Robot Learning from Embodied Human Data

Danfei Xu

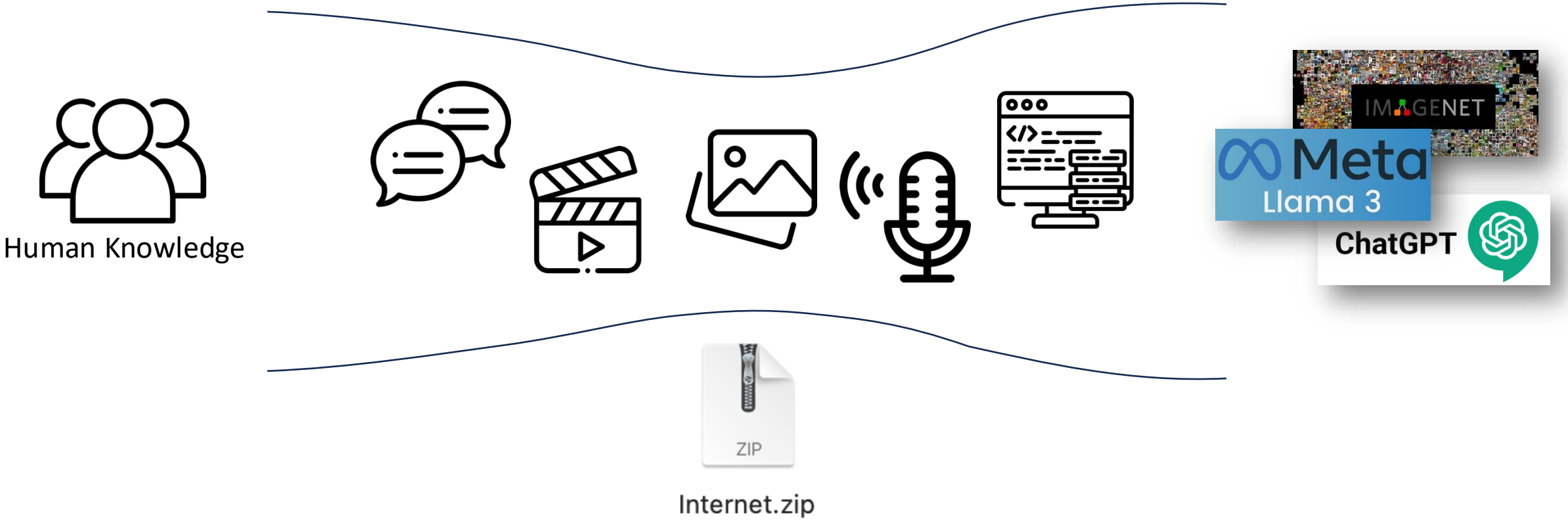
CoRL 2024 | 11.07.2024

Early Career Keynote

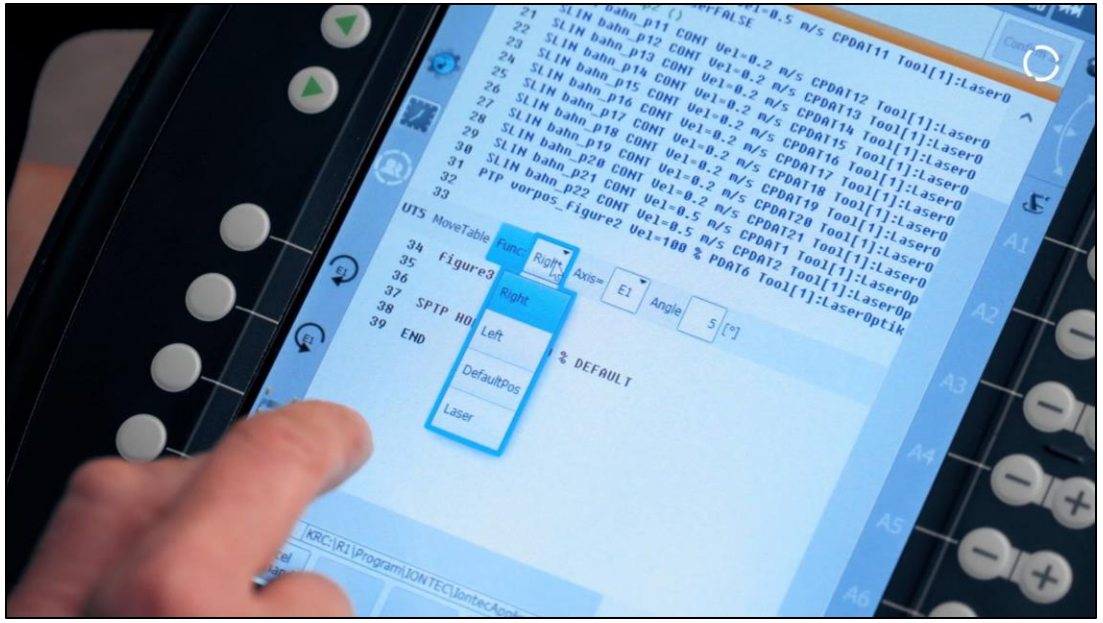
CoRL 2017

What makes today's AI so powerful?

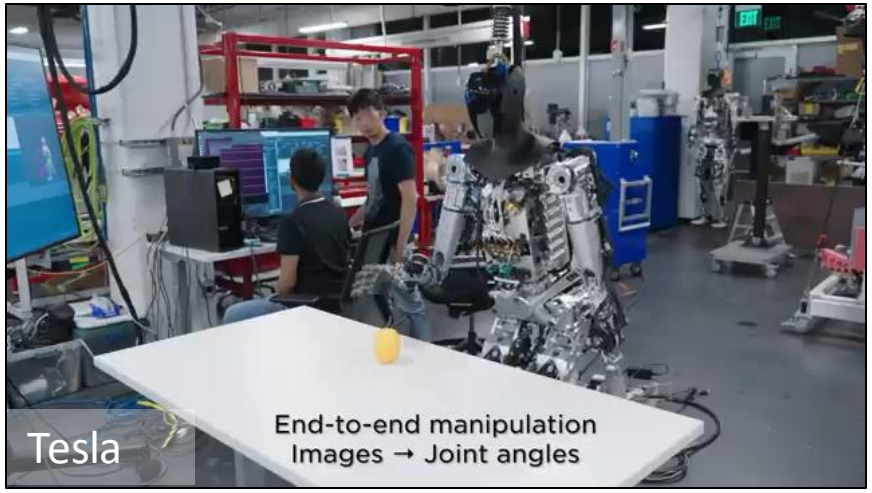
Scalable human-machine knowledge transfer via the Internet

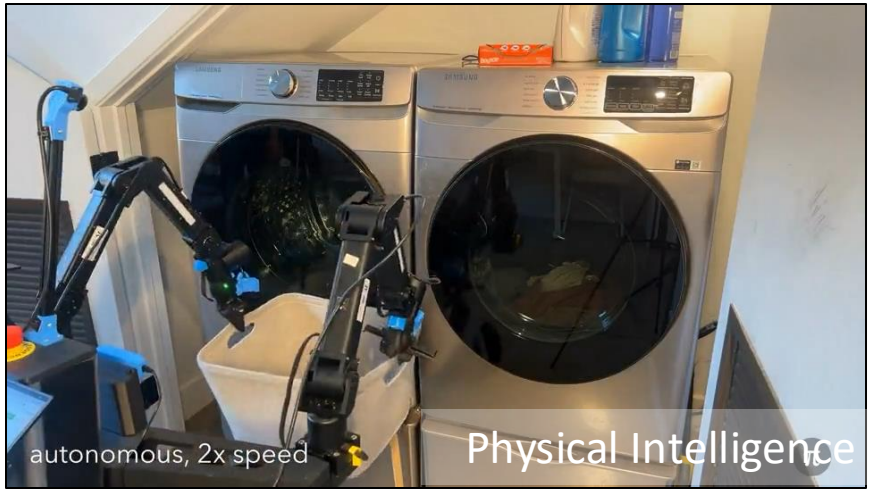


Transferring human knowledge to robots is painfully slow…



Transferring human knowledge to robot via teleoperation





Transferring sensorimotor knowledge via teleoperation

Mandlekar and Xu *et al.*, RSS'20

Robot Sensing

Robot Actions

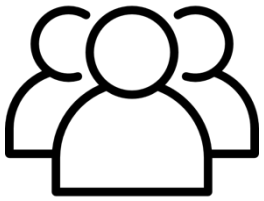
Demonstration Dataset

Robot Policy Models

Ajay Mandlekar

We transfer our sensorimotor knowledge by
perceiving and acting through the robot's body

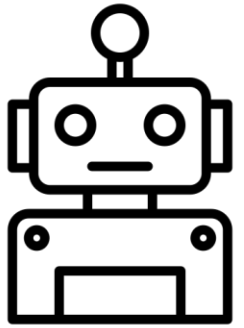
The knowledge transfer bottleneck



Human Knowledge

Teleoperation ≅ Playing claw machine games, except:

- It's not fun
- You don't win anything
- You have to put the toy back (reset)

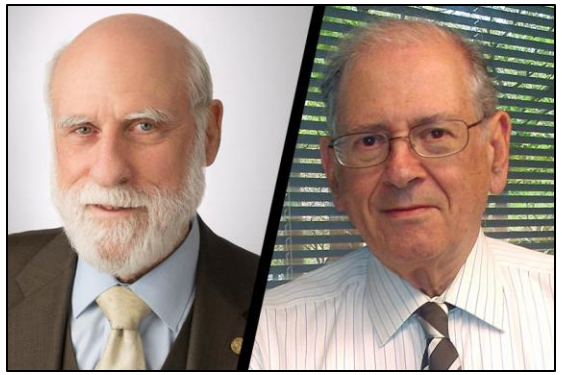


Robot Policies

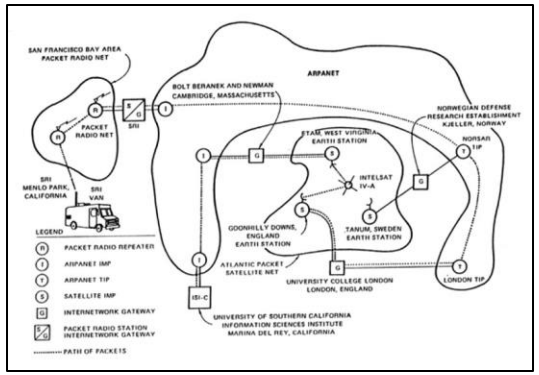
MIME Dataset (2018), RoboTurk Dataset (2019), Open X-Embodiment Dataset (2024), DROID Dataset (2024)

Lessons from the Internet

The Internet is an **unintended repository of AI training data.**
It captures human knowledge **without conscious effort** from its users,
with **frictionless** user experiences on **ubiquitous** devices

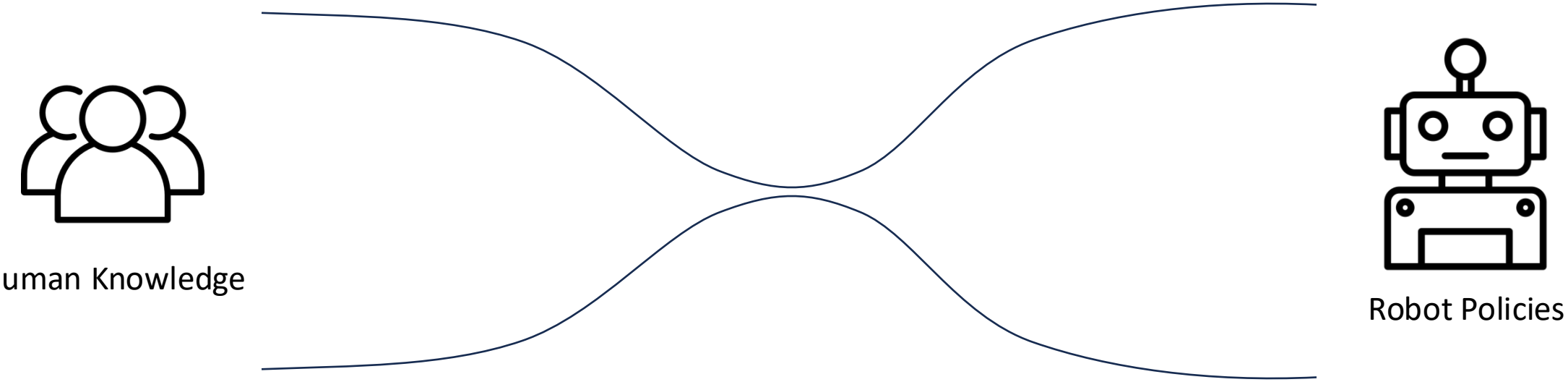


Cerf and Kahn, TCP/IP Protocol



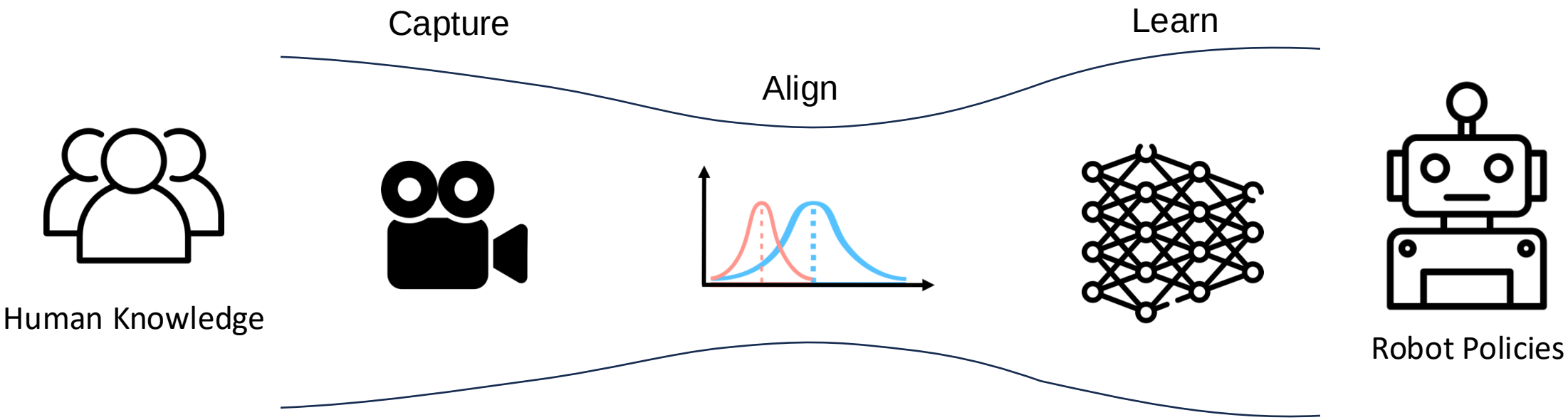
ARPANET, 1977

Breaking the Bottleneck



How can we collect robot learning data
without conscious human effort?

