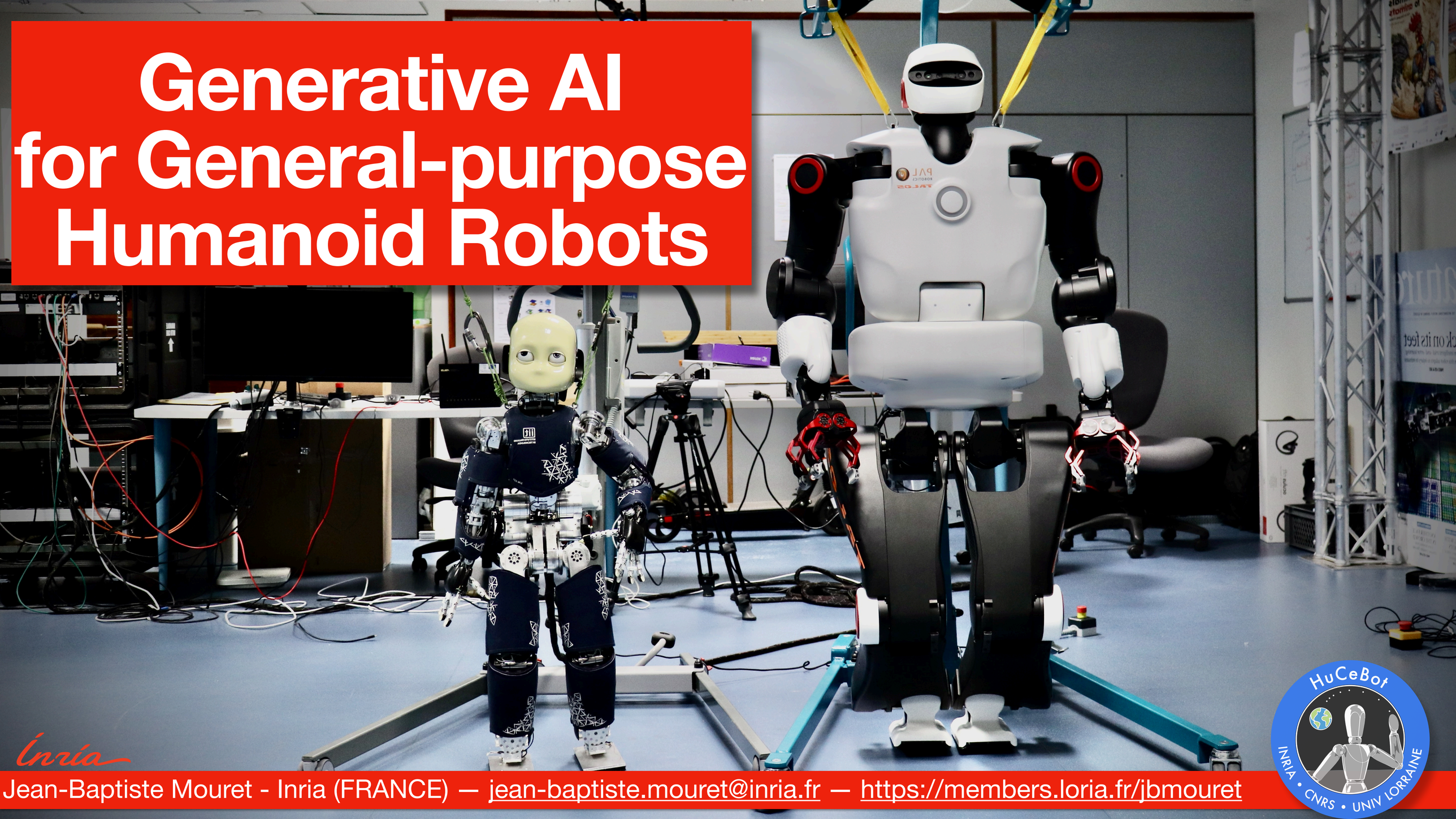
Generative AI for General-purpose Humanoid Robots



Jean-Baptiste Mouret - Inria (FRANCE) — jean-baptiste.mouret@inria.fr — https://members.loria.fr/jbmouret

