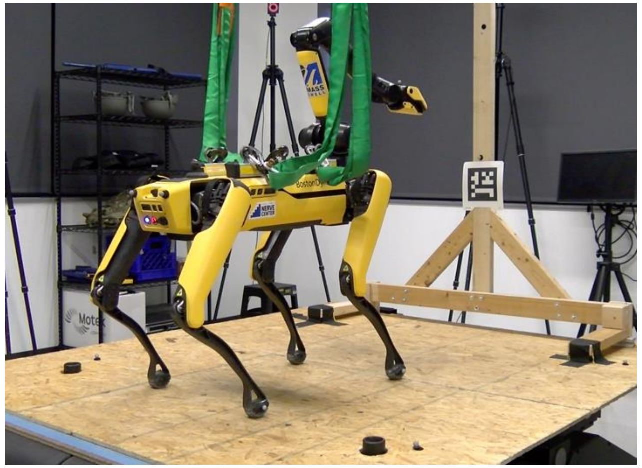
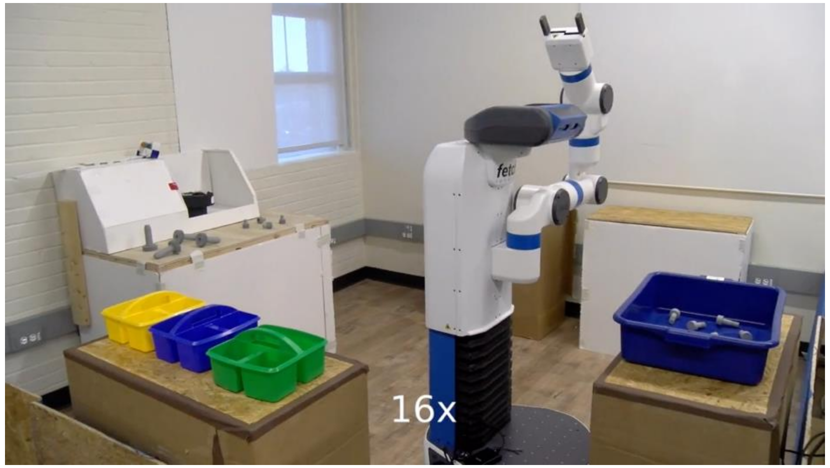
# Robotics Standards and Performance Measurement





**Adam Norton**
Co-Director, NERVE Center
University of Massachusetts Lowell
Lowell, MA, USA
adam_norton@uml.edu

# Adam Norton

adam_norton@uml.edu

Co-Director, NERVE Center

University of Massachusetts Lowell

ASTM International Board of Directors, Director (2024-2026 term)

ASTM F45 Committee Chair

ASTM F48 Standards development, Voting member

ASTM E54.09 Standards development, Voting member

ARM Institute Metrics and Evaluation Working Group, Lead

COMPARE Ecosystem, Community Facilitator

IEEE RAS TC for Performance Evaluation & Benchmarking of Robotic and Automation Systems (PEBRAS), Co-Chair

<u>Robotics research interests</u>: test methods, performance evaluation, metrics, benchmarking, reproducibility, and standardization

## University of Massachusetts Lowell

# New England Robotics Validation and Experimentation (NERVE) Center

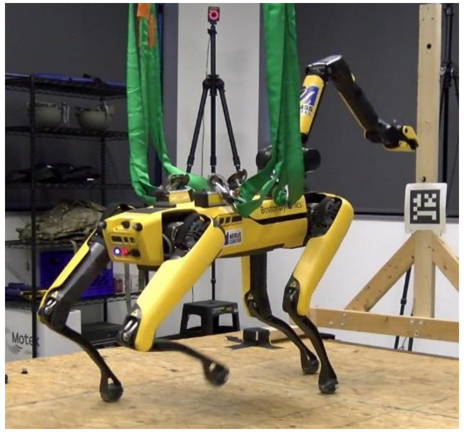
Interdisciplinary robotics testing, research, and training facility that evaluates robot capabilities, human performance, and human-robot interaction

UMASS LOWELL

UMASS LOWELL NERVE CENTER

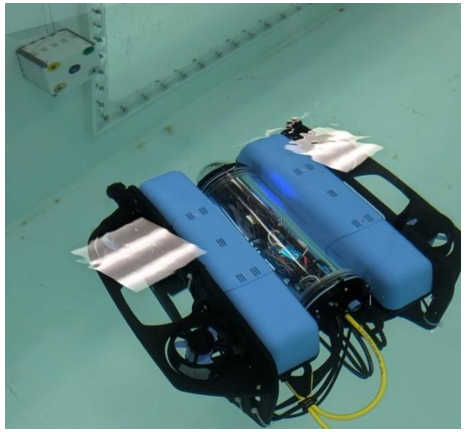

Exoskeletons and Wearable Robots


Human Performance

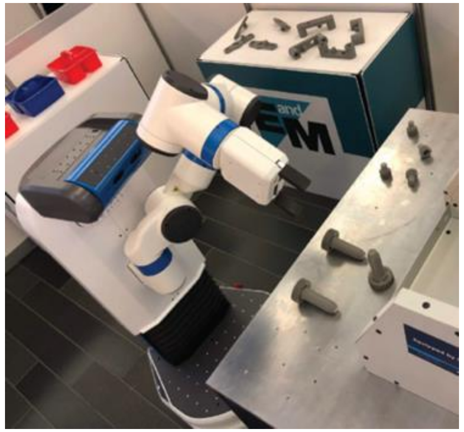

Legged Robots Systems


Human-Robot Teaming


Uncrewed Aerial Systems


Aquatic Robot Systems


Uncrewed Ground Vehicles


Grasping and Manipulation

Benchmarking

Standards

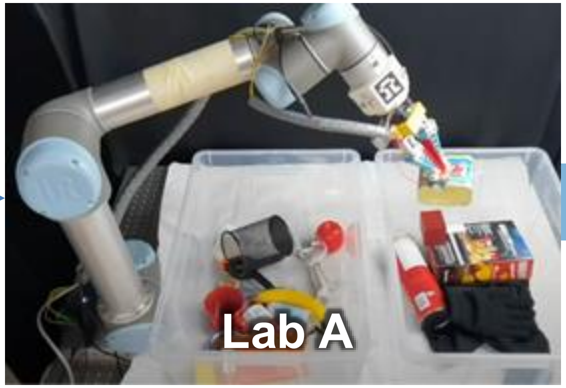
Reproducibility

Evaluation

Generalizability

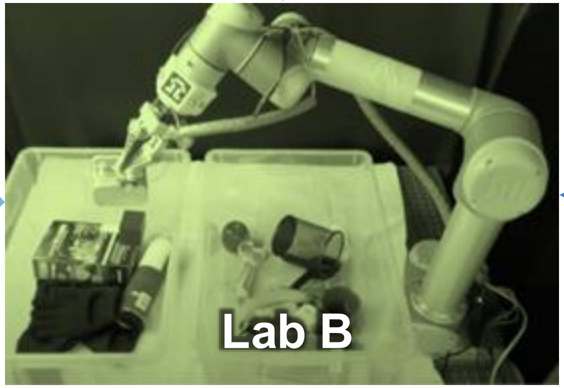
# Benchmarking and Reproducibility in Robotics

