Learning Contact-Rich Manipulation Skills Using Multisensory Inputs

Ryan Diaz (diaz0329@umn.edu)
Mason Hawver (hawve005@umn.edu)
Hanchen Cui (cui00191@umn.edu)

Introduction:

Multisensory inputs incorporating RGB images, depth images, and force-torque (F/T) data, can be effective for contact-rich robotic manipulation tasks. We leverage visuotactile pretraining to learn strong latent observation representations from which contact-rich manipulation policies can be learned.

Input/Output:

Pretraining Input: Masked RGB, depth, F/T
Pretraining Output: Unmasked reconstructions

Finetuning Input: Unmasked RGB and depth
Finetuning Output: 7-dim EEF actions (translation, rotation, gripper)

Framework:

During pretraining, we mask a large portion of the input (RGB, depth, force-torque) and reconstruct the masked parts, with loss computed only on the masked regions. In the finetuning phase, we perform policy training using solely RGB and depth input information. Inspired by MultiMAE [1].

Dataset Information:

Using MimicGen [2], we generate a rollouts of several tasks. During each rollout rgb, depth, and force-torque data is recorded and saved to a dataset. Part of this dataset will be used to train the encoder. The other part will be used for finetuning.

Evaluation:

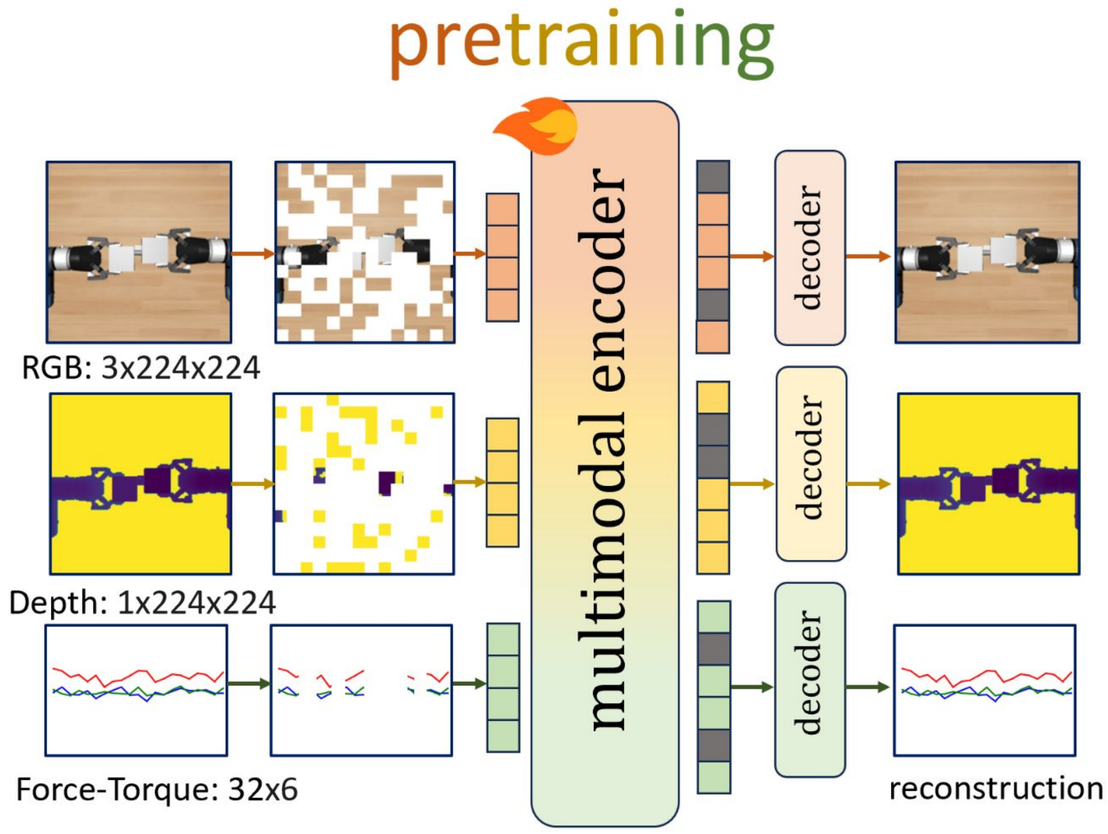
We evaluate our pipeline by measuring success rates in 20 policy rollouts for a peg-in-hole task and 3 tasks from MimicGen (not present in the pretraining dataset). We compare against two baselines: training policies from scratch and pretraining only on RGB and depth data.

