

Multimedia content for embodied AI

Andrea Cavallaro

10 October 2025

Agenda

Human-object-robot interaction

Reproducibility: protocol, data, models

Embodied multimodal language models

Natural language to condition perception

Human-object-robot interaction

CORSMAL project

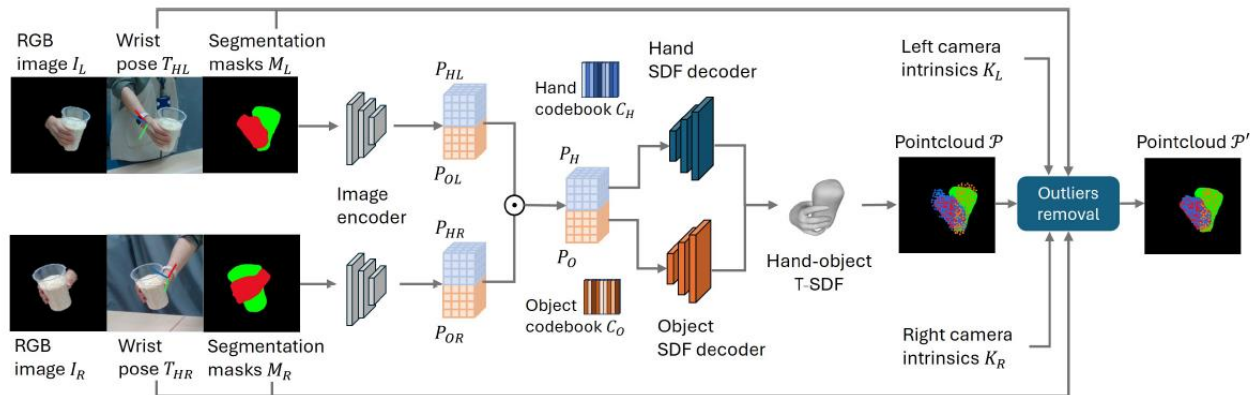
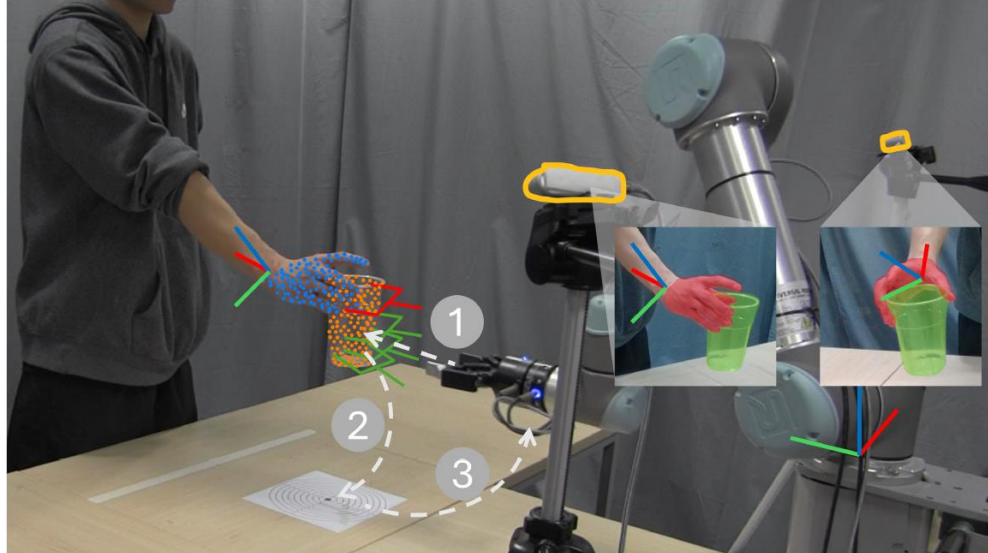
Open science 'success story'