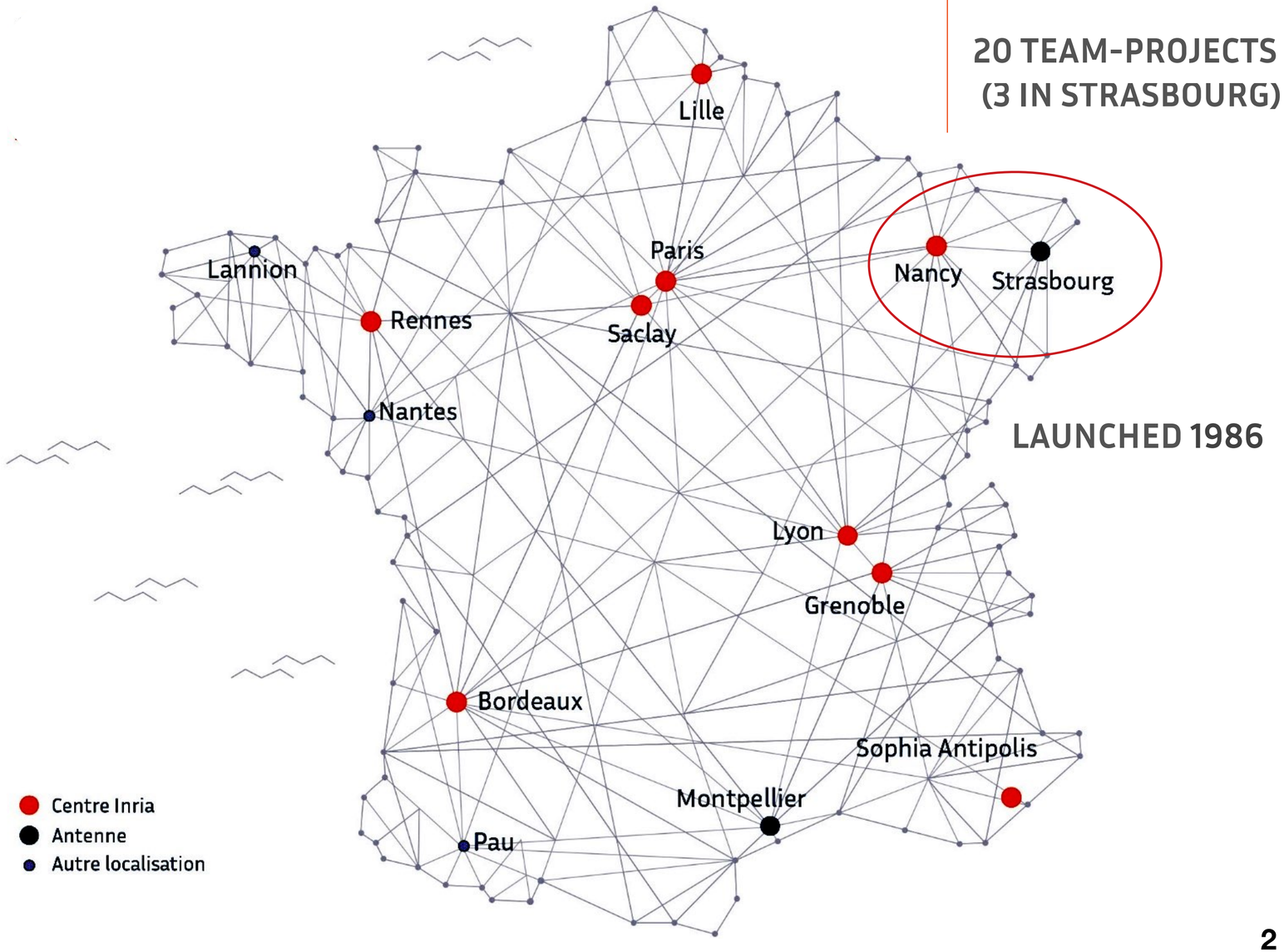
Jean-baptiste Mouret

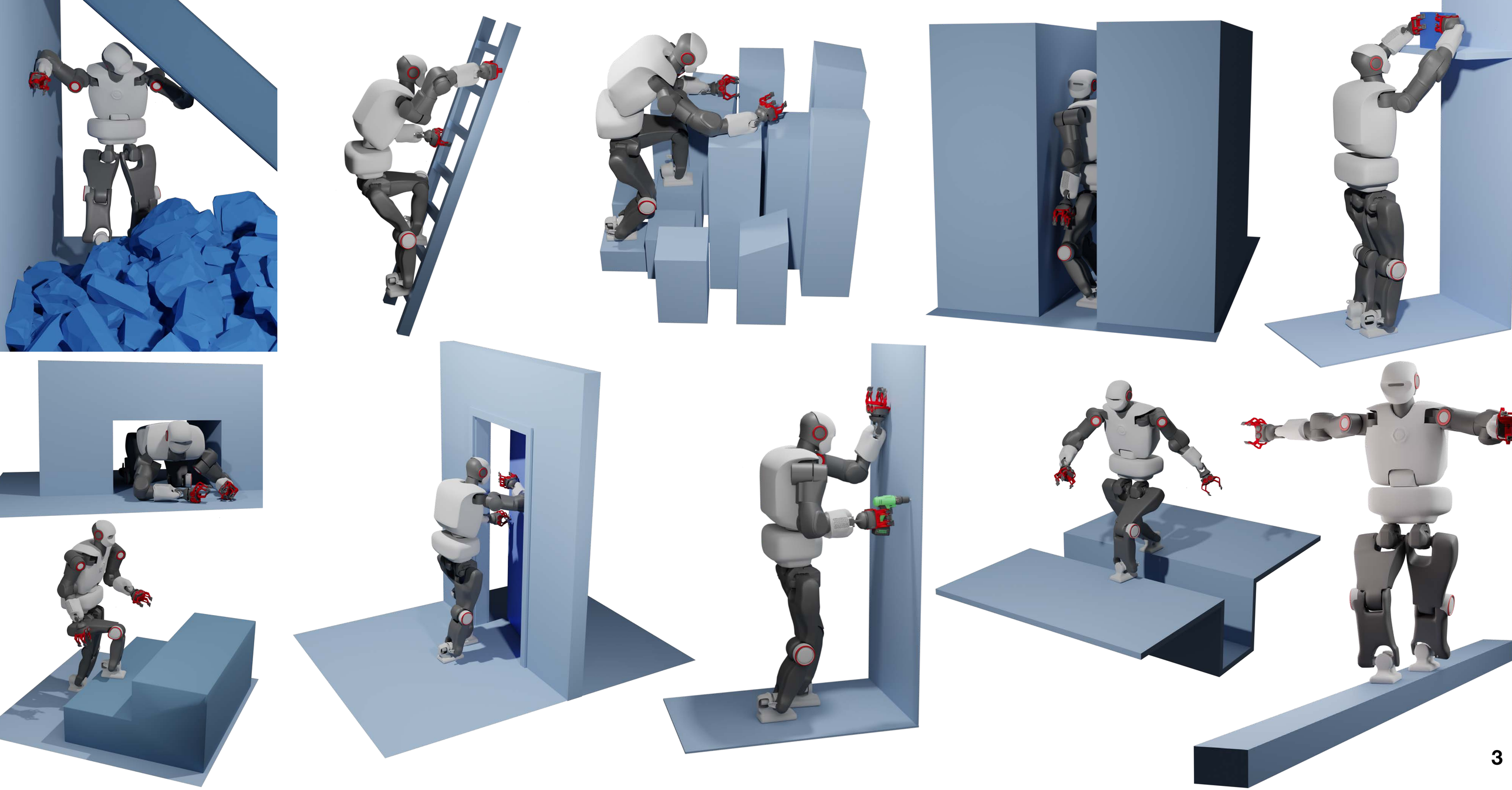
- "Head of science" / scientific director for the Inria center of University of Lorraine, Nancy, France

➔ 20 teams on CS & Maths speech, crypto, formal methods, ..

- Member of HuCeBot team — Human-Centered Robotics — headed by **Serena Ivaldi**

- **Main field:** leveraging AI in robotics (esp. humanoids)





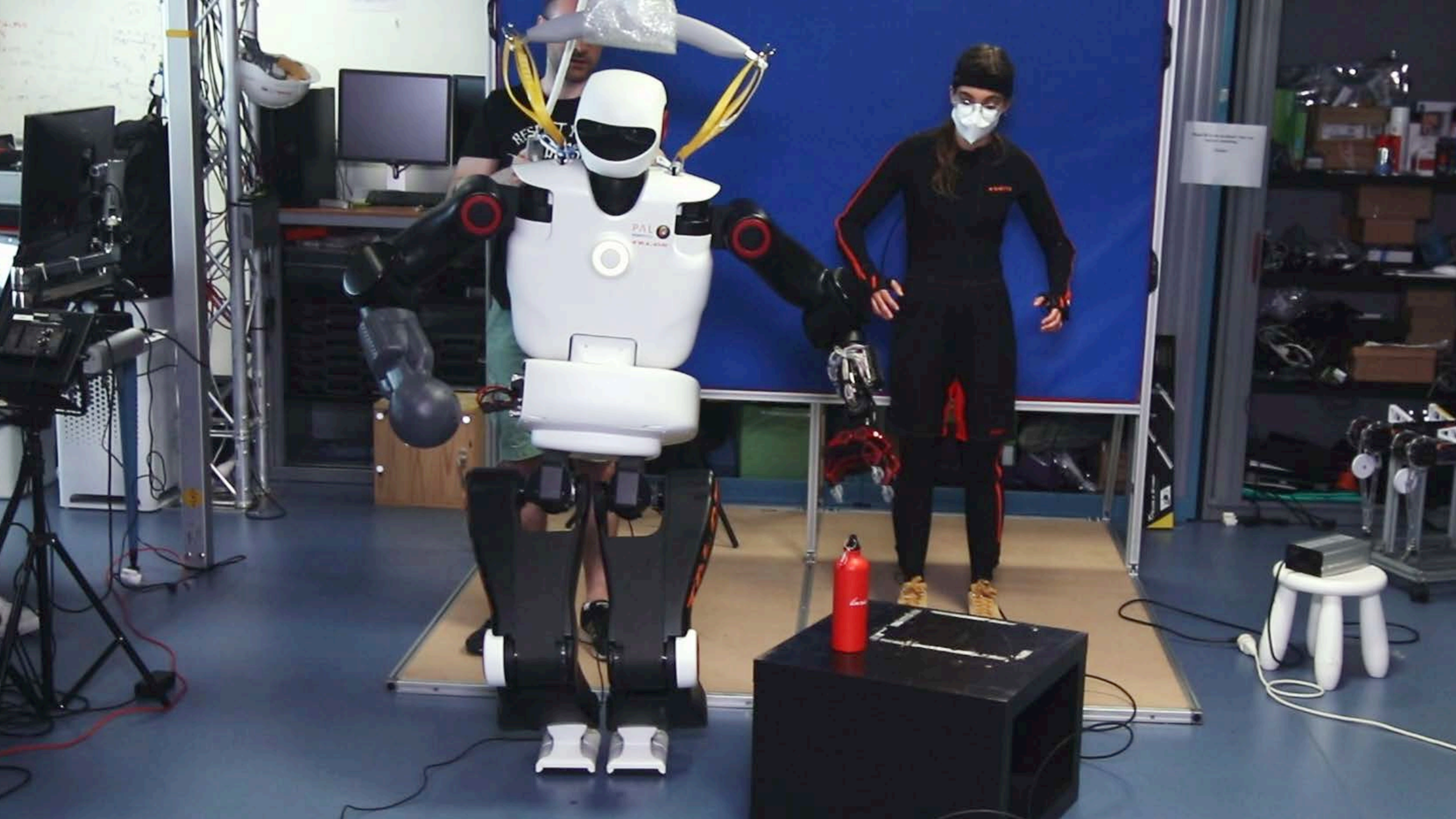
Humanoid robots are highly versatile machines

