

NaVIDA: 基于逆动力学增强的视觉—语言导航

Weiye Zhu^{1*} Zekai Zhang^{1*} Xiangchen Wang^{1*} Hewei Pan¹
Teng Wang¹ Tiantian Geng¹ Rongtao Xu^{2,3} Feng Zheng^{1,3†}

¹ 南方科技大学, 中国深圳

² 穆罕默德·本·扎耶德人工智能大学, 阿拉伯联合酋长国阿布扎比

³ SpatialTemporal AI, 中国深圳

摘要 视觉—语言导航 (Vision-and-Language Navigation, VLN) 要求智能体理解自然语言指令, 并在视觉信息丰富的环境中采取连贯行动。然而, 多数现有方法依赖反应式状态—动作映射, 并未显式建模动作落地视觉动力学。由于不了解动作如何改变后续视觉观测, 智能体难以合理规划行动, 导致行为不稳定、泛化能力弱, 并沿轨迹不断累积误差。为解决这些问题, 本文提出 NaVIDA (**N**avigation with **I**nverse **D**ynamics **A**ugmentation), 这是一个轻量级 VLN 框架, 将逆动力学监督 (IDS) 作为显式目标, 把动作落地视觉动力学融入策略学习。通过在共享表征与动作空间中联合优化视觉动力学和指令条件动作预测, NaVIDA 提供额外的结构化监督以正则化学习, 使导航更加稳定、一致。为组织这种监督并扩大有效规划范围, NaVIDA 采用层次概率动作分块 (HPAC), 将轨迹组织为多步动作块, 提供更具判别力的长程视觉变化线索。大量实验表明, 相较于最先进方法, NaVIDA 以更少参数 (3B 对 8B) 取得了更优的导航性能。真实世界机器人评测进一步验证了本文方法的实际可行性与有效性。代码见 <https://github.com/waynechu1021/NAVIDA>。

关键词: 视觉—语言导航 · 逆动力学

1 引言

视觉—语言导航 (VLN) [3, 17, 31] 是一项具有挑战性的多模态序列决策任务, 要求智能体仅根据自然语言指令在三维视觉环境中有效导航。因此, 有效的 VLN 不仅需要准确理解指令, 还需要长时域建模能力, 以及对视觉多样的未见场景具备稳健泛化能力。受视频多模态大语言模型 (MLLM) [4, 21, 46] 进展的推动, 近期研究 [7, 33, 37, 43, 44, 47] 将这类模型引入 VLN, 以利用其视觉感知、语义理解与跨场景泛化能力。

许多现有工作 [7, 43, 44] 通常将 VLN 建模为短视的状态到动作策略。如图 1 左图 (a) 所示, 这种“基础 VLN”范式直接将当前语言指令和历史视觉观测映射为有限的后续动作。然而, 这种反应式建模并未显式刻画动作落地视觉动力学: 当前视觉观测如何在动作影响下转变为下一视觉观测。智能体若未内化这种视觉动力学, 就只能学习指令与观测之间的浅层相关性, 无法预判所执行动作带来的视觉后果。导航过程因而极易受到微小感知偏差影响; 这些偏差在长时域中不断累积, 最终造成提前停止或无法纠正的轨迹漂移等严重不稳定现象。为

* 同等贡献

† 通讯作者

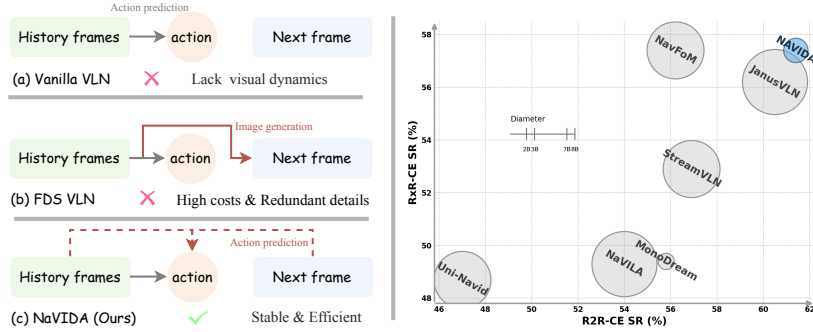


图 1: 左图: 三种 VLN 范式的对比。(a) 基础 VLN 依赖反应式状态—动作映射, 缺乏动作落地视觉动力学; (b) 前向动力学建模 (FDM) 增加了预测未来观测的图像生成任务, 但计算开销大且容易受到无关细节干扰; (c) NAVIDA 仅通过动作预测, 利用逆动力学监督学习动作落地视觉动力学, 使训练更高效、稳定。右图: 成功率与模型规模的对比。蓝色的 NAVIDA 以更小模型取得更优性能。每个圆的直径与模型参数量 (十亿参数) 成正比。

缓解该问题, 近期研究 [14, 33, 45] 探索了前向动力学建模 (FDM, 如图 1b 所示), 迫使智能体预测未来视觉观测。FDM 虽将环境动力学显式纳入策略学习, 但资源开销很大。生成高维未来帧通常需要复杂的辅助模块 (例如基于扩散的世界模型 [5, 8]), 并带来显著计算负担 [30, 45]。更关键的是, FDM 存在内在的目标错位: 重建稠密的像素级未来帧会迫使模型关注与导航完全无关的冗余外观细节 (例如光照或纹理), 而非优化结构化动作空间。

鉴于有效导航无须预测未来的每个像素, 本文提出 NAVIDA (图 1左图 c): 一个将逆动力学监督 (IDS) 作为辅助目标的 VLN 框架, 由此把动作落地视觉动力学引入策略训练。IDS 不再仅依赖指令—动作相关性, 而是训练模型根据视觉转变推断动作, 促使预测动作与其引起的可观测状态变化保持一致。这种动作落地监督可提供稠密且与语言无关的训练信号。通过在共享特征与动作空间中联合对齐 IDS 和基于指令的动作预测, NAVIDA 能够学习刻画动作如何驱动环境转变的视觉运动表征, 且无须稠密图像预测或访问未来帧。为更好地组织这种监督并延长有效规划时域, 本文提出层次概率动作分块 (HPAC) 机制, 逐步将低层动作合并为多层动作块。这种分层随机合并策略将短暂的微动作压缩为信息更丰富的训练目标, 为逆动力学学习提供更丰富、更具判别力的信号。HPAC 在动作块边界中引入受控随机性, 避免过于僵化的压缩, 在保留分段多样性的同时增强对轻微执行变化的鲁棒性。智能体由此能够在更长时间跨度上推理, 并保持稳定、一致的导航行为。大量实验验证了 NAVIDA 的有效性: 它以显著更少的参数在 VLN-CE [17] 基准上取得了最先进 (State-of-the-Art, SOTA) 性能 (直观对比见 图 1右图), 为 VLN 研究确立了一种简洁轻量的范式。

本文的主要贡献如下:

- 本文提出 NAVIDA, 一个利用逆动力学监督 (IDS) 所驱动的动作落地视觉动力学来增强策略学习的 VLN 框架。通过在共享特征与动作空间中对齐这两个目标, NAVIDA 实现了更稳定、一致的导航。

- 本文提出层次概率动作分块 (HPAC) 机制, 在时间维度上将智能体轨迹组织为变长语义单元。这一设计提高了对轻微执行变化的鲁棒性, 并为逆动力学目标提供更丰富的视觉转变线索, 有效延长规划时域。
- 全面实验表明, NAVIDA 以更少参数取得了 SOTA 性能。真实世界测试验证了其可行性与鲁棒性, 以及更快、更高效的执行能力, 体现了联合动作理解方法在物理环境中的实际优势。

2 相关工作

视觉—语言导航。 VLN 任务要求具身智能体根据自然语言指令进行感知、推理与行动。早期研究 [6, 13, 20] 关注以全景 RGB-D 为输入、在连通图上进行的离散导航。这些方法虽在域内取得了较好结果, 却依赖笨重传感器和预扫描视点。后续工作 [17, 38] 转向连续控制和低层动作预测, 以更贴近真实世界部署; 另有多项研究 [1, 12, 23, 28, 29, 34] 探索了单目 RGB 或 RGB-D 变体, 降低对全景感知的依赖。随着 MLLM 兴起, 仅使用单目视觉的 VLN 系统开始利用视觉—语言骨干网络增强跨场景迁移能力并降低硬件成本 [7, 37, 43, 44]。尽管取得这些进展, 多数现有 VLN 方法仍局限于反应式状态—动作映射, 忽视了对已执行动作与后续视觉转变关系的显式建模。缺乏这种动作理解能力时, 智能体难以预判短时域观测动力学, 往往出现转向不稳定、提前停止, 以及误差在长轨迹上持续累积等问题。与之不同, NAVIDA 在策略学习中引入逆动力学监督, 将其作为状态转变目标。通过在共享特征与动作空间中将该目标与基于指令的动作预测对齐, 本文方法提升了单目感知下的执行稳定性。

导航中的动力学模型。 反应式状态—动作建模将历史视觉观测和自然语言指令直接映射到动作空间, 并未显式刻画视觉转变与所执行动作之间的依赖关系。为缓解这一问题, 早期工作 [10, 19, 39] 探索了路径掩码预测、图像预测或图像补全等额外预训练任务。然而, 这些方法大多局限于离散环境, 并经常依赖扩展的上下文线索或语言指令。近期方法 [14, 33, 45] 尝试结合前向动力学建模并分解决策任务, 在预测动作的同时预测未来图像, 从而建模状态转变。前向模型虽然在部分设置下有效, 却通常需要额外模块完成高维预测, 增加计算量与内存开销, 还可能使学习偏向与决策空间关联较弱的冗余外观细节。NAVIDA 则在连续环境中采用逆动力学: 模型从两幅相邻图像直接推断引起该转变的动作, 与反应式策略共享输出空间, 无须额外生成器或解码器。这一设计降低了从图像空间映射到动作空间的优化难度, 提高数据效率, 并注入动作落地的转变先验约束以强化指令落地, 从而实现稳定的长时域导航。

