ロボットチャレンジと産業の接点を探る

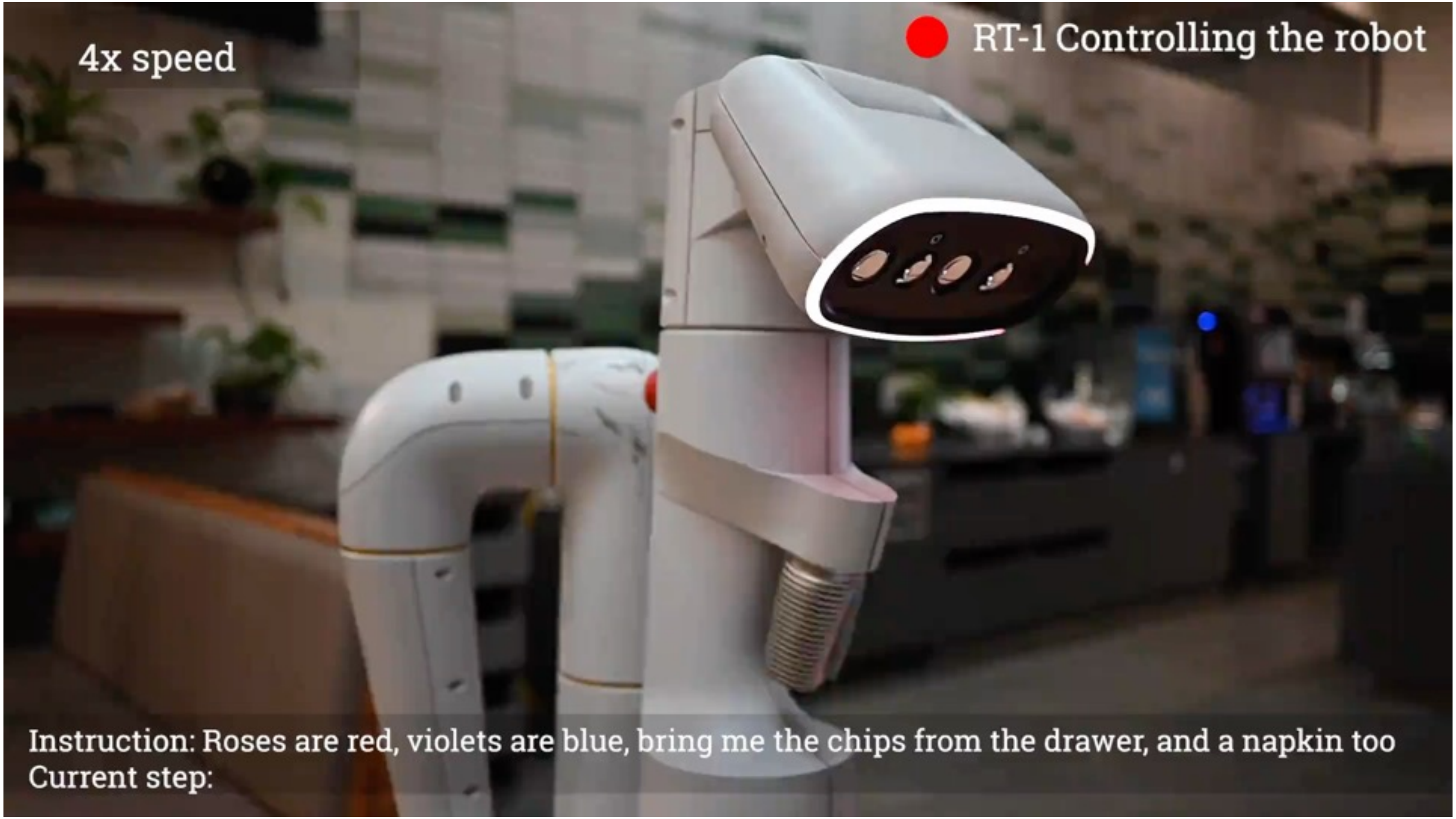
堂前幸康，産業技術総合研究所

マニピュレーション委員会・第一回シンポジウム
ロボット革命・産業IoTイニシアティブ協議会

The impact of foundation models in robotics

基盤モデルのインパクト
ーロボティクスにおける最新のチャレンジー

Robots meet Large-scale Language-Vision-Action models

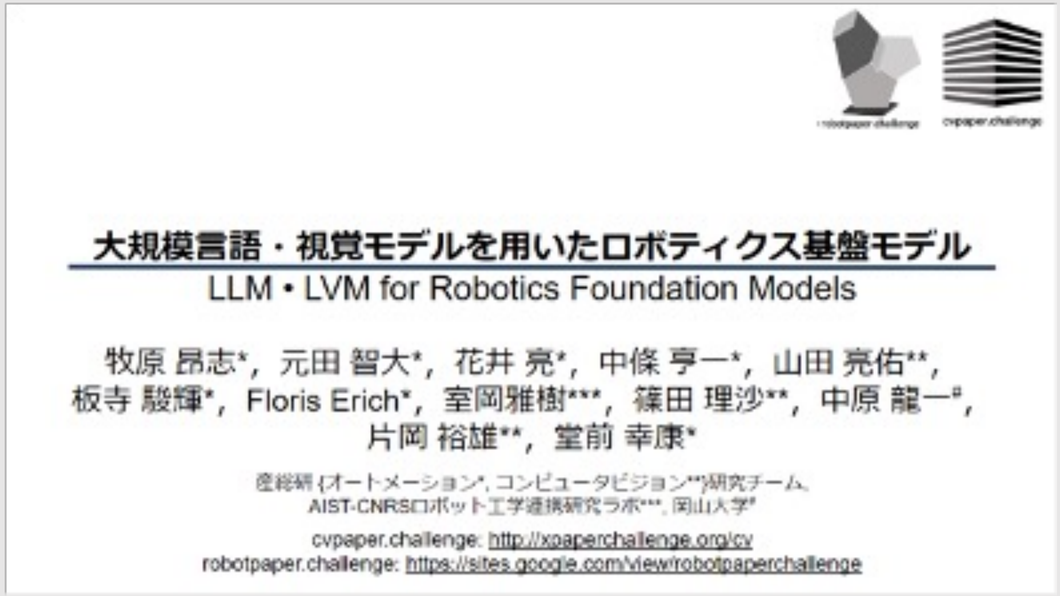


[RT-1: Robotics Transformer (robotics-transformer.github.io)](robotics-transformer.github.io)

Meta-survey of LLM, LVM for robotics

Posted on 22 June 2023.

95.3K Views (Jun, 2024)



全168ページ

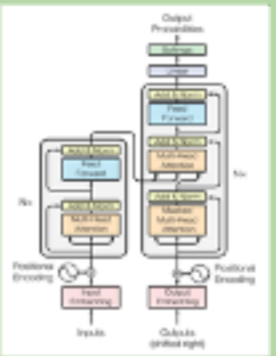
1. **イントロダクション**　page. 1 - page. 44
   → Transformer以前からCV/Robotics分野の歴史を紹介。直近で話題になったGPT-4についても紹介する。
2. **論文紹介**　page. 45 - page. 129
   → 大規模言語・視覚モデル分野で着目すべき論文をスライド１枚分に要約。2023年6月時点で最新の情報を提供する80本以上の論文が掲載されている。
3. **メタサーベイ**　page. 130 - page. 157
   → ロボティクスを議論の中心として、世界の研究者がどのような「戦略」をとっているのかについて、過去の研究からのトレンドに至るまでの流れを踏まえつつ、議論を展開している。
4. **著者紹介**　page. 158 -

NLP, CV, V&L after the introduction of Transformer

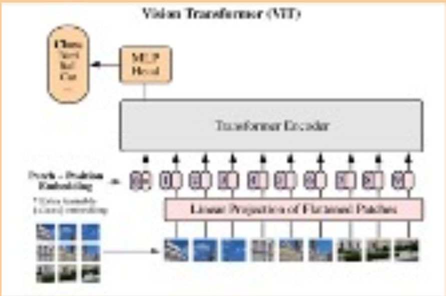
Transformer [Vaswani+, 2017]

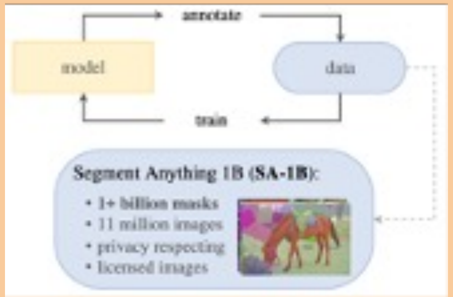
Transformers utilize only **attention mechanisms**. They have demonstrated superior performance compared to traditional models such as recurrent or convolutional models.

