LaViRA: Language-Vision-Robot Actions Translation for Zero-Shot Vision Language Navigation in Continuous Environments

LaViRA: 面向连续环境中零样本视觉语言导航的语言-视觉-机器人动作翻译

Hongyu Ding [1,*] , Ziming Xu[1,*] , Yudong Fang [2] , You Wu[1] , Zixuan Chen [1] ,

丁鸿宇 [1,*] ， 徐子明 Xu[1,*] ， 方宇东 [2] ， 吴有 Wu[1] ， 陈子轩 [1] ，

Jieqi Shi[2,†] , Jing Huo[1,†] , Yifan Zhang[3] , Yang Gao[2]

杰琦 Shi[2,†] ， 晶 Huo[1,†] ， 一帆 Zhang[3] ， 杨 Gao[2]

Abstract- Zero-shot Vision-and-Language Navigation in Continuous Environments (VLN-CE) requires an agent to navigate unseen environments based on natural language instructions without any prior training. Current methods face a critical trade-off: either rely on environment-specific waypoint predictors that limit scene generalization, or underutilize the reasoning capabilities of large models during navigation. We introduce LaViRA, a simple yet effective zero-shot framework that addresses this dilemma by decomposing action into a coarse-to-fine hierarchy: Language Action for high-level planning, Vision Action for middle-level perceptual grounding, and Robot Action for low-level control. This modular decomposition allows us to leverage the distinct strengths of different scales of Multimodal Large Language Models (MLLMs) at each stage, creating a system that is powerful in its reasoning, grounding and practical control. LaViRA significantly outperforms existing state-of-the-art methods on the VLN-CE benchmark, demonstrating superior generalization capabilities in unseen environments, while maintaining transparency and efficiency for real-world deployment. Project page: https: //robo-lavira.github.io/lavira-zs-vln/

摘要——连续环境中的零样本视觉语言导航 (VLN-CE) 要求智能体基于自然语言指令在未见过的环境中导航，且无需任何先验训练。现有方法面临关键权衡: 要么依赖环境特定的路径点预测器，限制了场景泛化能力；要么在导航过程中未充分利用大型模型的推理能力。我们提出 LaViRA，一种简单而有效的零样本框架，通过将动作分解为粗到细的层级: 语言动作用于高层规划，视觉动作用于中层感知定位，机器人动作用于低层控制。该模块化分解使我们能够在每个阶段充分发挥多模态大型语言模型 (MLLMs) 不同尺度的优势，构建出在推理、定位和实际控制方面均具备强大能力的系统。LaViRA 在 VLN-CE 基准上显著优于现有最先进方法，展示了在未见环境中的卓越泛化能力，同时保持了透明性和效率，适合实际部署。项目主页:https://robo-lavira.github.io/lavira-zs-vln/

I. INTRODUCTION

一、引言

Vision-and-Language Navigation (VLN) presents the challenge of grounding natural language instructions within visual observations to enable an embodied agent to navigate through previously unseen environments [1]. Early VLN research was primarily conducted in discrete, graph-based settings where navigation is simplified to selecting paths between predefined nodes. To bridge the gap to the real world, Vision-and-Language Navigation in

Continuous Environments (VLN-CE) [2] was introduced, removing the reliance on connectivity graphs and forcing agents to contend with realistic challenges like continuous visual perception and fine-grained motor control.

视觉语言导航 (VLN) 面临的挑战是将自然语言指令与视觉观察相结合，使具身智能体能够在之前未见过的环境中导航 [1]。早期 VLN 研究主要在离散的图结构环境中进行，导航简化为在预定义节点间选择路径。为缩小与现实世界的差距，提出了连续环境中的视觉语言导航 (VLN-CE)[2]，取消了对连通图的依赖，迫使智能体应对连续视觉感知和细粒度运动控制等现实挑战。

The recent success of large language models (LLMs) and multimodal large language models (MLLMs) has inspired a new frontier: zero-shot VLN-CE [3]-[6], where an agent navigates without any environment-specific training. Two main paradigms have emerged from this research. Waypoint

大型语言模型 (LLMs) 和多模态大型语言模型 (MLLMs) 的最新成功激发了一个新领域: 零样本 VLN-CE[3]-[6]，即智能体无需任何环境特定训练即可导航。该领域出现了两大主要范式。路径点

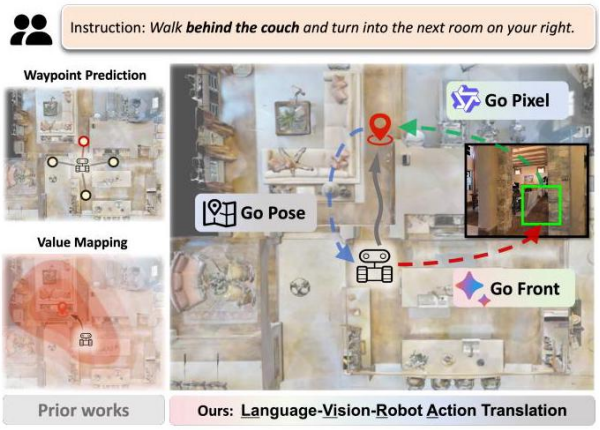

Fig. 1: Prior methods rely on pre-trained waypoint prediction or value mapping with limited online planning. Our LaViRA framework instead decomposes navigation into language-level planning (”Go Front”), vision-level grounding (”Go Pixel”), and robot-level control (”Go Pose”), fully leveraging MLLMs reasoning for efficient coarse-to-fine decision-making.

图 1: 先前方法依赖预训练的路径点预测或价值映射，在线规划能力有限。我们的 LaViRA 框架则将导航分解为语言层规划 (“向前走”)、视觉层定位 (“看像素”) 和机器人层控制 (“调整姿态”)，充分利用 MLLMs 的推理能力，实现高效的粗到细决策。

prediction with large model reasoning. These methods combine a pre-trained waypoint predictor with an LLM or MLLM [3], [4]. The predictor proposes navigable waypoints in the agent's surroundings, and the large model reasons over these discrete options to select the best next step. This design leverages the large model's high-level planning abilities but comes at a cost: the waypoint predictor requires separate pre-training and struggles to generalize to novel, unseen environments. Value mapping with vision language models. An alternative approach discards the waypoint predictor and instead uses a vision-language model (e.g., BLIP [7]) to generate a heatmap of semantic relevance over the visual scene. The agent navigates towards the highest-scoring region. While this avoids the need for predictor training, it typically relegates the powerful large models to an offline instruction-parsing role, leaving the vision-language model to handle online navigation alone. This underutilizes the large model's sophisticated reasoning and decision-making capabilities.

预测结合大型模型推理。这类方法将预训练的路径点预测器与 LLM 或 MLLM 结合 [3]，[4]。预测器提出智能体周围可导航的路径点，大型模型在这些离散选项上进行推理，选择最佳下一步。该设计利用了大型模型的高层规划能力，但代价是路径点预测器需要单独预训练，且难以泛化到新颖未见环境。基于视觉语言模型的价值映射。另一种方法舍弃路径点预测器，改用视觉语言模型 (如 BLIP[7]) 生成视觉场景的语义相关热图，智能体朝得分最高区域导航。虽然避免了预测器训练，但通常将强大的大型模型限制为离线指令解析角色，在线导航则由视觉语言模型单独承担，未充分利用大型模型复杂的推理和决策能力。

These two paradigms reveal a fundamental trade-off. Waypoint-based methods excel at high-level reasoning but are constrained by a separate, inflexible waypoint prediction module. Value-mapping methods are more perceptually grounded but lack dynamic, high-level reasoning during navigation. This leads us to a simple but critical question:

这两种范式揭示了一个根本权衡。基于路径点的方法擅长高层推理，但受限于独立且不灵活的路径点预测模块。价值映射方法感知基础更强，但导航过程中缺乏动态高层推理。这引出了一个简单而关键的问题:

Can we design a purely zero-shot VLN-CE framework that (1) removes the dependency on a pre-trained waypoint predictor and (2) fully harnesses the reasoning abilities of MLLMs for navigation decision making?

我们能否设计一个纯零样本的 VLN-CE 框架，(1) 取消对预训练路径点预测器的依赖，(2) 充分利用多模态大语言模型 (MLLMs) 的推理能力来进行导航决策?

We answer this question with LaViRA: Language-Vision-Robot Action Translation. Our key insight is to decompose navigation into a coarse-to-fine, multi-stage action space, progressively refined from language to vision to robot control. Rather than requiring a single model to produce low-level controls directly, we allocate each stage to a model scale that best matches its reasoning or perceptual demands, allowing the system to exploit the

* Equal Contribution, † Corresponding Author

* 贡献相等，† 通讯作者

[1] Hongyu Ding, Ziming Xu, You Wu, Zixuan Chen and Jing Huo are with the School of Computer Science, Nanjing University, China. Emails: {hongyuding, zimingxu, you}@smail.nju.edu.cn, {chenzx, huojing}@nju.edu.cn

[1] 丁鸿宇、徐子明、吴有、陈子轩和霍晶均来自南京大学计算机科学学院，中国。电子邮件:{hongyuding，zimingxu, you}@smail.nju.edu.cn，{chenzx, huojing}@nju.edu.cn

[2] Yudong Fang, Jieqi Shi and Yang Gao are with the School of Intelligence Science and Technology, Nanjing University, China. Emails: 231880023@smail.nju.edu.cn, isjieqi@nju.edu.cn, gaoy@nju.edu.cn

[2] 方宇东、石洁琦和高扬来自南京大学智能科学与技术学院, 中国。电子邮件:231880023@smail.nju.edu.cn, isjieqi@nju.edu.cn, gaoy@nju.edu.cn

[3] Yifan Zhang is with the Institute of Automation, Chinese Academy of Sciences, China. Emails: yfzhang@nlpr.ia.ac.cn

[3] 张一帆来自中国科学院自动化研究所，中国。电子邮件:yfzhang@nlpr.ia.ac.cn

complementary strengths of different models.

我们用 LaViRA(语言-视觉-机器人动作转换) 回答了这个问题。我们的核心见解是将导航分解为一个由粗到细的多阶段动作空间，逐步从语言到视觉再到机器人控制进行细化。我们不要求单一模型直接产生低级控制，而是将每个阶段分配给最适合其推理或感知需求的模型规模，使系统能够发挥不同模型的互补优势。

1) Language Action: A powerful MLLM acts as a high-level planner, analyzing the instruction, history, and current observation to produce a coarse strategic decision, such as which general direction to head, whether to backtrack, or when to stop.

