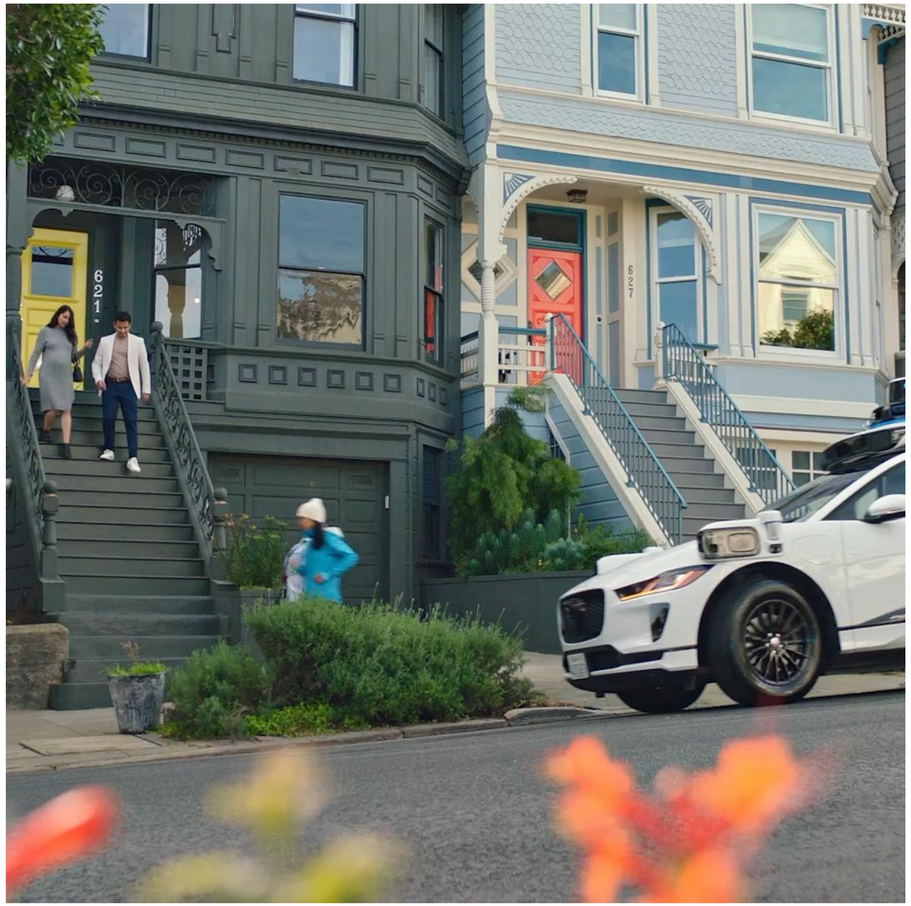
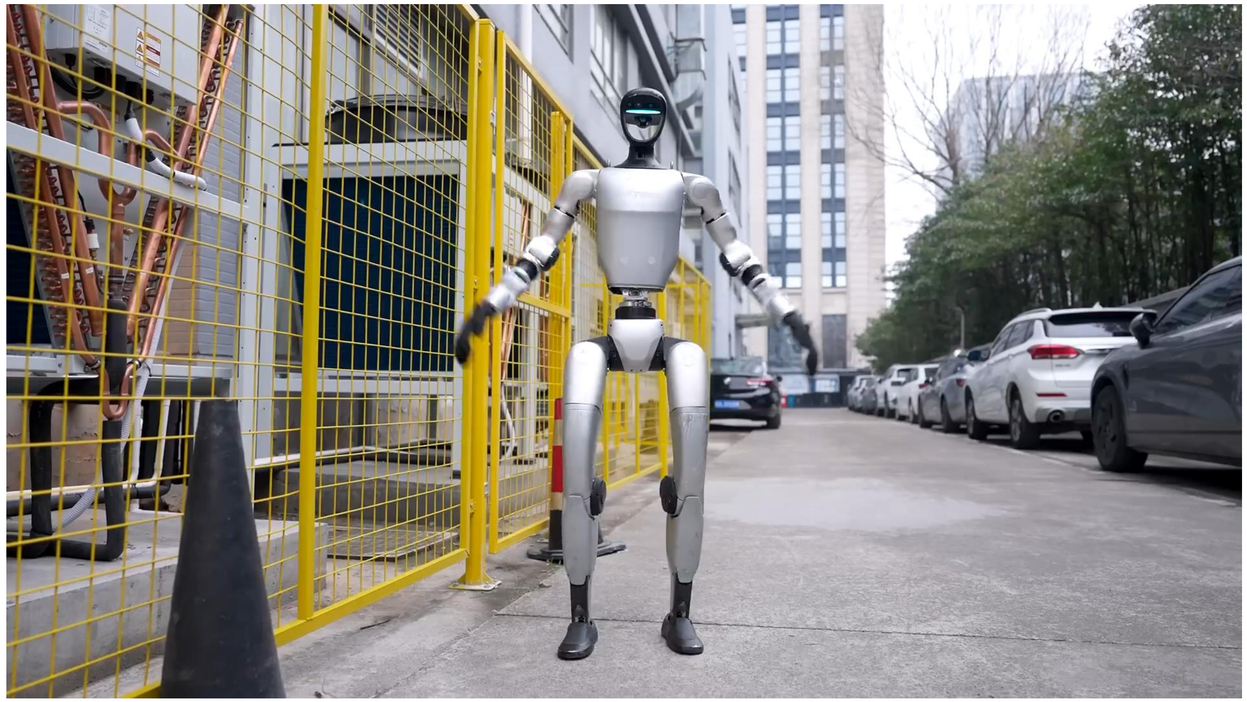
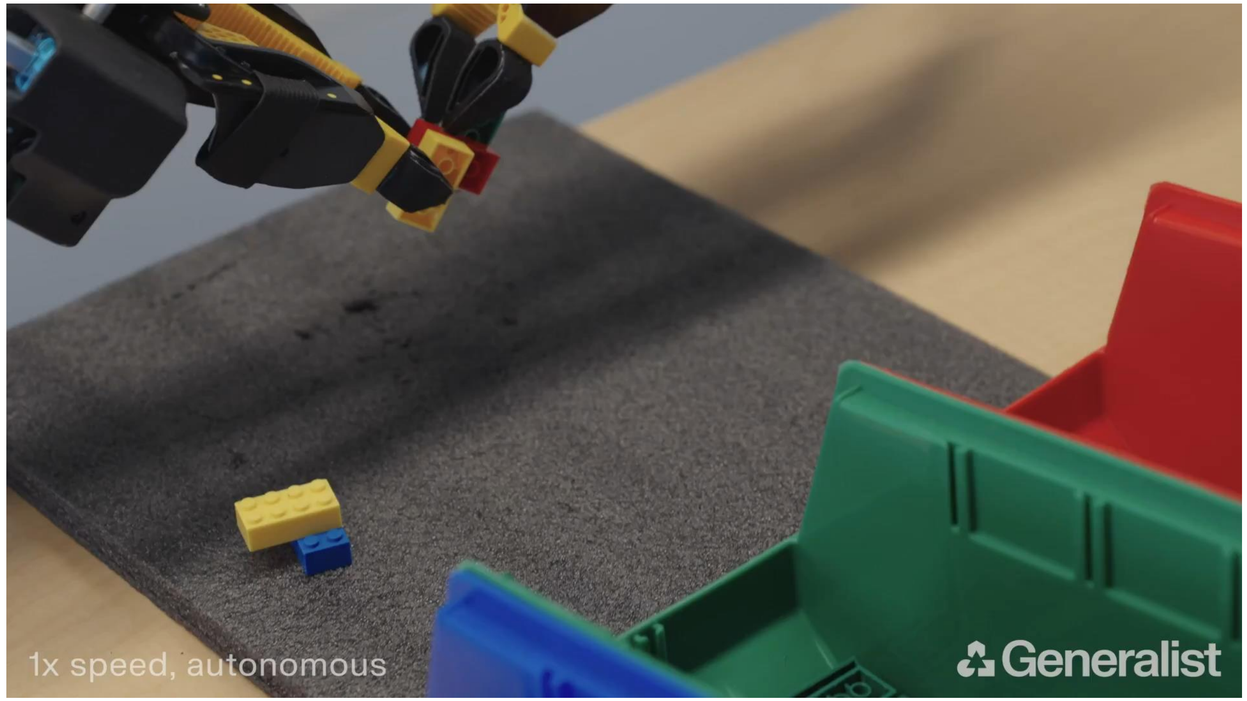
Scaling Up Robotic Data with Minimal Supervision

