arXiv:2602.02590v1 [cs.RO] 1 Feb 2026

StepNav：面向高效多模态视觉导航的结构化轨迹先验

罗旭博[1], 吴奥迪[1], 韩浩东[1],万雪[2*]，张伟[2],舒雷正[2] 和王瑞索[2]

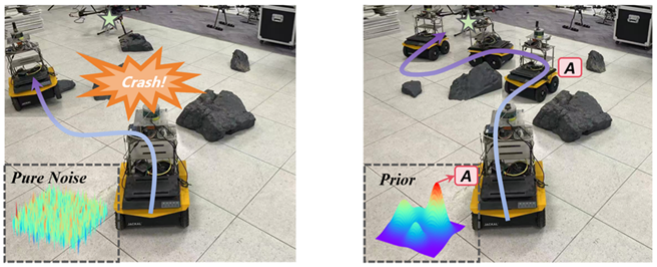
摘要——视觉导航是自主系统的基础，但在拥挤且不确定的环境中生成可靠轨迹仍是一项核心挑战。近期生成模型有望实现端到端合成，但它们依赖非结构化噪声先验，往往会产生不安全、低效或单模态的规划，无法满足实时性要求。我们提出了一种新颖框架 StepNav，通过引入基于变分原理的结构化多模态轨迹先验来弥合这一差距。StepNav 首先学习一个具备几何感知能力的成功概率场，以识别所有可行导航通道。随后利用这些通道构建显式的多模态混合先验，为条件流匹配过程提供初始化。该精化过程被表述为一个带有显式平滑性和安全性正则化的最优控制问题。通过以具有物理依据的候选方案取代非结构化噪声，StepNav 能以显著更少的步数生成更安全、更高效的规划。在仿真和真实世界基准测试中的实验表明，与最先进的生成式规划器相比，StepNav 在鲁棒性、效率和安全性方面均有稳定提升，推动了实用自主导航中可靠轨迹生成的发展。代码已发布于 https://github.com/LuoXubo/StepNav。

I. 引言

在复杂、非结构化环境（如杂乱的森林或城市街道）中导航的自主机器人，必须根据观测历史和目标图像等视觉输入生成平滑、可行的轨迹 [1], [2]。这些轨迹需要满足物理约束（例如连续性和避碰），同时还要考虑由遮挡、视觉歧义或多模态路径选择（例如在交叉口选择左转还是右转）带来的感知不确定性。传统方法往往将感知与规划分离，导致流程脆弱 [3]。相比之下，扩散策略 [4] 和条件流匹配方法 [5],等端到端生成模型，能够直接根据视觉数据合成轨迹，展现出良好的应用前景。然而，这些模型面临一个根本性挑战：它们难以高效地产生符合物理规律且能反映不确定性的轨迹，尤其是在存在视觉歧义时 [6]。

与能够明确建模状态空间和约束的传统运动规划问题不同，视觉导航存在一个固有的差距：高维、含噪的感知输入与低维可行轨迹流形之间

本研究得到国家自然科学基金委员会（项目编号：42171445）的资助。[1]罗旭博、吴奥迪和韩浩东就读于中国科学院大学，中国北京 100101(luoxubo23, wuaodi20, hanhaodong23)@mails.ucas.ac.cn[2]万雪、张伟、舒雷正和王瑞索就职于中国科学院空间应用工程与技术中心，中国北京 100094(wanxue, zhangwei, shuleizheng wangruisuo)@cas.ac.cn



(a) FlowNav (b) StepNav (Ours)

图 1。（a）FlowNav 在朝目标（无人机）导航时未能避开障碍物。（b）我们的 StepNav 估计成功概率场，以生成先验轨迹，从而加快优化并提高成功率。

弥合这一差距需要数学结构：模型必须同时（i）从原始视频中提取在时间上保持一致的动态，（ii）在存在感知歧义的情况下表示未来可能性的多模态特性，以及（iii）快速生成平滑、安全的轨迹，以满足闭环执行的速度要求。这三项要求往往相互冲突：强调多样性的方法会牺牲效率，强制保持连续性的方法会忽视不确定性，而能够实时运行的方法往往会在安全性方面作出妥协。

这一挑战体现在两个相互交织的局限性上。首先，轨迹生成通常从非结构化的高斯噪声先验开始，而该先验并不了解有效轨迹所在的流形。这种脱节导致需要经过漫长的优化过程（例如，超过 10 个扩散步数），才能将噪声塑造成连贯的路径，因此不适用于实时机器人技术 [7], [8]。其次，此类先验忽略了视觉输入中固有的几何结构，例如模糊走廊或被遮挡的障碍物，因而生成过度自信的轨迹，忽视替代路径，并危及动态环境中的安全性。尽管已有研究探索学习到的先验 [7], ，但这类方法往往需要额外的训练阶段，且缺乏直接从视觉线索中捕捉多模态路径选择的显式机制。

总之，核心难点在于构建动态一致、具备不确定性感知能力且可实时执行的轨迹生成器。现有方法依赖非结构化先验，难以权衡生成多样性与结构化、高效规划之间的关系。这促使我们提出一种有原则的方法：将关于导航可行性的结构化多模态知识直接注入生成过程，从而避免噪声驱动采样的低效与脆弱性。

为应对这一挑战，我们提出了**StepNav**，这是一种视觉

导航框架，将变分场估计与基于最优控制的生成式精化融为一体。我们的核心见解是，对可通行性的稠密学习表示可以作为构建结构化先验的有力基础。我们的贡献主要有三点：

- 我们提出一种几何感知成功概率场，该场通过双调和正则化偏微分方程学习得到，能够根据时间平滑视觉特征捕捉所有可行导航通道的拓扑。
- 我们提出一种方法，通过识别低能量路径，直接从该场中提取结构化的多模态轨迹候选混合，为生成式精化提供强有力的初始化。
- 我们将最终精化过程表述为一个正则化条件流匹配（Reg-CFM）问题，显式优化轨迹平滑度和安全性，从而以更少的积分步数生成质量更高的规划。

