2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

StepNav: Structured Trajectory Priors for Efficient and Multimodal Visual Navigation

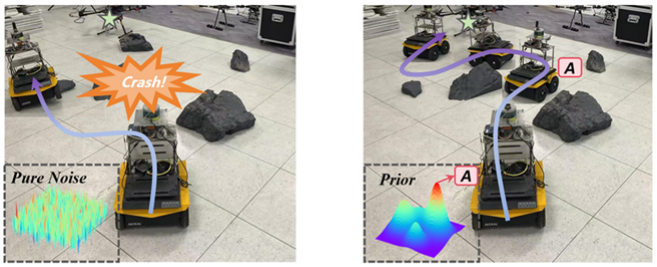
Xubo Luo[1], Aodi Wu[1], Haodong Han[1], Xue Wan[2*], Wei Zhang[2], Leizheng Shu[2] and Ruisuo Wang[2]

***Abstract*— Visual navigation is fundamental to autonomous systems, yet generating reliable trajectories in cluttered and uncertain environments remains a core challenge. Recent generative models promise end-to-end synthesis, but their reliance on unstructured noise priors often yields unsafe, inefficient, or unimodal plans that cannot meet real-time requirements. We propose StepNav, a novel framework that bridges this gap by introducing structured, multimodal trajectory priors derived from variational principles. StepNav first learns a geometry-aware success probability field to identify all feasible navigation corridors. These corridors are then used to construct an explicit, multi-modal mixture prior that initializes a conditional flow-matching process. This refinement is formulated as an optimal control problem with explicit smoothness and safety regularization. By replacing unstructured noise with physically-grounded candidates, StepNav generates safer and more efficient plans in significantly fewer steps. Experiments in both simulation and real-world benchmarks demonstrate consistent improvements in robustness, efficiency, and safety over state-of-the-art generative planners, advancing reliable trajectory generation for practical autonomous navigation. The code has been released at `https://github.com/LuoXubo/StepNav`.**

I. Introduction

Autonomous robots navigating complex, unstructured environments, such as cluttered forests or urban streets, must generate smooth, feasible trajectories from visual inputs like a history of observations and a goal image [1], [2]. These trajectories need to respect physical constraints (e.g., continuity and collision avoidance) while accounting for perceptual uncertainty arising from occlusions, visual ambiguities, or multi-modal routing choices (e.g., left vs. right at a junction). Classical approaches often decouple perception and planning, leading to brittle pipelines [3]. In contrast, end-to-end generative models, such as diffusion policies [4] and conditional flow matching approaches [5], have shown promise by directly synthesizing trajectories from visual data. However, these models are hampered by a fundamental challenge: they fail to produce physically plausible, uncertainty-aware trajectories efficiently, especially under visual ambiguity [6].

Unlike traditional motion planning problems where the state space and constraints can be explicitly modeled, visual navigation faces an inherent gap between high-dimensional, noisy perceptual inputs and the low-dimensional manifold of feasible trajectories. Bridging this gap requires mathematical structure: the model must simultaneously (i) extract temporally consistent dynamics from raw video, (ii) represent the multimodal nature of future possibilities under perceptual ambiguity, and (iii) generate smooth, safe trajectories fast enough for closed-loop execution. These three requirements often conflict: methods that emphasize diversity lose efficiency, those that enforce continuity ignore uncertainty, and those that are real-time often compromise on safety.



(a) FlowNav (b) StepNav (Ours)

Fig. 1. (a) FlowNav failed to avoid obstacles when navigating towards the goal (an UAV). (b) Our StepNav estimates the success probability field to generate the prior trajectory that can accelerate the refinement and improve the success rate.

This challenge manifests in two intertwined limitations. First, trajectory generation typically starts from an unstructured Gaussian noise prior, which is agnostic to the manifold of valid trajectories. This disconnect necessitates a long refinement process (e.g., more than 10 diffusion steps) to sculpt noise into a coherent path, rendering them unsuitable for real-time robotics [7], [8]. Second, such priors overlook the geometric structure inherent in visual inputs, such as ambiguous corridors or occluded obstacles, resulting in overconfident trajectories that ignore alternative paths and risk safety in dynamic environments. While learned priors have been explored [7], they often require auxiliary training stages and lack explicit mechanisms for capturing multi-modal path choices directly from visual cues.

In summary, the core difficulty lies in constructing trajectory generators that are *dynamically consistent, uncertainty-aware, and real-time feasible*. Existing approaches, by relying on unstructured priors, struggle with the fundamental trade-off between generative diversity and structured, efficient planning. This motivates our work: a principled method that injects structured, multimodal knowledge of navigation feasibility directly into the generative process, avoiding the inefficiency and brittleness of noise-driven sampling.

To tackle this challenge, we propose **StepNav**, a visual

This work was supported by the National Natural Science Foundation of China (Grant No. 42171445).

[1]Xubo Luo, Aodi Wu, and Haodong Han are with the University of Chinese Academy of Sciences, Beijing 100101, China (luoxubo23, wuaodi20, hanhaodong23)@mails.ucas.ac.cn

[2]Xue Wan, Wei Zhang, Leizheng Shu, and Ruisuo Wang are with the Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing 100094, China (wanxue, zhangwei, shuleizheng wangruisuo)@cas.ac.cn

navigation framework that unifies variational field estimation and optimal-control-based generative refinement. Our key insight is that a dense, learned representation of navigability can serve as a powerful foundation for constructing structured priors. Our contributions are threefold:

- We introduce a geometry-aware success probability field, learned via a biharmonic-regularized PDE, that captures the topology of all feasible navigation corridors from temporally-smoothed visual features.
- We propose a method to extract a structured, multi-modal mixture of trajectory candidates directly from this field by identifying low-energy paths, providing a powerful initialization for generative refinement.
- We formulate the final refinement as a regularized conditional flow-matching (Reg-CFM) problem, which explicitly optimizes for trajectory smoothness and safety, leading to higher-quality plans in fewer integration steps.