But humanoid robots are complex machines
➔ expensive, prone to failure, …

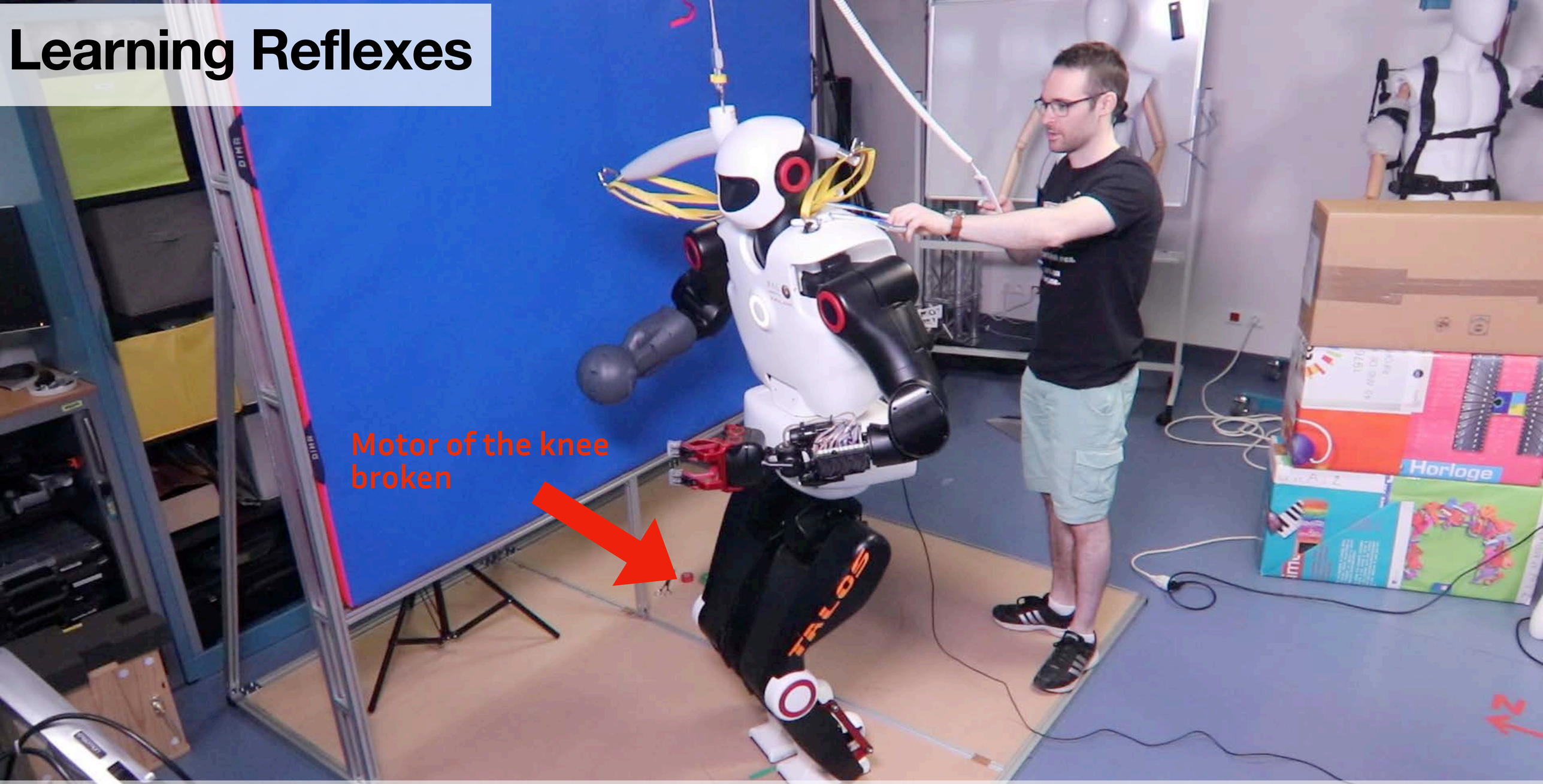
For now (and for a long time) humanoids for high-stakes tasks
➔ “improvisation”
➔ versatility is a requirement

High-stakes: Humans are in control
The wage of an operator is negligible
➔ (whole-body) **Teleoperation** (no fully autonomous robot)

In addition: good match between human morphology and humanoid robots



Learning Reflexes



What to do when a robot is damaged?

Anne T, Dalin E, Bergonzani I, Ivaldi S, Mouret JB. First do not fall: learning to exploit a wall with a damaged humanoid robot. Robotics and Automation Letters. 2022.


No compensation
Average round-trip delay 1s
Delayed
Without any delay compensation, it is very difficult to accomplish a given task

Large Language Models (LLMs) are <u>generalists</u>

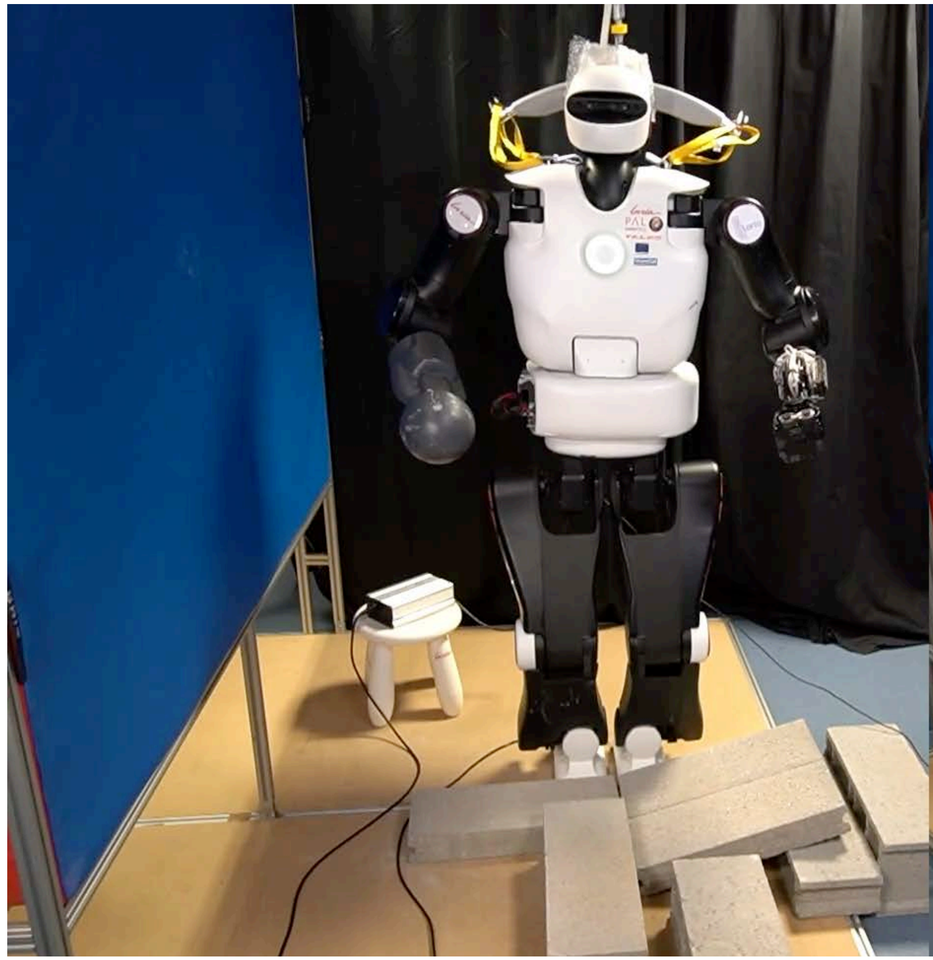
- not trained on (e.g.) cooking problem
- … but still gets results

LLMs have some “common sense” (cf frame problem)

