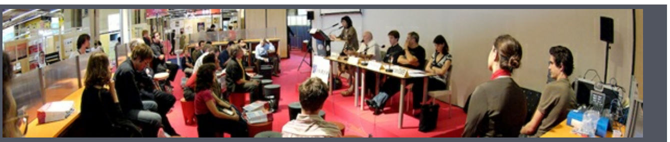
INVENTEURS DU MONDE NUMÉRIQUE

INTERNATIONAL RELATIONS



INTERNSHIP SUBJECT

2889 - Vision-Language-Action for humanoid robots

Vision-Language-Action (VLA) models are an emerging class of multi-modal large language models that enable robots to understand and execute complex verbal instructions [1,2,3]. These models integrate Vision-Language Models (VLMs), Large Language Models (LLMs), and imitation learning methods (often based on diffusion models) to combine language, vision, and action in a unified framework. The long-term goal is to develop generalist robots capable of performing virtually any task in any environment—ranging from factories to homes—based solely on natural language commands. For example, a user could instruct a robot to *“make breakfast with some eggs and coffee and bring it to me,”* or alternatively, *“put all these items in a box, close the package, and send it to a client,”* using the same robot and the same underlying policy.

The first objective of this internship is to evaluate open-source VLA systems (e.g., OpenVLA, Pi0) on real robotic platforms, starting with our Unitree G1 robot, integrating them with our whole-body stack, and potentially extending to our bimanual Tiago++ robot. These models will be fine-tuned using in-house data collected in the lab to improve performance in our specific scenarios.

The second objective is to enrich these models with force sensing data to improve the interactions with the environment.

Finally, the internship will explore any opportunity to push the state of the art in VLA-based robotics.

[1] Kim, Moo Jin, et al. 'OpenVLA: An open-source vision-language-action model.' arXiv preprint arXiv:2406.09246 (2024) - https://arxiv.org/pdf/2406.09246

[2] Black, Kevin, et al. '$\pi_0 $: A Vision-Language-Action Flow Model for General Robot Control.' arXiv preprint arXiv:2410.24164 (2024). - https://arxiv.org/pdf/2410.24164 - https://www.physicalintelligence.company/blog/pi0

[3] Pertsch, Karl, et al. 'Fast: Efficient action tokenization for vision-language-action models.' arXiv preprint arXiv:2501.09747 (2025) - https://arxiv.org/pdf/2501.09747

Required Skills

Required skills:

- good programming skills in Python
- software deployement: GNU/Linux, Docker (if possible)
- robotics : ROS1, ROS2

Languages: English (French is not required -- English is the main language of the team).

General Information

- **Research Theme :** Robotics and Smart environments
- **Locality :** Villers lès Nancy
- **Level :** Master
- **Period :** 5th January 2026 -> 30th June 2026 (6 months)

⚠ *These are approximative dates. Please contact the training supervisor to know the precise period.*

- **Deadline to apply :** 1st July 2025 (midnight)

Contacts

- **Training Supervisor :** Serena Ivaldi / serena.ivaldi@inria.fr
- **Second Training Supervisor :** DONOSO Clemente / clemente.donoso@inria.fr
- **Team Manager :** Francis Colas / Francis.Colas@inria.fr

More information

- **Inria Team :** LARSEN
- **Inria Center :** Centre Inria de l'Université de Lorraine