3 NaVIDA 框架

3.1 预备知识

任务定义。 连续环境中的视觉—语言导航任务定义如下。在每个导航回合开始时, 智能体收到自然语言指令 $\mathcal{I}$ 。随后, 在每个时间步 t , 智能体接收第一人称 RGB 观测 v_t ; 观测历史记为 $\mathcal{V}_{0:t} = \{v_0, \dots, v_t\}$ 。导航智能体的离散动作空间为: $\mathcal{A} = \{\text{FORWARD}, \text{TURNLEFT}, \text{TURNRIGHT}, \text{STOP}\}$, 选择动作 $a_t \in \mathcal{A}$ 后, 环境转移至新状态并产生下一观测 v_{t+1} 。当智能体发出 STOP 或达到最大步数

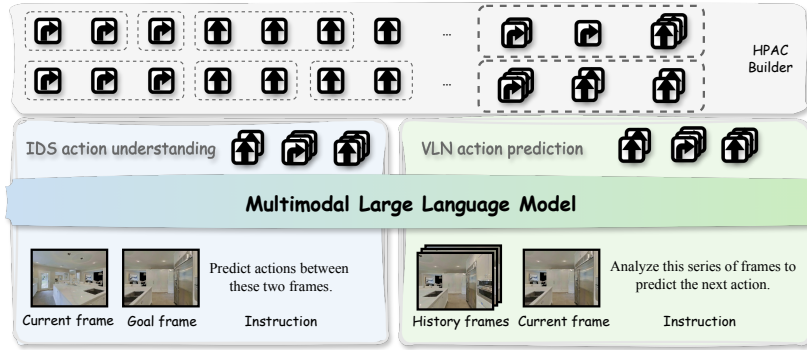


图 2: NAVIDA 概览。多模态语言模型骨干网络生成 HPAC 动作块。HPAC 机制将相邻动作随机合并为变长运动单元，在保留动作块多样性的同时增强对轻微执行变化的鲁棒性。导航动作预测与逆动力学监督 (IDS) 联合优化，从而在共享表征与动作空间中实现动作落地视觉动力学建模，并增强动作理解能力。

预算 $T_{\max}$ 时，回合终止。若智能体在距目标位置 3 米以内执行停止动作，则视为该回合成功。

框架概述。 NAVIDA 基于标准视觉—语言模型 (Vision-Language Model, VLM) 骨干网络构建，该骨干包含视觉编码器、视觉—语言投影器和 LLM 头 (图 2)。NAVIDA 在共享特征与动作空间中联合建模基于指令的动作预测和动作落地视觉动力学。为提供结构化且具有判别力的监督信号，NAVIDA 首先引入层次概率动作分块 (HPAC)，将时间相邻动作组织为多层动作块，并捕获更长程的视觉转变模式。在这种层次结构之上，本文将逆动力学监督 (IDS) 作为辅助目标，训练模型从视觉转变中推断动作；相较于简单的指令—动作相关性，这能增强动作理解和视觉—动作对齐。通过联合优化动作预测与 IDS，NAVIDA 在单目感知下实现了更稳定、一致的导航。

3.2 层次概率动作分块

NAVIDA 不再预测固定且少量的后续动作，而是用变长动作块表示控制序列，以便 (a) 放大动作作为逆动力学监督带来的视觉变化；(b) 在置信度较高时支持更远期的解码，同时保持输出与 VLN 动作空间一致。本文采用 HPAC 机制分两阶段合并原子动作。在第一阶段的概率合并中，以概率 p 将最多 3 个时间相邻、视觉转变一致且重复的同一原子动作合并为子动作块。该随机合并策略压缩微动作并缓解标签噪声，同时避免贪心方案造成僵化且可能过度的压缩。通过在动作块边界中引入受控随机性，它保留动作分段的多样性，并提高对轻微执行波动的鲁棒性。在第二阶段的确定性合并中，将最多 n 个连续子动作块合并为最终动作块，由此得到覆盖 n 至 $3n$ 个原子动作的序列。所得语法既保留细粒度短程决策，又允许在置信度较高时解码更远期动作；同时放大动作引起的外观与视点变化，产生更具判别力的监督信号。完整流程见算法 1。该动作块表征用作训练目标，为 IDS 和导航动作预测提供统一标签空间。

算法 1 层次概率动作分块**Require:** 原子动作 $\mathcal{A} = [a_1, \dots, a_N]$, 合并概率 p , 第二层最大动作块大小 n **Ensure:** 层次合并后的序列 $\mathcal{A}^{(2)}$

```

{第一层: 概率合并}
1:  $\mathcal{A}^{(1)} \leftarrow [ ]$ ,  $c \leftarrow [a_1]$ 
2: for  $i = 2$  到  $N$  do
3:   if  $a_i = a_{i-1}$  且  $\text{len}(c) < 3$  且  $\mathcal{U}(0, 1) \leq p$  then
4:     将  $a_i$  追加至  $c$ 
5:   else
6:     将  $c$  追加至  $\mathcal{A}^{(1)}$ ;  $c \leftarrow [a_i]$ 
7:   end if
8: end for
9: 将  $c$  追加至  $\mathcal{A}^{(1)}$ 
{第二层: 确定性合并}
10:  $\mathcal{A}^{(2)} \leftarrow [ ]$ ,  $c' \leftarrow [\mathcal{A}_1^{(1)}]$ 
11: for  $i = 2$  到  $|\mathcal{A}^{(1)}|$  do
12:   if  $\text{len}(c') < n$  then
13:     将  $\mathcal{A}_i^{(1)}$  追加至  $c'$ 
14:   else
15:     将  $c'$  追加至  $\mathcal{A}^{(2)}$ ;  $c' \leftarrow [\mathcal{A}_i^{(1)}]$ 
16:   end if
17: end for
18: 将  $c'$  追加至  $\mathcal{A}^{(2)}$ 
19: return  $\mathcal{A}^{(2)}$ 

```

3.3 逆动力学监督

为在不修改架构的情况下建模动作落地视觉动力学, 本文在数据层面引入逆动力学监督 (IDS)。给定一对由 HPAC 生成的帧 $(v_t, v_{t+\ell})$, 模型需要使用与 VLN 相同的动作语法 (即动作类型词元和数值词元), 预测二者之间变长动作块 $\mathbf{y}_t = (y_{t,1}, \dots, y_{t,n_t})$ 的词元序列。在这种定义下, IDS 是显式的动作理解目标, 促使模型以结构化动作表征解释视觉转变; VLN 则仍是指令条件动作预测任务。两个目标共享特征与动作空间, 因而相辅相成: IDS 利用视觉证据增强落地的动作理解, VLN 则利用这种表征进行语言引导决策。帧对与动作块层面的监督聚合多个微动作, 产生更大的视差和外观变化, 从而在视觉歧义场景 (例如对称结构或短暂遮挡) 中提供更具判别力的学习信号。遵循 StreamVLN [37] 和 JanusVLN [41], 本文使用 R2R [3]、RxR [18]、ScaleVLN [36] 和 DAgger [26] 在 Habitat [27] 中的轨迹构建监督数据。对于每条轨迹, 先用 HPAC 定义航路点, 再由相邻航路点确定互不重叠的子序列, 其中起始帧存为当前视图、结束帧存为目标视图, 二者之间的动作块作为监督标签。该过程生成约 2.4M 个 IDS 三元组, 以及约 ~ 2.8 M 个采用相同动作语法构建的分块 VLN 样本。

训练期间, 共享参数并使用任务特定提示词联合优化 IDS 与 VLN 目标。对于 IDS, 模型在与任务无关的提示词 $\mathcal{I}_{\text{IDS}}$ 下, 根据 $(v_t, v_{t+\ell})$ 预测 $\mathbf{y}_t$:

$$\mathcal{L}_{\text{IDS}}^t = - \sum_t \sum_{j=1}^{n_t} \log p(y_{t,j}^* | y_{t,<j}^*, \mathcal{I}_{\text{IDS}}, v_t, v_{t+\ell}), \quad (1)$$

对于 VLN，模型根据视觉历史和指令预测下一个动作块：

$$\mathcal{L}_{\text{VLN}}^{\tau} = - \sum_t \sum_{j=1}^{n_t} \log p(y_{t,j}^* | y_{t,<j}^*, \mathcal{I}_{\text{NAV}}, v_{<t}). \quad (2)$$

联合目标 $\mathcal{L} = \sum_{\tau \in D} (\mathcal{L}_{\text{VLN}}^{\tau} + \mathcal{L}_{\text{IDS}}^{\tau})$ 鼓励动作落地视觉动力学与指令条件动作预测保持一致，从而正则化共享表征。它将根据观测转变推断的动作与语言引导下解码的动作对齐，在保留原始 VLN 动作空间的同时，无须预测高维未来视觉外观。

表 1: 在 VLN-CE R2R Val-Unseen 划分上与当前最优方法的比较。所用观测包括全景视图 (Pano.)、里程计 (Odo.)、深度传感器 (Depth) 和单目 RGB 相机 (S.RGB)。所有结果均引自相应论文。* 表示使用航点预测器的方法。最优结果以粗体标出，次优结果以下划线标出。