COMPARE




Paper states X performance


Reproduce in another lab




Build on top of Lab A solution

COMPARE

COMPARE

*What can be reproduced?*

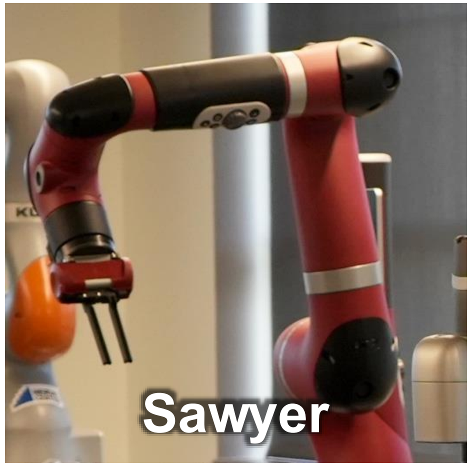
**Context → Functionality → Results**

*What can enable reproducibility?*

**Open-Source | Standards**

# Benchmarking Robot Manipulation

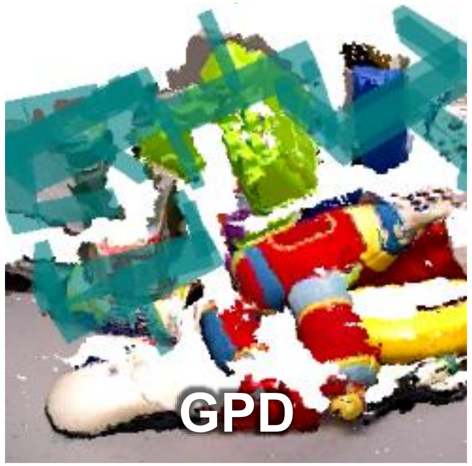
## HARDWARE




**Robot**

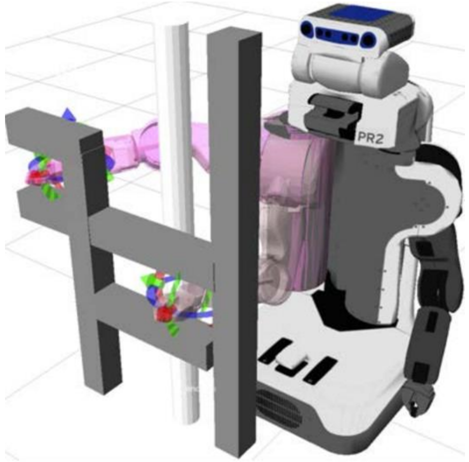
## SOFTWARE


**Perception**



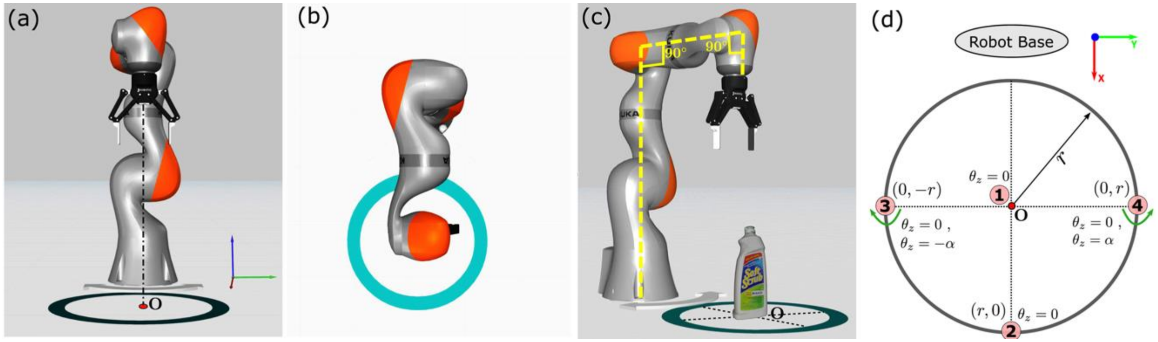

**Grasp Planning**

[Ten Pas et al., 2017]




**Motion Planning**

## EVALUATION




**Benchmarking Protocol**

=

## METRICS

| Trial # | Object name (YCB ID) | Pickup order | Selected (total Feasible) | Time (seconds) | Lift Test | Rotational Test | Shaking Test |
|---|---|---|---|---|---|---|---|
| 1 | Yellow Cup (56) | 4 | 19 (100) | 2.874 | ✔ | ✔ | ✔ |
| | Racquete ball (53) | 2 | 6 (60) | 6.083 | ✔ | ✔ | ✔ |
| | Scrub Cleanser (20) | 3 | 7 (69) | 6.665 | ✔ | ✔ | ✔ |
| | Flat Screwdriver (43) | 6 | 1 (10) | 0.229 | ✔ | ✔ | ✔ |
| | Big Clamp (46) | 5 | 21 (100) | 2.36 | ✔ | ✔ | ✔ |
| | Toy plane (67) | 1 | 1 (82) | 5.891 | ✔ | ✔ | ✔ |
| 2 | Yellow Cup (56) | 2 | 54 (100) | 6.121 | ✔ | ✔ | ✔ |
| | Racquete ball (53) | 3 | 65 (73) | 5.185 | ✔ | ✔ | ✔ |
| | Scrub Cleanser (20) | 4 | 51 (52) | 5.106 | ✔ | ✔ | ✔ |
| | Flat Screwdriver (43) | 6 | 1 (90) | 1.073 | ✔ | ✔ | ✔ |
| | Big Clamp (46) | 5 | 63 (87) | 4.36 | ✔ | ✔ | ✔ |
| | Toy plane (67) | 1 | 54 (100) | 6.424 | ✔ | ✔ | ✔ |

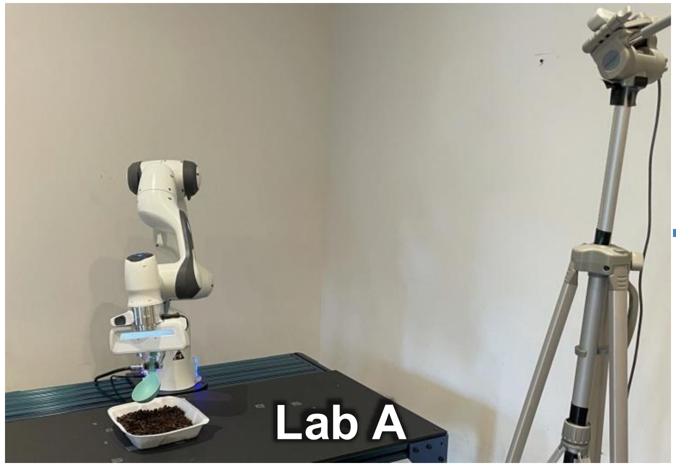
[Bekiroglu et al., 2020]

# Benchmarking and Reproducibility in Robot Manipulation

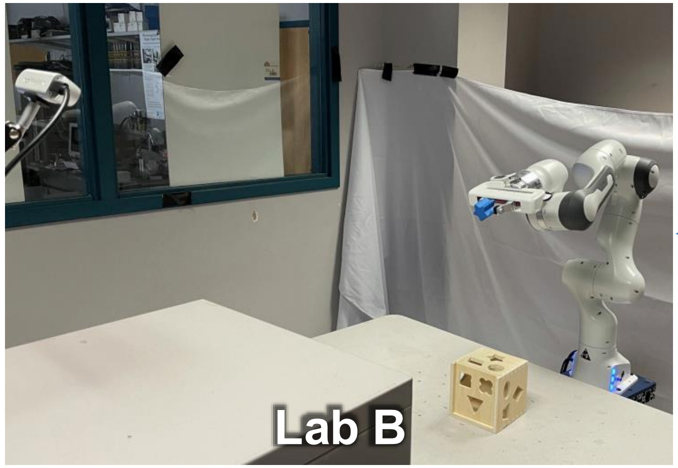
### RB2: Robotic Manipulation Benchmarking with a Twist

[Dasari et al., 2022]

- Compared five learning approaches run "identically" in two different labs to perform pouring, scooping, insertion, and zipping tasks
- ~20% variation in performance results observed when reproducing the experiments across the two labs
- *"…building precisely reproducible robotic setups is **impossible** and therefore absolute performance numbers on a benchmark task are **meaningless**…"*




VS




REPRODUCE

# Benchmarking and Reproducibility in Robot Manipulation

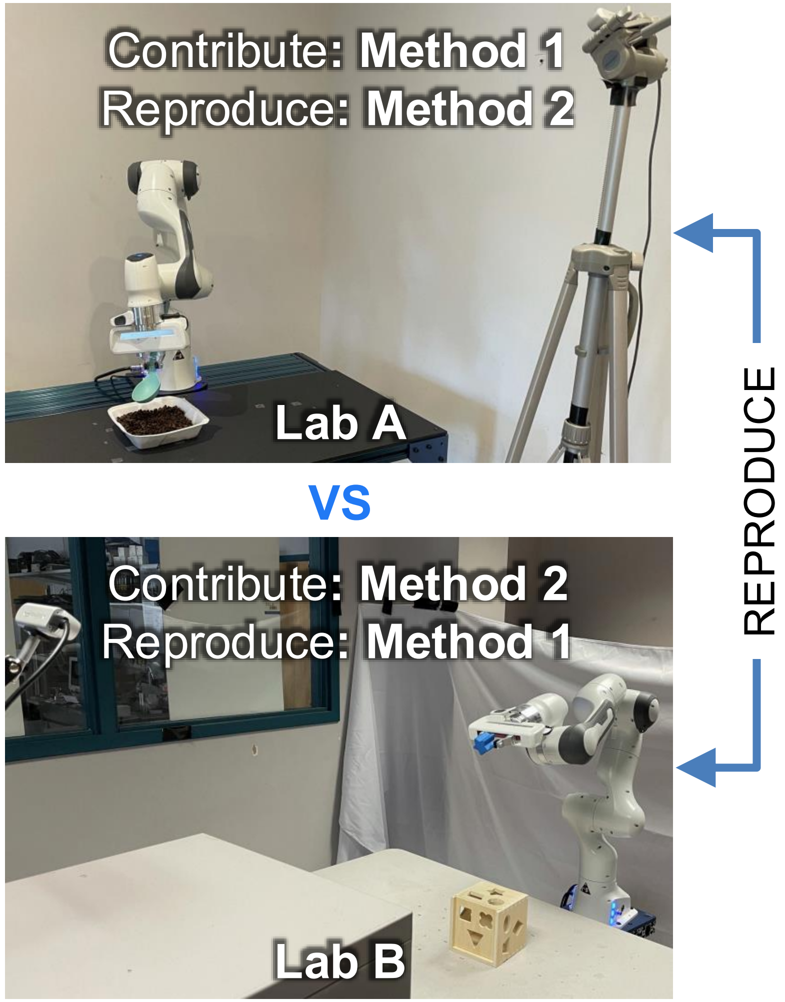
**RB2: Robotic Manipulation Benchmarking with a Twist**