Yue Wang

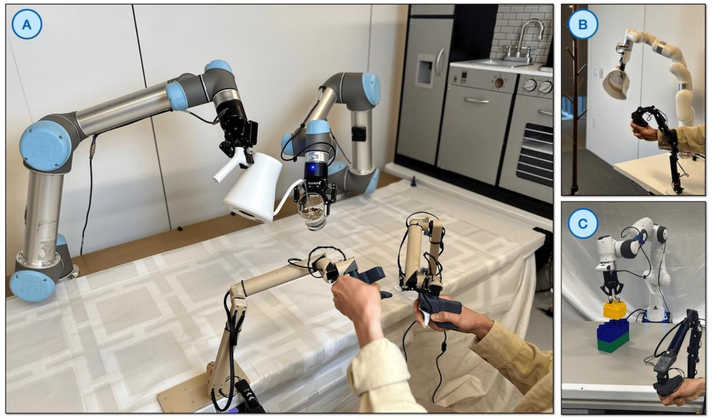
Space Robotics Workshop | July 29th, 2025







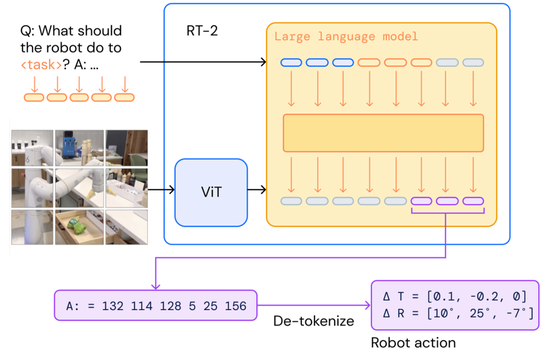
Cambrian Explosion of Robotics



Data

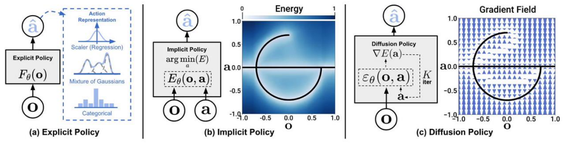
Hardwa
re

Algorith
m



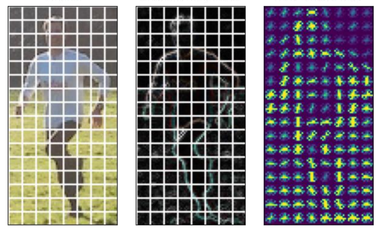
Diffusion Policy

Visuomotor Policy Learning via Action Diffusion

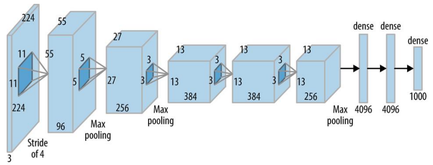


Data is the key to artificial intelligence.