1) 语言动作: 一个强大的多模态大语言模型作为高级规划者，分析指令、历史和当前观察，生成粗略的战略决策，如前进的大致方向、是否回溯或何时停止。

2) Vision Action: A smaller, efficient MLLM takes this high-level decision and grounds it in the visual scene, identifying a specific object or region to move towards.

2) 视觉动作: 一个较小且高效的多模态大语言模型接收该高级决策，并将其落地到视觉场景中，识别具体的目标物体或区域以移动。

3) Robot Action: A simple, rule-based controller executes the low-level movement to the identified visual target.

3) 机器人动作: 一个简单的基于规则的控制器执行到指定视觉目标的低级运动。

This hierarchical decomposition offers three main advantages. First, it is purely zero-shot, removing the dependency on pre-trained waypoint predictors. Second, it fully engages MLLM reasoning at multiple granularities, from high-level planning to middle-level perceptual grounding and then to low-level control. Third, its modular design ensures transparency and practicality, enabling straightforward adaptation to both simulated and real-world robots.

这种分层分解带来三大优势。首先，它完全零样本，消除了对预训练路径点预测器的依赖。其次，它在多个粒度层面充分调动多模态大语言模型的推理能力，从高级规划到中级感知落地再到低级控制。第三，其模块化设计确保了透明性和实用性，便于在模拟和真实机器人上直接应用。

Our contributions are as follows:

我们的贡献如下:

- We propose a general action decomposition strategy for zero-shot VLN-CE, separating navigation into language-level planning, vision-level grounding, and robot-level control, enabling flexible integration of reasoning and perception modules.

- 我们提出了一种通用的动作分解策略用于零样本 VLN-CE，将导航分为语言层规划、视觉层落地和机器人层控制，实现推理与感知模块的灵活集成。

- We implement this strategy in LaViRA, a practical Language-Vision-Robot Action framework that leverages different scales of MLLMs in a fully zero-shot manner.

- 我们在 LaViRA 中实现了该策略，这是一种实用的语言-视觉-机器人动作框架，利用不同规模的多模态大语言模型，完全零样本。

- We achieve state-of-the-art zero-shot performance on the VLN-CE benchmark while preserving effectiveness and efficiency for real-world deployment.

- 我们在 VLN-CE 基准上实现了最先进的零样本性能，同时保持了面向真实部署的有效性和效率。

II. RELATED WORK

二、相关工作

A. Vision-and-Language Navigation

A. 视觉与语言导航

Early research in Vision-and-Language Navigation (VLN) [1], where an agent follows instructions to navigate unseen environments, predominantly focused on discrete, graph-based settings [8]-[10]. Such environments enable high-level decision-making but fail to capture the continuous control demands of real-world scenarios. To address this, VLN in Continuous Environments (VLN-CE) [2] removes reliance on connectivity graphs and requires agents to perform fine-grained control actions like moving forward, rotating, and avoiding obstacles. This transition introduces new challenges in scene analysis, generalization, and low-level control.

视觉与语言导航 (VLN)[1] 的早期研究中，智能体根据指令在未知环境中导航，主要集中于离散的图结构环境 [8]-[10]。此类环境支持高级决策，但无法反映现实场景中连续控制的需求。为此，连续环境下的视觉与语言导航 (VLN-CE)[2] 取消了对连通图的依赖，要求智能体执行如前进、旋转和避障等细粒度控制动作。这一转变带来了场景分析、泛化能力和低级控制的新挑战。

Numerous VLN methods in both discrete and continuous environments have improved performance through enhanced cross-modal alignment [11], [12], reinforcement learning [13], data augmentation [14]-[16], and map-based representations [17]-[19]. However, these learning-based approaches require substantial environment-specific training, which limits their applicability for zero-shot deployment-motivating recent interest in training-free VLN-CE solutions.

众多 VLN 方法在离散和连续环境中通过增强跨模态对齐 [11]、[12]、强化学习 [13]、数据增强 [14]-[16] 和基于地图的表示 [17]-[19] 提升了性能。然而，这些基于学习的方法需要大量特定环境的训练，限制了其零样本部署的适用性，促使人们对无训练 VLN-CE 解决方案产生了新的兴趣。

B. Zero-Shot VLN with Foundation Models

B. 基于基础模型的零样本视觉语言导航 (VLN)

The rise of powerful foundation models like Large Language Models (LLMs) [20], [21] and Multimodal Large Language Models (MLLMs) [7], [22], [23] has spurred a new wave of zero-shot VLN-CE research, which can be categorized by how foundation models are integrated with robot control. One dominant paradigm is waypoint-based navigation [3], [4]. These methods use a LLM/MLLM to select from waypoints proposed by a pre-trained predictor [24]. However, this creates a critical dependency on the predictor, which may fail to generalize to new scenes and limits backtracking flexibility [4]. Another approach is heuristic value-mapping [5], [6], where a Vision-Language Model (VLM) generates a semantic heatmap to guide the agent. In these frameworks, a powerful LLM is only used for offline instruction parsing, underutilizing its dynamic reasoning capabilities during navigation. Furthermore, the reliance on hard constraints for progress estimation can introduce rigidity and limit adaptability in complex scenarios [5].

强大的基础模型如大型语言模型 (LLMs)[20]，[21] 和多模态大型语言模型 (MLLMs)[7]，[22]，[23] 的兴起，推动了新一波零样本视觉语言导航-连续环境 (VLN-CE) 研究，这些研究可根据基础模型与机器人控制的集成方式进行分类。一种主流范式是基于路径点的导航 [3]，[4]。这些方法使用 LLM/MLLM 从预训练预测器 [24] 提出的路径点中选择。然而，这导致对预测器的关键依赖，预测器可能无法泛化到新场景，且限制了回溯的灵活性 [4]。另一种方法是启发式价值映射 [5]，[6]，其中视觉语言模型 (VLM) 生成语义热图以引导智能体。在这些框架中，强大的 LLM 仅用于离线指令解析，未充分利用其在导航过程中的动态推理能力。此外，依赖硬性约束进行进度估计可能引入刚性，限制复杂场景中的适应性 [5]。

Our LaViRA framework bridges the gap between these approaches, introducing a Language-Vision-Robot Action Translation that progressively refines action from language-level planning to vision-level grounding to robot-level execution. This design allows MLLMs reasoning to directly influence decisions at multiple granularities, enabling fully zero-shot, interpretable, and adaptable navigation in continuous environments.

我们的 LaViRA 框架弥合了这些方法之间的差距，提出了一种语言-视觉-机器人动作转换机制，逐步将动作从语言层面的规划细化到视觉层面的定位，再到机器人层面的执行。该设计使 MLLM 的推理能力能够直接影响多个粒度的决策，实现了在连续环境中完全零样本、可解释且适应性强的导航。

C. Hierarchical Structure in VLN

C. 视觉语言导航中的层级结构

Hierarchical architectures have emerged as a powerful paradigm for addressing complex, long-horizon navigation tasks. This strategy effectively decomposes the problem into high-level planning and low-level execution, allowing different modules to specialize. Current hierarchical approaches often combine high-level semantic planning with low-level motion control. One common strategy involves dividing the environment into distinct topological regions and having a high-level planner select a sequence of zones to traverse, while a low-level controller navigates within each zone [25], [26]. Other methods learn policies for both the manager and worker through techniques like imitation learning [27] or reinforcement learning [28]. A recent example, Nav A^3 [29], uses a global

policy to identify target regions and a pre-trained model for fine-grained navigation.

层级架构已成为解决复杂长时域导航任务的有效范式。该策略将问题有效分解为高层规划和低层执行，使不同模块得以专注于各自任务。当前的层级方法通常结合高层语义规划与低层运动控制。一种常见策略是将环境划分为不同的拓扑区域，由高层规划器选择一系列区域路径，低层控制器则在各区域内导航 [25]，[26]。其他方法通过模仿学习 [27] 或强化学习 [28] 同时学习管理者和执行者的策略。一个最新例子，Nav A^3 [29]，使用全局策略识别目标区域，并利用预训练模型进行细粒度导航。

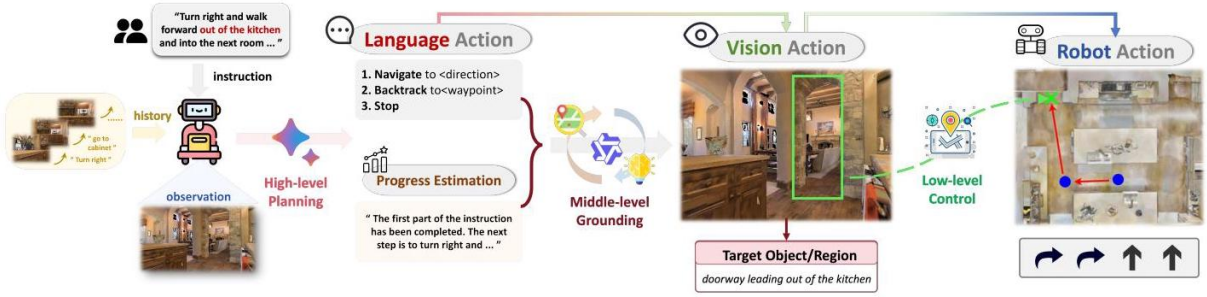

Fig. 2: The LaViRA Pipeline. Our framework decomposes navigation into three sequential stages. (1) Language Action: A large MLLM processes the instruction, history, and current observation to generate a high-level plan, deciding whether to move forward, turn, backtrack, or stop. (2) Vision Action: A smaller, more efficient MLLM takes the high-level plan and progress estimation to identify a specific visual target in the chosen direction, outputting its bounding box and description. (3) Robot Action: The target's pixel coordinates are projected onto a global map, and a rule-based controller navigates the robot to the destination. This hierarchical process enables generalized and robust zero-shot vision-language navigation.

图 2:LaViRA 流程。我们的框架将导航分解为三个连续阶段。(1) 语言动作: 大型 MLLM 处理指令、历史和当前观察，生成高层计划，决定前进、转向、回溯或停止。(2) 视觉动作: 较小且高效的 MLLM 根据高层计划和进度估计，识别所选方向上的具体视觉目标，输出其边界框和描述。(3) 机器人动作: 将目标的像素坐标投影到全局地图上，基于规则的控制器导航机器人到达目的地。该层级过程实现了通用且鲁棒的零样本视觉语言导航。

However, a common limitation of these methods is their reliance on extensive, environment-specific training, which restricts generalization and hinders zero-shot deployment. In contrast, LaViRA's Language-Vision-Robot action decomposition is fully zero-shot. By leveraging MLLM reasoning and a rule-based controller, it eliminates the need for any training, which simplifies deployment and enhances modularity.