[Dasari et al., 2022]

- **Local relative ranking (LRR)**: each lab establishes a baseline in their lab for comparison by reproducing the contribution of the other lab and running experiments

- Developed for the evaluation of software/algorithms as they can be easily shared (as opposed to hardware)

- The authors also propose a method to **globally rank** the contributed LRRs from multiple labs (see paper)



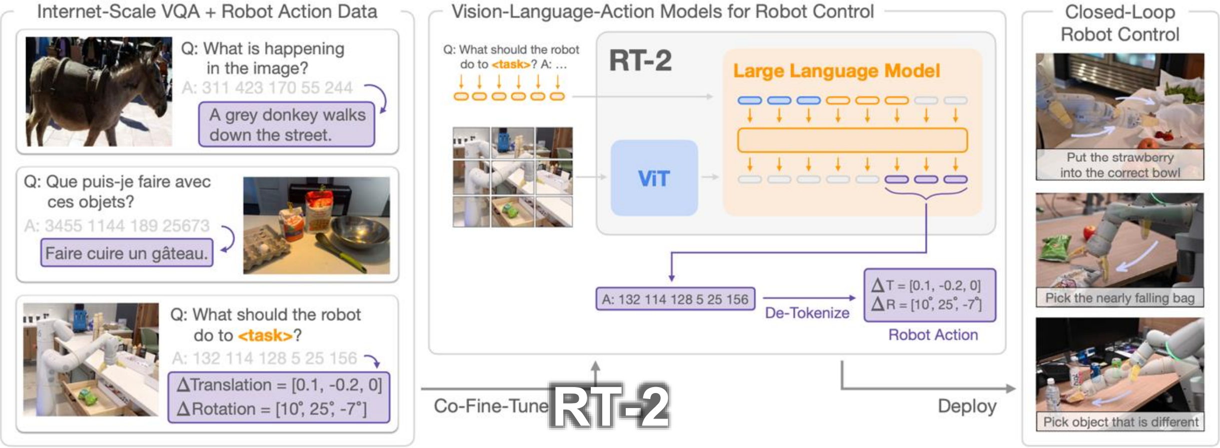
# Benchmarking Robot Manipulation

## HARDWARE


**Robot**

+

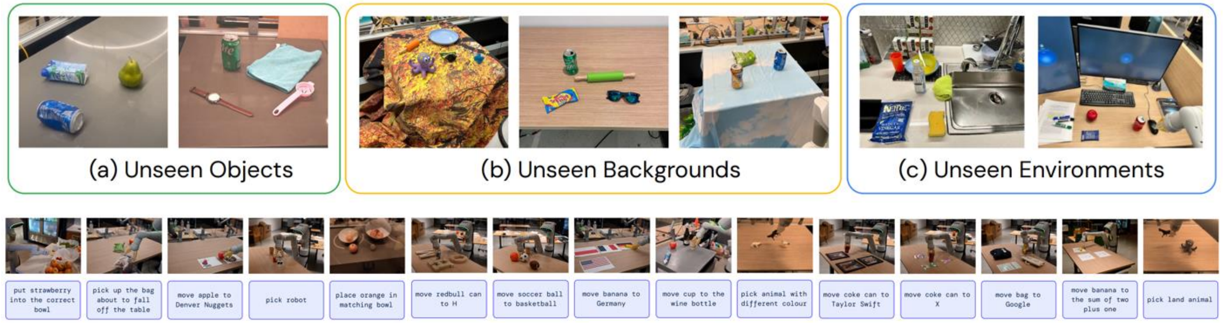
## SOFTWARE




[Zitkovich et al., 2023]