II. 相关工作

A. 端到端视觉导航

视觉导航已从将建图、定位和规划分离的经典模块化流程，发展为端到端的学习方法。早期学习方法通常依赖强化学习（RL） [9], [10], ，但可能难以实现较高的样本效率和泛化能力。较新的方法利用大规模预训练和复杂架构来提升鲁棒性。例如，ViNT [11] 通过在多样化数据集上进行预训练，为视觉导航引入了一种基础模型。其他研究则着重于纳入高斯泼溅等显式三维表示，以便为导航策略提供更好的空间定位 [12], [13]。尽管这些方法推动了技术发展，但它们通常会生成确定性规划，因而难以应对环境固有的模糊性。

B. 轨迹规划生成模型

为了解决确定性规划器的局限，生成模型，尤其是基于分数的扩散模型和流匹配模型，已成为轨迹生成的强大范式 [14], [15]。这些模型能够捕捉可行路径的复杂多模态分布。No-MaD [4] 是一项开创性工作，它直接根据视觉输入应用扩散模型生成导航动作序列。然而，这类模型面临的一个根本挑战是依赖无信息先验，通常是各向同性高斯噪声。从纯噪声出发，需要经过漫长且计算开销巨大的反向扩散过程才能生成结构化轨迹，这可能会损害其物理合理性和实时性能。

后续研究尝试通过使用信息量更丰富的先验来缓解这一问题。例如，NaviBridger [7],使用一种以单独训练的运动先验初始化的去噪扩散桥接模型，以缩短生成过程。NaviD [16] 引入深度信息，以更好地约束生成的路径。为了获得更快的推理速度，FlowNav [5] 利用CFM与Depth Anything-v2的深度先验的组合。然而，这些学习到的先验需要额外的训练阶段，并未明确建模视觉感知与物理空间对齐过程中产生的几何不确定性。一个常见的局限依然存在：这些先验要么缺乏结构，要么缺少严谨的几何基础。

我们的工作直接针对这一关键缺口。我们不再依赖非结构化噪声或单独训练的网络，而是提出一种统一且具有几何感知能力的先验，该先验基于从经时间细化的视频特征中计算出的学习到的成功概率场，并在运行时构建。通过提取显著峰值并沿该场中的低能量通道进行追踪，我们生成紧凑的多模态候选轨迹混合分布，为轻量级条件流匹配细化提供信息丰富的初始化。这种设计能够捕捉感知不确定性，避免代价高昂的辅助预训练，并与以噪声初始化的扩散方法相比，大幅缩短细化时间，从而实现高效、安全且多样化的实时视觉导航轨迹生成。

III. 方法

我们提出 **StepNav**，这是一个三阶段框架，可将原始视觉观测转化为高效且感知不确定性的轨迹，适用于实时导航：**(1)** 受动力学启发的视频特征精炼（DIFP）；**(2)** 由统一成功场驱动的结构化先验合成；**(3)** 正则化条件流匹配（Reg-CFM）精炼。形式化地，设机器人轨迹为一条连续曲线

$$\tau : [0,1] \to \mathbb{R}^d, \quad \tau(0) = x_{\text{init}}, \ \ \tau(1) = x_{\text{goal}}. \tag{1}$$

我们定义一个复合代价泛函

$$J(\tau) = \int_0^1 \Big(\lambda_g \, \ell_g(\tau(t)) + \lambda_s \, \ell_{\text{smooth}}(\dot{\tau}(t), \ddot{\tau}(t)) + \lambda_c \, \ell_{\text{coll}}(\tau(t)) \Big) \, dt, \tag{2}$$

其中， ℓ_g 衡量目标对齐， ℓ_{smooth} 惩罚曲率和加加速度， ℓ_{coll} 编码碰撞风险。StepNav 通过逐步增强结构化的阶段，逐步构造 $J(\tau)$ 的极小化解的近似。

A. 特征提取与细化（*DIFP*）

我们首先使用预训练的 V-JEPA2 编码器对输入图像历史 $\{I_t\}_{t=1}^{T}$ 和目标图像 I_{goal} 进行编码，得到特征

$$Z = [z_1, \ldots, z_T] \in \mathbb{R}^{T \times D}.$$

为减轻闪烁并确保时间一致性，我们将细化表述为一个投影问题：

$$\tilde{Z} = \arg\min_{Z'} \Big(\|Z' - Z\|_F^2 + \|LZ'\|_F^2 + \|Z' - z_c \mathbf{1}^\top\|_F^2 \Big), \tag{3}$$

其中， L 是用于约束平滑性的时间拉普拉斯算子， $z_c = \sum_{t=1}^{T} w_t z_t$ 是具有 $w_t \propto \exp(v_t)$的全局运动上下文。