然而，这些方法的一个常见限制是依赖大量环境特定的训练，限制了泛化能力并阻碍了零样本部署。相比之下，LaViRA 的语言-视觉-机器人动作分解完全零样本。通过利用 MLLM 的推理能力和基于规则的控制器，消除了任何训练需求，简化了部署并增强了模块化。

III. Proposed Method

III. 提出的方法

Our approach, LaViRA, introduces a novel framework for zero-shot Vision-and-Language Navigation in Continuous Environments (VLN-CE). The core idea is simple yet effective: we decompose the complex navigation task into a coarse-to-fine hierarchy of actions: Language Action, Vision Action, and Robot Action. This modular decomposition allows us to leverage the distinct strengths of different scales of Multimodal Large Language Models (MLLMs) at each stage, creating a system that is powerful in its reasoning, grounding and practical for real-world deployment, all without requiring any environment-specific training.

我们的方法 LaViRA 提出了一个用于连续环境中零样本视觉语言导航 (VLN-CE) 的新框架。核心思想简单而有效: 将复杂导航任务分解为粗到细的层级动作: 语言动作、视觉动作和机器人动作。该模块化分解使我们能够在每个阶段利用不同规模的多模态大型语言模型 (MLLM) 的独特优势，构建一个在推理、定位方面强大且适合实际部署的系统，且无需任何环境特定训练。

The overall pipeline is illustrated in Figure 2. In the VLN-CE task, an agent must follow a natural language instruction $\mathcal{I}$ through an unseen environment. At each timestep t , it uses an egocentric observation I_t to choose its next action $\mathcal{A}_t$ from a continuous space. To address this, our method decomposes the navigation process into a sequence of three hierarchical actions: a high-level directional plan (Language Action), the grounding of this plan into a specific visual target (Vision Action), and finally, the low-level movement to reach it (Robot Action). We detail each stage below.

整体流程如图 2 所示。在 VLN-CE 任务中，智能体必须根据自然语言指令 $\mathcal{I}$ 穿越未知环境。在每个时间步 t ，它使用自我中心观察 I_t 从连续动作空间中选择下一步动作 $\mathcal{A}_t$ 。为此，我们的方法将导航过程分解为三个层级动作序列: 高层方向规划 (语言动作)、将该规划定位到具体视觉目标 (视觉动作)，以及最终的低层移动以到达目标 (机器人动作)。以下详细介绍每个阶段。

A. Language Action: High-Level Planning

A. 语言动作: 高层规划

The first stage of our framework addresses the question: Where should I generally go next? To answer this, we employ a powerful, large-scale MLLM (e.g., GPT-4o) that functions as a high-level planner. This model is responsible for interpreting the full context of the navigation task and producing a coarse, directional command.

我们框架的第一阶段解决的问题是: 我下一步大致应该去哪里？为此，我们采用了一个强大的大规模多模态大语言模型 (MLLM)(例如，GPT-4o) 作为高级规划器。该模型负责解读导航任务的完整上下文，并生成一个粗略的方向性指令。

Specifically, the model receives three types of input:

具体来说，模型接收三种类型的输入:

- Language Instruction $\mathcal{I}$: The given natural language instruction provided at the start of the task.

- 语言指令 $\mathcal{I}$: 任务开始时给出的自然语言指令。

- Current Observation $\mathcal{O}_t$: A set of four images corresponding to the agent's front, left, right, and back views $\{I_{\text{front}}, I_{\text{left}}, I_{\text{right}}, I_{\text{back}}\}$, providing an informative understanding of the current location.

  - 当前观察 $\mathcal{O}_t$ ：一组四张图像，分别对应智能体的前方、左侧、右侧和后方视角 $\{I_{\text{front}}, I_{\text{left}}, I_{\text{right}}, I_{\text{back}}\}$，提供对当前位置的丰富理解。

- Navigation History $\mathcal{H}_t$: A structured summary of past observations and actions, formatted as a sequence like $\mathcal{H}_t = \{(\mathcal{O}_0, \mathcal{A}_0), (\mathcal{O}_1, \mathcal{A}_1), \ldots, (\mathcal{O}_{t-1}, \mathcal{A}_{t-1})\}$. This provides crucial context on what has already been accomplished.

  - 导航历史 $\mathcal{H}_t$ ：过去观察和动作的结构化摘要，格式为类似序列的形式 $\mathcal{H}_t = \{(\mathcal{O}_0, \mathcal{A}_0), (\mathcal{O}_1, \mathcal{A}_1), \ldots, (\mathcal{O}_{t-1}, \mathcal{A}_{t-1})\}$。这为已完成的内容提供了关键上下文。

Given this rich multimodal context, the MLLM performs two tasks simultaneously. First, it generates a Progress Estimation $\mathcal{P}_t$, a natural language assessment of how much of the instruction has been completed. This explicit reasoning step forces the model to track its progress against the overall instruction. Second, based on its analysis, the model selects a Language Action $\mathcal{A}_t^{\text{lang}}$ from a discrete set:

基于这一丰富的多模态上下文，MLLM 同时执行两项任务。首先，它生成一个进度评估 $\mathcal{P}_t$ ，即对指令完成程度的自然语言评估。这个显式推理步骤迫使模型跟踪其相对于整体指令的进展。其次，基于分析结果，模型从一个离散集合中选择一个语言动作 $\mathcal{A}_t^{\text{lang}}$ ：

- navigate to <direction>: Move forward, left, right, or behind.

  - 导航至 <direction>: 向前、左、右或后方移动。

- backtrack to <waypoint>: Return to a previously visited waypoint location.

  - 回溯至 <waypoint>: 返回先前访问过的路径点位置。

- stop: Terminate the navigation.

  - 停止: 终止导航。

This process can be formulated as:

该过程可形式化表示为:

$$\left(\mathcal{A}_t^{\text{lang}}, \mathcal{P}_t\right) = \text{MLLM}_{\text{large}}\left(\mathcal{I}, \mathcal{H}_t, \mathcal{O}_t\right) \tag{1}$$

This stage effectively abstracts the continuous environment into a few high-level choices, allowing the large MLLM to focus its powerful reasoning capabilities on strategic, long-term planning.

该阶段有效地将连续环境抽象为少数几个高级选项，使大型 MLLM 能够将其强大的推理能力集中于战略性、长期规划。

B. Vision Action: Perceptual Grounding

B. 视觉动作: 感知定位

Once a high-level Language Action is determined, the next stage must answer: What specific thing should I move towards in that direction? This is the role of the Vision Action stage, where we ground the abstract plan into a concrete perceptual target.

一旦确定了高级语言动作，下一阶段必须回答: 我应该朝那个方向具体移动到什么目标？这就是视觉动作阶段的作用，我们将抽象计划落地为具体的感知目标。

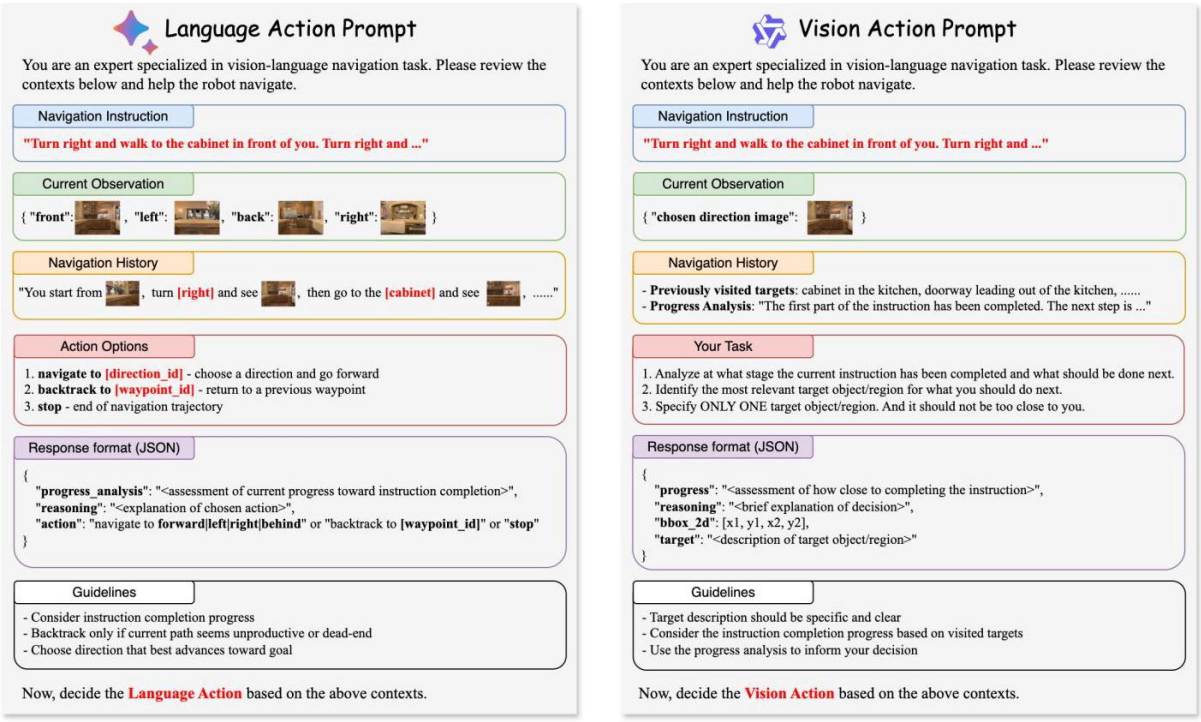

Fig. 3: Prompts for Language and Vision Actions. (Left) The prompt for the Language Action model, which takes in full context to decide on a high-level direction. (Right) The prompt for the Vision Action model, which uses the output from the first stage to ground the decision in a specific visual target.

图 3: 语言动作和视觉动作的提示。(左) 语言动作模型的提示, 输入完整上下文以决定高级方向。(右) 视觉动作模型的提示，利用第一阶段的输出将决策定位到具体视觉目标。

For this task, we use a smaller, more efficient MLLM (e.g., Qwen2.5-VL-32B). This choice is deliberate: grounding is a more focused perception task that does not require the same extensive world knowledge as high-level planning, making a smaller model more suitable and computationally efficient. As our ablation studies in Section IV-C confirm, pairing a powerful planner with a specialized, efficient grounding model yields optimal performance. The model is prompted with:

在此任务中，我们使用一个更小、更高效的 MLLM(例如，Qwen2.5-VL-32B)。这一选择是有意为之: 定位是一个更聚焦的感知任务，不需要像高级规划那样广泛的世界知识，因此更小的模型更合适且计算效率更高。正如我们在第四节 C 部分的消融研究所证实，将强大的规划器与专门的高效定位模型配对能获得最佳性能。模型的提示包括:

- Language Instruction $\mathcal{I}$: The original instruction.

