

具身导航基础模型

Jiazhao Zhang^{1,2,*} Anqi Li^{1,2,*} Yunpeng Qi^{3,4,*} Minghan Li^{2,*} Jiahang Liu²
Shaoan Wang¹ Haoran Liu^{1,2} Gengze Zhou⁵ Yuze Wu⁶ Xingxing Li⁶
Yuxin Fan⁶
Wenjun Li⁶ Zhibo Chen³ Fei Gao^{6,7} Qi Wu⁵ Zhizheng Zhang^{2,4,†} He Wang^{1,2,4,†}

¹ 北京大学 ² 银河通用 ³ 中国科学技术大学 ⁴ 北京智源人工智能研究院

⁵ 阿德莱德大学 ⁶ 浙江大学 ⁷ Differential Robotics

项目主页: <https://pku-epic.github.io/NavFoM-Web/>

摘要

导航是具身智能的一项基础能力，体现了在物理环境中进行感知与交互所需的智能。为获得此类智能，近期研究借助视觉—语言模型（Vision-Language Model, VLM）；这类模型具有很强的泛化能力，其建模形式也很适合导航。然而，现有方法大多仍局限于狭窄的任务设定和具身平台特定的架构。本文提出跨具身、跨任务的导航基础模型（Navigation Foundation Model, NavFoM）。该模型基于八百万条导航样本训练，覆盖四足机器人、无人机、轮式机器人和车辆，以及视觉—语言导航（Vision-and-Language Navigation, VLN）、目标搜索、目标跟踪和自动驾驶等多种任务。NavFoM采用统一架构，能够处理来自不同相机配置和不同导航时间跨度的多模态导航输入。为适配多样的相机配置与时间跨度，NavFoM 引入指示词元，将具身平台的相机视点信息与任务的时间上下文嵌入其中。为满足真实世界部署需求，NavFoM 还在有限的词元长度预算内采用动态调整的采样策略，控制全部观测词元。在七个公开基准上的广泛评估表明，无需针对特定任务微调，本模型便能在不同导航任务与具身平台上取得最先进或极具竞争力的性能。更多真实世界实验进一步验证了本方法的强泛化能力和实际适用性。

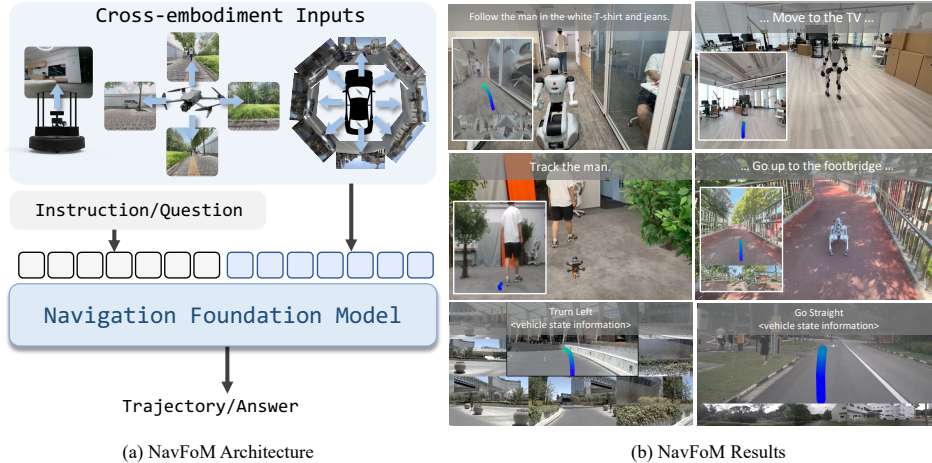


图 1: 本文展示了模型架构（左）和真实世界实验结果（右）。实验结果涵盖本方法的跨任务与跨具身性能。

* 共同第一作者 † 通讯作者

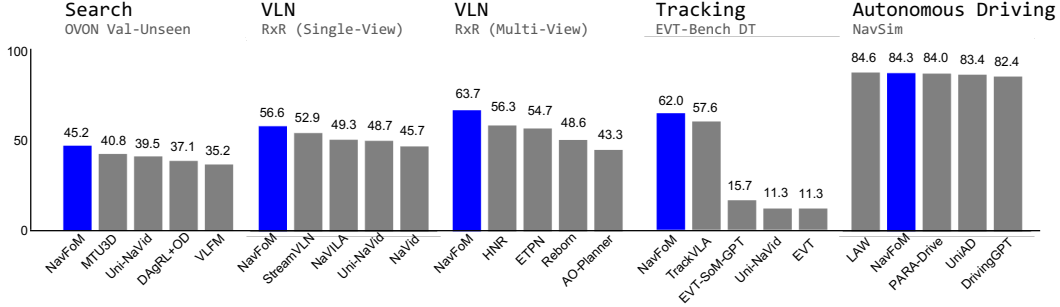


图 2: NavFoM 的基准性能: 在每个基准上将 NavFoM 与最先进 (State-of-the-Art, SOTA) 基线进行比较。更多细节参见第 4 节。

1 引言

对具身智能体和人类而言, 导航都是一项基础能力, 使其能够在物理环境中智能移动并完成指定任务 (Shah et al., 2023a; Bar et al., 2025; Zhang et al., 2024b)。稳健导航要求深入理解环境上下文和任务指令; 这些信息通常以视觉与语言观测的形式呈现, 与视觉-语言模型 (Vision-Language Model, VLM) 的输入形式相似。近年来, VLM 从大规模开放世界数据中学习, 在检索、分类和图像描述等任务上展现出显著的零样本泛化能力, 且无须依赖领域特定微调 (Liu et al., 2023a; Yang et al., 2024a; Guo et al., 2025)。相比之下, 具身导航 (Savva et al., 2019a; Deitke et al., 2022) 仍受限于狭窄的任务领域、具身平台特定的架构以及受约束的指令形式。

通用导航日益受到研究关注 (Zhang et al., 2024a; Cheng et al., 2025; Shah et al., 2023a; Long et al., 2024), 但既有研究受限的设计和有限的领域适用性阻碍了相关进展。在跨任务导航中, 先前方法 (Zhang et al., 2025a; Yin et al., 2025; Zhu et al., 2025) 通常假设机器人采用一致的相机配置, 并统一视觉-语言导航、目标搜索和目标跟踪等多种任务。在跨具身导航中, 现有方法 (Eftekhari et al., 2024; Hirose et al., 2023) 隐式学习具身平台物理外形的先验, 但往往局限于特定导航任务。导航任务与具身平台之间现存的割裂表明, 领域内仍缺少能够在多种具身平台上处理不同任务的导航基础模型。

本文致力于构建跨任务、跨具身的具身导航基础模型 NavFoM, 并使用覆盖多种具身平台与任务的八百万条导航样本进行训练。人类主要依靠视觉感知输入即可完成广泛的导航任务, 近期纯视觉导航方法也取得了成功 (Shah et al., 2023a; Zeng et al.)。受此启发, 本文将通用导航建模为: 处理自我中心视频 (由安装在机器人上的一个或多个相机采集) 与语言指令, 并预测后续轨迹以完成指令。该形式与多数现有导航任务设定兼容 (Contributors, 2023; Wang et al., 2024a)。

为在不同相机配置下对齐可泛化的具身平台, 本文提出时间-视点指示 (Temporal-Viewpoint Indicator, TVI) 词元, 同时标识相机配置的视点与导航时间跨度的时间信息。通过动态调整 TVI 词元, 本方法能够跨不同相机配置进行联合微调, 并支持与图像问答和视频问答样本共同训练 (Shen et al., 2024; Li et al., 2023)。为应对硬件内存开销和推理速度等实际部署约束, 本文进一步提出预算感知时间采样 (Budget-Aware Temporal Sampling, BATS) 策略, 根据受词元预算约束的遗忘曲线动态采样导航历史词元。这一词元采样方法在性能与推理速度之间取得平衡, 增强了真实世界部署的实用性。

本文汇集了一个全面且多样的导航数据集, 共包含 802 万条样本, 来源包括公开导航数据集 (Savva et al., 2019a; Wang et al., 2025c; Contributors, 2023; Wang et al., 2024a) 和 Web 视频伪导航数据 (Li et al., 2025a)。该数据集包含来自四足机器人、无人机、轮式机器人和车辆的跨具身轨迹, 覆盖视觉-语言导航、目标搜索、目标跟踪和自动驾驶等广泛任务。这些导航样本包含多样的指令和需要多项技能的场景, 使 NavFoM 能够获得泛化的导航能力。本文还收集了 476 万条开放世界知识样本 (Shen et al., 2024; Li et al., 2023), 这些样本来自图像问答和视频问答任务。遵循 (Zhang et al., 2024a) 的方法, 以端到端方式对导航数据及图像、视频问答数据进行联合微调, 从而支持 NavFoM 的大规模综合训练。

实验表明, NavFoM 显著推进了通用导航研究。无需针对特定任务微调, NavFoM 便在面向多种具身平台的各类公开基准上达到最先进 (State-of-the-Art, SOTA) 或具有竞争力的性能。在 VLN-CE RxR (Ku et al., 2020) 上, 与先前基线相比, NavFoM 在多相机设定下

将成功率 (SR) 从 56.3% 提升至 64.4%，在单相机设定下则从 51.8% 提升至 57.4%。在 HM3D-OVON (Yokoyama et al., 2024b) 上，本方法在零样本设定下达到 45.2% 的 SR，超过先前经微调的 SOTA 方法 (43.6% SR)。在目标搜索、跟踪和自动驾驶的多个基准上也取得了同样优异的结果。本文进一步在多种真实机器人平台上验证 NavFoM，包括人形机器人、四足机器人、无人机和轮式机器人。这些结果凸显了模型的强泛化能力，也表明通用导航研究取得了可喜进展。

2 相关工作

用于导航的大模型。将大语言模型 (Large Language Model, LLM) 和视觉—语言模型 (VLM) 融入机器人导航，推动该领域从传统的学习方法转向利用预训练知识获得开放世界理解与强泛化能力。一条直接的技术路线 (Zhou et al., 2023; Shah et al., 2022; Qiao et al., 2023) 是以零样本方式使用现成的 LLM。这类工作通过思维链机制 (Pan et al., 2023; Long et al., 2023; Lin et al., 2025) 和结构化推理框架 (Chen et al., 2024b; Qiao et al., 2025) 增强可解释性。然而，将稠密视觉信息抽象为文本会造成环境观测稀疏，且仅适用于静态环境。另一条路线 (Cheng et al., 2025; Zhang et al., 2024a; 2025a; Wei et al., 2025; Wang et al., 2025c) 使用导航数据对基于视频或图像 (Zhou et al., 2025; Zheng et al., 2024; Zhang et al., 2025c) 的 VLM 进行端到端微调，使其掌握导航能力。然而，现有方法大多关注同构具身平台，忽视了不同具身平台与任务之间潜在的训练协同效应。本文初步尝试将导航策略拓展至更广泛的跨具身、跨任务导航领域。

跨具身导航。开发能够在形状、尺寸和传感器配置各异的具身平台之间泛化的导航模型，仍是具身智能领域的一项重大挑战。近期研究 (Shah et al., 2023a;b; Yang et al., 2024b; Wang et al., 2020; Eftekhari et al., 2024; Hirose et al., 2023; Putta et al., 2024; Curtis et al., 2024; Wang et al., 2025a; Zhang et al., 2025b) 表明，在大规模跨具身数据集上训练的基于 Transformer 的策略，无需人工对齐观测空间与动作空间，便有望在多种机器人平台上取得稳健性能。然而，这些模型处理多模态输入时通常未纳入明确的时空线索；由于不同具身平台对数据的几何解释各异，这可能引发歧义。该方式还可能造成数据利用效率低下，并限制模型向分布外具身平台的泛化能力。(Eftekhari et al. (2024); Wang et al. (2025b)) 相比之下，NavFoM 引入了编码观测配置的时间—空间指示词元，使模型能够更准确地理解来自不同具身平台的多模态输入。

跨任务导航。具身智能领域的最新进展 (O’Neill et al., 2024; Team et al., 2024; Kim et al.; Bjorck et al., 2025; Black et al., 2024; Intelligence et al., 2025; Bu et al., 2025b;a; Qu et al., 2025) 表明，基于基础模型构建的通用模型能够在多种任务之间有效迁移知识。在导航领域，先前研究 (Zhou et al., 2024; Wang et al., 2022; Long et al., 2024; Song et al., 2025; Zhang et al., 2025a; Gao et al., 2025; Yin et al., 2025; Ruan et al., 2025) 表明，整合不同类型导航任务的数据可以提升模型在各类导航场景中的性能。早期工作 VIENNA (Wang et al., 2022) 使用强化学习在模拟器中训练智能体。近期，Uni-Navid (Zhang et al., 2025a) 基于视频视觉—语言模型开发了通用模型，通过视觉—语言导航、目标物体导航、具身问答和人员跟随四类任务上的跨任务学习，掌握泛化的导航技能。然而，这些方法局限于受约束的环境（例如室内可控环境），本文则拓展至更广泛的场景（例如自动驾驶和无人机导航），并以统一形式涵盖所有任务。在该框架中，模型以 RGB 视频和自然语言指令为输入，输出可执行轨迹。