Breaking the Bottleneck with Embodied Human Data



- Collected from humans, not robots
- Contain embodied sensorimotor experiences
- Captured without conscious effort

Robot Learning from Embodied Human Data

***MimicPlay*:**
Learning from “Unembodied” Human Data

EgoMimic:
Embodied Human Data as Robot Data

***MimicTouch*:**
Learning from Multimodal Embodied Experiences

Beyond Scaling Up:
Robot Learning from a Human Perspective

Robot Learning from Embodied Human Data

***MimicPlay*:**
What makes learning from human data so challenging?

***EgoMimic*:**
Embodied Human Data as Robot Data

***MimicTouch*:**
Learning from Multimodal Embodied Experiences

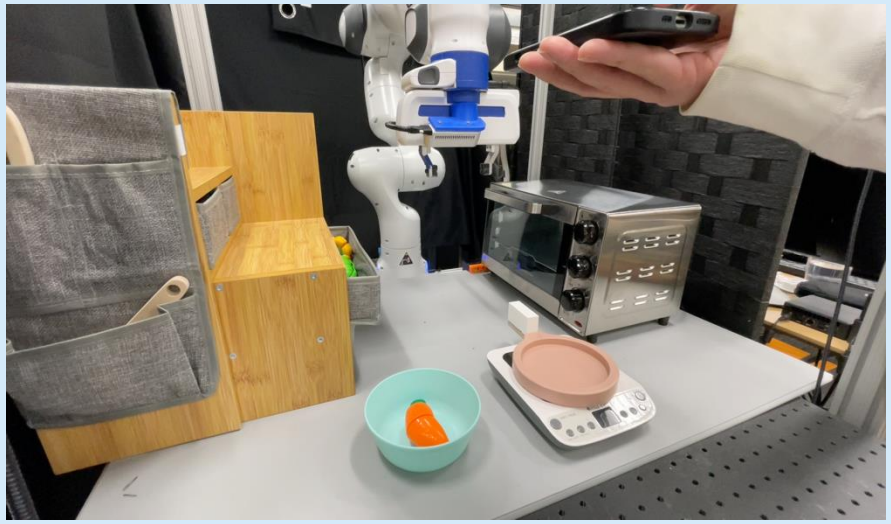
Beyond Scaling Up:
Robot Learning from a Human Perspective

What can robot learn from human videos?

Human Video Data

+ Fast to collect
- Large embodiment gap

Robot Teleop Data



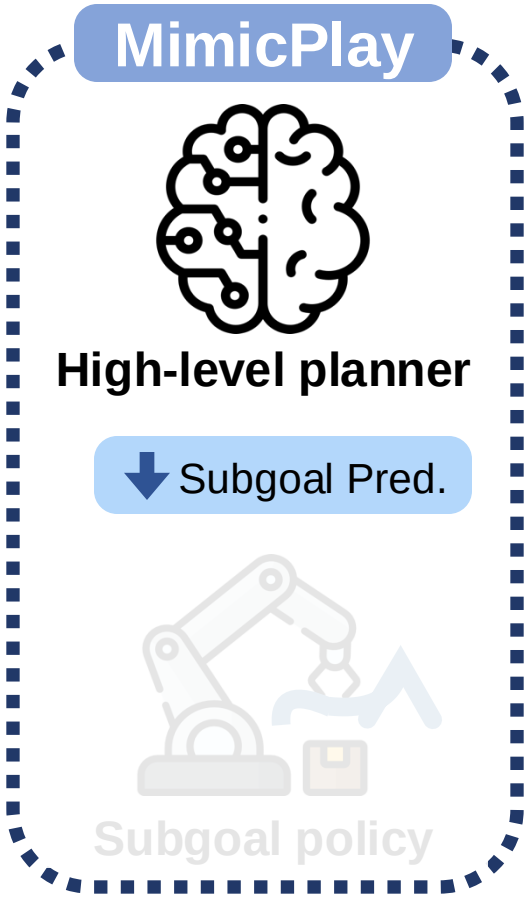
- Slow to collect
+ Minimal embodiment gap

Chen Wang

Key Insight: Learn "what to do" from human videos, and "how to do" from robot teleoperation data.

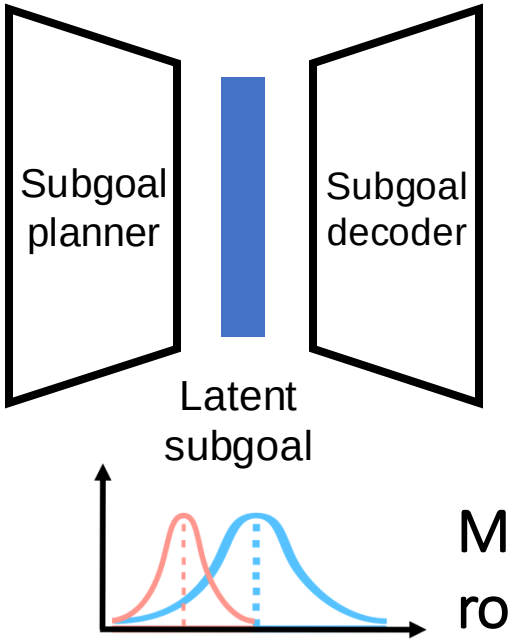
Wang et al. "MimicPlay: Long-Horizon Imitation Learning by Watching Human Play", CoRL'23 (Best Paper Finalist)

MimicPlay: Decoupled learning of "what" and "how"



Current image (human)

Goal image (robot)



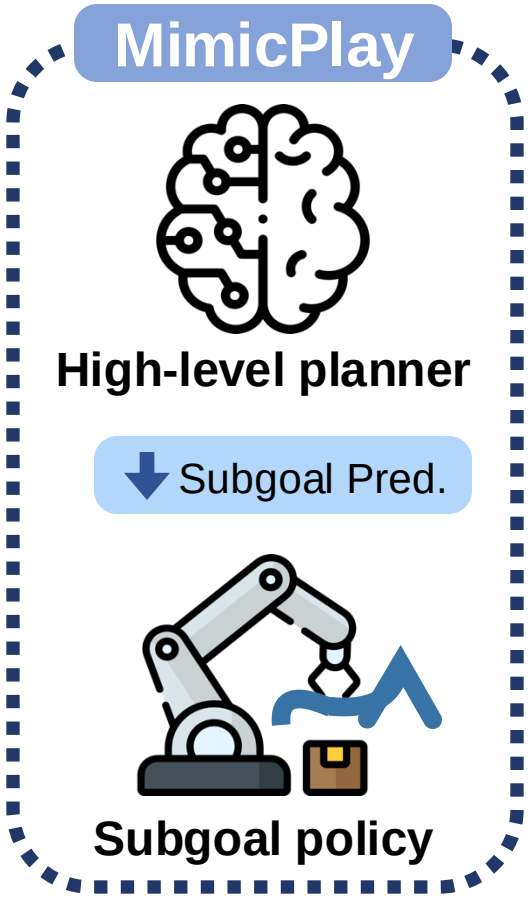
Planning: Predicting future 3D hand trajectories

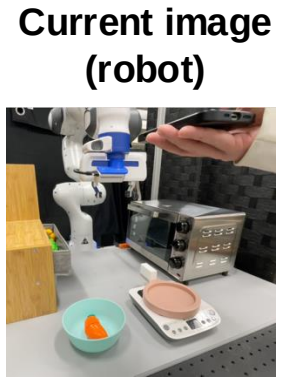
Minimize KL between human and robot latent plans

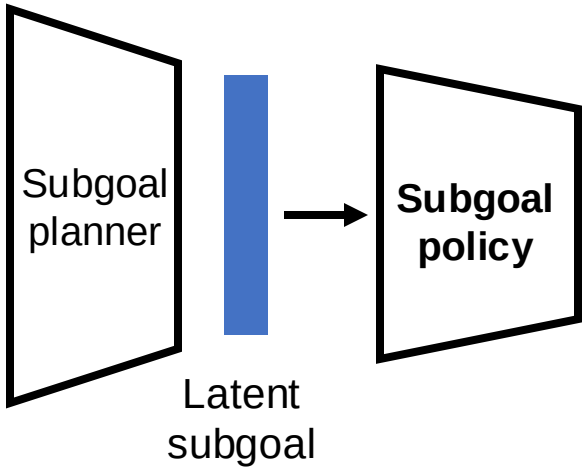
Learn *what* to do next (planning) from human video data.

Wang et al. "MimicPlay: Long-Horizon Imitation Learning by Watching Human Play", CoRL'23 (Best Paper Finalist)

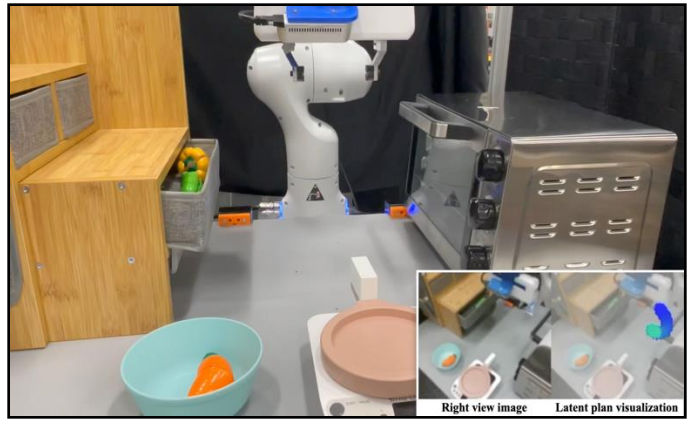
MimicPlay: Decoupled learning of "what" and "how"







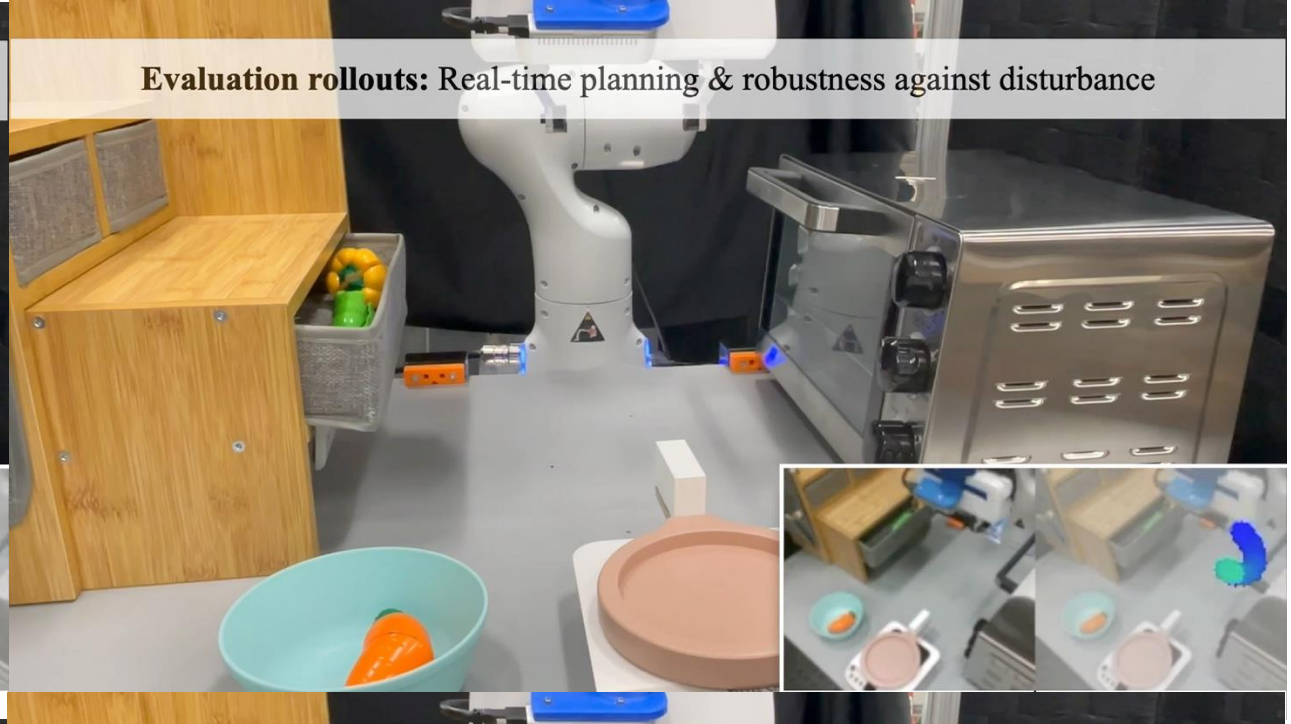
Execution: Real-time planning + closed-loop policy rollout



Learn *how* to reach the next subgoal from teleoperated robot data.