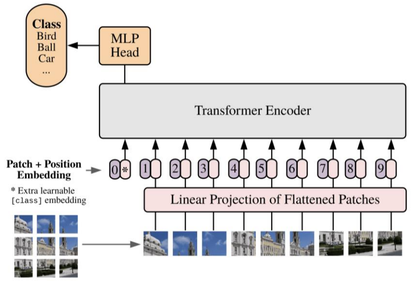
Physical AI



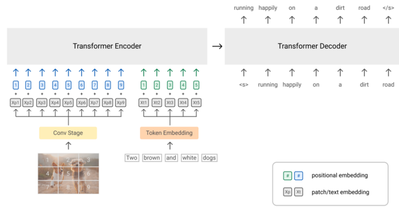
HOG+SIFT+SVM



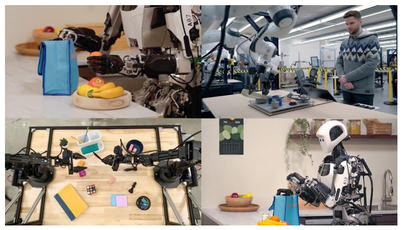
CNNs

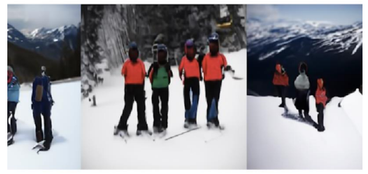


Transformers

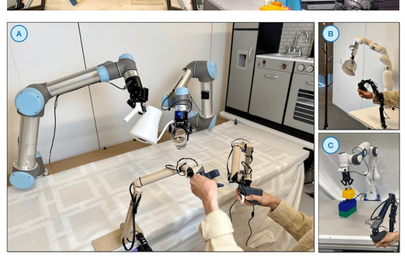


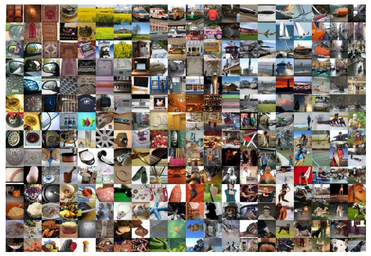
VLMs

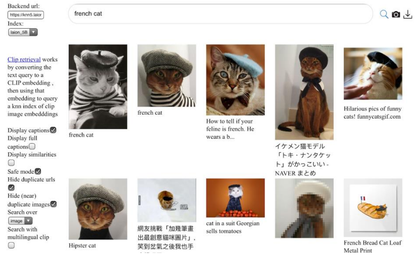




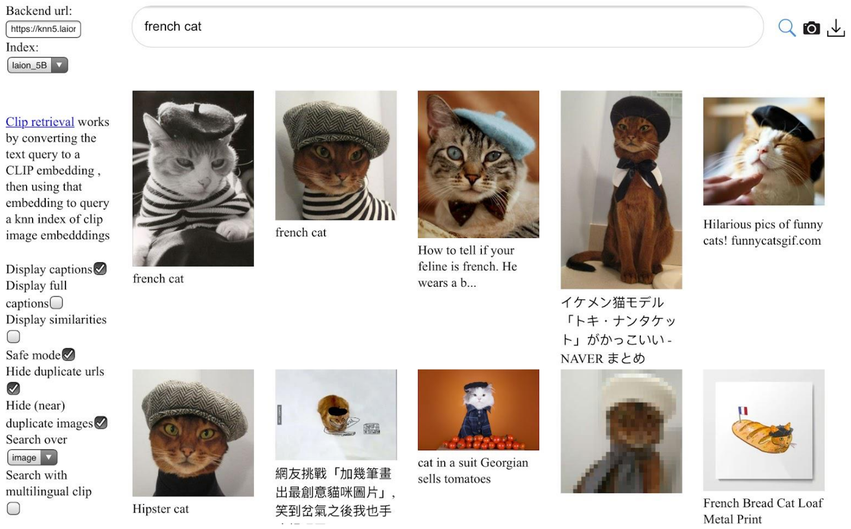
three people standing next to each other wearing skis and standing o