NLP

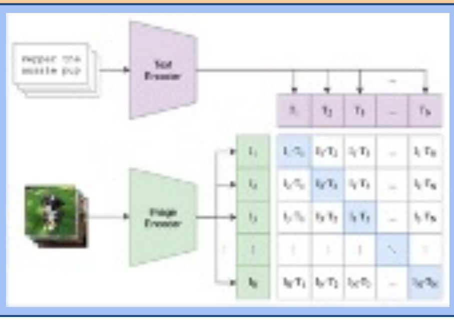


CV





Vision and Language



Cross-modalization

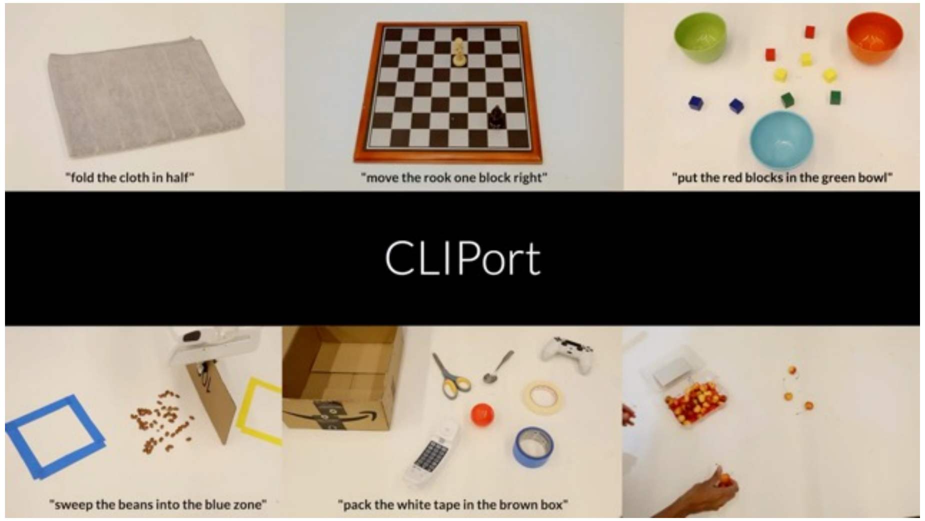
2017 → 2024

Large-scaleization

Robotic applications of Language-Vision Models 1

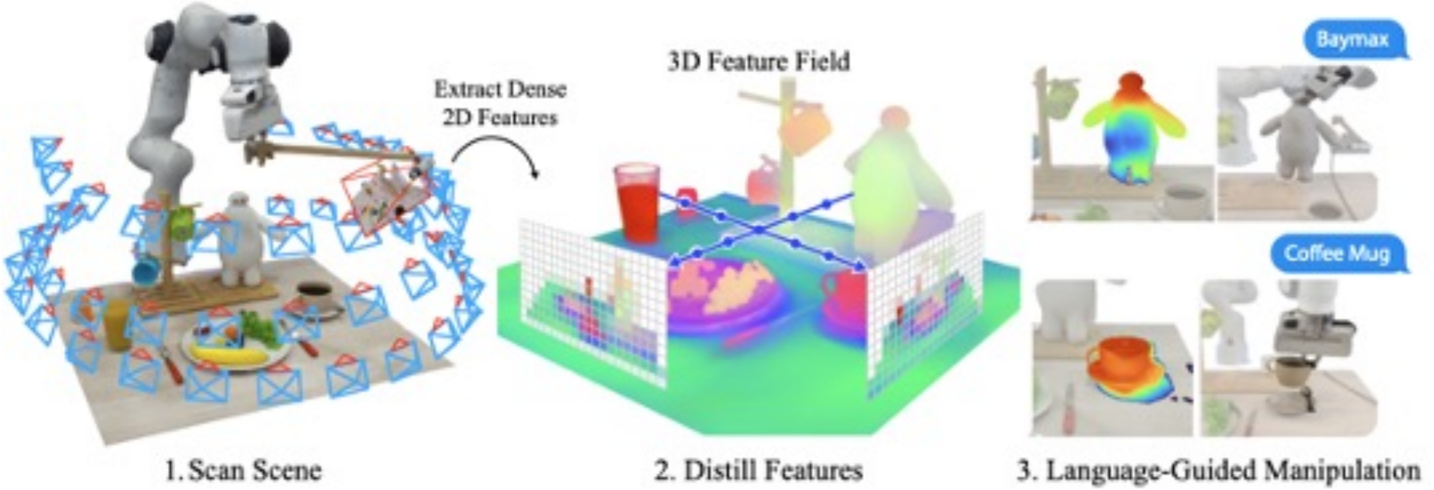
CLIP (Contrastive Language-Image Pre-Training) is a neural network model trained on various pairs of text and images. It can be applied to various tasks in a zero-shot manner. There are already numerous applications of CLIP in robotics.

CLIPort [Shridhar+, CoRL2021]



Prediction of affordances in robotic task by CLIP.

Distilled Feature Fields [W. Shen+, CoRL2023]



Designing a feature space that connects 2D image features to 3D geometry, enabling few-shot language-to-6DoF grasping.

Robotic applications of Language-Vision Models 2

Embedding knowledge into 3-D representations

LERF: Language Embedded Radiance Fields[Kerr+, 2023]

https://www.lerf.io/

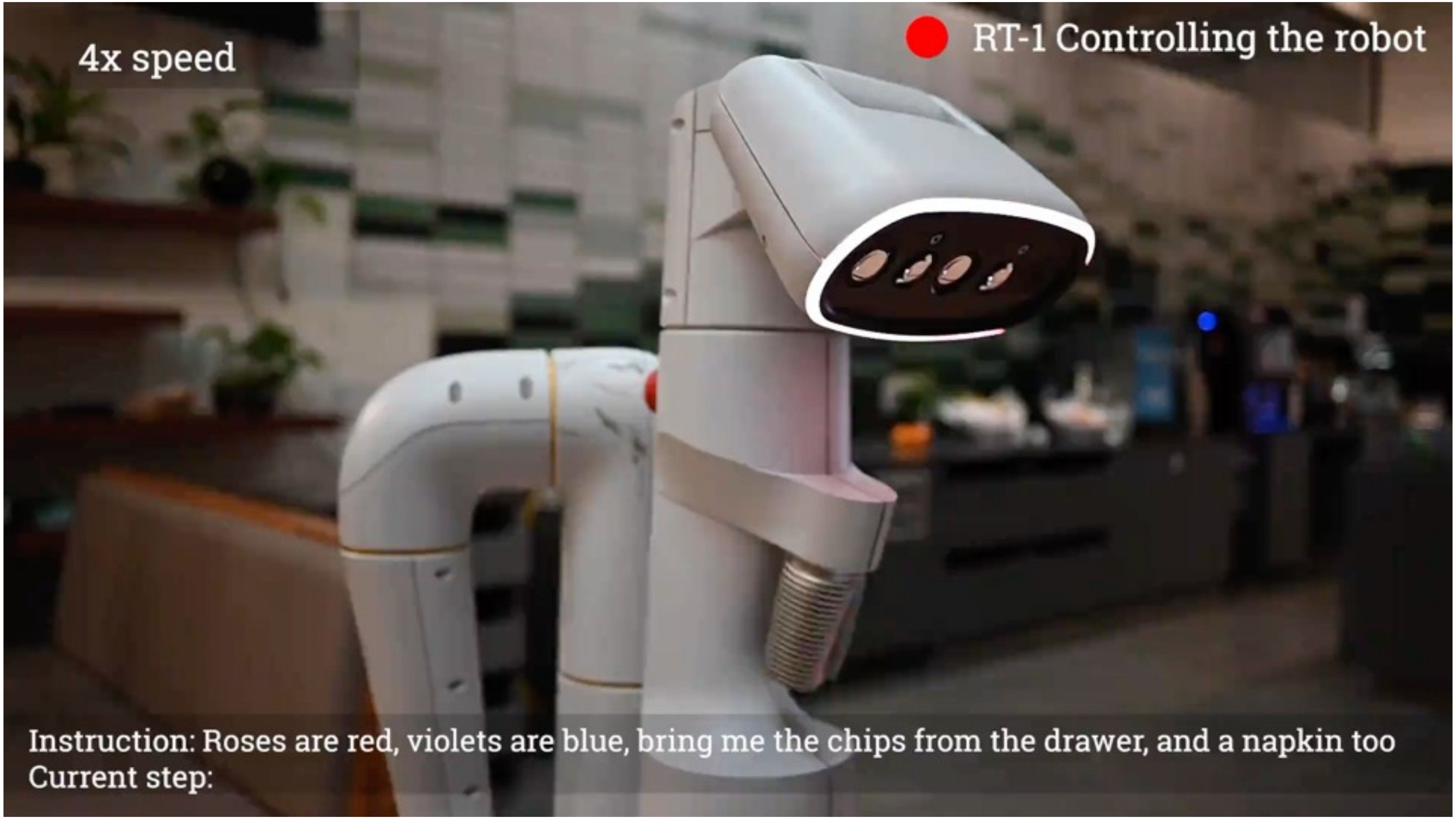
Connecting NERF and CLIP to embed text in a three-dimensional space.

LLM-Grounder [Yang, 2024]

[Chat with NeRF (chat-with-nerf.github.io)](https://chat-with-nerf.github.io)

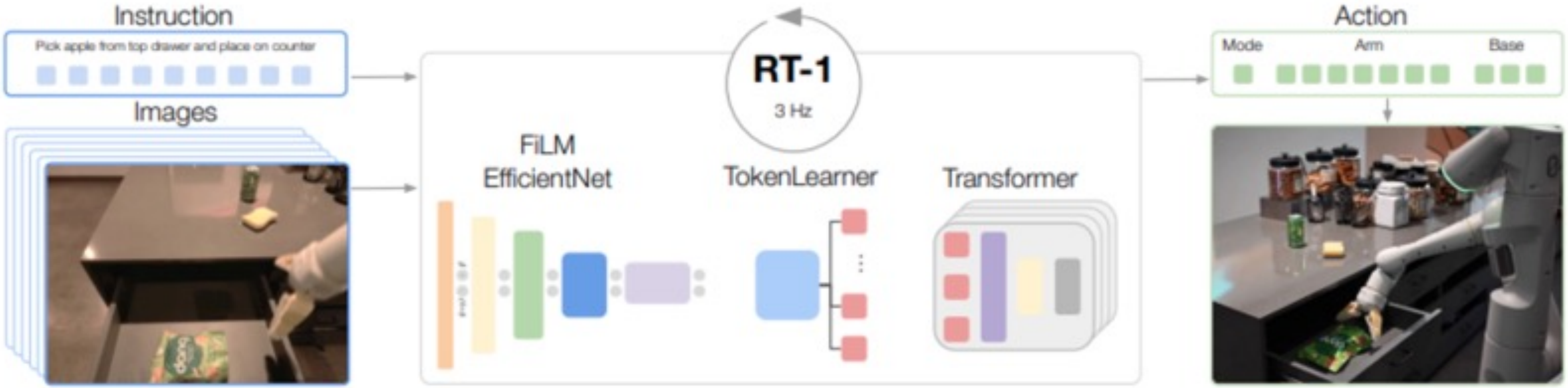
Combining the technologies of GPT-4/LLaVA/BLIP-2/NeRF Studio/LERF enables dialogue about 3D scene environments/objects.

Language-Vision-Action model

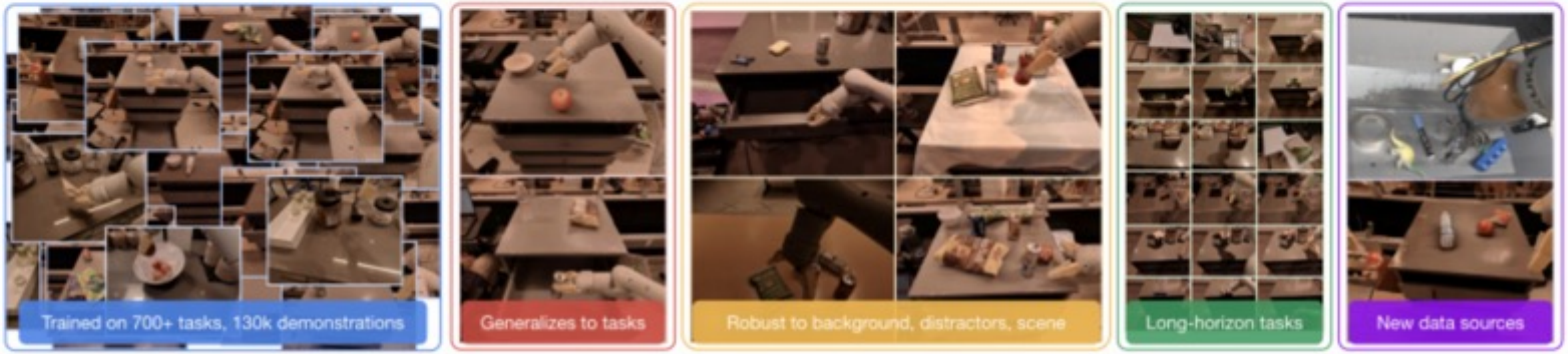


[RT-1: Robotics Transformer (robotics-transformer.github.io)](robotics-transformer.github.io)

RT-1: Robotics Transformer (robotics-transformer.github.io)



(a) RT-1 takes images and natural language instructions and outputs discretized base and arm actions. Despite its size (35M parameters), it does this at 3 Hz, due to its efficient yet high-capacity architecture: a FiLM (Perez et al., 2018) conditioned EfficientNet (Tan & Le, 2019), a TokenLearner (Ryoo et al., 2021), and a Transformer (Vaswani et al., 2017).



