视觉与语言导航(VLN)研究前沿综述:从奠基性工作到大模型驱动的泛化与适应

I. 具身导航与视觉与语言导航(VLN)引言

1.1 视觉与语言导航(VLN)的界定

视觉与语言导航(Vision-and-Language Navigation, VLN)是实现具身智能(Embodied Intelligence)的关键研究路径之一[1]。具身智能的核心目标是提高智能体对环境的感知、理解和交互能力。在VLN任务中,智能体的核心挑战在于如何有效接收和理解人类提供的自然语言指令,并完全依赖自身视觉信息(例如第一人称RGB-D图像)在复杂的3D环境中执行准确的连续行动[1]。

VLN并非一个单一学科的任务,它综合了人工智能(AI)、自然语言处理(NLP)、计算机视觉(CV)和机器人学(Robotics)等多个领域的技术[1]。因此,VLN被视为评估具身智能体鲁棒性和通用性的关键基准。

VLN任务的内在难度在于其交叉模态的复杂性,它要求智能体克服两大核心挑战:首先是跨模态对齐(Cross-Modal Alignment),即如何将抽象的语言指令(例如"走到厨房,停在水壶旁边")与动态变化的视觉场景特征建立精确的对应关系。其次是该任务本质上是一个部分可观察马尔可夫决策过程(Partially Observable Markov Decision Process, POMDP)[2]。这意味着智能体必须依赖有限的、局部的视觉观察,结合历史行动序列和指令,做出时间依赖的决策,最终实现从高层语言理解到低层动态动作执行的平稳过渡[1]。如果一个智能体在VLN任务中失败,错误可能来源于语言理解的偏差、视觉识别的错误、或者对环境状态的跟踪不力,这表明VLN必须在多个计算维度上都具备高度的稳健性和模块化能力。

1.2 具身任务的演进

VLN的出现，标志着具身AI研究从静态、单一模态任务（如图像分类、文本问答）向动态、多模态决策任务的迈进。早期的视觉语言任务通常只需要回答关于单个图像或视频的静态问题。相比之下，VLN要求智能体在复杂的、非结构化的室内环境中，根据语言指令进行长期的、序列化的决策制定和执行。

这种从被动感知到主动交互的转变，催生了对更先进智能体架构的需求。成功解决VLN要求智能体具备地图构建、目标识别、路径规划和错误修正等一系列高级认知能力。这一演进验证了VLN在促进具身AI发展中的核心地位，并使其成为评估Sim2Real迁移性能的关键基准。

II. 奠基性基准与离散VLN（VLN 1.0）

2.1 里程碑：Room-to-Room (R2R) 数据集（2018）

VLN领域的奠基性工作始于2018年由Peter Anderson等人提出的Room-to-Room (R2R) 数据集 [3]。R2R是首个在真实的、建筑级3D环境中进行视觉地面化自然语言导航的基准数据集 [4]。

环境与任务设置：
R2R任务要求自主智能体遵循人类生成的导航指令，在先前未见的真实建筑中导航 3。R2R数据集依托于Matterport3D模拟器 3。Matterport3D提供了一个高保真、真实世界的3D环境，智能体可以通过视觉信息（RGB-D图像）与之交互 3。
离散导航与局限性：
R2R任务采用离散导航图（Discrete Navigation Graph）的设置 7。在这个设置中，环境被抽象为一系列预定义的视点（即图中的节点），智能体只能在这些节点之间移动 7。
早期的R2R实验中，路径通常被设定为起点到终点之间的最短路径 [5]。虽然R2R的出现是革命性的，但这种“最短路径偏见”很快被发现是一个关键限制。由于大多数路径是拓扑图上的最短路径，智能体可以学习一种

目标寻求（Goal-Seeking）策略，即优先使用视觉信息预测目标位置，而不是严格遵循指令中提到的中间步骤和地标（路径依从性） [5]。这种结构性偏见，降低了对智能体真正语言理解和路径推理能力的考验。