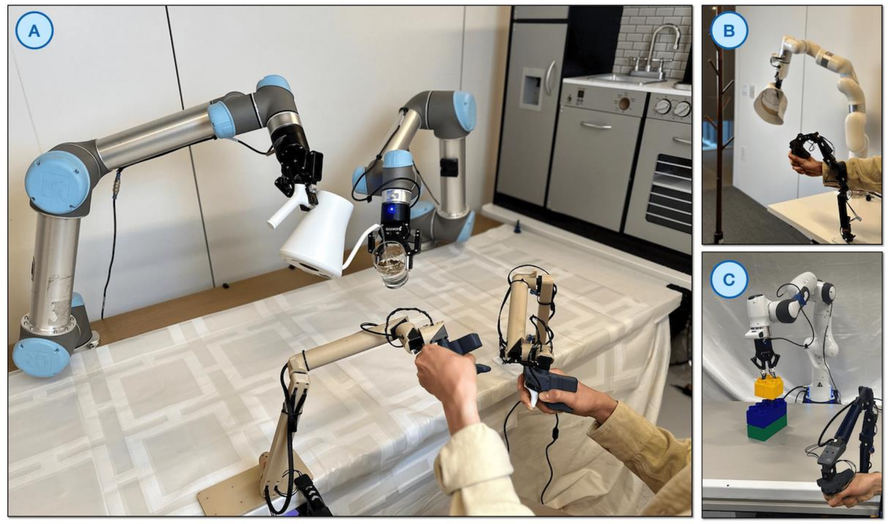






Little Data | Curated | Web Scale | Noisy Multimodal | Embodied Data





< 1s

Ubiquitous

$0.01 per data point

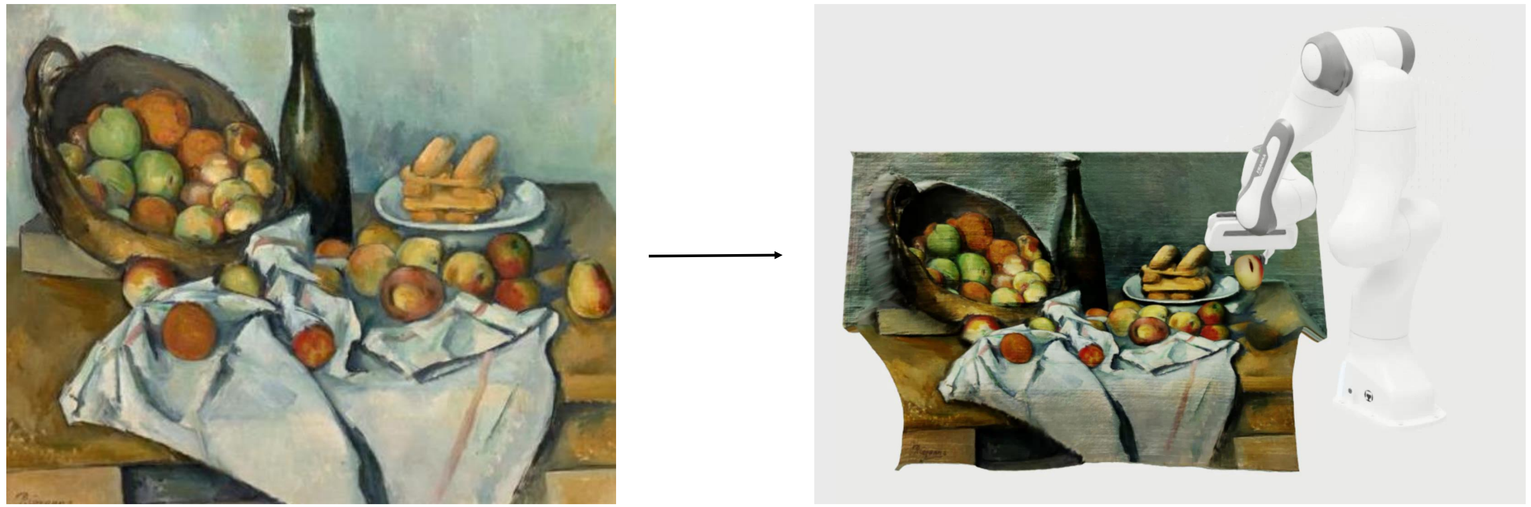
> 60s

Confined to lab environments

$5 per data point

For space robots, data collection is even more challenging

Can we convert a single image into a robotic data?

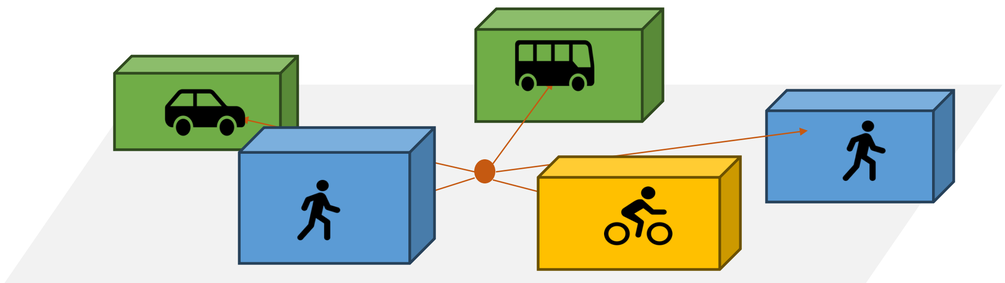


[In submission 2025] “Robot Learning from Any Images.” Zhao et al.

