ChatGPT

2020年以来视觉语言导航（VLN）研究综述

视觉语言导航（Vision-and-Language Navigation, VLN）要求智能体根据自然语言指令在未知环境中移动。自2020年以来，研究者在数据集扩展、模型结构、学习范式和应用场景方面取得了显著进展，本文综述这些发展并按主题归纳代表性论文。

核心任务与数据集演变

R2R 及衍生数据集

- **Room-to-Room (R2R)** – 最早的室内VLN数据集，由 Matterport3D 场景和短指令组成。后续研究通过延长路径、增添复杂指令等方式增强任务难度。
- **R4R / XL-R2R / FGR2R / RxR** – 为解决R2R简单路径和指令不够细粒的不足，研究者推出一系列扩展：R4R 增加路径长度和视角变换；XL-R2R 和 FGR2R 提供更复杂或细粒度指令；RxR（Room-across-Room）含有跨房间长路径，并支持多语言，丰富了跨语言导航任务 [1] 。
- **Landmark-RxR、R6R/R8R、BnB** – 近期出现的更长路径和真实世界场景扩展。Landmark-RxR 在 RxR 上提供观测标记，R6R/ R8R 将路径长度扩展至六个或八个房间，BnB 将AirBnB房源图像用于导航 [1] 。
- **VLN-CE / RoboVLN** – 传统R2R采用离散动作空间；VLN-CE引入连续控制和物理碰撞模型，RoboVLN利用机器人仿真实现连续控制，挑战现实部署 [2] 。

户外与互动任务

- **TOUCHDOWN / StreetLearn / StreetNav** – 将任务迁移到城市街景，指令描述道路和地标 [2] 。
- **REVERIE、AI2-THOR** – 从纯路径导航转向目标搜索或交互任务，如REVERIE要求在目标房间与物品互动。AI2-THOR结合虚拟机器人抓取任务，探索人与物的交互 [1] 。
- **GSA-VLN 数据集（2025）** – 新近提出的**General Scene Adaptation**任务要求代理在某个场景内长期学习并适应；作者使用大型语言模型生成多样指令，并设计了记忆增强模型GR-DUET，显著提升长期适应性 [3] 。

视频与多模态指令数据

- **YouTube-VLN（2023）** – 利用房屋参观视频构建大规模预训练数据集。论文通过熵选择帧并使用动作感知生成器产生指令，提出布局推理预训练任务，提升模型在R2R与REVERIE中的泛化能力 [4] 。
- **VLN-MP（2024）** – 传统指令仅含文本。该工作引入**多模态提示**，在指令中插入目标或中间视觉提示图片，并设计处理不同类型图像的模块。实验表明适当的视觉提示可以显著提升导航成功率，同时兼容纯文本输入 [5] [6] 。

模型方法：多模态融合与视觉-语言对齐

预训练与跨模态融合

早期VLN模型通常基于循环网络结合注意力机制。2020年以来，研究者引入Transformer和预训练框架，利用大规模图文数据提高表示能力。

- **VLN-BERT (CVPR 2020)** – 提出视觉-语言Transformer预训练模型。首先在大规模图文配对数据上预训练，学习评估指令与一系列全景图像之间的兼容性，然后在VLN任务上微调，成功率提高约4% [7] 。

- **PREVALENT (CVPR 2020)** – 首次采用“预训练–微调”范式，使用图片-文字-动作三元组自监督训练通用表示。模型将R2R成功率从47%提升至51%，并能迁移到对话导航等任务 [8]。
- **SOAT (NeurIPS 2021)** – 为解决对象描述和场景理解不足，引入场景分类网络和物体检测器生成双视觉特征，通过Transformer对齐语言描述与场景/对象，使得含有较多对象指令的样本性能提升更大 [9]。