方法	会议/期刊	VLM 参数量	观测					R2R Val-Unseen			
			全景	里程计	深度	单目	RGB	NE↓	OS↑	SR↑	SPL↑
HPN+DN* [15]	ICCV2021	-	✓	✓	✓			6.31	40.0	36.0	34.0
CMA* [12]	CVPR2022	-	✓	✓	✓			6.20	52.0	41.0	36.0
VLN $\odot$ BERT* [12]	CVPR2022	-	✓	✓	✓			5.74	53.0	44.0	39.0
Sim2Sim* [16]	ECCV2022	-	✓	✓	✓			6.07	52.0	43.0	36.0
GridMM* [34]	ICCV2023	-	✓	✓	✓			5.11	61.0	49.0	41.0
ETPNav* [2]	TPAMI2024	-	✓	✓	✓			4.71	65.0	57.0	49.0
ScaleVLN* [36]	ICCV2023	-	✓	✓	✓			4.80	-	55.0	51.0
InstructNav [22]	CoRL2024	-	-	-	-	-		6.89	-	31.0	24.0
LAW [25]	EMNLP2021	-		✓	✓	✓		6.83	44.0	35.0	31.0
CM2 [11]	CVPR2022	-		✓	✓	✓		7.02	41.5	34.3	27.6
ETPNav + FF [35]	CoRL2024	-		✓	✓	✓		5.95	55.8	44.9	30.4
Seq2Seq [17]	ECCV2020	-			✓	✓		7.77	37.0	25.0	22.0
NavMorph [40]	ICCV2025	-			✓	✓		5.75	56.9	47.9	33.2
NaVid [44]	RSS2024	7B					✓	5.47	49.1	37.4	35.9
NaVILA [7]	RSS2025	8B					✓	5.22	62.5	54.0	49.0
Uni-NaVid [43]	RSS2025	7B					✓	5.58	53.3	47.0	42.7
Aux-Think [32]	NeurIPS2025	8B					✓	6.08	60.0	54.8	46.9
StreamVLN [37]	ICRA2026	7B					✓	4.98	64.2	56.9	51.9
MonoDream [33]	AAAI2026	2B					✓	5.45	61.5	55.8	49.1
InternVLA-N1(S2) [9]	-	7B					✓	4.89	60.6	55.4	52.1
InternVLA-N1(S1+S2) [9]	-	8B				✓	✓	4.83	63.3	58.2	54.0
NavFoM [42]	ICLR2026	7B					✓	5.01	64.9	56.2	51.2
JanusVLN [41]	ICLR2026	8B					✓	<u>4.78</u>	<u>65.2</u>	<u>60.5</u>	56.8
NAVIDA (本文)	-	3B					✓	4.32	69.5	61.4	<u>54.7</u>

4 实验

4.1 实验设置

实现细节。本文在 NaVIDA 中采用 Qwen2.5-VL-3B [4]。训练期间冻结视觉编码器，并在混合视觉—语言导航与逆动力学数据上训练一轮，以 $2e-5$ 的学习率优化投影器和 LLM 主干。遵循 NaVILA [7]，将时间上下文限制为八个历史

表 2: 在 VLN-CE RxR Val-Unseen 划分上与当前最优方法的比较。所用观测包括全景视图 (Pano.)、里程计 (Odo.)、深度传感器 (Depth) 和单目 RGB 相机 (S.RGB)。所有结果均引自相应论文。最优结果以粗体标出, 次优结果以下划线标出。

方法	会议/期刊	VLM 参数量	观测					RxR Val-Unseen			
			全景	里程计	深度	单目	RGB	NE↓	SR↑	SPL↑	nDTW↑
NaVILA [7]	RSS2025	8B					✓	6.77	49.3	44.0	58.8
Uni-NaVid [43]	RSS2025	7B					✓	6.24	48.7	40.9	-
Aux-Think [32]	NeurIPS2025	8B					✓	6.24	52.2	40.2	-
StreamVLN [37]	ICRA2026	7B					✓	6.22	52.9	46.0	61.9
MonoDream [33]	AAAI2026	2B					✓	6.38	49.4	40.9	-
InternVLA-N1(S2) [9]	-	7B					✓	6.41	49.5	41.8	62.6
InternVLA-N1(S1+S2) [9]	-	8B			✓		✓	5.91	53.5	46.1	65.3
NavFoM [42]	ICLR2026	7B					✓	<u>5.51</u>	57.4	<u>49.4</u>	60.2
JanusVLN [41]	ICLR2026	8B					✓	6.06	<u>56.2</u>	47.5	62.1
NAVIDA (本文)	-	3B					✓	5.23	57.4	49.6	67.0

帧; 若序列超过此限制, 则均匀下采样为八帧, 以覆盖完整轨迹。除非另有说明, 层次概率动作分块 (HPAC) 采用合并概率 $p = 0.7$, 第 2 级分块最大大小为 $n = 3$ 。推理时仅执行预测出的前两个动作分块。更多超参数细节见补充材料。

评估基准。本文在 VLN-CE 上评估 NAVIDA, 采用 R2R-CE [3, 17] 和 RxR-CE [17, 18] 的 Val-Unseen 划分。遵循标准做法 [9, 32, 41], 报告导航误差 (NE)、成功率 (SR)、Oracle 成功率 (OS)、路径长度加权成功率 (SPL) 和归一化动态时间规整 (nDTW)。这些指标共同衡量目标完成情况、轨迹效率和路径保真度, 并支持与近期仅使用 RGB 及多传感器的方法直接比较。指标的详细说明见补充材料。

真实环境评估设置。真实环境实验在室内工作环境中使用 AgiBot X2 和 Unitree G1 两个人形机器人平台。AgiBot X2 配备两个单目鱼眼 RGB 相机; 实验仅使用其中一个, 并将其图像转换为与针孔相机采集结果等效的图像。Unitree G1 的内置相机朝下安装, 其 55.2° 视场 (FOV) 无法覆盖机器人正前方区域, 因此在其胸部另行安装 RealSense D455f 相机。机器人收到指令后, 将当前 RGB 图像上传至部署在配备 NVIDIA 4090 GPU 的远程服务器上的 NAVIDA。NAVIDA 在 0.93 秒内输出动作序列。机器人收到服务器发出的动作命令后, 通过运动 API 执行该命令。

4.2 主要结果

性能评估。如表 1 (R2R-CE) 和表 2 (RxR-CE) 所示, NAVIDA 仅使用单路 RGB 输入和 3B VLM 主干, 即在 Val-Unseen 划分上取得当前最优结果。在 R2R-CE 上, NAVIDA 的 NE、OS 和 SR 最优, SPL 次优; 相较于最强的 8B 基线 JanusVLN [41], NE、OS 和 SR 分别提升 -0.46、+4.3 和 +0.9。在 RxR-CE 上, NAVIDA 取得并列最优 SR (57.4), 同时获得最优 NE、SPL 和 nDTW; 相较于 StreamVLN 和 JanusVLN, SR 提升 +1.2 至 +4.5, SPL 提升 +2.1 至 +3.6, NE 提升 -0.83 至 -0.99。在轨迹更长、语言更丰富的 RxR-CE 基准上, IDS 与层次分块带来的持续增益表明, 模型具有更强的长程对齐能力, 并能减少漂移。与 ETPNav 和 NavMorph 等多传感器方法相比, NAVIDA 的单 RGB 策略在核心指标上仍具有竞争力或更优。其参数量仅为其他仅使用 RGB 方法的 35%, 且不使用辅助生成器或空间编码器, 却取得了最优性能。总体而言, 这些

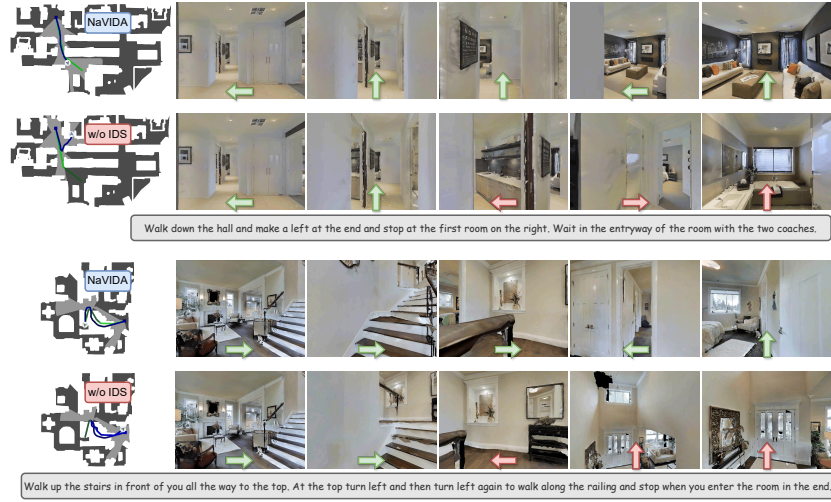


图 3: NAVIDA 在 VLN-CE R2R Val-Unseen 划分上的定性结果。正确动作以绿色箭头标出，错误动作以红色箭头标出。

表 3: VLN-CE R2R Val-Unseen 划分上的 IDS 数据消融实验。

R2R	RxR	Dagger	NE↓	OS↑	SR↑	SPL↑
○			7.86	44.5	32.0	27.6
●			6.99	51.5	40.5	35.5
●	●		5.72	57.4	47.7	41.5
●	●	○	5.25	60.9	53.7	47.0
○	○	○	5.29	61.9	52.6	47.1
●	●	●	4.91	63.2	55.5	49.6

表 4: VLN-CE R2R Val-Unseen 划分上的 IDS 设计变体消融实验。

形式	NE↓	OS↑	SR↑	SPL↑
无 IDS	5.29	61.9	52.6	47.1
MCQ	5.82	56.2	49.8	45.7
加噪 IDS	6.21	53.9	46.8	42.7
PD	5.27	60.1	52.3	47.8
本文方法	4.91	63.2	55.5	49.6