II. Related Work

A. End-to-End Visual Navigation

Visual navigation has evolved from classical modular pipelines, which separate mapping, localization, and planning, towards end-to-end learning-based approaches. Early learning methods often relied on reinforcement learning (RL) [9], [10], but could struggle with sample efficiency and generalization. More recent approaches leverage large-scale pretraining and sophisticated architectures to improve robustness. For instance, ViNT [11] introduced a foundation model for visual navigation by pre-training on diverse datasets. Other works have focused on incorporating explicit 3D representations, such as Gaussian Splatting, to provide better spatial grounding for the navigation policy [12], [13]. While these methods have advanced the state of the art, they often produce deterministic plans, limiting their ability to handle inherent environmental ambiguity.

B. Generative Models for Trajectory Planning

To address the limitations of deterministic planners, generative models, particularly score-based diffusion and flow-matching models, have emerged as a powerful paradigm for trajectory generation [14], [15]. These models can capture complex, multi-modal distributions of feasible paths. NoMaD [4] was a pioneering work that applied diffusion models to generate action sequences for navigation directly from visual inputs. However, a fundamental challenge in these models is their reliance on an uninformative prior, typically isotropic Gaussian noise. Starting from pure noise requires a long and computationally expensive reverse diffusion process to generate a structured trajectory, which can compromise physical plausibility and real-time performance.

Subsequent works have attempted to mitigate this issue by using more informative priors. NaviBridger [7], for example, uses a denoising diffusion bridge model initialized from a separately trained motion prior to shorten the generation process. NaviD [16] incorporates depth information to better constrain the generated paths. To obtain a faster inference speed, FlowNav [5] utilizes a combination of CFM and depth priors from Depth Anything-v2. However, these learned priors require additional training stages and do not explicitly model the geometric uncertainty that arises from the alignment between visual perception and physical space. A common limitation persists: the priors are either unstructured or lack a rigorous geometric foundation.

Our work directly addresses this critical gap. Instead of relying on unstructured noise or a separately learned network, we propose a unified, geometry-aware prior constructed on-the-fly from a learned success probability field computed over temporally refined video features. By extracting salient peaks and tracing low-energy corridors in this field we generate a compact, multi-modal mixture of candidate trajectories that serve as an informative initialization for a lightweight conditional flow-matching refinement. This design captures perceptual uncertainty, avoids costly auxiliary pretraining, and dramatically shortens refinement compared to noise-initialized diffusion methods, enabling efficient, safe, and diverse trajectory generation for real-time visual navigation.

III. Method

A. Problem Formulation

We formulate visual navigation as the problem of finding an optimal finite-time trajectory $\tau : [0,1] \to \mathbb{R}^d$ in the robot's local configuration space, given a history of visual observations $\mathcal{O} = \{I_1, \dots, I_T\}$ and a goal image I_{goal}. The objective is to minimize a composite cost functional $J(\tau)$ that balances goal reachability with physical feasibility:

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

where ℓ_g measures the likelihood of reaching the goal specified by I_{goal}, ℓ_{smooth} penalizes high curvature and jerk to ensure dynamic feasibility, and ℓ_{coll} encodes collision risk with the environment.

Directly minimizing (1) is intractable due to the high dimensionality of $\mathcal{O}$ and the non-convex nature of the collision cost. StepNav approximates the solution via a three-stage variational approach: (1) extracting temporally consistent features to estimate the geometry of ℓ_{coll} and ℓ_g, (2) finding coarse minimizers (priors) via a learned success field, and (3) refining these priors via regularized optimal control.

B. Dynamics-Inspired Feature Refinement (DIFP)

To accurately estimate the cost landscape from video, the visual features must be robust to noise and temporally consistent. We encode inputs using a pre-trained V-JEPA2 encoder to obtain raw features $Z = [z_1, \dots, z_T] \in \mathbb{R}^{T \times D}$. Raw features often contain high-frequency jitter uncorrelated with robot motion.

We propose a refinement scheme that is not merely low-pass filtering, but is *goal-conditional*. We compute a global motion context $z_c = \sum_{t=1}^{T} w_t z_t$ (where w_t are attention