Wang et al. "MimicPlay: Long-Horizon Imitation Learning by Watching Human Play", CoRL'23 (Best Paper Finalist)

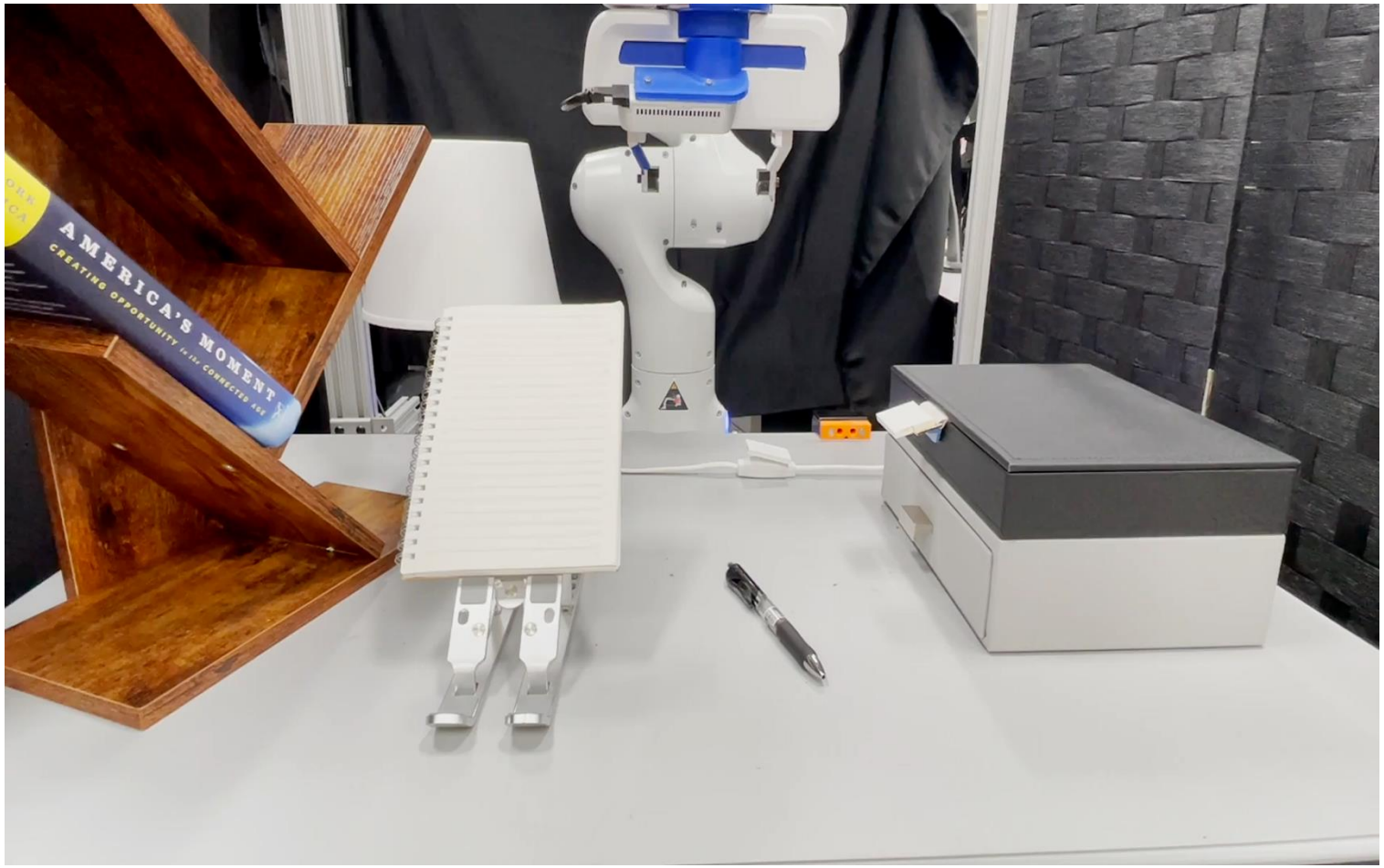

Evaluation rollouts: Real-time planning & robustness against disturbance
Input image (right view)
Generated plan


Evaluation rollouts: Real-time planning & robustness against disturbance


Evaluation rollouts: Real-time planning & robustness against disturbance
Input image (right view)
Generated plan


Evaluation rollouts: Real-time planning & robustness against disturbance
Input image (right view)
Generated plan

Policy “Prompting” with Human Video

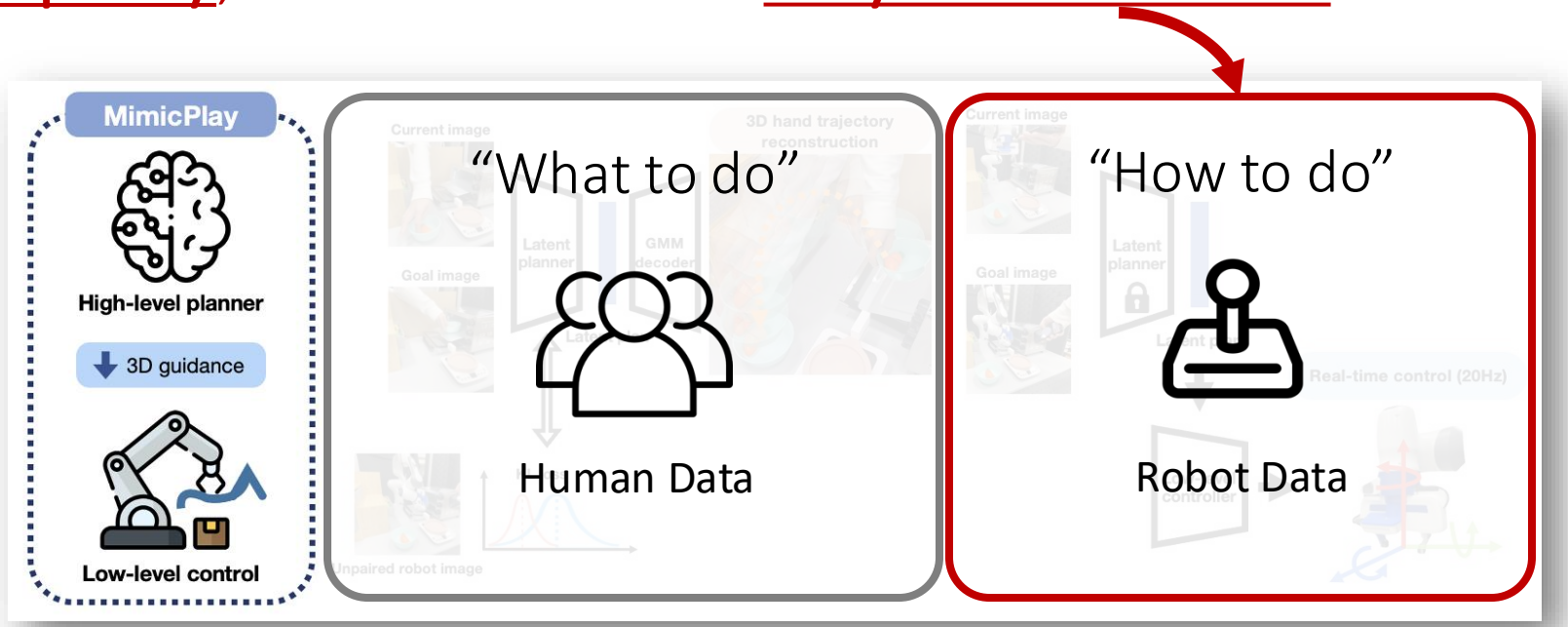


Follow new task sequence,
i.e., “do what I just did”.

Have we solved the problem?

✓ Learning **task goal -> subgoal guidance** from human data

✗ The hierarchical policy will **always be as good / bad as the low-level policy**, which is trained **only on robot data**



We are still bottlenecked by robot teleoperation data!

Can we go beyond learning the "what" from human data?

Common response: *No, Human data is not "**embodied enough!**"*

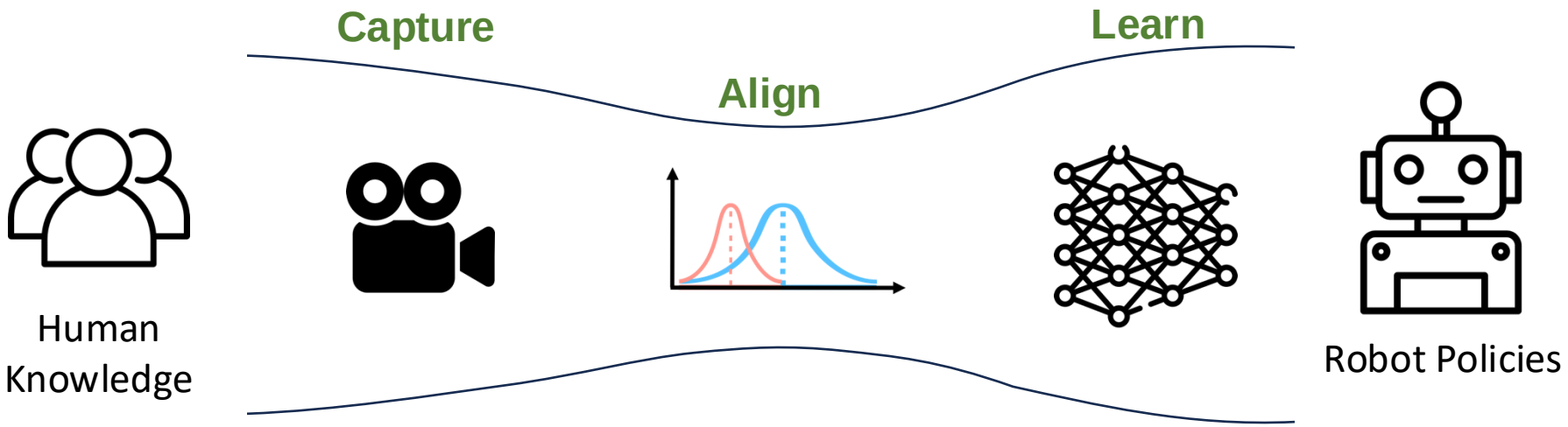
Human data should be treated differently than robot data because:

- **Sensing**: Third-person videos ≠ embodied experience.
- **Actions**: Human data don't have action labels.
- **Appearance & Kinematics:** Humans and robots look different and have different kinematic structures.

Can we **make human data *embodied,***
so robot can learn both the "what" and the "how"?

Can we make human data “embodied”?