R2R的工作在2018年CVPR大会上获得了Spotlight口头报告的认可，确立了其在VLN研究中的奠基性地位 [3]。

2.2 初始架构基线

在R2R数据集上，最初的基线模型采用了标准的序列到序列（Seq2Seq）神经网络[4]。这些模型通常使用循环神经网络（RNN）或长短期记忆网络（LSTM）来编码语言指令和视觉历史。然而，由于R2R环境的视觉和序列复杂性，这些早期的Seq2Seq模型在泛化到未见环境时性能有限。

为了解决数据稀疏性和泛化能力不足的问题，研究人员很快转向更复杂的、包含数据增强机制的架构，其中**“说话者-跟随者”（Speaker-Follower）模型**成为后续研究的重要基础（详见V.1节）。

III. 扩展研究边界：多语言与细粒度基准

3.1 Room-Across-Room (RxR) 数据集（2020）

RxR数据集作为R2R的直接继任者，旨在解决R2R在规模、语言多样性和路径设计上的固有缺陷[5]。RxR的引入标志着VLN研究对智能体推理能力提出了更高的要求。

规模与路径多样性：
RxR极大地扩展了数据集的规模。它包含126K条指令（R2R为22K）和16.5K条路径（R2R为7K）5。RxR路径的平均长度更长（平均7.0个拓扑图节点，14.9米），相比之下R2R平均只有5.0个节点，9.4米 5。
更重要的是，RxR路径设计了机制来反制已知偏见[5]。

44.5%的RxR路径不是起点到终点的最短路径[5]。这种设计迫使智能体必须严格遵循中间指令和地标，要求智能体具备卓越的视觉-语言地面化能力，而不是仅仅依赖目标位置的先验知识。

多语言特性：
RxR是第一个多语言VLN数据集，包含了三种具有不同语系特征的语言：英语（English）、印地语（Hindi）和泰卢固语（Telugu）5。关键在于，这些指令是
由母语使用者从零开始创建的，而非通过机器翻译得到[5]。语言的差异性对于VLN至关重要，因为不同的语言在编码空间和时间信息时存在独特的模式。例如，英语部分还进一步区分了来自美国（en-US）和印度（en-IN）说话者的指令[5]。

密集时空地面化（Dense Spatiotemporal Grounding）：
RxR的一大技术贡献是为每条指令增强了密集时空地面化 5。在数据收集过程中，标注工具记录

了指导者在说话和移动时的
3D姿态，并将完整的姿态轨迹与指令中的每个单词进行时间对齐[5]。这种细粒度的对齐，为智能体提供了极其丰富的监督信号，明确了指令中的每一个语言地标在3D空间中何时何地被提及。这种信息远比R2R中缺乏的精细地面化信息要丰富[5]。

RxR的出现，将VLN的关注点从粗粒度的目标预测，转向了对细粒度动作推理和路径依从性的要求。成功指标也相应地从简单的终点成功率（End-point Success Rate, ESR）演变为强调路径质量的路径长度加权成功率（Success weighted by Path Length, SPL）。

Table 1: 奠基性VLN基准比较

特性	**R2R (Room-to-Room)**	**RxR (Room-Across-Room)**	**VLN-CE (连续环境导航)**
年份/会议	2018 (CVPR Spotlight) [6]	2020 (EMNLP) [5]	2020/2021 (ECCV/CVPR) [8]
导航空间	离散（图节点）	离散（图节点）	连续（度量动作） [7]
模拟器平台	Matterport3D 模拟器 [3]	Matterport3D 模拟器 [5]	Habitat 平台 [9]
指令规模	22K 条指令	126K 条指令 [5]	支持 R2R/RxR 指令
多语言性	仅英语 [5]	3 种语言 (英语, 印地语, 泰卢固语) [5]	支持多语言指令 (通过 RxR)
路径偏见	最短路径先验 (平均 9.4m) [5]	反偏见 (平均 14.9m, 44.5% 非最短路径) [5]	非受限路径规划
地面化细节	缺乏细粒度地面化 [5]	密集时空地面化 (单词到 3D 姿态对齐) [5]	连续路径的密集地面化