LLMs benefit from the experience of “All the writers in the world”
… they have some “embodiment”
… their build own “model” of the world (emergent representation)

… this embodiment is from a human shape (experiences of humans)
➔ easier to embed in a humanoid robot

… a generalist model in a generalist robot
➔ (ideally) humanoid robot
➔ we still need good whole-body control



Rouxel et al. "Flow matching imitation learning for multi-support manipulation." *RA-L.* (2024).

good match between LLM & Humanoids

Key question: connecting language and action

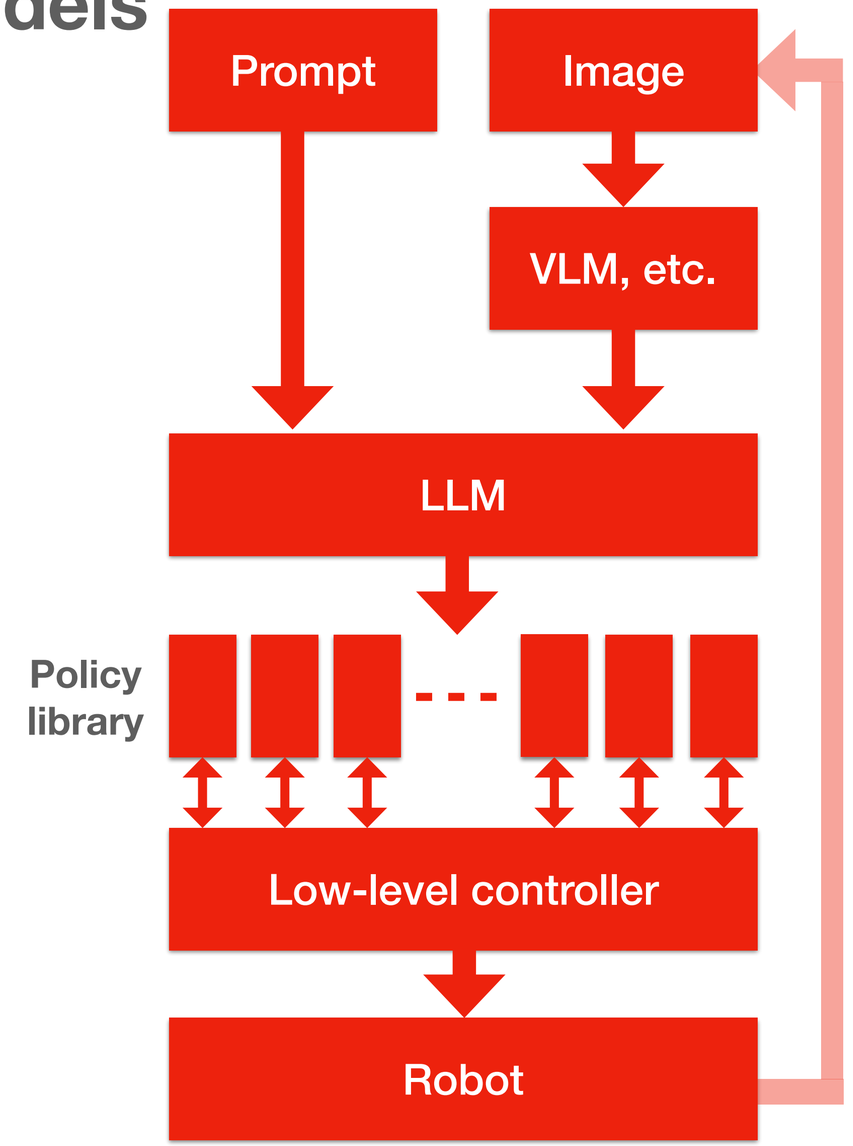
Approach 1: foundation models

➔ “Prompt engineering”:

1. describe the situation to the LLM with a VLM / vision
2. *prompt* what to do to achieve the goal
3. get a structured answer (code, JSON file, etc.)
4. activate pre-trained behaviors

Diffusion
(cf image generation)

How many policies do we need?



No need for (expensive) training: use pre-trained models
No need for training data (but still put examples in prompts!)
… but how far can we go with generalist models?