- 语言指令 $\mathcal{I}$: 原始指令。

- Progress Estimation $\mathcal{P}_t$: The text generated by the Language Action model.

- 进度估计 $\mathcal{P}_t$: 由语言动作模型生成的文本。

- Chosen Direction Image I_{dir} : The single image corresponding to the direction chosen in the previous stage $\mathcal{A}_t^{\text{lang}}$.

- 选定方向图像 I_{dir} : 对应于前一阶段选定方向的单张图像 $\mathcal{A}_t^{\text{lang}}$ 。

The model's task is to identify the most relevant object or region in the image that aligns with the next step of the instruction. It outputs a Vision Action $\mathcal{A}_t^{vis}$ in a structured format containing a bounding box and its description. This can be expressed as:

模型的任务是识别图像中与下一步指令最相关的物体或区域。它以结构化格式输出视觉动作 $\mathcal{A}_t^{vis}$ ，包含边界框及其描述。可表达为:

$$\mathcal{A}_t^{\text{vis}} = \text{MLLM}_{\text{small}}(\mathcal{I}, \mathcal{P}_t, I_{\text{dir}}) \tag{2}$$

The output $\mathcal{A}_t^{\text{vis}}$ is a dictionary containing:

输出 $\mathcal{A}_t^{\text{vis}}$ 是一个包含以下内容的字典:

- Bounding Box $bbox_{2d}$: The 2D coordinates $[\times 1, \text{y1}, \text{x2}, \text{y2}]$ localizing the target.

- 边界框 $bbox_{2d}$: 定位目标的二维坐标 $[\times 1, \text{y1}, \text{x2}, \text{y2}]$ 。

- Target Description: A textual description of the identified object or region.

- 目标描述: 对识别出的物体或区域的文本描述。

As shown in Figure 3, the prompt instructs the model to select a target that is not too close, encouraging meaningful progress. This stage effectively translates the high-level plan into a tangible, visually verifiable goal.

如图 3 所示，提示指示模型选择一个不过于接近的目标，以促进有意义的进展。此阶段有效地将高层计划转化为具体且可视觉验证的目标。

C. Robot Action: Low-Level Control

C. 机器人动作: 低级控制

The final Robot Action stage answers: How do I physically get there? It translates the identified visual target into low-level motor commands using a robust, rule-based controller. Pixel-to-World Projection. First, we select the bottom-center pixel of the target's bounding box and project it into a 3D point in the world frame. This involves unprojecting the 2D pixel to the camera's 3D coordinate system using the intrinsic matrix $\mathbf{K}$:

最终的机器人动作阶段回答: 我如何物理到达那里? 它使用稳健的基于规则的控制器，将识别出的视觉目标转化为低级运动指令。像素到世界投影。首先，我们选择目标边界框的底部中心像素，并将其投影到世界坐标系中的三维点。这涉及使用内参矩阵 $\mathbf{K}$ 将二维像素反投影到相机的三维坐标系:

$$[X_{\text{cam}}, Y_{\text{cam}}, Z_{\text{cam}}]^T = d \cdot \mathbf{K}^{-1} \cdot [u_{\text{target}}, v_{\text{target}}, 1]^T \tag{3}$$

where d is the depth. This 3D point is then transformed from the camera frame to the world frame using the agent's current pose $\mathbf{T}_{\text{agent}} = (x_{\text{agent}}, z_{\text{agent}}, \theta_{\text{agent}})$ to yield a target position $\mathbf{p}_{\text{world}} = (x_{\text{world}}, z_{\text{world}})$ on the 2D map:

其中 d 是深度。该 3D 点随后利用代理当前位姿 $\mathbf{T}_{\text{agent}} = (x_{\text{agent}}, z_{\text{agent}}, \theta_{\text{agent}})$ 从相机坐标系转换到世界坐标系，得到二维地图上的目标位置 $\mathbf{p}_{\text{world}} = (x_{\text{world}}, z_{\text{world}})$:

$$\begin{bmatrix} x_{\text{world}} \\ z_{\text{world}} \end{bmatrix} = \begin{bmatrix} x_{\text{agent}} \\ z_{\text{agent}} \end{bmatrix} + \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} Z_{\text{cam}} \\ -X_{\text{cam}} \end{bmatrix} \tag{4}$$

Path Planning and Control. Given the target position $\mathbf{p}_{\text{world}}$, the agent computes a short-term path using the Fast Marching Method (FMM). A low-level controller then executes this path with local obstacle avoidance. This deterministic final step grounds the reasoning chain in physical action, ensuring interpretability and making the system adaptable to different robot platforms by simply swapping the controller.

路径规划与控制。给定目标位置 $\mathbf{p}_{\text{world}}$ ，代理使用快速行进法 (Fast Marching Method, FMM) 计算短期路径。低级控制器随后执行该路径并进行局部障碍物避让。此确定性最终步骤将推理链落地为物理动作，确保可解释性，并通过简单替换控制器使系统适配不同机器人平台。

IV. Simulation Experiments

IV. 仿真实验

We validate LaViRA in simulation to answer three key questions: (1) How does it compare against state-of-the-art methods on the standard VLN-CE benchmark? (2) Does performance support our hypothesis of using different MLLM scales for different decision granularities? (3) What is the contribution of each component in our framework?

我们在仿真中验证 LaViRA 以回答三个关键问题:(1) 在标准 VLN-CE 基准上，它与最先进方法相比表现如何？(2) 性能是否支持我们关于使用不同规模多模态大语言模型 (MLLM) 处理不同决策粒度的假设？(3) 框架中各组件的贡献是什么?

A. Experimental Setup

A. 实验设置

Environment and Dataset. We use the Habitat simulator [30] with the VLN-CE dataset [2], which extends the R2R benchmark from Matterport3D (MP3D) [8] for continuous navigation. Following recent zero-shot works [3], [4], we report results on a standard 100-episode subset from the validation unseen split. An episode is successful if the agent stops within 3 meters of the target.

环境与数据集。我们使用 Habitat 模拟器 [30] 和 VLN-CE 数据集 [2]，该数据集基于 Matterport3D(MP3D)[8] 的 R2R 基准扩展，实现连续导航。遵循近期零样本工作 [3]，[4]，我们在验证集未见子集的标准 100 集上报告结果。若代理在目标 3 米范围内停止，则该集成功。

Evaluation Metrics. We use standard VLN metrics: Navigation Error (NE), the final distance to goal; Success Rate (SR), our primary metric for stopping within 3 m ; Oracle Success Rate (OSR), SR if stopping at the closest point on the path; and Success rate weighted by Path Length (SPL), which penalizes inefficient paths. To account for the inherent stochasticity in the outputs of MLLMs, we repeat 3 runs for each experiment over the 100-episode set and report the mean and standard deviation for all metrics.

评估指标。我们采用标准 VLN 指标: 导航误差 (NE)，即最终距离目标的距离；成功率 (SR)，我们主要的指标，表示在 3 m 范围内停止的比例；Oracle 成功率 (OSR)，即在路径上最近点停止的成功率；以及按路径长度加权的成功率 (SPL)，惩罚低效路径。为考虑 MLLM 输出的固有随机性，我们对 100 集数据集每个实验重复 3 次，报告所有指标的均值和标准差。

Implementation Details. Our zero-shot framework requires no environment-specific training. For the high-level Language Action stage, we evaluate two leading MLLMs: Gemini-2.5-Pro and GPT-4o. For the Vision Action stage, we primarily use the efficient Qwen2.5-VL-32B, with other models explored in our ablation studies. The agent's observation is composed of posed 640×480 RGB-D images, low-level path planning is executed using the Fast Marching Method (FMM) on a global map constructed from depth observations. All experiments were conducted on 8 NVIDIA RTX 4090 GPUs for parallel evaluation.

实现细节。我们的零样本框架无需针对特定环境进行训练。在高级语言动作阶段，我们评估了两款领先的多模态大语言模型 (MLLMs):Gemini-2.5-Pro 和 GPT-4o。在视觉动作阶段，我们主要使用高效的 Qwen2.5-VL-32B，其他模型则在消融研究中进行了探索。代理的观察由姿态 640×480 RGB-D 图像组成，低级路径规划通过在由深度观测构建的全局地图上使用快速行进法 (Fast Marching Method, FMM) 执行。所有实验均在 8 块 NVIDIA RTX 4090 GPU 上并行进行评估。

Inference Cost. To provide a clear picture of the computational cost, we report the token usage for GPT-4o + Qwen2.5- VL-32B over the 100-episode validation set. On average, each trajectory required approximately 32,682 tokens with 7.93 calls for the high-level planner (GPT-4o) and 8,050 tokens with 7.50 calls for the grounding

model (Qwen2.5- VL-32B). Based on current API pricing, the total inference cost is approximately $0.084 USD per episode. This highlights the efficiency of our hierarchical design, where an expensive, powerful model is used for high-impact decisions, while a lightweight model handles the low-level perceptual grounding task.

推理成本。为清晰展示计算成本，我们报告了 GPT-4o + Qwen2.5-VL-32B 在 100 集验证集上的令牌使用情况。平均而言，每条轨迹约需 32,682 个令牌，高级规划器 (GPT-4o) 调用次数为 7.93 次，定位模型 (Qwen2.5-VL-32B) 令牌数为 8,050，调用次数为 7.50 次。基于当前 API 定价，总推理成本约为每集 0.084 美元。这凸显了我们分层设计的高效性，即用昂贵且强大的模型处理高影响力决策，而轻量级模型负责低级感知定位任务。

TABLE I: Main results on the VLN-CE benchmark. LaViRA significantly outperforms all previous zero-shot methods. Best and second-best zero-shot results are highlighted.