(b) RT-1's large-scale, real-world training (130k demonstrations) and evaluation (3000 real-world trials) show impressive generalization, robustness, and ability to learn from diverse data.

Large-scale real-world data

130,000 robot action sequences data

10 robots x 17 months

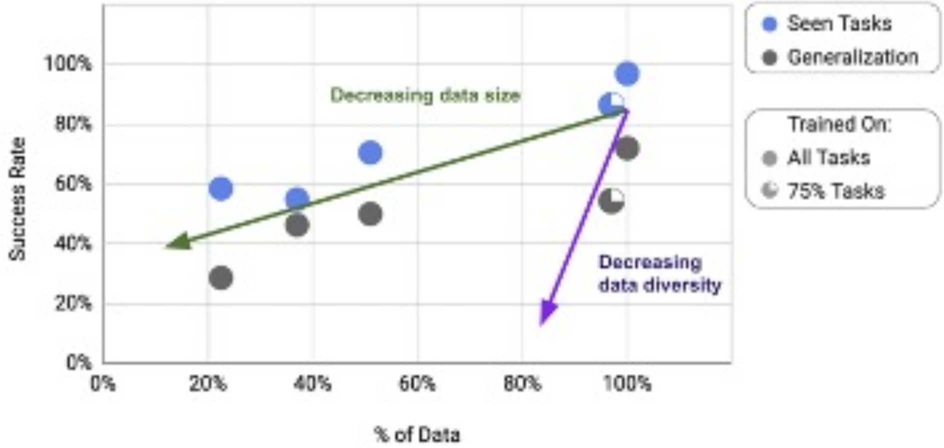
The necessity of large-scale data in foundational models

The scaling law [Kalpan, 2020]

The scale of data and the performance of the model are directly proportional.

Even in robotics, this law has begun to be demonstrated **[Padalkar, 2023]**.

Models	% Tasks	% Data	Seen Tasks	Generalization			
				All	Unseen Tasks	Distractors	Backgrounds
Smaller Data							
RT-1 (ours)	100	100	97	73	76	83	59
RT-1	100	51	71	50	52	39	59
RT-1	100	37	55	46	57	35	47
RT-1	100	22	59	29	14	31	41
Narrower Data							
RT-1 (ours)	100	100	97	73	76	83	59
RT-1	75	97	86	54	67	42	53



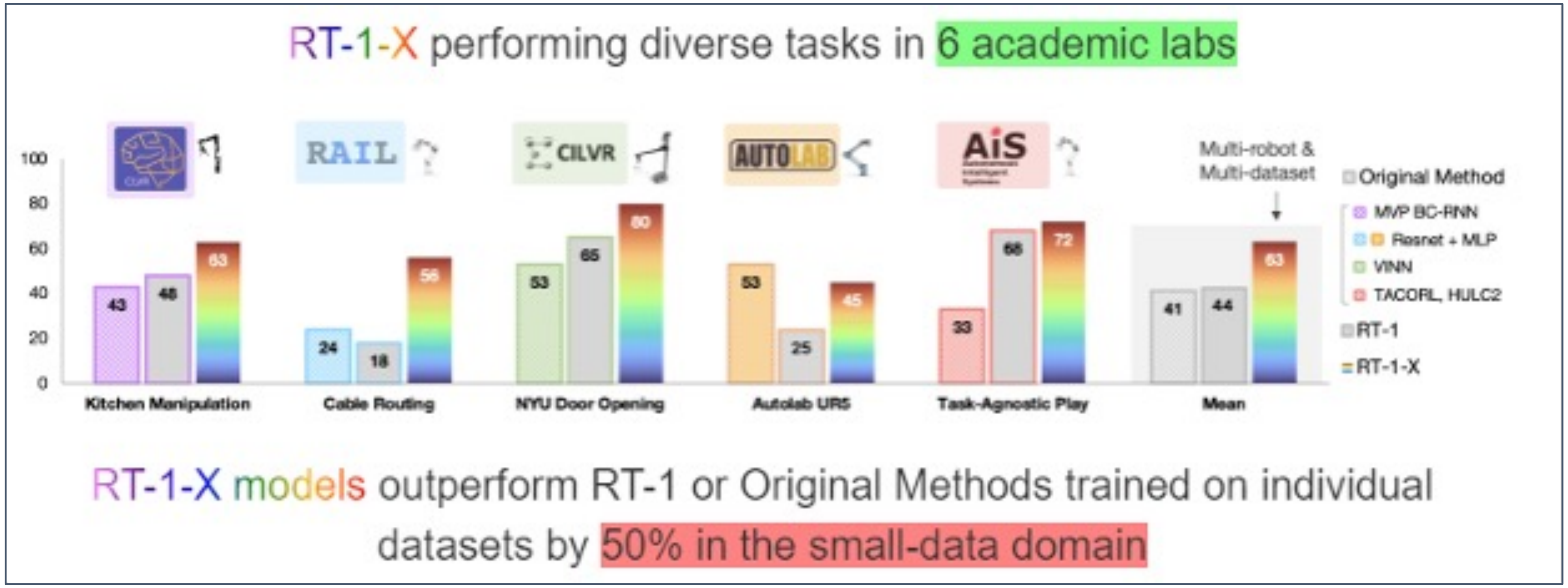
Open-X-Embodiment [Padalkar, et al, 2023]

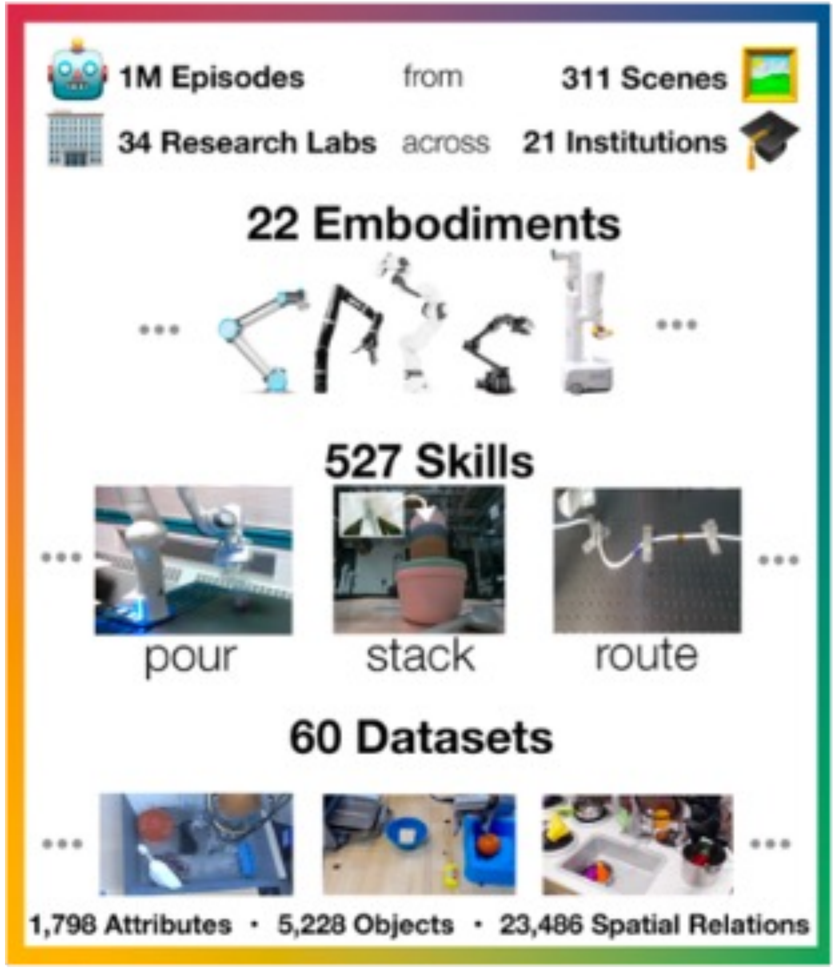
- Increasing the amount of data improved generalization performance
- Task success rates improved as data diversity increased

the more tasks a robot learns,
the better it becomes at performing its job.

Collecting real-world large-scale data through brute force methods

Open-X-Embodiment:
1M episodes, 527 skills (**160,266 tasks**)
22 different robots, 21 institutions, 60 datasets

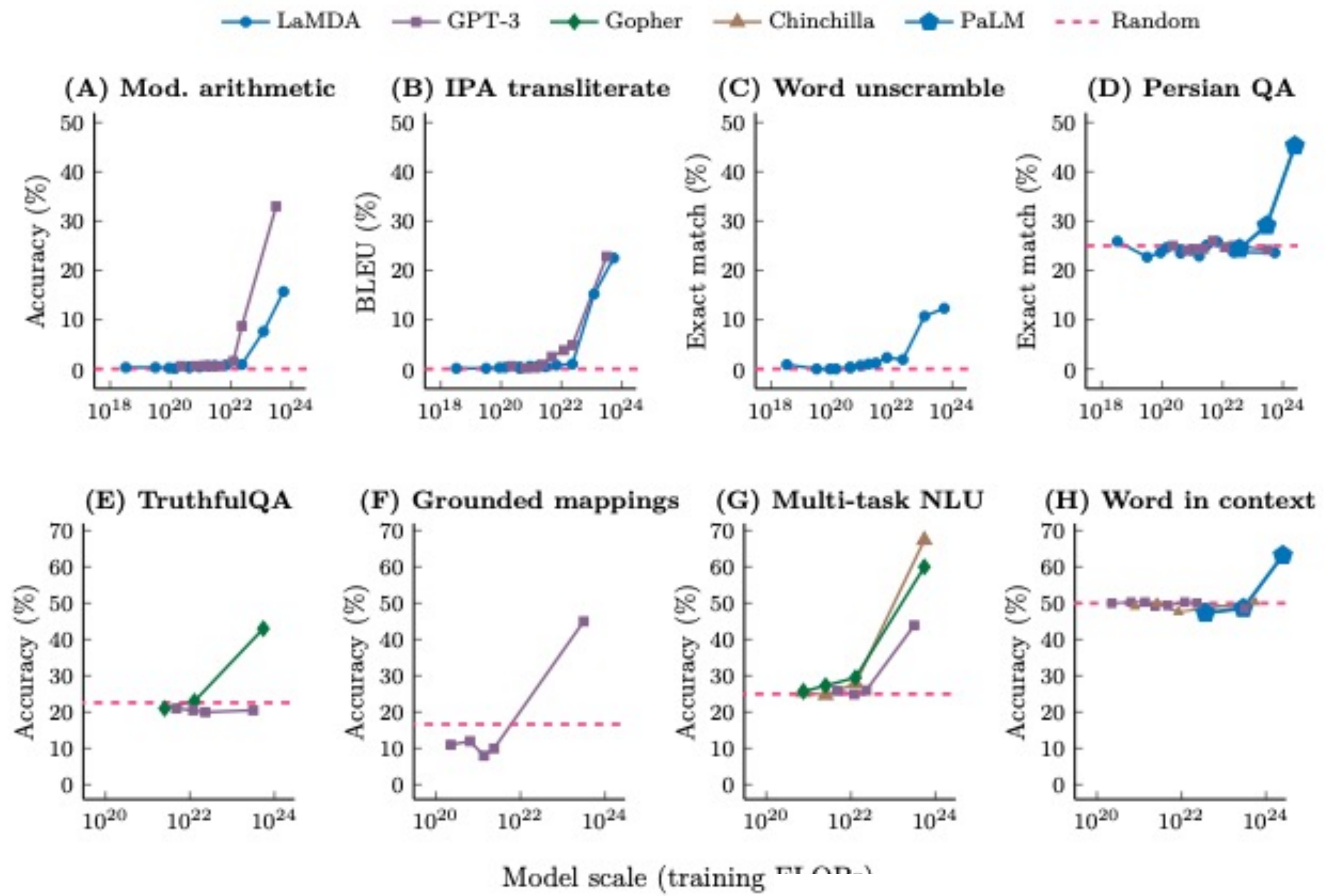




Open-X-Embodiment [Padalkar, et al, 2023]