欧拉-拉格朗日条件给出了闭式解，它平衡了对原始特征的保真度（第一项）

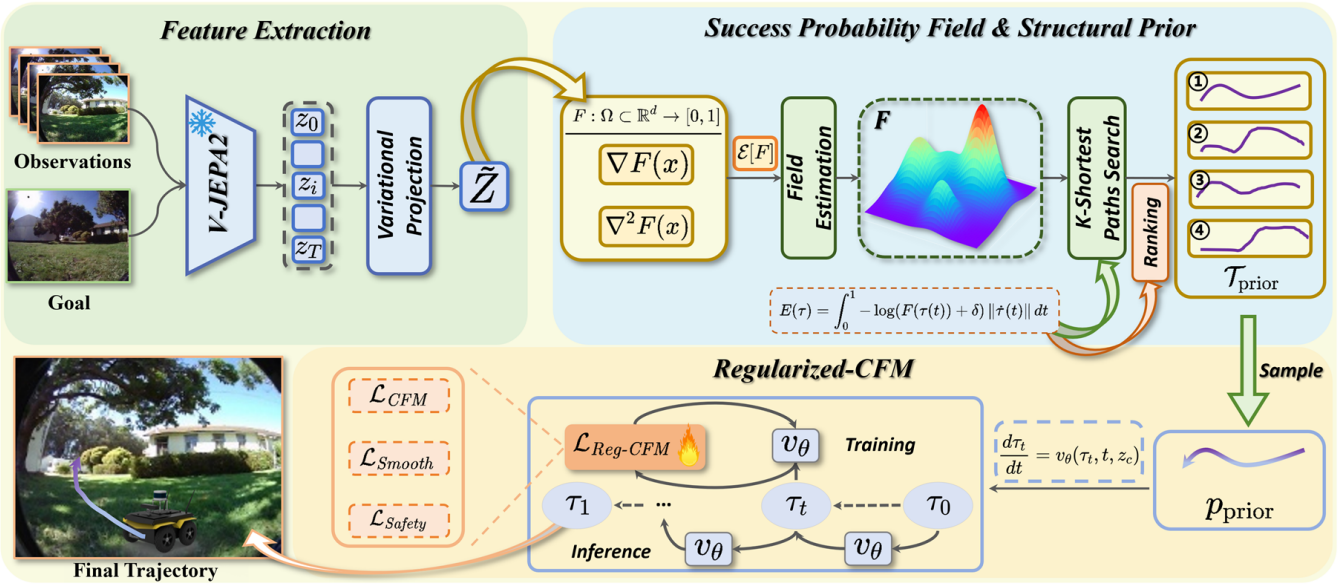


图 2. StepNav 概览。给定一系列输入图像和一张目标图像，我们首先通过 V-JEPA2 提取时序特征，并使用 DIFP 模块对其进行精炼。利用这些精炼后的特征预测成功概率场 F。基于该场，我们估计若干条先验轨迹，并根据能量景观 $E(\tau)$ 对先验轨迹进行采样。随后，通过 Reg-CFM 将该先验精炼为平滑且可行的轨迹。

、时间平滑性（第二项）以及与全局运动上下文的对齐程度（第三项）：

$$\tilde{Z} = (2I + L^\top L)^{-1}(Z + z_c \mathbf{1}^\top), \tag{4}$$

确保精炼嵌入 $\tilde{z}_t$ 在时间上平滑且能够感知目标。

B. 成功概率场与结构化先验

根据 $(Z,\ z_c)$，轻量级卷积解码器头预测连续的成功概率场 $F : \Omega \subset \mathbb{R}^d \to [0,1]$，用于表征局部可通行性。该场作为一种经过学习、具有空间感知能力的碰撞代价逆函数近似，$1 - \ell_{\text{coll}}(x)$。我们将 F 定义为变分能量的极小化解

$$\mathcal{E}[F] = \int_\Omega \Big((F(x) - y(x))^2 + \mu \|\nabla F(x)\|^2 + \nu \|\nabla^2 F(x)\|^2 \Big)\, dx, \tag{5}$$

其中，$y(x)$ 是来自专家演示的二元标签。相应的欧拉-拉格朗日方程为 F 导出一个经双调和正则化的泊松偏微分方程：

$$-\nu \Delta^2 F(x) + \mu \Delta F(x) + (F(x) - y(x)) = 0. \tag{6}$$

直观而言，$\|\nabla F\|^2$ 项用于保证平滑性，而双调和项 $\|\nabla^2 F\|^2$ 则会惩罚场中的高曲率，从而促进形成宽阔、平滑且成功概率较高的“走廊”，这非常适合路径规划。给定 F，我们定义一个能量景观

$$E(\tau) = \int_0^1 -\log(F(\tau(t)) + \delta)\, \|\dot{\tau}(t)\|\, dt, \tag{7}$$

用于衡量穿越路径 τ 的难度 。结构化候选先验是在边界条件 $\tau(0) = x_{\text{init}}, \tau(1) =$ 下，提取为 $E(\tau)$ 的近似极小值。x_{goal}。实际操作中，通过将 Ω 离散化为图并应用 K最短路径算法来求解。

为保留多模态性，我们采用贪心最大最小豪斯多夫准则，选取一个具有多样性的子集$\mathcal{T}_{\text{prior}} = \{\tau^{(m)}\}_{m=1}^{M}$ 。由此得到一个混合先验分布：

$$p_{\text{prior}}(\tau) = \sum_{m=1}^{M} \pi_m\, \delta(\tau - \tau^{(m)}), \quad \pi_m \propto \exp\Big(\tfrac{S(\tau^{(m)})}{T}\Big), \tag{8}$$

其中，$S(\tau^{(m)})$对成功概率、路径长度和曲率进行评分。

C. 正则化条件流匹配（*Reg-CFM*）

结构化先验虽然能提供可行的初始轨迹，但无法充分保证平滑性或安全性。因此，我们使用经过正则化以遵循物理约束的条件流匹配模型来优化样本。

给定初始轨迹$\tau_0 \sim p_{\text{prior}}$，我们将优化过程表述为一个最优控制问题：

$$\frac{d\tau_t}{dt} = v_\theta(\tau_t, t, z_c), \quad t \in [0,1], \tag{9}$$

其中 v_θ 是学习到的流场，z_c 提供全局上下文。

标准流匹配使 v_θ 与来自随机插值的目标流 u_t对齐。为促进可行导航，我们在此损失中加入平滑性和安全性正则项：

$$\min_{v_\theta}\ \mathbb{E}_{\tau_0,\tau_1,t}\Big[\|v_\theta(\tau_t,t,z_c) - u_t(\tau \mid \tau_0,\tau_1)\|^2 + \rho \|\dddot{\tau}_t\|^2 + \kappa \phi(\tau_t) \Big], \tag{10}$$

其中，第二项用于惩罚较大的加加速度，而 $\phi(\tau_t)$ =–log($d(\tau_t) - \epsilon$) 是一个对数障碍函数，通过抑制靠近障碍物的状态来防止碰撞。

在实现中，我们对离散化的航点 $\{x_k\}$ 使用两个显式项：

$$\mathcal{L}_{\text{smooth}} = \sum_{k=1}^{K-3} \|x_{k+3} - 3x_{k+2} + 3x_{k+1} - x_k\|_2^2, \quad (11)$$

