Policy Distillation from a Model-Based Expert for Non-Prehensile Manipulation

Denis Shcherba and Marc Toussaint

I. Introduction

While significant advancements have been made in prehensile pick-and-place operations, achieving human-level dexterity requires further advancements in non-prehensile manipulation. Strategies such as pushing, sliding, or toppling remain formidable challenges, as they require the system to reason about complex contact dynamics and frictional forces without the stability of a firm grasp.

Existing methods encompass multiple planning and learning paradigms. Classical optimization-based planners provide smooth paths but depend on privileged state information (exact poses and geometries) rarely available in the real world. Conversely, Imitation Learning (IL) from human demonstrations enables reactive sensorimotor control from raw sensor observations, yet requires extensive teleoperated data collection and considerable hardware cost.

This work bridges this gap through Policy Distillation from a Model-Based Expert. We leverage an optimization-based planner in simulation to generate an abundance of optimal trajectories for complex non-prehensile tasks. This privileged knowledge is then distilled into a sensorimotor student policy, allowing for a systematic comparison between different observation modalities and policy architectures. Finally, we investigate the sim-to-real transition and validate our approach through physical robot experiments.

II. Background

A. K-Order Markov Optimization (KOMO)

To solve the constrained motion planning problem, we utilize k-order Markov Path Optimization (KOMO) [1], which transcribes continuous trajectories into a structured nonlinear program (NLP). By assuming k-order Markov dynamics, the state at step t depends only on the previous k configurations. The objective with constrains are solved using Augmented Lagrangian method:

$$\mathcal{L}(x,\lambda,\mu) = \sum_{t=0}^{T} f(x_t) + \lambda^\top h(x) + \mu^\top g(x) + \rho\Phi(h,g) \quad (1)$$

where $f(x_t)$ typically penalizes squared accelerations $\ddot{x}(t)^2$ for smoothness. The resulting sparse, block-banded Hessian enables efficient second-order Newton updates.

III. Related Work

A. Modern Imitation Learning

Recent literature has seen the emergence of several expressive Imitation Learning paradigms. ACT [2] employs a Conditional VAE (CVAE) with a transformer-based encoder-decoder to predict sequences, or "chunks" of future actions. Diffusion Policy [3] represents the policy as a Conditional Denoising Probabilistic Model (DDPM). This paradigm excels at capturing multi-modal action distributions and high-dimensional trajectories. While these state-of-the-art models are highly expressive, they are fundamentally designed for learning from expensive, human-collected demonstrations rather than self-generated expert plans.

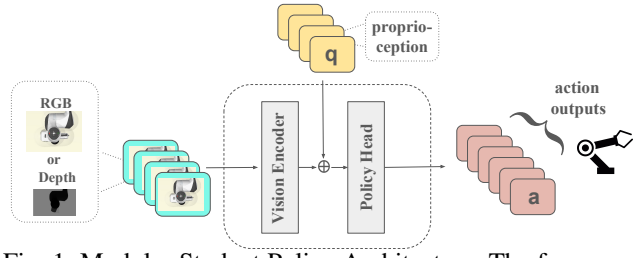

Fig. 1: Modular Student Policy Architecture. The framework utilizes distilled expert knowledge to process raw visual input in real-time.

B. Distilling a model-based expert

Recent work has explored distilling model-based planners into neural policies. [4] utilize Transformers to distill Task and Motion Planning (TAMP) experts, though they focus on high-level logical sequencing rather than fine-grained geometric constraints. Recently, [5] introduced SPIN, which employs a sampling-based planner (Skill-RRT) and RL-trained "connectors" to bridge state gaps between primitive skills before distilling them into a Diffusion Policy. While SPIN uses sampling and RL-glue to handle long-horizon sequences, our work leverages optimization-based planning to generate the precise, contact-rich trajectories required for complex non-prehensile tasks.

IV. Methodology

This work adopts the *Learning by Cheating* paradigm [6], a form of policy distillation [7]. We decompose our pipeline into two stages, where we (*i*) utilize KOMO programs to generate optimal trajectories using the full state space s_t, and (*ii*) distill these into a student policy π_θ trained to mimic the expert using only raw sensor observations o_t. This allows for distilling global geometric reasoning into a local, reactive sensorimotor policy capable of real-time execution. In our evaluation, we compare policy and vision encoder

TABLE I: Comparative analysis of policy architectures across task domains evaluated on 500 rollouts.

Backbone	Policy Head	Params (M)	Success ↑ (%)
Tabletop–Push Domain			
ResNet18 (RGB)	Transformer	14.6	85.2
ResNet18 (Depth)	Transformer	14.6	**88**
DINOv2 (Patch)	Transformer	4.5	63.6
ResNet18 (RGB)	Diffusion (DDPM)	58.3	83.8
ResNet18 (Depth)	Diffusion (DDPM)	58.3	63.6
DINOv2 (Patch)	Diffusion (DDPM)	48.2	43.3
Shelf–Retrieval Domain			
ResNet18 (RGB)	Transformer	12.1	**92.6**
ResNet18 (Depth)	Transformer	12.1	80.2
DINOv2 (Patch)	Transformer	4.5	84.2
ResNet18 (RGB)	Diffusion (DDPM)	58.3	62.8
ResNet18 (Depth)	Diffusion (DDPM)	58.3	30
DINOv2 (Patch)	Diffusion (DDPM)	48.2	54.6

performance, evaluate sim-to-real transfer from purely simulated data, and decouple failure modes by training on a real observation dataset, alongside multiple further ablations.