3 方法

通用导航任务。本文考虑一种通用导航设置：移动具身接收文本指令 L ，以及在时间步 $\{1, \dots, T\}$ 由 N 个不同相机在线采集的图像序列 $I_{1:T}^{1:N} \in \mathbb{R}^{W \times H \times 3}$ 。给定这些观测与指令，模型 π 需要预测导航轨迹 $\tau = \{\mathbf{a}_1, \mathbf{a}_2, \dots\}$ ，其中每个 $\mathbf{a} \in \mathbb{R}^4 = (x, y, z, \theta)$ 表示一个包含位置与朝向的航点。 z 仅用于无人机 (UAV) 具身， θ 表示偏航角；由于本任务不要求敏捷飞行动作，使用偏航角即可。模型按照映射 $\pi(L, I_{1:T}^{1:N}) \mapsto \tau_T$ 驱动移动具身完成指令。

基本架构。我们将基础的视觉—语言模型 (Vision-Language Model, VLM) (Li et al., 2023; Shen et al., 2024) 扩展为同时处理导航与问答的双分支架构 (Wang et al., 2025c)。对于导航任务，首先使用视觉编码器和跨模态投影器 (Cross-Modality Projector) (Liu et al., 2023a) 对观测图像 $I_{1:T}^{1:N}$ 编码，得到视觉词元 $E_{1:T}^{1:N}$ 。随后遵循现有语言模型的常见做法 (Liu et al., 2023a) 嵌入指令，生成语言词元 E_L 。视觉词元经由时间—视点指示 (Temporal-Viewpoint Indicator, TVI) 词元 (第 3.1.1 节) 与预算感知时间采样 (Budget-Aware Temporal Sampling, BATS) (第 3.1.2 节) 组织后，与语言词元拼接并输入大语言

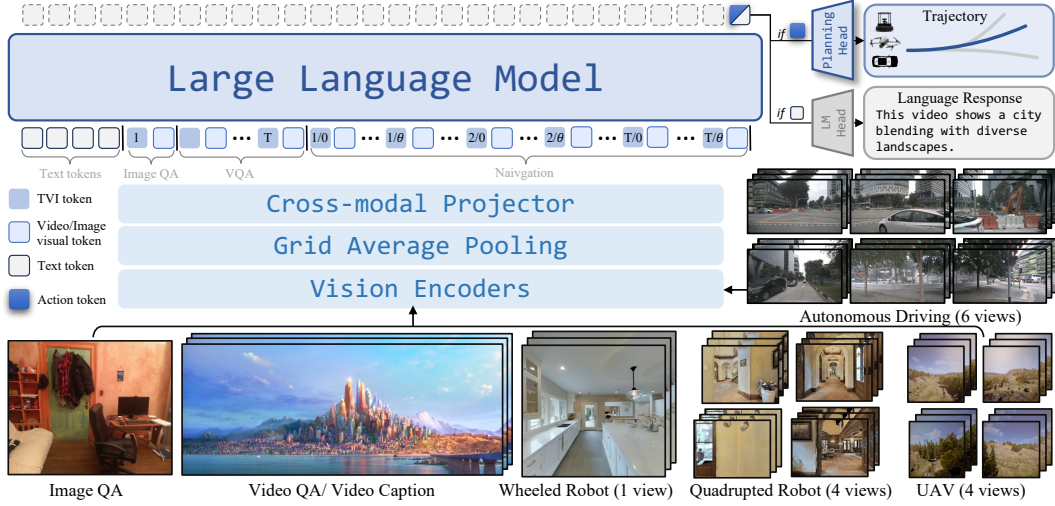


图 3: **NavFoM** 的流程。本方法为图像问答、视频问答和导航等多种任务提供了统一框架。如第 3.1.3 节所述, 我们使用时间—视点指示 (Temporal-Viewpoint Indicator, TVI) 词元 (第 3.1.1 节) 组织文本词元与视觉词元。对于问答任务, 模型采用常规语言建模头进行自回归预测; 对于导航任务, 则使用规划头直接预测轨迹。

模型 (Large Language Model, LLM), 以预测动作词元。该词元随后由规划模型解码, 生成基于航点的轨迹。

$$\begin{aligned} E_T^A &= \text{LLM}(E_{1:T}^{1:N}, E_L), \\ \tau_T &= \text{ActionModel}(E_T^A). \end{aligned} \quad (1)$$

对于问答任务, 遵循现有方法 [Liu et al. \(2023a\)](#), 以自回归方式进行下一词元预测。与已有工作 ([Zhang et al., 2024a; 2025a; Wang et al., 2025c; Cheng et al., 2025](#)) 一致, 本模型支持对导航与问答样本进行联合微调。

3.1 导航基础模型

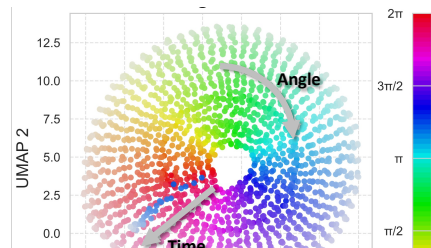
观测编码。 给定时间步 T 从 N 个相机视点采集的自我中心 RGB 序列 $I_{1:T}^{1:N} \in \mathbb{R}^{W \times H \times 3}$, 我们采用预训练视觉编码器 (DINOv2 ([Oquab et al., 2023](#)) 与 SigLIP ([Zhai et al., 2023](#)), 这是广泛使用的组合 ([Kim et al.; Tong et al., 2024](#))) 提取视觉特征 $\mathbf{V}_{1:T}^{\text{dino/SigLIP}} \in \mathbb{R}^{P \times C}$, 其中 P 表示图像块数量 (设为 576), C 表示嵌入维度。为节省词元并提高计算效率, 我们沿通道维直接拼接 $\mathbf{V}_{1:T}^{\text{dino}}$ 与 $\mathbf{V}_{1:T}^{\text{siglip}}$, 并将所得表示记为 $\mathbf{V}_{1:T}$ 。导航过程中在线采集的视频会产生大量帧, 进而形成庞大的视觉特征集合。为此, 我们对视觉特征采用网格平均池化策略 ([Zhang et al., 2024a; 2025a](#)) (图 3, Grid Average Pooling), 生成更紧凑的表示。具体而言, 使用两个分辨率尺度:

$$\mathbf{V}^{\text{fine/coarse}} = \text{GridPool}(\mathbf{V}, \frac{64}{P} \text{ or } \frac{4}{P}), \quad (2)$$

其中, $\mathbf{V}^{\text{fine}} \in \mathbb{R}^{64 \times C}$ 提供细粒度观测, $\mathbf{V}^{\text{coarse}} \in \mathbb{R}^{4 \times C}$ 提供粗粒度观测。对于最新导航观测和图像问答 (时间步 T), 采用细粒度特征 $\mathbf{V}^{\text{fine}}$; 对于导航历史与视频数据 (跨时间步 $1:T$), 则采用粗粒度特征。最后, 遵循已有 VLM ([Liu et al., 2023a; Li et al., 2023](#)), 使用跨模态投影器 $\mathcal{P}(\cdot)$ (一个两层多层感知机, 即 MLP) 将视觉特征投影至 LLM 的潜在空间: $\mathbf{E}_T^V = \mathcal{P}(\mathbf{V}_{1:T}^{\text{fine/coarse}})$ 。

3.1.1 时间—视点指示 (TEMPORAL-VIEWPOINT INDICATOR, TVI) 词元。

视觉词元本身不包含视点和时间信息, 因此, 多视点导航模型面临的关键挑战是使 LLM 能够辨别哪些词元对应不同时间步或相机视点。以往方法或局限于特定相机配置或具身 ([Long et al., 2024; Gao et al.,](#)



2025), 或仅拼接所有视点图像的词元 (Zheng et al., 2024; Fu et al., 2025b), 未能发挥 LLM 词元组织的灵活性。为灵活处理任意相机布局, 我们提出 TVI 词元; 其设计受专用词元在时间、模态及任务标识方面有效性的启发 (Guo et al., 2025; Chen et al., 2023), 而此类设计已被广泛证实有助于 LLM 学习。在本文设置中, 指示词元用于图像问答、视频问答和导航等多种任务, 并应满足三个重要属性:

- **视点感知**: 词元的角度嵌入必须保持方位角的环状连续性 (例如, $0 \equiv 2\pi$), 使嵌入间的距离度量能够反映几何邻近性 (例如, 当 $\epsilon \neq \pi$ 时, $d(0, \epsilon) < d(0, \pi)$)。
- **时间感知**: 词元必须能够唯一标识所有相机视点下各帧的时间顺序, 同时对不规则采样间隔保持稳健。
- **可分离性**: 指示词元可以仅编码视点或时间信息 (用于视频问答), 也可以完全不编码这些信息 (用于图像问答)。

为满足这些要求, TVI 词元 $\mathbf{E}_{\text{TVI}} \in \mathbb{R}^C$ (时间步和视角分别记为 t 与 ϕ) 由三类嵌入组成: 角度嵌入 $\text{AnglePE}(\phi) \in \mathbb{R}^C$ 、时间嵌入 $\text{TimePE}(t) \in \mathbb{R}^C$, 以及可学习基础嵌入 $\mathbf{E}_{\text{Base}} \in \mathbb{R}^C$:

$$\mathbf{E}_{\text{TVIT}} = \begin{cases} \mathbf{E}_{\text{Base}} + \mathcal{P}_{\text{time}}(\text{TimePE}(t)) + \mathcal{P}_{\text{angle}}(\text{AnglePE}(\phi)), & \text{if Navigation} \\ \mathbf{E}_{\text{Base}} + \mathcal{P}_{\text{time}}(\text{TimePE}(t)), & \text{if Video QA} \\ \mathbf{E}_{\text{Base}}, & \text{if Image QA} \end{cases} \quad (3)$$

$\text{AnglePE}(\phi)$ 的实现方式是: 分别对方位角的余弦值与正弦值应用正弦位置编码 (Vaswani et al., 2017), 再将二者拼接; $\text{TimePE}(t)$ 则实现为 t 的正弦位置编码。 $\mathcal{P}_{\text{time}}$ 与 $\mathcal{P}_{\text{angle}}$ 均采用两层 MLP, 其设计与 Liu et al. (2023a) 类似。针对不同任务和 TVI 词元, 我们采用不同的指示词元组件组合来表示各类视觉词元的属性。导航任务同时包含时间与视点信息; 视频问答任务包含时间信息; 图像问答任务仅使用 $\mathbf{E}_{\text{Base}}$, 用于指示其后的词元为视觉词元。该策略能够灵活组织差异显著的样本类型, 并有助于 LLM 学习 (第 3.1.3 节)。图 4 展示了 TVI 词元的聚类结果 (McInnes et al., 2018): 各词元可按视点 θ (以彩虹色条表示) 和时间步 t (以颜色明度表示) 彼此区分。

尽管还可采用位置编码等其他方式引入多视点信息 (Zheng et al., 2024), 实验结果 (表 6) 表明, 使用额外指示词元在训练期间最为稳健, 并在评估中表现良好。我们认为, 这是因为添加此类词元不会扰动现成的视觉词元空间。相关文献 (Guo et al., 2025; Chen et al., 2023) 也报告了这一发现。

3.1.2 预算感知时间采样 (BUDGET-AWARE TEMPORAL SAMPLING, BATS)。

导航过程中在线采集的视频会产生过多视觉词元, 增加推理与训练时间, 并阻碍真实世界部署。以往方法主要通过两种方式应对这一挑战: (1) 词元合并 (Token Merging) (Zhang et al., 2025a), 会在训练期间引入额外计算开销, 并导致评估时的推理速度不一致; (2) 均匀采样 (Uniform Sampling) (Cheng et al., 2025), 由于缺乏短期上下文, 往往无法充分捕获近期观测。此外, 在相机视点数量可变的场景中 (此时帧数显著增加), 两种策略都需要额外调整。

为此, 本文提出预算感知时间采样 (*Budget-Aware Temporal Sampling, BATS*), 其设计目标包括: (a) 满足实际需求, 即限制最大词元长度, 以适应推理速度和 GPU 显存限制; (b) 保留更多近期信息, 在维持充足导航历史上下文的同时增强理解与规划能力; (c) 直接适配不同的相机数量。具体而言, 给定词元预算 B_{token} 和多视点视频序列 $I_{1:T}^{1:N} \in \mathbb{R}^{W \times H \times 3}$, 采用受“遗忘曲线” (Forgetting Curve) 启发的指数增长采样概率 $P(t)$ 。当已采集帧对应的词元数量超过词元预算时, 计算每一帧的采样概率:

$$P(t) = (1 - \epsilon)e^{k(t-T)/T} + \epsilon, \quad k > 0, \quad (4)$$

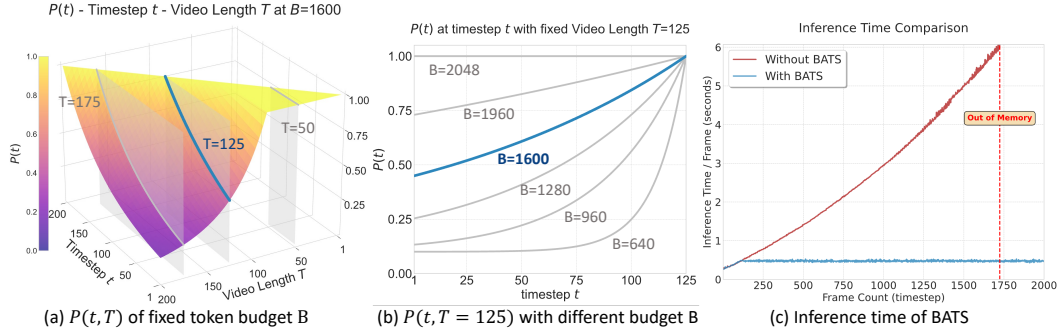


图 5: BATS 及其相应时间开销的可视化。(a) 在固定词元预算 $B = 1600$ 下, 展示最新时间步为 T 时不同时间步 t 的采样概率。(b) 在最大时间步 $T = 125$ 下, 绘制不同词元预算 B 对应的各时间步 t 采样概率。(c) 对比使用 BATS 与不使用 BATS (保留所有帧) 时的推理时间。