3.2 细粒度 R2R 与子指令意识

在RxR之前，细粒度R2R数据集（EMNLP 2020）已开始探索对R2R指令的精细化。该数据集将指令

分解为子指令，并将每个子指令与其对应的视点路径配对[10]。这一工作旨在强调分步骤地遵循指令，为后来RxR的密集时空地面化概念奠定了基础，并强化了对智能体步进式指令依从性的要求。

IV. 转向连续环境（VLN-CE）

4.1 连续环境视觉与语言导航 (VLN-CE)

尽管R2R和RxR在理解语言和视觉对应关系方面取得了进展，但它们仍受限于离散的导航图结构。这种离散性与现实世界中机器人必须执行的流畅、度量级的控制动作存在显著差异[7]。

VLN-CE（Vision and Language Navigation in Continuous Environments）的引入，解决了从理想化的图导航向逼真的连续动作控制的过渡问题[9]。

任务与平台：
VLN-CE是一个指令引导的导航任务，其特点是具有真实的环境（通过Habitat平台中的Matterport3D数据实现）和非受限的智能体导航 7。智能体接收第一人称（自我中心）视角和人类生成的指令，并必须输出
度量控制动作，例如："向前移动 0.25米"，"向左转 15 度"[7]。VLN-CE通过消除原始VLN任务的离散假设，将模拟智能体更接近真实世界的机器人操作环境[7]。VLN-CE通常使用Habitat平台进行实现，该平台以其高性能和逼真的模拟能力而闻名[9]。

4.2 VLN中的持续学习（CVLN）

将VLN智能体部署到现实世界中，必须考虑环境并非静态不变这一事实——建筑可能会随着时间推移进行翻新或布局改变[11]。因此，研究人员提出了

持续视觉与语言导航（Continual Vision-and-Language Navigation, CVLN）范式[11]。

CVLN旨在训练和评估VLN智能体在不丢失先前环境知识的前提下，学习和适应新环境的能力（即持续学习，CL）。它涵盖了两种设置来处理不同类型的指令：基于初始指令的CVLN（I-CVLN）和基于对话的CVLN（D-CVLN）[11]。

持续学习机制：

CVLN中的一个主要挑战是顺序决策过程固有的灾难性遗忘问题。为了解决这个问题，研究提出了简单而有效的排练机制（rehearsal mechanism）基线方法，例如：

1. 困惑度重放（**Perplexity Replay, PerpR**）：根据情节难度选择重放情节，以增强对困难场景的记忆 [11]。
2. 情节自我重放（**Episodic Self-Replay, ESR**）：在训练过程中存储并重新访问单个情节步骤中的动作逻辑值（action logits），以细化学习 [11]。

实验表明，这些基于排练的方法优于现有的持续学习方法，能够更有效地处理CVLN中的序列决策需求 [11]。CVLN是连接实验室模型与动态操作机器人之间的桥梁，验证了当前研究重心已从静态环境下的性能表现，转向动态鲁棒性和终身学习能力。