**Vision-Language-Action Model (VLA)**

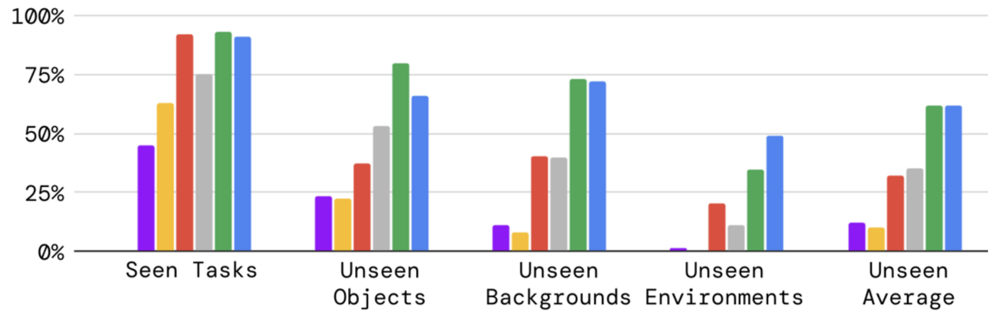
## EVALUATION




**Benchmarking Protocol**

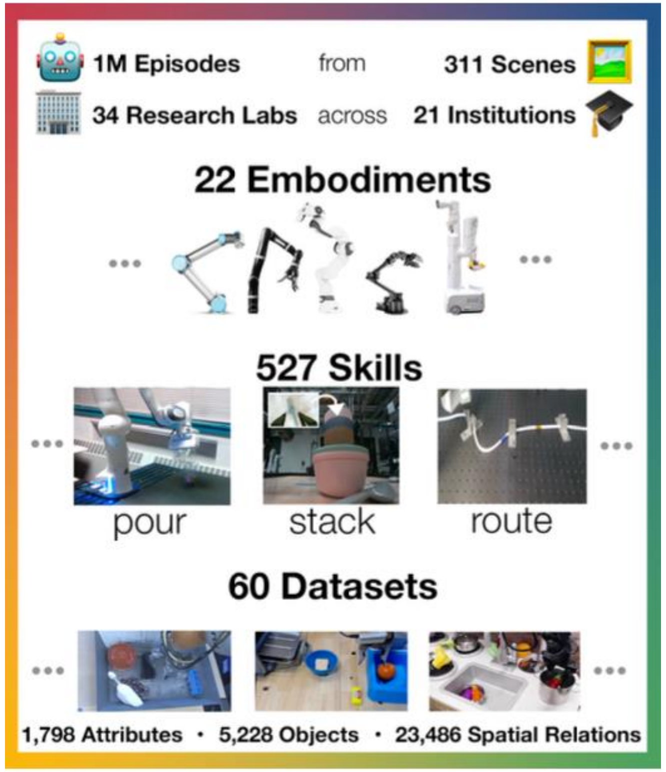
=

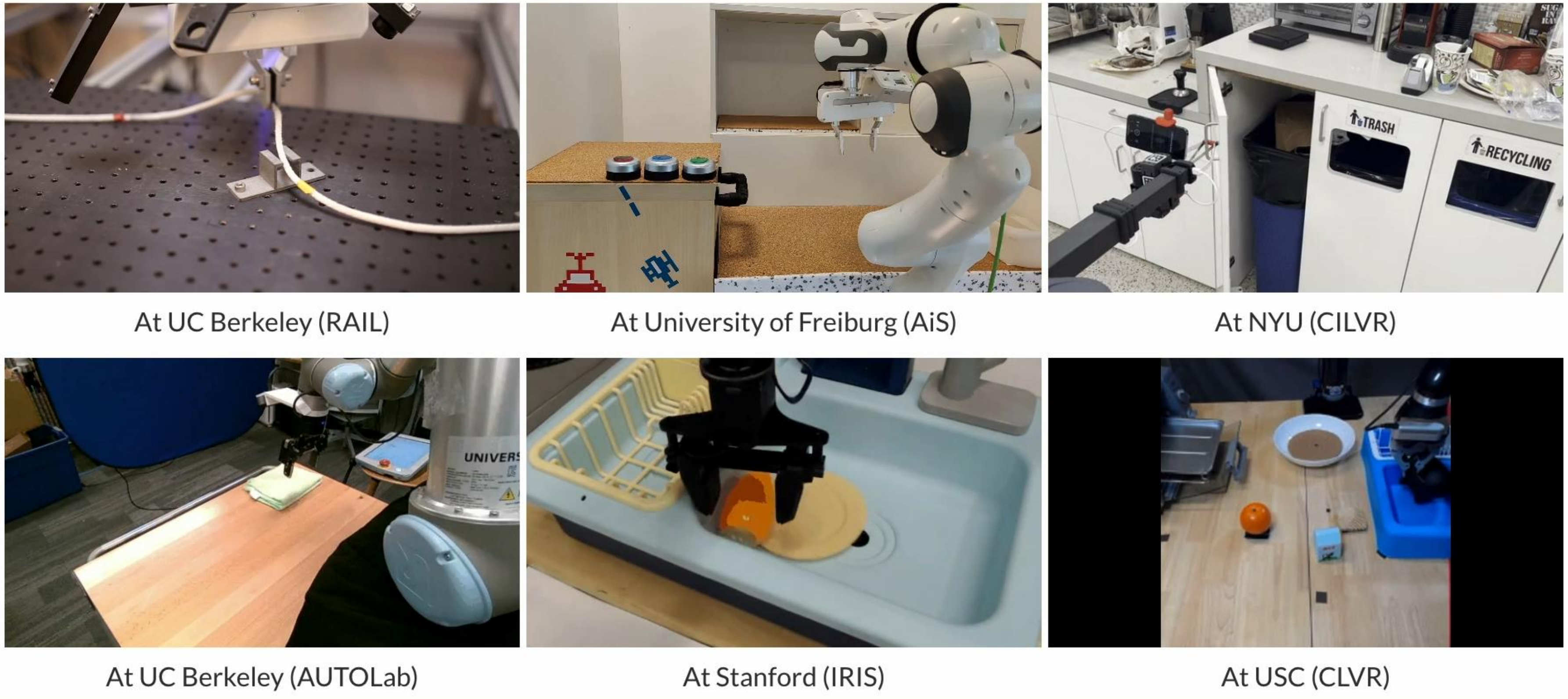
## METRICS




[Zitkovich et al., 2023]

# Benchmarking and Reproducibility in Robot Manipulation




*Cross-embodiment training results in generalist models that outperform specialist models on their own tasks*

**Open X-Embodiment** [O'Neill et al., 2024] https://robotics-transformer-x.github.io/

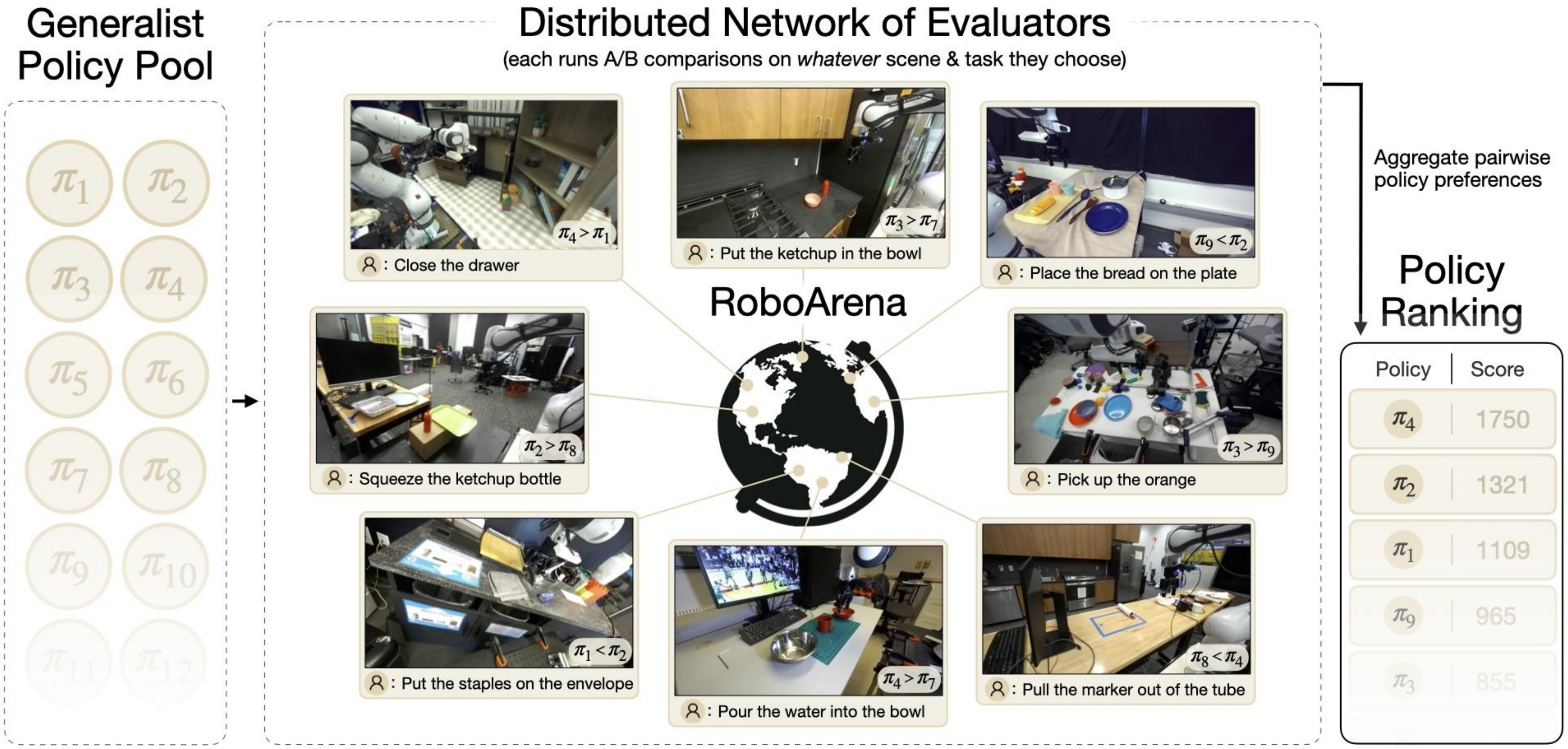
# Benchmarking and Reproducibility in Robot Manipulation

- Sergey Levine, "*Benchmarks, Surrogate Objectives, and Cross-Embodiment*," RSS 2025 Workshop on Benchmarking Robot Manipulation: Improving Interoperability and Modularity
  - Shared benchmarks can be very challenging, so benchmarks are essentially "surrogate functions" (very local surrogate functions)
  - ***Benchmarks are local!*** →

    [Dasari et al., 2022]

    - *"…building precisely reproducible robotic setups is* ***impossible*** *and therefore absolute performance numbers on a benchmark task are* ***meaningless****…"*
  - To be practical, researchers should be able to run their own evaluations
  - Require more honesty and transparency from researchers