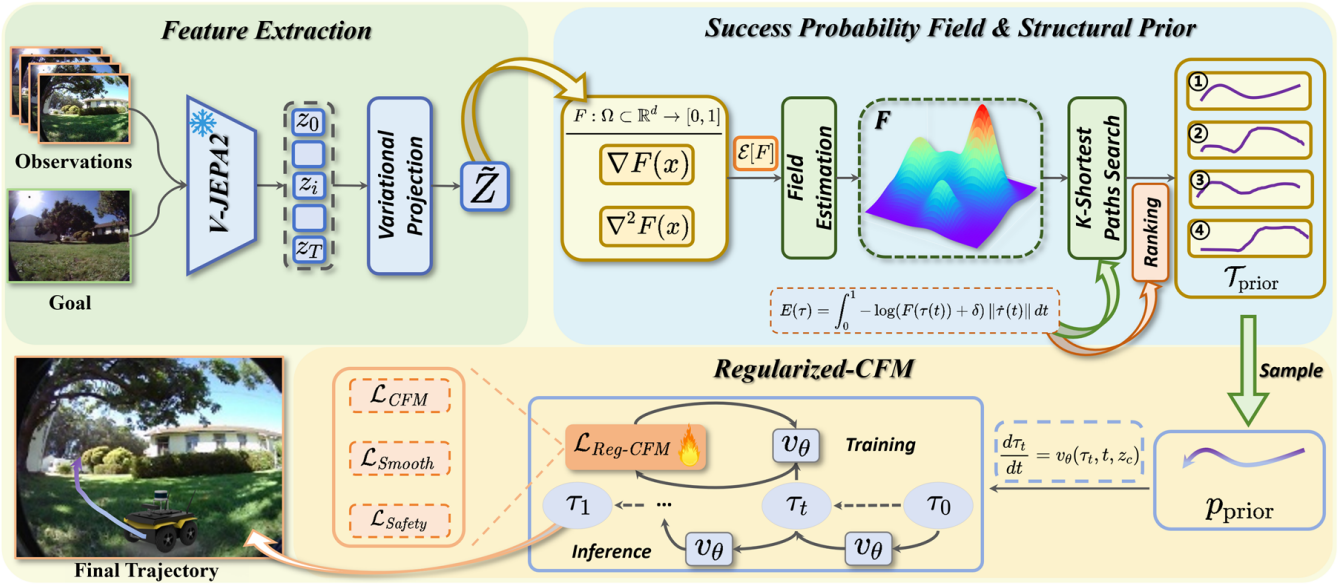


Fig. 2. Overview of StepNav. Given a sequence of input images and a goal image, we first extract temporal features via V-JEPA2 and refine them using our DIFP module. These refined features are used to predict a success probability field F. From this field, we estimate several prior trajectories and sample the prior trajectory according to the energy landscape $E(\tau)$. This prior is then refined into a smooth, feasible trajectory via Reg-CFM.

weights) representing the semantic intent of the episode. The refined features $\tilde{Z}$ are the solution to:

$$\tilde{Z} = \arg\min_{Z'} \Big(\underbrace{\|Z' - Z\|_F^2}_{\text{fidelity}} + \underbrace{\|LZ'\|_F^2}_{\text{smoothness}} + \underbrace{\|Z' - z_c \mathbf{1}^\top\|_F^2}_{\text{goal alignment}} \Big), \tag{2}$$

where L is the temporal Laplacian. The third term forces the feature trajectory to lie close to the global goal context, suppressing task-irrelevant visual variations. The closed-form solution is:

$$\tilde{Z} = (2I + L^\top L)^{-1}(Z + z_c \mathbf{1}^\top). \tag{3}$$

This mechanism ensures that the downstream field estimation focuses on the stable, goal-oriented structure of the environment rather than transient visual noise.

C. Success Probability Field and Structured Prior

We approximate the inverse cost landscape via a success probability field $F : \Omega \subset \mathbb{R}^d \to [0, 1]$. A lightweight convolutional head predicts F from $(\tilde{Z}, z_c)$. We define F as the minimizer of the variational energy:

$$\mathcal{E}[F] = \int_\Omega \Big((F(x) - y(x))^2 + \mu \|\nabla F(x)\|^2 + \nu \|\nabla^2 F(x)\|^2 \Big)\, dx, \tag{4}$$

where $y(x)$ are binary labels from expert demonstrations. The Euler-Lagrange equation yields a biharmonic-regularized Poisson PDE:

$$-\nu \Delta^2 F(x) + \mu \Delta F(x) + (F(x) - y(x)) = 0. \tag{5}$$

The biharmonic term $\|\nabla^2 F\|^2$ is crucial: it penalizes sharp transitions, effectively "widening" the narrow feasible paths found in demonstrations into smooth navigable corridors.

We then define an energy landscape $E(\tau)$ related to the path integral of this field:

$$E(\tau) = \int_0^1 -\log(F(\tau(t)) + \delta)\, \|\dot{\tau}(t)\|\, dt. \tag{6}$$

Minimizing $E(\tau)$ corresponds to finding paths that maximize the cumulative probability of success while minimizing length. We solve this on a discretized graph using K-shortest path search to identify distinct local minima (modes). To preserve multi-modality, we select a diverse subset $\mathcal{T}_{\text{prior}} = \{\tau^{(m)}\}_{m=1}^M$ using a greedy max-min Hausdorff criterion. This yields a mixture prior:

$$p_{\text{prior}}(\tau) = \sum_{m=1}^{M} \pi_m\, \delta(\tau - \tau^{(m)}), \quad \pi_m \propto \exp\Big(\tfrac{S(\tau^{(m)})}{T}\Big). \tag{7}$$

D. Regularized Conditional Flow-Matching (Reg-CFM)

The structured prior provides a coarse solution to (1). We perform fine-grained refinement using a Conditional Flow-Matching (CFM) formulation interpreted as optimal control. Given $\tau_0 \sim p_{\text{prior}}$, we define the flow $\frac{d\tau_t}{dt} = v_\theta(\tau_t, t, z_c)$. We train v_θ to match the conditional vector field u_t while explicitly imposing the smoothness and safety constraints from (1):

$$\mathcal{L}_{\text{Reg-CFM}} = \mathbb{E}_{\tau,t}\Big[\|v_\theta(\tau_t, t, z_c) - u_t\|^2 + \rho \|\ddot{\tau}_t\|^2 - \kappa \sum_k \log(\max(d(x_k) - \epsilon, 0)) \Big], \tag{8}$$

where $d(x_k)$ is the distance to obstacles approximated via $1 - F(x_k)$. Inference becomes an integration of v_θ starting from the structured prior τ_0. Because τ_0 is already close to

the solution manifold, StepNav requires significantly fewer integration steps ($N = 5$) compared to noise-based diffusion ($N > 10$).