ICRA 2025 Competition

Robot Grasping and Manipulation

IROS 2025 Workshop

Benchmarking via Competitions



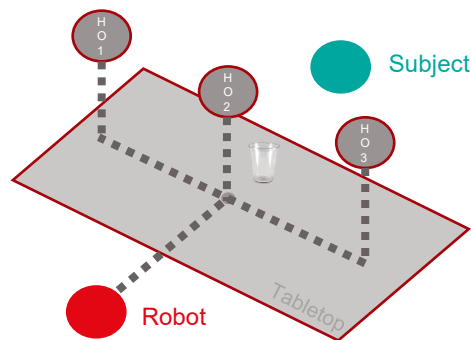
⊙ Element-wise product

Stereo hand-object reconstruction for human-to-robot handover

Pang, Xompero, Oh, Cavallaro. *IEEE Robotics and Automation Letters*, 2025

Reproducibility: protocol, multimedia data, models

Benchmark and baseline algorithms for human-to-robot handovers distributed to the community



Multi-modal datasets: containers manipulated by a human or by a robot



Machine learning models distributed to (and complemented by) the community



Platform for participants to submit the results of the **CORSMAL Challenge**

Team	Description	Task 1	Task 2	Task 3	Public	Private	Overall
Because It's Tacti...	GRU+ Random Forest for filling properties estimation. LODE with RGB-D-IR data from selected frames in a video for volume estimation.	✓	✓	✓	64.98	65.15	65.06
HVRL	Log-Mel spectrogram-based audio features as input to VGG-based CNN and LSTM for filling properties estimation. Container volume from the shape approximation as cuboid of the 3D point cloud obtained with RGB-D data and object detection with Mask R-CNN.	✓	✓	✓	63.32	61.01	62.16
Concatenation	Multi-modal learning with audio features and prior of container categories through object detection for inferring container capacity and fluid properties.	✓	✓	✓	52.80	54.14	53.47
NTNU-ERC	MFCC features in a 20s-window + neural network to classify filling type. Object detection and selection of the closest contours (up to 700 mm) in the depth data + regression with a CNN for container capacity.		✓	✓	38.56	39.80	39.18
Random	Baseline with random estimations for each task.	✓	✓	✓	38.47	31.65	35.06
Challengers	Sound-based classification of filling type and level with STFT and 5-layers fully connected neural network.	✓	✓		29.25	23.21	26.23
SCC-Net	Sound-based hierarchical ensemble of DNNs to jointly classify filling type and level.	✓	✓		28.02	22.92	25.47
Mask R-CNN + R...	Vision baseline for filling properties estimation.	✓	✓		19.46	9.59	14.53
Mask R-CNN + R...	Vision baseline for filling properties estimation.	✓	✓		15.15	9.96	12.56
Mask R-CNN + R...	Vision baseline for filling properties estimation.	✓	✓		17.28	6.99	12.14
Mask R-CNN + R...	Vision baseline for filling properties estimation.	✓	✓		12.95	10.25	11.60

CORSMAL Protocol and Code

github.com/CORSMAL/Benchmark
ieeexplore.ieee.org/document/8968407
IEEE RA-L



CORSMAL Datasets

doi.org/10.17636/101CORSMAL1
doi.org/10.17636/corsmal2
doi.org/10.5281/zenodo.6372438
doi.org/10.5281/zenodo.4642577