Yes, but we need a full-stack solution



EgoMimic

Scaling Imitation Learning via Egocentric Video


Simar Kareer


Dhruv Patel*


Ryan Punamiya*


Pranay Mathur*

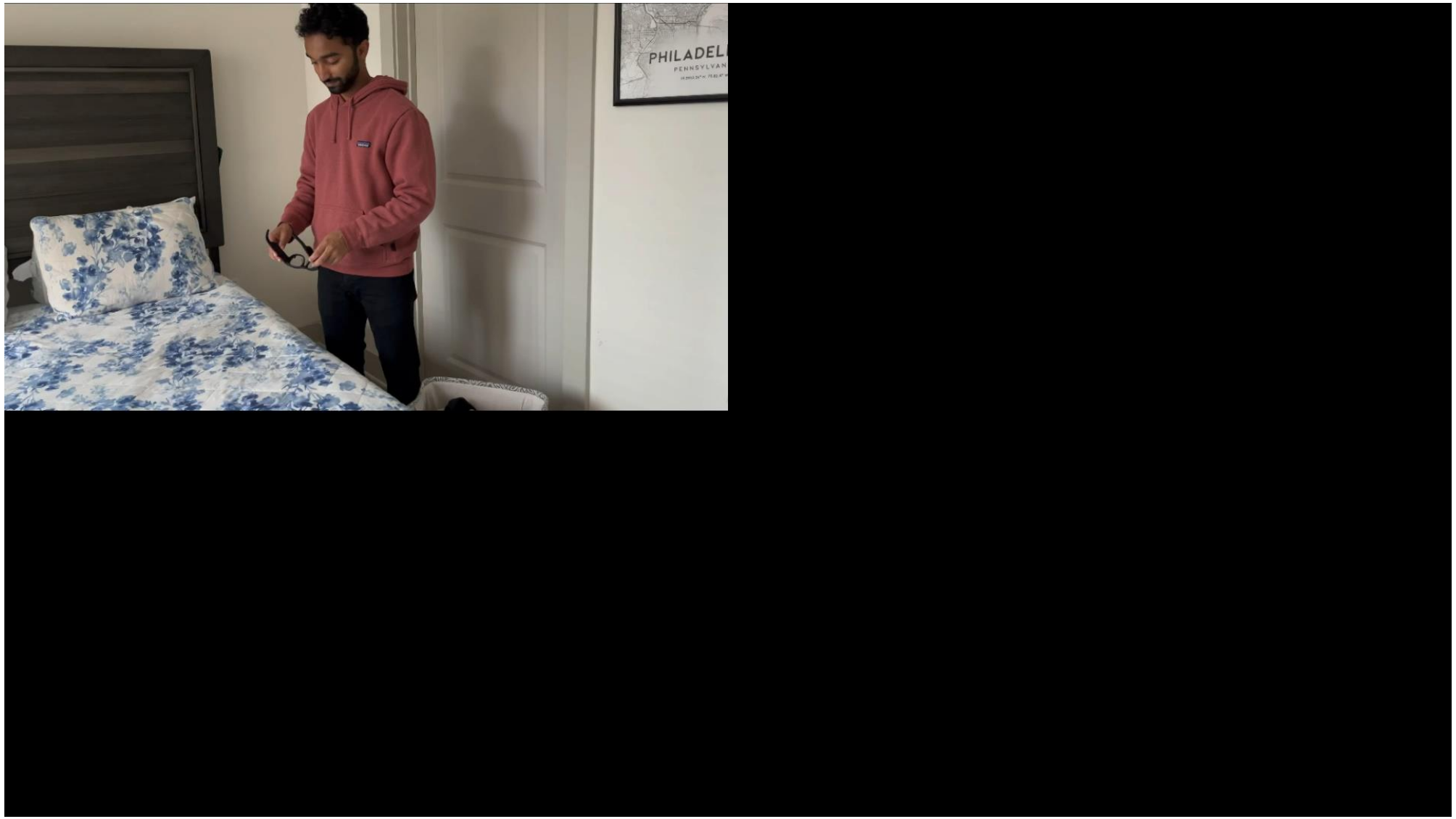

Shuo Cheng


Chen Wang


Judy Hoffman†


Danfei Xu†

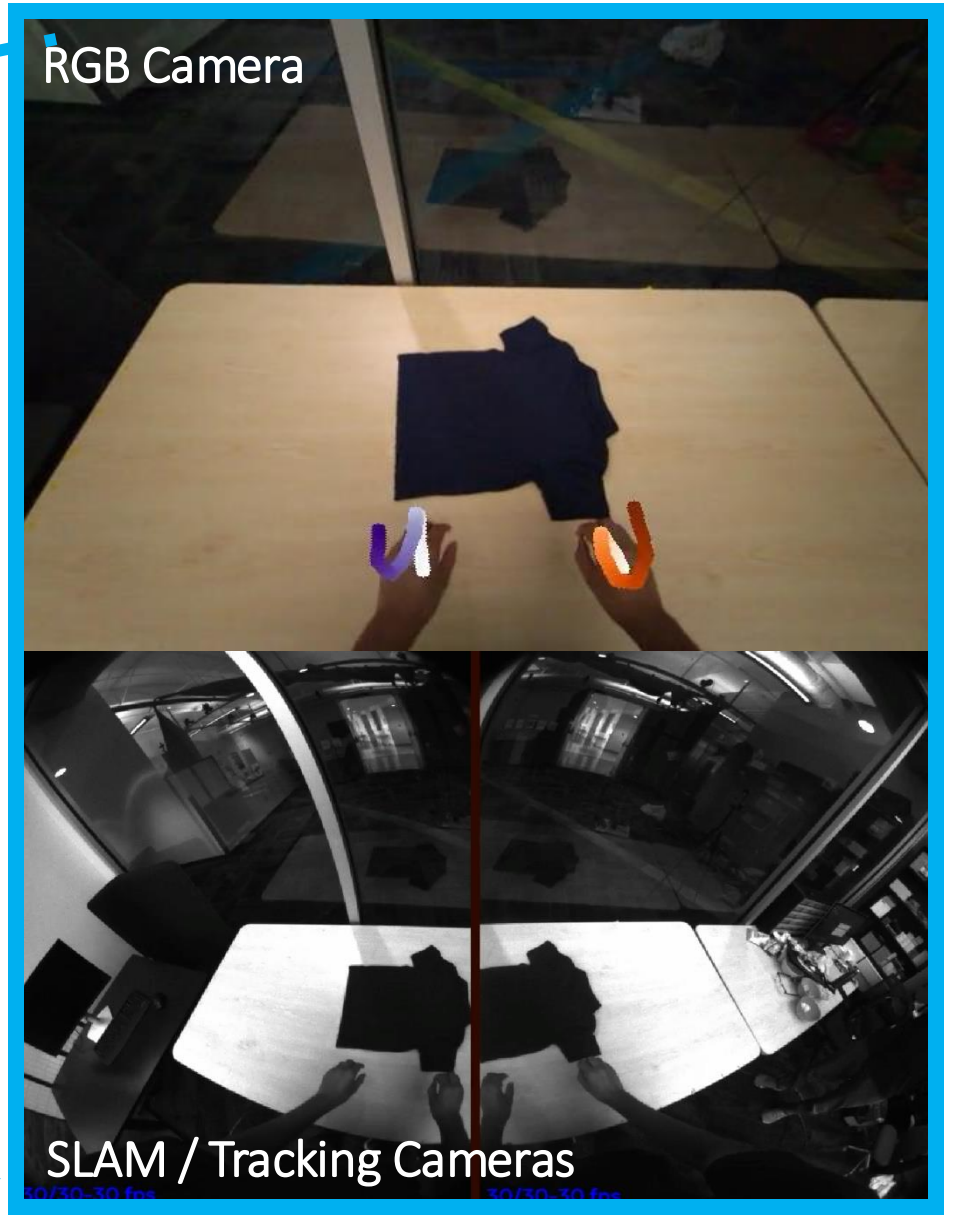
Capturing Embodied Experience from an Egocentric View



Kareer et al. “EgoMimic: Scaling Imitation Learning via Egocentric Video”, egomimic.github.io

Capture Egocentric Data with Smart Glasses

Project Aria Glasses

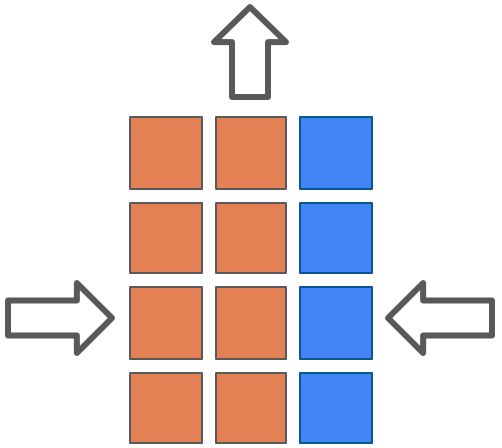


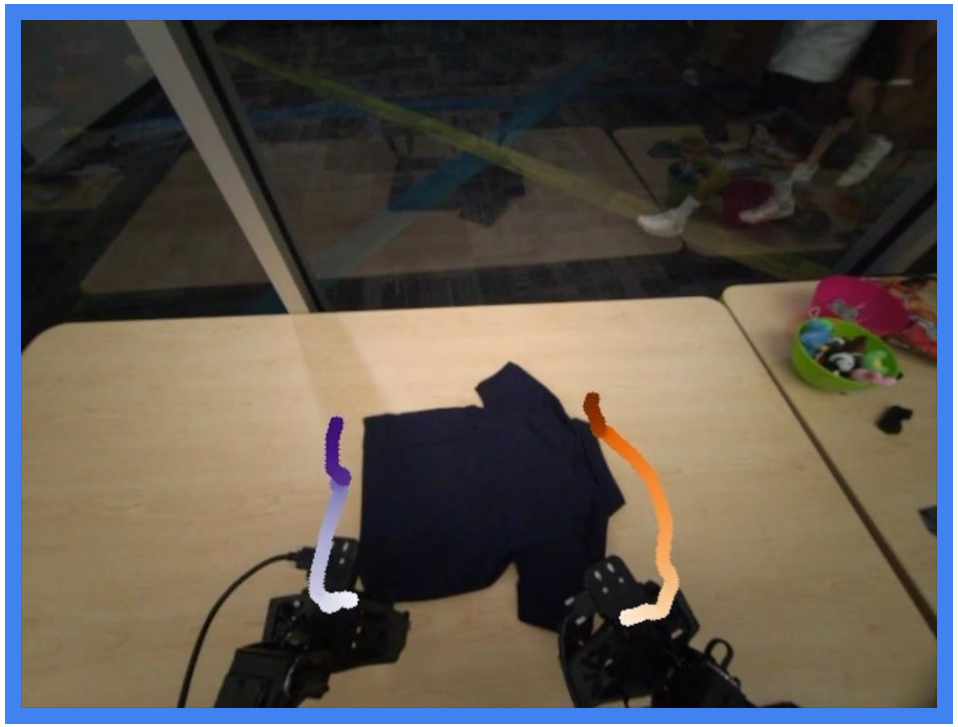
Project Aria Glasses:

- Ergonomic and light (75g)
- 5 Cameras (RGB, scene)
- SLAM + 3D hand tracking

Scaling Imitation Learning with Embodied Human Data

EgoMimic





Human Data

Robot Data

- Treating human data and robot data as *equals* by bridging sensing, action, and kinematic gaps.
- Joint policy learning that scales with human data.

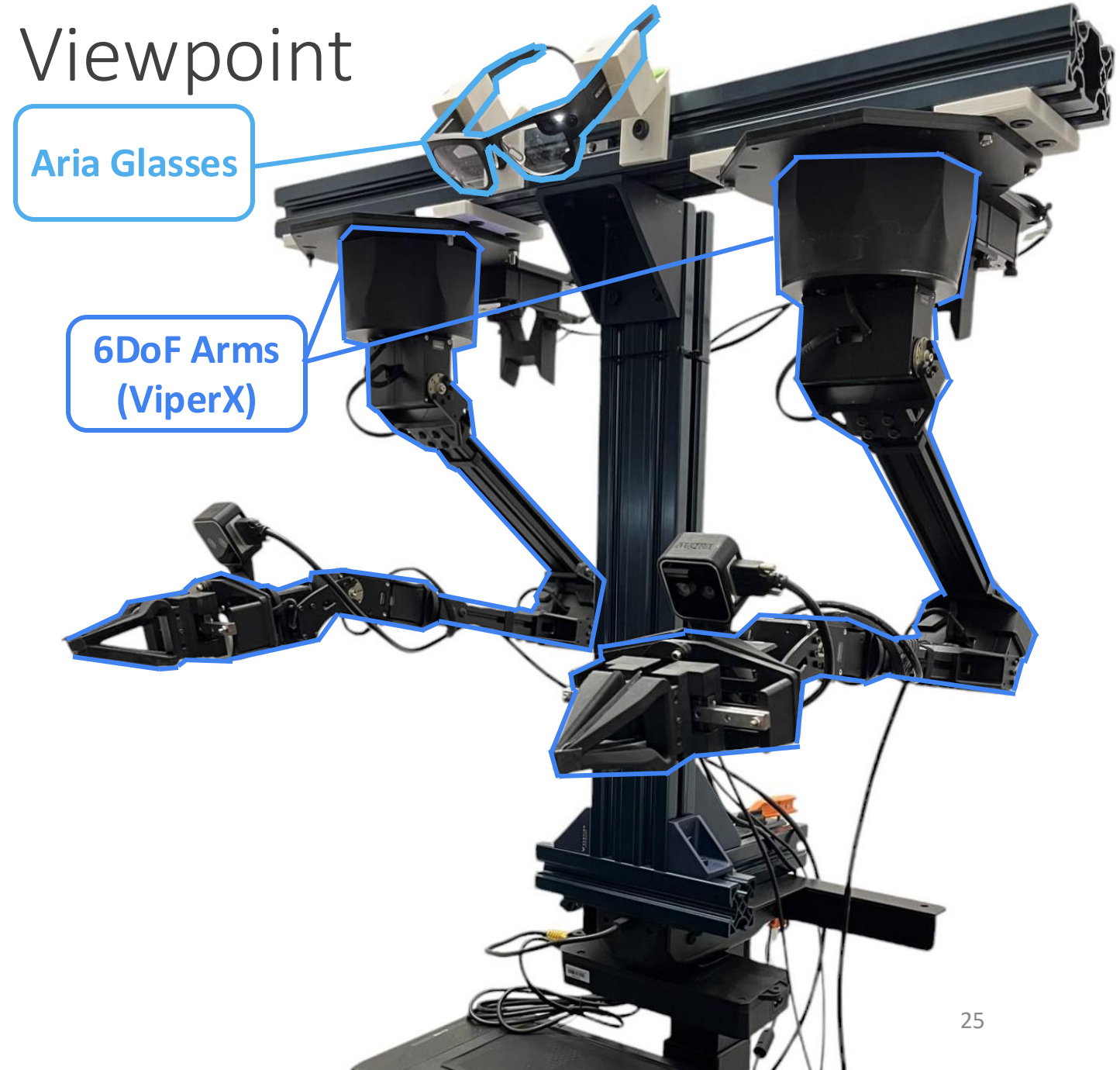
Kareer et al. “EgoMimic: Scaling Imitation Learning via Egocentric Video”, egomimic.github.io

Aligning Embodiment & Viewpoint

Embodiment:
- Anthropomorphism matters!
- Lightweight & lean
- Similar sizes in egocentric view

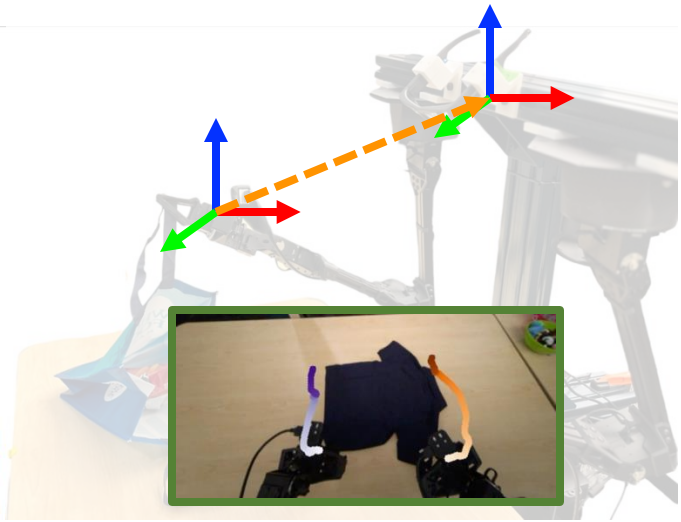
Viewpoint:
- Robot wearing Aria
- Eliminates inter-device differences
- Similar hand-eye configuration

Cost: ~$13,000 (excluding Aria)
BOM available at egomimic.github.io

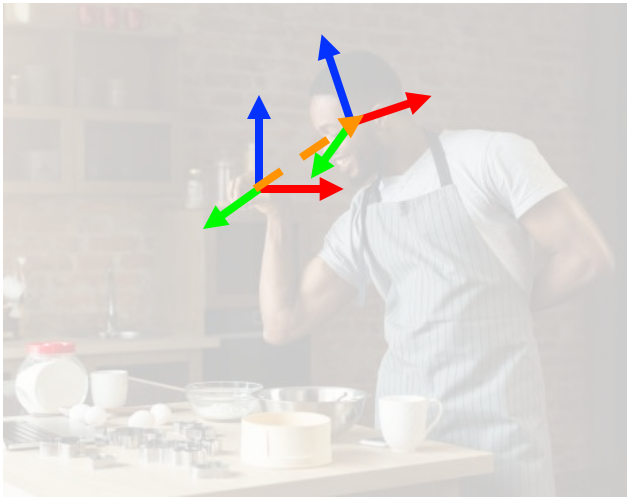


Stabilizing Egocentric Action Reference Frames

Solution: SLAM!