IV. Experiments

We conduct a comprehensive set of experiments to validate the performance of StepNav in a range of challenging and diverse environments. Our evaluation is designed to rigorously answer three central questions corresponding to our core contributions:

Overall Efficacy: Does StepNav outperform other SOTA visual navigation methods on challenging, standard benchmarks in terms of success, safety, and path quality?

Component Necessity: Are the key components of our framework (i.e., the DIFP feature refinement, the structured multi-modal prior, and the Reg-CFM) all critical to its performance?

Real-Time Feasibility: Is the proposed method efficient enough for potential real-time deployment on robotic hardware, and how does its latency compare to other generative models?

A. Experimental Setup

Datasets. Following [4], we aggregate the training datasets including RECON [23], SCAND [24], GoStanford [25], and SACSoN [26]. These datasets encompass over 1,450 total scenes, covering varied challenges like occlusions, dynamic obstacles, and lighting variations.

To ensure a robust evaluation, our primary quantitative comparisons are performed on two standard and challenging benchmarks that represent distinct domains: *Indoor (Stanford 2D-3D-S) [27]* and *Outdoor (Gazebo citysim) [28]*. This collection presents challenges such as severe occlusions, dynamic obstacles, ambiguous corridors, and varied lighting.

Baselines. We compare StepNav against a representative set of SOTA methods, including both generative and deterministic approaches: *ViNT* [11]: A strong deterministic visual navigation model based on a pretrained foundation model. *NoMaD* [4]: A pioneering diffusion-based policy for navigation that starts from an unstructured Gaussian noise prior. *NaviBridger* [7]: An improved diffusion model that uses a separately trained motion prior to initialize a denoising diffusion bridge, aiming for higher efficiency. *FlowNav* [5]: A conditional flow matching model that refines trajectories from Gaussian noise, without structured priors or regularization.

For fair comparison, all learning-based methods are trained on our specific dataset splits using the authors' official codebases and recommended hyperparameters.

Tasks and Protocols. We evaluate all methods on two distinct tasks, following the protocol established in [4]: *Basic Task:* Standard point-goal navigation where agents are spawned in seen environments (from the training distribution of scenes) and must navigate to a specified goal.*Adaptation Task:* A more challenging zero-shot transfer task where agents must navigate in entirely unseen environments with significant domain shifts (e.g., different lighting, weather conditions, and object textures) to test generalization. An episode is considered successful if the agent reaches within 0.2 meters of the goal in under 500 steps.

Metrics. We report four standard metrics to comprehensively assess the performance:

- *Success Rate (SR, % ↑):* The percentage of episodes where the agent reaches the goal within a predefined tolerance. The primary measure of task completion.
- *Success Weighted by Path Length (SPL, ↑):* A standard navigation metric [29], defined as $\mathrm{SPL}_i = S_i \cdot L_i / \max(P_i, L_i)$, where $S_i = 1$ if the goal is reached (within 0.2 m, $\leq$500 steps), else 0; L_i is the shortest path length, P_i the executed path length. Reported SPL is the mean over episodes.
- *Collision Rate (%, ↓):* The percentage of episodes ending in a collision, directly measuring safety.
- *Minimum Snap (MS, m^2/s^7 ↓):* A measure of the integral of the squared snap (fourth derivative of position), which quantifies trajectory smoothness. Lower values indicate more dynamically feasible and comfortable paths.

Implementation Details. All models were trained on NVIDIA RTX 4090. We will release our code and evaluation framework upon publication to ensure full reproducibility. We deployed StepNav on a real-world Clearpath Jackal robot equipped with an NVIDIA Jetson AGX Orin, demonstrating real-time navigation in both indoor and outdoor settings. The robot uses a forward-facing RGB camera for visual input and relies on onboard computation for trajectory generation and execution.

B. Comparison with SOTA Methods

As summarized in Table I, StepNav demonstrates a clear performance advantage over all baselines across both indoor and outdoor navigation tasks. On the indoor *Basic Task*, StepNav achieves an SR of 95%, outperforming the next best baseline (NaviBridger) by 3 percentage points. This performance gain is even more pronounced in the more challenging Adaptation Task. Crucially, the improvement is not just in the success rate. StepNav also achieves the highest SPL scores, indicating that its successful paths are more efficient. Furthermore, it records the lowest Minimum Snap and is tied for the lowest Collision Rate, demonstrating its ability to generate trajectories that are both safer and smoother. This superior performance stems from the structured prior, which provides a strong, geometry-aware initialization. Unlike NoMaD, which must expend many refinement steps to shape unstructured noise, or NaviBridger, whose separately trained prior is inherently unimodal, StepNav's on-the-fly, multi-modal prior allows it to robustly handle ambiguous scenarios (e.g., junctions), leading to more effective and safer plans.

The efficacy of our multi-modal prior is particularly evident in ambiguous scenarios. Fig. 4 provides a qualitative visualization of StepNav's internal mechanism when encountering a challenging T-junction. At this junction, the learned success probability field correctly identifies two viable corridors: one straight ahead and another to the right.