V. VLN智能体架构的演进

VLN智能体架构的演进是不断克服数据稀缺性和状态跟踪两大核心挑战的历史。

5.1 奠基性架构：序列到序列（Seq2Seq）与说话者-跟随者模型

早期的VLN模型以标准Seq2Seq网络为基础，用于执行指令。然而，由于数据的获取成本高昂，数据稀缺性一直是VLN领域的一个严峻挑战 [12]。

说话者-跟随者（Speaker-Follower）范式：
这一开创性工作旨在通过数据增强来解决数据稀缺性问题 13。

1. 说话者模型（ ）的训练：说话者模型首先在真实路径和人类描述上进行训练 [13]。

2. 数据增强：该模型随后被用于对随机采样的路径 进行逆向

翻译，生成合成的指令 [12]。这些合成的路线-指令对被用来扩充跟随者模型的训练数据集，从而允许跟随者模型学习如何在新的路线上导航 [13]。

3. 语用推理（**Pragmatic Inference**）：在推断阶段，说话者模型还能帮助跟随者解释模糊的指令，选择最佳路径。跟随者通过选择最有可能导致说话者模型产生给定描述的路径

来实现语用推理（即 ）[13]。这种机制允许智能体进行反事实推理，并利用全局语境来修正跟随者路径中的错误 [13]。

5.2 基于 Transformer 的架构（VLN-BERT 时代）

随着 Transformer 架构在自然语言处理和视觉语言任务中取得巨大成功，VLN研究开始尝试将其

引入。然而，标准的 Vision-and-Language BERT 架构难以直接适应VLN任务的固有复杂性。主要原因在于VLN的序列决策过程和部分可观察性（POMDP）要求模型具备历史依赖的注意力和决策能力 [2]。

VLN-BERT (CVPR 2021)：
VLN-BERT模型通过引入一个循环机制（Recurrent Function）来解决历史依赖问题，从而成为一个时间感知的循环 BERT 模型 2。

- 状态维护：该模型使 BERT 能够维护跨模态状态信息（ ），该状态信息在导航的每一步都包含着视觉和语言的历史上下文 [2]。
- 性能提升：这种结构有效地解决了 VLN 中固有的 POMDP 挑战，使其能够取代更复杂的编码器-解码器模型，并在 R2R 和 REVERIE 等数据集上取得了最先进的成果 [2]。

VLN-BERT 的成功奠定了基于 Transformer 的 VLN 架构的基础，后续工作如 Airbert [8]、History Aware Multimodal Transformer [8] 和 Episodic Transformer [8] 均在此基础上进一步优化了上下文意识和记忆机制。这一架构上的转变，反映了领域对有效状态跟踪和跨模态注意力机制的不断精细化。

Table 2: VLN智能体架构分类

时代	关键模型范式	解决的核心 **VLN** 挑战	代表性技术/模型
奠基性 **(2018-2019)**	编码器-解码器/RL	动作预测，视觉地面化，数据稀缺性	Seq2Seq, 说话者-跟随者 (语用推理/数据增强) [12]
Transformer (2020-2022)	循环视觉-语言 BERT (V&L BERT)	历史依赖性 (POMDP)，鲁棒的跨模态注意力	VLN-BERT (循环机制) [2], Airbert [8], 情节 Transformer [8]
泛化 **(2023-至今)**	基础模型增强/预训	对未见环境的泛化	RAM 范式 (无模拟

	练	(OOD)，视觉-语言对齐	器重写) [14], BEVBert [8]
推理与规划 (2024-至今)	大语言模型 (LLM) 编排	复杂、粗粒度规划，高层推理，连续动作执行	RAGNav (RAG/PEOA) [15], NavGPT-2 [8], VLN-R1 (LVLM-to-Action) [16]

VI. 基础模型在现代 VLN 中的应用

近年来，大规模语言模型（LLMs）和视觉-语言模型（LVLMs）的崛起，使VLN研究焦点从针对特定任务的模型优化，转向开放域指令的泛化 [17]。基础模型凭借其强大的世界知识和复杂推理能力，成为解决VLN中高层规划问题的核心工具。

6.1 LLM在高层规划与指令精炼中的作用

LLM被广泛用于生成更具多样性和风格的指令，以提高智能体的适应性。例如，通用场景适应VLN (General Scene Adaptation for VLN, GSA-VLN) 任务就利用LLM设计了三阶段指令编排流程，并采用角色扮演技术将指令改写成不同的说话风格，模拟单个用户在家庭机器人应用中可能出现的持续且一致的语言特征 [18]。

