# Multimodal Robot Intelligence with Vision-Language-Action Models

**Faculty advisor: Prof. Chi-Guhn Lee**

Future robots will need to understand instructions, perceive complex scenes, and turn multimodal information into reliable physical actions. Recent progress in vision-language models and vision-language-action policies suggests a new generation of robots that can use language, images, demonstrations, and sensor feedback to perform useful tasks.

This project investigates multimodal robot learning, including how robots can interpret human instructions, use visual context, and generalize across objects, tasks, and environments. Students may explore vision-language-action models, multimodal prompting, imitation learning, data collection, evaluation, and real-robot testing. The project may involve the Trossen Stationary AI Robot, UR5 robot, and other real robots, alongside simulation-based experimentation.

### Preferred Foundation / Experience

- Solid background in deep learning and Python
- Experience with PyTorch and modern ML workflows
- Familiarity with vision-language models, transformers, or multimodal learning
- Interest in robotics, imitation learning, or language-guided manipulation
- Comfort working with datasets, experiments, and model evaluation
- Experience with robot simulation or real robot systems is a strong plus

### Why this project?

You'll gain experience at the intersection of foundation models and robotics, one of the most active areas in embodied AI. The project offers a strong portfolio opportunity for students interested in robot learning, multimodal AI, applied ML, or PhD-level research.

**Contact: Ramy ElMallah, ramy.elmallah@mail.utoronto.ca; Chongyu Zhu, chongyu.zhu@mail.utoronto.ca**