通过有限差分惩罚加加速度，并且

$$\mathcal{L}_{\text{safe}} = -\sum_{k=1}^{K} \log(\max(d(x_k) - \epsilon, 0)), \quad (12)$$

它起到屏障的作用，防止进入不安全区域；其中，到障碍物的距离 $d(x_k)$根据成功场估算为 $d(x_k) \propto 1 - F(x_k)$。

因此，总体训练损失为

$$\mathcal{L}_{\text{Reg-CFM}} = \mathcal{L}_{\text{FM}} + \rho\mathcal{L}_{\text{smooth}} + \kappa\mathcal{L}_{\text{safe}}. \quad (13)$$

在推理时，优化过程简化为对学习到的流进行积分：

$$\tau_{t+\Delta t} = \tau_t + \Delta t\, v_\theta(\tau_t, t, z_c), \qquad \Delta t = \tfrac{1}{N}, \quad (14)$$

初始值为 $\tau_0 \sim p_{\text{prior}}$。由于先验本身已具有结构且支持多模态，只需少量积分步数 N 即可实现实时轨迹生成。

IV. 实验

我们开展了一系列全面的实验，以验证 StepNav 在各种具有挑战性且多样化的环境中的性能。我们的评估旨在严谨地回答与核心贡献相对应的三个关键问题：

总体有效性： 在具有挑战性的标准基准测试中，StepNav 在成功率、安全性和路径质量方面是否优于其他最先进的视觉导航方法?

组件必要性：我们框架的关键组件（即 DIFP 特征精炼、结构化多模态先验和 Reg-CFM）是否都对其性能至关重要?

实时可行性：所提出的方法是否足够高效，能够在机器人硬件上实现潜在的实时部署？其延迟与其他生成模型相比如何?

A. 实验设置

数据集。 Following [4], 我们汇总了训练数据集，包括 RECON [23], SCAND [24], GoStanford [25],和 SACSoN [26]。这些数据集共包含超过 1,450 个场景，涵盖遮挡、动态障碍物和光照变化等多种挑战。

为确保评估具有稳健性，我们主要在两个代表不同领域、标准且具有挑战性的基准测试上进行定量比较：室内（*Stanford 2D-3D-S*） *[27]*和室外（*Gazebo citysim*） *[28]*。这些场景带来了严重遮挡、动态障碍物、模糊走廊以及多变光照等挑战。**基线方法**。我们将 StepNav 与一组具有代表性的最先进方法进行比较，其中包括生成式和确定性方法：*ViNT* [11]: 一种基于预训练基础模型的强大确定性视觉导航模型。*NoMaD* [4]: 一种开创性的基于扩散的导航策略，从无结构的高斯噪声先验开始。*NaviBridger* [7]: 一种改进的扩散模型，使用单独训练的运动先验来初始化去噪扩散桥，旨在提高效率。*FlowNav* [5]: 一种条件流匹配模型，从高斯噪声中细化轨迹，不使用结构化先验或正则化。

为确保公平比较，所有基于学习的方法均使用作者提供的官方代码库和推荐的超参数，在我们特定的数据集划分上进行训练。

任务与评估协议。我们按照 [4]确立的协议，在两项不同的任务上评估所有方法：基本任务： 标准的点目标导航，智能体被部署在已见过的环境中（来自训练分布中的场景），并须导航至指定目标。适应任务： 更具挑战性的零样本迁移任务，智能体须在完全未见过且存在显著领域偏移的环境中导航（例如，光照、天气状况和物体纹理各不相同），以检验泛化能力。如果智能体在 500 步内到达距目标 0.2 米以内的位置，则视为该回合成功。

评估指标。 我们报告四项标准指标，以全面评估性能：

- 成功率（*SR*，% ↑）： 智能体在预先设定的容差范围内到达目标的回合所占百分比，是衡量任务完成情况的主要指标。
- 路径长度加权成功率（*SPL*， ↑）： 一种标准导航指标， [29], 定义为 $\text{SPL}_i = S_i \cdot L_i/\max(P_i, L_i)$，其中，若到达目标（在 0.2 米范围内， ≤500 步），则为 $S_i = 1$ ，否则为 0； L_i 为最短路径长度， P_i 为执行路径长度。报告的 SPL 为所有回合的平均值。
- 碰撞率 *(%，* ↓*)*：以碰撞结束的回合所占的百分比，直接衡量安全性。
- 最小跃度 *(MS，* 米^2/秒^7 ↓*)*：衡量跃度平方（位置的四阶导数）积分的指标，用于量化轨迹平滑度。数值越低，表示路径在动力学上越可行、行驶体验越舒适。

实现细节。所有模型均在 NVIDIA RTX 4090 上训练。为确保完全可复现，我们将在论文发表后发布代码和评估框架。我们在配备 NVIDIA Jetson AGX Orin 的真实 Clearpath Jackal机器人上部署了 StepNav，展示了其在室内和室外环境中的实时导航能力。该机器人使用前向 RGB摄像头获取视觉输入，并依靠机载计算进行轨迹生成和执行。

B. 与最先进方法的比较

如表 I 所总结，StepNav 在室内和室外导航任务中均展现出相较所有基线方法的明显性能优势。在室内基本任务中，StepNav 的成功率达到 95%，比表现次优的