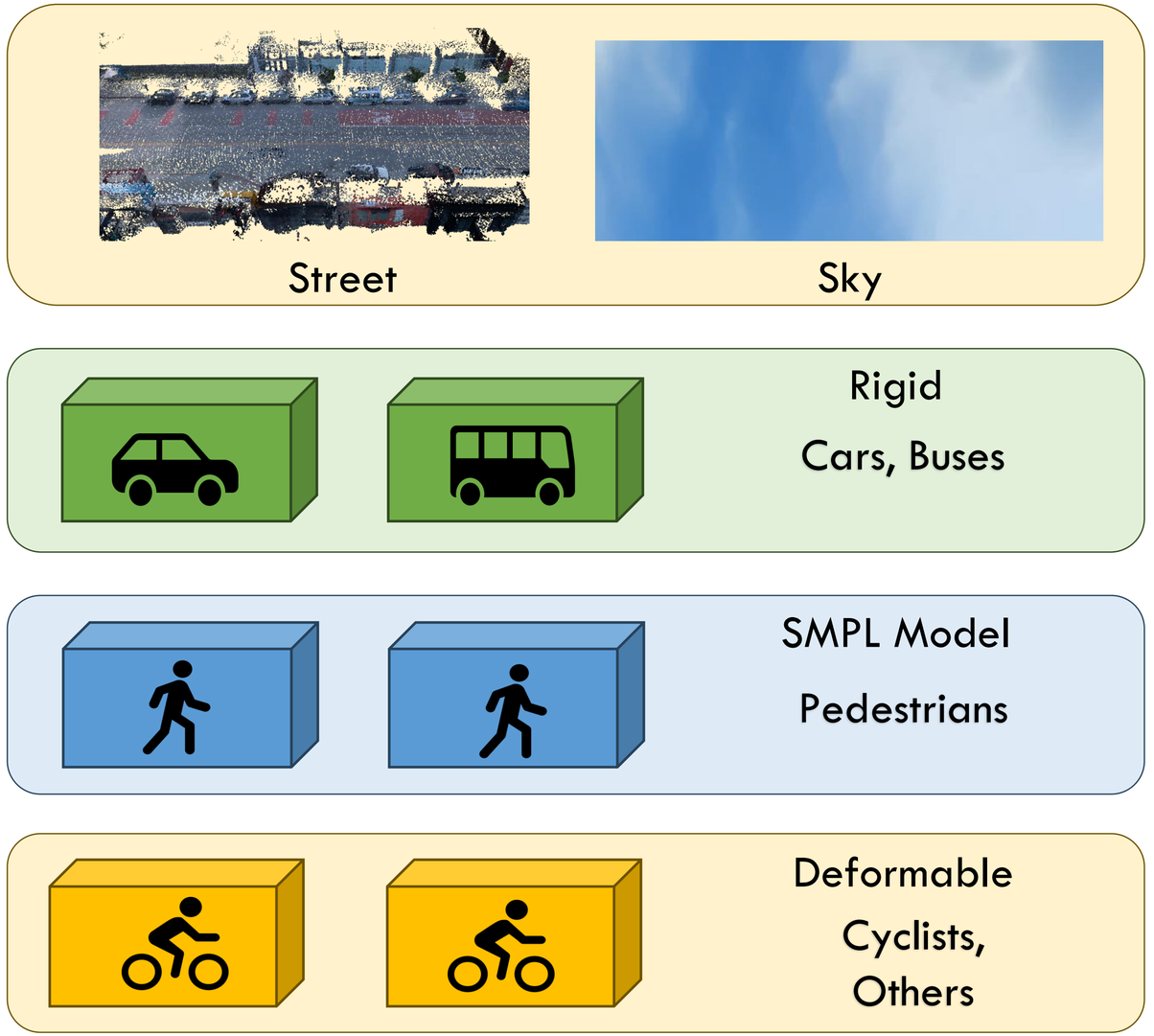
Omni Urban Scene Understanding

[ICLR 2025] “Omni Urban Scene Reconstruction.” Chen et al.

Scene Modeling



Gaussian Scene Graph



Applications

Let People Dance!

Applications

Driving Simulation

Applications

Bullet time

Robot learning from non-robotic data

Manipulate unseen objects in unseen environments with unseen embodiments.

Learn manipulation from costly in-domain demonstrations.

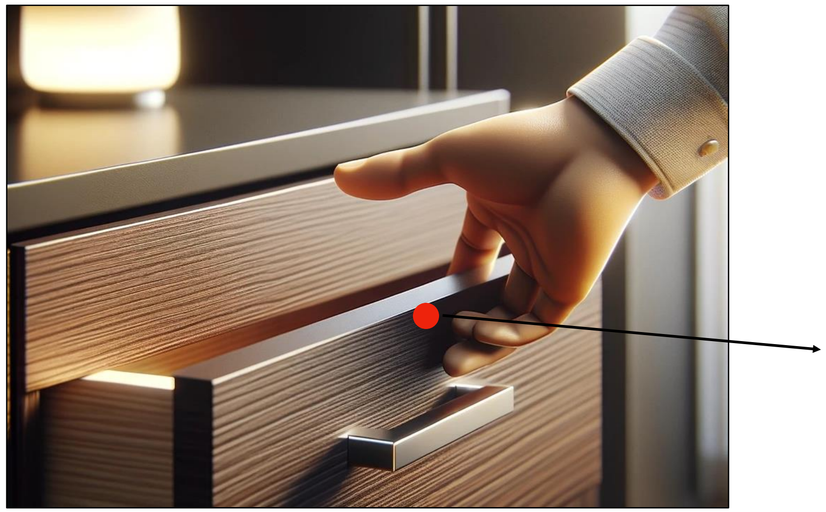
Acquire versatile manipulation capabilities from abundant out-of-domain data.

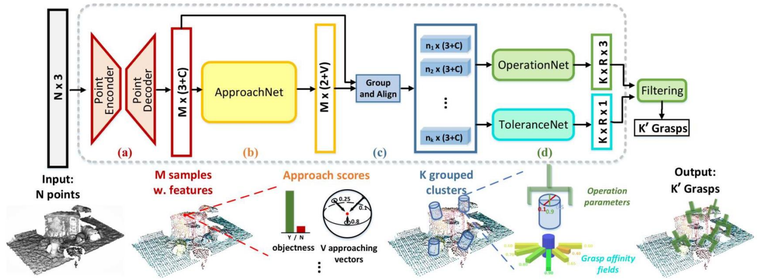
...

RAM: Retrieval-Based Affordance Transfer for Generalizable Zero-Shot Robotic Manipulation

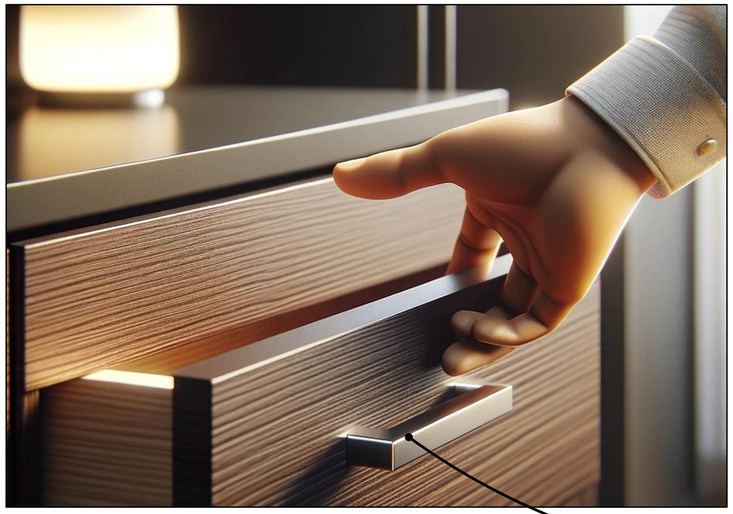
- Represent the actionable knowledge as transferrable affordance, i.e. ‘where’ and ‘how’ to act
  - ‘where’ to act: 3D contact point
  - ‘how’ to act: 3D post-contact direction

- Off-the-shelf grasp generators and motion planners for execution.