Fixed Camera (Aria) ->
Future gripper poses (actions) are relative to a fixed camera frame

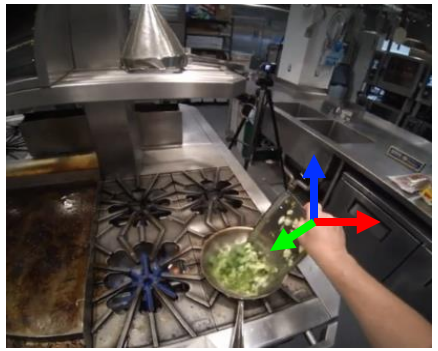


Moving Head (Aria) ->
Future hand poses are relative to a moving reference frame!

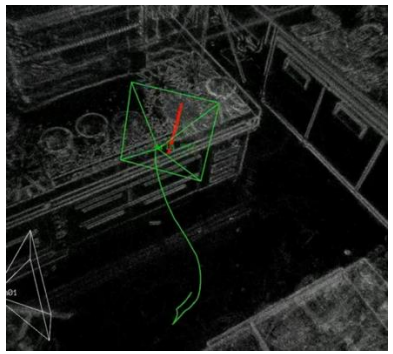
Need to stabilize the reference frame to use human hand tracks as actions!

Kareer et al. “EgoMimic: Scaling Imitation Learning via Egocentric Video”, egomimic.github.io

Stabilizing Egocentric Reference Frames

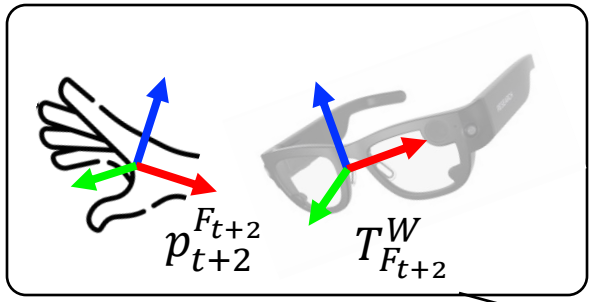


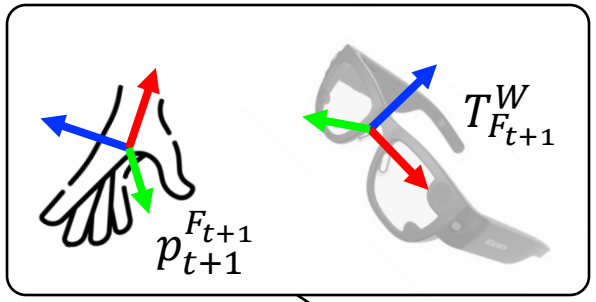
Hand Poses $p_t^{F_t}$



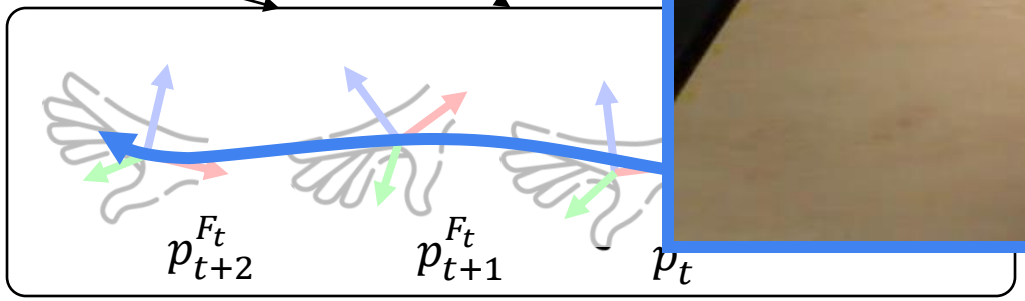
Device Tracks $T_{F_t}^W$

Goal: Transfer **future hand positions (actions)** into the **current camera frame** using device tracks





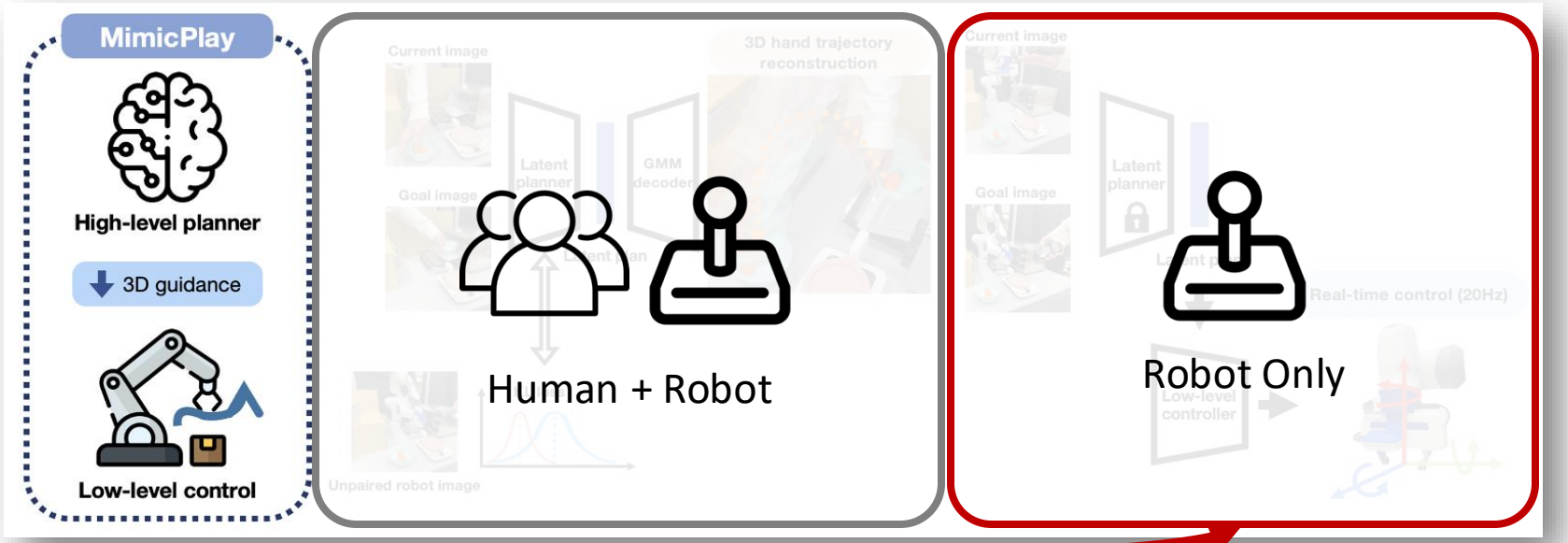
Future hand poses $p_{t:t+H}^{F_t}$ in the current camera frame F_t



Hand Tracks with Stable Reference Frame

Joint Human-Robot Policy Learning

MimicPlay Architecture: Hierarchical policy to learn “what” and “how”



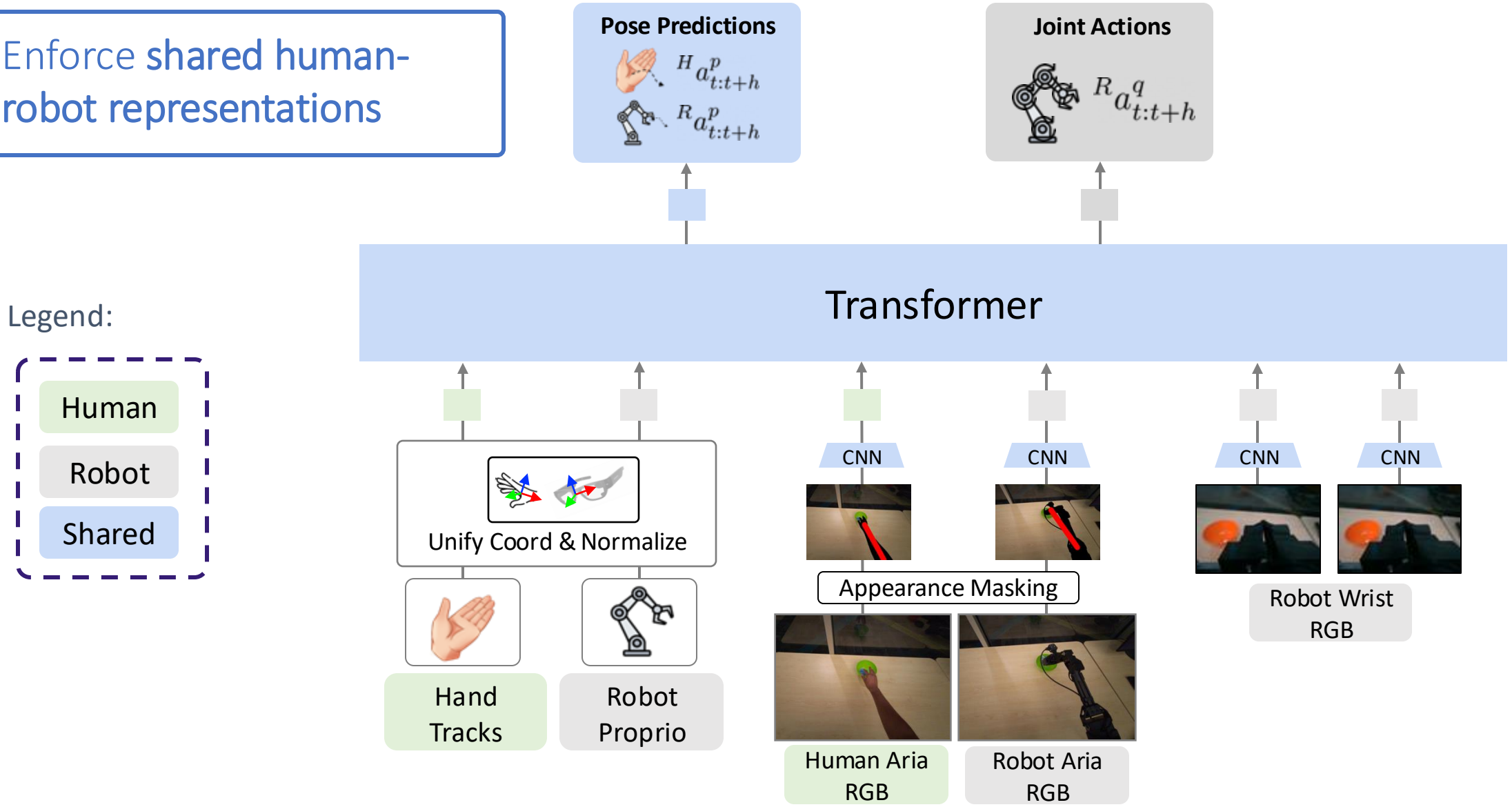
Low-level goal conditional policy is bottlenecked by robot data.

Key Insight: With embodied human data, we can use a simple, unified architecture that scales with human data.

Kareer et al. “EgoMimic: Scaling Imitation Learning via Egocentric Video”, egomimic.github.io