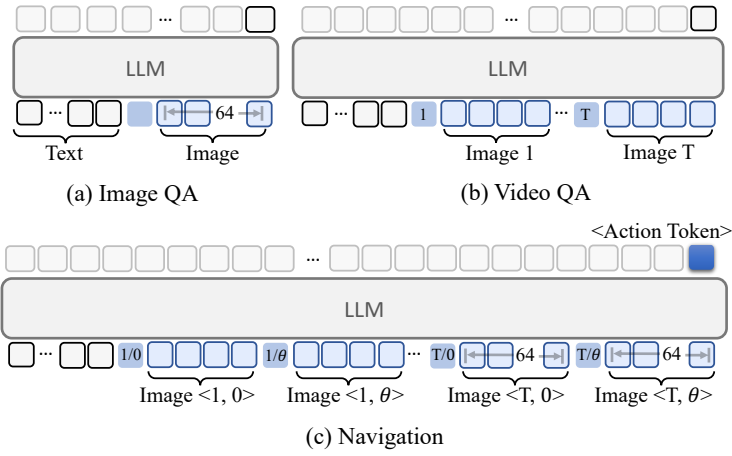


图 6: NavFoM 在不同任务中的词元组织策略。(a) 对于图像问答, 采用细粒度视觉词元, 并仅引入 TVI 词元的基础嵌入。(b) 对于视频问答, 采用粗粒度视觉词元, 其中同时包含 TVI 词元的基础嵌入和时间嵌入。(c) 对于导航, 同时采用粗粒度和细粒度视觉词元, 并融合 TVI 词元的基础嵌入、时间嵌入与角度嵌入。

其中, ϵ (本文取 $\epsilon = 0.1$) 确保采样概率下界处于合理范围, k 表示指数衰减率。因此, 采样帧数的期望可计算为:

$$\mathbb{E}_{\text{frames}} \approx \int_0^T P(t) dt = (1 - \epsilon) \frac{1 - e^{-k}}{k} T + \epsilon T \quad (5)$$

我们约束词元数量的期望 $((4 + 1)\mathbb{E}_{\text{frame}} + (64 + 1))N$ 不超过 B_{token} 。由此可得 $\mathbb{E}_{\text{frame}} \leq \frac{B_{\text{token}} - (64 + 1)N}{(4 + 1)N}$; 当帧数 T 足够大时, 采样帧数将收敛于其期望 (图 5 (c))。可以采用 Brent 方法 (Brent, 2013) 离线计算不同 T 对应的 k , 进而得到相应的 $P(t)$ (式 4)。由于设置了概率下界 ϵ , 当 T 非常大时, 式 5 可能无解 (例如, 在四相机设置且词元预算 $B_{\text{token}} = 2048$ 时, $T = 1120$)。不过, 这种情况很少出现 (对于图 2 所列任务), 因为 VLN-CE RxR (Ku et al., 2020) 中大多数轨迹约为 122 个时间步。附录 A.2 给出了 BATS 的使用细节。

图 5 展示了时间步采样概率的分布与时间效率。可以看出, 本方法在不同词元预算 B 和时间步 T 下均能平滑地得到合理的 $P(t)$ 。词元预算较高时, BATS 会自适应采样更多历史词元; 即便词元数量较少, 该策略仍能维持合理的概率下界。此外, BATS 在整个导航过程中都保持稳定的推理速度。

3.1.3 LLM 前向传播

词元组织。 获得视觉词元 $E_{1:T}^{1:N}$ (通过 BATS 采样, 见第 3.1.2 节) 与语言词元 E_L 后, 使用 TVI 词元 (第 3.1.1 节) 对其进行组织, 再通过 LLM 前向传播。图 6 详细展示了不

同任务的词元组织策略。对于图像问答，使用 E_{Base} 与细粒度视觉词元（每幅图像 64 个词元）共同表示图像。对于视频问答，引入 $E_{\text{Base}} + \mathcal{P}_{\text{time}}(\text{TimePE}(t))$ 为每一帧编码时间信息，并采用粗粒度视觉词元（每帧 4 个词元）以避免词元数量过多。对于导航，使用 $E_{\text{Base}} + \mathcal{P}_{\text{time}}(\text{TimePE}(t)) + \mathcal{P}_{\text{angle}}(\text{AnglePE}(\phi))$ 同时表示时间和视点信息。其中，细粒度视觉词元用于最新观测，粗粒度词元用于历史观测。该词元组织策略增强了 LLM 对输入词元的理解，并支持统一处理图像问答、视频问答与导航任务。

轨迹预测。对于导航任务，给定 LLM 前向传播所得的预测动作隐状态 E_T^A ，使用规划模型 $\mathcal{A}_\theta$ （实现为三层 MLP）提取轨迹信息 τ_T 。原始轨迹的尺度可能从数米（室内导航）到数十米（自动驾驶和无人机）不等，直接预测原始轨迹可能导致航点分布发散。因此，遵循以往方法 [Shah et al. \(2023a\)](#)，使用任务特定的缩放因子 α_{task} ，将轨迹航点归一化至 $[-1, 1]$ 分布。对于室内导航、无人机和汽车，分别采用三个不同的缩放因子，详见附录 A.1。轨迹预测可表示为：

$$\tau_T = \{\mathbf{a}_1, \dots, \mathbf{a}_M\}_T = \alpha_{\text{task}} \cdot \mathcal{A}_\theta(E_T^A), \quad (6)$$

其中， M 设为 8；将归一化轨迹乘以 α_{task} ，即可重新缩放至绝对值。轨迹损失采用均方误差（Mean Squared Error, MSE）计算，即 $L_{\text{nav}} = \text{MSE}(\tau_{\text{idx}}^{\text{idx}}, \tau_{\text{gt}}^{\text{idx}})$ ，其中 idx 表示有效动作索引。对于轮式机器人或汽车具身， $\mathbf{a}^{\text{idx}} = (x, y, \theta)$ ；对于无人机， $\mathbf{a}^{\text{idx}} = (x, y, z, \theta)$ 。对于问答任务，在下一词元预测监督框架下采用交叉熵损失 L_{QA} 。给定同时包含导航与问答样本的批次，总损失定义为 $L = \beta L_{\text{nav}} + L_{\text{QA}}$ 。其中， β 是用于放大导航损失的常数缩放因子（设为 10）；由于导航损失源于均方误差，其数值通常较小。

3.2 数据

为微调 NavFoM，我们收集并处理了大规模且全面多样的训练样本，共计 1270 万条，其中包括 802 万条导航样本、315 万条图像问答样本和 161 万条视频问答样本。训练样本规模超过以往方法 [\(Zhang et al., 2024a; 2025a\)](#)。导航样本覆盖多种移动具身（轮式机器人、四足机器人、无人机和汽车）及多类任务，包括视觉—语言导航（Vision-and-Language Navigation, VLN）、目标物体导航、主动视觉跟踪和自动驾驶。所有导航数据均以统一方式采集，包括此前录制的视频（单相机与多相机）、指令和预测轨迹航点。对于问答样本，我们参照现有视频视觉—语言模型（Vision-Language Models, VLMs）[\(Shen et al., 2024; Li et al., 2023\)](#)，从现成数据集中收集图像问答和视频问答数据。导航样本的具体情况如下：

视觉—语言导航（VLN）（337 万条）要求智能体理解自然语言指令和第一人称视觉观测，将指令与视觉输入对齐，并规划后续动作以到达指令描述的地标。依据 VLN 的广义定义 [\(Zheng et al., 2024; Wang et al., 2025b; Zhou et al., 2024\)](#)，我们同时考虑室内环境（如 R2R [\(Krantz et al., 2020\)](#) 和 RxR [\(Ku et al., 2020\)](#) 上的 VLN-CE）与室外环境（如 OpenUAV [\(Wang et al., 2024a\)](#)），并分别部署于机器人和无人机（Unmanned Aerial Vehicle, UAV）上。对于 R2R 和 RxR 上的 VLN-CE（294 万条），在机器人沿真实路径移动时采集多视角 RGB 视频、标注指令和轨迹数据。多视角 RGB 配置由固定前视相机和随机采样的周视相机组成（相机总数为 1 至 8 个）。相机高度在 0.6 m 至 1.5 m 之间随机取值，水平视场角（Horizontal Field of View, HFoV）在 75° 至 120° 之间变化。对于 OpenUAV 数据集（42.9 万条），所有序列均记录前、左、右、后四个视角的相机流，其余随机化策略与 VLN-CE 任务一致。

目标物体导航（102 万条）要求机器人探索未见环境并识别指令描述的目标。对于目标物体导航数据集，我们沿用 [\(Zhang et al., 2025a\)](#) 的方法，从 L3MVN [\(Yu et al., 2023\)](#) 中收集成功回合；L3MVN 是一种显式建模探索与识别阶段的启发式方法。数据来自 HM3D ObjectNav [\(Savva et al., 2019a\)](#) 回合，要求智能体定位预定义类别集合中的物体（如沙发、椅子和床）。尽管如此，表 3 的实验表明，本方法可泛化至最先进的开放词汇目标物体搜索。为保持与 L3MVN 相同的视觉观测配置，此处未采用多相机或相机随机化。

主动视觉跟踪（89.7 万条）[\(Islam et al., 2019; Francis et al., 2023; Wang et al., 2025c\)](#) 要求机器人在动态、拥挤的环境中辨别目标。目标通过文本指令指定，例如“跟随穿蓝色 T 恤的男子”。智能体必须识别人类外观，根据指令跟随正确的人，并在避障的同时保持适当距

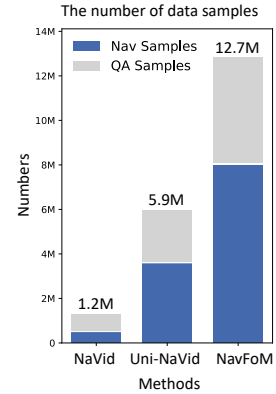


图 7: 与以往方法的训练样本数量比较。

离。与 (Wang et al., 2025c) 一致, 此任务采用 EVT-Bench 数据, 其中包含多样的室内环境、数百个虚拟人物及其对应描述; 同时采用 VLN 数据采集流程所述的相机随机化策略。

自动驾驶 (68.1 万条) (Hu et al., 2023; Liao et al., 2024b) 要求智能体生成安全、舒适且满足运动学约束的轨迹, 以在复杂动态的真实环境中行驶。该任务评估智能体持续感知周围环境、预测其他交通参与者未来运动, 并作出稳健的序列决策, 从而在驶向目的地的同时避免碰撞。我们分别处理来自 nuScenes (Caesar et al., 2020a) 和 OpenScene (Contributors, 2023) 的 2.7 万条与 65.4 万条样本, 直接记录数据集中的原始多视角图像、指令和车辆状态信息。与常见自动驾驶基线 (Chen et al., 2024d; Li et al., 2024a) 不同, 我们不采集车道细节等显式周边信息。

网络视频导航。(203 万条) 我们还采用 Sekai 数据集 (Li et al., 2025a), 该数据集收录约 18.2 万个 YouTube 视频, 并提供相应指令 (由 VLM (Bai et al., 2025) 生成) 和轨迹 (由同步定位与建图系统 (Simultaneous Localization and Mapping, SLAM) (Li et al., 2025b) 生成)。尽管这些导航样本的指令和轨迹并不完美, 但它们仍有助于引入真实世界导航场景; (Cheng et al., 2025; Wei et al., 2025) 也报告了类似发现。

开放世界问答。(476 万条) 参照现有视频 VLM (Li et al., 2023; Shen et al., 2024; Wang et al., 2025c), 我们收集了 315 万条图像问答 (Question Answering, QA) 样本和 161 万条视频 QA 样本, 涵盖开放世界理解所需的丰富且全面的知识。

3.3 实现细节

训练配置。模型在配备 56 块 NVIDIA H100 GPU 的集群服务器上训练约 72 小时, 总计 4032 GPU 小时。对于问答数据, 所有帧均以 1 FPS 采样, 以减少连续帧间的冗余。对于离散导航数据 (如 Habitat 环境 Savva et al. (2019a)), 机器人每执行一次离散动作后采样一步 (离散动作转换为轨迹的细节见附录 A.1)。对于连续导航环境 (如 EVT-Bench Wang et al. (2025c)、自动驾驶 (Caesar et al., 2020b; Contributors, 2023)), 数据以 2 FPS 采样以减少冗余。训练时, 视觉编码器 (DINOv2 Oquab et al. (2023) 和 SigLIP Zhai et al. (2023)) 与大语言模型 (Qwen2-7B Yang et al. (2024a)) 均采用默认预训练权重初始化。遵循 VLM (Liu et al., 2023a) 的训练范式, 仅对指定的可训练参数微调一个轮次。

通过缓存视觉特征加速训练。由于视频的时间跨度较长（数百帧），以大批量方式在线编码所有图像会产生高昂计算开销。为缓解这一问题，我们采用视觉特征缓存机制 (Yan et al., 2022) 并构建视觉特征数据库（见图 8）。由于单个导航回合通常会产生数十个视频，我们仅缓存粗粒度视觉词元（每帧 4 个词元），相比存储完整视频可显著减少磁盘占用。对于图像 QA 与导航中的最新观测，仍使用视觉编码器在线提取细粒度视觉词元（每帧 64 个词元）。该方案将训练速度提高至 2.9 倍，并将 GPU 显存占用降至原来的 1/1.8。