结果表明，结合 HPAC 的逆动力学增强可作为传统地图建模或前向世界建模的一种有效轻量替代方案。

定性结果。图 3 对比了 NAVIDA 与不使用 IDS 的变体，其中绿色箭头表示正确动作，红色箭头表示错误动作。NAVIDA 能够正确延迟转向直至越过楼梯，并选择预期路口；消融模型则出现过早转向和路口误判。这些现象是典型的误差累积：模型缺乏对动作落地视觉动力学的理解，无法内化自身动作造成的视觉后果。

图 4 展示了真实环境实验示例。NAVIDA 能够在狭窄转角处完成急转，并顺利通过单人宽门口；即使前方有行人，也能抵抗干扰并保持稳定朝向。仿真与真实环境证据共同表明，逆动力学监督（IDS）是实现上述稳定性和场景泛化能力的关键，使模型优于不使用逆动力学监督的变体。

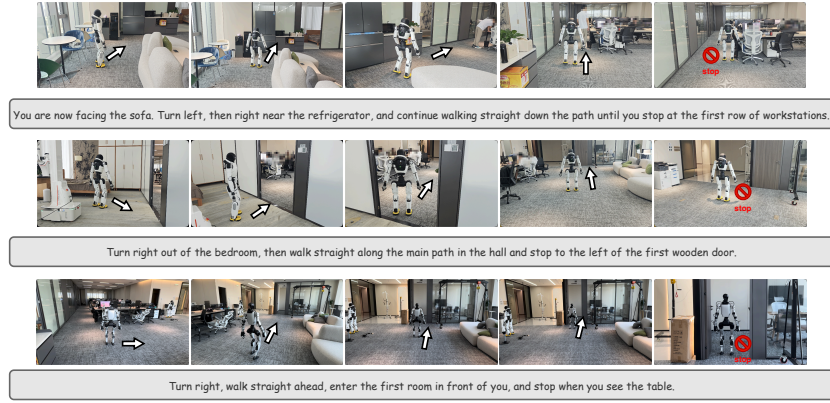


图 4: NAVIDA 的真实环境结果。图中箭头表示机器人的运动方向。

表 5: VLN-CE R2R Val-Unseen 划分上的 HPAC 消融实验。

分块大小	策略	NE↓	OS↑	SR↑	SPL↑
1	概率式	5.08	62.2	54.4	48.8
2	概率式	4.94	60.8	54.2	48.9
3	概率式	4.91	63.2	55.5	49.6
3	基于 DINOv2	6.29	56.1	44.7	39.5
4	概率式	5.23	60.2	52.5	46.4

表 6: VLN-CE R2R Val-Unseen 划分上的动力学建模消融实验。

动力学方法	NE↓	OS↑	SR↑	SPL↑
基线	5.29	61.9	52.6	47.1
使用 FDS (第 1 轮)	6.13	56.0	44.6	38.7
使用 FDS (第 2 轮)	5.06	60.1	51.0	45.2
使用 FDS (第 3 轮)	5.06	62.4	53.9	48.2
使用 IDS	4.91	63.2	55.5	49.6

4.3 消融实验

本节对 IDS、HPAC 及不同动力学建模方法的有效性进行消融实验。为提高实验效率，采用不含 ScaleVLN 的数据开展实验，并报告 R2R-CE 基准上的性能。为便于分析，本节执行所有预测动作分块。

IDS 数据的可扩展性。 表 3 改变 R2R、RxR 和 DAgger 轨迹是否提供 IDS 数据对（提供为 ●，不提供为 ○）。引入 IDS 在所有指标和数据配置下均能持续提升性能。仅使用 R2R 时，加入 IDS 使 SR 从 32.0 提升至 40.5；扩展至 R2R+RxR 后，SR 进一步升至 47.7。不使用 IDS 而引入 DAgger 时，SR 为 52.6；对全部三个数据源启用 IDS 监督后，SR 达到 55.5，同时 NE 降至 4.91，SPL 升至 49.6。性能增益在数据稀缺设置下最为明显，并随轨迹来源增加而持续累积。这种扩展规律表明，IDS 提供的是互补训练信号，而非仅仅增强模仿监督。总体来看，IDS 为模型注入了可迁移且与语言无关的动作落地视觉动力学概念。更重要的是，该能力可与 DAgger 等现有基于轨迹展开的范式无缝结合，无需改变底层模仿学习框架即可获得叠加增益。

IDS 设计消融。 为深入理解 IDS 如何促进动作预测，本文对其构建策略进行消融实验（表 4）。首先，在数据构建过程中用随机动作替换真实动作以破坏 IDS（加噪 IDS），从而打破动作与视觉转换之间的对齐，同时保持视觉—语言导航数据和训练规模不变。噪声监督为动作标记引入不一致的训练信号，使导航性能显著下降。这证实 IDS 的主要作用是强化细粒度动作—视觉动力学一致性，而

不仅增加数据规模。随后，将 IDS 改写为多项选择题 (MCQ)：模型不再直接生成动作标记，而是从四个候选项中选择正确动作。尽管保留了动作落地视觉动力学，导航性能仍略有下降。MCQ 格式允许模型依赖候选项间的简单比较或排除，而生成式 IDS 要求智能体仅根据视觉变化直接生成动作标记。因此，与视觉—语言导航解码共享同一生成式动作空间至关重要，这使 IDS 学到的转换知识可直接用于导航过程中的动作预测，从而产生更稳健的行为。本文还实现了特权蒸馏 (PD) 变体：先以 IDS 监督训练教师模型，再通过最小化师生动作分布之间的 KL 散度，将其知识蒸馏到执行视觉—语言导航任务的学生模型中。尽管增加了这种监督，PD 的表现仍与不含动力学监督的基线相近，且逊于 NAVIDA。这一有限增益源于动作落地视觉动力学理解的传递方式。PD 仅在输出分布层面对齐，促使学生模仿教师的最终预测。这种事后监督将视觉动力学知识视为训练后注入的固定动作信号。相比之下，NAVIDA 在统一框架内联合优化 IDS 与视觉—语言导航，使动作—视觉对齐能够通过共享表示持续学习和强化。这种更紧密的耦合使动作落地视觉动力学直接塑造导航推理，学习效果优于基于蒸馏的方法。

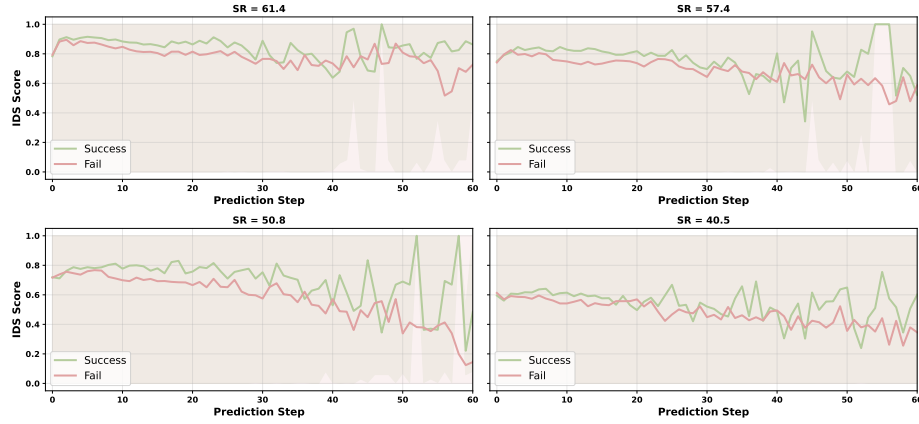


图5: 成功与失败轨迹的 IDS 分数随预测步数的变化。成功回合的 IDS 分数始终更高且更稳定，失败回合则逐渐下降，表明准确的动作—状态对齐对长程导航至关重要。

IDS 与导航性能的相关性。为量化 IDS 推断动作序列与真实动作序列的一致性，本文根据二者产生的仿真器状态差异定义 **IDS 分数**。分别执行预测序列 $\hat{\mathbf{a}}$ 和真实序列 $\mathbf{a}$ ，得到最终位姿 $(\hat{\mathbf{p}}, \hat{\theta})$ 和 $(\mathbf{p}, \theta)$ ，其中 $\mathbf{p} \in \mathbb{R}^3$ 表示位置， θ 表示偏航角。位置与角度差异计算如下：

$$\Delta_p = \|\hat{\mathbf{p}} - \mathbf{p}\|_2, \quad \Delta_\theta = \min \left(\left| \hat{\theta} - \theta \right|, 2\pi - \left| \hat{\theta} - \theta \right| \right). \quad (3)$$

IDS 分数定义为：

$$\text{IDS}_{\text{Score}} = \exp \left(-\frac{\Delta_p}{\sigma_p} \right) \cdot \exp \left(-\frac{\Delta_\theta}{\sigma_\theta} \right), \quad (4)$$

其中, σ_p 和 σ_θ 分别控制对位置与角度偏差的敏感度。实现中设定 $\sigma_p = 0.1$ 、 $\sigma_\theta = \pi/12$ 。

图 5 展示了在不同总体成功率下, 成功与失败轨迹的 IDS 分数随预测步数的变化, 呈现出若干一致趋势。在早期阶段, 成功与失败回合的 IDS 分数相近, 表明两种情况下的短期动作预测通常都较可靠。随着导航推进, 两条曲线逐渐分离: 成功轨迹保持更高、更稳定的 IDS 分数, 而失败轨迹的分数持续下降。差距不断扩大, 说明对动作引起的视觉转换进行理解和准确预测, 对长程导航愈发重要。总体成功率更高的设置始终具有更高、更稳定的 IDS 分数。这种强相关性表明, IDS 的准确性反映了模型理解自身动作视觉动力学的能力。理解越准确, 动作序列随时间越能保持连贯, 导航性能也越好。相反, 失败轨迹中较低的 IDS 分数意味着模型对动作—视觉动力学理解较弱, 微小偏差不断累积并最终导致失败。