表 I:VLN-CE 基准测试的主要结果。LaViRA 显著优于所有先前的零样本方法。最佳和次佳零样本结果已高亮显示。

Method	NE↓	OSR↑	SR↑	SPL↑
SupervisedLearning				
CMA [31]	6.92	45	37	32.2
RecBERT [31]	5.80	57	48	43.2
ETPNav [18]	5.15	58	52	52.2
BEVBert [32]	5.13	64	60	53.4
Zero-Shot				
Random	8.63	12	2	1.5
NavGPT-CE [33]	8.37	26.9	16.3	10.2
DiscussNav-CE [33]	7.77	15	11	10.5
MapGPT-CE [34]	8.16	21	7	5.0
Open-Nav [3]	6.70	23	19	16.1
SmartWay [4]	7.11	51	29	22.5
InstructNav [6]	6.89	47	31	24.0
CA-Nav [5]	7.58	48.0	25.3	10.8
GC-VLN [35]	7.30	41.8	33.6	16.3
LaViRA(GPT-4o)	6.43 ± 0.28	43.3 ± 3.2	36.0 ± 1.7	28.3 ± 0.8
LaViRA(Gemini-2.5-Pro)	6.54 ± 0.27	48.7 ± 2.1	38.3 ± 0.6	28.3 ± 0.9

Method	NE↓	OSR↑	SR↑	SPL↑
SupervisedLearning				
CMA [31]	6.92	45	37	32.2
RecBERT [31]	5.80	57	48	43.2
ETPNav [18]	5.15	58	52	52.2
BEVBert [32]	5.13	64	60	53.4
零样本				
随机	8.63	12	2	1.5
NavGPT-CE [33]	8.37	26.9	16.3	10.2
DiscussNav-CE [33]	7.77	15	11	10.5
MapGPT-CE [34]	8.16	21	7	5.0
Open-Nav [3]	6.70	23	19	16.1
SmartWay [4]	7.11	51	29	22.5
InstructNav [6]	6.89	47	31	24.0
CA-Nav [5]	7.58	48.0	25.3	10.8
GC-VLN [35]	7.30	41.8	33.6	16.3
LaViRA(GPT-4o)	6.43 ± 0.28	43.3 ± 3.2	36.0 ± 1.7	28.3 ± 0.8
LaViRA(Gemini-2.5-Pro)	6.54 ± 0.27	48.7 ± 2.1	38.3 ± 0.6	28.3 ± 0.9

B. Main Results

B. 主要结果

We compare LaViRA with a range of existing methods on the VLN-CE benchmark. As detailed in Table 1, our method sets a new state-of-the-art for zero-shot VLN-CE.

我们在 VLN-CE 基准上将 LaViRA 与多种现有方法进行了比较。如表 1 所示，我们的方法在零样本 VLN-CE 任务中创下了新的最先进水平。

The LaViRA variant using Gemini-2.5-Pro as the high-level planner achieves a SR of 38.3% and an SPL of 28.3% . This represents a significant improvement over the previous leading zero-shot method, InstructNav [6], with a 7.3-point gain in SR and a 4.3-point gain in SPL. Furthermore, the low standard deviations across multiple runs underscore the robustness and stability of our framework, a key advantage in real-world applications where consistency is critical. Notably, our framework's performance surpasses even supervised methods in SR, underscoring the powerful reasoning and generalization capabilities unlocked by our hierarchical decomposition. The GPT-4o variant of LaViRA also shows strong and stable results. These findings validate our central hypothesis: by decomposing the navigation task and leveraging the advanced reasoning of MLLMs, we can achieve superior and robust zero-shot performance without relying on a pre-trained waypoint predictor.

使用 Gemini-2.5-Pro 作为高层规划器的 LaViRA 变体实现了 38.3% 的成功率 (SR) 和 28.3% 的路径效率 (SPL)。这较之前领先的零样本方法 InstructNav [6] 在 SR 上提升了 7.3 个百分点，在 SPL 上提升了 4.3 个百分点。此外，多次运行中较低的标准差凸显了我们框架的鲁棒性和稳定性，这在实际应用中对一致性至关重要。值得注意的是，我们框架的 SR 表现甚至超过了监督学习方法，凸显了通过分层分解所释放的强大推理和泛化能力。LaViRA 的 GPT-4o 变体也表现出强劲且稳定的结果。这些发现验证了我们的核心假设: 通过分解导航任务并利用多模态大语言模型 (MLLMs) 的先进推理能力，我们能够在无需预训练路径点预测器的情况下，实现卓越且稳健的零样本性能。

C. Ablation Studies

C. 消融研究

We performed a series of ablation studies to analyze LaViRA's performance and quantify the contribution of its core components. Our baseline for these studies is LaViRA (GPT-4o + Qwen2.5-VL-32B). Although the Gemini-2.5-Pro variant delivered superior performance, we used the GPT- 4o variant for ablations due to documented stability issues with the Gemini-2.5-Pro API during our experiments, which could have compromised the consistency of iterative testing. The GPT-4o model provided the necessary reliability for a rigorous and reproducible analysis.

我们进行了系列消融研究，以分析 LaViRA 的性能并量化其核心组件的贡献。这些研究的基线是 LaViRA(GPT-4o + Qwen2.5-VL-32B)。尽管 Gemini-2.5-Pro 变体表现更优，但由于实验中 Gemini-2.5-Pro API 存在已知的稳定性问题，可能影响迭代测试的一致性，我们在消融研究中采用了 GPT-4o 变体。GPT-4o 模型提供了进行严谨且可复现分析所需的可靠性。

Model Selection. Our framework is model-agnostic, allowing flexible combinations of MLLMs. As shown in Table II, our experiments confirm that pairing models of different scales according to task granularity is crucial. Using a powerful MLLM (e.g., GPT-4o) for high-level Language Action (LA) is key; replacing it with the smaller Qwen2.5-VL-72B leads to a 7.0-point drop in SPL. For the more focused Vision Action (VA) stage, an efficient model like Qwen2.5-VL- 32 B proves highly effective. This aligns with the principles of hierarchical VLN discussed in our related work II-C where specialized modules handle different levels of the task. Interestingly, using a powerful model like GPT-4o for both stages significantly degrades performance (SPL drops from 28.3% to 16.8%). This suggests that simply using the largest model is not optimal, the best results-pairing a top-tier planning model with an efficient grounding model-validate our core hypothesis and demonstrate a key advantage of our zero-shot hierarchical approach.

模型选择。我们的框架对模型无特定依赖，支持灵活组合多模态大语言模型。正如表 II 所示，实验确认根据任务粒度匹配不同规模模型至关重要。使用强大的 MLLM(如 GPT-4o) 处理高层语言动作 (LA) 是关键；若用较小的 Qwen2.5-VL-72B 替代，SPL 将下降 7.0 个百分点。对于更聚焦的视觉动作 (VA) 阶段，使用高效模型如 Qwen2.5-VL- 32 B 效果显著。这与我们相关工作 II-C 中讨论的分层 VLN 原则一致，即专门模块处理任务的不同层级。有趣的是，若两个阶段均使用强大模型如 GPT-4o，性能显著下降 (SPL 从 28.3% 降至 16.8%)。这表明单纯使用最大模型并非最优，最佳结果是将顶级规划模型与高效定位模型配对，验证了我们的核心假设并展示了零样本分层方法的关键优势。

TABLE II: Ablation on model selection. Performance is maximized when a powerful MLLM handles high-level Language Actions (LAM) and an efficient MLLM handles focused Vision Actions (VAM). Qwen-32B and Qwen-72B are short for Qwen2.5-VL-32B and Qwen2.5-VL-72B.

表 II: 模型选择的消融实验。当强大的 MLLM 负责高层语言动作 (LAM), 高效的 MLLM 负责聚焦视觉动作 (VAM) 时，性能达到最大化。Qwen-32B 和 Qwen-72B 分别是 Qwen2.5-VL-32B 和 Qwen2.5-VL-72B 的简称。

LAM	**VAM**	**NE** ↓	OSR↑	SR↑	SPL↑
Qwen-32B	Qwen-32B	8.04 ± 0.28	25.3 ± 2.1	19.7 ± 3.5	14.3 ± 2.6
Qwen-72B	Qwen-32B	7.18 ± 0.18	33.3 ± 2.1	27.7 ± 2.3	21.3 ± 2.7
GPT-40	Qwen-32B	6.43 ± 0.28	43.3 ± 3.2	36.0 ± 1.7	$\mathbf{28.3 \pm 0.8}$
GPT-4o	Qwen-72B	6.78 ± 0.11	39.3 ± 2.1	32.3 ± 1.2	23.8 ± 0.7
GPT-4o	GPT-40	7.47 ± 0.24	26.0 ± 7.5	20.7 ± 5.7	16.8 ± 4.9

LAM	**VAM**	**NE** ↓	OSR↑	SR↑	SPL↑
Qwen-32B	Qwen-32B	8.04 ± 0.28	25.3 ± 2.1	19.7 ± 3.5	14.3 ± 2.6
Qwen-72B	Qwen-32B	7.18 ± 0.18	33.3 ± 2.1	27.7 ± 2.3	21.3 ± 2.7
GPT-40	Qwen-32B	6.43 ± 0.28	43.3 ± 3.2	36.0 ± 1.7	$\mathbf{28.3 \pm 0.8}$
GPT-4o	Qwen-72B	6.78 ± 0.11	39.3 ± 2.1	32.3 ± 1.2	23.8 ± 0.7
GPT-4o	GPT-40	7.47 ± 0.24	26.0 ± 7.5	20.7 ± 5.7	16.8 ± 4.9

TABLE III: Ablation on framework design. A three-stage pipeline, rich visual history, and flexible backtracking are all crucial for robust navigation.

表 III: 框架设计消融实验。三阶段流程、丰富的视觉历史和灵活的回溯机制对于实现稳健导航均至关重要。

Configuration	**NE** ↓	OSR↑	SR↑	SPL↑
LaViRA (Full)	6.43 ± 0.28	43.3 ± 3.2	36.0 ± 1.7	28.3 ± 0.8
Framework Decomposition				
w/o LA	8.94 ± 0.53	13.0 ± 3.0	6.7 ± 0.6	4.4 ± 1.2
w/o VA	7.28 ± 0.23	34.0 ± 4.0	23.0 ± 5.2	13.9 ± 3.4
w/o LA+VA	9.21 ± 0.11	1.7 ± 0.6	0.0 ± 0.0	0.0 ± 0.0
History Representation				
text obs. + act.	6.99 ± 0.32	37.7 ± 5.5	31.0 ± 3.6	23.0 ± 2.8
only act.	6.54 ± 0.11	38.3 ± 1.2	32.7 ± 2.1	25.1 ± 1.8
only visual obs.	6.64 ± 0.23	36.0 ± 5.6	29.3 ± 3.5	22.5 ± 2.2
only text obs.	6.96 ± 0.24	35.7 ± 4.0	27.7 ± 2.5	21.8 ± 1.8
w/o history	6.90 ± 0.46	36.3 ± 7.0	27.0 ± 5.6	19.4 ± 7.3
Backtracking Mechanism				
w/o backtrack	6.92 ± 0.32	42.0 ± 3.0	30.0 ± 4.4	22.2 ± 4.0
last waypoint only	6.65 ± 0.16	41.7 ± 6.0	31.7 ± 1.5	23.5 ± 1.0
Pixel Point Selection				
direct output point	7.00 ± 0.84	45.3 ± 3.1	31.0 ± 5.6	19.4 ± 5.3
bbox median depth	6.73 ± 0.09	51.0 ± 2.6	34.0 ± 1.0	21.8 ± 1.2