4 实验

4.1 实验设置

为评估 NavFoM 的性能，我们围绕三个关键问题开展了大量实验与消融实验：（1）NavFoM 在不同基准的多种导航任务上表现如何？（2）NavFoM 在真实环境中的表现如何？（3）本方法的关键设计组件是否有效？在各基准上，我们均与强基线进行比较。

基准。我们在 VLN、搜索、跟踪和自动驾驶等多种导航任务上评估本方法，这些任务涵盖不同具身配置（如第一人称视角、四相机、六相机和八相机配置）。在所有基准中，NavFoM 仅以在线采集的第一人称视频（部分来自多视角源）和一条指令为输入，预测供机器人执行的轨迹。鉴于这些基准横跨多个环境和模拟器，我们仔细核验场景划分，确保所有基准的训练场景与验证场景互不重叠。

- **视觉—语言导航 (VLN)：**在 VLN-CE R2R (Krantz et al., 2020) 和 RxR (Ku et al., 2020) 基准的 VAL-Unseen 划分上评估本方法，这些基准要求机器人在未见室内环境中遵循指令。我们还在 Open-UAV 基准 (Wang et al., 2024a) 上评估，该基准要求无人机在未见室外环境中遵循指令。
- **目标物体导航：**遵循以往方法 (Zhang et al., 2025a; Zhu et al., 2025)，在开放词汇目标物体导航基准 HM3D-OVON (Yokoyama et al., 2024b) 上，以零样本方式评估目标物体导航的泛化能力。
- **主动视觉跟踪：**在 EVT-Bench (Wang et al., 2025c) 上评估本方法。该高难度基准要求机器人在拥挤环境中辨别并跟随目标。
- **自动驾驶：**在主流基准 nuScenes (Caesar et al., 2020a) 和 NAVSIM (Dauner et al., 2024b) 上分别进行开环评估和伪模拟评估。

指标。为评估导航性能，我们遵循标准评估指标 (Anderson et al., 2018)，包括成功率 (SR)、预言成功率 (OS)、路径长度加权成功率 (SPL)、归一化动态时间规整 (nDTW) 和导航误差 (NE)。对于跟踪任务，采用跟踪率 (TR) (Puig et al., 2023)，即成功跟踪步数占总步数的比例。由于不同导航任务的成功判据各异，我们采用各基准的默认成功判据。自动驾驶评估中，开环规划 (Caesar et al., 2020a) 报告 L2 距离和碰撞率 (CR)。NAVSIM 闭环评估采用 PDM 分数 (PDMS) (Dauner et al., 2024b)，它是多个子指标的加权组合：无责任碰撞 (NC)、可行驶区域合规性 (DAC)、碰撞时间 (TTC)、舒适度 (Comf.) 和自行车行驶进度 (EP)。

模拟器部署。针对每项导航任务，我们遵循以往工作 (Krantz et al., 2020; Savva et al., 2019b; Das et al., 2018; Islam et al., 2019) 的默认设置。本方法以数量可变的相机在线采集的 RGB 视频（每执行一步后，各相机采集一帧）和文本指令为输入，输出下一段轨迹（公式 6）。对于 Habitat-Lab 连续环境中使用离散动作的某些基准（如 FORWARD、LEFT、RIGHT 和 STOP），我们将这些离散动作替换为基于轨迹的动作，更多细节见附录。

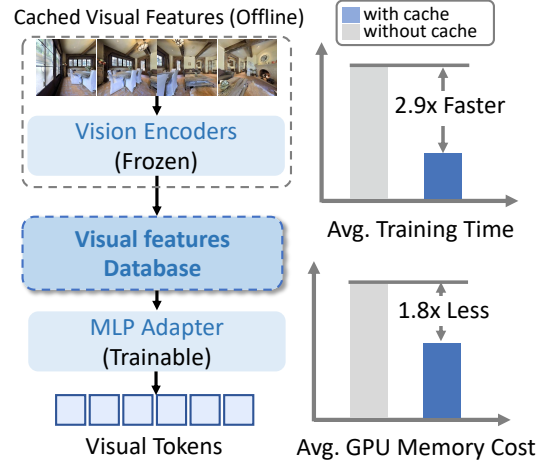


图 8: 离线视觉特征缓存。预先计算视频帧和导航历史，并保存为粗粒度视觉词元。

表 1: **VLN-CE 单视角与多视角设置下的比较**。S.RGB 和 M.RGB 分别表示单视角与多视角配置。符号 * 表示采用 (Hong et al., 2022) 航点预测器的方法。本方法无需针对特定相机配置微调, 也不需要深度或里程计等额外输入, 便在单视角和多视角设置下均达到 SOTA 性能。

方法	观测				R2R Val-Unseen				RxR Val-Unseen			
	S.RGB	M.RGB	深度	里程计	NE ↓	OS ↑	SR ↑	SPL ↑	NE ↓	SR ↑	SPL ↑	nDTW ↑
AG-CMTP (多视角) (Chen et al., 2021a)	✓	✓	✓		7.90	39.0	23.0	19.0	-	-	-	-
R2R-CMTP (多视角) (Chen et al., 2021a)	✓	✓	✓		7.90	38.0	26.0	22.0	-	-	-	-
HPN+DN (多视角) * (Krantz et al., 2021)	✓	✓	✓		6.31	40.0	36.0	34.0	-	-	-	-
CMA (多视角) * (Hong et al., 2022)	✓	✓	✓		6.20	52.0	41.0	36.0	8.76	26.5	22.1	47.0
VLN $\odot$ BERT (多视角) * (Hong et al., 2022)	✓	✓	✓		5.74	53.0	44.0	39.0	8.98	27.0	22.6	46.7
Sim2Sim* (Krantz & Lee, 2022)	✓	✓	✓		6.07	52.0	43.0	36.0	-	-	-	-
AO-Planner (Chen et al., 2024a)	✓	✓			5.55	59.0	47.0	33.0	7.06	43.3	30.5	50.1
GridMM* (Wang et al., 2023b)	✓	✓	✓		5.11	61.0	49.0	41.0	-	-	-	-
Ego ² -Map* (Hong et al., 2023)	✓	✓	✓		5.54	56.0	47.0	41.0	-	-	-	-
DreamWalker* (Wang et al., 2023a)	✓	✓	✓		5.53	59.0	49.0	44.0	-	-	-	-
Reborn* (An et al., 2022)	✓	✓	✓		5.40	57.0	50.0	46.0	5.98	48.6	42.0	63.3
ETPNav* (An et al., 2024)	✓	✓	✓		4.71	65.0	57.0	49.0	5.64	54.7	44.8	61.9
HNR* (Wang et al., 2024b)	✓	✓	✓		4.42	67.0	61.0	51.0	5.50	56.3	46.7	63.5
BEVBert* (An et al., 2023)	✓	✓	✓		4.57	67.0	59.0	50.0	-	-	-	-
HAMT+ScaleVLN* (Wang et al., 2023c)	✓	✓	✓		4.80	-	55.0	51.0	-	-	-	-
NavFoM (四视角)	✓				4.61	72.1	61.7	55.3	4.74	64.4	56.2	65.8
LAW (Raychaudhuri et al., 2021)	✓		✓	✓	6.83	44.0	35.0	31.0	10.90	8.0	8.0	38.0
CM2 (Georgakis et al., 2022)	✓		✓	✓	7.02	41.0	34.0	27.0	-	-	-	-
WS-MGMap (Chen et al., 2022)	✓		✓	✓	6.28	47.0	38.0	34.0	-	-	-	-
Seq2Seq (Krantz et al., 2020)	✓		✓		7.77	37.0	25.0	22.0	12.10	13.9	11.9	30.8
CMA (Krantz et al., 2020)	✓		✓		7.37	40.0	32.0	30.0	-	-	-	-
RGB-Seq2Seq (Krantz et al., 2020)	✓				10.10	8.0	0.0	0.0	-	-	-	-
RGB-CMA (Krantz et al., 2020)	✓				9.55	10.0	5.0	4.0	-	-	-	-
NaVid (Zhang et al., 2024a)	✓				5.72	49.2	41.9	36.5	5.72	45.7	38.2	-
Uni-NaVid (Zhang et al., 2025a)	✓				5.58	53.3	47.0	42.7	6.24	48.7	40.9	-
NaVILA (Cheng et al., 2025)	✓				5.22	62.5	54.0	49.0	6.77	49.3	44.0	58.8
StreamVLN-RGB-only (Wei et al., 2025)	✓				5.10	64.0	55.7	50.9	6.16	51.8	45.0	62.1
NavFoM (单视角)	✓				5.01	64.9	56.2	51.2	5.51	57.4	49.4	60.2

真实环境部署。真实世界部署采用配备 NVIDIA RTX 4090 GPU 的远程服务器运行 NavFoM。系统处理观测和文本指令, 并向本地机器人发送动作命令。实验使用四足机器人、人形机器人、无人机和轮式机器人执行轨迹。在 1600 个词元的预算下, NavFoM 生成包含 8 个航点的轨迹最多需要 0.5 秒。导航过程中, 机器人在执行动作的同时异步压缩最新观测并上传至模型。对于不同机器人, 采用其现成的本地规划器沿预测轨迹行驶。真实世界机器人设置的详细说明见附录。

4.2 基准结果

VLN: VLN-CE 上的性能 (Krantz et al., 2020; Ku et al., 2020)。首先在应用最广泛的视觉—语言指令基准 VLN-CE R2R 和 VLN-CE RxR 上评估本方法, 结果见表 1。我们报告单相机与四相机设置 (360° 观测) 下的性能。模型并未针对任何特定相机配置微调, 而是直接采用时间—视点指示 (TVI) 词元组织视觉词元 (图 6)。本方法在两项基准的不同相机设置下均达到最先进 (SOTA) 性能。在难度最高的单视角 VLN-CE RxR 条件下, 本方法将成功率 (SR) 从 51.8% 提升至 57.4%。在多相机设置中, 仅使用四个 RGB 相机便达到 64.4% 的 SR, 超过依赖 RGB-D 相机和里程计信息的既有 SOTA 方法 (SR 为 56.3%), 表明本方法确实有效。从单视角切换至多视角设置后, R2R-CE 和 RxR-CE 的性能分别显著提升 5.5% 和 7.0%, 说明多视角导航基础模型是值得探索的未来方向。此外, 尽管 RxR 的时间跨度更长、指令更复杂, 本方法在 RxR 上的表现却优于 R2R, 这与其他基线不同。我们认为, RxR 指令中的描述更加具体详尽, 提供了更强的上下文线索, 帮助模型更好地区分目标地标。

VLN: OpenUAV 上的性能 (Wang et al., 2024a)。我们报告本方法在高难度无人机场景中的性能 (表 2)。该场景要求无人机遵循自然语言指令, 执行平均长度为 200 米的长时域轨迹, 并在室外环境中到达指令描述的目标。由于缺乏可像 ObjectNav 数据采集那样收集专家轨迹的强基线, 本方法直接采用从 TravelUAV (Wang et al., 2024a) 训练集采集的轨迹 (模拟真实轨迹)。即便如此, 本方法在不依赖既有方法所用下视相机的情况下, 仍超越 TravelUAV 等面向无人机的既有基线并达到最先进性能 (未来将为相机配置引入更多自由度)。这进一步表明本方法有效, 也体现了从多样导航任务中学习的优势 (图 12)。

表 2: 采用 L1 级辅助的 OpenUAV 基准完整结果。Seen 表示已见划分，UO 和 UM 分别表示测试未见物体集与测试未见地图集。DA 表示采用基于回溯采样的数据聚合训练的模型。**粗体**和下划线 分别表示**最佳**与**次佳** 结果。

方法	测试集	全部				简单				困难			
		NE↓	SR↑	OSR↑	SPL↑	NE↓	SR↑	OSR↑	SPL↑	NE↓	SR↑	OSR↑	SPL↑
OpenUAV 已见集													
人类	Seen	14.15	94.51	94.51	77.84	11.68	95.44	95.44	76.19	17.16	93.37	93.37	79.85
随机动作	Seen	222.20	0.14	0.21	0.07	142.07	0.26	0.39	0.13	320.12	0.00	0.00	0.00
固定动作	Seen	188.61	2.27	8.16	1.40	121.36	3.48	11.48	2.14	270.69	0.79	4.09	0.49
CMA (Wang et al., 2024a)	Seen	135.73	8.37	18.72	7.90	84.89	11.48	24.52	10.68	197.77	4.57	11.65	4.51
TravelUAV (基线) (Wang et al., 2024a)	Seen	106.28	16.10	44.26	14.30	68.78	18.84	47.61	16.39	152.04	12.76	40.16	11.76
TravelUAV-DA	Seen	98.66	17.45	48.87	15.76	66.40	20.26	51.23	18.10	138.04	14.02	45.98	12.90
NavFoM (四视角)	Seen	93.05	29.17	49.24	25.03	58.98	32.91	53.16	27.87	143.83	23.58	43.40	20.80
OpenUAV 未见集													
Random Action	UO	260.14	0.16	0.16	0.16	174.10	0.48	0.48	0.48	302.96	0.00	0.00	0.00
Fixed Action	UO	212.84	3.66	9.54	2.16	151.66	6.70	13.88	3.72	243.29	2.14	7.38	1.38
CMA (Wang et al., 2024a)	UO	155.79	9.06	16.06	8.68	102.92	14.83	22.49	13.90	182.09	6.19	12.86	6.08
TravelUAV (Wang et al., 2024a)	UO	118.04	22.42	46.90	20.51	86.12	24.40	49.28	22.03	134.03	21.43	45.71	19.75
NavFoM (四视角)	UO	108.04	29.83	47.99	27.20	70.51	32.54	50.72	29.54	133.01	28.03	46.18	25.64
Random Action	UM	202.98	0.00	0.00	0.00	158.46	0.00	0.00	0.00	265.88	0.00	0.00	0.00
Fixed Action	UM	180.47	0.52	2.61	0.39	132.89	0.89	4.28	0.67	247.72	0.00	0.25	0.00
CMA (Wang et al., 2024a)	UM	141.68	2.30	10.02	2.16	102.29	3.57	14.26	3.33	197.35	0.50	4.03	0.50
TravelUAV (Wang et al., 2024a)	UM	138.80	4.18	20.77	3.84	102.94	4.63	22.82	4.24	189.46	3.53	17.88	3.28
NavFoM (四视角)	UM	125.10	6.30	18.95	5.68	102.41	6.77	20.07	6.04	170.58	5.36	15.71	4.97

