Training and Perception for Nomad Navigation

Authors Mathematics and Computing
Indian Institute of Science

April 9, 2025

Abstract

This report presents our work on implementing and analyzing the training and perception components of a visual navigation system based on diffusion policies, as adapted from the NOMAD (Navigation with Goal Masked Diffusion) framework. Our approach combines a visual perception backbone with a trajectory diffusion model to support both goal-directed and exploratory navigation. We trained our model on the SACSoN dataset, which features diverse real-world trajectories across various environments and robot platforms. We leverage a conditional diffusion model to generate multimodal waypoint predictions, enabling the agent to reason about complex, uncertain navigation scenarios. Our contributions include a detailed breakdown of the model architecture, training methodology, and evaluation metrics—particularly focusing on waypoint alignment through cosine similarity. We have left out the deployment aspects.

1 Introduction

Robotic learning for navigation in unfamiliar environments requires the ability to perform both task-oriented navigation (i.e., reaching a known goal) and task-agnostic exploration (i.e., searching for a goal in a novel environment). Traditionally, these functionalities are tackled by separate systems — for example, using subgoal proposals, explicit planning modules, or distinct navigation strategies for exploration and goal-reaching.

What is NoMaD?

NoMaD is a transformer-based diffusion policy designed for long-horizon, memory-based navigation, that can:

- Explore unknown places on its own (goal-agnostic behavior).
- Go to a specific place or object when given a goal image (goal-directed behavior).

Our project involves implementing the NoMaD Policy adapting its Transformer-based architecture and conditional diffusion decoder to learn from a rich, multimodal dataset (SACSoN) composed of real-world trajectories. Unlike traditional latent-variable models or methods that rely on separate generative components for subgoal planning, the unified diffusion policy exhibits superior generalization and robustness in unseen environments, while maintaining a compact model size. In this report, we focus on the perception and training components of this policy, emphasizing how a strong visual encoder combined with a diffusion-based decoder leads to improved alignment of predicted and ground-truth waypoints. We analyze the training dynamics, present key quantitative metrics such as cosine similarity and distance loss, and highlight the model's ability to generalize across diverse scenarios.

Overview of NoMaD Architecture

Refer to the Appendix A for preliminaries.

2 Implementation Details

2.1 Environment Setup

2.2 Data Pipeline

We used a pre-collected dataset of trajectories containing RGB observations, actions, and ground-truth waypoints. Data augmentations were not used in our initial experiments.

2.3 Training Procedure

Training was done on a single NVIDIA GPU using a batch size of 64. The training loop involved:

- Calculating diffusion loss from predicted vs ground-truth waypoints.
- Waypoint cosine similarity.
- Auxiliary action prediction losses.

Training checkpoints were saved every epoch. EMA models were also stored.

3 Perception Module

The ResNet18 backbone encodes RGB frames, while a transformer-based encoder maintains temporal context. This allows the policy to act based on history, crucial for long-horizon navigation.

3.1 Cosine Similarity Metrics

We track waypoint cosine similarity to evaluate how well the predicted and ground-truth waypoints align. Early training epochs show increasing cosine similarity, indicating improved waypoint alignment.

4 Results

4.1 Training Metrics

- Final training loss: ~1.11
- Cosine similarity: ~0.47 (multi-action waypoints)
- Distance loss: ~128

4.2 Observations

Loss plateaued after around 5,000 batches. Training logs show improvement in cosine similarity and reduction in loss. Action losses remained stable across UC and GC branches.

5 Challenges and Debugging

6 Conclusion and Future Work

We successfully trained the NOMAD policy and analyzed the perception module. Future work could involve domain randomization, hyperparameter tuning, and evaluating transfer to real-world or simulated environments.