6.2 检索增强生成（RAG）在导航中的应用（RAGNav）

RAGNav方法旨在解决连续环境（VLN-CE, RxR-CE）中智能体知识不足和避障困难的问题 [15]。该方法利用检索增强生成（RAG）机制进行精确的导航规划，并引入了一种提示增强避障（PEOA）策略来提高鲁棒性。

规划组件（RAG）：
RAGNav构建了一个包含区域和对象通用概念知识的导航知识库 15。

1. 指令分析：一个外部LLM组件（ ）分析输入指令（ ），提取对象参数（ ）和房间参数（ ）[15]。

2. 知识检索：通过 RAG 机制，利用提取的参数（ ）和指令（

) 从知识库中检索最相关的知识 (

) [15]。

3. 粗粒度目标生成：检索到的知识 ( ) 与场景图描述 (

) 一同输入给规划 LLM ( ),

用于生成准确的粗粒度导航目标 ( ) [15]。

执行组件(PEOA):
RAGNav引入的提示增强避障(PEOA)策略是其提高鲁棒性的关键 15。

- 动作生成: PEOA 模块利用结构化提示，将粗粒度目标 ( )、场

景图描述 ( ) 和深度图(DepthMap)作为输入，指导LLM生成

细粒度的动作执行 ( ) [15]。

- 性能：PEOA 策略显著增强了智能体的避障能力，在R2R-CE和RxR-CE数据集上，成功率至少提高了2%和2.32% [15]。

RAGNav 模型展示了现代 VLN 架构的分层控制趋势：LLM 充当高级、符号推理的规划者，负责知识整合和目标设定，而 PEOA 等专业模块则处理低级、度量执行和即时避障。

6.3 LVLM到动作模型

最新的研究正在探索如何利用大规模视觉-语言模型（LVLMs）直接驱动导航。例如，VLN-R1 框架利用 LVLM 将自我中心视频流直接转换为连续导航动作，采用基于 GRPO 的训练机制 [16]。这些方法旨在将整个导航过程视为一个复杂的跨模态序列生成任务（如 NavGPT-2 [8]），从而释放其固有的导航推理能力。

VII. 关键研究前沿与泛化挑战

7.1 VLN 中的泛化挑战

VLN 面临的核心挑战是要求智能体能够在训练过程中从未见过的环境中进行零样本（Zero-shot）执行 [3]。数据稀缺性是阻碍泛化的主要因素 [14]。传统方法试图通过利用额外的模拟器数据或网络收集的图像来改进泛化，但这往往导致多样性有限或数据噪声大 [14]。

7.2 无模拟器数据增强：RAM 范式

Rewriting-driven AugMentation (RAM) 范式是 VLN 领域的一个创新方法，它提供了一种无模拟器且省力的数据增强方案，通过重写人类标注的训练数据来生成未见的观测-指令对[14]。RAM 范式的引入，标志着 VLN 数据集多样性的扩展开始摆脱对昂贵的 3D 环境重建的依赖，转向

生成式合成。

RAM 范式巧妙地结合了基础模型（VLM、LLM 和 T2IM）来实现数据合成：

模块 1：富对象观测重写（**Object-Enriched Observation Rewriting**）：

- 生成场景描述：VLMs 和 LLMs 协同工作，从原始场景描述（ ）

  中推导出包含新对象和空间布局信息的重写场景描述（ ）[14]。

- 合成新观测：重写后的描述（ ）被输入给文本到图像生成模型（T2IMs，例如 MultiDiffusion）。通过全景到视点生成策略（panorama-to-view generation），