表 3: 目标物体导航。在 HM3D-OVON Yokoyama et al. (2024b) 上的比较。* 表示零样本评估。报告本方法在第一人称视角和四视角设置下的性能。**粗体**和下划线 分别表示**最佳**与**次佳** 结果。

方法	VAL SEEN		VAL SEEN SYNONYMS		VAL UNSEEN	
	SR↑	SPL↑	SR↑	SPL↑	SR↑	SPL↑
BC (行为克隆)	11.1	4.5	9.9	3.8	5.4	1.9
DAgger (数据集聚合)	11.1	4.5	9.9	3.8	5.4	1.9
RL (强化学习)	18.1	9.4	15.0	7.4	10.2	4.7
BCRL	39.2	18.7	27.8	11.7	18.6	7.5
DAGRL	41.3	21.2	29.4	14.4	18.3	7.9
VLFM* (Yokoyama et al., 2024a)	35.2	18.6	32.4	17.3	35.2	19.6
DAGRL+OD (Yokoyama et al., 2024b)	38.5	21.1	39.0	21.4	37.1	19.8
Uni-NaVid* (Zhang et al., 2025a)	41.3	21.1	43.9	21.8	39.5	19.8
MTU3D (Zhu et al., 2025)	55.0	23.6	<u>45.0</u>	14.7	40.8	12.1
NavFoM * (单视角)	37.7	<u>25.5</u>	43.3	<u>29.9</u>	43.6	31.3
NavFoM * (四视角)	40.1	27.1	45.4	32.6	45.2	31.9

不过，所有方法在 Unseen-Map 划分上的表现均较差。该划分要求无人机平均穿越 300 米的复杂街区以到达未见目标，需要更高级的导航能力，例如高效探索大规模环境，而这又依赖质量更高的无人机数据。

搜索：OVON 上的性能 (Yokoyama et al., 2024b)。遵循先前工作 (Zhang et al., 2025a; Zhu et al., 2025)，我们在零样本设置下评估开放词汇基准上的搜索能力。表 3 给出了单相机和四相机配置的性能。在单相机零样本评估设置下，本方法在 VAL SEEN 和 VAL SEEN SYNONYMS 两个划分上均达到与最先进方法 (Zhu et al., 2025) 相当的性能。在难度更高的 VAL UNSEEN 划分上，本方法超越 SOTA，将成功率 (SR) 从 40.8% 提升至 43.6%。从单相机切换至四相机设置后，所有划分和指标均得到一致提升。值得注意的是，模型仅使用单相机搜索样本训练，这表明跨不同相机配置的联合微调能够增强模型对多种相机设置的泛化能力。

跟踪：EVT-Bench 上的性能 (Wang et al., 2025c)。我们在单视角和四视角相机设置下，于 EVT-Bench 的 Single Target 与 Distracted Target 两个划分上评估本方法 (表 4)。模型仅在单视角设置下训练，并以零样本方式在四视角设置下评估。结果表明，本方法在单视角设置下达到最先进 (SOTA) 性能，超过专门针对跟踪数据微调的既有基线 TrackVLA (Wang et al., 2025c)。将相机配置以零样本方式从单视角扩展至四视角后，性能进一步提升。不过，与 VLN 上的提升 (VLN-CE RxR 的 SR 提升 6.8% ↑) 相比，此处增益较小 (SR 提升 0.6% ↑)。原因可能在于 EVT-Bench 的关键假设：大多数目标均生成于机器人前方。未来将通过模拟和方法改进进一步研究这一问题，例如在机器人周围随机放置目标。

自动驾驶：NAVSIM (Dauner et al., 2024a) 与 nuScenes 上的性能 (Caesar et al., 2020b)。我们在六视角和八视角设置下评估本方法，且不针对特定配置微调。NAVSIM 和 nuScenes 的结果分别见表 5 与附录表 8。在不显式建模车道线、附近车辆等驾驶相关信息及其他上下

表 4: **EVT-Bench** 上的性能。†: 采用 GroundingDINO (Liu et al., 2023b) 作为开放词汇检测器。
‡: 采用 SoM (Yang et al., 2023)+GPT-4o (OpenAI, 2024) 作为视觉基础模型。**粗体**和下划线 分别表示**最佳**与**次佳** 结果。

方法	单目标		干扰目标	
	SR↑	TR↑	SR↑	TR↑
IBVS (基线) † (Gupta et al., 2016)	42.9	56.2	10.6	28.4
PoliFormer (基线) † (Zeng et al.)	4.67	15.5	2.62	13.2
EVT (Zhong et al., 2024)	24.4	39.1	3.23	11.2
EVT‡ (Zhong et al., 2024)	32.5	49.9	15.7	35.7
Uni-NaVid (Zhang et al., 2025a)	25.7	39.5	11.3	27.4
TrackVLA (Wang et al., 2025c)	85.1	78.6	57.6	63.2
NavFoM (单视角)	85.0	<u>80.5</u>	<u>61.4</u>	68.2
NavFoM (四视角)	88.4	80.7	62.0	<u>67.9</u>

表 5: 采用闭环指标在面向规划的 **NAVSIM navtest** 划分上进行比较。 $\mathcal{V}_{8192}$ 表示 8192 个锚点。**粗体**和下划线 分别表示**最佳**与**次佳** 结果。

方法	观测与结构			指标					
	相机	激光雷达	基于 VLM	NC ↑	DAC ↑	TTC ↑	Comf. ↑	EP ↑	PDMS ↑
人类驾驶	-	-	-	100	100	100	99.9	87.5	94.8
恒定速度	-	-	-	69.9	58.8	49.3	100	49.3	21.6
自车状态 MLP	-	-	-	93.0	77.3	83.6	100	62.8	65.6
LTF (Chitta et al., 2022)	✓	✓	-	97.4	92.8	92.4	100	79.0	83.8
Transfuser (Chitta et al., 2022)	✓	✓	-	97.7	92.8	92.8	100	79.2	84.0
VADv2- $\mathcal{V}_{8192}$ (Chen et al., 2024c)	✓	✓	-	97.2	89.1	91.6	100	76.0	80.9
Hydra-MDP- $\mathcal{V}_{8192}$ (Li et al., 2024b)	✓	✓	-	97.9	91.7	92.9	100	77.6	83.0
DiffusionDrive (Liao et al., 2024a)	✓	✓	-	98.2	96.2	<u>94.7</u>	100	82.2	88.1
DRAMA (Yuan et al., 2024)	✓	✓	✓	<u>98.0</u>	<u>93.1</u>	94.8	100	<u>80.1</u>	<u>85.5</u>
UniAD (Hu et al., 2023)	✓	-	-	97.8	91.9	92.9	100	78.8	83.4
PARA-Drive (Weng et al., 2024)	✓	-	-	<u>97.9</u>	92.4	<u>93.0</u>	99.8	79.3	84.0
LAW (Li et al., 2024a)	✓	-	-	96.4	95.4	88.7	99.9	81.7	84.6
DrivingGPT (Chen et al., 2024d)	✓	-	✓	98.9	90.7	94.9	95.6	79.7	82.4
NavFoM (八视角)	✓	-	✓	97.7	<u>93.5</u>	92.3	100	79.6	<u>84.3</u>

文要素的情况下，本方法在两项基准上均取得与 SOTA 方法相当的性能。与其他基线类似，将场景描述作为提示或可进一步改进本方法。我们也希望在 (Dosovitskiy et al., 2017) 等闭环自动驾驶模拟器中评估该模型。

基准结果可视化。图 9 展示了 NavFoM 在上述基准上的可视化结果，其中绘制了预测轨迹、相机视图及对应指令。

4.3 真实世界结果

110 个可复现测试案例上的真实世界性能。为评估本方法的真实世界性能，我们设计了一系列覆盖不同能力的导航测试案例，包括 50 个 VLN 样本、30 个搜索样本和 30 个跟踪样本。具体而言，我们搭建了一个 $5\text{m} \times 5\text{m}$ 的空间，并记录每个测试案例中机器人、障碍物和目标的位置。图 10 给出了 NavFoM 在这些导航能力的复杂场景中的定性与定量结果。结果表明，NavFoM 具有较强的真实世界性能：它能够正确理解周围环境，并规划合适轨迹以完成任务。与强基线 Uni-NaVid (Zhang et al., 2025a) 相比，本方法在两项任务上均显著提升，体现了更优的真实环境性能。

高难度跨任务、跨具身真实世界实验的可视化结果。我们还在使用不同具身（四足机器人、人形机器人、无人机和轮式机器人）的更具挑战性场景中开展了大量实验。图 11 表明，本方法能够应对复杂真实环境并完成长时域指令。配套视频提供了更直观的演示。

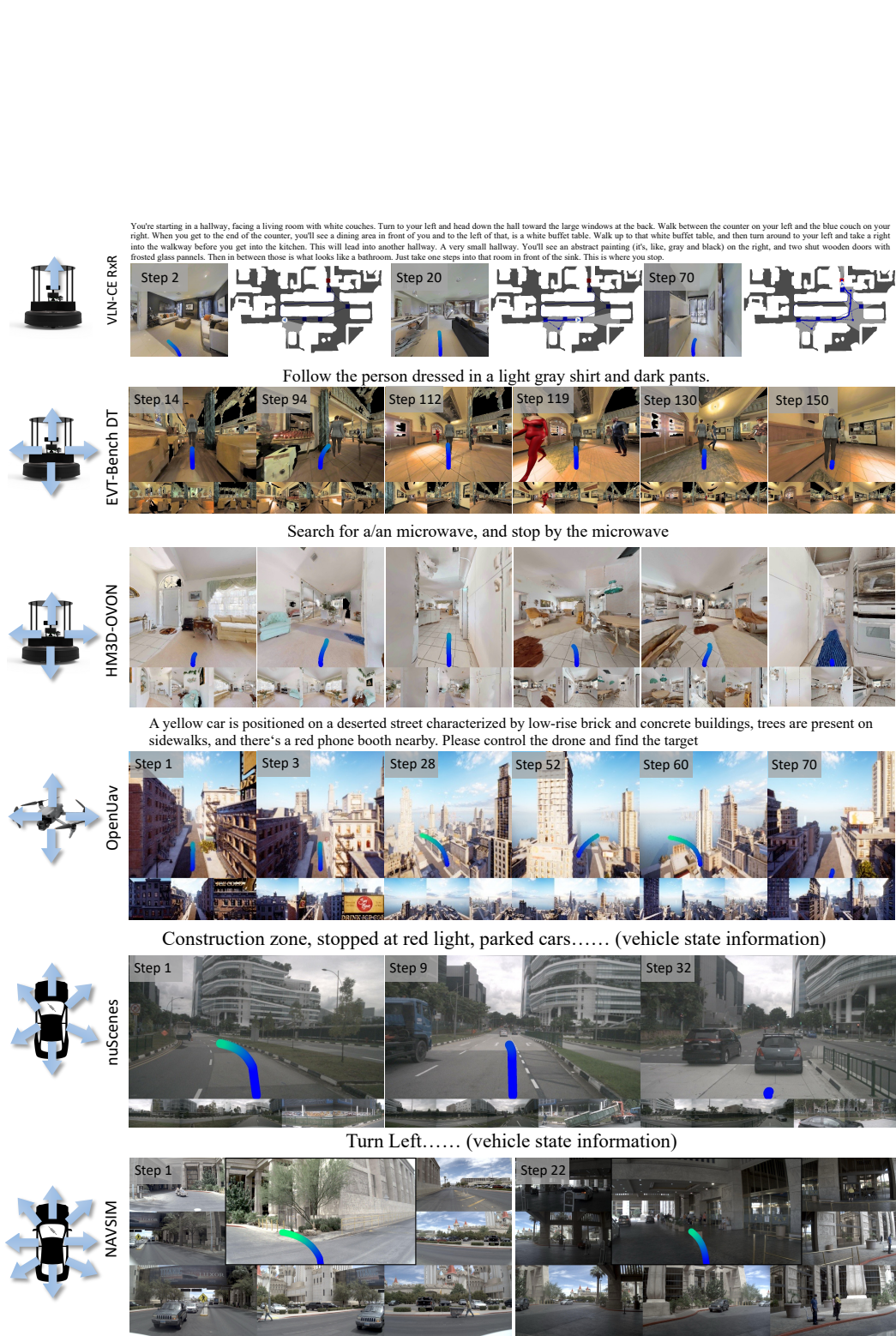


图 9: 基准性能可视化。展示 NavFoM 在 VLN-CE RxR (单视角)、EVT-Bench Distracted Targets (四视角)、OpenUAV (四视角)、NeuScenes (六视角) 和 OpenScenes (八视角) 上的可视化结果。