References:

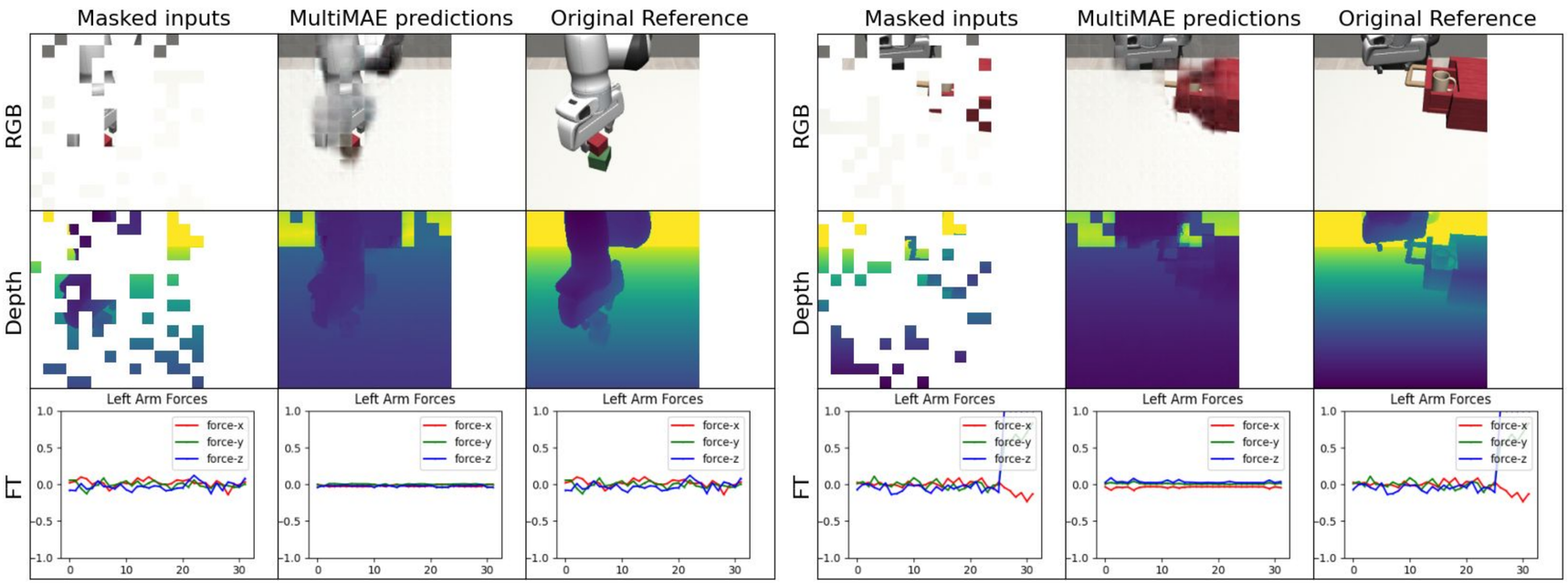
[1] Bachmann, Roman, et al. "Multimae: Multi-modal multi-task masked autoencoders." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.
[2] Mandlekar, Ajay, et al. "Mimicgen: A data generation system for scalable robot learning using human demonstrations." arXiv preprint arXiv:2310.17596 (2023).