  T2IM 合成新的、多样化的观测图像（ ），这个过程通过全景图离散化技术自然地确保了视点之间的一致性 [14]。

模块 **2**：观测对比指令重写（**Observation-Contrast Instruction Rewriting**）：

- 指令对齐：LLMs 被要求通过推理原始观测与新合成观测之间的差异，来生成与新观测精确

  对齐的重写指令（ ）[14]。
- 指令增强：LLMs 还会被提示改变动作描述和语法结构，使得重写后的指令在表述上更具信息性和灵活性 [14]。

RAM 范式成功地利用基础模型的生成能力，创造出分布外（Out-of-Distribution, OOD）数据，这对于提高智能体对未见环境的泛化能力具有重要意义。

7.3 适应性与持续学习范式

通用场景适应 (GSA-VLN)：
这一新任务场景提出，现实世界中的导航机器人往往在一个相对固定的环境中长期运行，拥有固定的物理布局和特定的指令风格 18。GSA-VLN 要求智能体在一个特定场景中执行指令的同时，能够持续适应这个场景以提高长期性能 18。GSA-VLN 的目标是解决 OOD 数据不足和每个场景的指令数量及风格多样性有限的问题，强调智能体对特定环境的持续适应性。
持续学习方法：
针对 CVLN，除了 PerpR 和 ESR 等排练机制 11 之外，未来研究必须探索更高效、更具选择性的适应机制，以确保智能体在不断演进的环境中保持知识的稳定性，同时高效地吸收新的环境信息。

VIII. 动态环境与具身交互

8.1 人类感知VLN（HA-VLN）

传统的 VLN 框架通常依赖于环境是静态的这一简化假设 [19]。然而，真实的室内环境往往充满动态变化和移动的人类活动。

HA-VLN 的引入：
人类感知视觉与语言导航（HA-VLN）通过引入动态人类活动来扩展传统 VLN 19。

- **HA3D** 模拟器：结合了动态人类活动与 Matterport3D 数据集，模拟了在人类活动空间中的导航 [19]。
- **HA-R2R** 数据集：基于 R2R 扩展了人类活动描述，用于评估智能体在拥挤室内空间中的鲁棒性 [19]。

HA-VLN 带来的挑战包括多人类动态和部分可观察性，要求智能体不仅要遵循指令，还要预测和避开动态障碍物 [19]。这标志着 VLN 任务从“单独服从指令”转向了“社交机器人”所需的环境感知和安全规划能力。

8.2 多智能体通信与新兴对话

另一前沿研究方向是探索多智能体设置下的 VLN 任务，例如“游客”（Tourist，具身智能体）和“导游”（Guide，拥有全局视野的智能体）的协作导航任务 [21]。

- 多轮对话：该任务要求智能体通过多轮的新兴对话实现协作，引导游客到达目标地点 [21]。
- 新兴语言：通过协同多智能体强化学习方法，智能体可以从零开始学习生成和理解一种新兴语言，并制定实现长期目标的最佳对话决策 [21]。

这项工作超越了固定、预标注的人类指令，关注于智能体能否在任务导向的目标下，自发地发展出沟通和协作的能力，从而实现更具交互性和灵活性的具身智能系统 [21]。

8.3 部署中的安全、伦理与计算挑战

随着 VLN 研究转向更真实的连续和动态环境，部署挑战也浮出水面 [22]。

- 实时推理约束和计算需求：复杂的 LVLMs 和 RAG 系统在边缘计算设备上实现实时动作生成，面临巨大的计算复杂性和延迟挑战 [22]。
- 多模态动作表示与安全保障：尤其是在动态的、有人类存在的环境中，必须确保智能体的动作表示安全可靠，避免与人类发生碰撞或意外交互 [19]。
- 鲁棒性和伦理挑战：智能体在面对未知或预期外情况时必须保持鲁棒性。此外，在涉及人类活动的具身系统中，必须考虑隐私和伦理挑战 [22]。