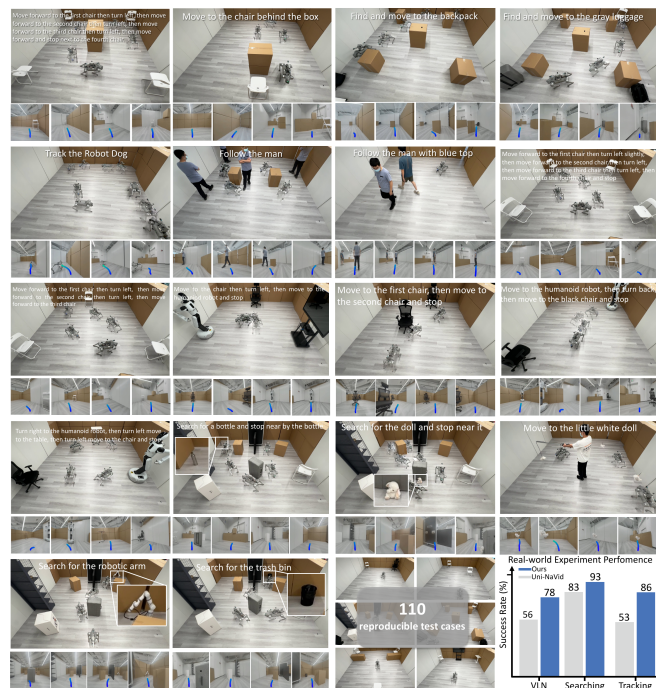


图 10: 真实世界实验。报告 NavFoM 在不同导航能力复杂场景中的定性与定量结果。



图 11: 跨任务、跨具身设置下真实世界实验的可视化结果。

4.4 消融实验

多导航任务训练的协同作用。我们比较单任务训练与加入其他导航任务额外数据（比例为 50% 和 100%）的联合微调，以研究多导航任务训练的协同效应。此处，“VLN”指 VLN-CE RxR 四视角设置，“Searching”指 OVON 四视角设置（以零样本方式评估），“Tracking”指 EVT-Bench 四视角设置，“Driving”指 NavSIM 八视角设置。结果显示，使用多种导航任务数据联合微调后，各任务性能均随其他任务数据比例从 50% 增至 100% 而持续提升。搜索（从 10.3% 提升至 45.2%）和跟踪（从 12.6% 提升至 62.0%）的增益尤为显著。我们认为这是因为其训练条件（以单视角和闭集目标类别为主）与多视角、开放词汇的评估设置存在差异。通过学习其他任务中包含多样目标的多视角数据，搜索和跟踪任务均增强了多视角与开放集导航能力，因而性能大幅提升。结果表明，跨多项导航任务训练有助于缓解对任务特定导航模式的过拟合。

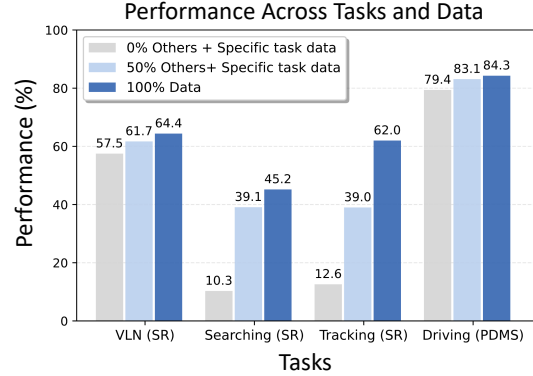


图 12: 多导航任务训练的消融实验。报告不同训练数据组合的性能，包括仅使用特定任务数据、特定任务数据加 50% 其他数据，以及特定任务数据加 100% 其他数据。† 搜索任务采用零样本测试。

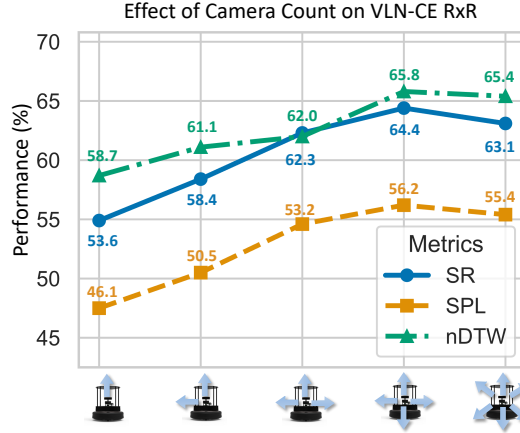


图 13: VLN-CE RxR 中相机数量的消融实验。在相同词元预算 ($B = 2048$) 下，报告五种不同相机配置的性能（从左至右依次为一、二、三、四和六相机设置）。

不同相机数量下的性能。我们在 VLN-CE RxR 上评估导航任务中增加相机的有效性；该基准提供了较为全面的视觉—语言导航挑战。表 13 比较了机器人周围安装一、二、三、四和六个相机的配置，以获得更宽视场。相机数量从一个增至四个时，性能持续提升，验证了更充分的环境观测有助于提高导航性能。然而，扩展至六个相机后性能略有下降。原因可能在于，相比四个相机，六个相机并未显著扩大观测覆盖范围，而增加的视图词元会压缩用于编码历史帧的容量（公式 5），削弱导航历史与指令之间的对齐。采用自适应多视角词元编码策略或可缓解该问题；为保持本文的一致性，我们将其留待未来研究。

BATS 与 TVI 词元的有效性。我们通过消融实验评估历史词元组织策略和视觉—时间历史建模等关键设计的有效性。实验采用 VLN-CE RxR 四相机设置，结果见表 6。在 1024 和 2048 两种词元预算下测试不同词元策略，预算感知时间采样（BATS）在两种设置下均优于其他策略。特别是在直接衡量导航轨迹与真实轨迹对齐程度的 nDTW 指标上，BATS 的性能仅小幅下降（1.4% ↓），优于均匀采样（Cheng et al., 2025）（6.0% ↓）和线性概率采样（5.2% ↓）。我们还将时间—视点指示（TVI）词元与其他常见替代方案比较，包括独立学习的特殊词元，以及公式 3 中不含 $\mathcal{P}_{\text{angle/time}}$ 的手工词元。结果显示，TVI 词元的性能显著更优。如图 4 所示，这一提升得益于学习充分的时间和视点信息。相比之下，常见的历史—视点位置嵌入方法（Chen et al., 2021b）出现明显性能下降。我们认为，这是由于视觉

类型	RxR Val-Unseen			
	NE ↓	SR ↑	SPL ↑	nDTW ↑
$B = 1024$, 均匀采样 *	5.33	59.7	49.6	57.9
$B = 1024$, 线性概率采样	5.28	61.2	50.9	58.9
$B = 1024$, 预算感知时间采样 (BATS)	4.98	62.5	53.9	64.1
$B = 2048$, 词元合并 (Zhang et al., 2025a)	5.01	63.2	54.9	64.4
$B = 2048$, 均匀采样 *	4.90	62.4	54.0	63.9
$B = 2048$, 线性概率采样	4.89	63.0	54.6	64.8
$B = 2048$, 预算感知时间采样 (BATS)	4.74	64.4	56.2	65.8
视点—历史位置嵌入 †	6.27	52.3	46.3	58.7
独立学习的特殊词元 ‡	5.52	59.1	52.0	59.6
手工词元 (公式 3, 不含 $\mathcal{P}_{\text{angle/time}}$)	6.06	53.6	46.1	58.0
时间—视点指示 (TVI) 词元 (公式 3)	4.74	64.4	56.2	65.8

表 6: 历史词元组织策略与指示词元的消融实验。对于线性概率采样和均匀采样, 所有方法均从包含 2048 个词元的基础集合中采样。* 均匀采样指 (Cheng et al., 2025) 采用的技术。† 采用 Chen et al. (2021b) 中传统 VLN 方法广泛使用的常见位置编码技术。‡ 为数据集中的每个视点引入预定义的可学习特殊词元。

词元中额外的嵌入分量增加了大语言模型 (LLM) 表征学习的复杂度。结果表明, 如何有效融入视点与时间信息仍需新的技术, 而 TVI 词元提供了一个具有潜力的起点。

5 讨论与结论

本文提出 NavFoM, 旨在拓展导航能力边界, 并探索从跨具身、跨任务导航数据中习得的智能。本文引入时间—视点指示 (Temporal-Viewpoint Indicator, TVI) 词元, 增强大语言模型 (Large Language Model, LLM) 对不同相机配置与导航时间跨度的理解, 同时支持导航数据和问答数据的联合训练。本文进一步采用预算感知时间采样 (Budget-Aware Temporal Sampling, BATS) 策略, 在多种相机配置与任务时间跨度下统一采样词元, 兼顾导航性能与效率。公共基准和真实环境中的大量实验表明, NavFoM 性能优异, 采用更先进技术或更高质量数据后仍有较大提升潜力。

本文认为, NavFoM 只是迈向导航基础模型的起点。希望本工作推动研究社区更加关注以智能为核心的导航, 并催生新一代技术、数据集与基准。

参考文献

- Dong An, Zun Wang, Yangguang Li, Yi Wang, Yicong Hong, Yan Huang, Liang Wang, and Jing Shao. 1st place solutions for rxr-habitat vision-and-language navigation competition (cvpr 2022). *arXiv preprint arXiv:2206.11610*, 2022. 10
- Dong An, Yuankai Qi, Yangguang Li, Yan Huang, Liang Wang, Tieniu Tan, and Jing Shao. Bevbart: Multimodal map pre-training for language-guided navigation. In *ICCV*, 2023. 10
- Dong An, Hanqing Wang, Wenguan Wang, Zun Wang, Yan Huang, Keji He, and Liang Wang. Etpnav: Evolving topological planning for vision-language navigation in continuous environments. *IEEE TPAMI*, 2024. 10
- Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. 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. 9
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 8

- Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15791–15801, 2025. 2
- Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025. 3
- Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0 : A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024. 3
- Richard P Brent. *Algorithms for minimization without derivatives*. Courier Corporation, 2013. 6
- Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xuan Hu, Xu Huang, et al. Agibot world colosseum: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025a. 3
- Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Univla: Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2505.06111*, 2025b. 3
- Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020a. 8, 9
- Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020b. 8, 11
- Jiaqi Chen, Bingqian Lin, Xinmin Liu, Xiaodan Liang, and Kwan-Yee K Wong. Affordances-oriented planning using foundation models for continuous vision-language navigation. *arXiv preprint*, 2024a. 10
- Jiaqi Chen, Bingqian Lin, Ran Xu, Zhenhua Chai, Xiaodan Liang, and Kwan-Yee K Wong. Mapgpt: Map-guided prompting for unified vision-and-language navigation. *arXiv preprint arXiv:2401.07314*, 2024b. 3
- Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 5
- Kevin Chen, Junshen K Chen, Jo Chuang, Marynel Vázquez, and Silvio Savarese. Topological planning with transformers for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11276–11286, 2021a. 10
- Peihao Chen, Dongyu Ji, Kunyang Lin, Runhao Zeng, Thomas H Li, Mingkui Tan, and Chuang Gan. Weakly-supervised multi-granularity map learning for vision-and-language navigation. *arXiv preprint arXiv:2210.07506*, 2022. 10
- Shaoyu Chen, Bo Jiang, Hao Gao, Bencheng Liao, Qing Xu, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Vadv2: End-to-end vectorized autonomous driving via probabilistic planning. *arXiv preprint arXiv:2402.13243*, 2024c. 12
- Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. History aware multimodal transformer for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 34:5834–5847, 2021b. 16, 17

- Yuntao Chen, Yuqi Wang, and Zhaoxiang Zhang. Drivingsgpt: Unifying driving world modeling and planning with multi-modal autoregressive transformers. *arXiv preprint arXiv:2412.18607*, 2024d. 8, 12
- An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Xueyan Zou, Jan Kautz, Erdem Biyik, Hongxu Yin, Sifei Liu, and Xiaolong Wang. Navila: Legged robot vision-language-action model for navigation. In *RSS*, 2025. 2, 3, 4, 5, 8, 10, 16, 17
- Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *IEEE transactions on pattern analysis and machine intelligence*, 45(11):12878–12895, 2022. 12
- OpenScene Contributors. Openscene: The largest up-to-date 3d occupancy prediction benchmark in autonomous driving. <https://github.com/OpenDriveLab/OpenScene>, 2023. 2, 8
- Nimrod Curtis, Osher Azulay, and Avishai Sintov. Embodiment-agnostic navigation policy trained with visual demonstrations. *arXiv preprint arXiv:2412.20226*, 2024. 3
- Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–10, 2018. 9
- Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a. 11
- Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, et al. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. *Advances in Neural Information Processing Systems*, 37:28706–28719, 2024b. 9
- Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, et al. Procthor: Large-scale embodied ai using procedural generation. *arXiv preprint arXiv:2206.06994*, 2022. 2
- Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pp. 1–16. PMLR, 2017. 12
- Ainaz Eftekhari, Rose Hendrix, Luca Weihs, Jiafei Duan, Ege Caglar, Jordi Salvador, Alvaro Herrasti, Winson Han, Eli VanderBil, Aniruddha Kembhavi, et al. The one ring: a robotic indoor navigation generalist. *arXiv preprint arXiv:2412.14401*, 2024. 2, 3
- Anthony Francis, Claudia Pérez-d’Arpino, Chengshu Li, Fei Xia, Alexandre Alahi, Rachid Alami, Aniket Bera, Abhijat Biswas, Joydeep Biswas, Rohan Chandra, et al. Principles and guidelines for evaluating social robot navigation algorithms. *arXiv preprint arXiv:2306.16740*, 2023. 7
- Haoyu Fu, Diankun Zhang, Zongchuang Zhao, Jianfeng Cui, Dingkan Liang, Chong Zhang, Dingyuan Zhang, Hongwei Xie, Bing Wang, and Xiang Bai. Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation. *arXiv preprint arXiv:2503.19755*, 2025a. 28
- Haoyu Fu, Diankun Zhang, Zongchuang Zhao, Jianfeng Cui, Dingkan Liang, Chong Zhang, Dingyuan Zhang, Hongwei Xie, Bing Wang, and Xiang Bai. Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation. *arXiv preprint arXiv:2503.19755*, 2025b. 5