配置	**NE** ↓	OSR↑	SR↑	SPL↑
LaViRA(完整)	6.43 ± 0.28	43.3 ± 3.2	36.0 ± 1.7	28.3 ± 0.8
框架分解				
无 LA	8.94 ± 0.53	13.0 ± 3.0	6.7 ± 0.6	4.4 ± 1.2
无 VA	7.28 ± 0.23	34.0 ± 4.0	23.0 ± 5.2	13.9 ± 3.4
无 LA+VA	9.21 ± 0.11	1.7 ± 0.6	0.0 ± 0.0	0.0 ± 0.0
历史表示				
文本观测 + 行动	6.99 ± 0.32	37.7 ± 5.5	31.0 ± 3.6	23.0 ± 2.8
仅行动	6.54 ± 0.11	38.3 ± 1.2	32.7 ± 2.1	25.1 ± 1.8
仅视觉观测	6.64 ± 0.23	36.0 ± 5.6	29.3 ± 3.5	22.5 ± 2.2
仅文本观测	6.96 ± 0.24	35.7 ± 4.0	27.7 ± 2.5	21.8 ± 1.8
无历史	6.90 ± 0.46	36.3 ± 7.0	27.0 ± 5.6	19.4 ± 7.3
回溯机制				
无回溯	6.92 ± 0.32	42.0 ± 3.0	30.0 ± 4.4	22.2 ± 4.0
仅最后航点	6.65 ± 0.16	41.7 ± 6.0	31.7 ± 1.5	23.5 ± 1.0
像素点选择				
直接输出点	7.00 ± 0.84	45.3 ± 3.1	31.0 ± 5.6	19.4 ± 5.3
边界框中值深度	6.73 ± 0.09	51.0 ± 2.6	34.0 ± 1.0	21.8 ± 1.2

Framework Design Choices. We conducted further ablations to evaluate key design choices within our framework, as consolidated in Table III.

框架设计选择。我们进行了进一步的消融实验，以评估框架中的关键设计选择，结果汇总于表 III。

First, we ablated our hierarchical structure. An end-to-end baseline (”w/o LA+VA”) that directly predicts actions fails completely, achieving an SPL of 0% . Removing the high-level planner (”w/o LA”) and using a single MLLM to directly output a bounding box yields a low 4.4% SPL. Conversely, removing the perceptual grounding model (”w/o VA”) and having the planner select only a direction achieves 13.9% SPL. Our full framework surpasses this by 14.4 absolute percentage points in SPL, proving the coarse-to-fine decomposition is critical for generating efficient and successful paths.

首先，我们消融了层次结构。一个端到端的基线模型 (“无 LA+VA”) 直接预测动作，表现完全失败，SPL 为 0% 。去除高层规划器 (“无 LA”) 并使用单一多模态大语言模型 (MLLM) 直接输出边界框，SPL 也很低，为 4.4% 。相反，去除感知定位模型 (“无 VA”) 且规划器仅选择方向，SPL 达到 13.9%。我们的完整框架在 SPL 上超出该方法 14.4 个百分点，证明粗到细的分解对于生成高效且成功的路径至关重要。

Next, we analyzed history representation. Replacing visual history with textual descriptions (”text obs. + act.”) drops SPL by 5.3 points to 23.0%, showing MLLMs benefit more from raw visual data than summaries when planning efficient paths. Using only past actions (”only act.”, 25.1% SPL) or only textual observations (”only text obs.”, 21.8% SPL) is also less effective. This confirms that a combination of visual observations and actions provides the richest context for optimal navigation performance.

接着，我们分析了历史表示。用文本描述替代视觉历史 (“文本观测 + 动作”) 导致 SPL 下降 5.3 个百分点至 23.0%，表明 MLLM 在规划高效路径时更依赖原始视觉数据而非摘要。仅使用过去动作 (“仅动作”，25.1% SPL) 或仅文本观测 (“仅文本观测”，21.8% SPL) 效果也较差。这证实视觉观测与动作的结合为导航性能提供了最丰富的上下文。

We then evaluated our backtracking mechanism. Disabling it (”w/o backtrack”) causes a 6.1-point SPL drop. A restrictive policy allowing only backtracking to the last way-point (”last waypoint only”) is still 4.8 points worse in SPL than our default strategy, which allows selecting any previous waypoint. This demonstrates that flexible, long-range error correction is crucial for improving path efficiency.

然后，我们评估了回溯机制。禁用回溯 (“无回溯”) 导致 SPL 下降 6.1 个百分点。仅允许回溯到最后一个路径点的限制策略 (“仅最后路径点”) 比默认策略 (允许选择任意先前路径点) 低 4.8 个百分点 SPL。这表明灵活的长距离错误纠正对提升路径效率至关重要。

Finally, we ablated the pixel point selection method. Having the model directly output pixel coordinates (”direct output point”) resulted in a 19.4% SPL,8.9 points lower than our approach. By first identifying a bounding box and then applying a geometrically-grounded heuristic (bottom-center point), we achieve more reliable and efficient navigation.

最后，我们消融了像素点选择方法。模型直接输出像素坐标 (“直接输出点”) 导致 SPL 为 19.4% ，比我们的方法低 8.9 个百分点。通过先识别边界框再应用几何启发式 (底部中心点)，我们实现了更可靠高效的导航。

D. Qualitative Analysis

D. 定性分析

To offer qualitative insights into LaViRA’s decision-making, Figure 4 shows a successful navigation run and common failures. In the success case, LaViRA demonstrates its coarse-to-fine approach: it first makes a high-level directional choice (”navigate left”), then grounds this in a visual landmark (”Black door with glass panels”), and finally selects a precise waypoint. This hierarchical process validates our design.

为提供 LaViRA 决策过程的定性见解，图 4 展示了一次成功导航和常见失败案例。在成功案例中，LaViRA 展现了其粗到细的方法: 首先做出高层方向选择 (“向左导航”)，然后将其定位于视觉地标 (“带玻璃窗的黑色门”)，最后选定精确路径点。该层次过程验证了我们的设计。

The failure cases illustrate common errors: (1) A Language Action error from an ambiguous instruction. For example, when the instruction asks to ”walk to the door in the front” where there are several doors in agent’s observation, LAM fails to identify the door desired. (2) A Vision Action error where the correct object description is grounded to the wrong image region. Though VAM exhibit a strong capability to ground objects in a bounding box, it sometimes fails to locate a relatively larger area like ”hallway” or ”living room” and chooses a meaningless region in the observation. (3) A simulation-induced error where depth reconstruction artifacts cause incorrect 3D projection of the target. For instance, in the Habitat simulator, transparent objects like windows are not assigned depth values. This lack of data disrupts the agent’s spatial perception, causing it to mislocate the target.

失败案例展示了常见错误:(1) 语言动作错误，源于指令歧义。例如，当指令要求“走到前面的门”而观察中有多扇门时，语言动作模型未能识别目标门。(2) 视觉动作错误，正确的目标描述被定位到错误的图像区域。尽管视觉动作模型能较好地将对象定位于边界框，但有时无法准确定位较大区域如“走廊”或“客厅”，而选择了无意义的观察区域。(3) 仿真引发的错误，深度重建伪影导致目标的 3D 投影错误。例如，在 Habitat 仿真器中，透明物体如窗户未被赋予深度值，缺失数据干扰了智能体的空间感知，导致目标定位错误。

V. REAL-WORLD EXPERIMENTS

V. 真实世界实验

To validate LaViRA's practicality beyond simulation, we deployed it on two distinct real-world robots: a Unitree Go1 quadruped and an Agilex Cobot Magic wheeled platform. These experiments tested the framework's sim-to-real transferability, requiring only the replacement of the low-level robot controller. The Unitree Go1 was equipped with a Jetson Orin NX and an Intel RealSense D435i camera, using its native velocity controller to navigate. The Agilex Cobot Magic platform used a chest-mounted Orbbec Dabai camera and a 2-DOF velocity controller for its mobile base. In both deployments, the onboard computers called the respective MLLM APIs for Language and Vision Actions, transmitting the resulting target pixel coordinates to the robot's navigation system.

为验证 LaViRA 在仿真之外的实用性，我们将其部署在两种不同的真实机器人上:Unitree Go1 四足机器人和 Agilex Cobot Magic 轮式平台。这些实验测试了框架的仿真到现实迁移能力，仅需替换低层机器人控制器。Unitree Go1 配备 Jetson Orin NX 和 Intel RealSense D435i 摄像头，使用其原生速度控制器导航。Agilex Cobot Magic 平台配备胸部安装的 Orbbec Dabai 摄像头和 2 自由度速度控制器。在两种部署中，车载计算机调用相应的 MLLM API 执行语言和视觉动作，并将生成的目标像素坐标传输给机器人导航系统。

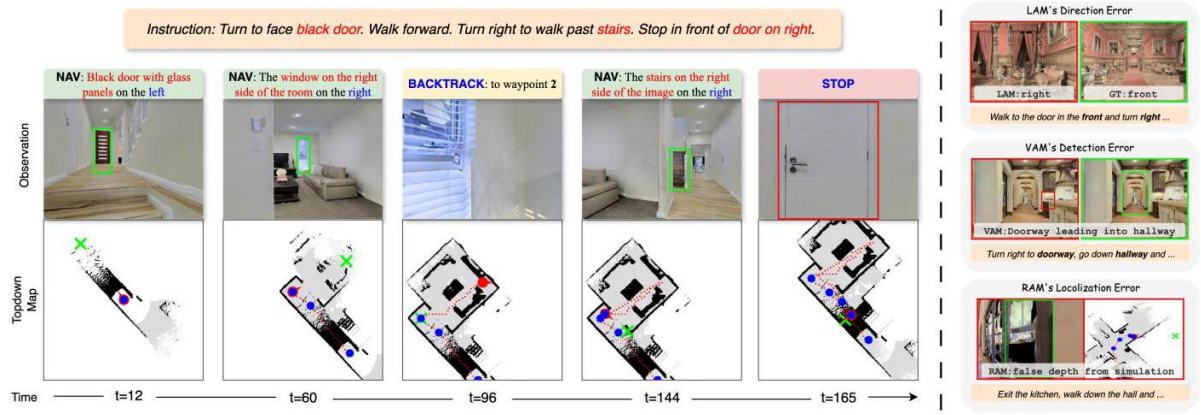

Fig. 4: Visualization examples. (Left) Navigation visualization: The outputs of Language Action model are denoted in blue text, Vision Action model outputs bounding boxes in green and target description in red. The robot's position and orientation are represented by a red dot with arrow, blue dots denote history waypoints, and the green cross marks the target position of the next waypoint. (Right) Failure cases visualization: Language action model misjudges direction due to ambiguous instruction; Vision action model gives correct target description but selects wrong region; Visual reconstruction errors in simulation cause incorrect target localization.