User: *Pick the bowl from the INRIA dishwasher and place it on the INRIA table.*

Selecting contacts with language

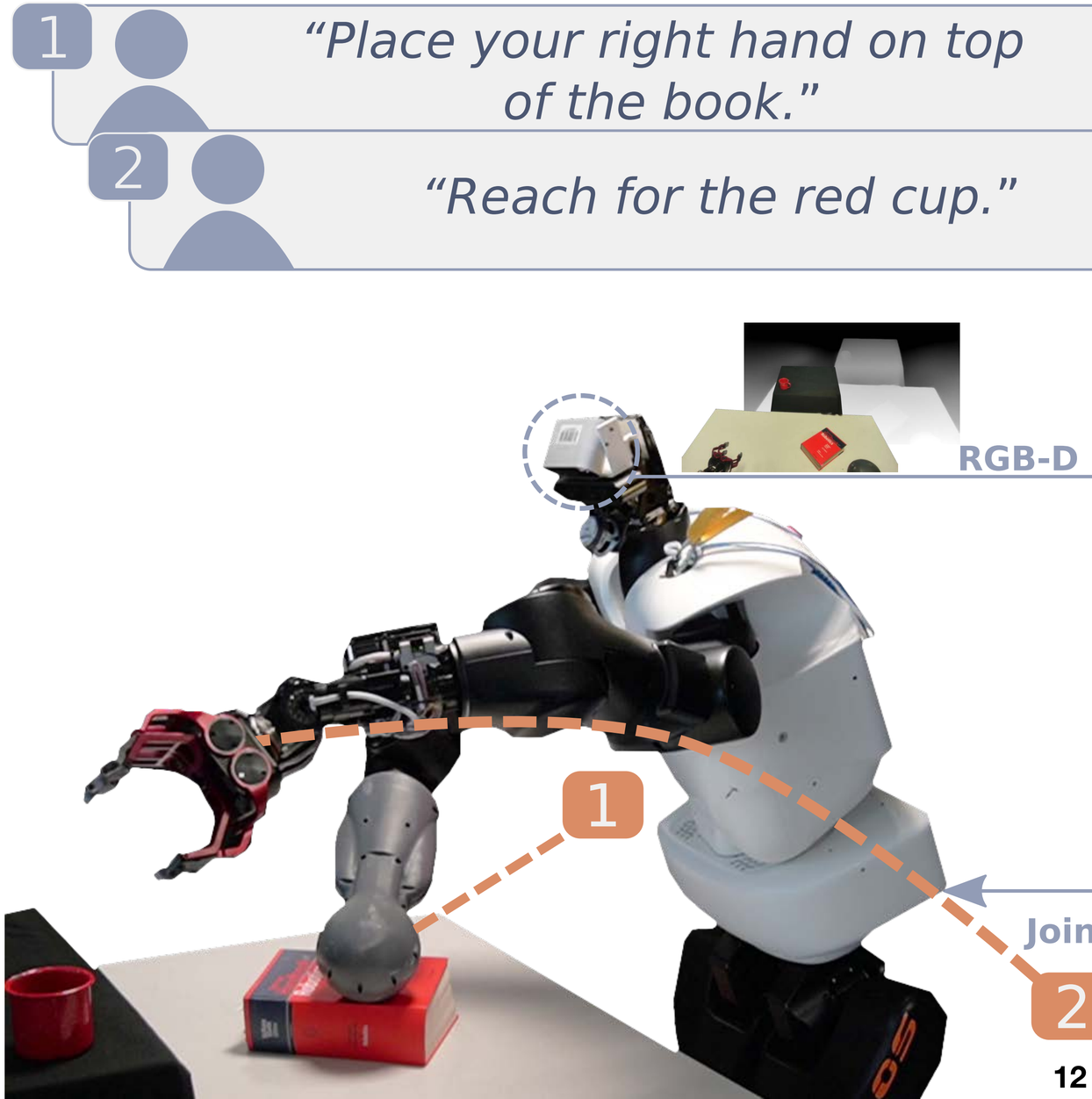
Challenges

Vision + language + actions

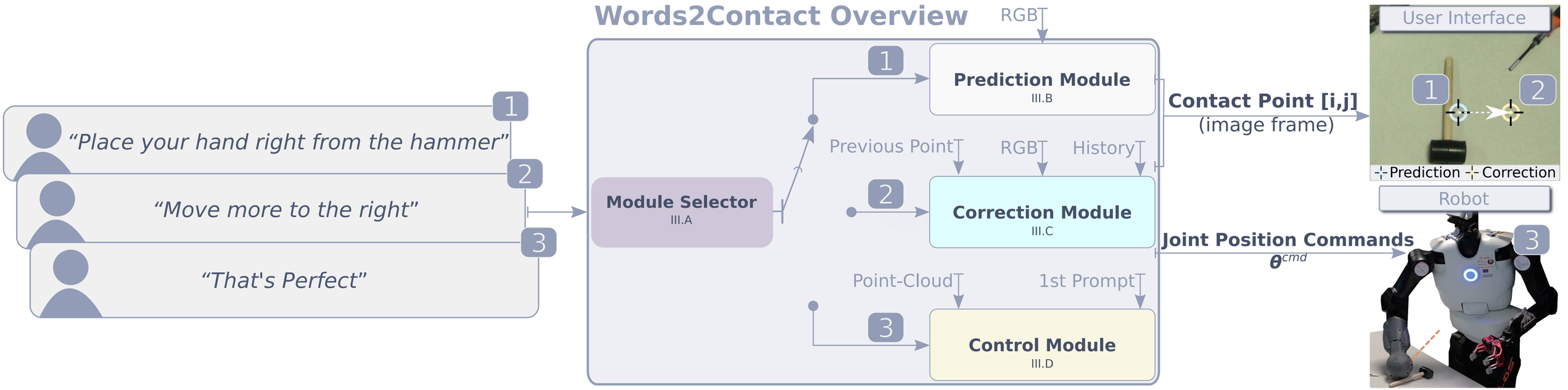
Synonyms, periphrases, etc.

Context-dependent

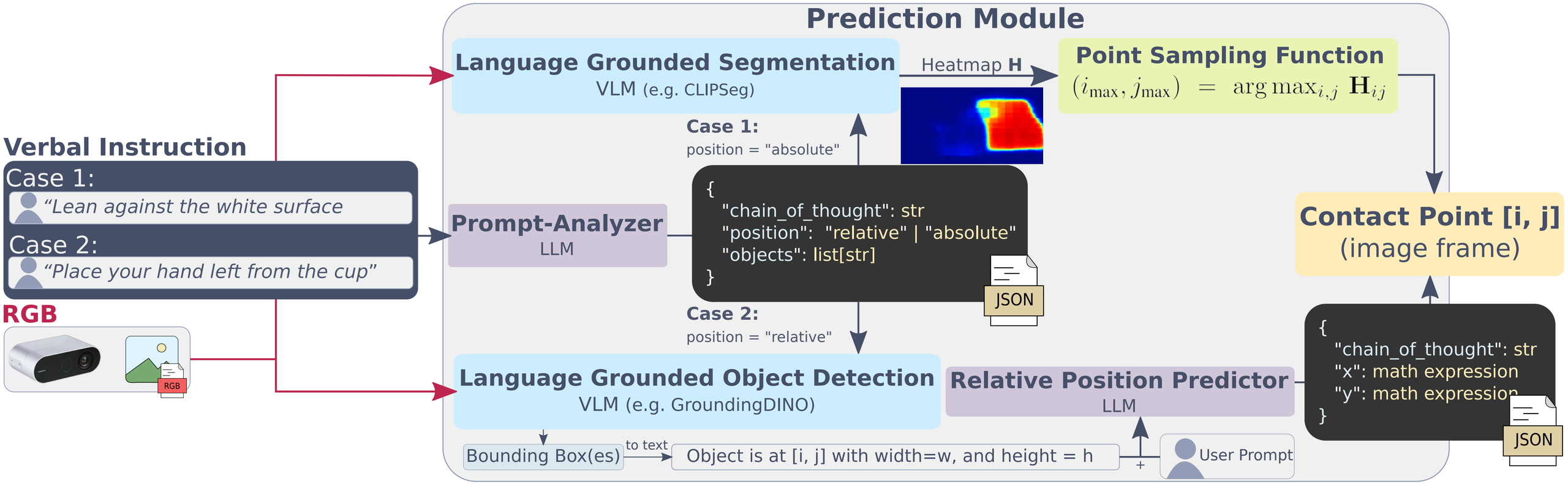
Many ways of saying the same thing



Totsila D, Rouxel Q, Mouret JB, Ivaldi S. (2024).
Words2Contact: Identifying Support Contacts from Verbal Instructions Using Foundation Models.
Proc. of IEEE Humanoids

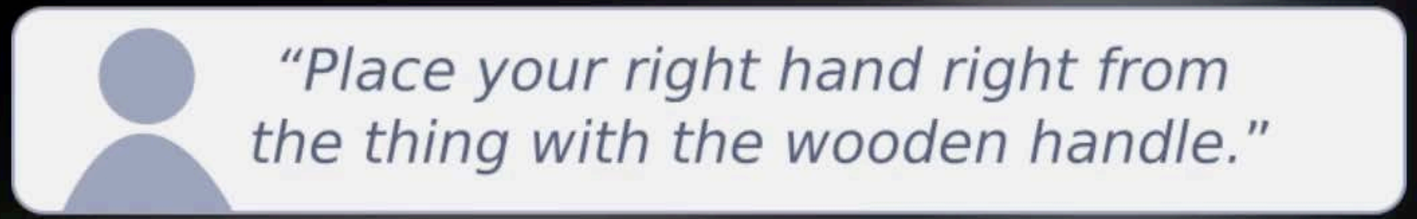


Totsila D, Rouxel Q, Mouret JB, Ivaldi S. (2024). Words2Contact: Identifying Support Contacts from Verbal Instructions Using Foundation Models.
Proc. of IEEE Humanoids



Totsila D, Rouxel Q, Mouret JB, Ivaldi S. (2024). Words2Contact: Identifying Support Contacts from Verbal Instructions Using Foundation Models.
Proc. of IEEE Humanoids

approach 1 — foundation models — no training



Totsila D, Rouxel Q, Mouret JB, Ivaldi S. (2024). Words2Contact: Identifying Support Contacts from Verbal Instructions Using Foundation Models.
Proc. of IEEE Humanoids

… but how to choose additional contacts?