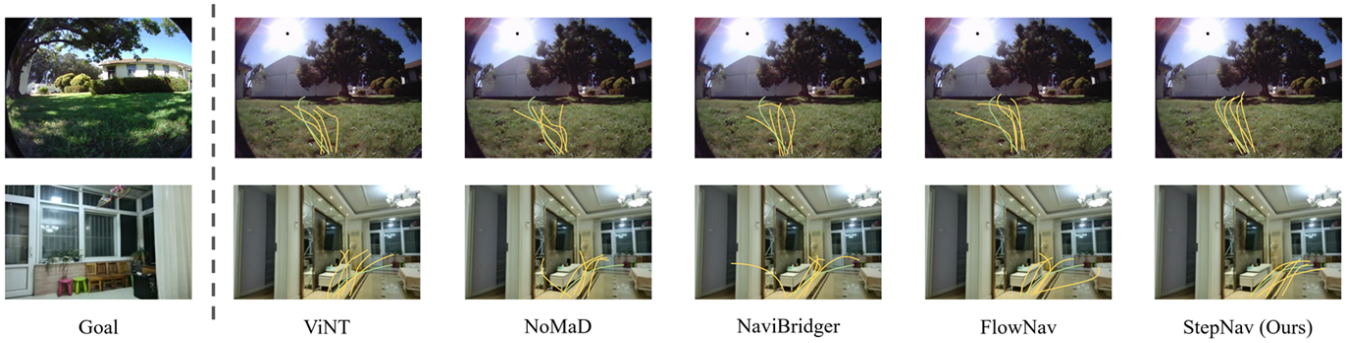


图 3. StepNav 在复杂室内与室外导航任务中的示例。上方一行展示了模拟城市环境中的室外导航，StepNav 成功与其他智能体一同通过多车道交叉路口。下方一行展示了杂乱房间中的室内导航，演示了对静态障碍物的避让。其中黄色轨迹为估计路径，绿色圆点表示真实航点。

表 I 室内（Stanford 2D-3D-S）与室外（Citysim）导航任务的定量结果，以均值 ±标准差报告。

Task	方法	室内（Stanford 2D-3D-S）				室外（Citysim）			
		成功率 (%) ↑	SPL ↑	碰撞率 (%) ↓	运动平滑度 (m^2/s^7) ↓	成功率 (%) ↑	SPL ↑	碰撞率 (%) ↓	运动平滑度 (m^2/s^7) ↓
基础 Task	ViNT	68±2.7	0.71±0.04	1.0±0.25	0.28±0.04	20±3.1	0.68±0.05	0.8±0.22	0.35±0.05
	NoMaD	86±2.5	0.69±0.04	0.7±0.24	0.30±0.04	22±3.4	0.65±0.05	0.6±0.21	0.37±0.05
	NaviBridger	92±1.9	0.73±0.03	0.6±0.18	0.26±0.03	44±3.6	0.69±0.04	0.6±0.20	0.33±0.04
	FlowNav	90±2.2	**0.81±0.02**	0.8±0.22	0.22±0.03	40±3.1	0.73±0.04	**0.5±0.18**	0.30±0.04
	StepNav	**95±1.1**	0.80±0.02	**0.6±0.12**	**0.20±0.02**	**57±3.0**	**0.76±0.03**	**0.5±0.10**	**0.28±0.03**
适应 Task	ViNT	28±3.4	0.63±0.05	1.6±0.35	0.33±0.04	38±3.8	0.60±0.06	0.4±0.13	0.40±0.05
	NoMaD	32±3.7	0.61±0.05	1.3±0.33	0.35±0.04	52±4.1	0.58±0.06	0.3±0.12	0.42±0.05
	NaviBridger	88±2.2	0.65±0.04	0.5±0.20	0.31±0.03	64±4.2	0.62±0.05	0.3±0.11	0.38±0.04
	FlowNav	85±2.5	0.71±0.04	0.6±0.21	0.33±0.03	65±4.0	0.70±0.05	0.4±0.12	0.36±0.04
	StepNav	**90±1.6**	**0.74±0.03**	**0.4±0.12**	**0.24±0.02**	**68±3.5**	**0.71±0.04**	**0.3±0.10**	**0.32±0.03**

基线方法（NaviBridger）高出 3 个百分点。在更具挑战性的适应任务中，这一性能提升更加显著。重要的是，这种改进并不仅体现在成功率上。StepNav 还取得了最高的 SPL 分数，表明其成功路径更加高效。此外，它的最小跃度最低，且碰撞率并列最低，证明其能够生成更加安全、平滑的轨迹。这一卓越性能源于结构化先验，该先验提供了强大的几何感知初始化。不同于 No-MaD 必须耗费大量精炼步骤来塑造非结构化噪声，也不同于其单独训练的先验本质上是单峰的 NaviBridger，StepNav 的在线多模态先验能够稳健地处理模糊场景（例如路口），从而制定出更有效、更安全的规划。

我们的多模态先验在模糊场景中的效果尤为明显。图 4 定性展示了 StepNav 遇到具有挑战性的 T型路口时的内部机制。在这个交叉口，学习到的成功概率场正确识别出两条可行的走廊：一条直行，另一条向右。随后，我们的先验提取方法为每条可能的走廊生成不同的候选轨迹（灰线）。接着，根据我们的能量函数选取得分最高的候选轨迹，并通过 Reg-CFM 将其优化为平滑的最终轨迹（黄线）。相比之下，基线方法在这种情况下，依赖单峰或非结构化先验的方法往往会过早地选定一条路径，而这条路径可能并不正确。



图4. StepNav 在模糊 T型路口处核心机制的定性可视化。灰色线条表示生成的先验轨迹，黄色线条表示 StepNav 和其他基线模型的最终轨迹。真实轨迹以绿色显示。

C. 消融研究

为了剖析 StepNav 的性能并验证我们的设计选择，我们在 Stanford 2D-3D-S 数据集上进行了一系列消融实验。该数据集包含多种室内挑战场景。

1. 结构化先验的影响。这是最关键的消融实验，旨在检验我们的成功场驱动先验的价值。我们将完整模型与两个变体进行比较：

- 高斯先验： 我们的 Reg-CFM 使用标准各向同性高斯噪声进行初始化，因此实质上类似于基于流的 NoMaD。
- 峰值先验：一种简化的单峰先验，仅利用成功概率场中检测到的峰值并将其连接起来构成。

表 II 中的结果显示出明显且显著的性能层级。高斯先验变体的表现最差，这证实了无信息先验难以生成成功且高效的轨迹。即使采用简单的结构化先验（峰值先验），也能带来巨大提升（例如， +12.8%成功率），这表明以成功场为基础构建初始轨迹的优势。最后，我们完整的多模态先验进一步带来了显著提升，尤其是在成功率和 SPL 方面，验证了通过低能量走廊追踪探索多种假设的重要性。

