COMSW4733 Computational Aspects of Robotics Project Proposal

Hua Hsuan Liang, Jaisel Singh and Edward Zhang

***Abstract*—Recent advances in generative modeling such as flow matching model can sample from the unknown target distribution [2]. In reinforcement learning, models can improve itself by maximizing objective function, such as Proximal Policy Optimization (PPO)[4] has established strong baselines in continuous control, but its performance often degrades in high-dimensional or deformable object domains. Flow Matching Policy Gradients proposed Flow Policy Optimization (FPO) [3] extends flow-based generative models into reinforcement learning, offering improved sample efficiency and expressive policy representations. Complementarily, DaXBench[1] provides a standardized platform for benchmarking deformable object manipulation, enabling systematic evaluation of policy learning algorithms in more complex, realistic environments. Together, these works set the stage for a new research direction: integrating flow-based generative policies with robust optimization frameworks to address the challenges of dynamic, non-rigid tasks. This integration not only promises gains in stability and learning efficiency but also opens opportunities for imitation learning and diffusion-based extensions in robotics and beyond.**

I. Problem Statement & Objectives

Deformable object manipulation is a long-standing challenge in robotics, with applications ranging from cloth folding and rope tying to surgical assistance. Unlike rigid-body control, deformable objects exhibit high-dimensional, nonlinear, and partially observable dynamics that are difficult to capture with standard reinforcement learning (RL) policies such as Gaussian-distributed multilayer perceptrons (MLPs). Recent benchmarks like DaXBench highlight the gap between existing policy learning algorithms and the requirements of real-world deformable manipulation.

Our project addresses this gap by exploring flow-based policy representations within reinforcement learning. Specifically, we aim to integrate Conditional Flow Matching (CFM) with policy gradient optimization to leverage the expressivity of generative flows while maintaining the stability of on-policy RL.

We define the following objectives for this project:

- Implement and train FPO for deformable manipulation tasks, ensuring actions remain valid and learnable within continuous action spaces.
- Evaluate whether FPO improves sample efficiency and training stability compared to standard PPO with MLP policies.
- Compare Imitation learning(Flow matching model) and reinforcement learning(FPO) performance in the deformable manipulation tasks.

A. Training Algorithm

We adopt the FPO surrogate [3], which preserves the PPO clipped objective but replaces the likelihood ratio with a ratio computed from per-sample conditional flow-matching (CFM) losses. Let o_t denote the observation (state) at time t and $\hat{A}_t$ the advantage (e.g., GAE). The surrogate objective is

$$\max_{\theta} \; E\Big[\min\Big(\hat{r}_{\text{FPO}}(\theta)\,\hat{A}_t,\ \text{clip}(\hat{r}_{\text{FPO}}(\theta),\,1-\varepsilon_{\text{clip}},\,1+\varepsilon_{\text{clip}})\,\hat{A}_t\Big)\Big]\,, \tag{1}$$

where the FPO ratio is

$$\hat{r}_{\text{FPO}}(\theta) = \exp\Big(\hat{L}_{\text{CFM},\,\theta_{\text{old}}}(a_t;o_t) \;-\; \hat{L}_{\text{CFM},\,\theta}(a_t;o_t)\Big). \tag{2}$$

The per-sample CFM loss is estimated with Monte Carlo:

$$\hat{L}_{\text{CFM},\,\theta}(a_t;o_t) = \frac{1}{N_{\text{mc}}}\sum_{i=1}^{N_{\text{mc}}} \ell_\theta(\tau_i,\epsilon_i), \tag{3}$$

with

$$\ell_\theta(\tau,\epsilon) \;=\; \tfrac{1}{2}\,\big\|\hat{v}_\theta(a_t^\tau,\tau;o_t) - (a_t-\epsilon)\big\|_2^2, \tag{4}$$

and the partially noised action

$$a_t^\tau = \alpha_\tau\, a_t + \sigma_\tau\,\epsilon, \qquad \tau \sim \mathcal{U}(0,1),\;\; \epsilon \sim \mathcal{N}(0,I). \tag{5}$$

We share the same (τ_i,ϵ_i) draws when evaluating $\hat{L}_{\text{CFM},\,\theta_{\text{old}}}$ and $\hat{L}_{\text{CFM},\,\theta}$ for variance reduction. Intuitively, decreasing the CFM/denoising loss for an action increases its ELBO and thus its effective likelihood under the flow policy; plugging the ratio in (2) into the clipped surrogate (1) yields a PPO-style, advantage-weighted update without computing exact log-likelihoods.