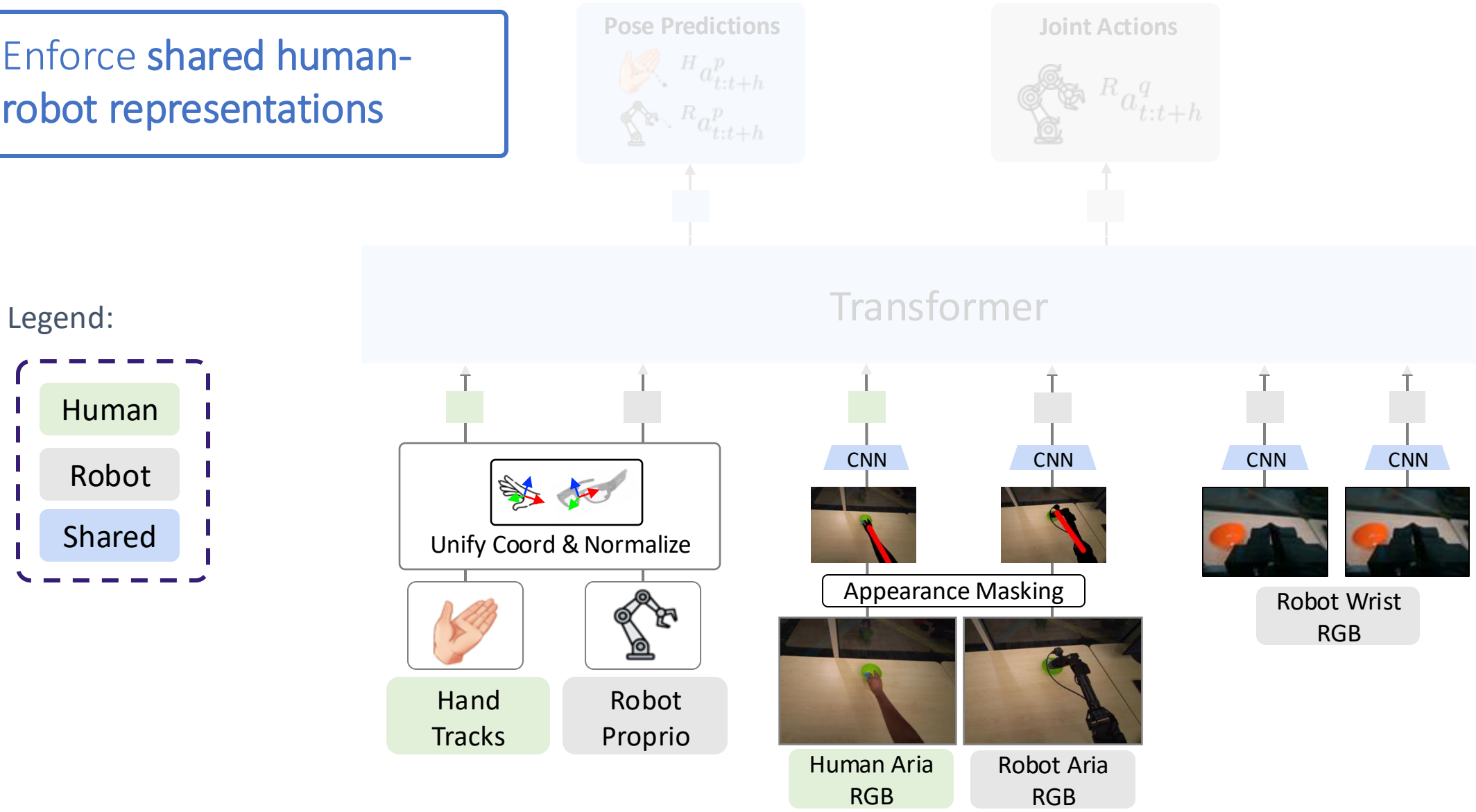
Joint Human-Robot Policy Learning

Enforce **shared human-robot representations**



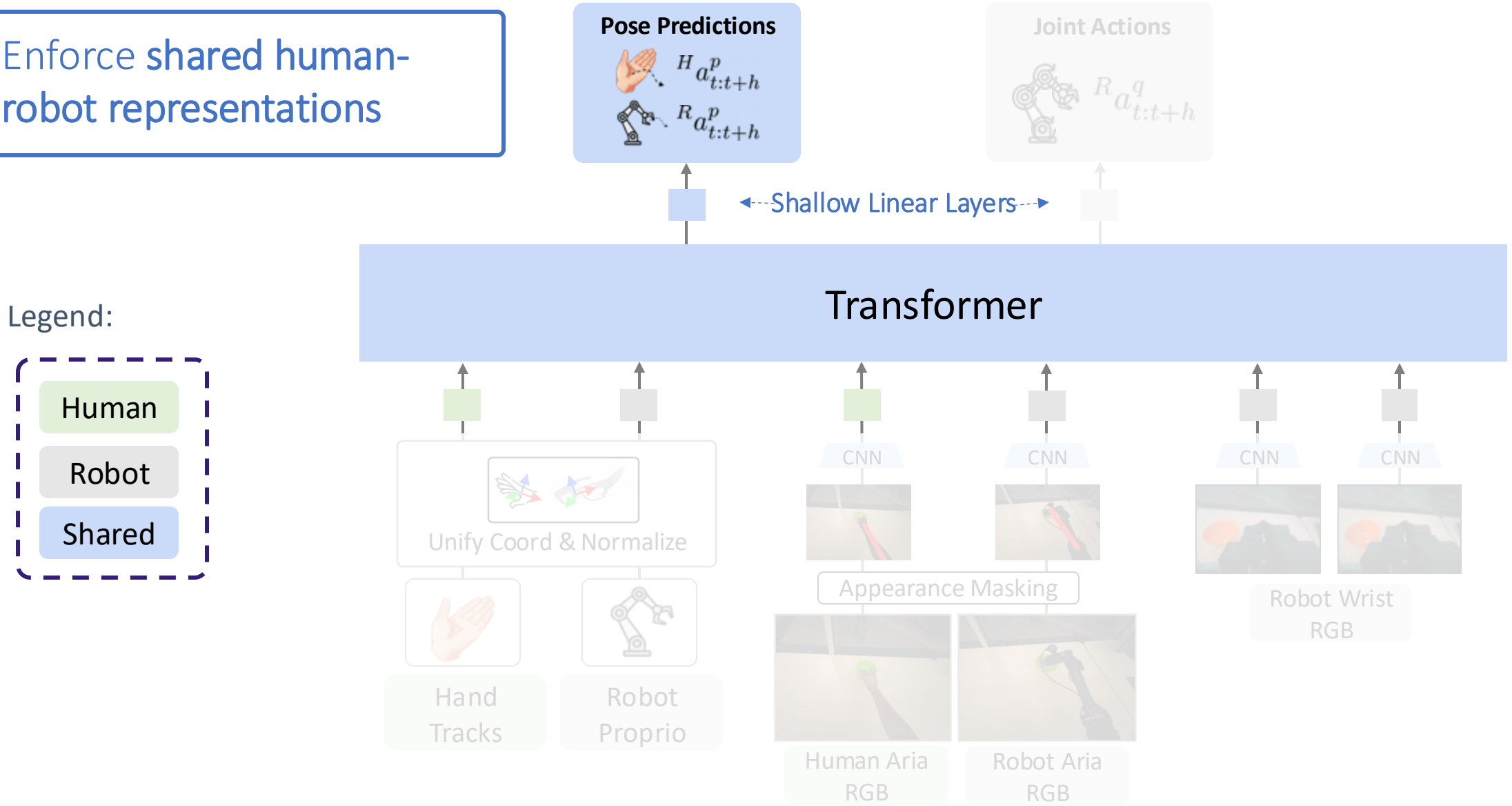
Joint Human-Robot Policy Learning

Enforce **shared human-robot representations**



Joint Human-Robot Policy Learning

Enforce **shared human-robot representations**



Does it work?

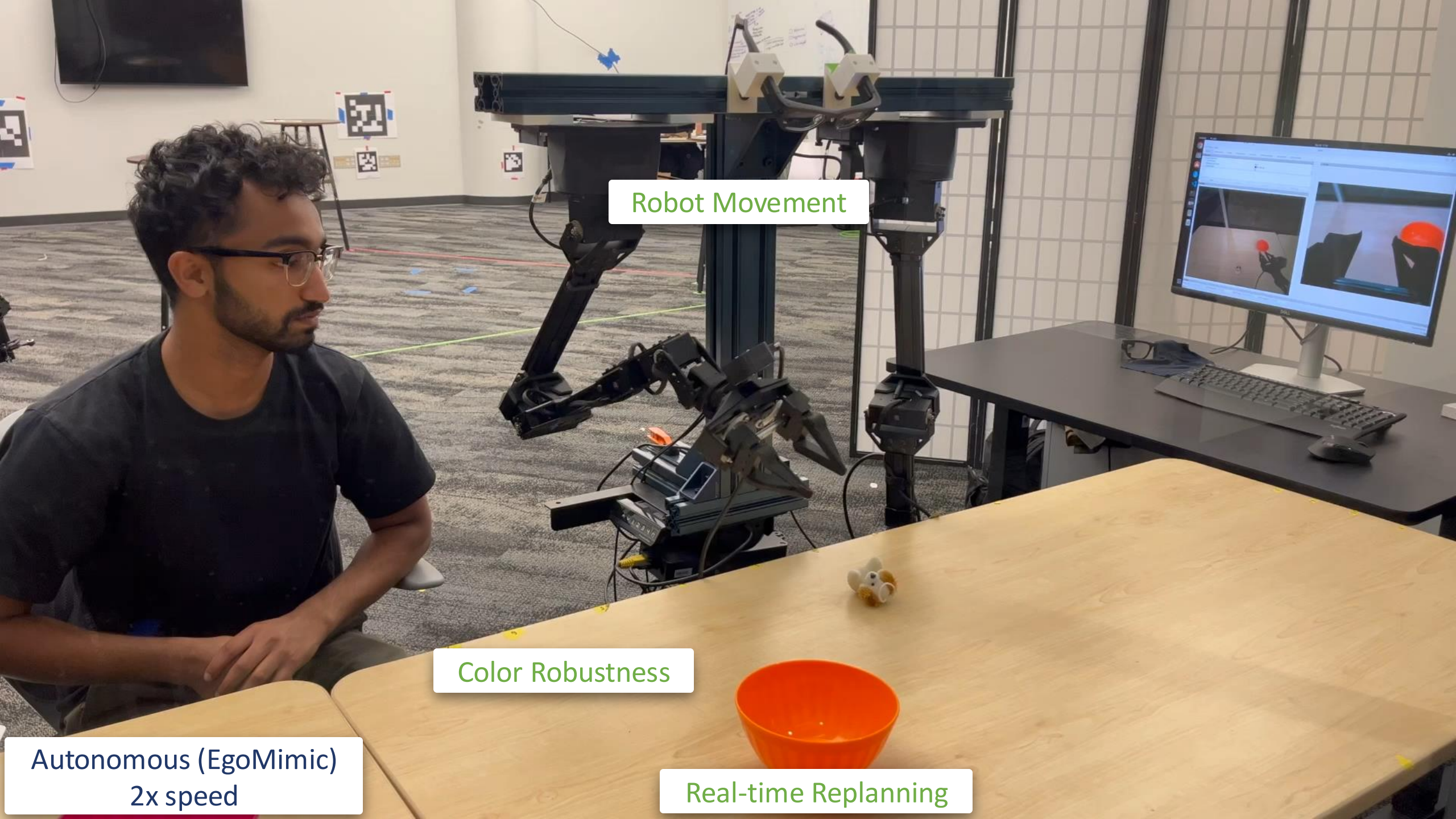
- What **tasks** can EgoMimic solve?
- How well does it **scale** with human data?
- Does it **generalize** beyond training scenarios?

Autonomous (EgoMimic)
4x speed

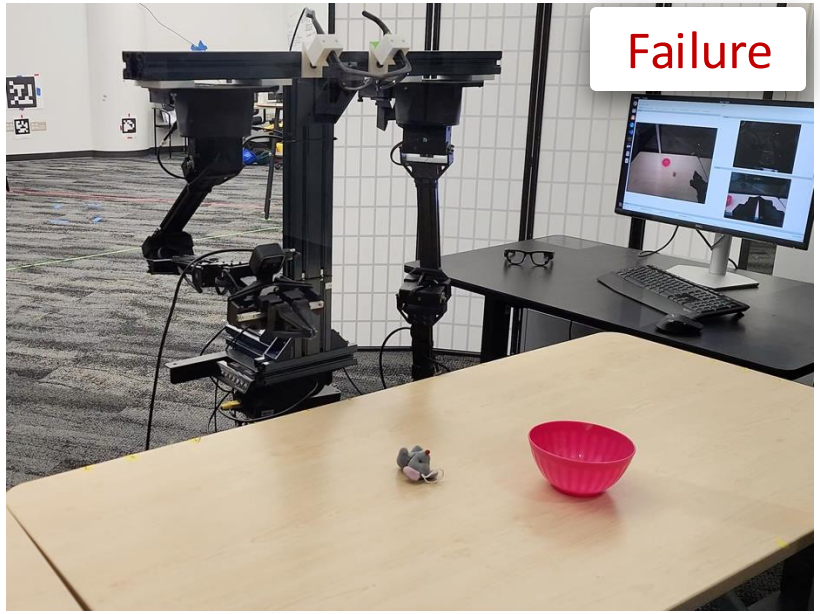

Autonomous (EgoMimic)
2x speed

Autonomous (EgoMimic)
2x speed

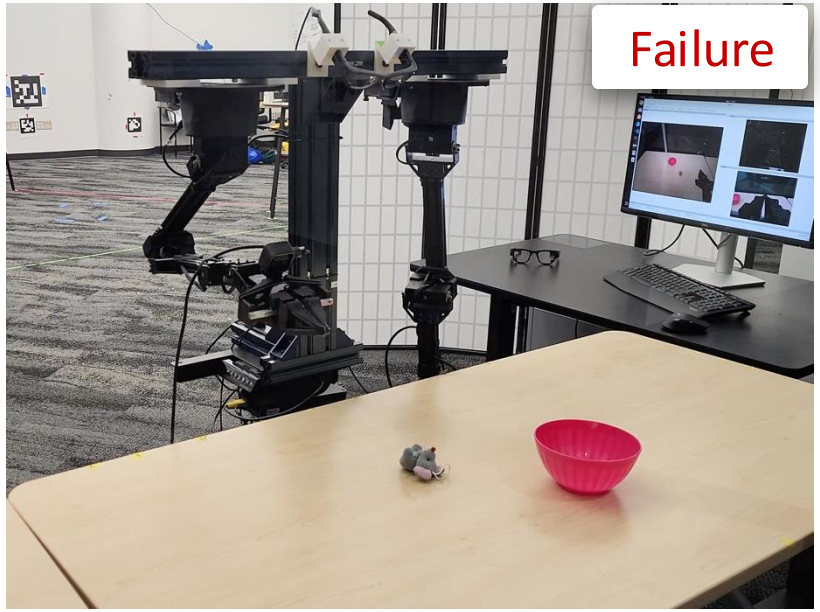

Autonomous (EgoMimic)
4x speed


Robot Movement
Color Robustness
Real-time Replanning
Autonomous (EgoMimic)
2x speed

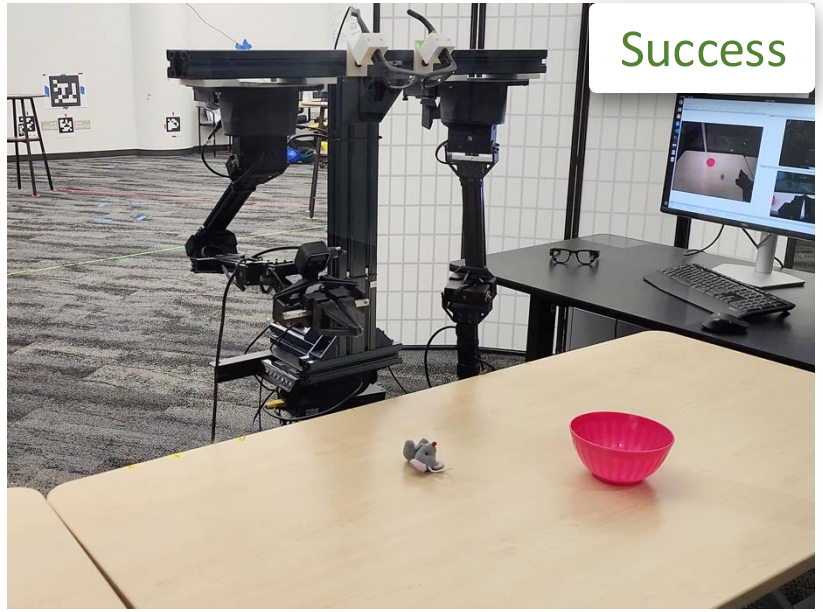
Does EgoMimic Scale with Human Data?