References

1. H. Janner et al., ”NOMAD: Planning with Diffusion for Visual Navigation,” 2022.

2. Diffusion Policy GitHub Repository: `https://github.com/wayveai/diffusion-policy`

Appendices

A Related Work and Contextual Foundations of NoMaD

Exploration in unfamiliar enviroments is approached as the problem of efficient mapping, typically formulated around information maximization to guide the robot toward unexplored regions.
We factorize the classical exploration problem into two categories:

- Local exploration strategies that rely on current observations. Objective is to learn control policies that can take diverse,short-horizon actions
- Global exploration strategies that utilize a map of the environment. Basically a high-level planner based on a topological graph that uses the policy for long-horizon goal-seeking

Robots exploring a new area are essentially trying to map it efficiently—this means covering as much area as possible, ideally without wasting time.
However, building detailed geometric maps, can be difficult without accurate depth perception.
Several prior approaches have investigated learning-based exploration policies. Some approaches use simulation data (training in virtual environments).
Others learn from real-world data directly. These models may use:

- Intrinsic rewards: Encouraging the robot to explore new things.
- Semantic prediction: Going to interesting or informative places.
- Latent variable models: Abstract models of how actions affect the world.

Yet, policies trained in simulation frequently struggle to transfer to real-world environments. Even real-world-trained models can underperform in complex indoor and outdoor settings.
Enter NoMaD : A New Method

The work most closely related to NoMaD is ViNT (refer Appendix X for more details), which combines a goal-conditioned policy with a separate subgoal proposal module. The subgoal proposals are generated using an image diffusion model, condiitioned on robot's current view.
NoMaD improves on this by:

- Not generating images.
- Directly predicting actions using diffusion models, which are typically used in image generation tasks but can model complex probabilities really well.

- This makes NoMaD more accurate and much lighter (needs 15x fewer parameters).

One of the core challenges in modeling robot exploration policies is the inherently multimodal nature of action sequences.
Observation-conditioned diffusion models have emerged as powerful tools because they can learn complex action distributions without needing without needing explicit state prediction.
Nomad builds upon this adding **goal conditioning** to diffusion-based action generation, meaning it is capable of both:

- Goal-directed exploration
- Undirected exploration

B Technical Preliminaries

The primary objective is to develop a visual navigation policy, denoted by π, that enables a robot to navigate using only RGB images from its onboard camera.
The policy devised should operates as follows:

- It receives a sequence of past and current observations: $o_t := o_{t-P:t}$.
- It predicts a distribution over future actions: $a_t := a_{t:t+H}$.
- Optionally, it can also condition on a goal image o_g, representing the desired destination.

Depending on whether a goal is provided, the policy behaves differently:

- **Goal-directed navigation:** When a goal image o_g is available, π generates actions that guide the robot toward the goal.
- **Exploratory behavior:** When no goal is given (as in pure exploration settings), π must still generate safe and purposeful actions—avoiding obstacles and staying on traversable paths—while efficiently covering the environment.

To handle long-horizon planning and complex environments, the system is further augmented with:

- A topological memory graph $\mathcal{M}$, which maintains a structured map of past visual observations.
- A high-level planner that leverages this memory to decide on intermediate goals and broader exploration strategies.

Visual Goal-Conditioned Policies: ViNT as the Backbone

NoMaD builds on the ViNT (Visual Navigation Transformer) architecture, a Transformer-based model tailored for goal-conditioned navigation.

Key Components of ViNT:

- **Visual Encoding:** Each observation is processed using an EfficientNet-B0 encoder to extract feature embeddings.
- **Goal Fusion:** The current and goal image features are combined using a goal fusion encoder.
- **Transformer Attention:** These fused features (tokens) are passed through a Transformer model to generate a context vector c_t.
- **Predictions:** The context vector is used to predict:
  - A distribution over future actions: $a_t = f_a(c_t)$.
  - An estimate of temporal distance to the goal: $d(o_t, o_g) = f_d(c_t)$.

These outputs are learned via supervised training, where the model is shown expert trajectories and learns to imitate them.

> *However, ViNT is inherently goal-conditioned—it cannot operate in the absence of a goal image, limiting its ability to explore autonomously.*