对齐与对比学习

- **RCM & SIL (CVPR 2019)** – 使用**强化跨模态匹配**(Reinforced Cross-modal Matching)。一个匹配评估器计算路径与指令的匹配程度作为内在奖励，强化学习代理执行策略；同时自监督模仿学习(SIL)通过回放历史轨迹模仿过去的好决策，提升在未知环境的泛化 [10] [11]。
- **DELAN (2024)** – 提出**双层对齐**(Dual-Level Alignment)，分别对指令历史与轨迹观察进行对比学习，然后再融合，以解决训练信号稀缺问题；在R2R、R4R和RxR等数据集上提升成功率 [12]。
- **Contrastive Instruction-Trajectory Learning (CITL, AAAI 2022)** – 采用粗粒度和细粒度对比学习，对完整轨迹与指令对齐，同时使用时序对比强化局部连续性，并通过样本重加权处理数据偏差，提升泛化 [13] [14]。
- **Bayesian Optimization for Fine-Grained Alignment (2024)** – 发现负例采样影响对齐质量，使用贝叶斯优化生成视觉负例，改进跨模态嵌入，使R2R和REVERIE任务上微调模型取得提升 [15]。

历史记忆与层次规划

- **HAMT (History Aware Multimodal Transformer, NeurIPS 2021)** – 使用层次视觉Transformer存储全部历史全景观察和动作，通过预训练任务（动作预测、空间关系预测）和强化学习微调，实现长序列决策，并在R2R、RxR等任务上取得新的SOTA [16] [17]。
- **DUET (Dual-scale Graph Transformer, CVPR 2022)** – 动态构建场景的拓扑图用于全局规划，同时编码当前视野的细粒度局部信息。模型利用图Transformer在全局和局部尺度间交替传播信息，实现长距离规划和精细语言 grounding，在R2R和REVERIE上成功率显著提高 [18]。
- **GR-DUET (GSA-VLN, 2025)** – 在泛化适应任务中，将记忆图保存在代理内部，并结合全局-局部双尺度规划，促进环境长期适应 [3]。

能量模型与物理场景适应

- **Energy-Based Policy (ENP, 2024)** – 提出以能量函数刻画专家策略的**能量模型**，通过最小化前向散度和噪声对比学习训练，使低能量对应专家可能采取的状态-动作对 [19]。该方法适用于连续场景如 REVERIE，并可结合 CLIP 特征改进物体定位 [20]。

多模态提示与外部视觉语言模型

- **VLN-MP (IJCAI 2024)** – 任务增强：在导航指令中嵌入图像提示，如目标或中间物品照片。作者设计多模态提示处理器并将其集成到现有模型中，实验证明视觉提示能显著提升导航成功率并兼容原有文本输入 [5] [6]。

此外，一些工作探索利用**CLIP**、**BLIP**等大规模视觉语言模型提取视觉语义特征以增强对齐能力；ENP证明使用CLIP特征有助于物体定位 [20]。随着**GPT-4V**等多模态大模型的普及，如何将其与VLN结合（如利用生成式模型解释和推理）成为研究趋势。

学习策略：模仿学习、强化学习与混合范式

强化学习

- **跨模态奖励设计** – RCM利用**匹配评估器**作为内在奖励，鼓励代理在视觉和语言双模态间保持一致 [10]。同时将环境评价指标（如成功率或距离）设为外在奖励，推动策略优化 [21]。

- **带搜索的策略改进** – 研究者采用beam search、回溯搜索和前瞻规划，缓解局部最优；并引入记忆网络或导航图进行长期规划 [22] 。

模仿学习与混合范式

- **自监督模仿学习 (SIL)** – 代理在未知环境中探索并记录成功轨迹，随后模仿这些轨迹以提升泛化能力 [11] 。
- **预训练-微调** – VLN-BERT和PREVALENT通过在图文数据上自监督预训练，之后在VLN任务上微调，将强化学习框架与模仿学习融合 [23] 。
- **混合优化** – CITL、DELAN等同时利用对比学习、强化学习和模仿学习，多层次对齐语言与视觉，并通过样本重加权和负例优化增强泛化。