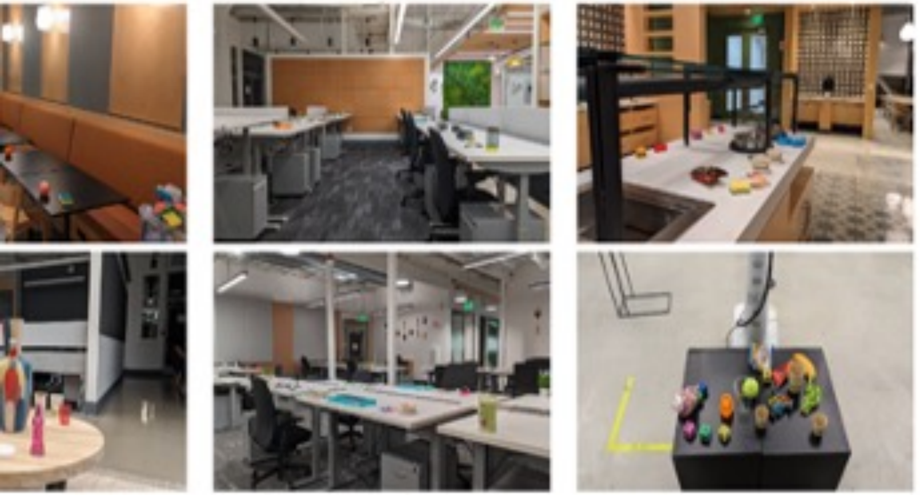
https://robotics-transformer-x.github.io/

Emergent Abilities of Large Language Models [Wei 2022]



Collecting real-world large-scale data through brute force methods

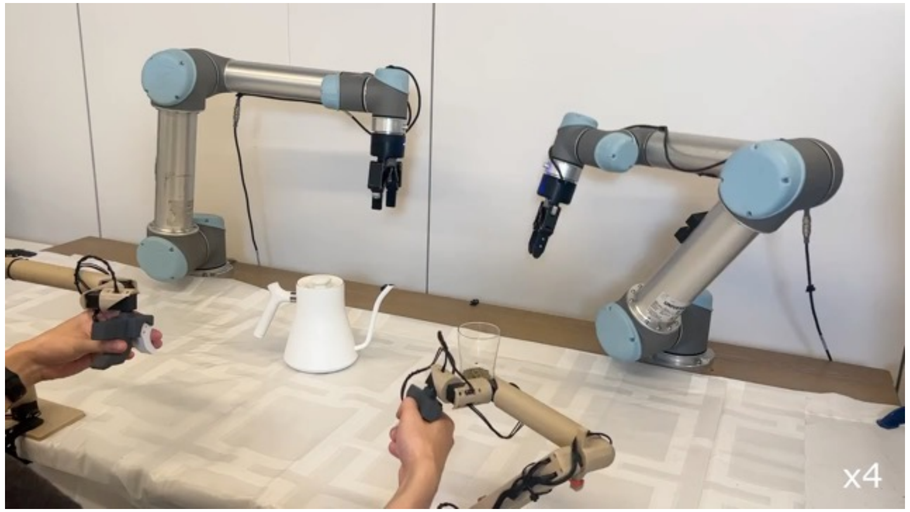
AutoRT: **77K episodes**,
6650 unique language instructions,
20 robots, 7 months

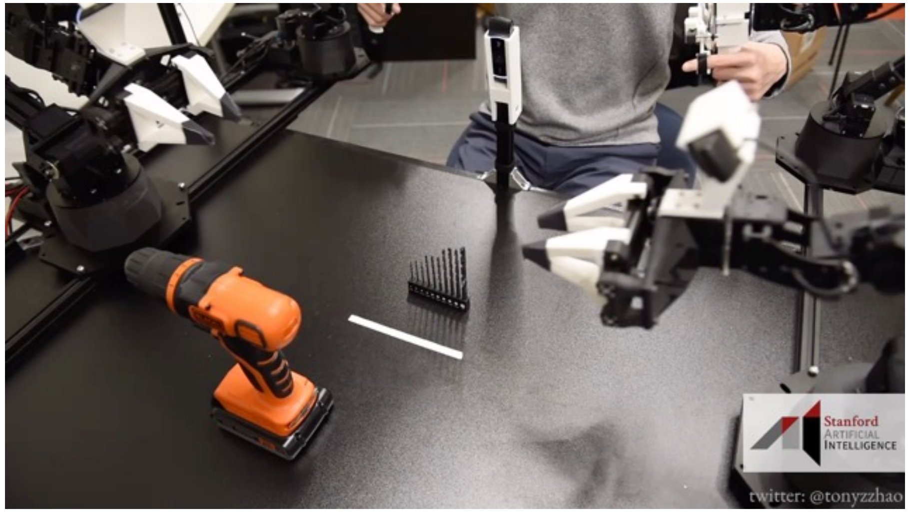


AutoRT [Google DeepMind, 2023]　　https://auto-rt.github.io/

Transferring human behaviors to robots

Leader-follower systems have emerged with **low cost, intuitiveness, and responsiveness**.