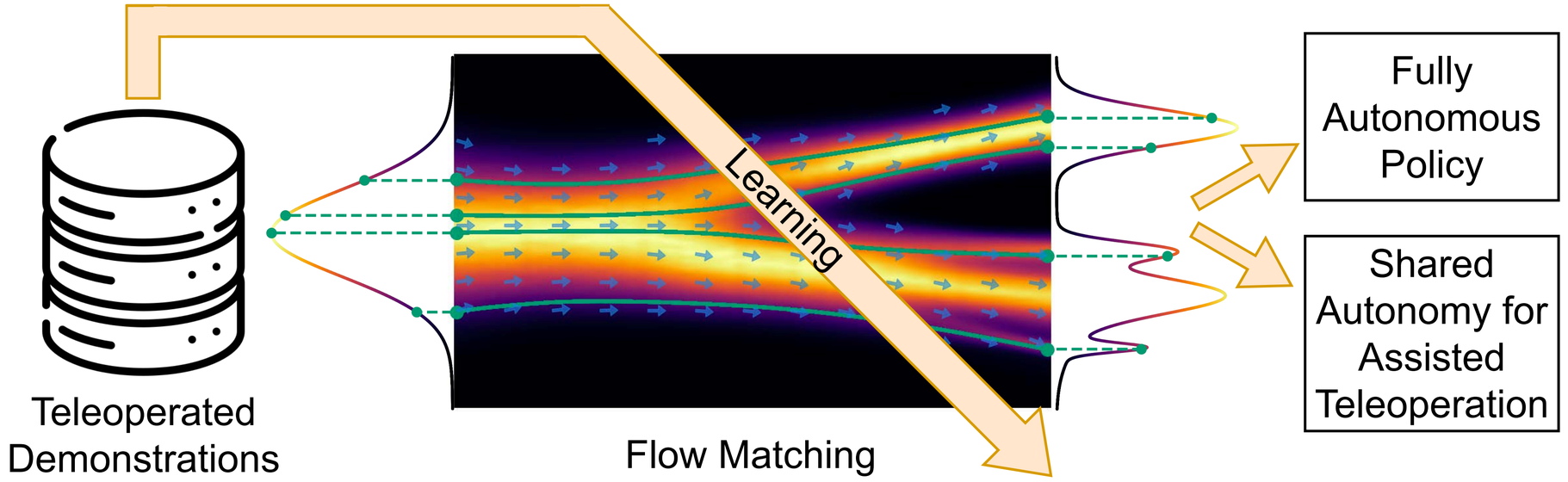
Concept: use imitation learning

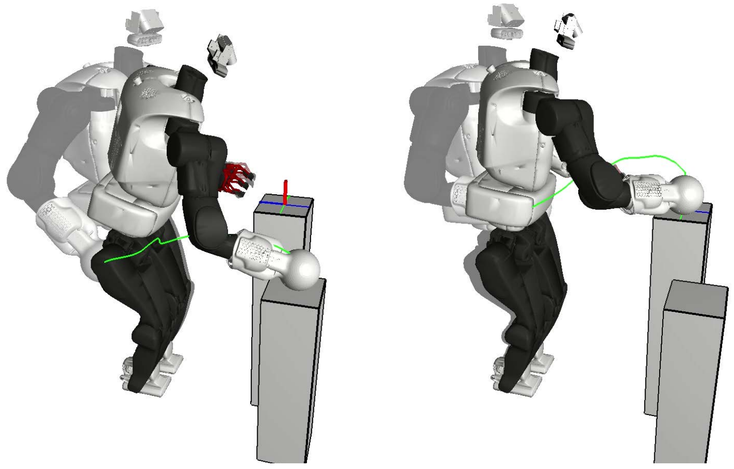
Example: the operator wants to reach with the right hand a bottle that is too far
… and the robot uses the left hand to add a contact

Why not automatic placement?

- difficult (many papers about this!)
- might make “common sense mistakes” (e.g., putting weight on a window)
- requires a very good perception / world model

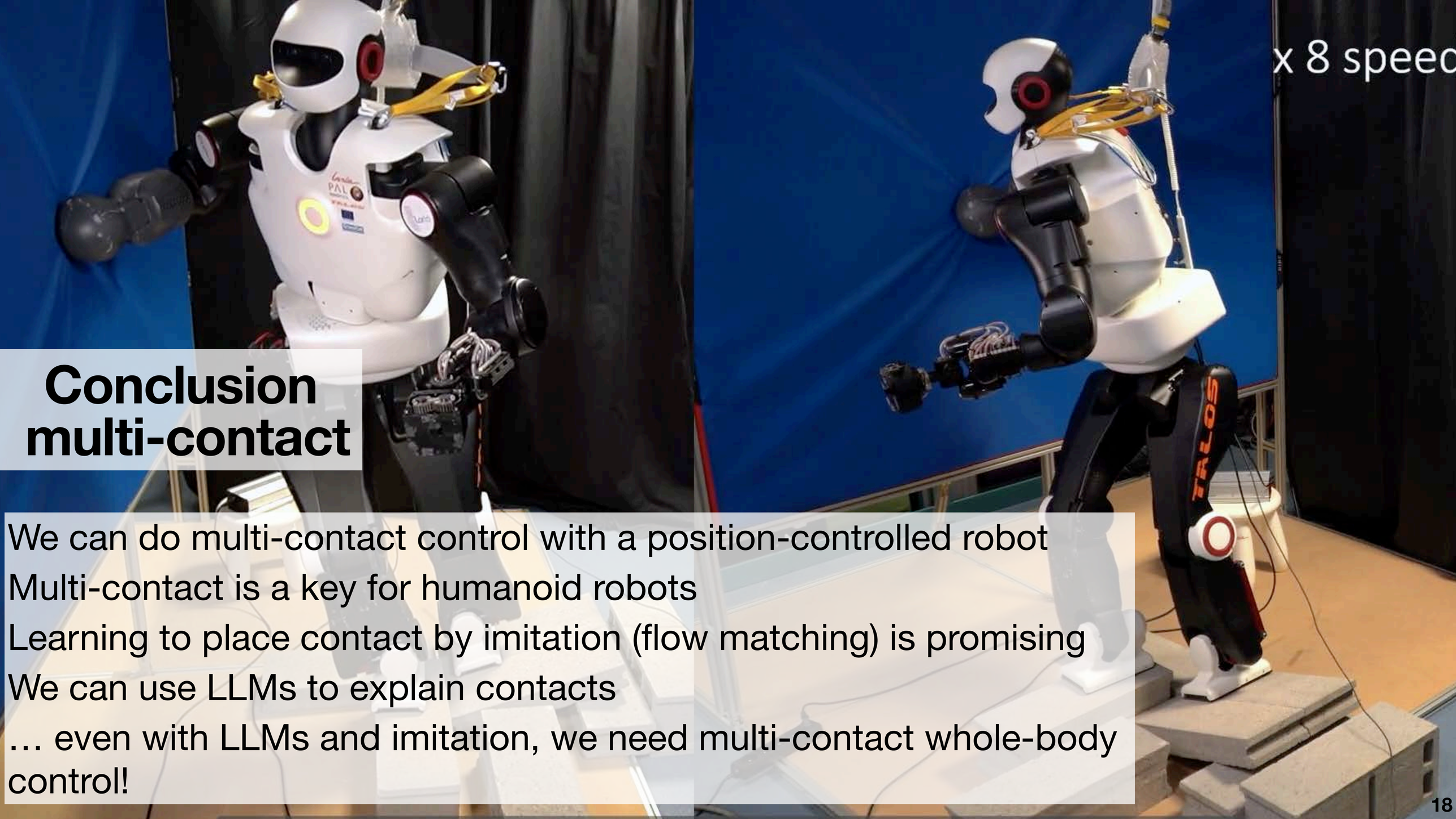
… and recent progresses in imitation learning (diffusion policies)





Method	Inference Time (ms)	In Distribution		Out of Distribution	
		Success Rate	Median±MAD Error (cm)	Success Rate	Median±MAD Error (cm)
Demonstrations	–	100%	1.3±0.6	–	–
Flow 20 steps	35±4	99%	1.4±0.5	78%	3.4±2.0
DDPM 100 steps	178±12	100%	1.5±0.5	69%	4.0±1.7
DDIM 20 steps	39±4	100%	1.4±0.5	67%	3.9±1.8
Supervised Learning	3±1	92%	4.1±1.4	52%	7.6±4.0

Rouxel Q, Ferrari A, Ivaldi S, Mouret JB. (2024). Flow matching imitation learning for multi-support manipulation
Proc. of IEEE Humanoids.