表 II 结构化先验各组成部分的消融研究。

先验类型	成功率（%）↑	SPL ↑	碰撞率（%）↓	MS ↓
高斯先验	72.5	0.64	1.3	0.38
峰值先验	85.3	0.71	0.8	0.31
StepNav（完整模型）	**94.8**	**0.80**	**0.6**	**0.20**

2. 特征精炼的影响。为验证 DIFP 模块的贡献，我们将完整模型与一个变体 不含 *DIFP* 进行比较；该变体直接使用原始 V-JEPA2 特征进行成功场预测。如表 III 所示，移除 DIFP 会导致所有指标显著下降，成功率降低了 15%。这证实了我们的假设：原始视频特征包含时间不一致性和噪声，会对规划产生不利影响。因此，DIFP 模块通过强制实现运动连贯性，对下游性能至关重要。

表 III DIFP 特征精炼模块的消融研究。

变体	成功率（%）↑	SPL ↑	碰撞率（%）↓	MS ↓
不含 DIFP	79.8	0.61	1.9	0.29
StepNav（完整）	**94.8**	**0.80**	**0.6**	**0.20**

3. 成功概率场消融。 为证明我们学习得到的稠密成功概率场是更优的先验生成表示，我们将其与更简单的替代方案进行比较，同时保持所有其他框架组件（DIFP、路径提取逻辑、CFM 优化器）固定不变。

- 直接航点回归： 该变体完全移除了场的概念。模型头被修改为直接回归一系列关键航点，以检验稠密的中间表示是否必要。
- 深度场：我们使用预训练的单目深度估计器（DepthAnything）来创建一个简单的几何可通行性场，其中较高的值对应更远的空间（即开阔空间）。这用于检验简单的几何线索是否足够，而不必使用学习得到的面向任务的场。
- *StepNav*（学习场）：我们提出的完整方法。

表 IV 表明，我们的学习场显著优于两种替代方案。直接回归表现较差，因为它缺乏稠密场丰富的拓扑上下文，因而较为脆弱。深度场表现更好，但仍不及我们的学习场。这是因为我们的场不仅感知“空旷空间”，而且经过端到端训练，能够识别属于“通往目标的成功路径”的区域，从而嵌入纯几何所缺乏的关键任务导向语义。该实验证实，学习稠密且具有任务感知能力的成功场是我们方法的一项关键创新。值得注意的是，深度场在碰撞率方面略有更优表现，表明显式深度线索可能提供互补收益。这一发现促使未来工作将深度信息更直接地整合到我们的框架中。

表 IV 用于先验生成的表示形式消融研究。

先验生成方法	成功率 (%) ↑	SPL ↑	碰撞率 (%) ↓
航点回归	74.2	0.66	1.6
深度场	79.5	0.73	0.5
StepNav	**94.8**	**0.80**	**0.6**

4. 细化策略的影响。 最后，我们验证了选择 Reg-CFM 进行细化的合理性。我们以完整的结构化先验作为固定起点，比较以下方法：

- 原始 *DDIM* [30]（基于标准扩散的细化方法）。
- 原始 *CFM* [31]（不使用额外的正则化项）。
- *CFM*+$\mathcal{L}_{\text{平滑}}$（仅使用平滑性惩罚）。
- *CFM*+$\mathcal{L}_{\text{安全}}$（仅使用安全屏障）。
- *Reg-CFM*（同时 $\mathcal{L}_{\text{平滑}}$和 $\mathcal{L}_{\text{安全}}$）。

如表 V 所示，与原始 CFM 相比，仅加入 $\mathcal{L}_{\text{smooth}}$ 项即可提升所有指标，尤其是将最小跃度（MS 从 0.30 降至 0.23）显著降低，并使碰撞率减半以上（从 2.1% 降至 0.8%）。相反，仅加入 $\mathcal{L}_{\text{safe}}$ 项可将碰撞率显著降低至 0.6%，但对平滑性没有影响。同时使用两个正则化项时，我们的方法整体表现最佳，成功率和 SPL 进一步提升，且碰撞率和 MS 同时达到最低。这表明这两种约束具有互补性。即使初始先验很强，显式的任务相关正则化仍能提升可部署性。

我们还展示了经过 2、5 和 10 个精炼步骤后轨迹的演变情况。我们将 StepNav 与标准 CFM 基线进行比较，后者采用相同的架构，但以高斯噪声初始化。如图 5 所示，StepNav 的结构化先验已经能够捕捉场景的全局拓扑

表 V 精炼策略消融（均由同一结构化先验初始化；除非另有说明，精炼步骤为 10 步）。

策略	成功率 (%) ↑	SPL ↑	碰撞率 (%) ↓	MS ↓
DDIM（10 步）	84.1	0.68	1.9	0.33
CFM（10 步）	84.9	0.71	2.1	0.30
CFM+$\mathcal{L}_{\text{smooth}}$	89.0	0.77	0.8	0.23
CFM+$\mathcal{L}_{\text{safe}}$	87.1	0.75	0.6	0.30
Reg-CFM（我们的方法）	**94.8**	**0.80**	**0.6**	**0.20**

，使其能够仅用几个步骤就快速精炼出高质量轨迹。相比之下，CFM 从无结构先验开始，需要更多步骤才能收敛到合理的路径。即使经过 10 个步骤，CFM 的轨迹仍不够理想，且比 StepNav 的轨迹更不平滑。这一可视化结果证实，我们的结构化先验显著加快了收敛速度，并带来了更优的最终轨迹。