0% hand

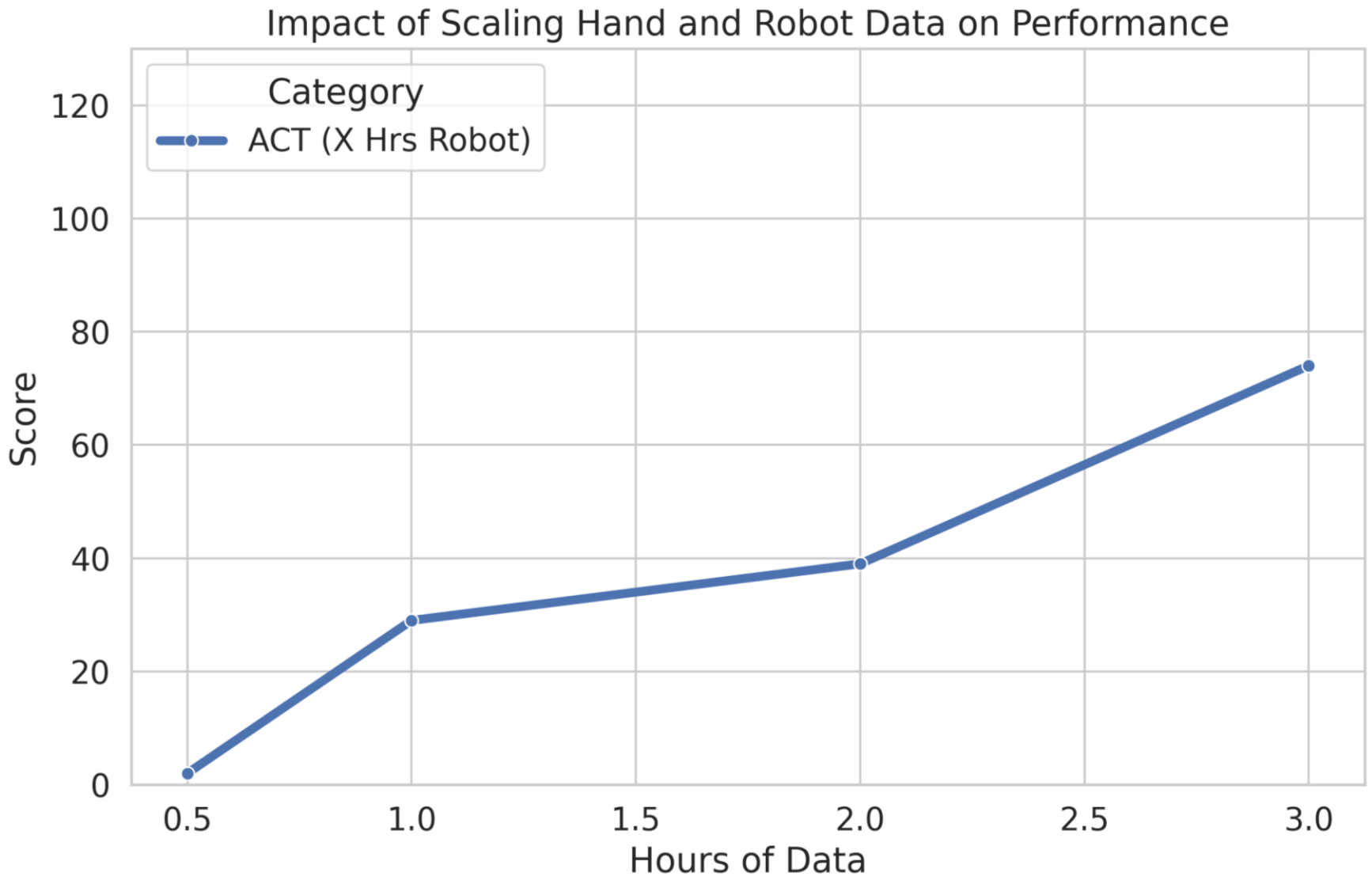


50% hand

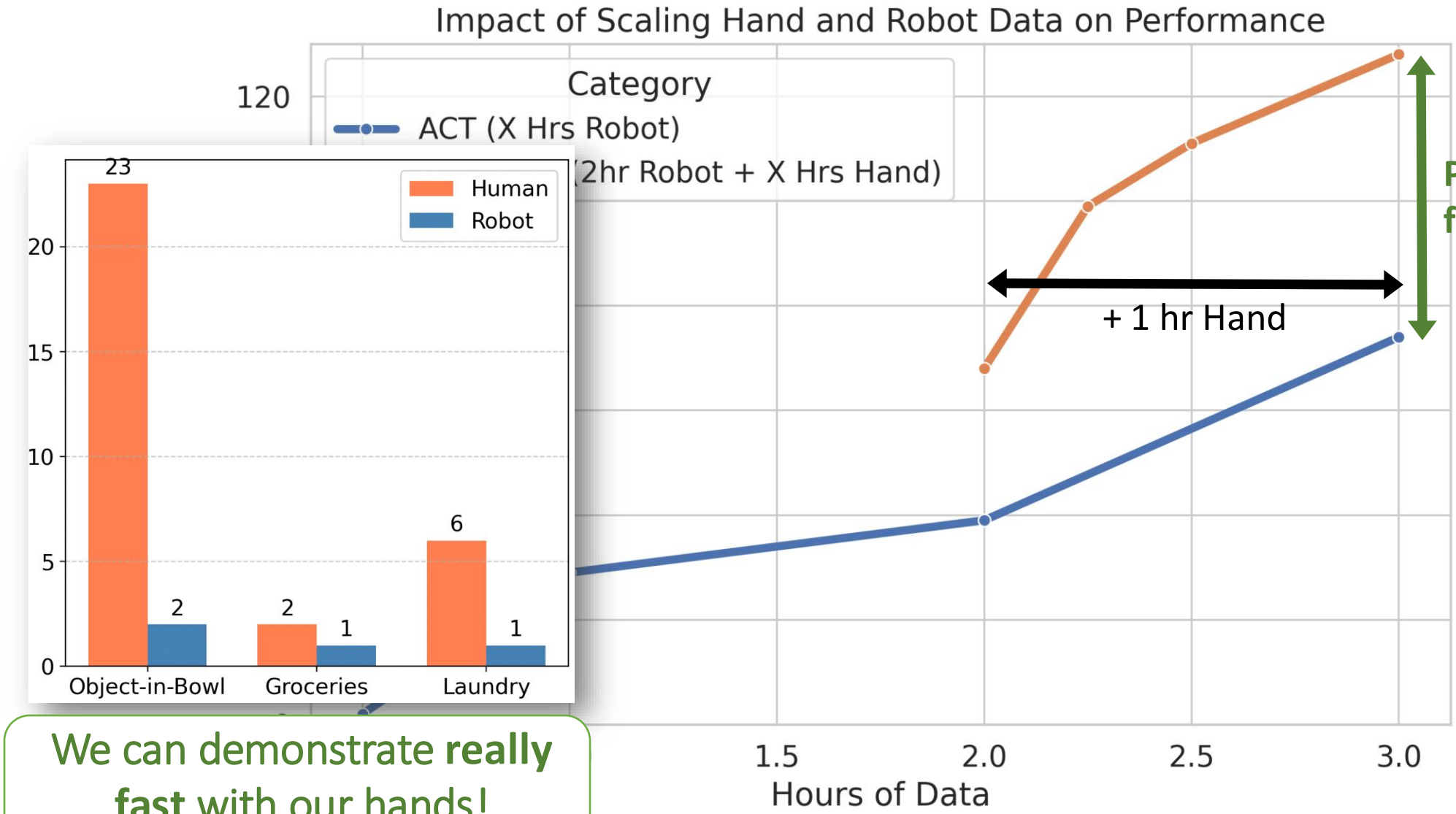


100% hand

Does EgoMimic Scale with Human Data?

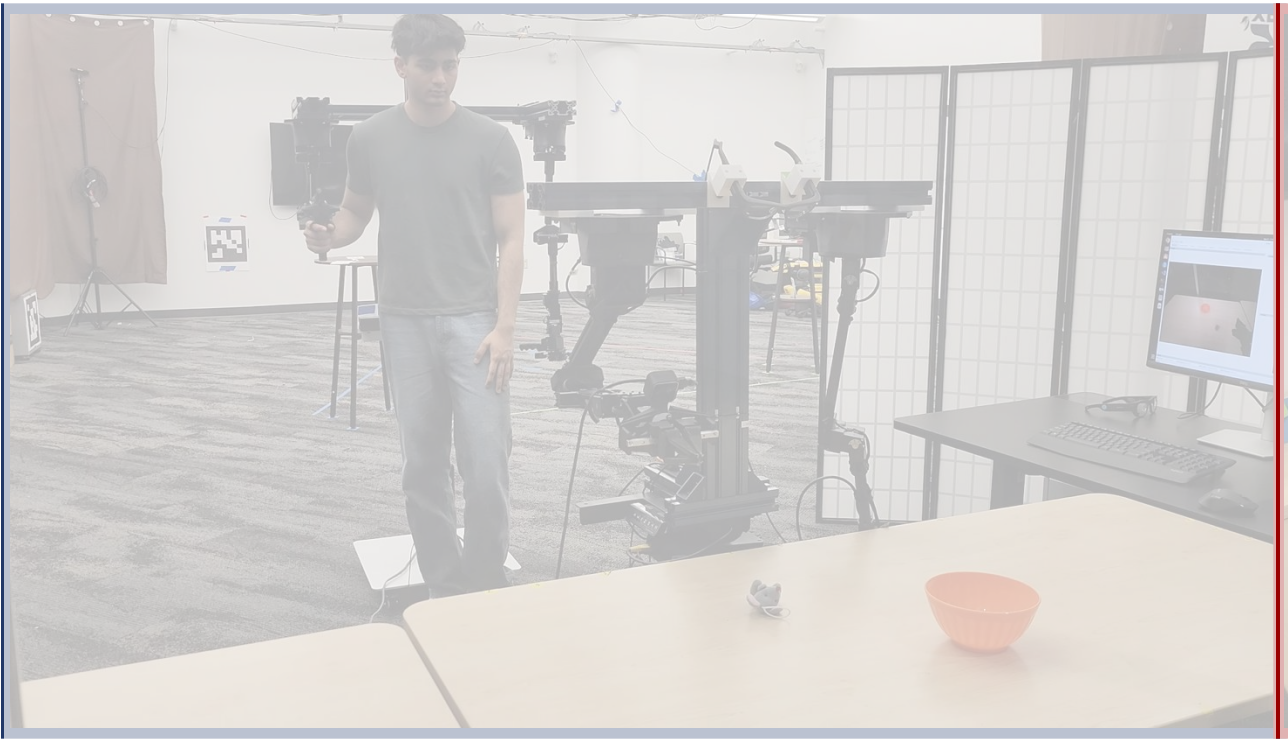


Does EgoMimic Scale with Human Data?



We can demonstrate **really fast** with our hands!

Scene Generalization

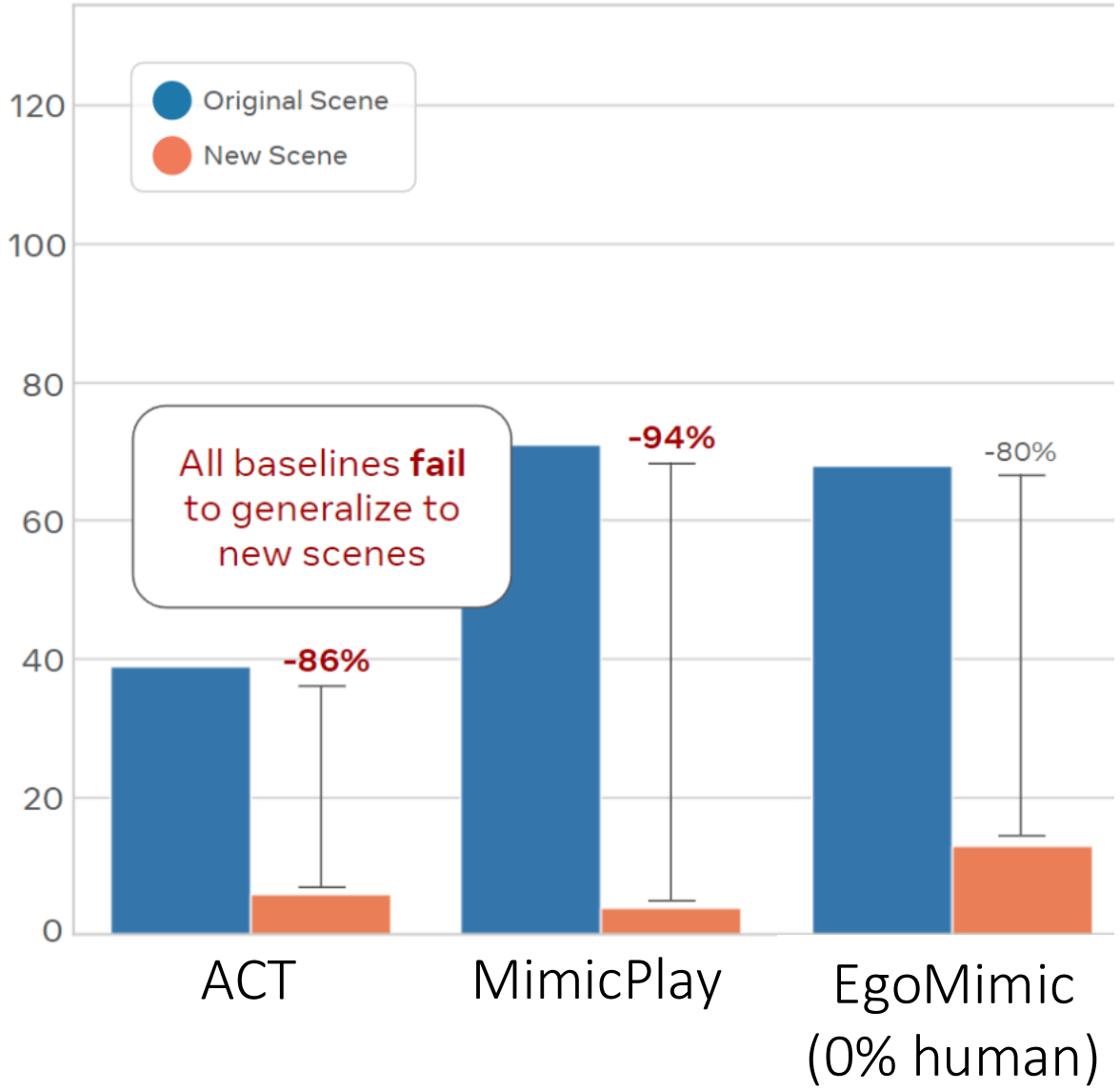


Robot data in original scene

Hand data in new scene


Unseen from Robot Data
Autonomous (EgoMimic)
2x speed

Generalization to New Scenes (Quantitative)