- Chen Gao, Liankai Jin, Xingyu Peng, Jiazhao Zhang, Yue Deng, Annan Li, He Wang, and Si Liu. Octonav: Towards generalist embodied navigation. *arXiv preprint arXiv:2506.09839*, 2025. [3](#), [4](#)
- Georgios Georgakis, Karl Schmeckpeper, Karan Wanchoo, Soham Dan, Eleni Miltsakaki, Dan Roth, and Kostas Daniilidis. 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. [10](#)
- Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025. [2](#), [5](#)
- Meenakshi Gupta, Swagat Kumar, Laxmidhar Behera, and Venkatesh K Subramanian. A novel vision-based tracking algorithm for a human-following mobile robot. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 47(7):1415–1427, 2016. [12](#)
- Wencheng Han, Dongqian Guo, Cheng-Zhong Xu, and Jianbing Shen. Dme-driver: Integrating human decision logic and 3d scene perception in autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3347–3355, 2025. [28](#)
- Noriaki Hirose, Dhruv Shah, Ajay Sridhar, and Sergey Levine. Exaug: Robot-conditioned navigation policies via geometric experience augmentation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4077–4084. IEEE, 2023. [2](#), [3](#)
- Yicong Hong, Zun Wang, Qi Wu, and Stephen 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*, pp. 15439–15449, 2022. [10](#)
- Yicong Hong, Yang Zhou, Ruiyi Zhang, Franck Deroncourt, Trung Bui, Stephen Gould, and Hao Tan. Learning navigational visual representations with semantic map supervision. In *ICCV*, 2023. [10](#)
- Shengchao Hu, Li Chen, Penghao Wu, Hongyang Li, Junchi Yan, and Dacheng Tao. St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In *European Conference on Computer Vision*, pp. 533–549. Springer, 2022. [28](#)
- Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17853–17862, 2023. [8](#), [12](#), [28](#)
- Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, et al. Emma: End-to-end multimodal model for autonomous driving. *arXiv preprint arXiv:2410.23262*, 2024. [28](#)
- Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025. [3](#)
- Md Jahidul Islam, Jungseok Hong, and Junaed Sattar. Person-following by autonomous robots: A categorical overview. *The International Journal of Robotics Research*, 38(14): 1581–1618, 2019. [7](#), [9](#)
- Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8350, 2023. [28](#)
- Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, et al. Openvla: An open-source vision-language-action model. In *8th Annual Conference on Robot Learning*. [3](#), [4](#)

- Jacob Krantz and Stefan Lee. Sim-2-sim transfer for vision-and-language navigation in continuous environments. In *European Conference on Computer Vision*, pp. 588–603. Springer, 2022. 10
- Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *ECCV*, 2020. 7, 9, 10
- Jacob Krantz, Aaron Gokaslan, Dhruv Batra, Stefan Lee, and Oleksandr Maksymets. Waypoint models for instruction-guided navigation in continuous environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15162–15171, 2021. 10
- Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4392–4412, 2020. 2, 6, 7, 9, 10
- Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. *arXiv preprint arXiv:2311.17043*, 2023. 2, 3, 4, 7, 8
- Yingyan Li, Lue Fan, Jiawei He, Yuqi Wang, Yuntao Chen, Zhaoxiang Zhang, and Tieniu Tan. Enhancing end-to-end autonomous driving with latent world model. *arXiv preprint arXiv:2406.08481*, 2024a. 8, 12
- Zhen Li, Chuanhao Li, Xiaofeng Mao, Shaoheng Lin, Ming Li, Shitian Zhao, Zhaopan Xu, Xinyue Li, Yukang Feng, Jianwen Sun, et al. Sekai: A video dataset towards world exploration. *arXiv preprint arXiv:2506.15675*, 2025a. 2, 8
- Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. Megasam: Accurate, fast and robust structure and motion from casual dynamic videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10486–10496, 2025b. 8
- Zhenxin Li, Kailin Li, Shihao Wang, Shiyi Lan, Zhiding Yu, Yishen Ji, Zhiqi Li, Ziyue Zhu, Jan Kautz, Zuxuan Wu, et al. Hydra-mdp: End-to-end multimodal planning with multi-target hydra-distillation. *arXiv preprint arXiv:2406.06978*, 2024b. 12
- Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, et al. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. *arXiv preprint arXiv:2411.15139*, 2024a. 12
- Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, et al. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. *arXiv preprint arXiv:2411.15139*, 2024b. 8, 28
- Bingqian Lin, Yunshuang Nie, Ziming Wei, Jiaqi Chen, Shikui Ma, Jianhua Han, Hang Xu, Xiaojun Chang, and Xiaodan Liang. Navcot: Boosting llm-based vision-and-language navigation via learning disentangled reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 3
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023a. 2, 3, 4, 5, 8
- Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023b. 12
- Yuxing Long, Xiaoqi Li, Wenzhe Cai, and Hao Dong. Discuss before moving: Visual language navigation via multi-expert discussions. *arXiv preprint arXiv:2309.11382*, 2023. 3

- Yuxing Long, Wenzhe Cai, Hongcheng Wang, Guanqi Zhan, and Hao Dong. Instructnav: Zero-shot system for generic instruction navigation in unexplored environment. *arXiv preprint arXiv:2406.04882*, 2024. 2, 3, 4
- Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 4, 5
- OpenAI. Introducing 4o image generation. <https://openai.com/index/introducing-4o-image>, 2024. Accessed: 2025-04-29. 12
- Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 4, 8
- Abby O’ Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903. IEEE, 2024. 3
- Bowen Pan, Rameswar Panda, SouYoung Jin, Rogerio Feris, Aude Oliva, Phillip Isola, and Yoon Kim. Langnav: Language as a perceptual representation for navigation. *arXiv preprint arXiv:2310.07889*, 2023. 3
- Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander William Clegg, Michal Hlavac, So Yeon Min, et al. Habitat 3.0: A co-habitat for humans, avatars and robots. *arXiv preprint arXiv:2310.13724*, 2023. 9
- Pranav Putta, Gunjan Aggarwal, Roozbeh Mottaghi, Dhruv Batra, Naoki Yokoyama, Joanne Truong, and Arjun Majumdar. Embodiment randomization for cross embodiment navigation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5527–5534. IEEE, 2024. 3
- Yanyuan Qiao, Yuankai Qi, Zheng Yu, Jing Liu, and Qi Wu. March in chat: Interactive prompting for remote embodied referring expression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15758–15767, 2023. 3
- Yanyuan Qiao, Wenqi Lyu, Hui Wang, Zixu Wang, Zerui Li, Yuan Zhang, Mingkui Tan, and Qi Wu. Open-nav: Exploring zero-shot vision-and-language navigation in continuous environment with open-source llms. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6710–6717. IEEE, 2025. 3
- Delin Qu, Haoming Song, Qizhi Chen, Zhaoqing Chen, Xianqiang Gao, Xinyi Ye, Qi Lv, Modi Shi, Guanghui Ren, Cheng Ruan, et al. Embodiedonevision: Interleaved vision-text-action pretraining for general robot control. *arXiv preprint arXiv:2508.21112*, 2025. 3
- Sonia Raychaudhuri, Saim Wani, Shivansh Patel, Unnat Jain, and Angel Chang. Language-aligned waypoint (law) supervision for vision-and-language navigation in continuous environments. 2021. 10
- Shouwei Ruan, Liyuan Wang, Caixin Kang, Qihui Zhu, Songming Liu, Xingxing Wei, and Hang Su. From reactive to cognitive: brain-inspired spatial intelligence for embodied agents. *arXiv preprint arXiv:2508.17198*, 2025. 3
- Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. *ICCV*, 2019a. 2, 7, 8, 27

- Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9339–9347, 2019b. 9
- Dhruv Shah, Blazej Osinski, Brian Ichter, and Sergey Levine. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. *arXiv preprint arXiv:2207.04429*, 2022. 3
- Dhruv Shah, Ajay Sridhar, Arjun Bhorkar, Noriaki Hirose, and Sergey Levine. Gnm: A general navigation model to drive any robot. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7226–7233. IEEE, 2023a. 2, 3, 7
- Dhruv Shah, Ajay Sridhar, Nitish Dashora, Kyle Stachowicz, Kevin Black, Noriaki Hirose, and Sergey Levine. ViNT: A foundation model for visual navigation. In *7th Annual Conference on Robot Learning*, 2023b. URL <https://arxiv.org/abs/2306.14846>. 3
- Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024. 2, 3, 7, 8
- Xinshuai Song, Weixing Chen, Yang Liu, Weikai Chen, Guanbin Li, and Liang Lin. Towards long-horizon vision-language navigation: Platform, benchmark and method. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12078–12088, 2025. 3
- Wenchao Sun, Xuewu Lin, Yining Shi, Chuang Zhang, Haoran Wu, and Sifa Zheng. Sparsedrive: End-to-end autonomous driving via sparse scene representation. *arXiv preprint arXiv:2405.19620*, 2024. 28
- Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024. 3
- Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024. 28
- Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024. 4
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5
- Haitong Wang, Aaron Hao Tan, Angus Fung, and Goldie Nejat. X-nav: Learning end-to-end cross-embodiment navigation for mobile robots. *arXiv preprint arXiv:2507.14731*, 2025a. 3
- Hanqing Wang, Wei Liang, Luc V Gool, and Wenguan Wang. Towards versatile embodied navigation. *Advances in neural information processing systems*, 35:36858–36874, 2022. 3
- Hanqing Wang, Wei Liang, Luc Van Gool, and Wenguan Wang. Dreamwalker: Mental planning for continuous vision-language navigation. In *ICCV*, 2023a. 10
- Liuyi Wang, Xinyuan Xia, Hui Zhao, Hanqing Wang, Tai Wang, Yilun Chen, Chengju Liu, Qijun Chen, and Jiangmiao Pang. Rethinking the embodied gap in vision-and-language navigation: A holistic study of physical and visual disparities. *arXiv preprint arXiv:2507.13019*, 2025b. 3, 7

- Shaoan Wang, Jiazhaio Zhang, Minghan Li, Jiahang Liu, Anqi Li, Kui Wu, Fangwei Zhong, Junzhi Yu, Zhizheng Zhang, and He Wang. Trackvla: Embodied visual tracking in the wild. *arXiv preprint arXiv:2505.23189*, 2025c. [2](#), [3](#), [4](#), [7](#), [8](#), [9](#), [11](#), [12](#)
- Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M Alvarez. Omnidrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning. *arXiv preprint arXiv:2504.04348*, 2025d. [28](#)
- Xiangyu Wang, Donglin Yang, Ziqin Wang, Hohin Kwan, Jinyu Chen, Wenjun Wu, Hongsheng Li, Yue Liao, and Si Liu. Towards realistic uav vision-language navigation: Platform, benchmark, and methodology. *arXiv preprint arXiv:2410.07087*, 2024a. [2](#), [7](#), [9](#), [10](#), [11](#)
- Xin Eric Wang, Vihan Jain, Eugene Ie, William Yang Wang, Zornitsa Kozareva, and Sujith Ravi. Environment-agnostic multitask learning for natural language grounded navigation. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pp. 413–430. Springer, 2020. [3](#)
- Zihan Wang, Xiangyang Li, Jiahao Yang, Yeqi Liu, and Shuqiang Jiang. Gridmm: Grid memory map for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15625–15636, 2023b. [10](#)
- Zihan Wang, Xiangyang Li, Jiahao Yang, Yeqi Liu, Junjie Hu, Ming Jiang, and Shuqiang Jiang. Lookahead exploration with neural radiance representation for continuous vision-language navigation. In *CVPR*, 2024b. [10](#)
- Zun Wang, Jialu Li, Yicong Hong, Yi Wang, Qi Wu, Mohit Bansal, Stephen Gould, Hao Tan, and Yu Qiao. Scaling data generation in vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12009–12020, 2023c. [10](#)
- Meng Wei, Chenyang Wan, Xiqian Yu, Tai Wang, Yuqiang Yang, Xiaohan Mao, Chenming Zhu, Wenzhe Cai, Hanqing Wang, Yilun Chen, et al. Streamvln: Streaming vision-and-language navigation via slowfast context modeling. *arXiv preprint arXiv:2507.05240*, 2025. [3](#), [8](#), [10](#)
- Xinshuo Weng, Boris Ivanovic, Yan Wang, Yue Wang, and Marco Pavone. Para-drive: Parallelized architecture for real-time autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15449–15458, 2024. [12](#)
- Shen Yan, Tao Zhu, Zirui Wang, Yuan Cao, Mi Zhang, Soham Ghosh, Yonghui Wu, and Jiahui Yu. Videococa: Video-text modeling with zero-shot transfer from contrastive captioners. *arXiv preprint arXiv:2212.04979*, 2022. [9](#)
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yaqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024a. [2](#), [8](#)
- Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023. [12](#)
- Jonathan Yang, Catherine Glossop, Arjun Bhorkar, Dhruv Shah, Quan Vuong, Chelsea Finn, Dorsa Sadigh, and Sergey Levine. Pushing the limits of cross-embodiment learning for manipulation and navigation. *arXiv preprint arXiv:2402.19432*, 2024b. [3](#)