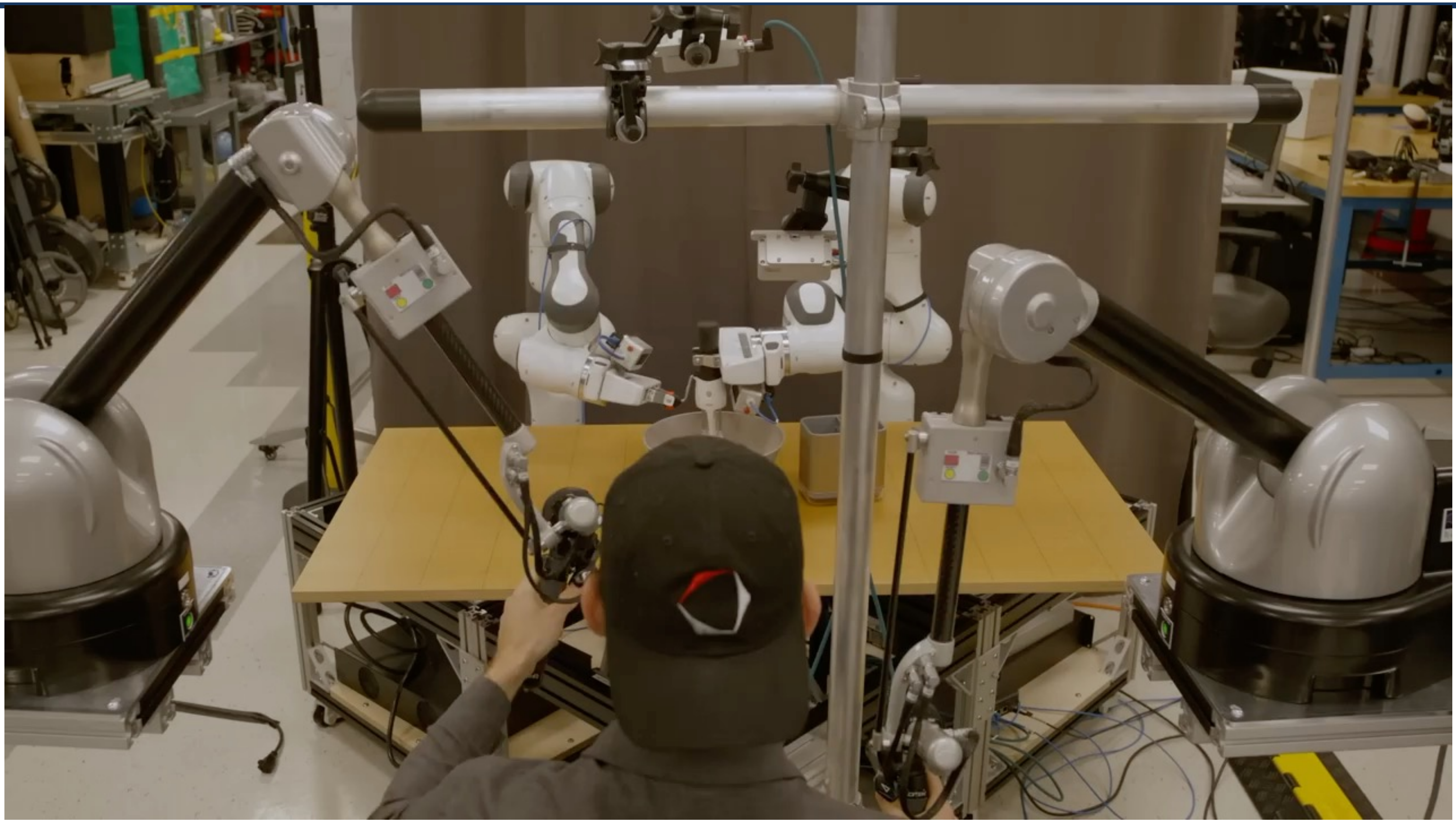


Mobile ALOHA [Z. Fu+, 2024]
https://mobile-aloha.github.io/

GELLO [P. Wu+, arXiv 2023]
https://wuphilipp.github.io/gello_site/

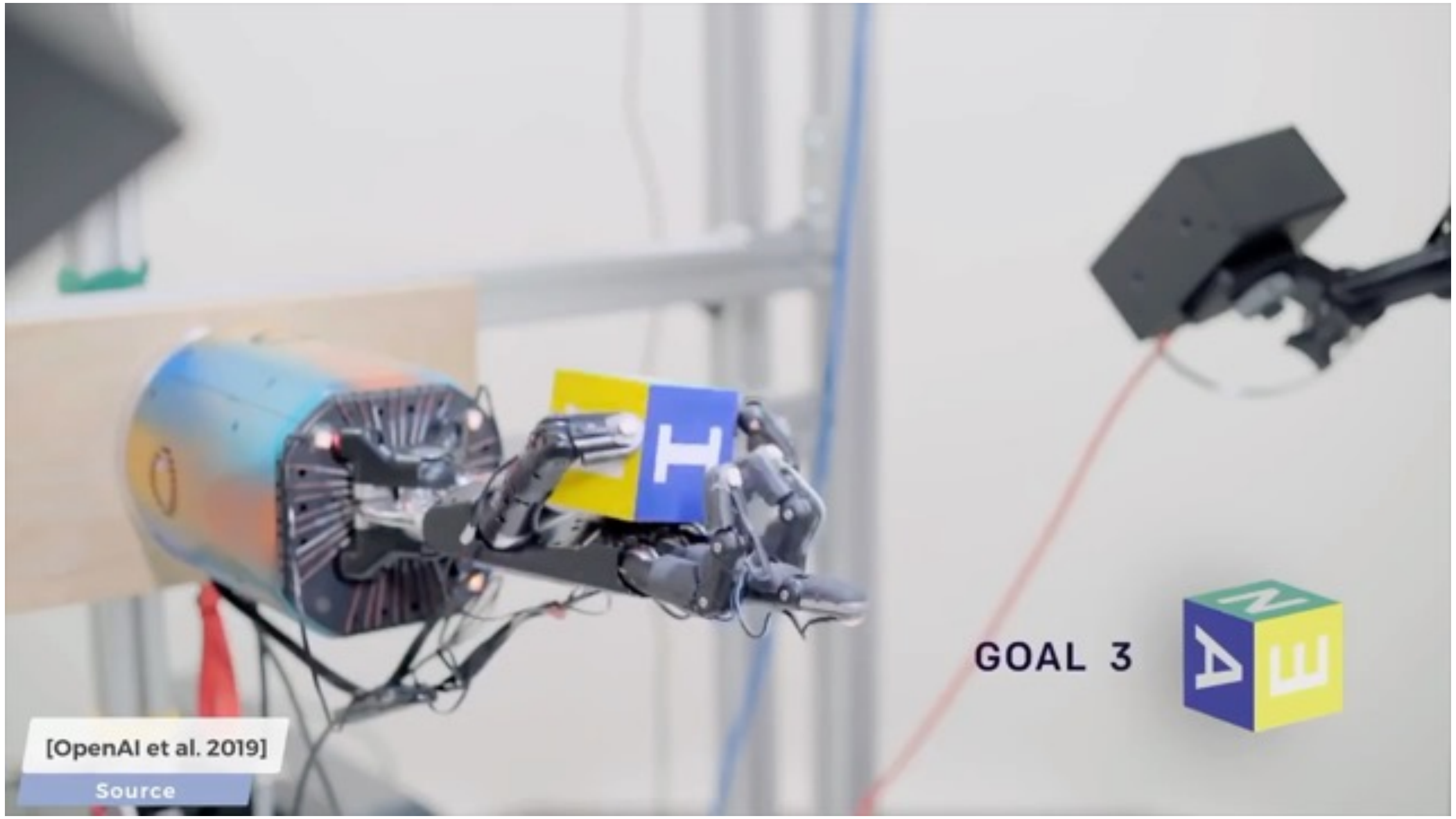
ALOHA [T. Shao+, RSS2023]
https://tonyzhaozh.github.io/aloha/

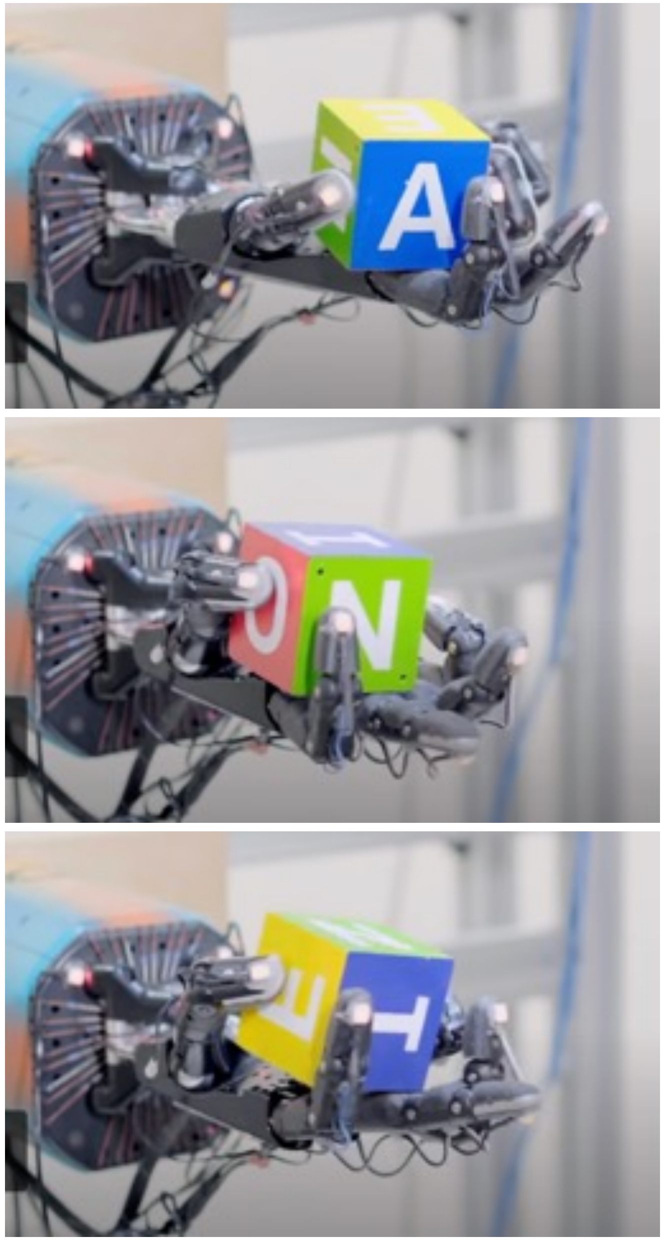
Imitation of human dexterity



Teaching Robots New Behaviors [TRI, Youtube 2023]

Randomization of “motion”





Randomize parameters related to the **physical dynamics** such as mass and dimensions of objects or robots, friction, control gains (PID), observation noise, joint constraints, etc.

[1808.00177] Learning Dexterous In-Hand Manipulation (arxiv.org)

Generative AI augments robot's experience

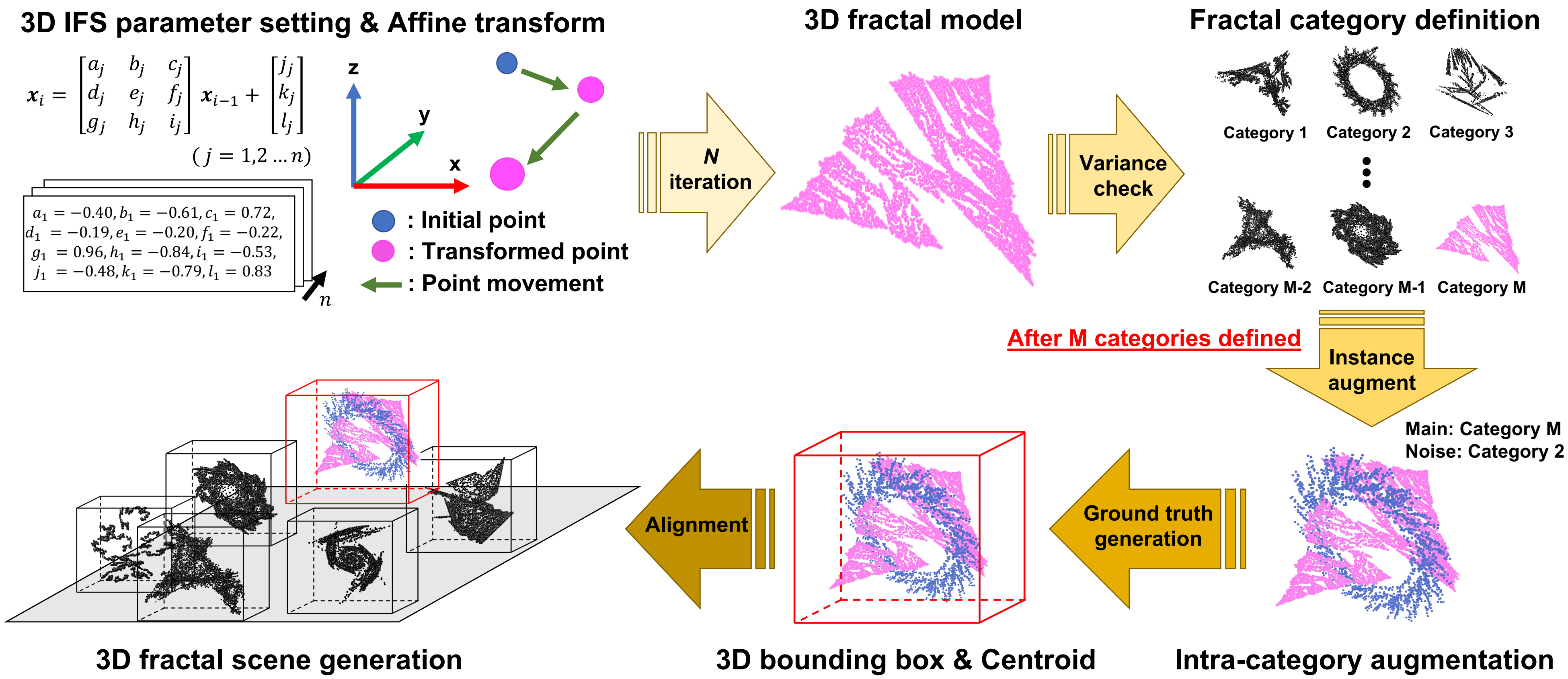
Unreal experience!

13 robots,17 months,
to collect 130k demonstrations

+ Generative models?

[Scaling Robot Learning with Semantically Imagined Experience (diffusion-rosie.github.io)](diffusion-rosie.github.io)

Randomization of object shapes based on mathematical equations



Unreal objects!

Randomly generate 3D shapes using fractals and place them in the scene. Utilizing a small amount of data for pretraining improves the performance of object detection (by VoteNet) from 3D point clouds.