Conclusion multi-contact

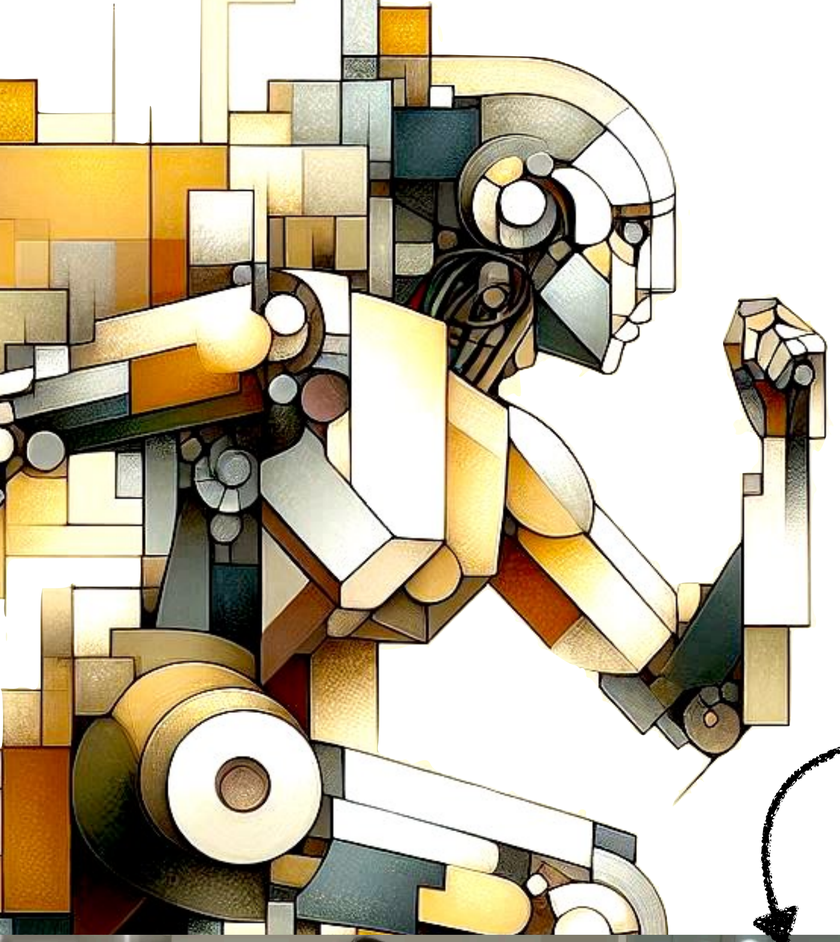
We can do multi-contact control with a position-controlled robot

Multi-contact is a key for humanoid robots

Learning to place contact by imitation (flow matching) is promising

We can use LLMs to explain contacts

… even with LLMs and imitation, we need multi-contact whole-body control!



The next questions

1. We are back to symbols and open-loop plans

➔ How can we blend (vs activate) motion with language?
➔ Continuous interaction / interruptions / etc.
➔ Other sensors (IMU, Force, skin, …)
➔ Physics consistency? Hallucinations?

2. Prompt engineering: a new programming language?

➔ automate? find techniques?

3. We need data with language (for training <u>and</u> evaluation)

- What data? with or without robot? Others?
- <u>Collective effort?</u> (all the labs unite, cf Wikipedia)
- Who pays for it? (storage, curation)
- Youtube videos? existing datasets? annotations?
- Will we have enough data?

4. We need pre-trained LLMs, VLMS, etc.

- train academic models? specialized for robotics?
- collective effort?
- How to reward something already done by others?

Conclusion

LLMs: a new era for robotics

robots that can understand verbal instructions (and speak)
robots with common sense
… combined with vision + voice (deep learning)
… but not (for now) smarter than humans!

A good match between humanoid robots and LLMs

generic models for generic tasks (vs specialized)
➔ *a generalist humanoid robot?*

We still need good whole-body control

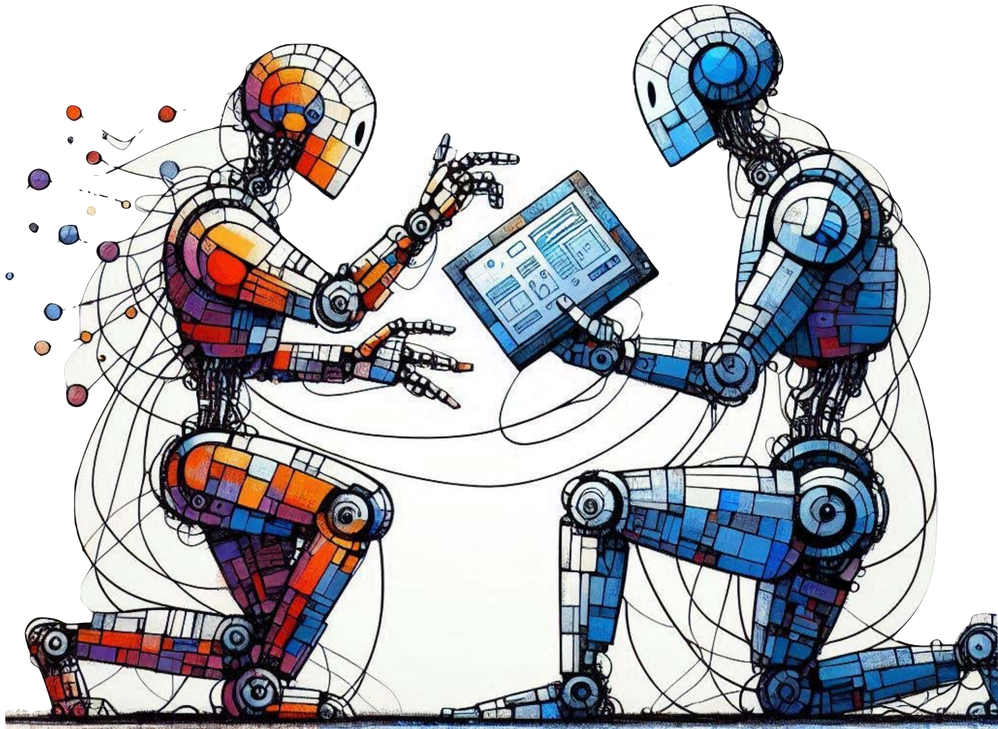
LLMs do not replace control

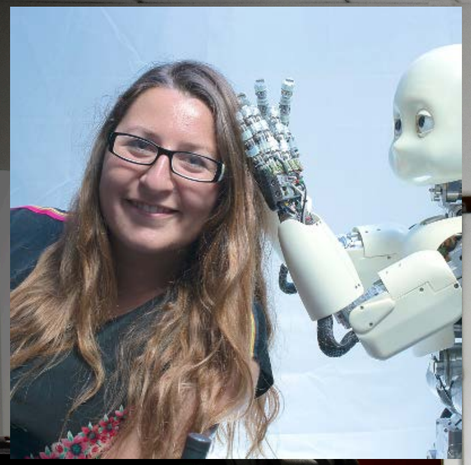
We need data (including for evaluation)

collective organization? simulation?

We need models

collective organization?





with Serena Ivaldi

QUESTIONS

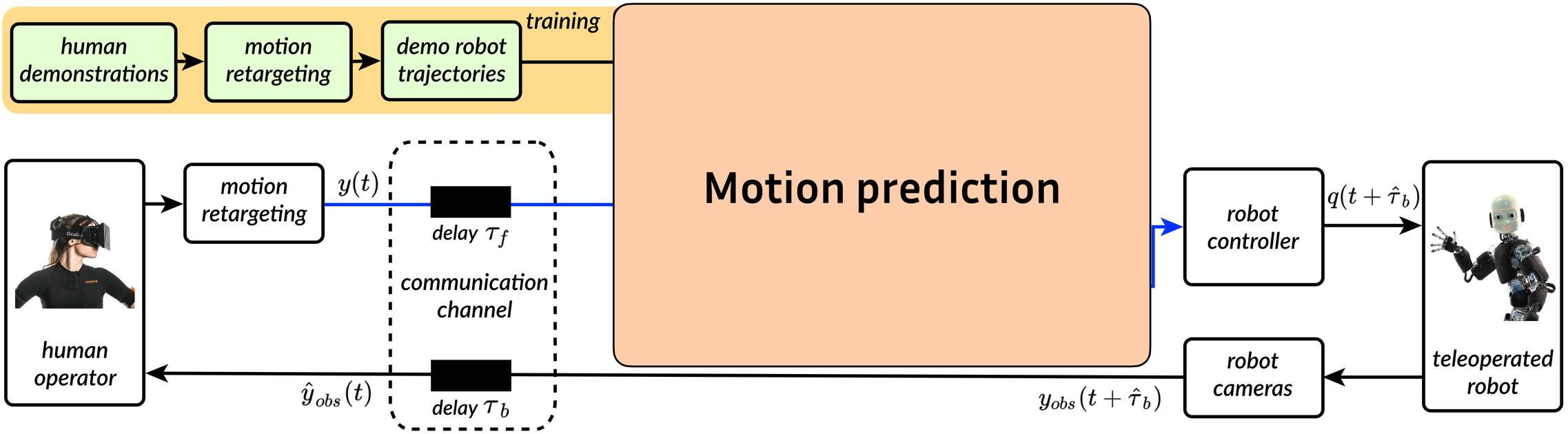
Jean-Baptiste Mouret - Inria (FRANCE) — jean-baptiste.mouret@inria.fr — https://members.loria.fr/jbmouret