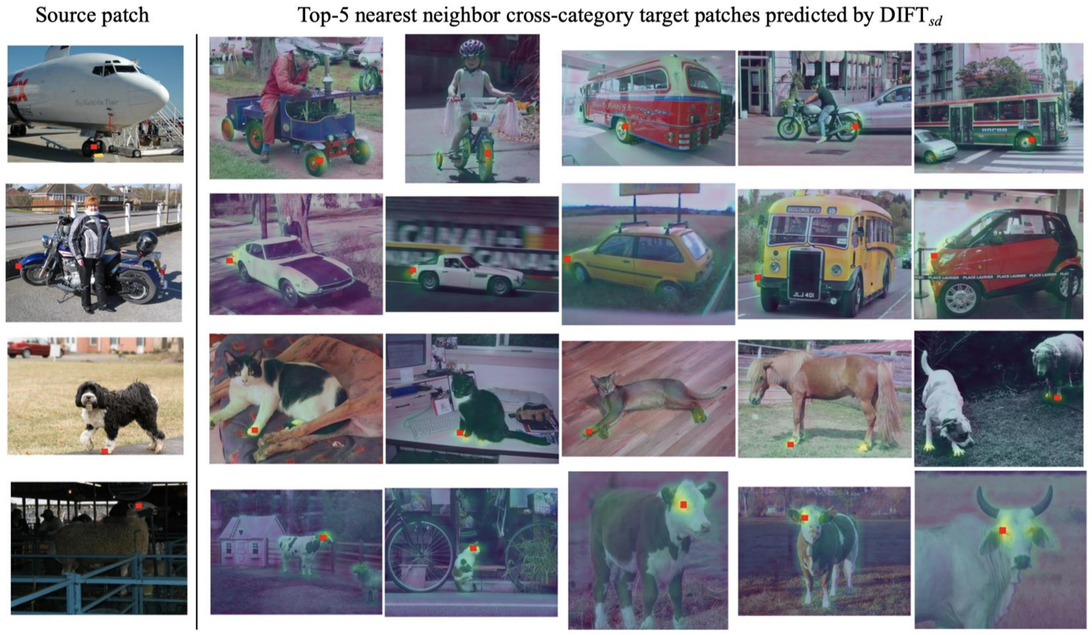




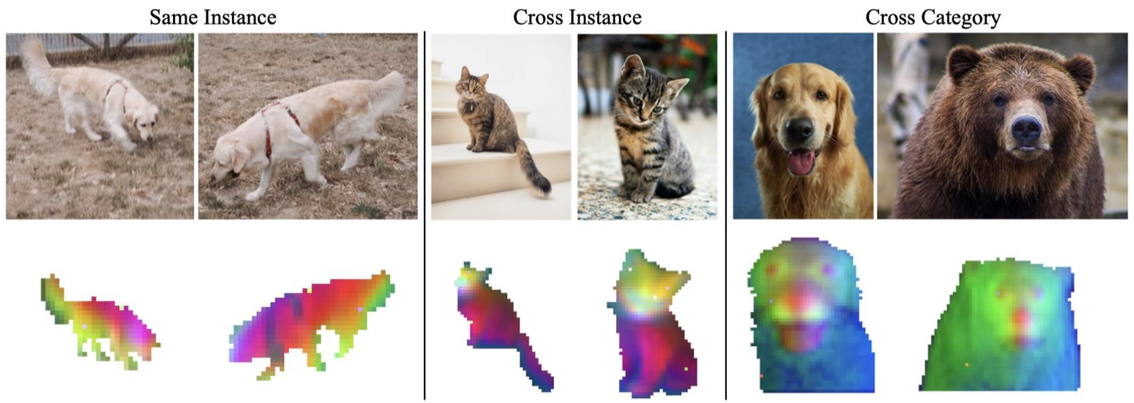
[CoRL 2024] “RAM: Retrieval-Based Affordance Transfer for Generalizable Zero-Shot Robotic Manipulation.” Kuang et al.



How to match points with the same semantic concepts?