这些非学术性的技术难题正成为 VLN 领域成熟和实际部署的关键瓶颈。

IX. 结论与新手研究人员建议

9.1 VLN 发展轨迹总结

视觉与语言导航（VLN）领域的发展轨迹清晰地展示了从简化到真实的渐进过程。

最初的奠基性工作 R2R [3] 在离散的、最短路径的拓扑图上建立了任务范式。随后，RxR [5] 通过引入更大的规模、多语言特性和密集的时空地面化，成功地克服了 R2R 中最短路径和语言单一性的局

限，将研究重点转移到精确的路径依从性上。

架构层面，研究从基本的 Seq2Seq 模型，通过 Speaker-Follower 范式 [13] 解决了数据稀缺性，再通过 VLN-BERT 的循环机制 [2] 解决了 POMDP 中的状态跟踪问题。

最新的研究则聚焦于将 VLN 推向现实：VLN-CE [7] 引入了连续动作空间，而 CVLN [11] 和 GSA-VLN [18] 解决了持续学习和环境适应性。RAGNav [15] 等大模型驱动的系统，则实现了高层规划与低层控制的解耦，并利用 RAM 范式 [14] 突破了模拟器环境多样性的限制，转向生成式数据合成。

VLN 的目标已经从最初的“在模拟器中找到目标”，演变为“在动态、有人类参与的开放域环境中安全地进行长时期的、协作性的导航”。

9.2 关键开放挑战与未来研究方向

尽管取得了显著进展，VLN 仍面临多项关键挑战，这些挑战构成了未来研究的沃土：

1. 基础模型数据合成的保真度与噪声：RAM 范式虽然具有前景，但 T2IMs 生成的数据中固有的噪声和偏差如何有效地被缓解，以及如何确保合成环境的逼真度能支持高可靠的训练，仍需深入研究 [14]。
2. 鲁棒的 **Sim2Real** 迁移：在 HA-VLN 等动态、人类感知环境中，模拟器（如 Habitat）与真实世界机器人部署之间的性能差距仍然巨大 [19]。需要更先进的领域随机化和适应性技术来提升迁移能力。
3. 多模态知识推理的规模化：当前的 RAG 系统依赖于预先构建的知识库。如何构建更通用、更灵活的导航知识基，以支持对复杂指令（例如涉及否定、反事实或长期记忆）的鲁棒、多语言推理。
4. 灾难性遗忘与持续适应：尽管 PerpR/ESR [11] 有效，但对于长期、大规模的持续适应（如 CVLN 或 GSA-VLN [18]），需要开发更具理论基础和计算效率的排练和知识整合机制。

9.3 对初学者的建议

对于希望进入 VLN 领域的新手研究人员，建议将研究重点放在以下交叉领域，这些领域当前处于研究的前沿和瓶颈：

1. 探索 **LLM** 在 **VLN** 中的分层控制作用：关注如何优化 LLM 的高层规划（例如，改进 RAG 知识检索的准确性 [15]）与低层执行（例如，提升 PEOA 策略在复杂环境中的鲁棒性 [15]）之间的衔接机制。
2. 深耕生成式数据增强和泛化：研究如何改进 RAM 范式 [14]，例如开发新的损失函数或训练策

略，以更好地管理 T2IM 引入的合成数据噪声，从而提高智能体在未见环境中的零样本性能。

3. 着手动态或协作性任务：关注 HA-VLN [19] 或多智能体新兴通信 [21] 等任务，这些领域涉及安全关键的规划、预测动态障碍物和社交交互，是 VLN 迈向实用性的关键所在。