图 4: 可视化示例。(左) 导航可视化: 语言动作模型输出以蓝色文本表示，视觉动作模型输出边界框为绿色，目标描述为红色。机器人位置和朝向用带箭头的红点表示，蓝点表示历史路径点，绿色十字标记下一个路径点的目标位置。(右) 失败案例可视化: 语言动作模型因指令歧义误判方向；视觉动作模型给出正确目标描述但选错区域；仿真中的视觉重建错误导致目标定位错误。

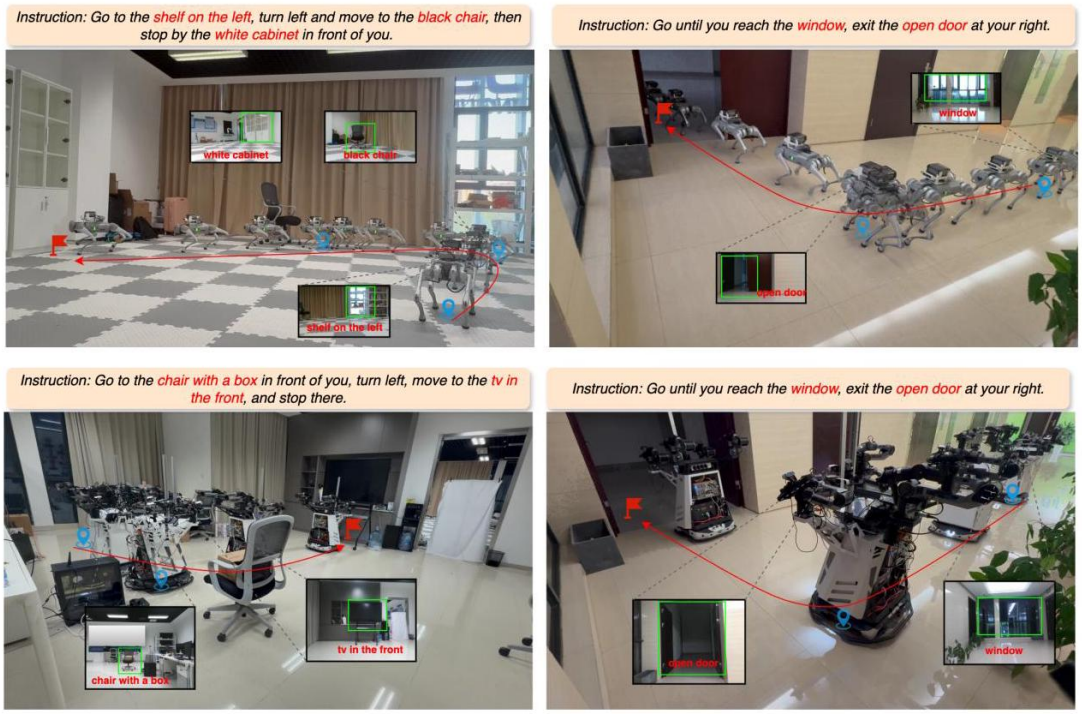

Fig. 5: Real-world experiment examples. LaViRA guides a Unitree Go1 quadruped (top) and an Agilex Cobot Magic wheeled robot (bottom) in an office. The visualization shows the third-person view of the robot's trajectory alongside the agent's ego view and targets, demonstrating successful real-world adaptation on diverse platforms.

图 5: 真实世界实验示例。LaViRA 在办公室中引导 Unitree Go1 四足机器人 (上方) 和 Agilex Cobot Magic 轮式机器人 (下方)。可视化展示了机器人轨迹的第三人称视角以及智能体的自我视角和目标，展示了在多样平台上的成功真实世界适应。

As shown in Figure 5, both platforms successfully executed navigation tasks in complex indoor environments. This demonstrates LaViRA's modularity and robustness, as the high-level reasoning and perceptual grounding components functioned effectively across different robot morphologies and control systems without modification, confirming the framework's real-world applicability. More details will be provided in the video.

如图 5 所示，两种平台均成功执行了复杂室内环境中的导航任务。这证明了 LaViRA 的模块化和鲁棒性，高层推理和感知定位组件在不同机器人形态和控制系统中无需修改即可有效运行，验证了该框架的真实世界适用性。更多细节将在视频中展示。

VI. CONCLUSION

六、结论

We introduce LaViRA, a novel, hierarchical framework for zero-shot Vision-and-Language Navigation in continuous environments. By decomposing the navigation process into a coarse-to-fine hierarchy of actions, LaViRA eliminates the need for pre-trained waypoint predictors and fully leverages the multi-granularity reasoning of MLLMs. Our experiments show that pairing a powerful MLLM for high-level planning with an efficient one for perceptual grounding sets a new state-of-the-art on the VLN-CE benchmark. LaViRA significantly outperforms existing zero-shot methods, demonstrating the potential of structured, model-driven reasoning for complex embodied navigation.

我们提出了 LaViRA，一种新颖的分层框架，用于连续环境中的零样本视觉语言导航 (Vision-and-Language Navigation，VLN)。通过将导航过程分解为粗到细的动作层级，LaViRA 消除了对预训练路径点预测器的需求，充分利用了多粒度大规模多模态语言模型 (MLLMs) 的推理能力。实验表明，将强大的 MLLM 用于高层规划与高效的 MLLM 用于感知定位相结合，在 VLN-CE 基准上创下了新的最先进水平。LaViRA 显著优于现有零样本方法，展示了结构化、模型驱动推理在复杂具身导航中的潜力。

Despite its performance, LaViRA's success is dependent on the underlying MLLMs, which can fail to interpret ambiguous instructions or ground descriptions of large areas. Future work will focus on enhancing robustness by fine-tuning the MLLMs on in-domain navigation data and improving the system's resilience to sim-to-real gaps, pushing LaViRA towards reliable real-world deployment. REFERENCES

尽管表现优异，LaViRA 的成功依赖于底层 MLLMs，这些模型可能无法准确理解模糊指令或定位大范围描述。未来工作将聚焦于通过在领域内导航数据上微调 MLLMs 提升鲁棒性，并增强系统对仿真到现实差异的适应能力，推动 LaViRA 实现可靠的真实世界部署。参考文献

[1] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. Van Den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 3674-3683.

P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. Van Den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 3674-3683.

[2] J. Krantz, E. Wijmans, A. Majumdar, D. Batra, and S. Lee, "Beyond the nav-graph: Vision and language navigation in continuous environments," in Proceedings of the European Conference on Computer Vision, 2020, pp. 104-120.

J. Krantz, E. Wijmans, A. Majumdar, D. Batra, and S. Lee, "Beyond the nav-graph: Vision and language navigation in continuous environments," in Proceedings of the European Conference on Computer Vision, 2020, pp. 104-120.

[3] Y. Qiao, W. Lyu, H. Wang, Z. Wang, Z. Li, Y. Zhang, M. Tan, and Q. Wu, "Open-nav: Exploring zero-shot vision-and-language navigation in continuous environment with open-source llms," arXiv preprint arXiv:2409.18794, 2024.

Y. Qiao, W. Lyu, H. Wang, Z. Wang, Z. Li, Y. Zhang, M. Tan, and Q. Wu, ”Open-nav: Exploring zero-shot vision-and-language navigation in continuous environment with open-source llms,” arXiv preprint arXiv:2409.18794, 2024.

[4] X. Shi, Z. Li, W. Lyu, J. Xia, F. Dayoub, Y. Qiao, and Q. Wu, ”Smart-way: Enhanced waypoint prediction and backtracking for zero-shot vision-and-language navigation,” arXiv preprint arXiv:2503.10069, 2025.

X. Shi, Z. Li, W. Lyu, J. Xia, F. Dayoub, Y. Qiao, and Q. Wu, ”Smart-way: Enhanced waypoint prediction and backtracking for zero-shot vision-and-language navigation,” arXiv preprint arXiv:2503.10069, 2025.

[5] K. Chen, D. An, Y. Huang, R. Xu, Y. Su, Y. Ling, I. Reid, and L. Wang, ”Constraint-aware zero-shot vision-language navigation in continuous environments,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025.

K. Chen, D. An, Y. Huang, R. Xu, Y. Su, Y. Ling, I. Reid, and L. Wang, ”Constraint-aware zero-shot vision-language navigation in continuous environments,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025.

[6] Y. Long, W. Cai, H. Wang, G. Zhan, and H. Dong, ”Instructnav: Zero-shot system for generic instruction navigation in unexplored environment,” arXiv preprint arXiv:2406.04882, 2024.

Y. Long, W. Cai, H. Wang, G. Zhan, and H. Dong, ”Instructnav: Zero-shot system for generic instruction navigation in unexplored environment,” arXiv preprint arXiv:2406.04882, 2024.

[7] J. Li, D. Li, S. Savarese, and S. Hoi, ”BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models,” in International Conference on Machine Learning, 2023, pp. 19730-19742.

J. Li, D. Li, S. Savarese, and S. Hoi, ”BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models,” in International Conference on Machine Learning, 2023, pp. 19730-19742.

[8] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, ”Matterport3d: Learning from rgb-d data in indoor environments,” arXiv preprint arXiv:1709.06158, 2017.

A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, ”Matterport3d: Learning from rgb-d data in indoor environments,” arXiv preprint arXiv:1709.06158, 2017.

[9] Y. Qi, Q. Wu, P. Anderson, X. Wang, W. Y. Wang, C. Shen, and A. van den Hengel, ”Reverie: Remote embodied visual referring expression in real indoor environments,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9979-9988.

Y. Qi, Q. Wu, P. Anderson, X. Wang, W. Y. Wang, C. Shen, and A. van den Hengel, ”Reverie: Remote embodied visual referring expression in real indoor environments,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9979-9988.

[10] A. Ku, P. Anderson, R. Patel, E. Ie, and J. Baldridge, ”Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding,” arXiv preprint arXiv:2010.07954, 2020.