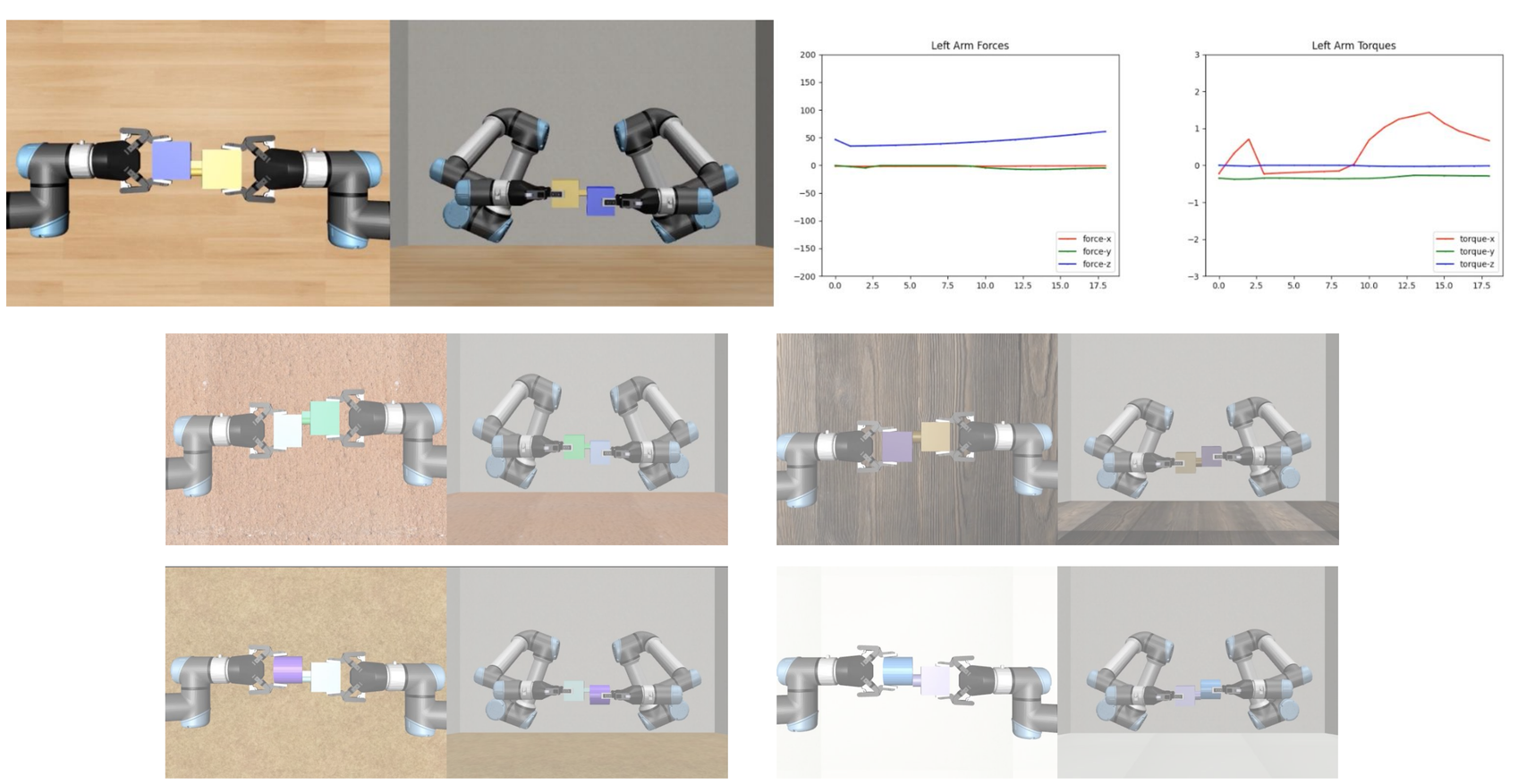
computer science & engineering
MINNESOTA ROBOTICS INSTITUTE
UNIVERSITY OF MINNESOTA

*We can **improve contact-rich** robotic manipulation policies using **multisensory pretraining***



$$\mathcal{L}_{Total} = \alpha\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\mathbf{x}_i^{RGB} - \hat{\mathbf{x}}_i^{RGB}\right\|^2\right) + \beta\left(\frac{1}{M}\sum_{i=1}^{M}\left\|\mathbf{x}_i^{Depth} - \hat{\mathbf{x}}_i^{Depth}\right\|^2\right) + \gamma\left(\frac{1}{K}\sum_{i=1}^{K}\left\|\mathbf{x}_i^{FT} - \hat{\mathbf{x}}_i^{FT}\right\|_1\right) \qquad \mathcal{L}_{policy} = \frac{1}{N}\sum_{i=1}^{N}\left\|\mathbf{a}_i - \hat{\mathbf{a}}_i\right\|^2$$





Model	From Scratch	RGBD Pretrain	RGBD+FT Pretrain
Success Rate (20 rollouts)	0%	35%	50%