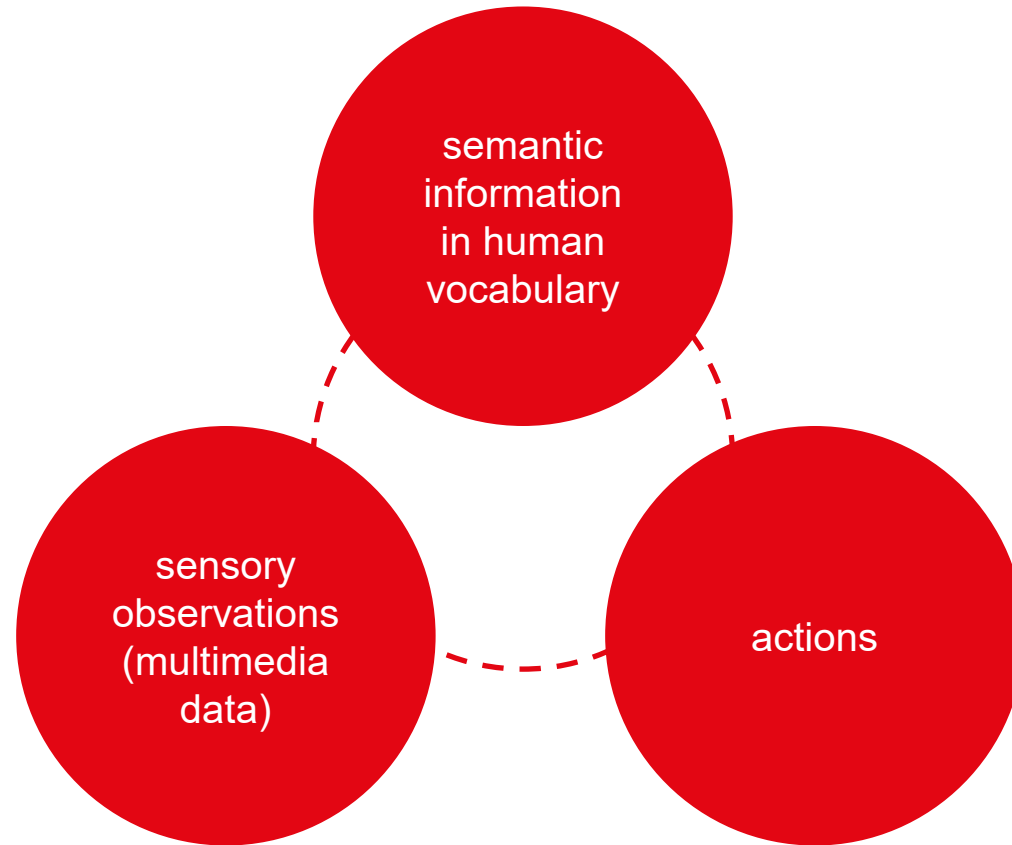
CORSMAL Pre-trained Models

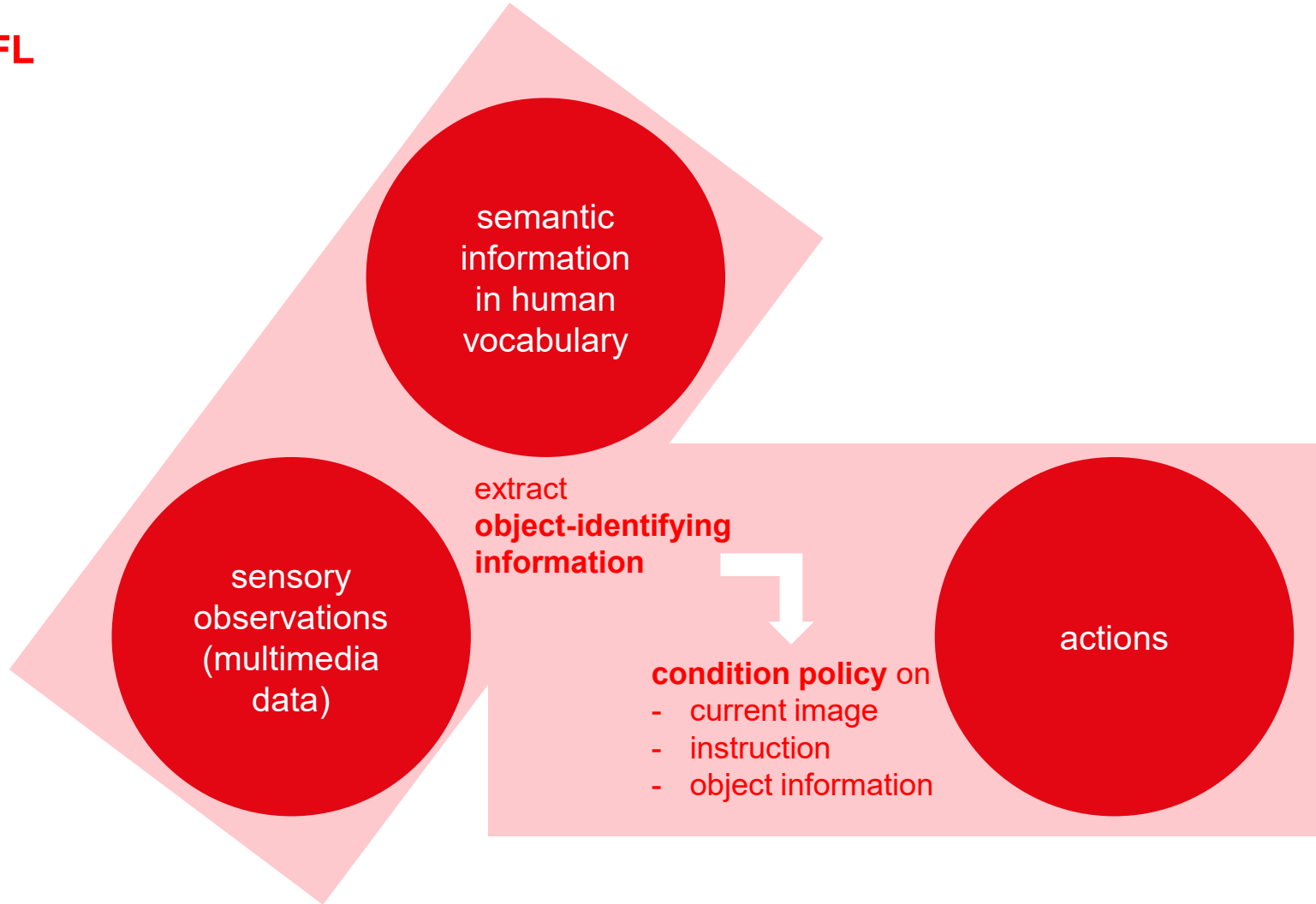
doi.org/10.5281/zenodo.4518950
doi.org/10.5281/zenodo.5525332
doi.org/10.5281/zenodo.4770061
doi.org/10.5281/zenodo.4518951