Ryosuke Yamada, et el., “Point Cloud Pre-training with Natural 3D Structure”, CVPR 2022

cvpaper.challenge

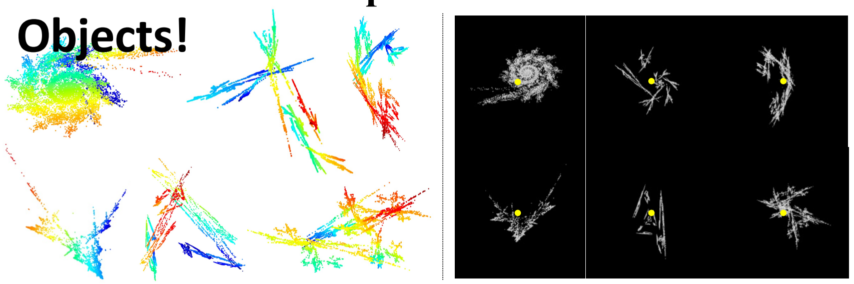
Randomization of object shapes using mathematical equations

Pretraining with object database generated by fractals

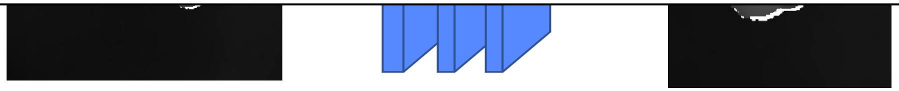
Grasp-FractalDB Pre-training

Unreal Objects!

Grasp-FractalDB



Dex-Net 2.0 Fin



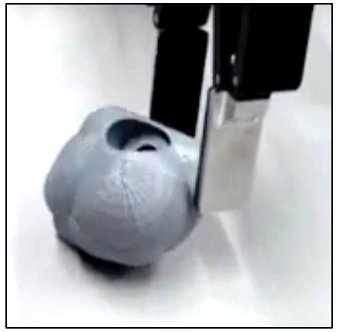
Depth image

GQ-CNN

Estimate grasp quality for parallel jaw gripper

Grasping experiments

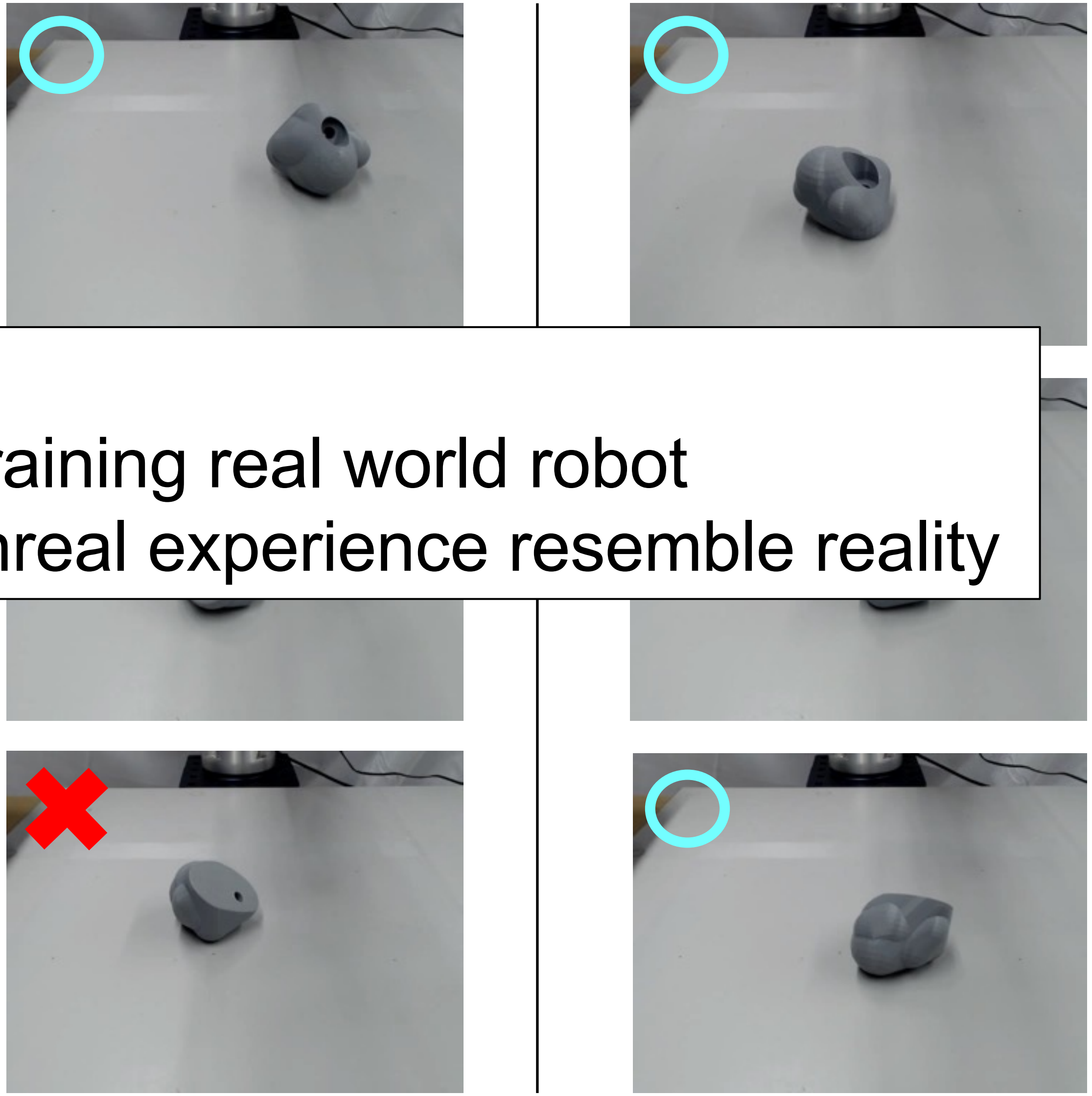
Before(far from CoM)



with Grasp-FractalDB →

After (more robust grasp)

Dex-Net 2.0 | Grasp-FractalDB (Ours)



In some cases,

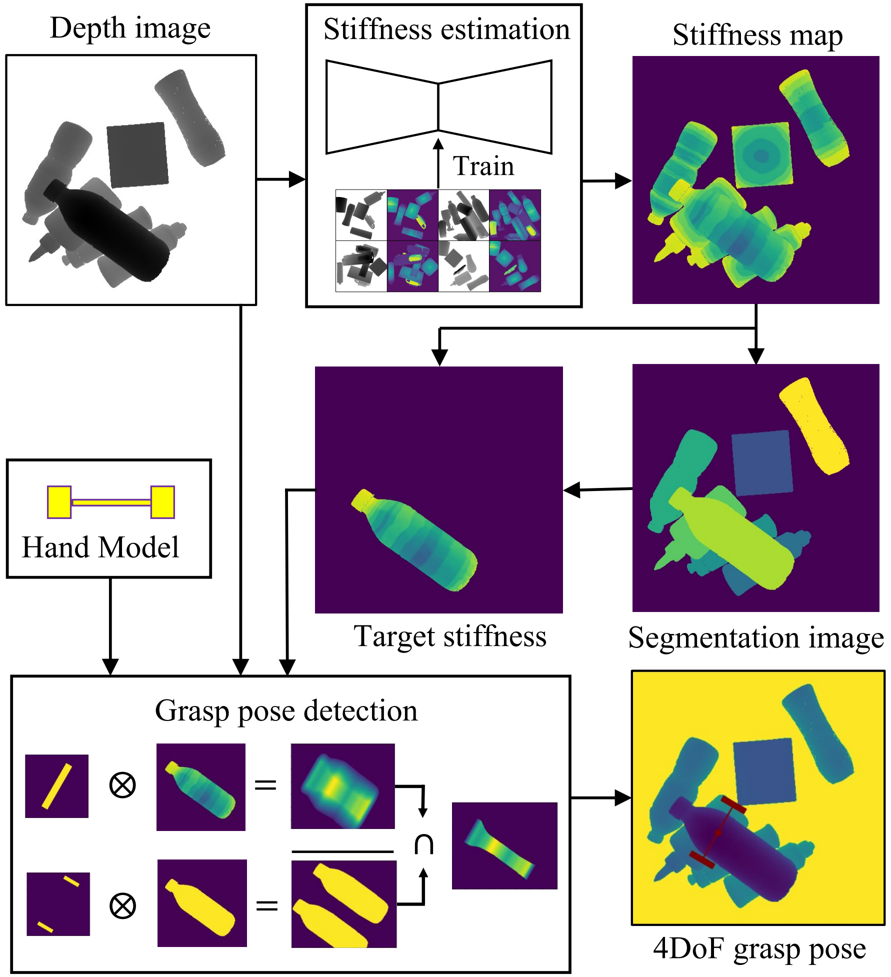
- we don’t need real world data for training real world robot
- we don't have to try to make the unreal experience resemble reality

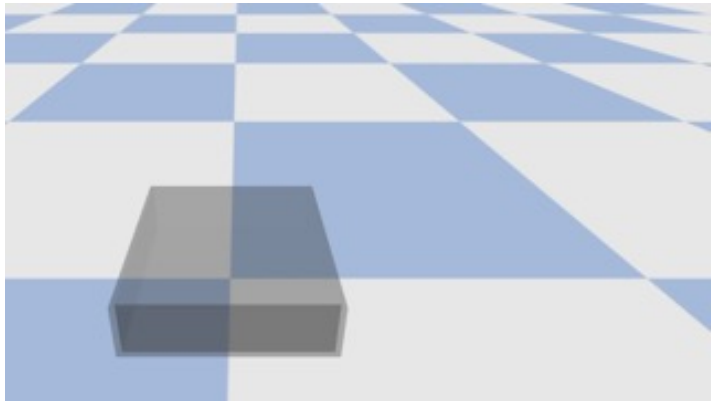
Yamada, SSII2022

Acquisition of cross-modal ability from unreal experience

Learning the relationship between vision(depth) and **object softness** through simulation

Picking by visual information
- Missing to grasp soft objects.
- Picking errors

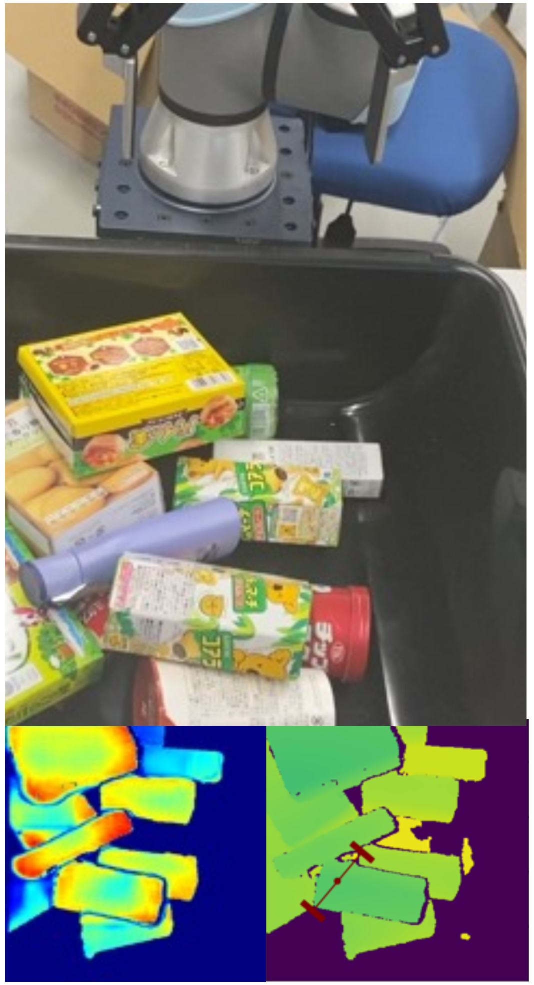




Experiecne augmentation based on "depth-to-softness" simulation

K. Makihara, et. al., "Grasp pose detection for deformable daily items by pix2stiffness estimation", Advanced Robotics, 2022.