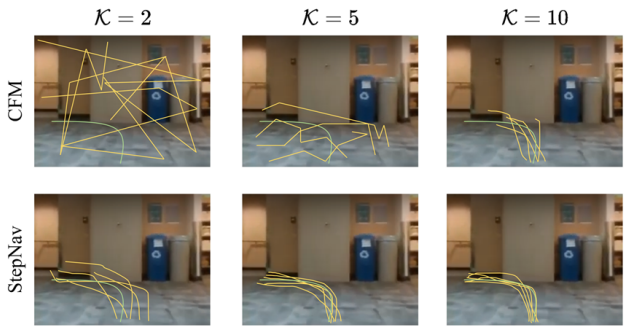


图 5. 在相同的结构化先验初始化下比较精炼策略。与从高斯噪声开始的 CFM 相比，StepNav 收敛更快，生成的轨迹也更平滑。

5. 精炼步骤消融。我们评估了精炼步骤对可扩展性的影响。如图6所示，StepNav 能够快速提升，并在五次迭代内收敛，成功率超过 95%，且碰撞率可忽略不计。相比之下，FlowNav 和 NaviBridger 需要十多步才能达到相当的成功率，同时碰撞率始终更高。这些结果表明，StepNav 通过更少的精炼步骤即可实现更快的收敛和更可靠的导航。

D. 真实世界部署

我们将 StepNav 部署在搭载 NVIDIA Jetson AGX Orin 的 Clearpath Jackal机器人上，展示了其在复杂室内环境中的实时导航能力。机器人使用前置 RGB摄像头获取视觉输入，并依靠 Orin AGX 进行轨迹生成与执行。该系统的运行频率为 8.5Hz，高于 FlowNav 的 6.9Hz，证实其适用于实时应用。图 7 展示了机器人成功穿越杂乱环境、避开障碍物并抵达指定目标。真实世界测试证实，StepNav 的

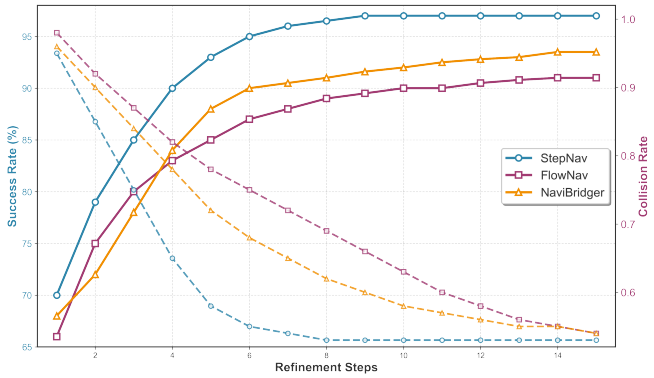


图 6. 成功率（SR，虚线）和碰撞率（Coll.，实线）随精炼步骤数的变化。StepNav 在五步内收敛，而 FlowNav 和 NaviBridger 则需要更多迭代才能达到相近的性能。

高效推理与稳健规划能力能够有效地从仿真迁移到实体部署。

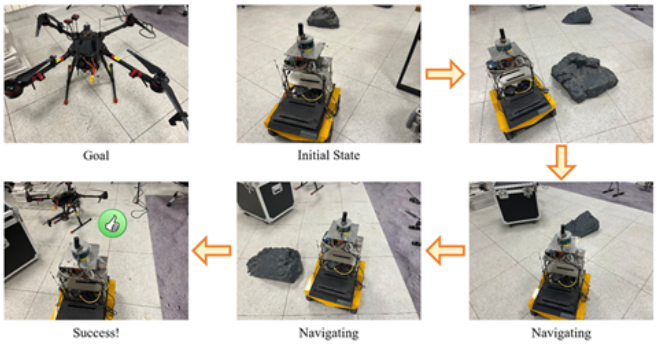


图 7. StepNav 部署于 Clearpath Jackal 机器人，在室内环境中导航。

V. 结论

本文提出了 StepNav，这是一种用于视觉导航的新型框架，结合了学习到的成功概率场、结构化多模态先验和正则化条件流匹配模型。我们的方法通过提供强大的几何感知初始化，并融入与任务相关的正则化，解决了先前生成式导航方法的关键局限。在具有挑战性的室内和室外基准测试上进行的大量实验表明，StepNav 在成功率、路径效率、安全性和平滑性方面显著优于最先进的基线方法。消融研究证实了各核心组件的必要性，而在机器人平台上的真实世界部署则验证了该方法的实用可行性。未来工作将重点通过把短期运动预测直接整合到成功概率场中来提升动态环境下的鲁棒性，并探索半监督或自监督方法，以减少对专家演示的依赖。