CORSMAL Challenge and Evaluation Toolkit

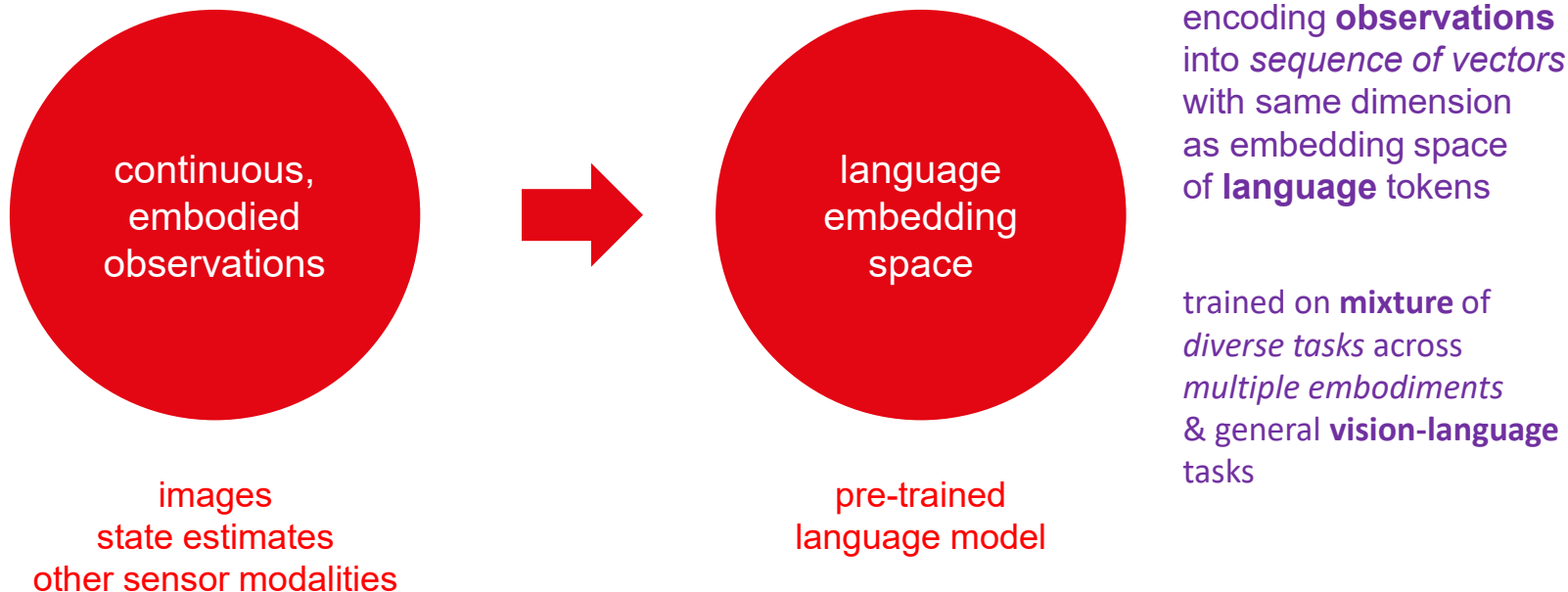
<https://corsmal.github.io/challenge>
github.com/CORSMAL/CORSMALChallengeEvalToolkit