Acquisition of cross-modal ability from unreal experience



The robot "pushes aside soft objects" to pick up the target.

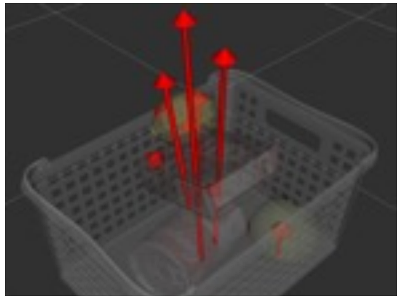
Without cross-modal ability

With cross-modal ability

K. Makihara, et. al., "Grasp pose detection for deformable daily items by pix2stiffness estimation", Advanced Robotics, 2022.

Acquisition of cross-modal ability from unreal experience

simulation



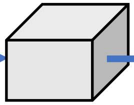
- Domain Randomization

- Kernel Density Estimation
- moving average on time

Learning the relationship between vison and “**forces acting between objects**” through simulation.

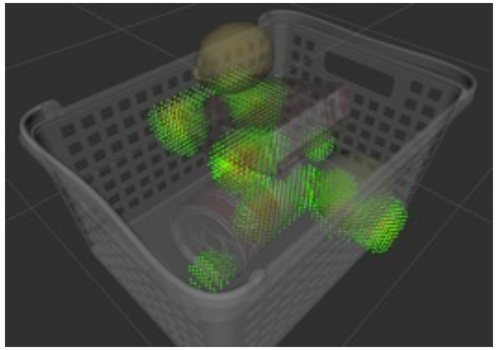
ResNet50

Encoder



ResNet based Decoder

Decoder



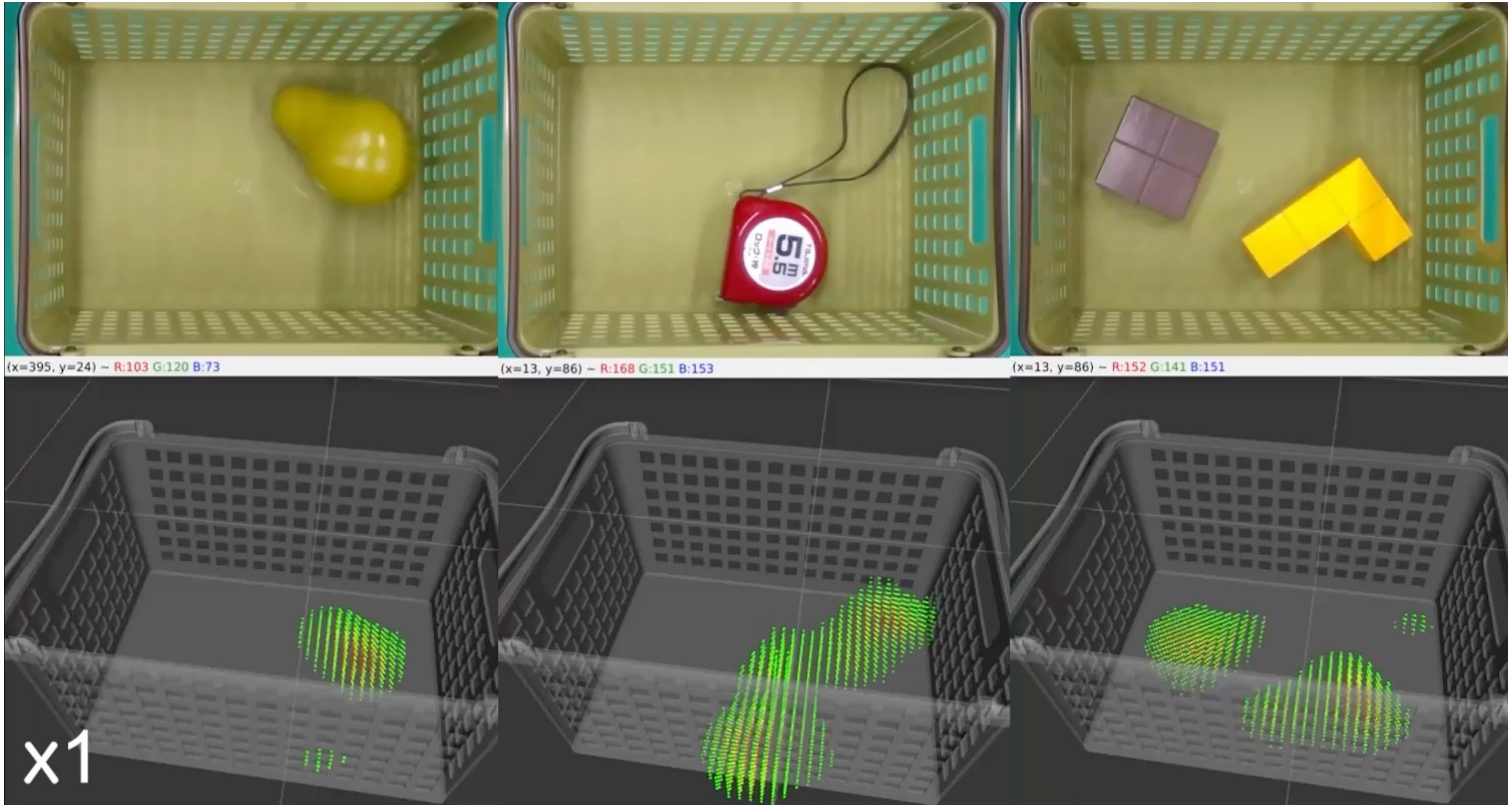
Forcemap
(contact force label)

[Hanai, IROS2023]

[Force Map: Learning to Predict Contact Force Distribution from Vision (ryhanai.github.io)](ryhanai.github.io)

Acquisition of cross-modal ability from unreal experience

We have successfully achieved real-time visualization of the forces between objects in 3D, using only a single RGB image

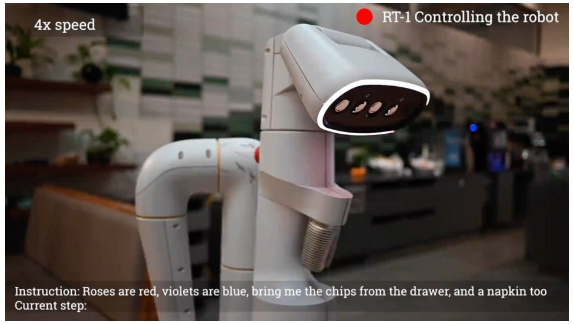


[Force Map: Learning to Predict Contact Force Distribution from Vision (ryhanai.github.io)](https://ryhanai.github.io) (R.Hanai, IROS2023)

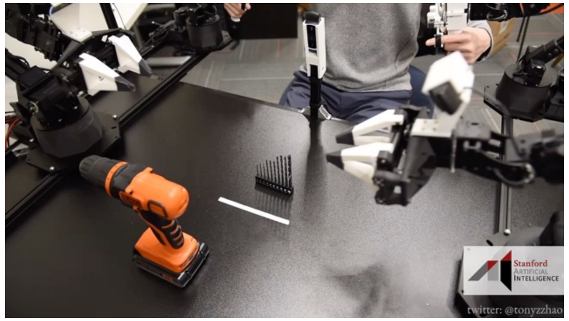
Conclusion : Learning from reality / unreality

Learning from reality

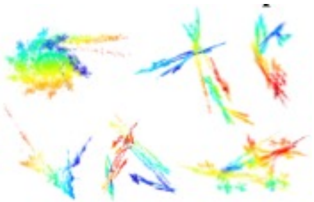
Accurate experience



Imitation of human behaviors



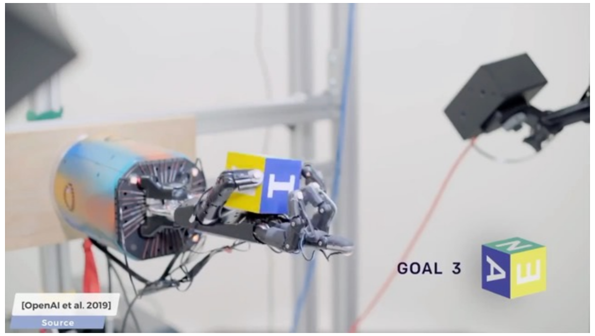
Domain gap
- Randomization
- Curriculum

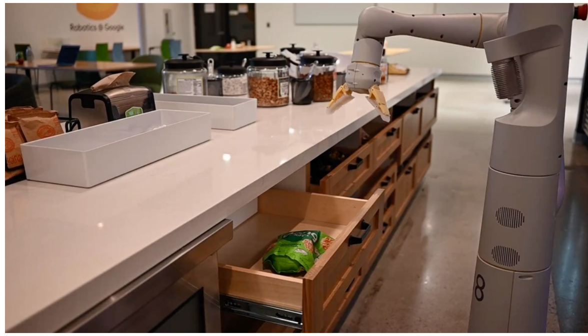


There are cases where it's not necessary to fill the gap!

