出了SSMG-Nav，一种基于语义骨架记忆图（SSMG）的新型终身多模态物体导航框架，可将历史观测整合为具有空间依据、可复用的记忆。通过将语义锚定到拓扑关键点，并将邻近实体聚类为子图，SSMG生成紧凑、可直接用于决策的候选目的地。在此表示的基础上，我们合成多模态提示，以从视觉语言模型中引出目标置信度，并使用长时域规划器（长时程规划器）进行规划，在信念与可通行性之间取得平衡，从而最小化期望路径长度。实验表明，我们的方法在终身多模态物体导航和标准物体导航上均超越了强基线方法，并在具有挑战性的终身设定中取得了显著提升。我们进一步证实了其在真实世界中的可行性。

然而，仍存在若干局限。首先，该方法依赖大型模型，这会带来延迟，并限制真实世界部署中的运行速度；因此，机器人的运行速度相对较低。其次，当前方法无法处理多楼层场景。一个颇有前景的方向是将基于体素的表示引入SSMG，以构建分层SSMG。

参考文献

[1] C. Lu, M. Liu, Z. Luan, Y. He, and B. Chen, “面向物体目标导航的多视角空间上下文与状态约束,” *IEEE*机器人与自动化快报, pp. 1–8, 2025年。 [2] R. Ramrakhya, E. Undersander, D. Batra, and A. Das, “Habitat-web：大规模从人类示范中学习具身物体搜索策略,” in *IEEE*计算机视觉与模式识别会议论文集, 2022年, pp. 5163–5173. [3] S. Y. Gadre, M. Wortsman, G. Ilharco, L. Schmidt, and S. Song, “Cows on pasture：语言驱动的零样本物体导航的基线与基准测试,” in *IEEE*计算机视觉与模式识别会议论文集, 2023年6月, pp. 23 171–23 181. [4] N. Yokoyama, S. Ha, D. Batra, J. Wang, and B. Bucher, “VLFM：用于零样本语义导航的视觉语言前沿地图,” in *IEEE*机器人与自动化国际会议论文集, 2024 年。 [5] Y. Long, W. Cai, H. Wang, G. Zhan, and H. Dong, “InstructNav：面向未探索环境中的通用指令导航的零样本系统,” *arXiv*预印本 *arXiv:2406.04882*, 2024年。 [6] H. Yin, X. Xu, L. Zhao, Z. Wang, J. Zhou, and J. Lu, “UniGoal：迈向通用零样本目标导向导航,” in *IEEE*计算机视觉与模式识别会议论文集, 2025年。 [7] M. Chang, T. Gervet, M. Khanna, S. Yenamandra, D. Shah, S. Y. Min, K. Shah, C. Paxton, S. Gupta, D. Batra, R. Mottaghi, J. Malik, and D. S. Chaplot, “GOAT：前往任意物体,” in 机器人学：科学与系统, 2023年。 [8] M. Khanna*, R. Ramrakhya*, G. Chhablani, S. Yenamandra, T. Gervet, M. Chang, Z. Kira, D. S. Chaplot, D. Batra, and R. Mottaghi, “GOAT-Bench：多模态终身导航基准测试,” in *IEEE*计算机视觉与模式识别会议论文集, 2024年。 [9] S. K. Ramakrishnan, D. S. Chaplot, Z. Al-Halah, J. Malik, and K. Grauman, “PONI：通过无交互学习实现物体目标导航的势函数,” in *IEEE*计算机视觉与模式识别会议论文集, 2022年6月, pp. 18 868–18 878. [10] K. Zhou, K. Zheng, C. Pryor, Y. Shen, H. Jin, L. Getoor, and X. E. Wang, “ESC：结合软常识约束的零样本物体导航探索,” in 国际机器学习会议论文集, vol. 202, 美国夏威夷州檀香山, 2023年7月, pp. 42 829–42 842. [11] Y. Kuang, H. Lin, and M. Jiang, “OpenFMNav：借助视觉语言基础模型实现开放集零样本物体导航,” in 计算语言学协会论文集：*NAACL 2024*, 墨西哥墨西哥城, 2024年6月, pp. 338–351. [12] H. Yin, X. Xu, Z. Wu, J. Zhou, and J. Lu, “SG-Nav：用于基于大语言模型的零样本物体导航的在线三维场景图提示,” in 神经信息处理系统进展论文集, vol. 37, 2024年, pp. 5285–5307. [13] Z. Zhan, L. Yu, S. Yu, and G. Tan, “Mc-gpt：借助记忆地图与推理链增强视觉语言导航,” *arXiv*预印本 *arXiv:2405.10620*, 2024年。 [14] H. Niu, X. Ji, L. Zhang, F. Wen, R. Ying, and P. Liu, “面向复杂未知环境探索的骨架拓扑规划器,” in *IEEE*机器人与自动化国际会议论文集, 2025年, pp. 11 766– 11 772.[15] S. Zhang, X. Yu, X. Song, X. Wang, and S. Jiang, “Imagine before go：用于物体目标导航的自监督生成式地图,” in *IEEE*计算机视觉与模式识别会议论文集, 美国华盛顿州西雅图, 2024年6月, pp. 16 414–16 425. [16] A. Majumdar, G. Aggarwal, B. Devnani, J. Hoffman, and D. Batra, “ZSON：利用多模态目标嵌入实现零样本物体目标导航,” in 神经信息处理系统进展论文集, vol. 35, 2022年, pp. 32 340–32 352. [17] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, 等, “通过自然语言监督学习可迁移视觉模型,” in 国际机器学习会议论文集 , 2021 年 , pp. 8748–8763. [18] S. Nasiriany, F. Xia, W. Yu, T. Xiao, J. Liang, I. Dasgupta, A. Xie, D. Driess, A. Wahid, Z. Xu, Q. Vuong, T. Zhang, T.-W. E. Lee, K.- H. Lee, P. Xu, S. Kirmani, Y. Zhu, A. Zeng, K. Hausman, N. Heess, C. Finn, S. Levine, and B. Ichter, “PIVOT：迭代式视觉提示可为视觉语言模型提取可操作知识,” in 国际机器学习会议论文集, 2024年。[19] D. Goetting, H. G. Singh, and A. Loquercio, “基于视觉语言模型的端到端导航：将空间推理转化为问答,” in神经信息处理系统进展论文集, 2025年7月, pp. 22–35. [20] Z. Ji、H. Lin 和 Y. Gao，“Dynavlm：一种具有动态视点和自完善图记忆的零样本视觉语言导航系统”，arXiv 预印本 arXiv:2506.15096，2025年。[21] D. Nie、X. Guo、Y. Duan、R. Zhang 和 L. Chen，“WMNav：将视觉语言模型集成到世界模型中以实现物体目标导航”，载于《IEEE智能机器人与系统国际会议论文集》，2025年，第2392–2399页。[22] B. Yu、H. Kasaei 和 M. Cao，“L3MVN：利用大语言模型进行视觉目标导航”，载于《IEEE智能机器人与系统国际会议论文集》，2023年，第3554–3560页。[23] J. Wei、X. Wang、D. Schuurmans、M. Bosma、F. Xia、E. Chi、Q. V. Le、D. Zhou 等，“思维链提示能够激发大语言模型中的推理能力”，载于《神经信息处理系统进展》，第35卷，2022年，第24 824–24 837页。[24] Q. Gu、A. Kuwajerwala、S. Morin、K. M. Jatavallabhula、B. Sen、A. Agarwal、C. Rivera、W. Paul、K. Ellis、R. Chellappa、C. Gan、C. M. De Melo、J. B. Tenenbaum、A. Torralba、F. Shkurti 和 L. Paull，“Conceptgraphs：用于感知和规划的开放词汇三维场景图”，载于《IEEE机器人与自动化国际会议论文集》，2024年5月，第5021–5028页。[25] A. Rajvanshi、K. Sikka、X. Lin、B. Lee、H.