

World-Paired Demonstration

Behavior captured from the embodied perspective and surrounding world.

Sanjay Mukhyala and Evan Rose

Robots will live in a world built for humans. Why should we have to demonstrate every skill again to a robot?

Human performs task

(knowledge trapped in human behavior)

≠

Robot repeats task

Teleoperate robot

Collect robot demos

Train robot

Learning from demonstration (LfD) with wrist cams or teleop is not going to be done in your house or in a factory.

Dyna: Dyna-2



Generalist: 1.5



Skild: S1

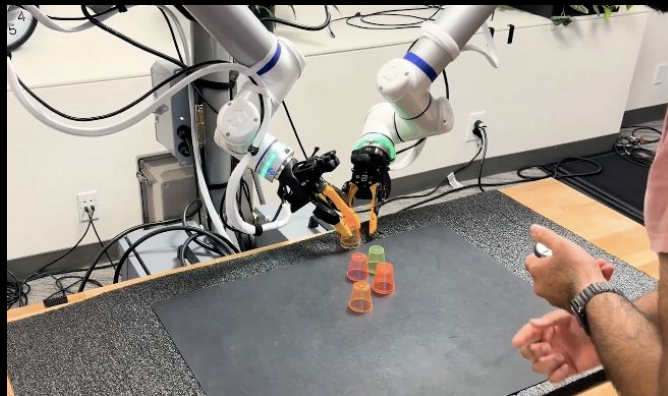


LfD through exocentric understanding is scalable

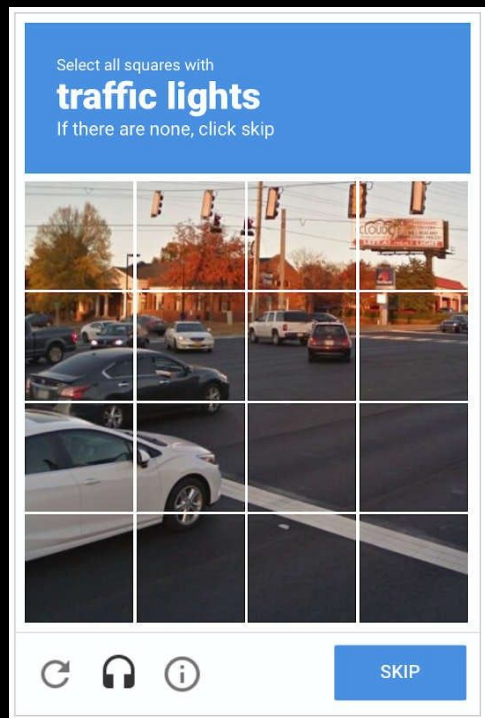
Matic



Gen 1.5



Waymo's Model



Internet info (general ego)



Successful autonomous systems already benefit from world context. Why shouldn't robotic manipulation?

Can world context make robotic learning more data-efficient?

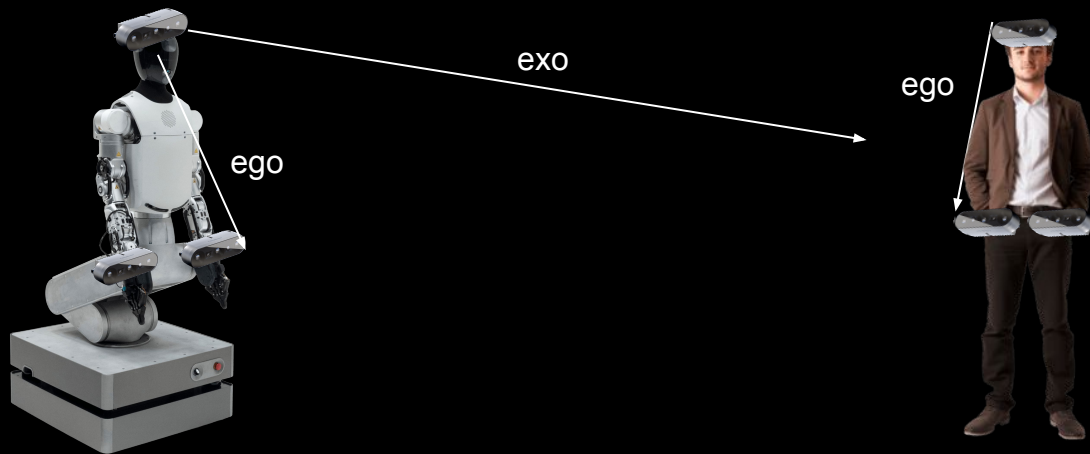
Hypothesis: pairing embodied views (ego) with world views (exo) helps robots learn new tasks more effectively from fewer demonstrations.