# Benchmarking and Reproducibility in Robot Manipulation



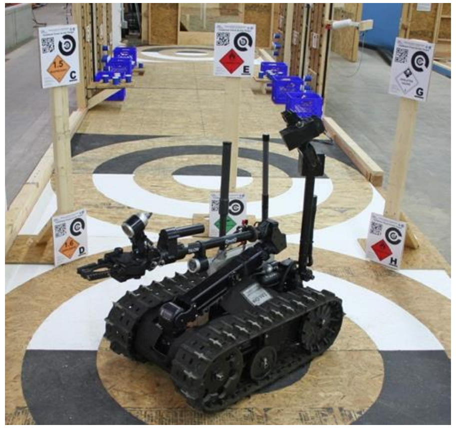

**RoboArena** [Atreya et al., 2025] https://robo-arena.github.io/

# Standard Test Methods Applicable to Humanoids

## ASTM E54 Committee on Homeland Security Applications
## ASTM E45.09 Subcommittee on Response Robots

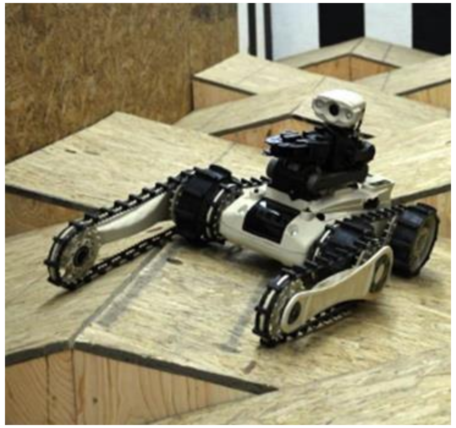
ASTM INTERNATIONAL


Mobility

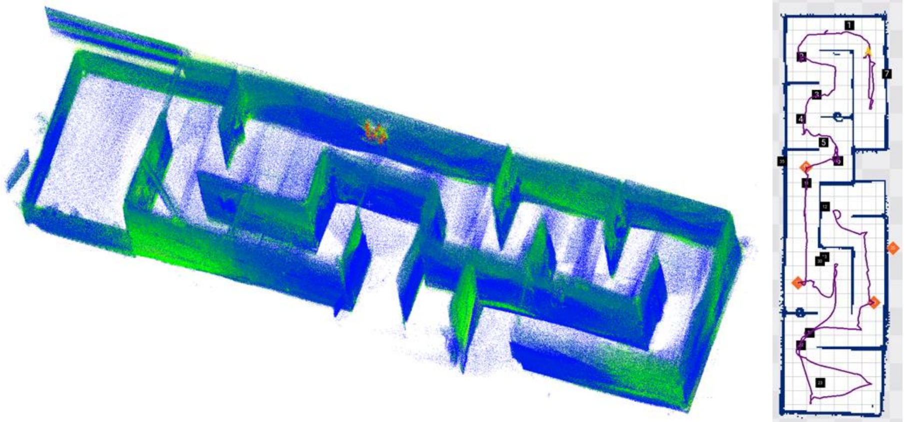


Sensors


Dexterity



Endurance

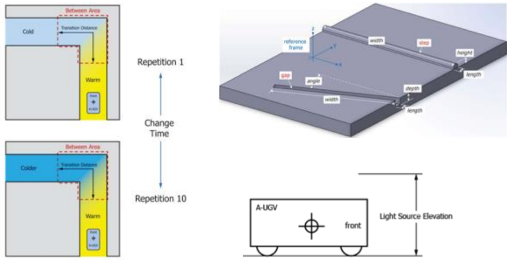


Mapping

# Standard Test Methods Applicable to Humanoids

## ASTM F45 Committee on Robotics, Automation, and Autonomous Systems

ASTM INTERNATIONAL

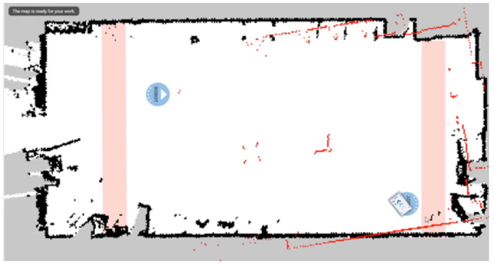



Environment

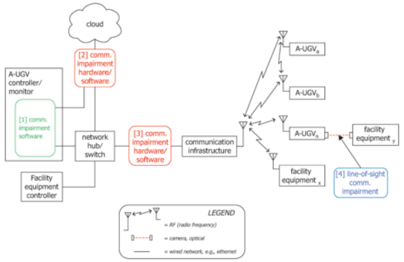

A-UGV Docking


A-UGV Navigation

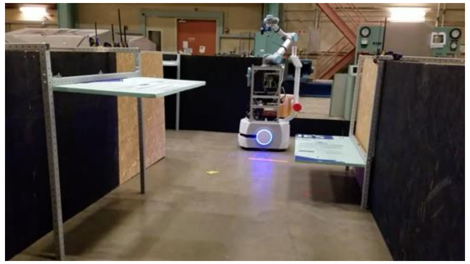
A-UGV Obstacle Avoidance



A-UGV Localization

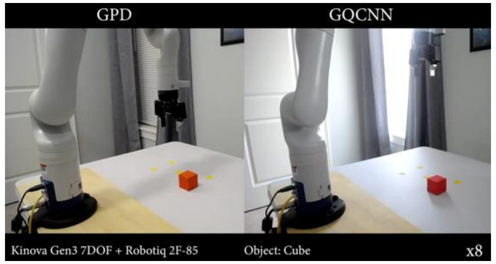



System Communication

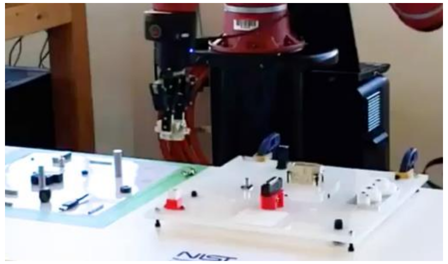


System Interoperability

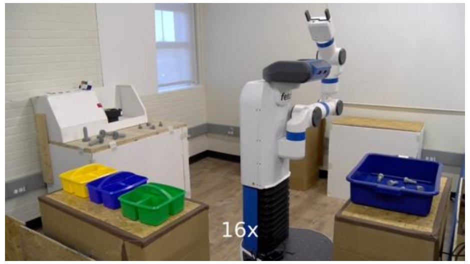

Robotic End-Effectors




Robotic Grasping



Robotic Manipulation



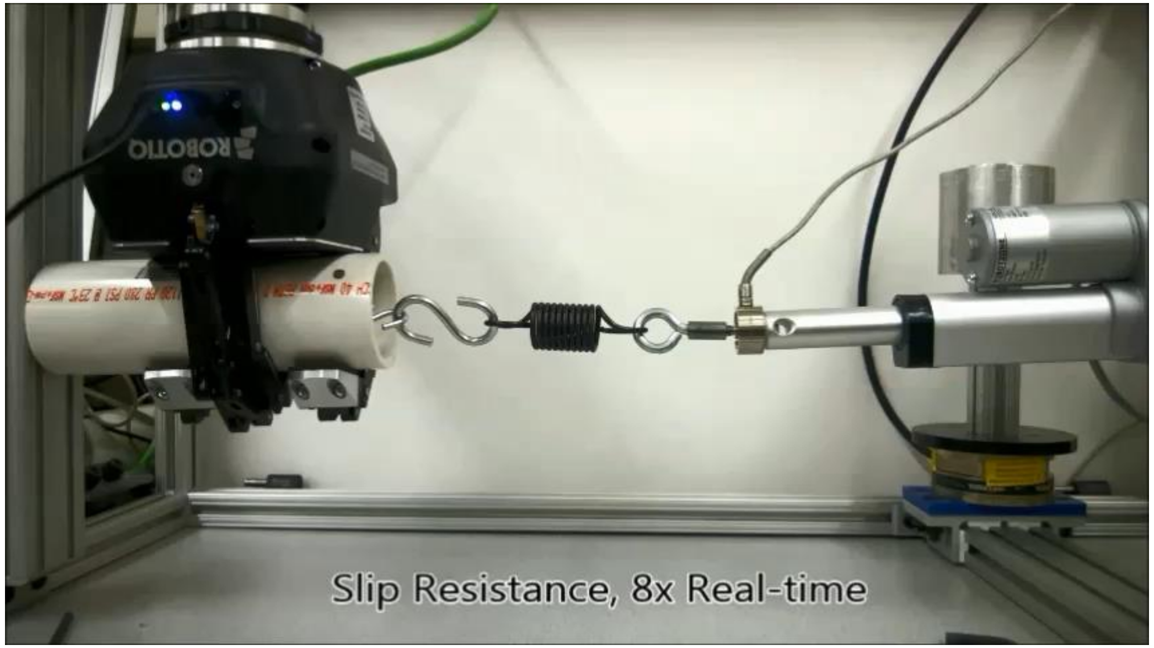

Mobile Manipulators

Legged Robots

# Standard Test Methods Applicable to Humanoids
# F45.05 Subcommittee on Grasping and Manipulation

ASTM INTERNATIONAL

## Evaluating Grasp-Type End-Effector Grasp Strength and Slip Resistance


ASTM F3756-25 Standard Practice for Grasp-Type Robot End-Effectors: Split Force Measurement Apparatus
ASTM WK83863 Grasp-Type Robot End-Effectors: Grasp Strength Performance
ASTM WK96472 Standard Test Method for Grasp-Type Robot End-Effectors: Slip Resistance
ASTM WK96417 Standard Test Method for Grasp-Type Robot End-Effectors: Cycle Time

NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

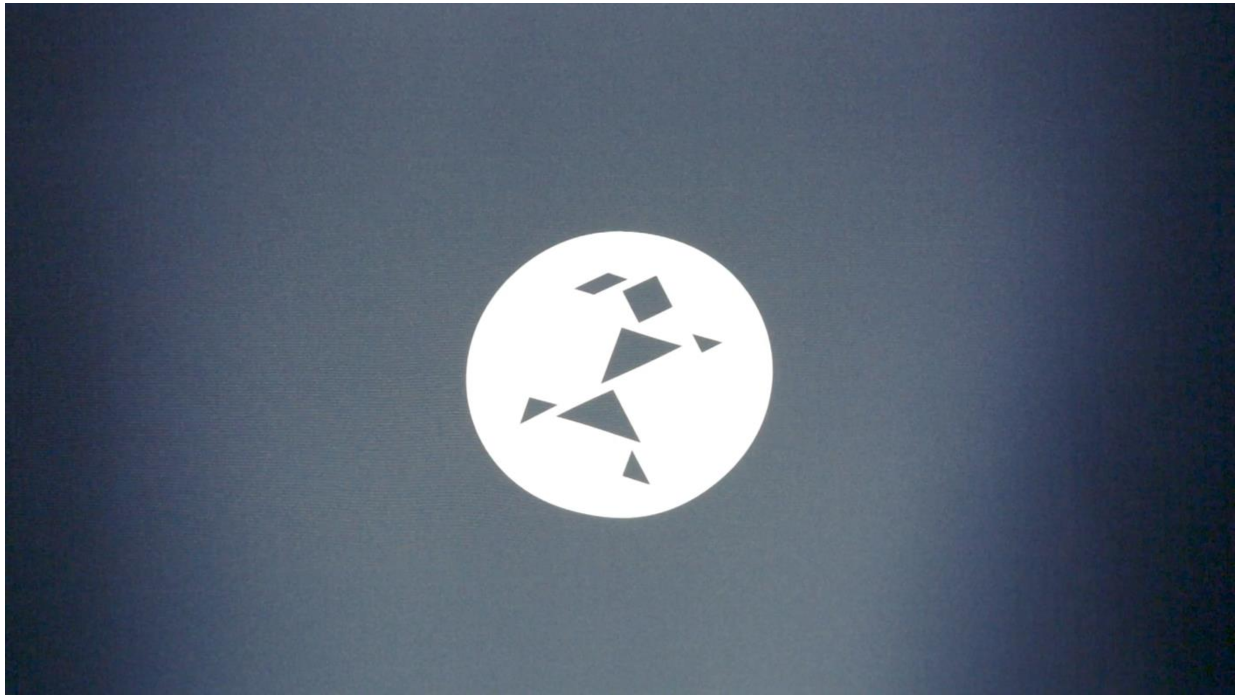
# Standard Test Methods Applicable to Humanoids
# F45.05 Subcommittee on Grasping and Manipulation

ASTM INTERNATIONAL

## Evaluating Robotic Assembly Tasks

Video: https://www.youtube.com/watch?v=Fy8dnS45YyA


Assembly Task Board



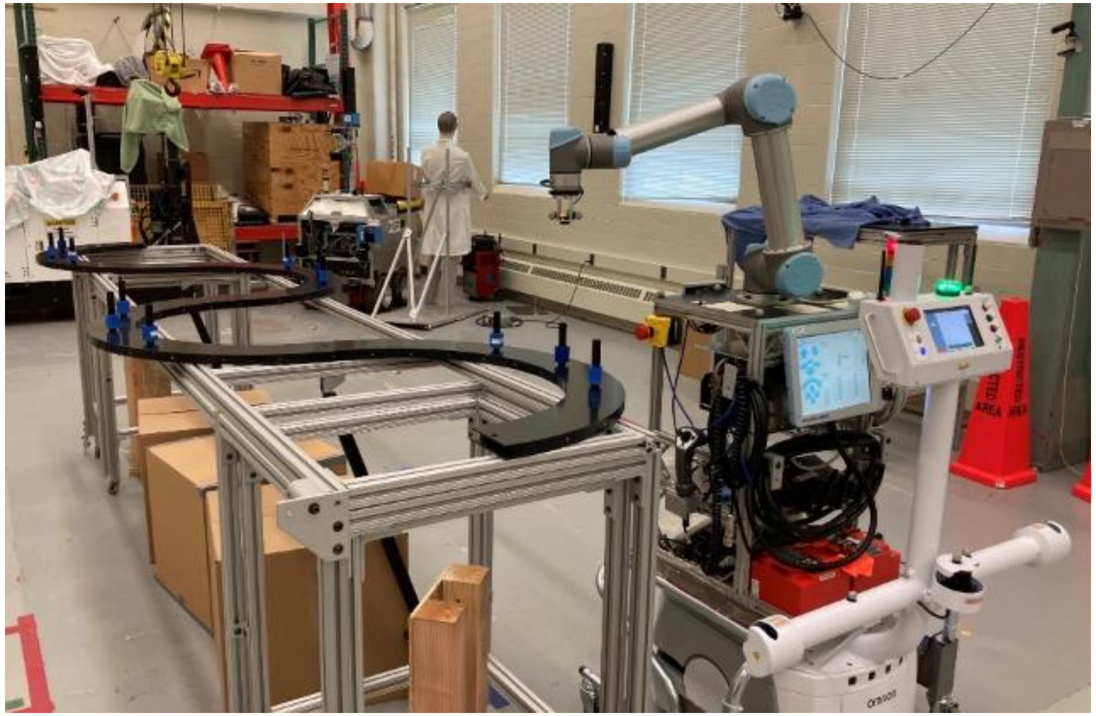
ASTM WK87214 Robotic Assembly Task Boards: Benchmarking performance
ASTM WK87213 Robotic Assembly Task Boards: Recording Assembly Test Configuration

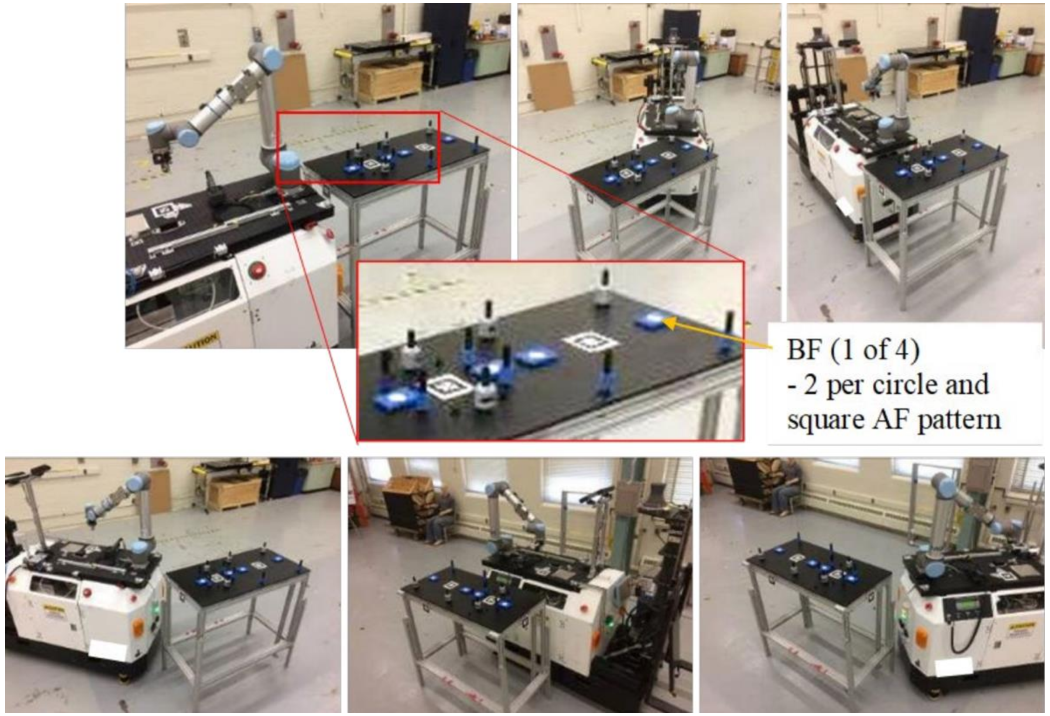
NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY U.S. DEPARTMENT OF COMMERCE

# Standard Test Methods Applicable to Humanoids
# F45.05 Subcommittee on Grasping and Manipulation

ASTM INTERNATIONAL

## Evaluation of Non-Continuous and Continuous Mobile Manipulator Tasks






ASTM F3713-25 Standard Practice for Measuring Mobile Manipulator Performance: Recording the Workpiece Configuration
ASTM WK89198 Mobile Manipulator Performance: Continuous Tasks
ASTM WK83858 Measuring Mobile Manipulator Performance: Non-continuous Tasks
ASTM WK92144 Measuring Mobile Manipulator Performance: Inducing Workpiece Disturbance Impairment

NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY U.S. DEPARTMENT OF COMMERCE