泛化与现实部署

跨环境迁移是VLN的重要挑战。R2R模型在未见环境性能下降明显，需要应对**观测差异**、**长尾指令**和**语义漂移**。

- **环境自适应** – GSA-VLN通过持续交互让代理适应单个复杂场景，并利用记忆图和GR-DUET模型学习长期策略 [3] 。
- **多语言与跨领域** – RxR支持多语言指令，研究者使用预训练语言模型处理跨语言；YouTube-VLN利用真实房屋视频构建数据，增强视觉多样性 [4] 。
- **连续控制** – VLN-CE和RoboVLN在仿真器中引入连续动作和物理碰撞模型，探索从模拟到现实的迁移 [2] 。ENP使用能量模型处理连续动作 [19] 。

最新趋势与未来展望

1. **大模型融合与知识迁移**：将CLIP、BLIP、GPT-4V等多模态大模型用于导航，可借助其丰富的视觉语义理解和推理能力。未来研究需设计高效的知识适配与安全控制机制。
2. **多模态与多任务协同**：任务扩展至包含指令中插入图片或视频提示（VLN-MP），以及结合问答、对话等多任务，使代理能够理解复合指令并与人类互动。
3. **数据集扩展与真实部署**：GSA-VLN和YouTube-VLN等数据集提供真实布局和长序列，使模型学习更强泛化。未来需采集更多真实场景数据，并探索从模拟到物理机器人的迁移。
4. **可靠性与解释性**：研究对比学习、能量模型等方法提升对齐和稳定性；结合大模型解释能力，使导航决策更透明。
5. **跨语言与文化适应**：随着RxR等多语言数据集的出现，模型需理解不同语言和文化的指令；LLM提供的跨语言翻译与生成能力将加速这一方向。

总结

2020年以来，VLN领域从小规模室内数据和简单模型向大型预训练、多模态对齐和现实部署迈进。各种数据集扩展、预训练模型（VLN-BERT、PREVALENT）、对比学习方法（DELAN、CITL）和层次规划模型（HAMT、DUET）推动了性能的持续提升。未来的研究将依托多模态大模型与更真实的数据环境，探索跨语言、多任务的智能导航，并在解释性和安全性上不断突破，以接近真正能在现实世界中自主导航的智能体。

---

1 2 21 22 23 Vision-Language Navigation with Embodied Intelligence: A Survey
https://arxiv.org/html/2402.14304v1

3 General Scene Adaptation for Vision-and-Language Navigation | OpenReview
https://openreview.net/forum

4 YouTubeVLN.pdf
https://peihaochen.github.io/files/publications/YouTubeVLN.pdf

5 6 Why Only Text: Empowering Vision-and-Language Navigation with Multi-modal Prompts
https://www.ijcai.org/proceedings/2024/0093.pdf

7 [2004.14973] Improving Vision-and-Language Navigation with Image-Text Pairs from the Web
https://arxiv.org/abs/2004.14973

8 [2002.10638] Towards Learning a Generic Agent for Vision-and-Language Navigation via Pre-training
https://arxiv.org/abs/2002.10638

9 [2110.14143] SOAT: A Scene- and Object-Aware Transformer for Vision-and-Language Navigation
https://arxiv.org/abs/2110.14143

10 11 Reinforced Cross-Modal Matching and Self-Supervised Imitation Learning for Vision-Language Navigation
https://openaccess.thecvf.com/content_CVPR_2019/papers/Wang_Reinforced_Cross-Modal_Matching_and_Self-Supervised_Imitation_Learning_for_Vision-Language_Navigation_CVPR_2019_paper.pdf

12 DELAN: Dual-Level Alignment for Vision-and-Language Navigation by Cross-Modal Contrastive Learning
https://arxiv.org/html/2404.01994v1

13 14 Contrastive Instruction-Trajectory Learning for Vision-Language Navigation
https://cdn.aaai.org/ojs/20050/20050-13-24063-1-2-20220628.pdf

15 [2411.14811] Fine-Grained Alignment in Vision-and-Language Navigation through Bayesian Optimization
https://arxiv.org/abs/2411.14811

16 17 [2110.13309] History Aware Multimodal Transformer for Vision-and-Language Navigation
https://ar5iv.labs.arxiv.org/html/2110.13309

18 2202.11742.pdf
https://arxiv.org/pdf/2202.11742.pdf

19 20 [2410.14250] Vision-Language Navigation with Energy-Based Policy
https://ar5iv.labs.arxiv.org/html/2410.14250