HPAC 消融。 表 5 给出了 VLN-CE R2R Val-Unseen 划分上的 HPAC 消融实验。首先分析动作分块最大数量的影响。当最大值设为 1 且合并概率固定为 1 时, HPAC 退化为先前工作 [7,44] 采用的确定性动作合并策略。在默认概率合并方案下, 允许最多 3 个动作分块时性能最佳; 分块更少或更多都会降低性能。这表明, 适度延长分块有助于长程规划, 而过长分块可能超出模型的有效感知范围并导致次优决策。概率合并策略还会在分块边界中引入受控随机性, 保持动作分块的多样性, 并增强对轻微执行波动的鲁棒性。本文还评估了基于 DINOv2 [24] 相似度的确定性替代方案, 其中依据视觉相似度对相邻帧分组。该变体总体表现最差, 原因在于 DINOv2 主要捕获高层语义相似性, 而导航依赖几何变化和视角变化。例如, 从不同视角观察同一物体时, 即使空间位移显著, DINOv2 相似度仍可能很高。这种由语义驱动的确定性合并会掩盖运动线索并降低分段多样性, 进而削弱动作监督并降低性能。

逆动力学与前向动力学。 为评估逆动力学监督的有效性, 本文将其与前向动力学监督 (FDS) 比较, 二者均旨在增强动作落地视觉动力学建模。遵循 MonoDream [33], FDS 使用同一 VLM 预测未来视觉特征, 不引入额外模块。考虑到其收敛较慢, 本文将 FDS 训练三轮, 以评估其相较于 IDS 的有效性和效率。如表 6 所示, IDS 始终优于 FDS 和不含动力学监督的基线, 而 FDS 在初期会导致明显的性能下降。这一差距源于输出空间对齐方式和优化难度的差异。预测稠密未来视觉特征需要建模高维外观变化, 这对预训练 VLM 而言较难, 通常需要更充分的训练 (例如, MonoDream 训练五轮)。相比之下, IDS 与视觉—语言导航中的动作预测共享同一输出空间, 因而与模型的预训练目标和下游目标对齐得更好。尽管 FDS 随训练增加而逐步改善, 并在三轮后最终超过基线, 其性能仍落后于仅训练一轮即取得更优结果的 IDS。这些结果表明, IDS 提供了更稳定且数据效率更高的监督, 而 FDS 需要更长时间的优化才能学到可靠的未来状态预测。

5 结论

本文提出 NAVIDA, 这是一个将策略学习与动作落地视觉动力学建模相结合的统一视觉—语言导航框架。通过基于分块的逆动力学监督, NAVIDA 实现了更具可解释性且更稳定的导航。所提出的层次概率动作分块机制将导航轨迹组织为多层级运动单元, 为逆动力学学习提供更丰富、更长程的监督。大量实验表

明，与当前最优方法相比，NAVIDA 以更少的参数实现了更优的导航性能；真实机器人评估进一步验证了其有效性和实际适用性。

A 不同任务的提示词

我们针对视觉—语言导航和逆动力学监督（IDS）使用特定于任务的提示词，二者均采用相同的动作语法（动作类型 + 数值词元）表示。

对于视觉—语言导航任务，我们使用以下提示词模板，引导模型预测相应的动作块：

假设你是一个为导航任务编程的机器人。现向你提供一段历史观测视频：< 图像 >, ..., < 图像 >，以及当前观测：< 图像 >。分配给你的任务是：[指令]。请分析这一系列图像以决定下一步动作，该动作可能包括向左或向右旋转特定角度，或向前移动一定距离。

其中，[指令] 是当前导航任务给出的语言指令。对于逆动力学监督（IDS），我们使用如下提示词模板：

假设你是一个为导航任务编程的机器人。现向你提供一张当前视图图像 < 图像 > 和一张目标视图图像 < 图像 >。请分析这两张图像，预测可使机器人从当前视点移动至目标视图的导航动作，该动作可能包括向左或向右旋转特定角度，或向前移动一定距离。

对于逆动力学监督（IDS）的多项选择题（MCQ），提示词如下：

假设你是一个为导航任务编程的机器人。现向你提供一张当前视图图像 < 图像 > 和一张目标视图图像 < 图像 >。请分析这两张图像，预测可使机器人从当前视点移动至目标视图的导航动作。请从四个选项（A、B、C 或 D）中选择正确答案。选项：\n A. < 动作 A>\n B. < 动作 B>\n C. < 动作 C>\n D. < 动作 D>

对于前向动力学监督（FDS），任务要求根据给定的历史帧和语言指令预测下一时间步的视觉观测特征。我们使用以下提示词：

假设你是一个为导航任务编程的机器人。现向你提供一段历史观测视频：< 图像 >, ..., < 图像 >，以及当前观测：< 图像 >。分配给你的任务是：[指令]。请分析这一系列图像以预测未来观测。

B 更多实现细节

B.1 数据细节

图 6a 展示了逆动力学（IDM）数据集中每个层次概率动作分块（HPAC）所含原子动作数量的分布。尽管大多数动作块包含 3 至 5 个原子动作，与先前工

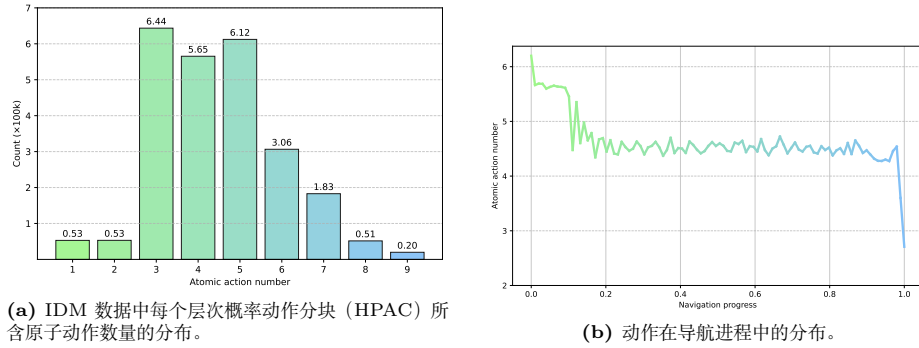


图 6: 数据集中层次概率动作分块 (HPAC) 的原子动作统计。

作 [7, 37, 44] 采用的固定动作块大小一致, 但仍有不可忽略的一部分样本包含更少或更多原子动作。较长的动作块编码更大幅度的第一人称视点变化与视差, 为逆动力学提供强且信息丰富的监督信号, 并通过扩展模型的时间上下文支持长时域规划。相比之下, 较短的动作块侧重细微、精细的视觉变化 (例如小幅转向或局部重新对齐), 有助于实现精确控制并减少短时域下的动作歧义。这种长短动作块的自然混合构成互补的课程: 层次概率动作分块 (HPAC) 既为规划提供全局结构线索, 也提供用于稳定控制的局部调整先验, 从而增强模型对动作条件视觉动态及动作的理解, 并提高执行过程中抵抗累积误差的鲁棒性。

我们还将导航进程归一化至 0 到 1 的范围, 图 6b 展示了动作块随导航推进的分布。可以看出, 在导航早期, 层次概率动作分块 (HPAC) 会合并大量原子动作; 到中期, 其数量略有下降并趋于稳定。在导航后期、接近目标时, 层次概率动作分块 (HPAC) 倾向于使用较少的原子动作来稳定机器人位置。

B.2 训练超参数

NAVIDA 基于 Qwen-2.5-VL-3B 实现; 视觉编码器被冻结, 仅投影器和 LLM 参与训练。我们采用 AdamW, 学习率 = $2e-5$ 、BF16、Flash-Attention v2、全局批大小 = 64; 其他超参数汇总于表 7。

B.3 评估指标细节

为形式化评估标准并阐明各指标的实际含义, 我们报告导航误差 (NE)、成功率 (SR)、Oracle 成功率 (OSR)、路径长度加权成功率 (SPL) 和归一化动态时间规整 (nDTW)。简而言之, NE 越低越好, 而 SR/OSR/SPL/nDTW 越高越好。**导航误差 (NE)**。NE 衡量智能体停止位置 $\mathbf{p}_T$ 与目标位置 $\mathbf{g}$ 之间的最终距离, 其定义为:

$$NE = d(\mathbf{p}_T, \mathbf{g}), \quad (5)$$

其中, $d(\cdot, \cdot)$ 表示导航环境中的测地距离。NE 越小, 说明终点精度越高。

表 7: 微调 Qwen2.5-VL 所用的训练超参数。仅投影器和 LLM 模块参与训练。

超参数	取值
训练轮数	1
可训练模块	投影器 + LLM
学习率	$2e - 5$
优化器	AdamW
AdamW β_1	0.9
AdamW β_2	0.999
Adam ϵ	$1e - 8$
学习率调度	线性
全局批大小	64
梯度累积	4
混合精度类型	BFloat16
注意力类型	Flash Attention v2
LLM 序列长度	64800
图像分辨率	308×252

成功率 (SR)。若智能体停止在距目标不超过阈值 τ 的范围内, 则该导航回合被视为成功。成功率定义为:

$$\text{SR} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}[d(\mathbf{p}_T^{(i)}, \mathbf{g}^{(i)}) \leq \tau], \quad (6)$$

