Master Thesis / Research Internship

Fine-Tuning a Robot Foundation Model for Contact-rich Manipulation

Key words: VLA; Foundation Model; Contact-rich Manipulation

Background

Recent progress in Vision-Language-Action (VLA) foundation models has opened up new opportunities in robotics, enabling generalization across tasks and environments []. While these models demonstrate impressive performance in vision-guided control, the role of tactile sensing—a crucial modality for contact-rich manipulation—remains underexplored. Only a limited number of studies have investigated how tactile information can be effectively integrated into such models [2-4], leaving a gap in leveraging this modality for improved robot robustness and dexterity.

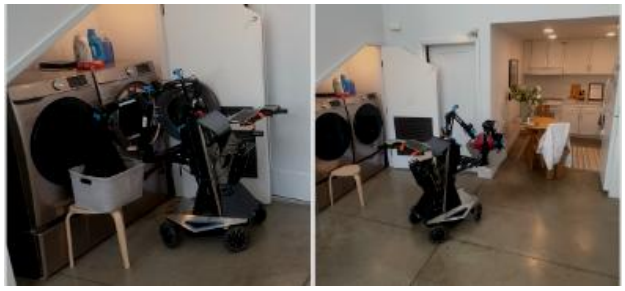
This thesis will investigate whether there exists a simpler yet effective way to enhance the performance of VLA foundation models on contact-rich manipulation tasks, by exploring fine-tuning a pre-trained strategy.

Your Tasks

1. Reproduce the SOTA open-source model.
2. Fine-tune the model on the contact-rich manipulation tasks.
3. Contribute towards a publication-oriented study. (The concrete details of the research direction will be discussed in person for confidentiality reasons.)

Requirement

- High self-motivation and interest in publication work.
- Background in deep learning.
- Interest in robotics and robotic manipulation.



Supervisor: Yansong Wu, Junan Li
yansong.wu@tum.de junnan.li@tum.de

TUM School of Computation, Information and Technology
Technische Universität München

Reference:

[1] Black K, Brown N, Driess D, et al. $\pi_0 $: A Vision-Language-Action Flow Model for General Robot Control[J]. arXiv preprint arXiv:2410.24164, 2024.
[2] Huang J, Wang S, Lin F, et al. Tactile-VLA: Unlocking Vision-Language-Action Model's Physical Knowledge for Tactile Generalization[J]. arXiv preprint arXiv:2507.09160, 2025.
[3] Yu J, Liu H, Yu Q, et al. ForceVLA: Enhancing VLA Models with a Force-aware MoE for Contact-rich Manipulation[J]. arXiv preprint arXiv:2505.22159, 2025.
[4] Cheng Z, Zhang Y, Zhang W, et al. OmniVTLA: Vision-Tactile-Language-Action Model with Semantic-Aligned Tactile Sensing[J]. arXiv preprint arXiv:2508.08706, 2025.