CS285 Assignment 1 Report
Imitation Learning for Push-T

XiaoLei Chu
SID: 3038525739

February 11, 2026

1 Overview

This report summarizes my Homework 1 experiments for CS285 (Push-T imitation learning) using:

- an MSE policy for action chunk prediction;
- a flow matching policy with Euler integration at inference time.

Both runs used the provided training and evaluation pipeline, and both include full WandB logs/videos/checkpoints in the submission package.

2 Experimental Setup

Shared training configuration (both runs):

- seed: 42
- chunk size: 8
- batch size: 128
- optimizer: Adam, learning rate 3×10^{-4}, weight decay 0
- hidden dimensions: $(256, 256, 256, 256)$
- epochs: 400
- eval interval: every 10,000 training steps
- flow denoising (Euler) steps: 10

Curve data source: The train loss curves are generated from the exported WandB CSV file `wandb_export_2026-02-11T18_37_24.546-08_00.csv`. The eval reward curves are taken from each run's `log.csv`.

MLP architecture used for MSE policy:

- input: normalized state (5-D);
- 4 fully connected hidden layers, each with 256 units;
- activation: ReLU after each hidden layer;
- output: flattened action chunk of size $8 \times 2 = 16$, reshaped to $(8, 2)$.

3 MSE Policy Results

3.1 Training Curves

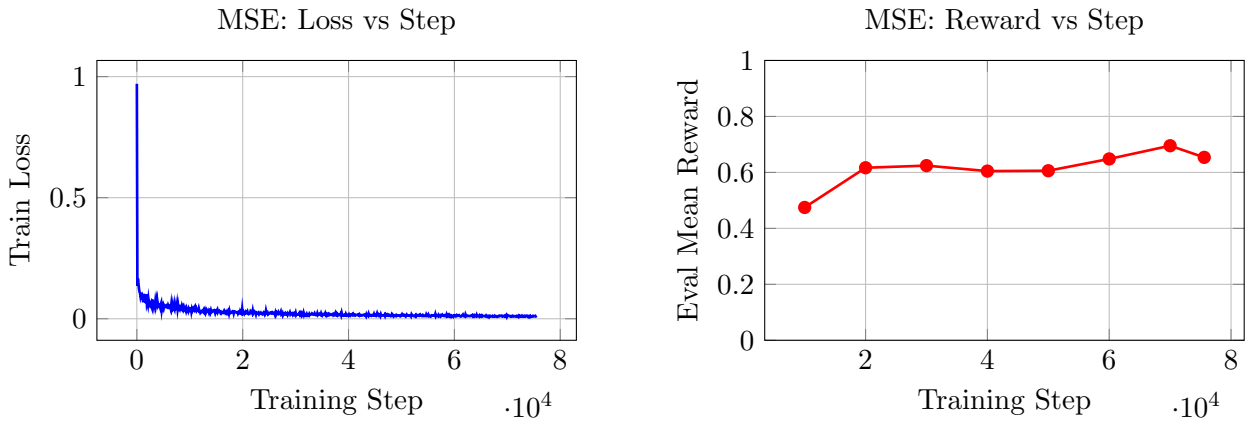

Figure 1: Training curves for the best MSE run.

Key metrics (MSE):

- best eval mean reward: 0.6953 at step 70,000;
- final eval mean reward: 0.6537 at step 75,600.

This meets the homework success threshold (≥ 0.5).

4 Flow Matching Policy Results

4.1 Training Curves

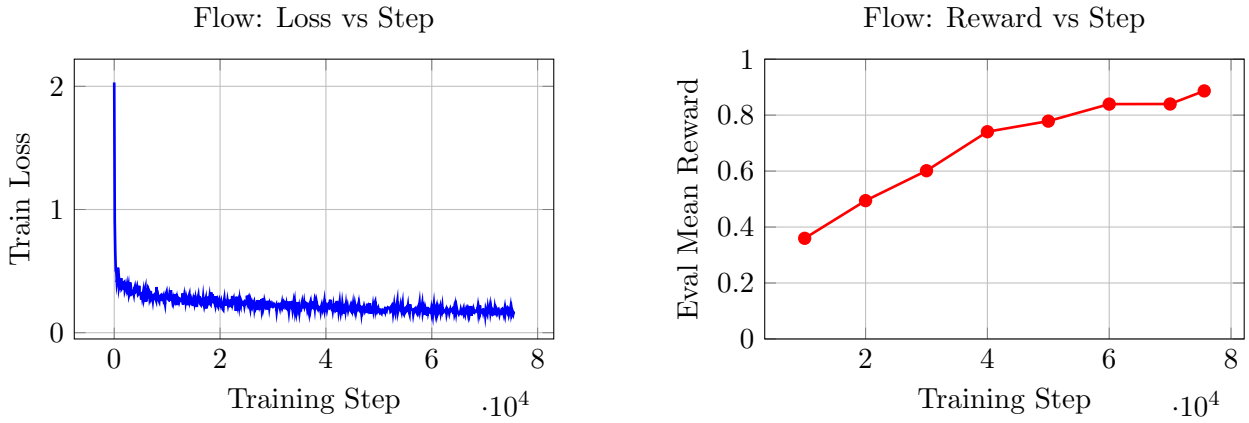

Figure 2: Training curves for the best flow matching run.

Key metrics (Flow):

- best eval mean reward: 0.8865 at step 75,600;

- final eval mean reward: 0.8865 at step 75,600.

This exceeds the homework success threshold (≥ 0.7).

5 Qualitative Comparison (from Evaluation Videos)

Figure 3: MSE rollout (episode 3): early, middle, and late frames.

Figure 4: Flow rollout (episode 3): early, middle, and late frames.

Behavioral observations:

- The MSE policy generally solves many cases but is less consistent at the end of trajectory execution.
- The flow policy produces smoother and more reliable progress toward goal completion across the trajectory.
- Consistent with these observations, flow achieves substantially higher final and peak rewards than MSE.

6 Conclusion

Both policies were implemented and trained successfully. The MSE policy reaches solid performance above the required threshold, while flow matching provides a clear improvement in both quantitative reward and qualitative rollout reliability.