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`