Prescient teleoperation

- Learn predictors of the operator's motion with machine learning

- Predict the motion of the operator (and therefore of the robot)
- Execute the prediction of the future (taking into account back & forth delay)
- … so that the visual feedback appears synchronous (but is actually delayed)
- Continuously update the prediction when commands are received

➡ **execute commands before having received them!**



Penco L, Mouret JB, Ivaldi S. **Prescient teleoperation of humanoid robots.** arXiv preprint arXiv:2107.01281. 2021 Jul 2.

We know:

- the posture of the robot
- the position of the wall wrt the robot

→ Where should the robot put the hand?

Decision in 100-200 ms max

1-Data Collection

Random Sampling with Rejection

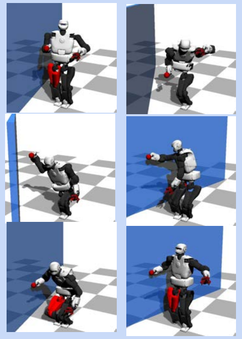
Hands positions

↓ **WBC**

Robot posture **q**

+

Wall configuration **(d, α)**



↓

Random Combination of Damage Conditions

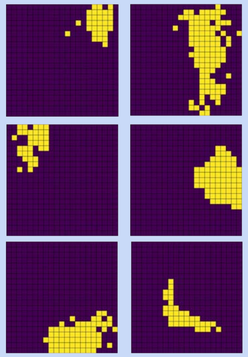
+

Contact Position

(21x21 discretization)

↓ **WBC**

Contact Map M



↓

Dataset **D** of 2000 tuples **(q, d, α, M)**

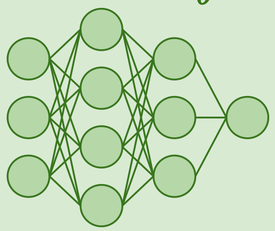
2-Training

Dataset **D** of 2000 tuples **(α, d, q, M)** random split

Training (37.5%) | Validation (12.5%) | Test (50%)

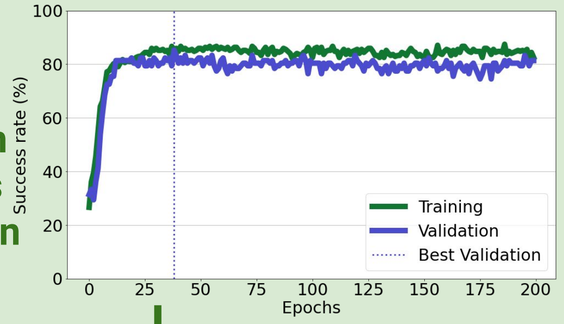
Classifier C_θ

Robot posture **q (37)**
Wall distance **d (1)**
Wall orientation **α (1)**
Contact **x** position **(1)**
Contact **y** position **(1)**



Confidence Score **c** [0,1]

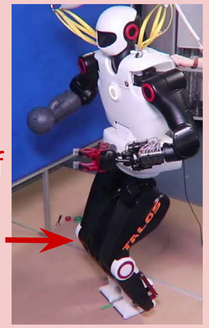
Supervised Learning Classification with **Success Rate Validation**



↓

Trained Neural Classifier C_{θ^*}

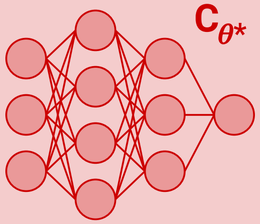
3-Inference



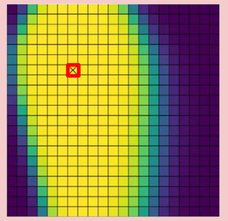
Loss of power in the knee

(q, d, α) →

Contact Map Estimation C_{θ^*}



↑ **M**



↓

Set **(x*, y*) = argmax M** as a **Contact Constraint**

WBC →

Fall Avoided

Key question: connecting language and action

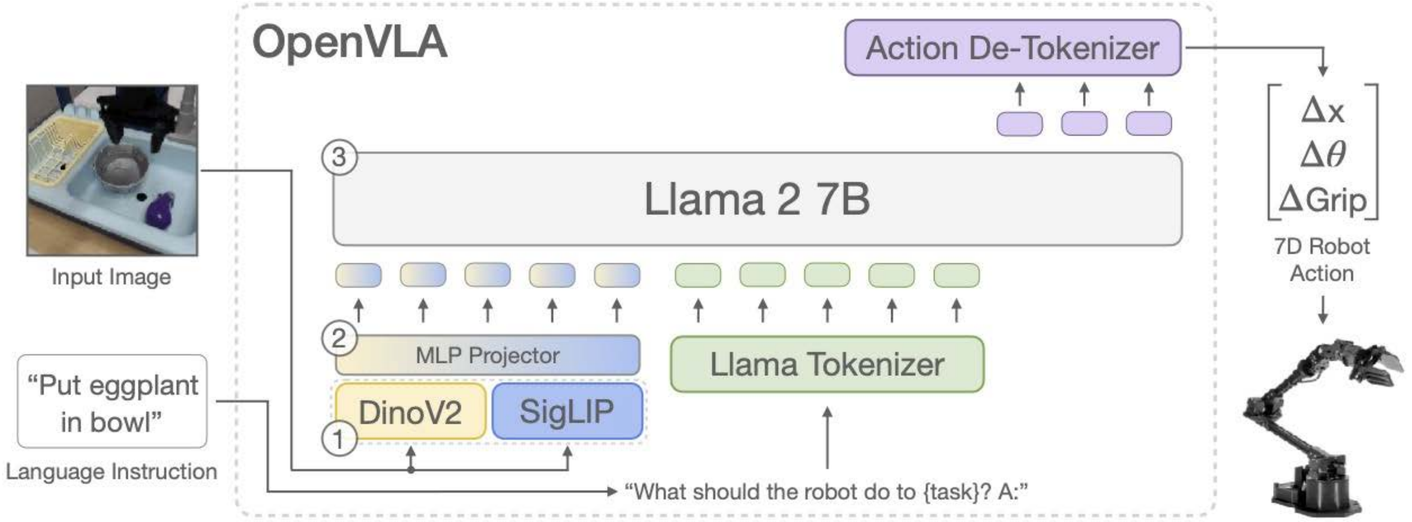
Approach 2: data-driven

Train "Visual-Language-Action" (VLA)

No model "from scratch" (usually blend of pretrained and trained)

Several attempts (manipulation) to get large datasets of image-language-motion (mostly US):

- Open X-Embodiment (RT-X): 22 robots, 21 institutions, 527 skills (160266 tasks).
- Aloha unleashed: 26,000 demonstrations for 5 tasks on a real robot



Zhao, Tony Z., et al. (2024) "Aloha unleashed: A simple recipe for robot dexterity." *CoRL*
Vuong, Quan, et al. (2023) "Open x-embodiment: Robotic learning datasets and RT-x models." arXiv preprint arXiv:2310.08864
Kim, M. J., et al. (2024). OpenVLA: An Open-Source Vision-Language-Action Model. arXiv preprint arXiv:2406.09246.