其中, N 为回合总数; 在大多数视觉—语言导航设定中, τ 设为 3。SR 越高, 表示完成目标的一致性越强。

Oracle 成功率 (OSR)。OSR 评估智能体在整条轨迹中是否曾进入成功半径 τ 的范围, 而不考虑其最终停止位置:

$$\text{OSR} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}\left[\min_t d(\mathbf{p}_t^{(i)}, \mathbf{g}^{(i)}) \leq \tau\right]. \quad (7)$$

其中, $\mathbf{p}_t^{(i)}$ 是机器人在第 i 个回合中的第 t 个位置。OSR 有助于理解可恢复性和过早停止现象。

路径长度加权成功率 (SPL)。SPL 衡量成功导航的效率, 对比最短测地路径长度 $L^{(i)}$ 与智能体实际行进的路径长度 $P^{(i)}$:

$$\text{SPL} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}[d(\mathbf{p}_T^{(i)}, \mathbf{g}^{(i)}) \leq \tau] \cdot \frac{L^{(i)}}{\max(P^{(i)}, L^{(i)})}. \quad (8)$$

对于失败回合, SPL 通过指示函数取值为 0; 当成功路径接近最优时, SPL 趋近于 1。

归一化动态时间规整 (nDTW)。nDTW 使用动态时间规整距离 $\text{DTW}(\cdot, \cdot)$ 衡量智能体路径 $\{\mathbf{p}_t\}_{t=1}^T$ 与最短路径参考轨迹 $\{\mathbf{q}_k\}_{k=1}^K$ 之间的轨迹相似度:

$$\text{nDTW} = \exp\left(-\frac{\text{DTW}(\mathbf{p}_{1:T}, \mathbf{q}_{1:K})}{\eta}\right), \quad (9)$$

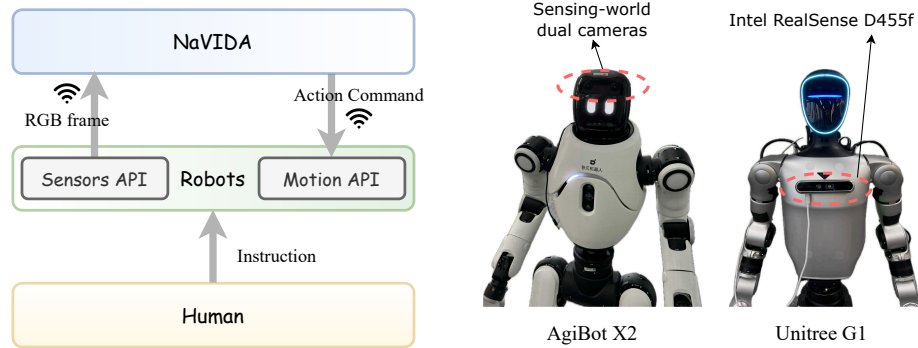
其中, η 是归一化常数 (通常取最短路径长度)。这种指数归一化将对齐代价转换为范围在 $(0, 1]$ 内的有界相似度分数; 数值越接近 1, 表示与最优参考轨迹在空间和时间上的一致性越高。直观而言, 即使轨迹在时序或步骤对应关系上存在细微差异, nDTW 也会奖励紧密遵循正确路线的轨迹。与简单的几何距离指标相比, nDTW 考虑了导航中常见的局部时间错位和路线重叠, 因而能以更灵活、更鲁棒的方式衡量路径跟随质量。

B.4 真实世界部署细节

图 7a 展示了真实世界部署系统的整体架构。具体而言, NAVIDA 部署在配备 NVIDIA RTX 4090 GPU 的远程计算服务器上。该系统采用客户端—服务器结构: 机器人作为客户端收集传感数据并执行运动命令, 远程服务器负责高层推理与动作预测。

运行过程中, 机器人持续通过前置相机采集 RGB 图像 (如图 7b 所示), 并经由局域网接收用户的自然语言或语音命令。随后, 传感数据及对应命令被打包并通过互联网传输至远程服务器。在服务器端, NAVIDA 模型处理这些输入, 执行视觉—语言推理, 并预测下一个导航动作。预测的运动命令随后返回机器人, 并通过机器人的运动控制 API 转换为线速度、角速度和运动时长等可执行参数。

此外, 我们还提供了真实世界部署的演示视频, 并随补充材料一并提交。视频中观察到的轻微延迟主要源于机器人与远程服务器之间的网络通信延迟。



(a) 用于真实世界评估的部署系统架构。RGB 帧和用户指令从机器人传输至运行 NAVIDA 的远程服务器; 服务器预测动作序列, 并返回运动命令, 由机器人的运动 API 执行。

(b) 机器人平台与传感配置。AgiBot X2 (左) 配备单个经校正的鱼眼 RGB 相机, Unitree G1 (右) 另配备 Intel RealSense D455f 前置相机, 从而为 NAVIDA 的导航策略提供一致的单目 RGB 输入。

图 7: 真实世界评估设置。

表 8: 在 VLN-CE R2R Val-Unseen 划分上微调视觉编码器的消融研究。

视觉编码器	NE↓	OS↑	SR↑	SPL↑
可训练	5.12	63.6	55.6	49.2
冻结	4.91	63.2	55.5	49.6

C 更多实验结果

C.1 更多消融实验

数据规模。 图 8 展示了 NAVIDA 在 R2R-CE Val-Unseen 上的性能随数据总量变化的趋势，其中数据总量同时包括逆动力学监督 (IDS) 数据与视觉—语言导航数据。我们还在 R2R-CE Val-Unseen 上构建了逆动力学监督 (IDS) 验证集，并绘制了 IDS 分数的变化曲线。随着数据总量增加，NAVIDA 的性能持续提升。在数据规模较小时，这种提升尤为显著，说明 NAVIDA 能够在低数据量条件下有效利用额外数据，学习更鲁棒的表示。然而，随着数据总量继续增长，性能增益逐渐趋于饱和。这一现象可能反映了边际收益递减：模型已经从现有数据中捕获了大部分有用模式，继续增加数据对性能的贡献随之减小。此外，性能趋稳也可能表明，在数据量达到某一阈值后，模型容量、任务复杂度或数据内在噪声等其他因素成为限制条件。

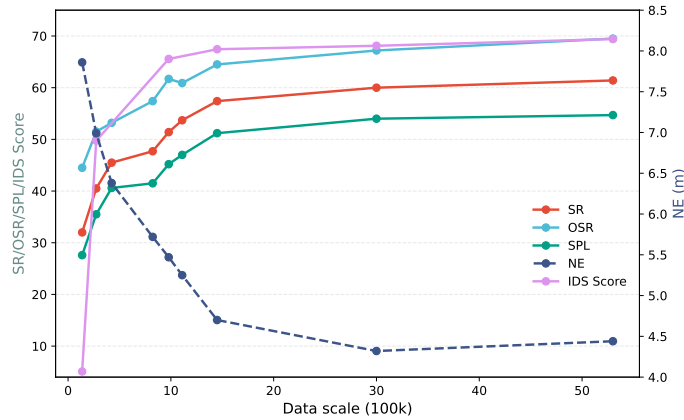


图 8: 随着数据规模增加，NAVIDA 在 VLN-CE R2R Val-Unseen 划分上的导航性能。随着数据量增长，SR/OSR/SPL/IDS 分数上升而 NE 下降；在大数据规模下，性能增益逐渐减弱。

视觉编码器：冻结与可训练。 我们开展消融研究，评估微调视觉编码器对整体导航性能的影响。如表 8 所示，与保持冻结相比，解冻视觉编码器并进行端到端训练仅带来微小提升。这一结果表明，预训练视觉表示已经为下游导航任务提供了足够的信息，进一步适配的收益有限。此外，微调会显著增加 GPU 显存占

用和训练时间，而精度收益很小。因此，在后续所有实验中，我们均保持视觉编码器冻结，使模型能够高效利用鲁棒的预训练视觉特征，同时在性能与计算效率之间取得良好平衡。

历史帧数量。我们开展消融研究，考察历史帧数量如何影响 R2R-CE Val-Unseen 划分上的导航性能。为保持一致性，所有模型均仅使用 R2R 训练数据进行训练。如表 9 所示，使用 8 个历史帧足以覆盖大多数指令时域，为模型的动作推理提供充分的时间上下文。将帧数增加至 16 并未带来显著性能提升，偶尔还会导致轻微下降，这可能是因为较长序列积累了冗余或噪声视觉信息。这一现象与 NaVILA [7] 的发现一致，表明过多的历史上下文可能削弱近期观测的相关性并妨碍时间定位。因此，我们采用 8 个历史帧，以在上下文丰富度与模型效率之间取得平衡。

表 9: VLN-CE R2R Val-Unseen 划分上历史帧数量的消融研究。

# 历史帧	NE↓	OS↑	SR↑	SPL↑
8	6.99	51.5	40.5	35.5
16	7.02	50.1	40.4	36.1

C.2 更多定性结果

我们在真实世界和仿真环境中展示了 NAVIDA 的大量定性结果，以进一步验证其有效性与泛化能力，如图 9、图 10 和图 11 所示。结果表明，NAVIDA 在仿真中能够生成平滑且符合指令的轨迹，并能在真实世界场景中鲁棒地执行导航命令。这些发现凸显了模型强大的跨域泛化能力、准确的语言定位能力，以及在复杂长时域导航任务中可靠的决策能力。