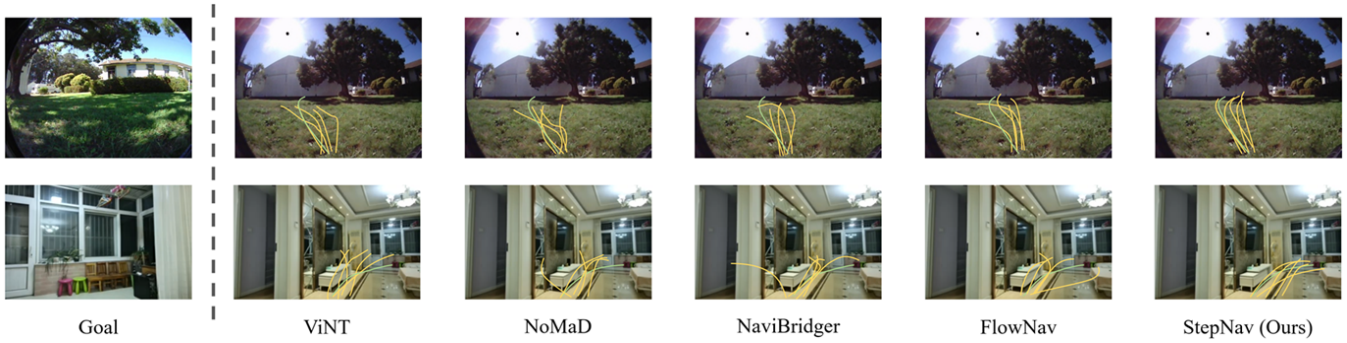


Fig. 3. Examples of StepNav in complex indoor and outdoor navigation tasks. The top row shows outdoor navigation in a simulated urban environment, where StepNav successfully navigates a multi-lane intersection with other agents. The bottom row illustrates indoor navigation in a cluttered room, demonstrating avoidance of static obstacles. The yellow trajectories are the estimated paths and the green dots represent the groundtruth waypoints.

TABLE I

QUANTITATIVE RESULTS REPORTED AS MEAN ± STANDARD DEVIATION ON INDOOR (STANFORD 2D-3D-S) AND OUTDOOR (CITYSIM) NAVIGATION TASKS.

Task	Method	Indoor (Stanford 2D-3D-S)				Outdoor (Citysim)			
		SR (%) ↑	SPL ↑	Coll. (%) ↓	MS (m^2/s^7) ↓	SR (%) ↑	SPL ↑	Coll. (%) ↓	MS (m^2/s^7) ↓
Basic 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**
Adaptation 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**

Subsequently, our prior extraction method generates distinct candidate trajectories for each potential corridor (gray lines). The highest-scoring candidate, based on our energy function, is then selected and refined via Reg-CFM into the smooth, final trajectory (yellow line). In contrast, baseline methods that depend on unimodal or unstructured priors often commit prematurely to a single, and potentially incorrect, path in such situations.

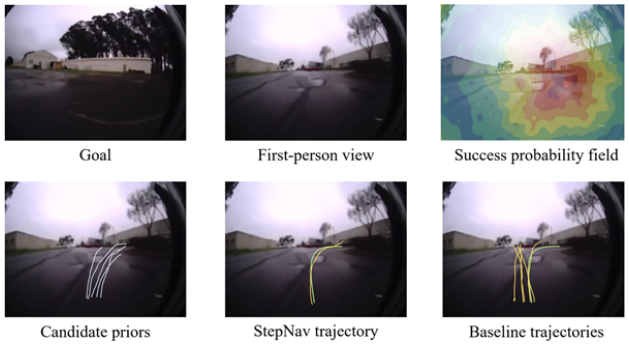


Fig. 4. A qualitative visualization of StepNav's core mechanism at an ambiguous T-junction. The gray lines indicate the generated prior trajectories, while the yellow lines represent the final trajectories of StepNav and other baseline models. The ground truth trajectories are shown in green.

C. Computational Performance

A key requirement for autonomous navigation is real-time performance. In Table II, we compare StepNav against baselines in terms of inference parameters, latency, and throughput on an NVIDIA AGX Orin. While ViNT is the fastest due to its deterministic nature, StepNav achieves a competitive 8.5 Hz, significantly outperforming other generative approaches like NoMaD (6.6 Hz). This efficiency is due to our structured prior, which reduces the required number of CFM integration steps to just 5, compared to 10+ steps for standard diffusion.

TABLE II

COMPUTATIONAL ANALYSIS ON NVIDIA AGX ORIN.

Method	Params	Lat. (ms)	FPS	Mem (GB)
ViNT	85	35	28.5	1.2
NoMaD	142	151	6.6	2.1
NaviBridger	150	148	6.7	2.2
FlowNav	145	145	6.9	2.2
StepNav	168	117	8.5	2.4

D. Ablation Studies

To deconstruct StepNav's performance and validate our design choices, we conduct a series of ablation experiments

on the Stanford 2D-3D-S dataset, chosen for its diverse set of indoor challenges.

1. Impact of the Structured Prior. This is the most critical ablation, designed to test the value of our success-field-driven prior. We compare our full model against two variants:

- *Gaussian Prior:* Our Reg-CFM initialized with a standard isotropic Gaussian noise, effectively making it similar to a flow-based version of NoMaD.
- *Peaks as Prior:* A simplified, unimodal prior formed by only using the detected peaks from the success probability field and connecting them.

Results in Table III show a clear and significant performance hierarchy. The *Gaussian Prior* variant performs worst, confirming that an uninformative prior struggles to generate successful and efficient trajectories. Using even a simple structured prior (*Peaks as Prior*) yields a massive improvement (e.g., +12.8% SR), demonstrating the power of grounding the initial trajectory in the success field. Finally, our full multimodal prior provides another substantial boost, particularly in SR and SPL, which validates the importance of exploring diverse hypotheses via low-energy corridor tracing.

TABLE III

ABLATION STUDY ON THE COMPONENTS OF THE STRUCTURED PRIOR.

Prior Type	SR (%) ↑	SPL ↑	Coll. (%) ↓	MS ↓
Gaussian Prior	72.5	0.64	1.3	0.38
Peaks as Prior	85.3	0.71	0.8	0.31
StepNav (Full)	**94.8**	**0.80**	**0.6**	**0.20**

2. Impact of Feature Refinement. To verify the contribution of the DIFP module, we compare our full model with a variant, *w/o DIFP*, which uses the raw V-JEPA2 features directly for success field prediction. As shown in Table IV, removing DIFP causes a stark degradation across all metrics, with SR dropping by 15%. This confirms our hypothesis that raw video features contain temporal inconsistencies and noise that are detrimental to planning. The DIFP module's ability to enforce motion-coherence is therefore crucial for downstream performance.