# Standard Test Methods Applicable to Humanoids

- “The Kick Test” / “The Hockey Stick Test”


https://www.youtube.com/watch?v=60Y3IEtCsDg

# Standard Test Methods Applicable to Humanoids
# F45.06 Subcommittee on Legged Robot Systems

ASTM INTERNATIONAL

## Disturbance Rejection Testing of Legged Robot Locomotion

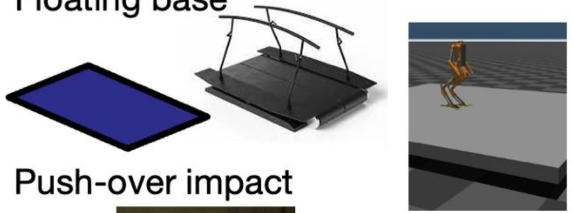

[Weng et al., 2024]

**Static** Disturbances

Terrain condition & Tripping hazards

**Dynamic** Disturbances

Floating base



Push-over impact

IOWA STATE UNIVERSITY

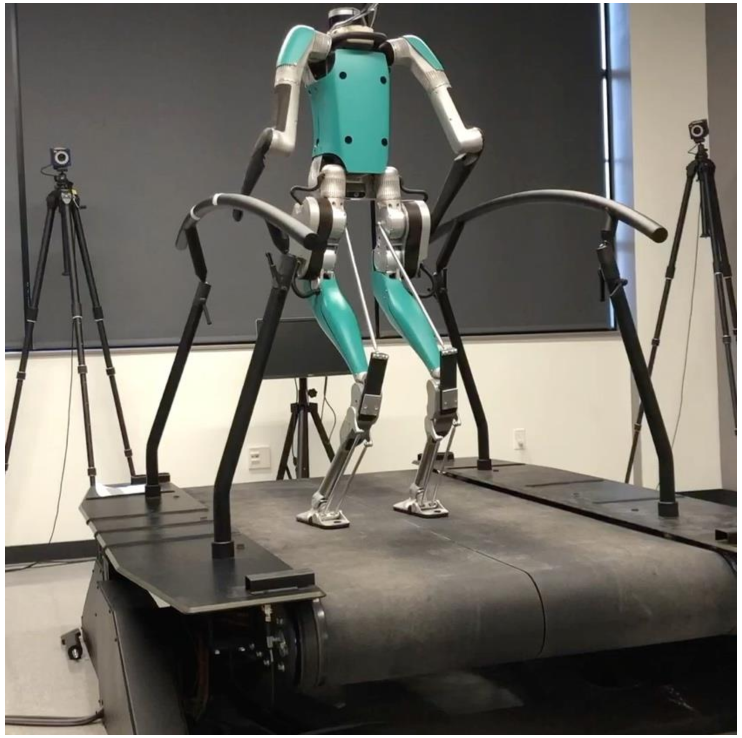
ASTM WK86916 Test Methods for Disturbance Rejection Testing of Legged Robots

# Standard Test Methods Applicable to Humanoids
# F45.06 Subcommittee on Legged Robot Systems

ASTM INTERNATIONAL

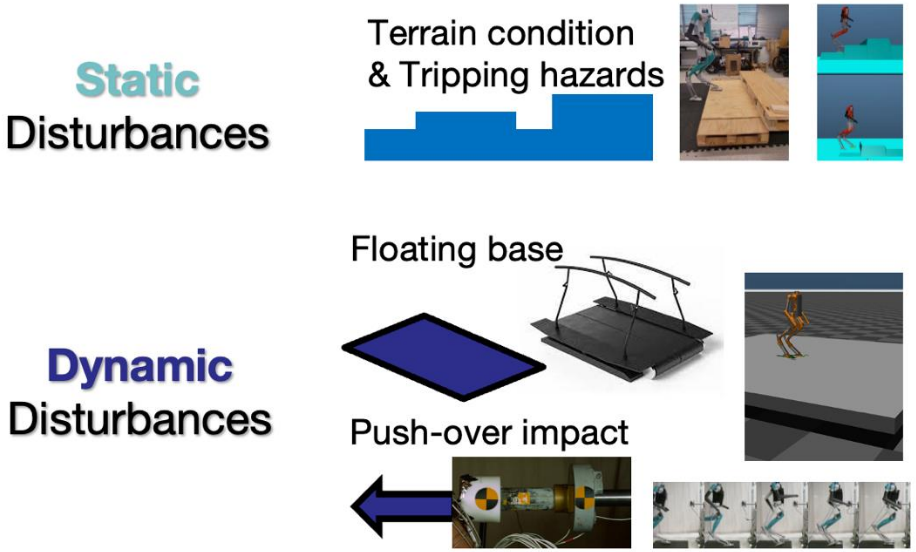
## Disturbance Rejection Testing of Legged Robot Locomotion

Default Agility Robotics controller



Custom research controller

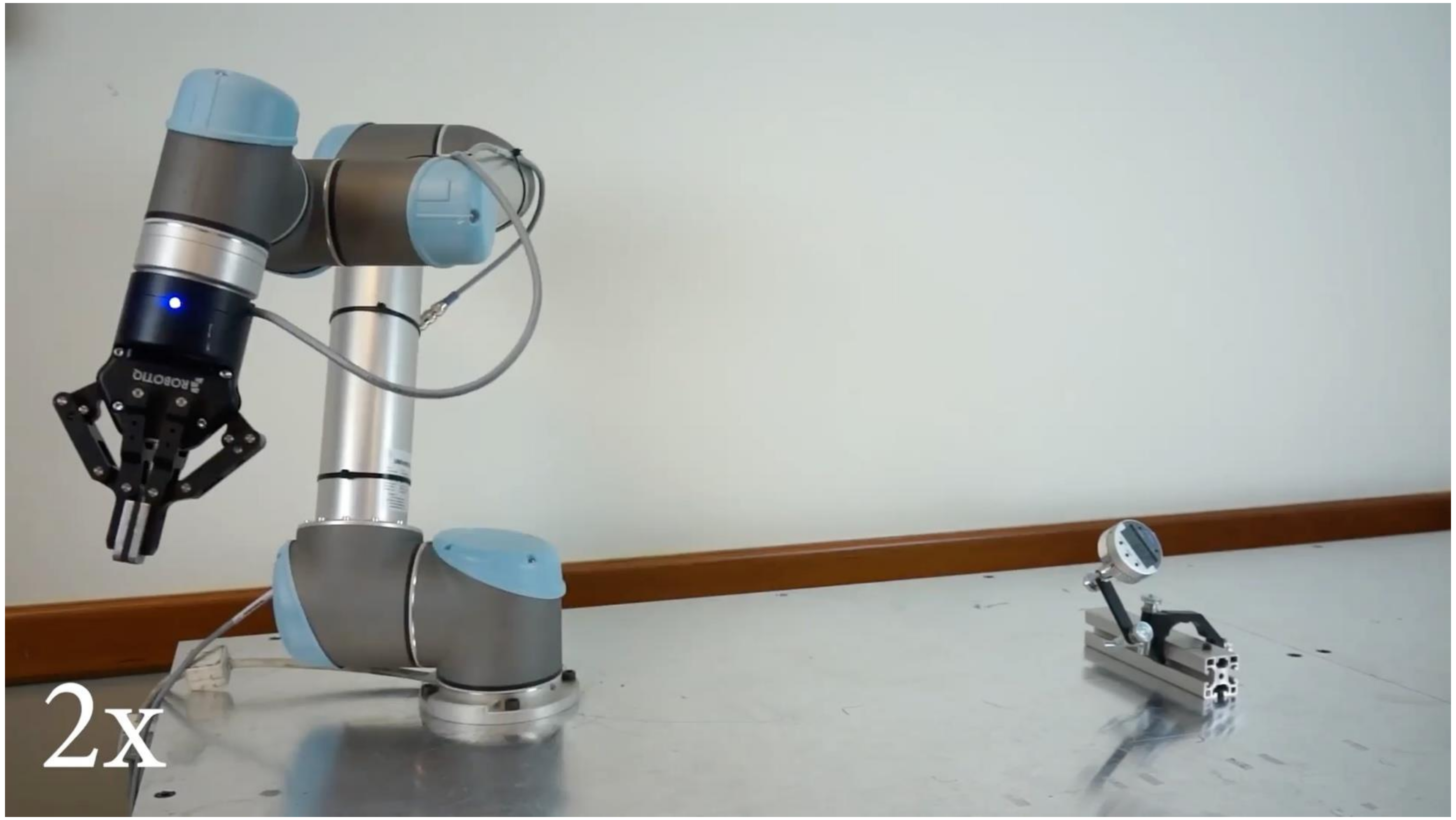
[Gao et al., 2022]