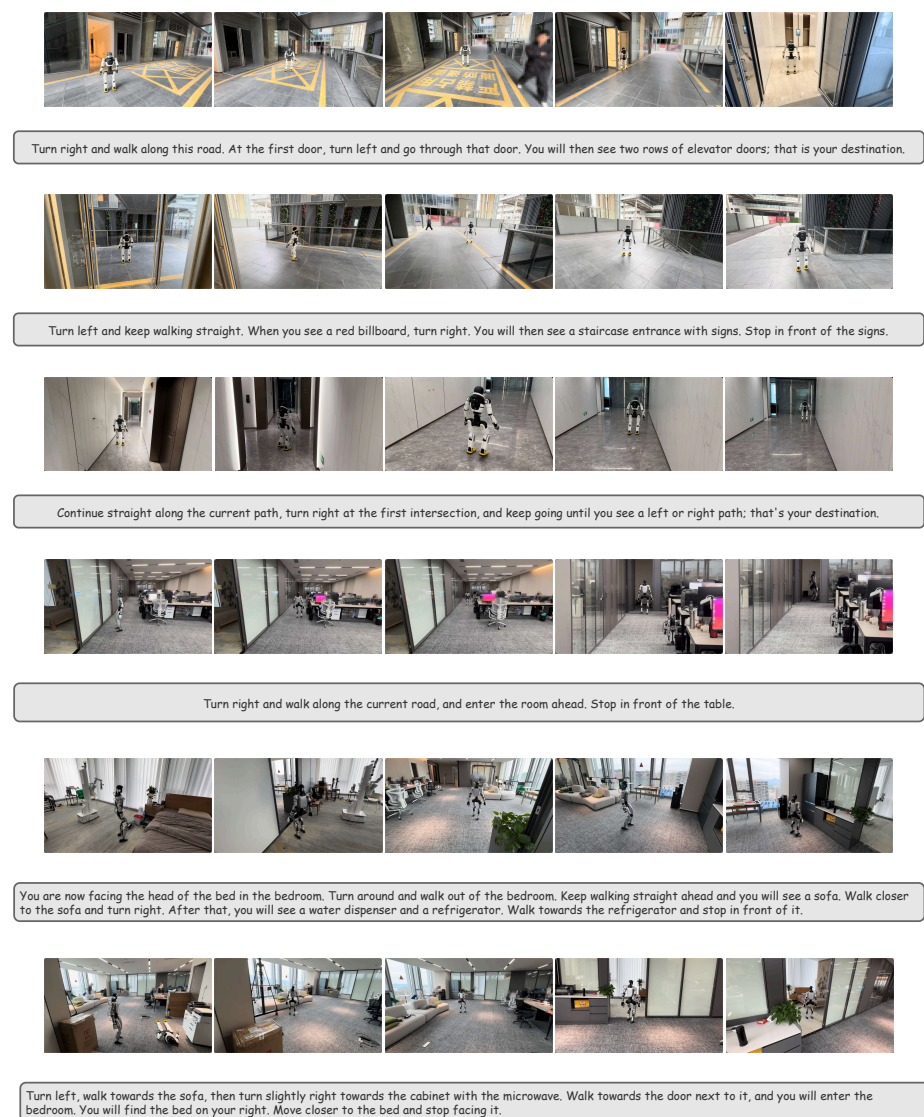


图9: NAVIDA 在真实世界中的定性结果。

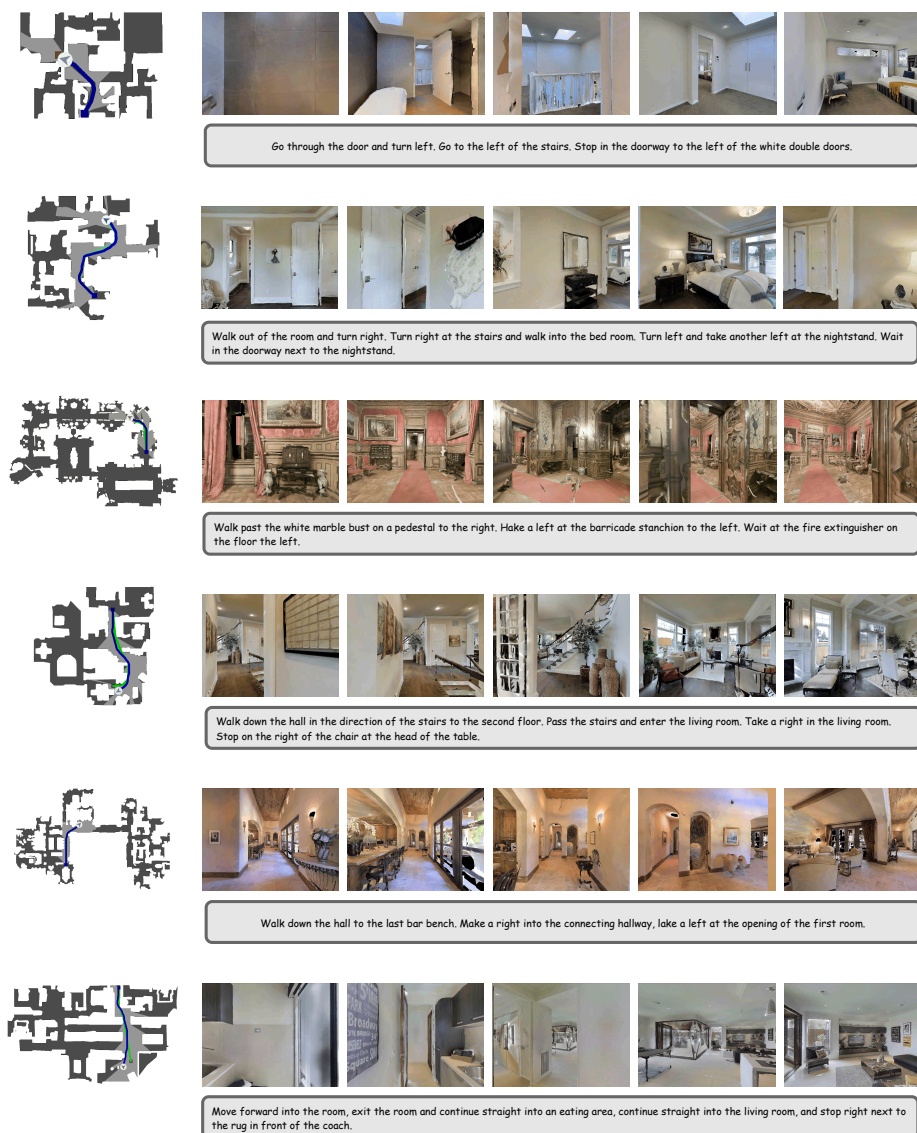


图 10: NAVIDA 在 VLN-CE R2R Val-Unseen 划分上的定性结果。

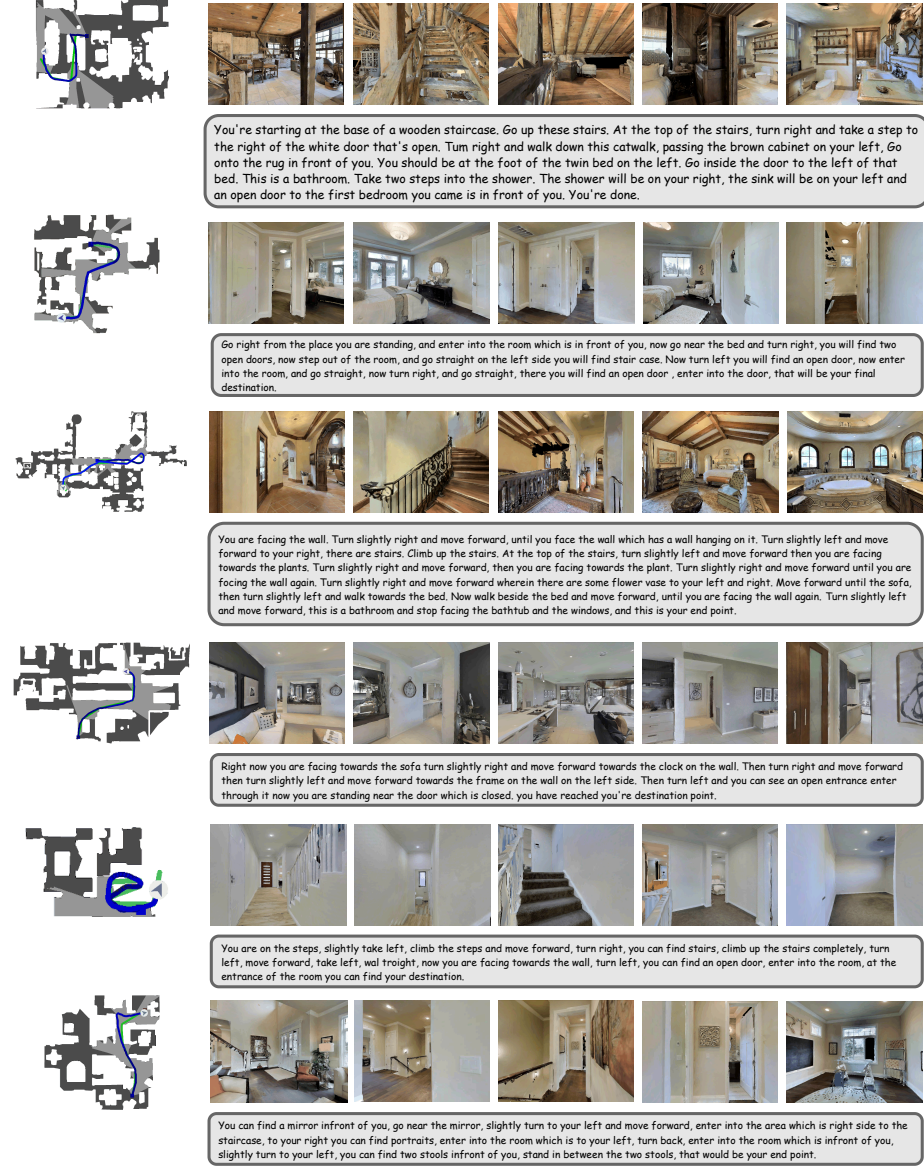


图 11: NAVIDA 在 VLN-CE RxR Val-Unseen 划分上的定性结果。

参考文献

1. An, D., Qi, Y., Li, Y., Huang, Y., Wang, L., Tan, T., Shao, J.: Bevbort: Multimodal map pre-training for language-guided navigation. arXiv preprint arXiv:2212.04385 (2022)
2. An, D., Wang, H., Wang, W., Wang, Z., Huang, Y., He, K., Wang, L.: Etpnav: Evolving topological planning for vision-language navigation in continuous environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024)
3. Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., Van Den Hengel, A.: Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 3674–3683 (2018)
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
5. Cai, W., Peng, J., Yang, Y., Zhang, Y., Wei, M., Wang, H., Chen, Y., Wang, T., Pang, J.: Navdp: Learning sim-to-real navigation diffusion policy with privileged information guidance. arXiv preprint arXiv:2505.08712 (2025)
6. Chen, S., Guhur, P.L., Schmid, C., Laptev, I.: History aware multimodal transformer for vision-and-language navigation. *Advances in neural information processing systems* **34**, 5834–5847 (2021)
7. Cheng, A.C., Ji, Y., Yang, Z., Zou, X., Kautz, J., Biyik, E., Yin, H., Liu, S., Wang, X.: Navila: Legged robot vision-language-action model for navigation. *Robotics: Science and Systems* (2025)
8. Chi, C., Xu, Z., Feng, S., Cousineau, E., Du, Y., Burchfiel, B., Tedrake, R., Song, S.: Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research* **44**(10-11), 1684–1704 (2025)
9. Contributors, I.: InternNav: InternRobotics’ open platform for building generalized navigation foundation models. <https://github.com/InternRobotics/InternNav> (2025)
10. Dou, Z.Y., Gao, F., Peng, N.: Masked path modeling for vision-and-language navigation. arXiv preprint arXiv:2305.14268 (2023)
11. Georgakis, G., Schmeckpeper, K., Wanchoo, K., Dan, S., Mitsakaki, E., Roth, D., Daniilidis, K.: Cross-modal map learning for vision and language navigation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 15460–15470 (2022)
12. Hong, Y., Wang, Z., Wu, Q., Gould, S.: 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*. pp. 15439–15449 (2022)
13. Hong, Y., Wu, Q., Qi, Y., Rodriguez-Opazo, C., Gould, S.: Vln bert: A recurrent vision-and-language bert for navigation. In: *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*. pp. 1643–1653 (2021)
14. Huang, Y., Wu, M., Li, R., Tu, Z.: Vista: Generative visual imagination for vision-and-language navigation. arXiv preprint arXiv:2505.07868 (2025)
15. Krantz, J., Gokaslan, A., Batra, D., Lee, S., Maksymets, O.: Waypoint models for instruction-guided navigation in continuous environments. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 15162–15171 (2021)

16. Krantz, J., Lee, S.: Sim-2-sim transfer for vision-and-language navigation in continuous environments. In: European conference on computer vision. pp. 588–603. Springer (2022)
17. Krantz, J., Wijmans, E., Majumdar, A., Batra, D., Lee, S.: Beyond the nav-graph: Vision-and-language navigation in continuous environments. In: European Conference on Computer Vision. pp. 104–120. Springer (2020)
18. Ku, A., Anderson, P., Patel, R., Ie, E., Baldridge, J.: Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. arXiv preprint arXiv:2010.07954 (2020)
19. Li, J., Bansal, M.: Improving vision-and-language navigation by generating future-view image semantics. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10803–10812 (2023)
20. Liu, R., Wang, X., Wang, W., Yang, Y.: Bird’s-eye-view scene graph for vision-language navigation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10968–10980 (2023)
21. Liu, Z., Zhu, L., Shi, B., Zhang, Z., Lou, Y., Yang, S., Xi, H., Cao, S., Gu, Y., Li, D., et al.: Nvila: Efficient frontier visual language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4122–4134 (2025)
22. Long, Y., Cai, W., Wang, H., Zhan, G., Dong, H.: Instructnav: Zero-shot system for generic instruction navigation in unexplored environment. arXiv preprint arXiv:2406.04882 (2024)
23. Nguyen, K., Dey, D., Brockett, C., Dolan, B.: Vision-based navigation with language-based assistance via imitation learning with indirect intervention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12527–12537 (2019)
24. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
25. Raychaudhuri, S., Wani, S., Patel, S., Jain, U., Chang, A.X.: Language-aligned waypoint (law) supervision for vision-and-language navigation in continuous environments. arXiv preprint arXiv:2109.15207 (2021)