<https://corsmal.github.io>





Embodied multimodal language model



Natural language to condition perception



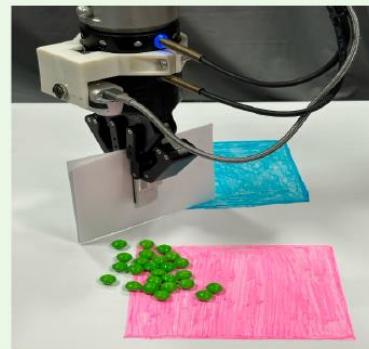
put the red blocks
in a brown bowl



pack the scissors
in the brown box



pack all the silver tape
in the brown box

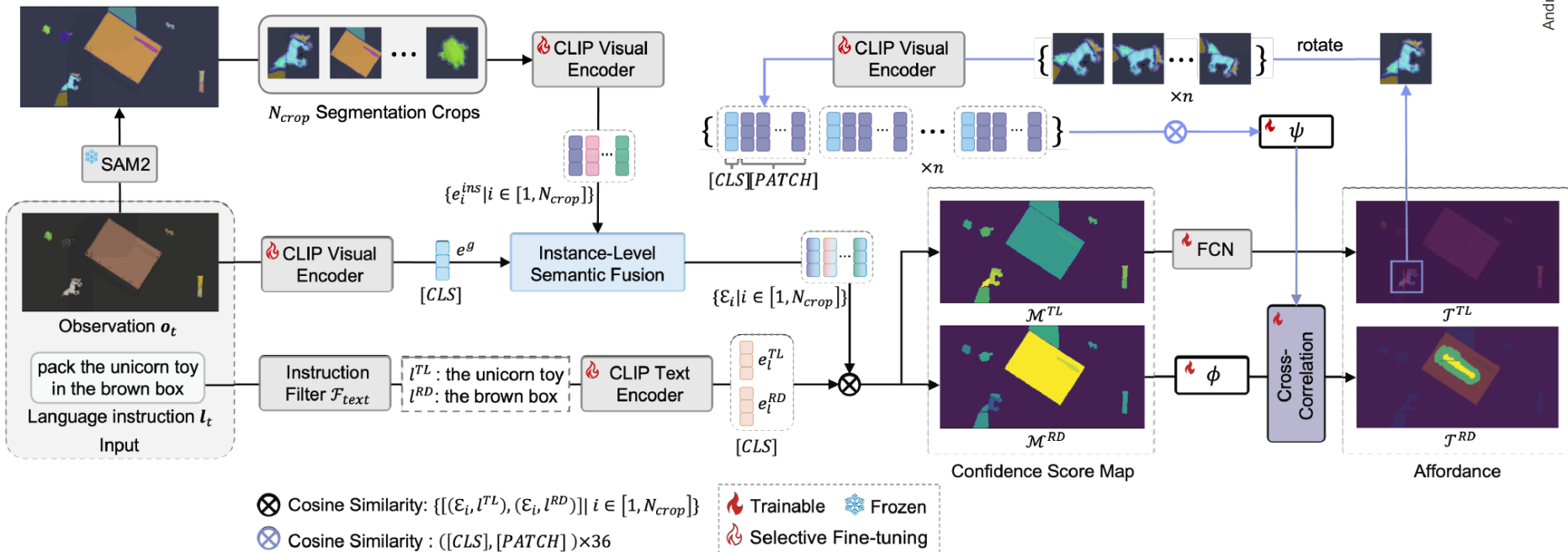


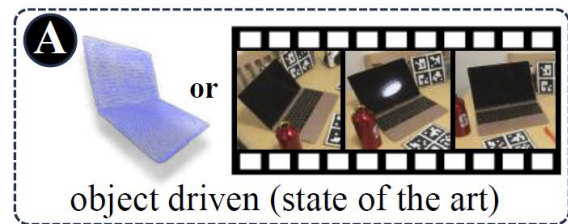
push the pile of green blocks
into the blue square