TABLE IV

ABLATION STUDY ON THE DIFP FEATURE REFINEMENT MODULE.

Variant	SR (%) ↑	SPL ↑	Coll. (%) ↓	MS ↓
w/o DIFP	79.8	0.61	1.9	0.29
StepNav (Full)	**94.8**	**0.80**	**0.6**	**0.20**

3. Ablation on the Success Probability Field. To demonstrate that our learned, dense success probability field is a superior representation for prior generation, we compare it against simpler alternatives, keeping all other framework components (DIFP, path extraction logic, CFM optimizer) fixed.

- *Direct Waypoint Regression:* This variant removes the concept of a field entirely. The model head is modified to directly regress a sequence of key waypoints, testing if a dense intermediate representation is necessary.
- *Depth Field:* We use a pre-trained monocular depth estimator (DepthAnything) to create a simple geometric traversability field where higher values correspond to more distant (i.e., open) space. This tests if simple geometric cues are sufficient, as opposed to a learned, task-oriented field.
- *StepNav (Learned Field):* Our full proposed method.

Table V shows that our learned field significantly outperforms both alternatives. Direct regression performs poorly, as it lacks the rich topological context of a dense field, making it brittle. The depth field performs better but is still inferior to our learned field. This is because our field is not just aware of "empty space" but is trained end-to-end to identify regions that are part of a "successful path to the goal," embedding crucial task-oriented semantics that pure geometry lacks. This experiment confirms that learning a dense, task-aware success field is a key innovation of our approach. Notably, the depth field achieves slightly better performance in terms of collision rate, suggesting that explicit depth cues may provide complementary benefits. This insight motivates future work on integrating depth information more directly into our framework.

TABLE V

ABLATION STUDY ON THE REPRESENTATION USED FOR PRIOR GENERATION.

Prior Generation Method	SR (%) ↑	SPL ↑	Coll. (%) ↓
Waypoint Regression	74.2	0.66	1.6
Depth Field	79.5	0.73	0.5
StepNav	**94.8**	**0.80**	**0.6**

4. Impact of Refinement Strategy. Finally, we validate our choice of the Reg-CFM for refinement. Using our complete structured prior as a fixed starting point, we compare:

- Vanilla *DDIM* [30] (standard diffusion-based refinement).
- Vanilla *CFM* [31] (no extra regularizers).
- *CFM+*$\mathcal{L}_{smooth}$ (only smoothness penalty).
- *CFM+*$\mathcal{L}_{safe}$ (only safety barrier).
- *Reg-CFM* (both $\mathcal{L}_{\text{smooth}}$ & $\mathcal{L}_{\text{safe}}$).

As shown in Table VI, compared to vanilla CFM, incorporating only the $\mathcal{L}_{\text{smooth}}$ term improves all metrics, most notably reducing the Minimum Snap (MS from 0.30 to 0.23) and more than halving the collision rate (from 2.1% to 0.8%). Conversely, adding only the $\mathcal{L}_{\text{safe}}$ term significantly lowers the collision rate to just 0.6%, but it has no effect on smoothness. When both regularizers are used together, our method achieves the best overall performance, with further improvements in SR and SPL, and the lowest Collision rate and MS simultaneously. This demonstrates that the two constraints are complementary. Even with a strong initial prior, explicit and task-relevant regularization yields improvements in deployable quality.

We also visualize the evolution of the trajectories across 2, 5, and 10 refinement steps. We compare StepNav with the standard CFM baseline, which uses the same architecture but

TABLE VI

REFINEMENT STRATEGY ABLATION (ALL INITIALIZED BY THE SAME STRUCTURED PRIOR; 10 REFINEMENT STEPS UNLESS OTHERWISE STATED).

Strategy	SR (%) ↑	SPL ↑	Coll. (%) ↓	MS ↓
DDIM (10 steps)	84.1	0.68	1.9	0.33
CFM (10 steps)	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 (Ours)	**94.8**	**0.80**	**0.6**	**0.20**

is initialized from Gaussian noise. As shown in Fig. 5, StepNav's structured prior already captures the global topology of the scene, allowing it to quickly refine to a high-quality trajectory within just a few steps. In contrast, CFM starts from an unstructured prior and requires many more steps to converge to a reasonable path. Even after 10 steps, CFM's trajectory remains suboptimal and less smooth compared to StepNav. This visualization confirms that our structured prior significantly accelerates convergence and leads to superior final trajectories.

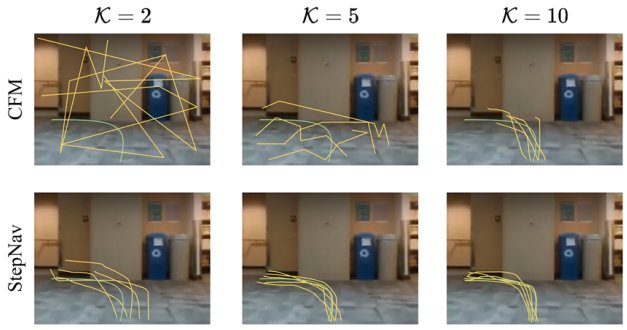


Fig. 5. Refinement strategies under identical structured prior initialization. StepNav converges faster and yields smoother trajectories compared to CFM, which starts from Gaussian noise.