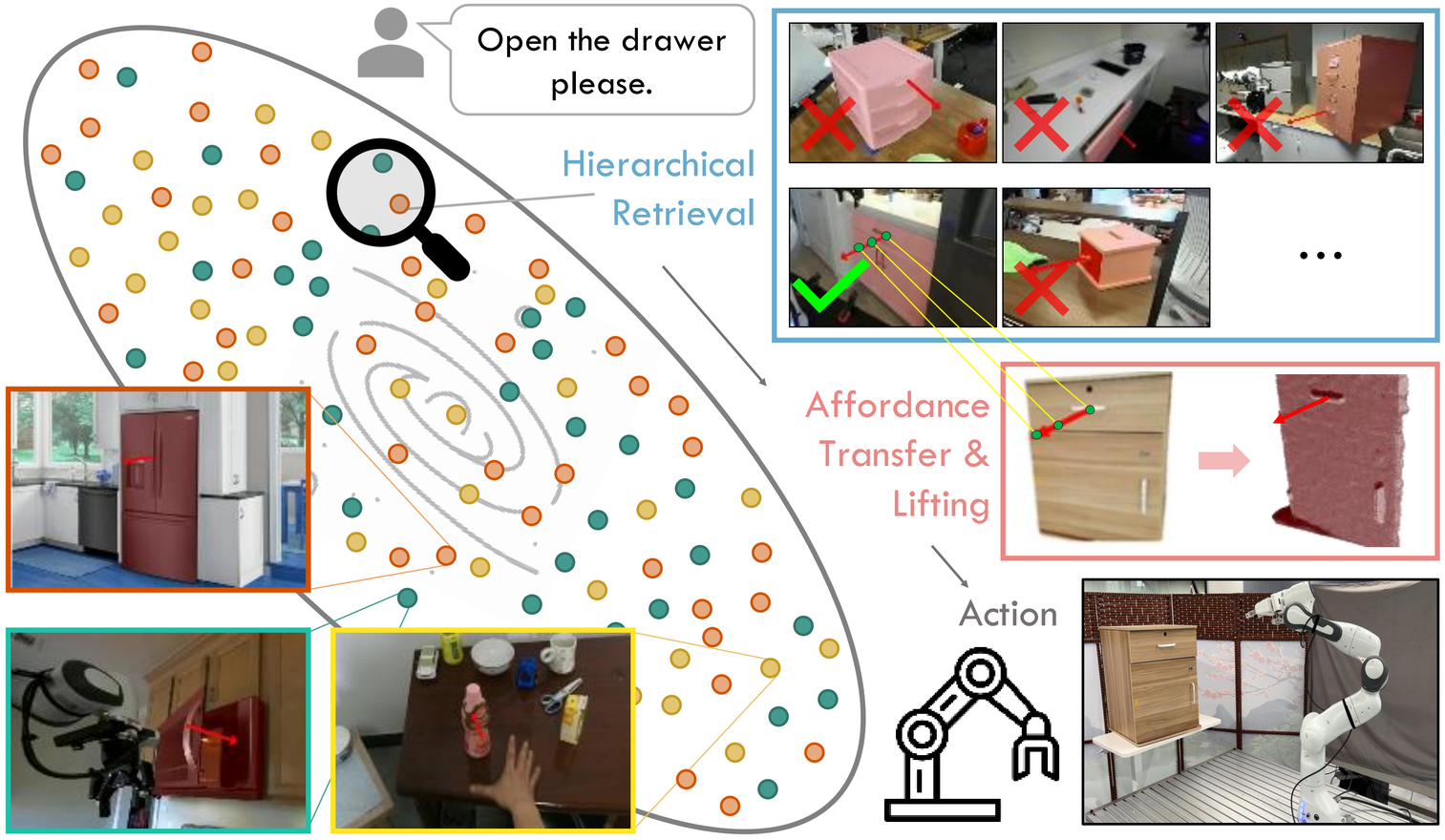


- Emergent dense correspondence of feature maps.
- Cross domain/instance/category generalization.



[NeurIPS 2023] “Emergent correspondence from image diffusion.” Tang et al.

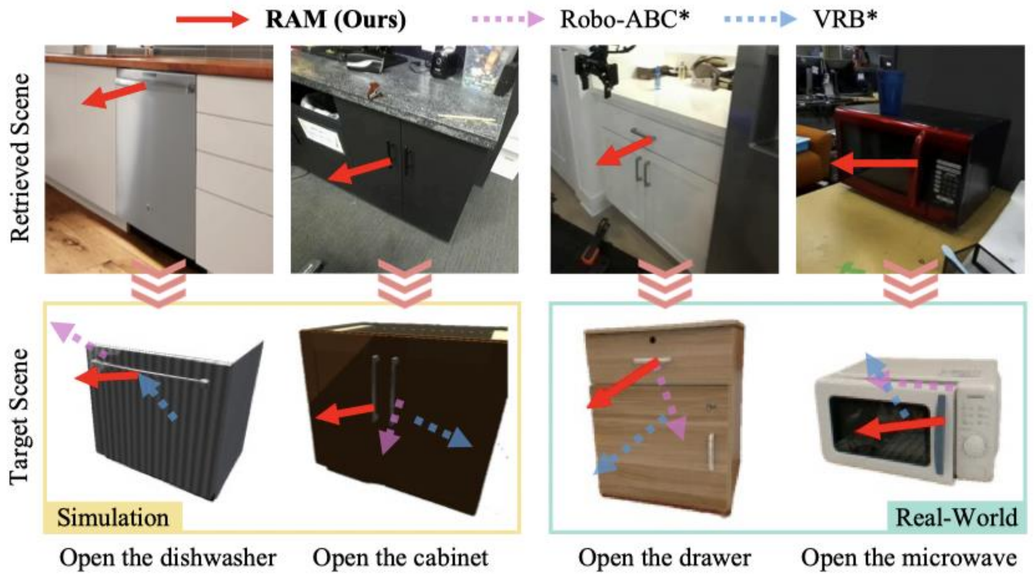
Overview



The Pipeline of RAM.

Experiment Results

Object														AVG
Task	O	C	O	C	O	C	O	O	O	P	P	P	P	/
Where2Act [14]	2	34	2	54	2	**68**	2	0	/	/	/	/	/	20.50
VRB* [12]	8	62	6	56	16	66	4	12	10	18	28	44	60	30.77
Robo-ABC* [44]	20	58	22	60	30	46	30	28	26	40	54	66	60	41.54
RAM (Ours)	**38**	**68**	**32**	**76**	**32**	50	**66**	**54**	**38**	**46**	**56**	**72**	**64**	**52.62**



Object							AVG
Task	O	O	O	P	P	P	/
Robo-ABC* [44]	2/5	1/5	1/5	**3/5**	**4/5**	4/5	50.0
RAM (Ours)	**3/5**	**2/5**	**3/5**	**3/5**	**4/5**	**5/5**	**66.7**

Zero-Shot Manipulation

"Open the microwave"

Retrieved Demonstration

3x

Zero-Shot Manipulation
"Open the cabinet"

Retrieved Demonstration

3x

Zero-Shot Manipulation
"Pick up the banana"

Retrieved Demonstration

3x

One-Shot Visual Imitation
"Pick up the tissue paper"

3x

One-Shot Visual Imitation

"Pick up the tissue box"

3x

One-Shot Visual Imitation
"Close the drawer"

3x

More DoFs
Not easily handled by affordance
Action retargeting is hard

How can we learn humanoid dexterity from Internet data?