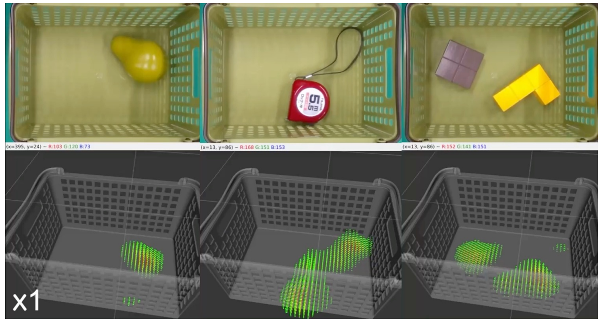
Learning from unreality

Experience augmentation





Acquisition of cross-modal ability



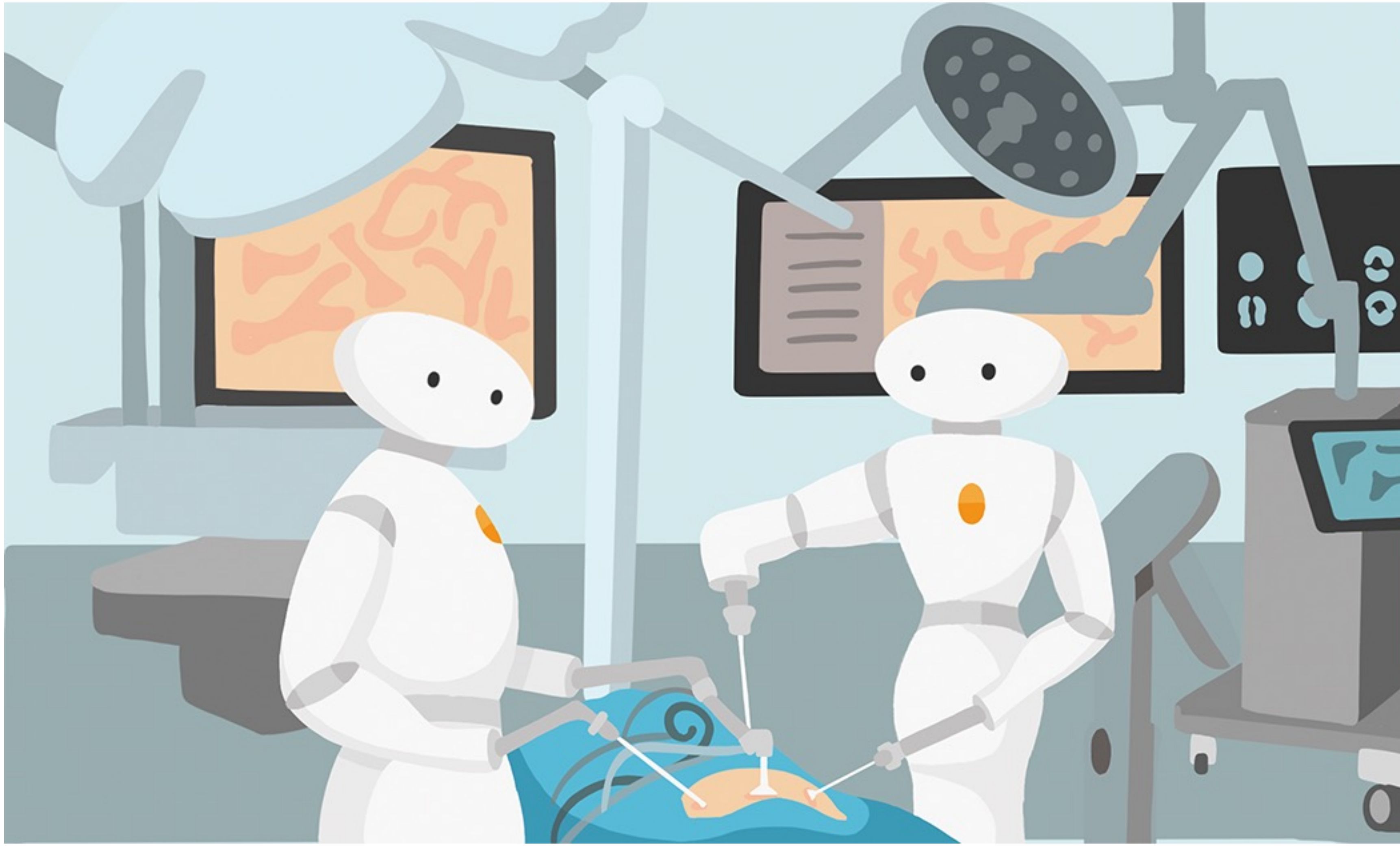
Future : Professional robot learned from experiences beyond reality

Learning from unreality

- Incidents that are unlikely to occur in reality

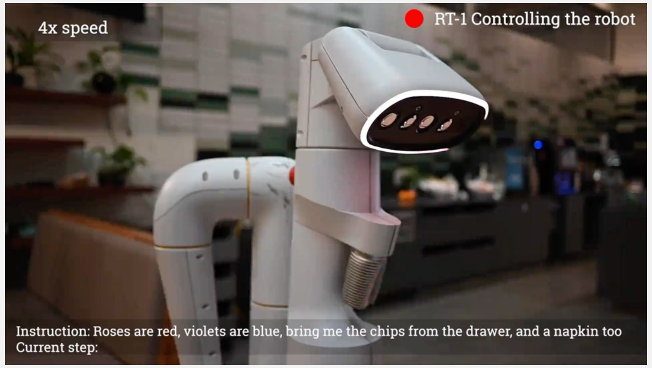
- Parameter randomization over a **slightly** broader range than reality

  Example:
  Experience in surgical vascular anastomosis slightly narrower than reality.



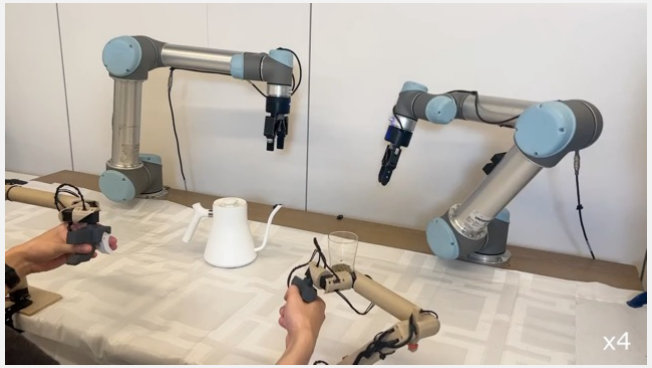
AIREC MOONSHOT

Next step : physical foundation models



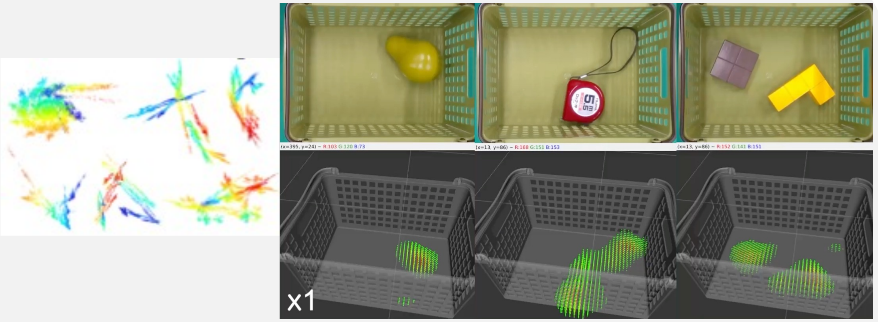
Vision-Language-Action model

+

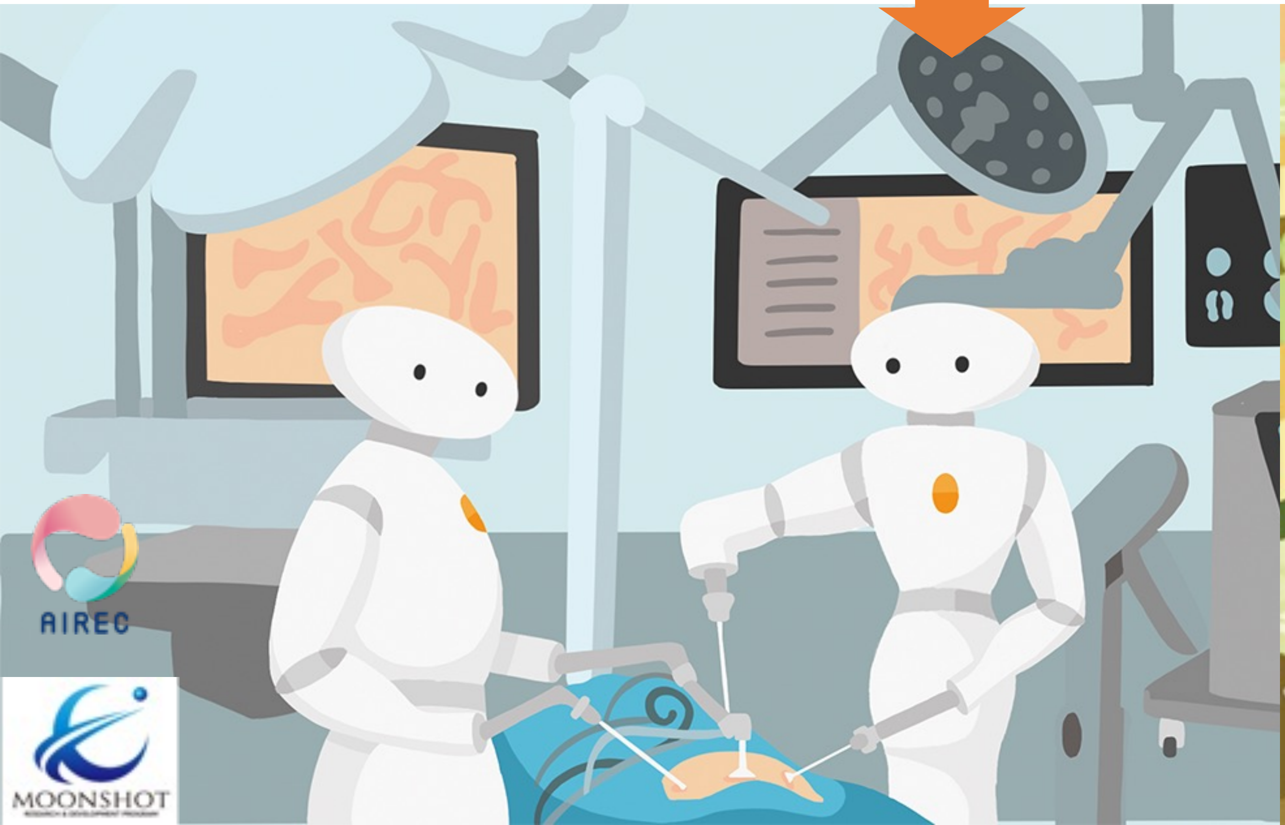


Dexterity

+



Unreal physical experience



Thank you for your attention! Special thanks to

Automaton Research Team, AIST

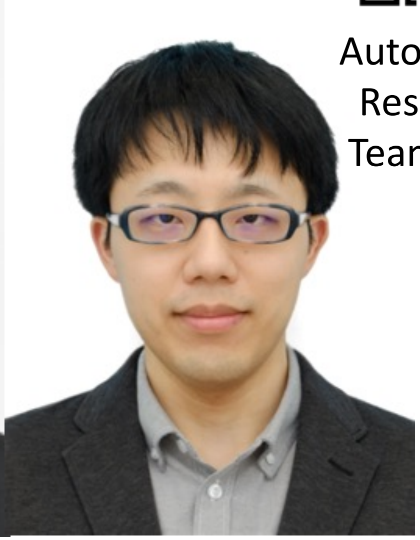
Ryo Hanai

Ixchel Ramirez

Koshi Makihara

Kensuke Harada

Tomohiro Motoda



Ryoichi Nakajo

MOONSHOT
RESEARCH & DEVELOPMENT PROGRAM

AIREC

Kei Kase

Tetsuya Ogata

cvpaper.challenge

Hirokatsu Kataoka

Ryosuke Yamada