-p. Chiu 和 A. Velasquez，“SayNav：为大语言模型提供接地能力，以实现在新环境中的动态规划导航”，载于《自动规划与调度国际会议论文集》，2024年3月。[26] Y. Tang、M. Wang、Y. Deng、Z. Zheng、J. Deng、S. Zuo 和 Y. Yue，“OpenIN：动态家庭环境中的开放词汇实例导向导航”，《IEEE机器人与自动化快报》，第10卷，第9期，第9256–9263页，2025年。[27] J. Li、D. Li、S. Savarese 和 S. Hoi，“BLIP-2：借助冻结图像编码器和大语言模型启动语言—图像预训练”，载于《国际机器学习会议论文集》，2023年，第19 730–19 742页。[28] S. Liu、Z. Zeng、T. Ren、F. Li、H. Zhang、J. Yang、Q. Jiang、C. Li、J. Yang、H. Su 等，“Grounding dino：将DINO与基于接地的预训练相结合，实现开放集目标检测”，载于《欧洲计算机视觉会议论文集》，Springer，2024年，第38–55页。[29] C. Zhang、D. Han、Y. Qiao、J. U. Kim、S.-H. Bae、S. Lee 和 C. S. Hong，“Faster segment anything：面向移动应用的轻量级SAM”，arXiv预印本 arXiv:2306.14289，2023年。[30] G. Zhao、G. Li、W. Chen 和 Y. Yu，“Over-nav：通过开放词汇检测和结构化表示提升迭代式视觉语言导航”，载于《IEEE计算机视觉与模式识别会议论文集》，美国华盛顿州西雅图，2024年6月，第16 296–16 306页。[31] B. Chen、Z. Xu、S. Kirmani、B. Ichter、D. Sadigh、L. Guibas 和 F. Xia，“Spatialvlm：赋予视觉语言模型空间推理能力”，载于《IEEE计算机视觉与模式识别会议论文集》，美国华盛顿州西雅图，2024年6月，第14 455–14 465页。[32] M. Aghzal、E. Plaku、G. J. Stein 和 Z. Yao，“大语言模型用于自动规划的综述”，arXiv预印本 arXiv:2502.12435，2025年。[33] J. Zhang、L. Tang、Y. Song、Q. Meng、H. Qian、J. Shao、W. Song、S. Zhu 和 J. Gu，“Fltrnn：利用大语言模型为机器人实现可靠的长时域任务规划”，载于《IEEE机器人与自动化国际会议论文集》，2024年，第6680–6686页。[34] J. Bai、S. Bai、S. Yang、S. Wang、S. Tan、P. Wang、J. Lin、C. Zhou 和 J. Zhou，“Qwen-vl：一种功能多样的视觉语言模型，可用于理解、定位、文本阅读等任务”，arXiv预印本 arXiv:2308.12966，2023年。[35] G. A. Croes，“一种求解旅行商问题的方法”，《运筹学》，第6卷，第6期，第791–812页，1958年。[36] S. K. Ramakrishnan、A. Gokaslan、E. Wijmans、O. Maksymets、A. Clegg、J. Turner、E. Undersander、W. Galuba、A. Westbury、A. X. Chang 等，“Habitat-Matterport三维数据集（HM3D）：用于具身人工智能的1000个大型三维环境”，arXiv预印本 arXiv: . . 3 08238, 2021 2109 d。[37] A. Szot、A. Clegg、E. Undersander、E. Wijmans、Y. Zhao、J. Turner、N. Maestre、M. Mukadam、D. S. Chaplot、O. Maksymets 等，“Habitat 2.0：训练家庭助手重新布置其栖息环境”，载于《神经信息处理系统进展》，第34卷，2021年，第251–266页。[38] A. Chang、A. Dai、T. Funkhouser、M. Halber、M. Niebner、M. Savva、S. Song、A. Zeng 和 Y. Zhang，“Matterport3D：从室内环境中的RGB-D数据中学习”，载于《国际三维视觉会议论文集》，2017年，第667–676页。[39] J. Zhang、C. Hu、R. G. Chadha 和 S. Singh，“Falco：快速似然碰撞规避及其在人类引导导航中的扩展”，《现场机器人学杂志》，第37卷，第8期，第1300–1313页，2020年12月。