- Hang Yin, Xiuwei Xu, Linqing Zhao, Ziwei Wang, Jie Zhou, and Jiwen Lu. Unigoal: Towards universal zero-shot goal-oriented navigation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19057–19066, 2025. [2](#), [3](#)
- Naoki Yokoyama, Sehoon Ha, Dhruv Batra, Jiuguang Wang, and Bernadette Bucher. Vlfm: Vision-language frontier maps for zero-shot semantic navigation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 42–48. IEEE, 2024a. [11](#)
- Naoki Yokoyama, Ram Ramrakhya, Abhishek Das, Dhruv Batra, and Sehoon Ha. Hm3d-ovon: A dataset and benchmark for open-vocabulary object goal navigation. *arXiv preprint arXiv:2409.14296*, 2024b. [3](#), [9](#), [11](#)
- Bangguo Yu, Hamidreza Kasaei, and Ming Cao. L3mvn: Leveraging large language models for visual target navigation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3554–3560. IEEE, 2023. [7](#)
- Chengran Yuan, Zhanqi Zhang, Jiawei Sun, Shuo Sun, Zefan Huang, Christina Dao Wen Lee, Dongen Li, Yuhang Han, Anthony Wong, Keng Peng Tee, et al. Drama: An efficient end-to-end motion planner for autonomous driving with mamba. *arXiv preprint arXiv:2408.03601*, 2024. [12](#)
- Kuo-Hao Zeng, Zichen Zhang, Kiana Ehsani, Rose Hendrix, Jordi Salvador, Alvaro Herrasti, Ross Girshick, Aniruddha Kembhavi, and Luca Weihs. Poliformer: Scaling on-policy rl with transformers results in masterful navigators. In *8th Annual Conference on Robot Learning*. [2](#), [12](#)
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023. [4](#), [8](#)
- Jiazhao Zhang, Kunyu Wang, Rongtao Xu, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang, and He Wang. Navid: Video-based vlm plans the next step for vision-and-language navigation. *Robotics: Science and Systems*, 2024a. [2](#), [3](#), [4](#), [7](#), [10](#)
- Jiazhao Zhang, Kunyu Wang, Shaoan Wang, Minghan Li, Haoran Liu, Songlin Wei, Zhongyuan Wang, Zhizheng Zhang, and He Wang. Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks. *Robotics: Science and Systems*, 2025a. [2](#), [3](#), [4](#), [5](#), [7](#), [9](#), [10](#), [11](#), [12](#), [17](#)
- Lingfeng Zhang, Xiaoshuai Hao, Yingbo Tang, Haoxiang Fu, Xinyu Zheng, Pengwei Wang, Zhongyuan Wang, Wenbo Ding, and Shanghang Zhang. nava3: Understanding any instruction, navigating anywhere, finding anything. *arXiv preprint arXiv:2508.04598*, 2025b. [3](#)
- Lingfeng Zhang, Xiaoshuai Hao, Qinwen Xu, Qiang Zhang, Xinyao Zhang, Pengwei Wang, Jing Zhang, Zhongyuan Wang, Shanghang Zhang, and Renjing Xu. MapNav: A novel memory representation via annotated semantic maps for VLM-based vision-and-language navigation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 13032–13056, July 2025c. [3](#)
- Yue Zhang, Ziqiao Ma, Jialu Li, Yanyuan Qiao, Zun Wang, Joyce Chai, Qi Wu, Mohit Bansal, and Parisa Kordjamshidi. Vision-and-language navigation today and tomorrow: A survey in the era of foundation models. *ArXiv*, abs/2407.07035, 2024b. [2](#)
- Duo Zheng, Shijia Huang, Lin Zhao, Yiwu Zhong, and Liwei Wang. Towards learning a generalist model for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13624–13634, 2024. [3](#), [5](#), [7](#)
- Fangwei Zhong, Kui Wu, Hai Ci, Churan Wang, and Hao Chen. Empowering embodied visual tracking with visual foundation models and offline rl. In *European Conference on Computer Vision*, pp. 139–155. Springer, 2024. [12](#)
- Gengze Zhou, Yicong Hong, and Qi Wu. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. *arXiv preprint arXiv:2305.16986*, 2023. [3](#)

- Gengze Zhou, Yicong Hong, Zun Wang, Chongyang Zhao, Mohit Bansal, and Qi Wu. Same: Learning generic language-guided visual navigation with state-adaptive mixture of experts. *arXiv preprint arXiv:2412.05552*, 2024. [3](#), [7](#)
- Gengze Zhou, Yicong Hong, Zun Wang, Xin Eric Wang, and Qi Wu. Navgpt-2: Unleashing navigational reasoning capability for large vision-language models. In *European Conference on Computer Vision*, pp. 260–278. Springer, 2025. [3](#)
- Ziyu Zhu, Xilin Wang, Yixuan Li, Zhuofan Zhang, Xiaojian Ma, Yixin Chen, Baoxiong Jia, Wei Liang, Qian Yu, Zhidong Deng, Siyuan Huang, and Qing Li. Move to understand a 3d scene: Bridging visual grounding and exploration for efficient and versatile embodied navigation. *International Conference on Computer Vision (ICCV)*, 2025. [2](#), [9](#), [11](#)

A 实现细节

A.1 动作规划模型

不同具身的轨迹尺度可能不同，例如室内机器人通常在米级范围内移动，而汽车的移动尺度可达数十米。如表 7 所示，我们通过乘以缩放因子 α_{task} ，将不同具身预测轨迹的各个维度归一化至 $[-1, 1]$ 。缩放因子并非取各维度的绝对最大值，而是采用各维度的第 99 百分位数，以避免异常值的影响。

具身	x(m)	y(m)	z(m)	$\theta(\text{rad})$
室内机器人 *	1.0	0.433	-	2.09
无人机 *	7.93	3.19	7.85	1.04
汽车 *	50.8	14.9	-	1.52

表 7: 不同具身预测轨迹各维度的缩放因子。

A.2 BATS 使用细节

导航开始时，若视觉词元数量未超过词元预算 B ，则保留全部视觉词元。一旦视觉词元超过预算 B ，便采用 BATS 基于遗忘曲线采样词元 (第 3.1.2 节)。实际使用中，我们针对给定词元预算 B 预计算 $P(t, T)$ ，以加速这一过程。若导航任务的时间跨度极长，例如达到数千步（这种情况很少发生），即使采用最小采样概率 ϵ ，视觉词元也可能超过预算。此时直接移除最旧的帧。

A.3 导航数据准备

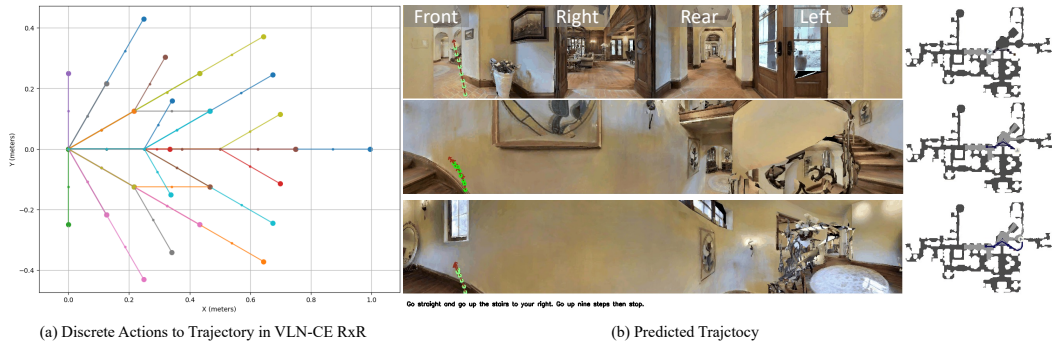


图 14: VLN-CE RxR 中 (a) 训练与 (b) 评估所用轨迹的可视化。

基于 Habitat 环境 (Savva et al., 2019a) 构建的导航任务使用 Move_Forward、Turn_Left、Turn_Right 和 Stop 等低层离散动作，但不同导航任务对这些离散动作的定义略有差异。例如，在 VLN-CE R2R 中，Turn_Left 表示转向 15 度，而在 VLN-CE RxR 和 HM3D-ObjNav 中则表示转向 30 度。为统一所有采用离散动作的导航任务，我们将前进 12.5 cm 或转向 15 度视为一次原子操作，再根据这些原子操作的累积构建轨迹。尽管所得轨迹可能呈锯齿状 (图 14)，但在所有导航数据集上微调后，本方法预测的轨迹平滑且能有效朝向目标。

A.4 真实世界部署系统

我们将本模型视为通用视觉—语言—动作 (Visual-Language-Action, VLA) 模型，可驱动不同具身完成多种导航任务。模型以一个或多个相机获取的视觉观测和指令为输入，直接预测轨迹；随后调用针对各类具身的现成 API（必要时可包含激光雷达或其他传感器），驱动机器人沿预测轨迹移动。

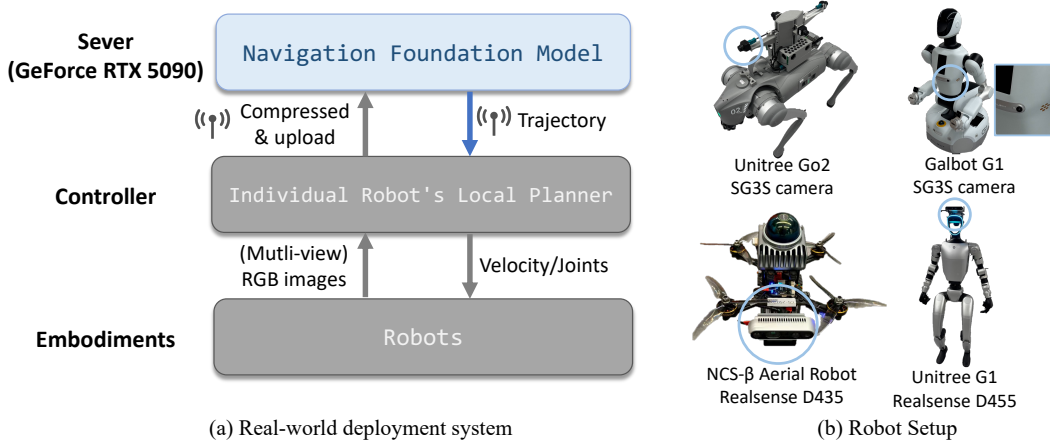


图 15: 真实世界部署设置。展示本方法的系统架构，以及本文测试使用的相应机器人。

图 15 展示了真实世界系统。具体而言，我们将模型部署在配备 GeForce RTX 5090 GPU 的远程服务器上，并通过互联网连接服务器与客户端（包括控制器和具身）。收到用户指令后，机器人压缩当前观测并传输至服务器；服务器处理观测与指令后输出轨迹。随后，各机器人的本地规划器处理该轨迹，并发送速度或关节控制等适当命令以驱动机器人。

B 补充实验

B.1 自动驾驶补充结果

表 8: 采用开环指标在面向规划的 nuScenes 数据集上进行比较。指标计算遵循 DiffusionDrive (Liao et al., 2024b)。粗体和下划线 分别表示最佳与次佳 结果。

方法	观测与结构		L2 (m) ↓				碰撞率 (%) ↓			
	相机	基于 VLM	1s	2s	3s	平均	1s	2s	3s	平均
ST-P3 (基线) (Hu et al., 2022)	✓	-	1.33	2.11	2.90	2.11	0.23	0.62	1.27	0.71
UniAD (基线) (Hu et al., 2023)	✓	-	0.45	0.70	1.04	0.73	0.62	0.58	0.63	0.61
VAD (Jiang et al., 2023)	✓	-	0.41	0.70	1.05	0.72	0.07	0.17	0.41	0.22
SparseDrive (Sun et al., 2024)	✓	-	0.29	0.58	0.96	0.61	<u>0.01</u>	0.05	<u>0.18</u>	0.08
DiffusionDrive (Liao et al., 2024b)	✓	-	0.27	0.54	0.90	0.57	0.03	0.05	0.16	0.08
DriveVLM (Tian et al., 2024)	✓	✓	0.18	0.34	0.68	0.40	0.10	0.22	0.45	0.27
EMMA (Hwang et al., 2024)	✓	✓	0.14	0.29	0.54	0.32	-	-	-	-
DME-Driver (Han et al., 2025)	✓	✓	0.45	0.91	1.58	0.98	0.05	0.28	0.55	0.29
Omni-Q (Wang et al., 2025d)	✓	✓	0.14	0.29	<u>0.55</u>	<u>0.33</u>	0.00	0.13	0.78	0.30
Omni-L (Wang et al., 2025d)	✓	✓	<u>0.15</u>	0.36	0.70	0.40	0.06	0.27	0.72	0.35
ORION (Fu et al., 2025a)	✓	✓	<u>0.17</u>	<u>0.31</u>	<u>0.55</u>	0.34	0.05	0.25	0.80	0.37
NavFoM (六视角)	✓	✓	0.26	0.39	0.60	0.42	0.07	<u>0.11</u>	<u>0.18</u>	<u>0.12</u>

表 8 报告了本方法在自动驾驶基准 nuScenes 上的性能。我们将其与专为自动驾驶设计的强基线比较；即使未显式建模驾驶相关信息，本方法仍取得与这些方法相当的性能。