26. Ross, S., Gordon, G., Bagnell, D.: A reduction of imitation learning and structured prediction to no-regret online learning. In: Proceedings of the fourteenth international conference on artificial intelligence and statistics. pp. 627–635. JMLR Workshop and Conference Proceedings (2011)
27. Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., et al.: Habitat: A platform for embodied ai research. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9339–9347 (2019)
28. Schumann, R., Zhu, W., Feng, W., Fu, T.J., Riezler, S., Wang, W.Y.: Velma: Verbalization embodiment of llm agents for vision and language navigation in street view. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 18924–18933 (2024)
29. Wang, H., Liang, W., Gool, L.V., Wang, W.: Towards versatile embodied navigation. *Advances in neural information processing systems* **35**, 36858–36874 (2022)
30. Wang, H., Liang, W., Van Gool, L., Wang, W.: Dreamwalker: Mental planning for continuous vision-language navigation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10873–10883 (2023)
31. Wang, L., Xia, X., Zhao, H., Wang, H., Wang, T., Chen, Y., Liu, C., Chen, Q., Pang, J.: Rethinking the embodied gap in vision-and-language navigation: A holis-

- tic study of physical and visual disparities. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9455–9465 (2025)
32. Wang, S., Wang, Y., Li, W., Cai, X., Wang, Y., Chen, M., Wang, K., Su, Z., Li, D., Fan, Z.: Aux-think: Exploring reasoning strategies for data-efficient vision-language navigation. arXiv preprint arXiv:2505.11886 (2025)
 33. Wang, S., Wang, Y., Li, W., Wang, Y., Chen, M., Wang, K., Su, Z., Cai, X., Jin, Y., Li, D., et al.: Monodream: Monocular vision-language navigation with panoramic dreaming. arXiv preprint arXiv:2508.02549 (2025)
 34. Wang, Z., Li, X., Yang, J., Liu, Y., Jiang, S.: Gridmm: Grid memory map for vision-and-language navigation. In: Proceedings of the IEEE/CVF International conference on computer vision. pp. 15625–15636 (2023)
 35. Wang, Z., Li, X., Yang, J., Liu, Y., Jiang, S.: Sim-to-real transfer via 3d feature fields for vision-and-language navigation. arXiv preprint arXiv:2406.09798 (2024)
 36. Wang, Z., Li, J., Hong, Y., Wang, Y., Wu, Q., Bansal, M., Gould, S., Tan, H., Qiao, Y.: Scaling data generation in vision-and-language navigation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12009–12020 (2023)
 37. Wei, M., Wan, C., Yu, X., Wang, T., Yang, Y., Mao, X., Zhu, C., Cai, W., Wang, H., Chen, Y., et al.: Streamvln: Streaming vision-and-language navigation via slowfast context modeling. arXiv preprint arXiv:2507.05240 (2025)
 38. Wu, Q., Gong, X., Xu, K., Manocha, D., Dong, J., Wang, J.: Towards target-driven visual navigation in indoor scenes via generative imitation learning. IEEE Robotics and Automation Letters **6**(1), 175–182 (2020)
 39. Wu, S., Fu, X., Wu, F., Zha, Z.J.: Vision-and-language navigation via latent semantic alignment learning. IEEE Transactions on Multimedia **26**, 8406–8418 (2024)
 40. Yao, X., Gao, J., Xu, C.: Navmorph: A self-evolving world model for vision-and-language navigation in continuous environments. arXiv preprint arXiv:2506.23468 (2025)
 41. Zeng, S., Qi, D., Chang, X., Xiong, F., Xie, S., Wu, X., Liang, S., Xu, M., Wei, X.: Janusvln: Decoupling semantics and spatiality with dual implicit memory for vision-language navigation. arXiv preprint arXiv:2509.22548 (2025)
 42. Zhang, J., Li, A., Qi, Y., Li, M., Liu, J., Wang, S., Liu, H., Zhou, G., Wu, Y., Li, X., et al.: Embodied navigation foundation model. arXiv preprint arXiv:2509.12129 (2025)
 43. Zhang, J., Wang, K., Wang, S., Li, M., Liu, H., Wei, S., Wang, Z., Zhang, Z., Wang, H.: Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks. Robotics: Science and Systems (2025)
 44. Zhang, J., Wang, K., Xu, R., Zhou, G., Hong, Y., Fang, X., Wu, Q., Zhang, Z., Wang, H.: Navid: Video-based vlm plans the next step for vision-and-language navigation. Robotics: Science and Systems (2024)
 45. Zhang, P., Su, Y., Wu, P., An, D., Zhang, L., Wang, Z., Wang, D., Ding, Y., Zhao, B., Li, X.: Cross from left to right brain: Adaptive text dreamer for vision-and-language navigation. arXiv preprint arXiv:2505.20897 (2025)
 46. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713 (2024)
 47. Zhang, Z., Zhu, W., Pan, H., Wang, X., Xu, R., Sun, X., Zheng, F.: Activevln: Towards active exploration via multi-turn rl in vision-and-language navigation. arXiv preprint arXiv:2509.12618 (2025)