ASTM WK86916 Test Methods for Disturbance Rejection Testing of Legged Robots

# Standard Test Methods Applicable to Humanoids

- **Repeatability**: ISO 9283:1998 Manipulating industrial robots — Performance criteria and related test methods




https://www.youtube.com/watch?v=9orN_aUDY7w

**Universal Robots UR5**:

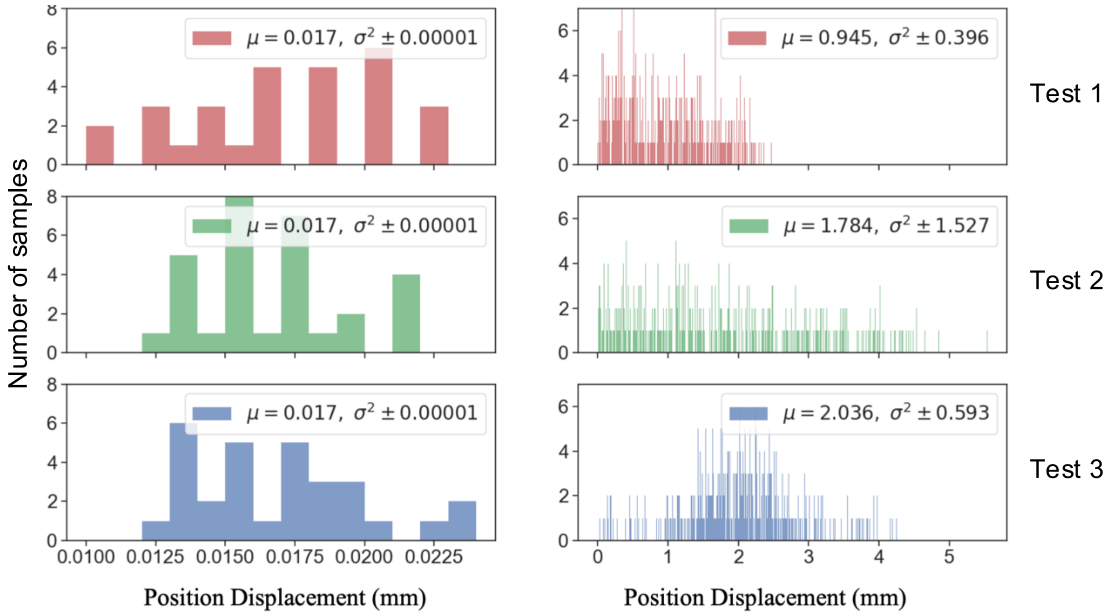
+/- 0.1 mm [0.004 in]

# Standard Test Methods Applicable to Humanoids

- **Repeatability**: ISO 9283:1998 Manipulating industrial robots — Performance criteria and related test methods

**UR10e** with commercial algorithm and software stack | **Unitree G1** with custom impedance whole-body control

IOWA STATE UNIVERSITY




*ISO 9283 procedure does not yield consistent and repeatable results for humanoids*

[Weng et al., 2025]

Benchmarking

Standards

If benchmarking results in research are local, can they be used for global comparison through standard test methods?

Reproducibility

How can we update the standards we have for humanoid robots that allows them to be trusted and repeatable?

Evaluation

Generalizability

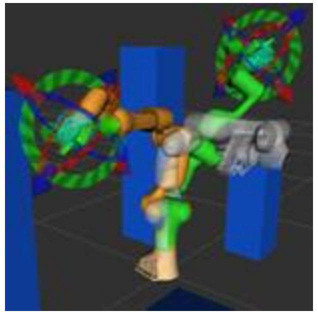
# robot-manipulation.org

- An online landing page for open-source and benchmarking in robot manipulation, including:



**Repositories** of open-source products and benchmarking assets

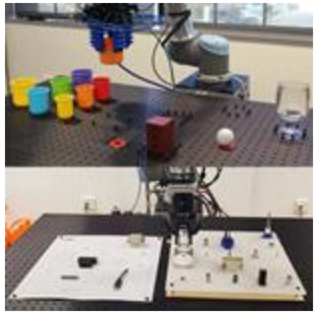
**Open-Source Products**



Software Components

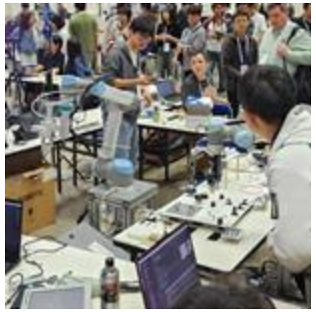
Hardware Designs

**Benchmarking Assets**



Objects, Artifacts, Tools, and Testbeds

Protocols and Leaderboards




**Event** listings for workshops, competitions, and webinars

- Repositories and event listings that can be filtered and sorted
- Google Forms available for users to submit content
- Google Calendar of events you can subscribe to
- Updates are communicated over Slack and Google Group

COMPARE Slack (discussion)

robot-manipulation.org

COMPARE Google Group (mailing list)

# Robotics Standards and Performance Measurement

# Thank you!

Download this presentation →

← Join the COMPARE Ecosystem!

**Adam Norton**
Co-Director, NERVE Center
University of Massachusetts Lowell
Lowell, MA, USA
adam_norton@uml.edu

UMASS LOWELL NERVE CENTER

# References

**[Atreya et al., 2025]** Atreya, Pranav, Karl Pertsch, Tony Lee, Moo Jin Kim, Arhan Jain, Artur Kuramshin, Clemens Eppner et al. "RoboArena: Distributed Real-World Evaluation of Generalist Robot Policies." *arXiv preprint arXiv:2506.18123* (2025).

**[Bekiroglu et al., 2020]** Bekiroglu, Yasemin, Naresh Marturi, Máximo A. Roa, Komlan Jean Maxime Adjigble, Tommaso Pardi, Cindy Grimm, Ravi Balasubramanian, Kaiyu Hang, and Rustam Stolkin. "Benchmarking protocol for grasp planning algorithms." IEEE Robotics and Automation Letters 5, no. 2 (2019): 315-322.

**[Dasari et al., 2022]** Dasari, Sudeep, Jianren Wang, Joyce Hong, Shikhar Bahl, Yixin Lin, Austin Wang, Abitha Thankaraj et al. "Rb2: Robotic manipulation benchmarking with a twist." *arXiv preprint arXiv:2203.08098* (2022).

**[Gao et al., 2022]** Gao, Yuan, Chengzhi Yuan, and Yan Gu. "Invariant filtering for legged humanoid locomotion on a dynamic rigid surface." *IEEE/ASME Transactions on Mechatronics* 27, no. 4 (2022): 1900-1909.

**[Gou et al., 2021]** Gou, Minghao, Hao-Shu Fang, Zhanda Zhu, Sheng Xu, Chenxi Wang, and Cewu Lu. "Rgb matters: Learning 7-dof grasp poses on monocular rgbd images." In 2021 IEEE International Conference on Robotics and Automation (ICRA), pp. 13459-13466. IEEE, 2021.

**[Levine, 2025]** Levine, Sergey. "Benchmarks, Surrogate Objectives, and Cross-Embodiment." RSS 2025 Workshop on Benchmarking Robot Manipulation: Improving Interoperability and Modularity, June 2025.

**[Ni et al., 2020]** Ni, Peiyuan, Wenguang Zhang, Xiaoxiao Zhu, and Qixin Cao. "Pointnet++ grasping: Learning an end-to-end spatial grasp generation algorithm from sparse point clouds." In 2020 IEEE International Conference on Robotics and Automation (ICRA), pp. 3619-3625. IEEE, 2020.

# References

**[O’Neill et al., 2024]** O’Neill, Abby, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley et al. "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0." In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892-6903. IEEE, 2024.

**[Sundermeyer et al., 2021]** Sundermeyer, Martin, Arsalan Mousavian, Rudolph Triebel, and Dieter Fox. "Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes." In 2021 IEEE International Conference on Robotics and Automation (ICRA), pp. 13438-13444. IEEE, 2021.

**[Ten Pas et al., 2017]** Ten Pas, Andreas, Marcus Gualtieri, Kate Saenko, and Robert Platt. "Grasp pose detection in point clouds." The International Journal of Robotics Research 36, no. 13-14 (2017): 1455-1473.

**[Weng et al., 2024]** Weng, Bowen, Guillermo A. Castillo, Yun-Seok Kang, and Ayonga Hereid. "Towards standardized disturbance rejection testing of legged robot locomotion with linear impactor: A preliminary study, observations, and implications." In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9946-9952. IEEE, 2024.

**[Weng et al., 2025]** Weng, Bowen, Linda Capito, Guillermo A. Castillo, and Dylan Khor. "Rethink Repeatable Measures of Robot Performance with Statistical Query." arXiv preprint arXiv:2505.08216 (2025).

**[Zitkovich et al., 2023]** Zitkovich, Brianna, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu et al. "Rt-2: Vision-language-action models transfer web knowledge to robotic control." In Conference on Robot Learning, pp. 2165-2183. PMLR, 2023.