What if robot demonstrations stop being the primary source of robot knowledge?

human data (exo) = knowledge source

robot data (ego) = robot grounding

} World-paired demonstration



Framing the experiment

EGO ONLY

Human ego

Encoder

Representation

WORLD-PAIRED

Human ego + Paired exo

Encoder

Representation



[< Prev](#)[Next >](#)

1 / 179 t = 0.0-5.0s

EXO — CEILING

the robot's view



EGO — HEAD-WORN

the human's view



Label what is actually happening, using BOTH views. The point of this data is to test each view separately afterwards, so the labels themselves must not favour either one. If an object is only legible in one view, still label it. Use **unclear** only when both views leave you genuinely unsure.

VERB — WHAT IS BEING DONE

1 pick up

2 put down

3 attach / fit

4 detach / remove

5 screw / tighten

6 insert

7 position / align

8 connect wire

9 inspect / check

0 idle / walk

NOUN — WHAT IS BEING HANDLED

a frame / chassis

s seat

d battery

f wheel / tyre

g handlebar

h mirror

j body panel

k footboard

l screw / bolt

z screwdriver

x wrench

c wire / harness

v connector

b carton / box

n sticker / label

m cable tie

q other

w none

idle / walk · none labelled

11 labelled · 168 left

1-90 verb

a-z noun

← →

window

space

replay

u

unclear

saves automatically in this browser

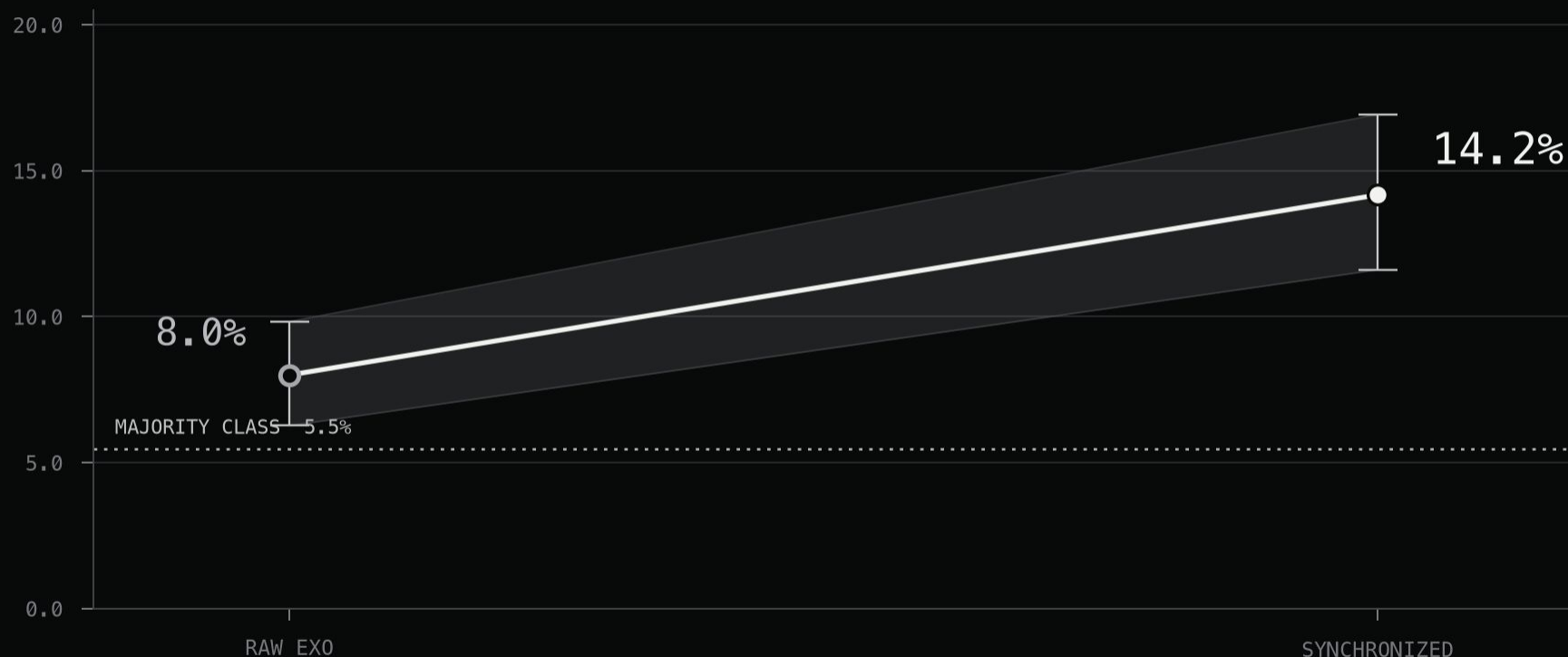
Results

Datasets: Assembly101 (Meta Reality Labs), World Context Dataset

- Exo-ego: +0.48 pp in task labeling on average compared to ego-ego (Assembly101)
- Exo-ego: 8.0% → 14.2% on held-out task labeling (Assembly101)
- Ceiling exo-cam provides 0 benefit in task labeling compared to ego-only task labeling (World Context 15 min exo-ego samples)

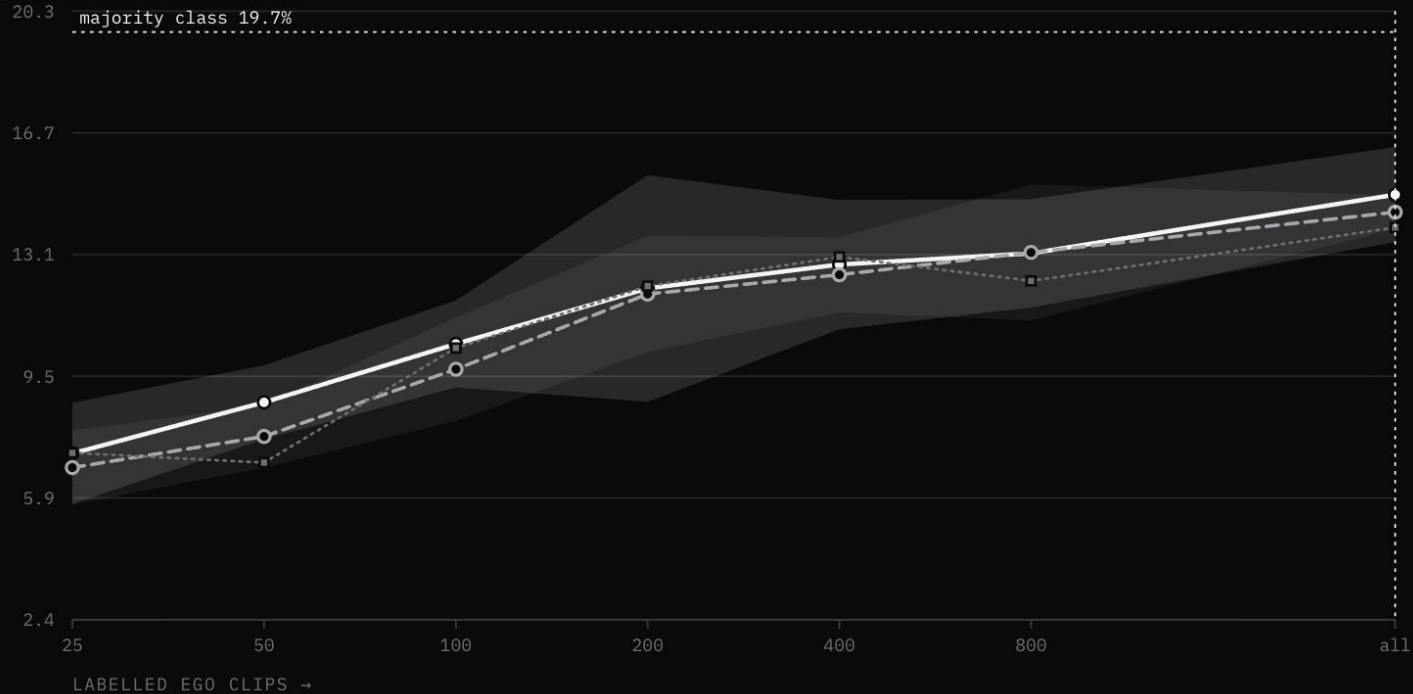
EGO-TRAINED HEAD / TASK RECOGNITION TOP-1 (%)

SYNCHRONIZED PAIRS TRANSFER THIRD-PERSON TASK SIGNAL



MEAN-PER-CLASS ACCURACY (%)

majority class 19.7%



all labelled clips

ego + exo 14.9 ± 1.4

ego only 14.4 ± 0.5

ego + ego (control) 13.9 ± 1.4

ego + exo

ego only

ego + ego

—●— ego + exo -○- ego only ...□... ego + ego (control)

If robots are going to live alongside humans, the human world has to become their training set.

World-**P**aired **D**emonstration