Works cited

1. Vision-Language Navigation with Embodied Intelligence: A Survey - arXiv, accessed October 1, 2025, https://arxiv.org/html/2402.14304v1
2. VLN BERT: A Recurrent Vision-and-Language ... - CVF Open Access, accessed October 1, 2025, https://openaccess.thecvf.com/content/CVPR2021/papers/Hong_VLN_BERT_A_Recurrent_Vision-and-Language_BERT_for_Navigation_CVPR_2021_paper.pdf
3. Bring Me A Spoon | Matterport3D Simulator and Room-to-Room ..., accessed October 1, 2025, https://bringmeaspoon.org/
4. Vision-and-Language Navigation: Interpreting Visually-Grounded Navigation Instructions in Real Environments - CVF Open Access, accessed October 1, 2025, https://openaccess.thecvf.com/content_cvpr_2018/papers/Anderson_Vision-and-Language_Navigation_Interpreting_CVPR_2018_paper.pdf
5. Room-Across-Room: Multilingual Vision-and ... - ACL Anthology, accessed October 1, 2025, https://aclanthology.org/2020.emnlp-main.356.pdf
6. [1711.07280] Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments - arXiv, accessed October 1, 2025, https://arxiv.org/abs/1711.07280
7. VLN-CE - Jacob Krantz, accessed October 1, 2025, https://jacobkrantz.github.io/vlnce/
8. zhangyuejoslin/VLN-Survey-with-Foundation-Models - GitHub, accessed October 1, 2025, https://github.com/zhangyuejoslin/VLN-Survey-with-Foundation-Models
9. jacobkrantz/VLN-CE: Vision-and-Language Navigation in Continuous Environments using Habitat - GitHub, accessed October 1, 2025, https://github.com/jacobkrantz/VLN-CE
10. Code and data of the Fine-Grained R2R Dataset proposed in the EMNLP 2021 paper Sub-Instruction Aware Vision-and-Language Navigation - GitHub, accessed October 1, 2025, https://github.com/YicongHong/Fine-Grained-R2R
11. Continual Vision-and-Language Navigation - arXiv, accessed October 1, 2025, https://arxiv.org/html/2403.15049v2
12. arXiv:1911.07308v4 [cs.CV] 17 Mar 2025, accessed October 1, 2025, https://arxiv.org/pdf/1911.07308
13. Speaker-Follower Models for Vision-and-Language Navigation, accessed October 1, 2025, http://papers.neurips.cc/paper/7592-speaker-follower-models-for-vision-and-language-navigation.pdf
14. Unseen from Seen: Rewriting Observation-Instruction Using ... - arXiv, accessed October 1, 2025, https://arxiv.org/abs/2503.18065

15. Enhancing Large Language Models with RAG for Visual Language ..., accessed October 1, 2025, https://www.mdpi.com/2079-9292/14/5/909
16. [2506.17221] VLN-R1: Vision-Language Navigation via Reinforcement Fine-Tuning - arXiv, accessed October 1, 2025, https://arxiv.org/abs/2506.17221
17. VLN-R1: Vision-Language Navigation via Reinforcement Fine-Tuning - arXiv, accessed October 1, 2025, https://arxiv.org/html/2506.17221v1
18. General Scene Adaptation for Vision-and-Language Navigation - arXiv, accessed October 1, 2025, https://arxiv.org/html/2501.17403v1
19. Human-Aware Vision-and-Language Navigation: Bridging Simulation to Reality with Dynamic Human Interactions, accessed October 1, 2025, https://proceedings.neurips.cc/paper_files/paper/2024/file/d8087deaf34bb07ddd41c65f8a9fe9b5-Paper-Datasets_and_Benchmarks_Track.pdf
20. [2503.14229] HA-VLN: A Benchmark for Human-Aware Navigation in Discrete-Continuous Environments with Dynamic Multi-Human Interactions, Real-World Validation, and an Open Leaderboard - arXiv, accessed October 1, 2025, https://arxiv.org/abs/2503.14229
21. Emergent Language based Dialog for Collaborative Multi-agent Navigation | OpenReview, accessed October 1, 2025, https://openreview.net/forum?id=WsHaBoucSG
22. Vision-Language-Action Models: Concepts, Progress, Applications and Challenges - arXiv, accessed October 1, 2025, https://arxiv.org/html/2505.04769v1