Improving generalization of language-conditioned robot manipulation

Cui, Zhu, Oh, Cavallaro. *IRIS 2025*

Natural language to condition perception





Textual prompt:
guides the *pose estimation* process

Reference:
single view

Anchor



Ground truth



SOTA1



SOTA2



Ours



(a) Prompt: Black lens camera

Anchor



Ground truth



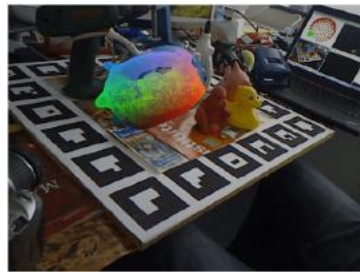
SOTA1



SOTA2



Ours



(c) Prompt: Light blue iron



(d) Prompt: Blue plastic water jug

High-resolution open-vocabulary object 6D pose estimation

Corsetti, Boscaini, Giuliani, Oh, Cavallaro, Poiesi. *CVPR 2024 + T-PAMI (under review)*

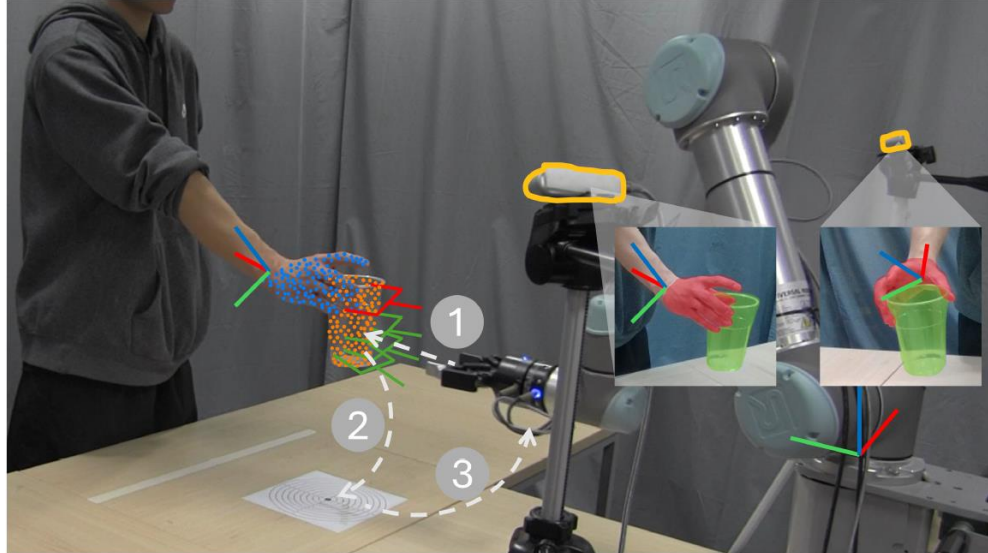
EPFL What's next?

Interoperability

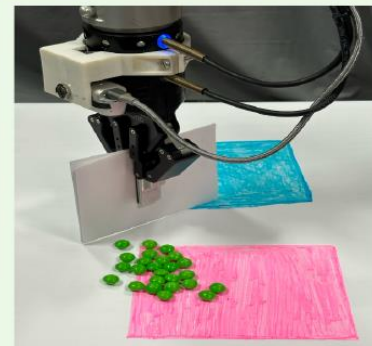
Transparency

Traceability

Compliance with fundamental rights and local norms when utilizing information extracted from multimedia data



pack all the silver tape
in the brown box



push the pile of green blocks
into the blue square

THANK YOU!

Summary

Human-object-robot interaction

Reproducibility: protocol, data, models

Embodied multimodal language models

Natural language to condition perception