Efficient and Robust VLA Models for Real-Time Robotic Manipulation: Week 9 Update

Shriyadita Medavaram

Faculty Mentor: Dr. Yuzhang Shang

Student Mentor: Lam Nguyen

Project Recap

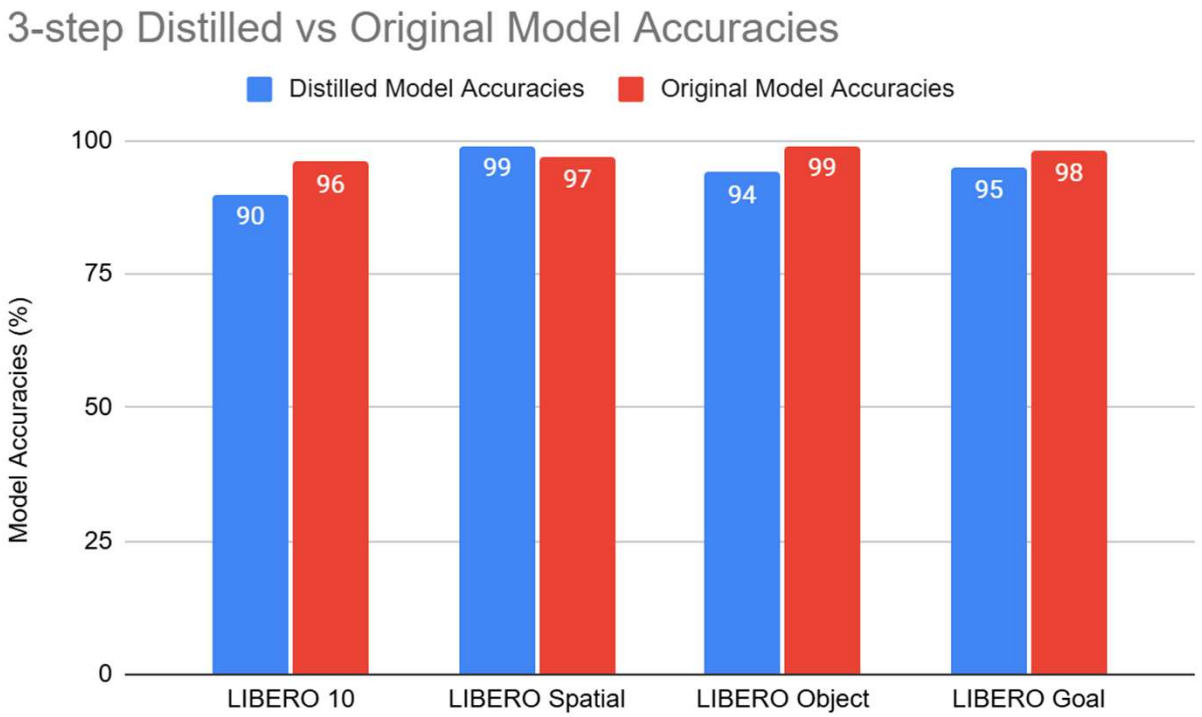
- Goal: Compress Pi0.5 to make smaller faster VLA models
- Two parts
  - Run Pi0.5 model on robot arm
  - Compress Pi0.5 LIBERO checkpoint



Timestep Distillation

- Used models defined by previous week, and did timestep distillation instead
  - This was our initial goal, we just did it slightly differently by distilling the entire thing till now since it takes much longer to train
- Kept overall $\pi 0.5$ architecture the same, but distilled 10-step diffusion model into 3-step diffusion model

LIBERO Evaluations



Latency and Memory Comparison

Metric	Original π0.5	3-step Distilled π0.5	Improvement
Full inference time	11.81 s	5.69 s	2.08x faster
Total CPU delta	672.73 MB	324.68 MB	2.07x lower
Total GPU delta	14 MB	8 MB	1.75x lower

Next Steps

- Finish poster and report
- Attempt 1 step distillation if time permits
- Attempt use onto robot arm if time permits