参考文献

[1] Wigness, M., Eum,S., Rogers, J., Han, D. & Kwon,H. 用于非结构化

室外环境中自主导航和视觉感知的RUGD数据集。2019 IEEE/RSJ国际智能机器人与系统会议（IROS）。第5000-5007页（2019） [2] Liang, Z., Fang, T., Dong, Z. & Li, J. 基于虚拟导航线的非结构化室外环境轮式机器人精准视觉导航方法。图像、视觉与智能系统国际会议（ICIVIS 2021）。第635-656页（2022） [3] Tang, Y., Zhao, C., Wang, J., Zhang, C., Sun, Q., Zheng, W., Du, W., Qian, F. & Kurths, J. 学习时代自主系统中的感知与导航：综述。《IEEE神经网络与学习系统汇刊》。34，9604-9624（2022） [4] Sridhar, A., Shah, D., Glossop, C. & Levine, S. NoMaD：用于导航与探索的目标掩码扩散策略。2024 IEEE国际机器人与自动化会议（ICRA）。第63-70页（2024） [5] Gode, S., Nayak, A., Oliveira, D., Krawez, M., Schmid, C. & Burgard, W. FlowNav：结合流匹配与深度先验的高效导航。（ 2025 2411 09524 ）， [6] https://arxiv.org/abs/ 。Janny, S., Poirier, H., Antsfeld, L., Bono, G., Monaci, G., Chidlovskii, B., Giuliari, F., Del Bue, A. & Wolf, C. 端到端训练智能体视觉导航中的推理：一种动力系统方法。计算机视觉与模式识别会议论文集。第12111-12121页（2025） [7] Ren, H., Zeng, Y., Bi, Z., Wan, Z., Huang, J. & Cheng, H. 先验确实重要：通过去噪扩散桥接模型实现视觉导航。计算机视觉与模式识别会议论文集。第12100-12110页（2025） [8] Song, Y., Dhariwal, P., Chen, M. & Sutskever, I. 一致性模型。（2023） [9] Zeng, F., Wang, C. & Ge, S. 基于深度强化学习的人工智能体视觉导航综述。《IEEE Access》。8，第135426-135442页（2020） [10] Kulh´anek, J., Derner, E. & Babu˘ska, R. 使用端到端深度强化学习在真实室内环境中进行视觉导航。《IEEE机器人与自动化快报》。6，4345-4352（2021） [11] Shah, D., Sridhar, A., Dashora, N., Stachowicz, K., Black, K., Hirose, N. & Levine, S. ViNT：用于视觉导航的基础模型。arXiv预印本 ArXiv:2306.14846。（2023） [12] Guo, W., Xu, X., Yin, H., Wang, Z., Feng, J., Zhou, J. & Lu, J. IGL-Nav：用于图像目标导航的增量式三维高斯定位。arXiv预印本 ArXiv:2508.00823。（2025） [13] Lei, X., Wang, M., Zhou, W. & Li, H. Gaussnav：用于视觉导航的高斯泼溅。《IEEE模式分析与机器智能汇刊》。（2025） [14] Carvalho, J., Le, A., Baierl, M., Koert, D. & Peters, J. 运动规划扩散：利用扩散模型学习和规划机器人运动。2023 IEEE/RSJ国际智能机器人与系统会议（IROS）。第1916-1923页（2023） [15] Ke, T., Gkanatsios, N. & Fragkiadaki, K. 3D Diffuser Actor：基于三维场景表示的策略扩散。arXiv预印本 ArXiv:2402.10885。（2024） [16] Zhang, J., Wang, K., Xu, R., Zhou, G., Hong, Y., Fang, X., Wu, Q., Zhang, Z. & Wang, H. NaviD：基于视频的视觉语言模型为视觉与语言导航规划下一步。arXiv预印本 ArXiv:2402.15852。（2024） [17] Jin, Y., Yuan, Z., Mu, Y. & Others 拥抱一致性：一种用于时空视频定位的单阶段方法。神经信息处理系统进展。35，第29192-29204页（2022） [18] Li, Q., Jia, X., Zhou, J., Shen, L. & Duan, J. 重新发现用于均匀分类的BCE损失。arXiv预印本 ArXiv:2403.07289。（2024） [19] Yen, J. 一种在一般网络中从所有源节点到给定目标寻找最短路径的算法。《应用数学季刊》。27，526-530（1970） [20] Nutanong, S., Jacox, E. & Samet, H. 一种增量式Hausdorff距离计算算法。《VLDB基金会会议录》。4，506-517（2011） [21] Lipman, Y., Chen, R., Ben-Hamu, H., Nickel, M. & Le, M. 用于生成建模的流匹配。（2023），https://arxiv.org/abs/ 2210 02747 。 [22] Consolini, L., Locatelli, M. & Minari, A. 一种用于加加速度受限速度规划的序列算法。《IEEE自动化科学与工程汇刊》。19，3192-3209（2021） [23] Shah, D., Li, J., Kar, A., Sutmann, H., Hagler, H., Schuhmacher, M., Snavely, N. & Levine, S. 面向敏捷空中使用大型模型进行探索。国际学习表征会议 *(ICLR)* (2021) [24] Karnan, C., Mathew, V. & Shah, D. 社会可接受导航数据集 (SCAND)。*IEEE*国际机器人与自动化会议论文集。(2022) [25] Hirose, T., Bedini, T., Smith, A., Maltz, D., Finn, C. & Levine, S. 深度视觉导航：全面综述。*IEEE*机器人与自动化快报。**4** (3) (2019) [26] Hirose, T., Wang, R., Dai, S., Lin, D., Finn, C. & Levine, S. SACSo N：语义与空间一致导航。*IEEE*机器人与自动化快报 **8**, 1155–1162 (2023) [27] Armeni, I., Sax, S., Zamir, A. & Savarese, S. 用于室内场景理解的联合二维-三维语义数据。*arXiv*预印本 *ArXiv:1702.01105*。(2017) [28] Koenig, N. & Howard, A. Gazebo（开源多机器人模拟器）的设计与使用范式。*2004 IEEE/RSJ*国际智能机器人与系统会议 *(IROS)(IEEE*目录号 *04CH37566)*。**3** 页 2149-2154 (2004) [29] Anderson, P., Chang, A., Chaplot, D., Dosovitskiy, A., Gupta, S., Koltun, V., Kosecka, J., Malik, J., Mottaghi, R., Savva, M. & 其他作者：关于具身导航智能体的评估。*arXiv*预印本 *ArXiv:1807.06757*。(2018) [30] Song, Y., Sohl-Dickstein, J., Kingma, D., Kumar, M., Ermon, S. & Poole, B. 通过随机微分方程进行基于分数的生成建模。国际学习表征会议 *(ICLR)*。(2021) [31] Lipman, Y., Rudi, A. & Bian, H. 用于生成建模的流匹配。*arXiv*预印本 *ArXiv:2210.02747*。(2022) [32] Ho, J., Jain, A. & Abbeel, P. 去噪扩散概率模型。神经信息处理系统进展。94, 991-1005 (2020) [33] Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Komeili, M., Muckley, M., Rizvi, A., Roberts, C., Sinha, K., Zholus, A., Mathieu, M. & Doersch, C. Vox jepa 2：用于三维表征学习的多视图掩码自编码器。*arXiv*预印本 *ArXiv:2506.12576*。(2025) [34] Gupta, S., Davidson, J., Levine, S., Sukthankar, R. & Malik, J. 视觉导航的认知地图构建与规划。*IEEE*计算机视觉与模式识别会议论文集。页 2616-2625 (2017)