B. Milestones

- Weeks 1–2: Literature review, environment setup.
- Weeks 3–6: Implement and test flow-based policy on rigid benchmarks.
- Weeks 7–8: Extend experiments to DaXBench deformable tasks.
- Weeks 9–10: Comparative study against PPO and diffusion baselines.
- Weeks 11–12: Results analysis, report, and presentation preparation.

II. Proposed Methods and Experiments

We propose to investigate the integration of Conditional Flow Matching (CFM)([2]) with policy gradient methods in reinforcement learning, focusing particularly on challenging domains involving deformable objects. While prior work termed Flow Matching Policy Gradient([3]) has demonstrated the potential of flow matching models in rigid-body control tasks, their applicability to more complex and dynamic environments remains underexplored. Flow Matching Policy Gradient([3]) leverages the representational power of conditional flow matching as a generative policy model, combined with the robust optimization framework of Proximal Policy Optimization (PPO)([4]). This combination is expected to provide two key advantages: (1) improved sample efficiency through structured policy learning, and (2) enhanced expressivity in capturing the intricate dynamics of deformable-object interactions. To evaluate this method, we will using a set of experiments (openai-gym and DaXBench([1])) comparing FMPG against baseline policies (e.g., MLP-based PPO([4]) and diffusion policies) across both rigid and deformable manipulation tasks. In particular, deformable object manipulation poses unique challenges such as high-dimensional state spaces, non-linear dynamics, and partial observability, making it an ideal testbed for assessing the strengths of flow-based policy learning.

III. Expected Outcomes & Evaluation

Deliverables:

- Open-source implementation of FMPG integrated with PPO-style training.
- Experimental results comparing flow-based policies to baselines (MLP-based PPO and diffusion policy).
- Analysis of sample efficiency, stability, and policy multimodality.
- Recorded rollouts of trained policies side-by-side with baseline results.

Prospective Tasks:

- **Pendulum-v1 (OpenAI Gym):** Learn to swing up and balance an underactuated pendulum.
- **Reacher-v2:** Control a 2-DoF arm to reach a target position in the plane.
- **Cloth Folding:** Fold a square piece of cloth to a target configuration.
- **Rope Straightening:** Manipulate a rope to match a straight-line target configuration.
- **Rope Shape Matching:** Arrange a rope to trace a target shape (e.g., 'U' or 'S' shape) - tests precision.

Evaluation Metrics:

- Average episode reward and success rate on DaXBench tasks.
- Sample efficiency measured by reward per environment step.
- Training stability measured by variance across random seeds.

Risks & Mitigation:

- *Risk:* High computational cost of training flow models. *Mitigation:* Use lightweight flow architectures and reduce task complexity during early experiments.
- *Risk:* Instability in early training phases. *Mitigation:* Gradually anneal noise schedule and use variance-reduced advantage estimates.
- *Risk:* In deformable manipulation, rewards may be sparse or delayed, which can hurt flow model convergence. *Mitigation:* Use reward shaping and semi-sparse manual labeling.

IV. Team Member Introduction

- **Hua Hsuan Liang (hl3811):**2nd Master in CS department.
  I have experience training real world RL fine-tuning models for finger-gaiting task, which is now under review by ICRA. I am currently doing research at Matei's ROAM lab at Columbia, and I am trying to combine the denoise style policy and the reinforcement learning to make the powerful model can explore the task and get a better performance.
- **Jaisel Singh (js6897):** First Year's Masters in Mechanical Engineering.
  I have experience working with RL training for manipulation tasks within MuJoCo where I am currently working on a PPO based training pipeline for the fine-tuning of a a foundation video model to be used as a dynamic model for MPC on a ur5 robot within the ROAM lab. I have also explored trajecotry optimization based policy training for various tasks within Gym and am interested in developing scalable algorithms for perception-guided manipulation that aim to maintain performance despite model uncertainty and environmental variability.
- **Edward Zhang (cz2874):** 3rd-year undergrad in CS. Researcher at ROAM Lab, working on dexterous teleoperation device for data collection.

References

[1] Siwei Chen, Cunjun Yu, Yiqing Xu, Linfeng Li, Xiao Ma, Zhongwen Xu, and David Hsu. Daxbench: Benchmarking deformable object manipulation with differentiable physics. *arXiv preprint arXiv:2210.13066*, 2022.

[2] Thibault Lipman, Jiaming Song, and Stefano Ermon. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[3] David McAllister, Songwei Ge, Brent Yi, Chung Min Kim, Ethan Weber, Hongsuk Choi, Haiwen Feng, and Angjoo Kanazawa. Flow matching policy gradients. In *arXiv preprint arXiv:2507.21053*, 2025.

[4] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. In *arXiv preprint arXiv:1707.06347*, 2017.