5. Ablation on Refinement Steps. We evaluate the effect of refinement steps on scalability. As shown in Fig. 6, StepNav achieves rapid gains and converges within five iterations, reaching over 95% SR with negligible collision rates. By contrast, FlowNav and NaviBridger require more than ten steps to attain comparable SR, while consistently exhibiting higher collision rates. These results indicate that StepNav yields faster convergence and more reliable navigation with fewer refinements.

E. Real-World Deployment

We deployed StepNav on a Clearpath Jackal robot equipped with an NVIDIA Jetson AGX Orin, demonstrating real-time navigation in complex indoor settings. The robot uses a forward-facing RGB camera for visual input and relies on Orin AGX for trajectory generation and execution. The system operates at 8.5Hz, surpassing FlowNav's 6.9Hz, thereby confirming its suitability for real-time applications. Fig. 7 shows the robot successfully navigating through cluttered environments, avoiding obstacles, and reaching

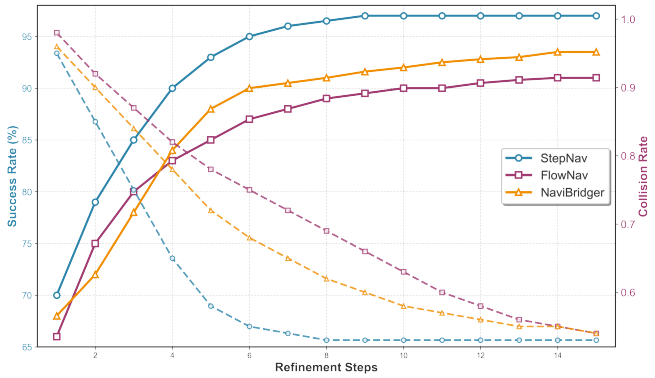


Fig. 6. Success rate (SR, dashed lines) and collision rate (Coll., solid lines) as functions of refinement steps. StepNav converges within five steps, whereas FlowNav and NaviBridger require more iterations to approach comparable performance.

specified goals. Real-world tests confirm that StepNav's efficient inference and robust planning capabilities effectively translate from simulation to physical deployment.

Fig. 7. Real-world deployment of StepNav on a Clearpath Jackal robot navigating through an indoor environment.

V. DISCUSSION AND LIMITATIONS

While StepNav demonstrates strong performance and efficiency, explicit limitations remain. First, the multi-stage architecture—encompassing encoder, refinement, field PDE, and flow matching—introduces engineering complexity compared to monolithic end-to-end Transformers. Errors in the success field estimation can propagate to the prior and final plan. Second, our approach relies on supervised binary labels from expert demonstrations to learn the success field. This limits scalability compared to self-supervised methods and may reduce transferability to environments drastically different from the training distribution. Third, our multimodal extraction assumes the existence of well-separated high-probability corridors. In wide, open spaces with uniform traversability, the energy landscape may become flat, leading to mode collapse or noisy prior initialization. Finally, while the system is real-time, it does not currently model the future intent of dynamic agents, treating them as dynamic obstacles rather than interactive entities.

VI. Conclusion

This paper presents StepNav, a novel framework for visual navigation that combines a learned success probability field, a structured multi-modal prior, and a regularized conditional flow matching model. Our approach addresses key limitations of prior generative navigation methods by providing a strong, geometry-aware initialization and incorporating task-relevant regularization. Extensive experiments on challenging indoor and outdoor benchmarks demonstrate that StepNav significantly outperforms state-of-the-art baselines in terms of success rate, path efficiency, safety, and smoothness. Ablation studies confirm the necessity of each core component, and real-world deployment on a robotic platform validates its practical feasibility. Future work will focus on improving robustness in dynamic environments by integrating short-term motion forecasts directly into the success probability field and exploring semi-supervised or self-supervised methods to reduce the reliance on expert demonstrations.

References

[1] Wigness, M., Eum, S., Rogers, J., Han, D. & Kwon, H. A rugd dataset for autonomous navigation and visual perception in unstructured outdoor environments. *2019 IEEE/RSJ International Conference On Intelligent Robots And Systems (IROS)*. pp. 5000-5007 (2019)

[2] Liang, Z., Fang, T., Dong, Z. & Li, J. An Accurate Visual Navigation Method for Wheeled Robot in Unstructured Outdoor Environment Based on Virtual Navigation Line. *The International Conference On Image, Vision And Intelligent Systems (ICIVIS 2021)*. pp. 635-656 (2022)

[3] Tang, Y., Zhao, C., Wang, J., Zhang, C., Sun, Q., Zheng, W., Du, W., Qian, F. & Kurths, J. Perception and navigation in autonomous systems in the era of learning: A survey. *IEEE Transactions On Neural Networks And Learning Systems*. **34**, 9604-9624 (2022)

[4] Sridhar, A., Shah, D., Glossop, C. & Levine, S. Nomad: Goal masked diffusion policies for navigation and exploration. *2024 IEEE International Conference On Robotics And Automation (ICRA)*. pp. 63-70 (2024)

[5] Gode, S., Nayak, A., Oliveira, D., Krawez, M., Schmid, C. & Burgard, W. FlowNav: Combining Flow Matching and Depth Priors for Efficient Navigation. (2025), https://arxiv.org/abs/2411.09524

[6] Janny, S., Poirier, H., Antsfeld, L., Bono, G., Monaci, G., Chidlovskii, B., Giuliari, F., Del Bue, A. & Wolf, C. Reasoning in visual navigation of end-to-end trained agents: a dynamical systems approach. *Proceedings Of The Computer Vision And Pattern Recognition Conference*. pp. 12111-12121 (2025)