A. Ku, P. Anderson, R. Patel, E. Ie, 和 J. Baldridge, “Room-across-room: 具有密集时空定位的多语言视觉与语言导航,”arXiv 预印本 arXiv:2010.07954, 2020。

[11] Y. Hong, Q. Wu, Y. Qi, C. Rodriguez-Opazo, and S. Gould, ”Vln bert: A recurrent vision-and-language bert for navigation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1643-1653.

Y. Hong, Q. Wu, Y. Qi, C. Rodriguez-Opazo, 和 S. Gould, “Vln bert: 用于导航的循环视觉与语言 BERT(BERT),”载于 IEEE/CVF 计算机视觉与模式识别会议论文集, 2021, 页 1643-1653。

[12] Y. Qi, Z. Pan, Y. Hong, M.-H. Yang, A. van den Hengel, and Q. Wu, ”The road to know-where: An object-and-room informed sequential bert for indoor vision-language navigation,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1635-1644.

Y. Qi, Z. Pan, Y. Hong, M.-H. Yang, A. van den Hengel, 和 Q. Wu, “知路之道: 基于对象和房间信息的室内视觉语言导航序列 BERT,”载于 IEEE/CVF 国际计算机视觉会议论文集, 2021, 页 1635-1644。

[13] X. Wang, Q. Huang, A. Celikyilmaz, J. Gao, D. Shen, Y.-F. Wang, W. Y. Wang, and L. Zhang, ”Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 6629-6638.

X. Wang, Q. Huang, A. Celikyilmaz, J. Gao, D. Shen, Y.-F. Wang, W. Y. Wang, 和 L. Zhang, “用于视觉语言导航的强化跨模态匹配与自监督模仿学习,”载于 IEEE/CVF 计算机视觉与模式识别会议论文集, 2019, 页 6629-6638。

[14] J. Li, H. Tan, and M. Bansal, ”Envedit: Environment editing for vision-and-language navigation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 15386-15396.

J. Li, H. Tan, 和 M. Bansal, “Envedit: 用于视觉与语言导航的环境编辑,”载于 IEEE/CVF 计算机视觉与模式识别会议论文集, 2022, 页 15386-15396。

[15] D. Fried, R. Hu, V. Cirik, A. Rohrbach, J. Andreas, L.-P. Morency, T. Berg-Kirkpatrick, K. Saenko, D. Klein, and T. Darrell, ”Speaker-follower models for vision-and-language navigation,” in Advances in Neural Information Processing Systems, 2018, pp. 3318-3329.

D. Fried, R. Hu, V. Cirik, A. Rohrbach, J. Andreas, L.-P. Morency, T. Berg-Kirkpatrick, K. Saenko, D. Klein, 和 T. Darrell, “视觉与语言导航的说话者-跟随者模型,”载于神经信息处理系统进展, 2018, 页 3318-3329。

[16] Z. Wang, J. Li, Y. Hong, Y. Wang, Q. Wu, M. Bansal, S. Gould, H. Tan, and Y. Qiao, ”Scaling data generation in vision-and-language navigation,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 12009-12020.

Z. Wang, J. Li, Y. Hong, Y. Wang, Q. Wu, M. Bansal, S. Gould, H. Tan, 和 Y. Qiao, “视觉与语言导航中的数据生成扩展,”载于 IEEE/CVF 国际计算机视觉会议论文集, 2023, 页 12009-12020。

[17] S. Chen, P.-L. Guhur, M. Tapaswi, C. Schmid, and I. Laptev, ”Think global, act local: Dual-scale graph transformer for vision-and-language navigation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16516-16526.

S. Chen, P.-L. Guhur, M. Tapaswi, C. Schmid, 和 I. Laptev, “全球思考，局部行动: 用于视觉与语言导航的双尺度图变换器,”载于 IEEE/CVF 计算机视觉与模式识别会议论文集, 2022, 页 16516-16526。

[18] D. An, H. Wang, W. Wang, Z. Wang, Y. Huang, K. He, and L. Wang, ”Etpnav: Evolving topological planning for vision-language navigation in continuous environments,” IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1-16, 2024.

D. An, H. Wang, W. Wang, Z. Wang, Y. Huang, K. He, 和 L. Wang, “Etpnav: 连续环境中视觉语言导航的演化拓扑规划,”IEEE 模式分析与机器智能汇刊, 页 1-16, 2024。

[19] Z. Wang, X. Li, J. Yang, Y. Liu, and S. Jiang, ”Gridmm: Grid memory map for vision-and-language navigation,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 15625-15636.

Z. Wang, X. Li, J. Yang, Y. Liu, 和 S. Jiang, “Gridmm: 用于视觉与语言导航的网格记忆地图,”载于 IEEE/CVF 国际计算机视觉会议论文集, 2023, 页 15625-15636。

[20] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., ”Language models are few-shot learners,” Advances in neural information processing systems, vol. 33, pp. 1877-1901, 2020.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, 等, “语言模型是少样本学习者,”神经信息处理系统进展, 卷 33, 页 1877-1901, 2020。

[21] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., ”Llama 2: Open foundation and fine-tuned chat models,” arXiv preprint arXiv:2307.09288, 2023.

H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, 等, “Llama 2: 开放基础模型与微调聊天模型,”arXiv 预印本 arXiv:2307.09288, 2023。

[22] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., ”Learning transferable visual models from natural language supervision,” in International Conference on Machine Learning, 2021, pp. 8748-8763.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, 等, “从自然语言监督中学习可迁移视觉模型,”载于国际机器学习会议, 2021, 页 8748-8763。

[23] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, et al., ”Gemini: a family of highly capable multimodal models,” arXiv preprint arXiv:2312.11805, 2023.

G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, 等, “Gemini: 一系列高性能多模态模型,”arXiv 预印本 arXiv:2312.11805, 2023.

[24] Y. Hong, Z. Wang, Q. Wu, and S. Gould, ”Bridging the gap between learning in discrete and continuous environments for vision-and-language navigation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 15418-15428.

Y. Hong, Z. Wang, Q. Wu, 和 S. Gould, “弥合视觉与语言导航中离散与连续环境学习的差距,”载于 IEEE/CVF 计算机视觉与模式识别会议论文集, 2022, 页 15418-15428.

[25] C. Gao, X. Peng, M. Yan, H. Wang, L. Yang, H. Ren, H. Li, and S. Liu, ”Adaptive zone-aware hierarchical planner for vision-language navigation,” in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14911-14920.

C. Gao, X. Peng, M. Yan, H. Wang, L. Yang, H. Ren, H. Li, 和 S. Liu, “适应性区域感知分层规划器用于视觉语言导航,”载于 2023 IEEE/CVF 计算机视觉与模式识别会议 (CVPR), 2023, 页 14911-14920.

[26] S. Zhang, X. Song, X. Yu, and et al., ”HOZ++: Versatile hierarchical object-to-zone graph for object navigation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025.

S. Zhang, X. Song, X. Yu, 等, “HOZ++: 多功能分层对象到区域图用于目标导航,”IEEE 模式分析与机器智能汇刊, 2025.

[27] M. Z. Irshad, C.-Y. Ma, and Z. Kira, ”Hierarchical cross-modal agent for robotics vision-and-language navigation,” in IEEE International Conference on Robotics and Automation (ICRA), 2021.

M. Z. Irshad, C.-Y. Ma, 和 Z. Kira, “用于机器人视觉与语言导航的分层跨模态代理,”载于 IEEE 国际机器人与自动化会议 (ICRA), 2021.

[28] F. Johnson, B. B. Cao, A. Ashok, and et al., ”Feudal networks for visual navigation,” arXiv preprint arXiv:2402.12498, 2024.

F. Johnson, B. B. Cao, A. Ashok, 等, “视觉导航的封建网络,”arXiv 预印本 arXiv:2402.12498, 2024.

[29] L. Zhang, X. Hao, Y. Tang, and et al.,”NavA ^{3} : Understanding any instruction, navigating anywhere, finding anything,” arXiv preprint arXiv:2508.04598, 2025.

L. Zhang, X. Hao, Y. Tang, 等, “NavA ^{3} : 理解任意指令，导航任意地点，寻找任意物品,”arXiv 预印本 arXiv:2508.04598, 2025.

[30] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik, et al., ”Habitat: A platform for embodied ai research,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9338-9346.

M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik, 等, “Habitat: 一个用于具身人工智能研究的平台,”载于 IEEE/CVF 国际计算机视觉会议论文集, 2019, 页 9338-9346.

[31] Y. Hong, Z. Wang, Q. Wu, and S. Gould, ”Bridging the gap between learning in discrete and continuous environments for vision-and-language navigation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 15439-15449.

Y. Hong, Z. Wang, Q. Wu, 和 S. Gould, “弥合视觉与语言导航中离散与连续环境学习的差距,”载于 IEEE/CVF 计算机视觉与模式识别会议论文集, 2022, 页 15439-15449.

[32] D. An, Y. Qi, Y. Li, Y. Huang, L. Wang, T. Tan, and J. Shao, ”Bevbert: Multimodal map pre-training for language-guided navigation,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 2737-2748.

D. An, Y. Qi, Y. Li, Y. Huang, L. Wang, T. Tan, 和 J. Shao, “Bevbert: 用于语言引导导航的多模态地图预训练,”载于 IEEE/CVF 国际计算机视觉会议论文集, 2023, 页 2737-2748.

[33] Y. Long, X. Li, W. Cai, and H. Dong, ”Discuss before moving: Visual language navigation via multi-expert discussions,” in IEEE International Conference on Robotics and Automation, 2023, pp. 17380-17387.

Y. Long, X. Li, W. Cai, 和 H. Dong, “行动前的讨论: 通过多专家讨论实现视觉语言导航,”载于 IEEE 国际机器人与自动化会议, 2023, 页 17380-17387.

[34] J. Chen, B. Lin, R. Xu, Z. Chai, X. Liang, and K.-Y. K. Wong, ”Mapgpt: Map-guided prompting with adaptive path planning for vision-and-language navigation,” arXiv preprint arXiv:2401.07314, 2024.

J. Chen, B. Lin, R. Xu, Z. Chai, X. Liang, 和 K.-Y. K. Wong, “Mapgpt: 基于地图引导的提示与自适应路径规划用于视觉与语言导航,”arXiv 预印本 arXiv:2401.07314, 2024.

[35] H. Yin, H. Wei, X. Xu, W. Guo, J. Zhou, and J. Lu, ”GC-VLN: Instruction as graph constraints for training-free vision-and-language navigation,” arXiv preprint arXiv:2509.10454, 2025.

H. Yin, H. Wei, X. Xu, W. Guo, J. Zhou, 和 J. Lu, “GC-VLN: 将指令作为图约束的无训练视觉与语言导航,”arXiv 预印本 arXiv:2509.10454, 2025.