UH-1: Learning from Massive Human Videos for Universal Humanoid Pose Control

Massive Internet Videos

YouTube

Youtube

DeepMind

Kinetics 700

AI2

Allen Institute for AI

Charades

Video Clip Extraction

3D Human Pose Estimation

Video Captioning

"A young woman is doing a workout in a living room, using her legs and arms to perform various exercises."

Goal-based Reinforcement Learning

Motion Retargeting from Humans to Humanoids

Text Representation

"A young woman is doing a workout in a living room, using her legs and arms to perform various exercises."

Humanoid Robot Actions

UH-1 Transformer

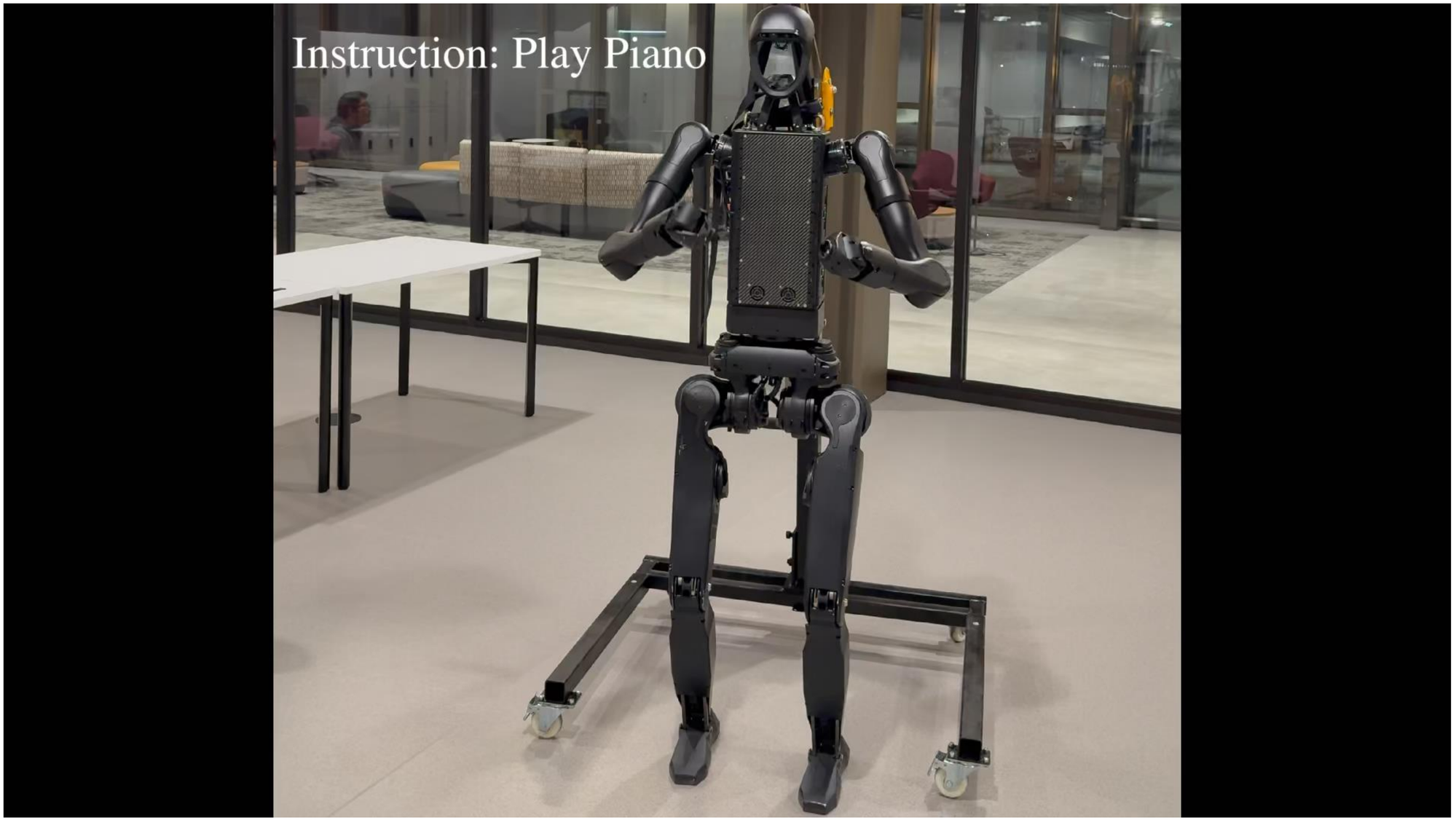
Humanoid Action Tokens

An automatic humanoid data engine

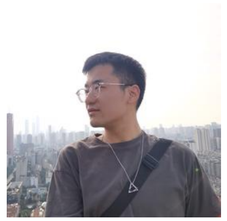
A unified whole-body control model

[In submission 2025] "Learning from Massive Human Videos for Universal Humanoid Pose Control." Mao et al.

Real-World Deployment of UH-1

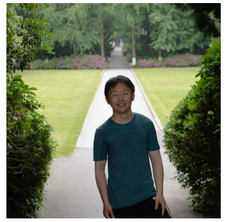


Acknowledgement


Junjie Ye


Jiageng Mao


Jiawei Yang


Siheng Zhao

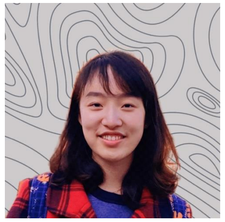

Cameron Smith


Yuxuan Kuang


Wei Zhou


Boris Ivanovic


Sanja Fidler


Congyue Deng


Zan Gojcic


Marco Pavone


Vitor Guizilini


Leo Guibas