[7] Ren, H., Zeng, Y., Bi, Z., Wan, Z., Huang, J. & Cheng, H. Prior does matter: Visual navigation via denoising diffusion bridge models. *Proceedings Of The Computer Vision And Pattern Recognition Conference*. pp. 12100-12110 (2025)

[8] Song, Y., Dhariwal, P., Chen, M. & Sutskever, I. Consistency models. (2023)

[9] Zeng, F., Wang, C. & Ge, S. A survey on visual navigation for artificial agents with deep reinforcement learning. *IEEE Access*. **8** pp. 135426-135442 (2020)

[10] Kulhánek, J., Derner, E. & Babuška, R. Visual navigation in real-world indoor environments using end-to-end deep reinforcement learning. *IEEE Robotics And Automation Letters*. **6**, 4345-4352 (2021)

[11] Shah, D., Sridhar, A., Dashora, N., Stachowicz, K., Black, K., Hirose, N. & Levine, S. ViNT: A foundation model for visual navigation. *ArXiv Preprint ArXiv:2306.14846*. (2023)

[12] Guo, W., Xu, X., Yin, H., Wang, Z., Feng, J., Zhou, J. & Lu, J. IGL-Nav: Incremental 3D Gaussian Localization for Image-goal Navigation. *ArXiv Preprint ArXiv:2508.00823*. (2025)

[13] Lei, X., Wang, M., Zhou, W. & Li, H. Gaussnav: Gaussian splatting for visual navigation. *IEEE Transactions On Pattern Analysis And Machine Intelligence*. (2025)

[14] Carvalho, J., Le, A., Baierl, M., Koert, D. & Peters, J. Motion planning diffusion: Learning and planning of robot motions with diffusion models. *2023 IEEE/RSJ International Conference On Intelligent Robots And Systems (IROS)*. pp. 1916-1923 (2023)

[15] Ke, T., Gkanatsios, N. & Fragkiadaki, K. 3d diffuser actor: Policy diffusion with 3d scene representations. *ArXiv Preprint ArXiv:2402.10885*. (2024)

[16] 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. *ArXiv Preprint ArXiv:2402.15852*. (2024)

[17] Jin, Y., Yuan, Z., Mu, Y. & Others Embracing consistency: A one-stage approach for spatio-temporal video grounding. *Advances In Neural Information Processing Systems*. **35** pp. 29192-29204 (2022)

[18] Li, Q., Jia, X., Zhou, J., Shen, L. & Duan, J. Rediscovering bce loss for uniform classification. *ArXiv Preprint ArXiv:2403.07289*. (2024)

[19] Yen, J. An algorithm for finding shortest routes from all source nodes to a given destination in general networks. *Quarterly Of Applied Mathematics*. **27**, 526-530 (1970)

[20] Nutanong, S., Jacox, E. & Samet, H. An incremental Hausdorff distance calculation algorithm. *Proceedings Of The VLDB Endowment*. **4**, 506-517 (2011)

[21] Lipman, Y., Chen, R., Ben-Hamu, H., Nickel, M. & Le, M. Flow Matching for Generative Modeling. (2023), https://arxiv.org/abs/2210.02747

[22] Consolini, L., Locatelli, M. & Minari, A. A sequential algorithm for jerk limited speed planning. *IEEE Transactions On Automation Science And Engineering*. **19**, 3192-3209 (2021)

[23] Shah, D., Eysenbach, B., Kahn, G., Rhinehart, N. & Levine, S. Rapid exploration for open-world navigation with latent goal models. *ArXiv Preprint ArXiv:2104.05859*. (2021)

[24] Karnan, H., Nair, A., Xiao, X., Warnell, G., Pirk, S., Toshev, A., Hart, J., Biswas, J. & Stone, P. Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation. *IEEE Robotics And Automation Letters*. **7**, 11807-11814 (2022)

[25] Hirose, N., Xia, F., Martin-Martin, R., Sadeghian, A. & Savarese, S. Deep visual mpc-policy learning for navigation. *IEEE Robotics And Automation Letters*. **4**, 3184-3191 (2019)

[26] Hirose, N., Shah, D., Sridhar, A. & Levine, S. Sacson: Scalable autonomous control for social navigation. *IEEE Robotics And Automation Letters*. **9**, 49-56 (2023)

[27] Armeni, I., Sax, S., Zamir, A. & Savarese, S. Joint 2d-3d-semantic data for indoor scene understanding. *ArXiv Preprint ArXiv:1702.01105*. (2017)

[28] Koenig, N. & Howard, A. Design and use paradigms for gazebo, an open-source multi-robot simulator. *2004 IEEE/RSJ International Conference On Intelligent Robots And Systems (IROS)(IEEE Cat. No. 04CH37566)*. **3** pp. 2149-2154 (2004)

[29] Anderson, P., Chang, A., Chaplot, D., Dosovitskiy, A., Gupta, S., Koltun, V., Kosecka, J., Malik, J., Mottaghi, R., Savva, M. & Others On evaluation of embodied navigation agents. *ArXiv Preprint ArXiv:1807.06757*. (2018)

[30] Song, Y., Sohl-Dickstein, J., Kingma, D., Kumar, M., Ermon, S. & Poole, B. Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations (ICLR)*. (2021)

[31] Lipman, Y., Rudi, A. & Bian, H. Flow matching for generative modeling. *ArXiv Preprint ArXiv:2210.02747*. (2022)

[32] Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Advances In Neural Information Processing Systems*. 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: Multi-view masked autoencoders for 3d representation learning. *ArXiv Preprint ArXiv:2506.12576*. (2025)

[34] Gupta, S., Davidson, J., Levine, S., Sukthankar, R. & Malik, J. Cognitive mapping and planning for visual navigation. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 2616-2625 (2017)