A. Task Domains and Expert Data Generation

We evaluate our approach on two contact-rich tasks: shelf-retrieval, requiring narrow-hooking extraction of a cuboid using a hook, and cylinder-pushing, requiring precise sliding of a puck to a target. Expert trajectories are generated using KOMO. We produce a dataset of 1,000 trajectories by executing these plans via position control and recording time-aligned, multi-modal sensor data (RGB, depth, and proprioception) at a uniform sampling frequency.

B. Student Policy Architecture

The student policy (Fig. 1) is designed to be modality-agnostic and modular and consists of (*i*) a *vision encoder* where a sequence of m observations are processed into a sequence of feature embeddings and (*ii*) a *policy head* which fuses the vision embedding with a history of m proprioceptive states to predict a chunk of n actions. We compare a ResNet18 [8] trained from scratch against a frozen DINOv2 backbone [9], augmented with a spatial softmax for geometry-aware feature extraction. We evaluate two policy heads: a transformer-encoder-based model and a DDPM based policy.

V. Results and Discussion

Evaluation across both domains (Table I) reveal that the transformer-based policies yield the highest success rates. We attribute this superior performance over the Diffusion Policy to the unimodal, deterministic nature of the optimization-based expert; the transformer's attention mechanism more effectively captures the precise temporal dependencies inherent in these singular, optimal action sequences.

We evaluated sim-to-real transfer for the transformer-based policy using two strategies. First, we trained for both vision backbones trained on a domain-randomized and augmented dateset from simulation; however, simulation-only training proved insufficient to capture the geometric precision required for real-world tasks. To isolate observation shifts, we instead executed the expert planner directly on the physical hardware to collect a small, a real-world dataset (≈250 samples). This approach, supplemented by data augmentation, successfully bridged the gap; physical experiments (50 trials) demonstrated success rates comparable to simulation, with minor degredation due to environmental noise II.

(a) Shelf–Retrieval (b) Tabletop–Pushing

Fig. 2: The physical experimental setup for both domains.

TABLE II: Success rates for policies trained on real-world observation data.

Task	Trials	Success Rate
Push Cylinder	50	68%
Shelf Withdrawal	50	40%

VI. Limitations and Future Research

a) Unimodality and Distribution Shift: The student policy is inherently unimodal due to the optimization-based expert. As the planner lacks recovery behaviors or sub-optimal variations, the policy is susceptible to distribution shift when minor execution errors occur. Future work could incorporate NLP-sampling [10] or hybrid planning approaches [11] to diversify the training distribution, thereby yielding more robust policies with higher task success rates.

b) Expert Engineering and Scalability: The current expert relies on manually crafted constraints and cost weights, creating a significant bottleneck for scaling to novel geometries. Future research will investigate Automated Expert Synthesis, leveraging language models and program search [12] to autonomously generate KOMO programs.

c) Hierarchical Control: Predicting dense, fixed-length action chunks at a high frequency presents heavy computational and policy learning challenges. We aim to pivot toward hierarchical control paradigms, where policies predict high-level keyframes or waypoints [13][14] rather than high-frequency raw actions. Coupling these sparse targets with low-level motion primitives could increase performance and robustness in long-horizon, contact-rich manipulation.

ACKNOWLEDGMENT

This research was funded by the Amazon Fulfillment Technologies and Robotics team. This work has been supported by the German Federal Ministry of Research, Technology and Space (BMFTR) under the Robotics Institute Germany (RIG).

REFERENCES

[1] M. Toussaint, "Newton methods for k-order markov constrained motion problems," 2014.

[2] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," 2023.

[3] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," 2024.

[4] M. Dalal, A. Mandlekar, C. Garrett, A. Handa, R. Salakhutdinov, and D. Fox, "Imitating task and motion planning with visuomotor transformers," 2023.

[5] H. Jung, D. Lee, H. Park, J. Kim, and B. Kim, "SPIN: distilling Skill-RRT for long-horizon prehensile and non-prehensile manipulation," 2025.

[6] D. Chen, B. Zhou, V. Koltun, and P. Krähenbühl, "Learning by cheating," 2019.

[7] A. A. Rusu, S. G. Colmenarejo, C. Gulcehre, G. Desjardins, J. Kirkpatrick, R. Pascanu, V. Mnih, K. Kavukcuoglu, and R. Hadsell, "Policy distillation," in *International Conference on Learning Representations (ICLR)*, 2016.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[9] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "Dinov2: Learning robust visual features without supervision," 2024.

[10] M. Toussaint, C. V. Braun, and J. Ortiz-Haro, "Nlp sampling: Combining mcmc and nlp methods for diverse constrained sampling," 2024.

[11] V. N. Hartmann, A. Orthey, D. Driess, O. S. Oguz, and M. Toussaint, "Long-horizon multi-robot rearrangement planning for construction assembly," *IEEE Transactions on Robotics*, vol. 39, p. 239–252, Feb. 2023.

[12] D. Shcherba, E. Cobo-Briesewitz, C. V. Braun, and M. Toussaint, "Meta-optimization and program search using language models for task and motion planning," 2025.

[13] B. Akgun, M. Cakmak, K. Jiang, and A. L. Thomaz, "Keyframe-based learning from demonstration," *International Journal of Social Robotics*, vol. 4, pp. 343–355, 2012.

[14] L. X. Shi, A. Sharma, T. Z. Zhao, and C. Finn, "Waypoint-based imitation learning for robotic manipulation," in *Conference on Robot Learning*, 2023.