Scaling Imitation with Embodied Human Data: Summary

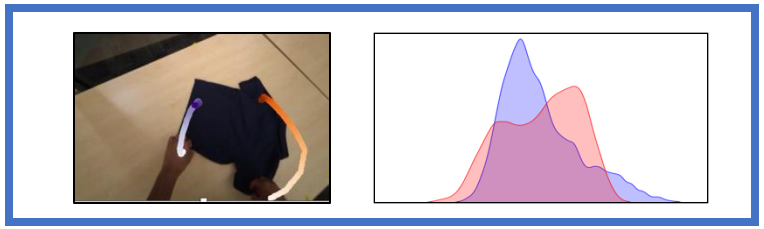
Human data is not *embodied* because:

- **Sensing**: Third-person videos ≠ embodied experience.
- **Actions**: Human data don't have action labels.
- **Appearance & Kinematics**: Humans and robots look different and have different kinematic structures.

Project Aria Glasses

- Video at eye level
- Ergonomic, lightweight (75g)
- SLAM + hand tracking
- Robot wearing aria

Hand Tracks + Action Alignment



- Hand tracking from Aria
- Stabilize reference frame with SLAM
- Action distribution normalization

Human-like Robot

- Light-weight, bimanual robot
- Human-like size & workspace
- Domain alignment based on image masking

EgoMimic: Scaling Imitation with Embodied Human Data


Simar Kareer

Dhruv Patel*


Ryan Punamiya*


Pranay Mathur*


Shuo Cheng


Chen Wang


Judy Hoffman†


Danfei Xu†

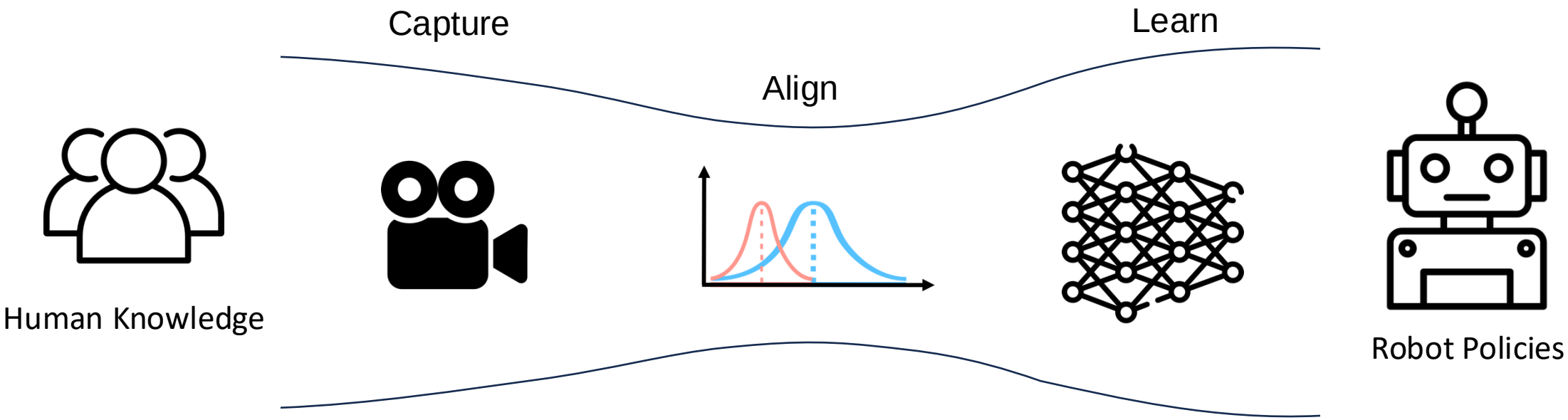
CoRL

Come talk to us!

- Workshop on X-Embodiment Robot Learning
- Workshop on Fine and Dexterous Manipulation

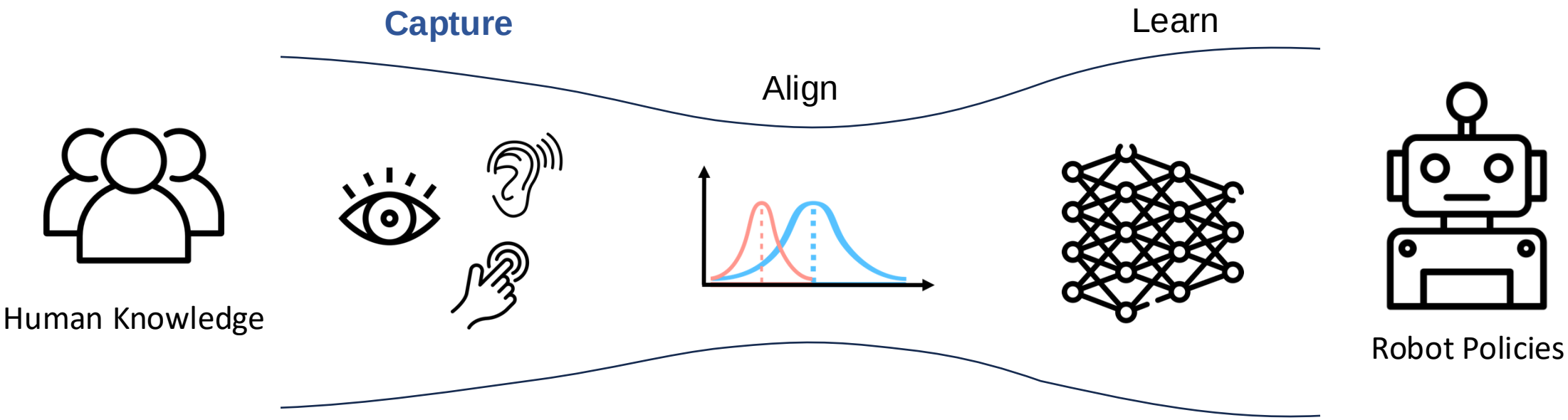
Visit egomimic.github.io to learn more!

Robot Learning from Embodied Human Data



- **Collected from humans**, not robots
- Contain **embodied sensorimotor** experiences
- Captured with **ubiquitous** devices

Robot Learning from Multimodal Embodied Human Data



- **Collected from humans**, not robots
- Contain **embodied sensorimotor** experiences
- Captured with **ubiquitous** devices

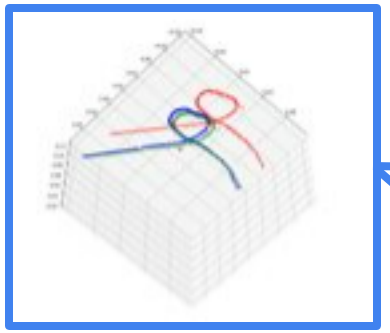
MimicTouch: Learning from Human Tactile Data

Kelin Yu

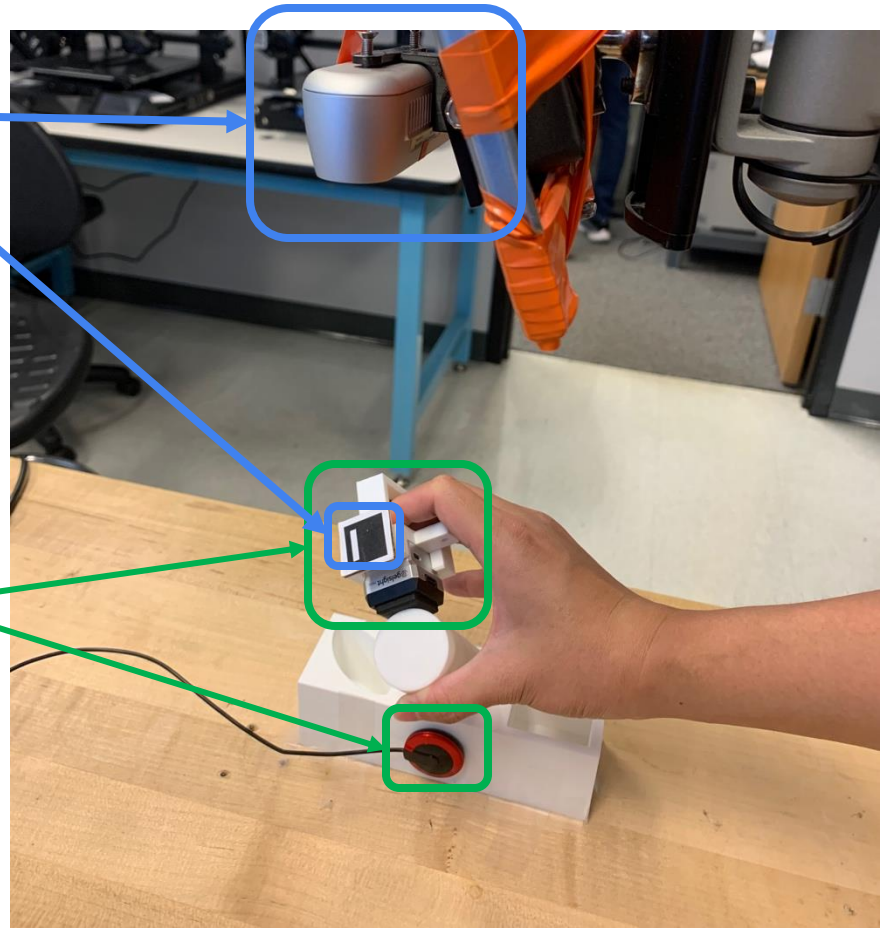
Yunhai Han

CoRL

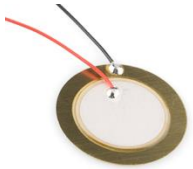
Poster Session #4
Friday 4pm



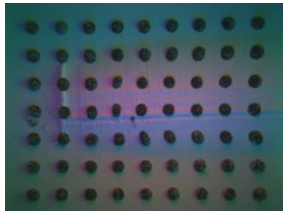
Device Tracking



GelSight Mini

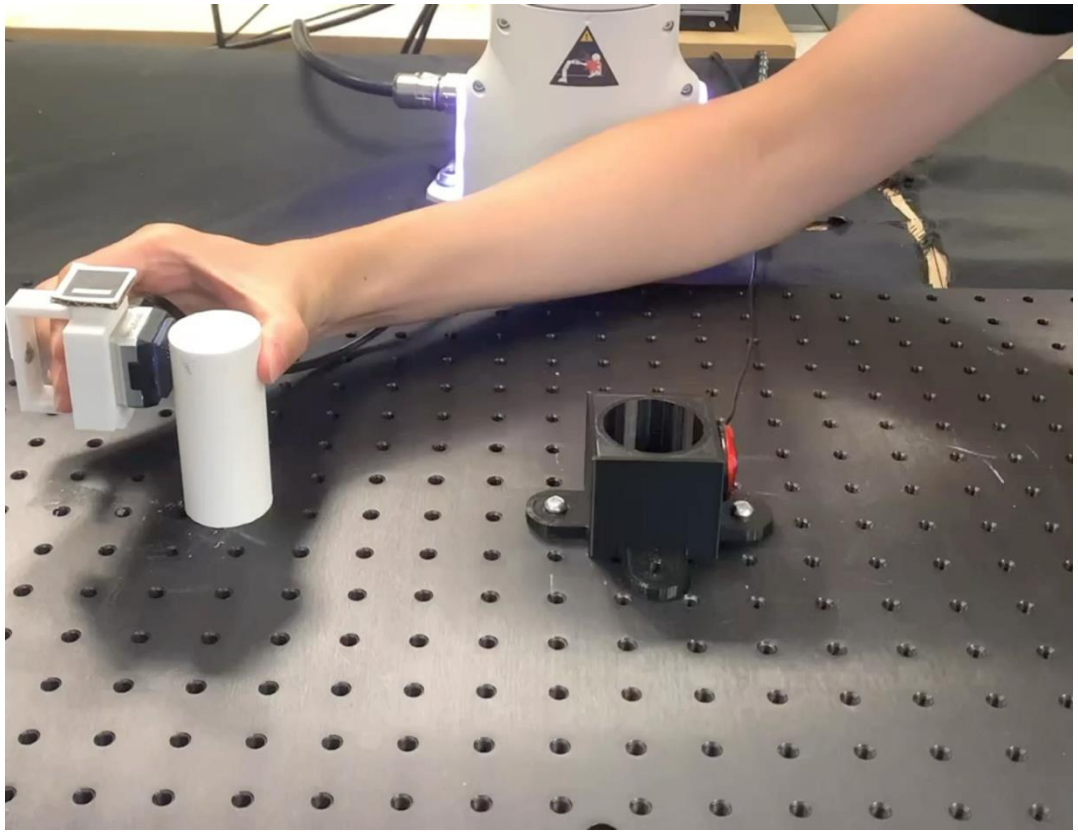


Piezoelectric
Contact Mic

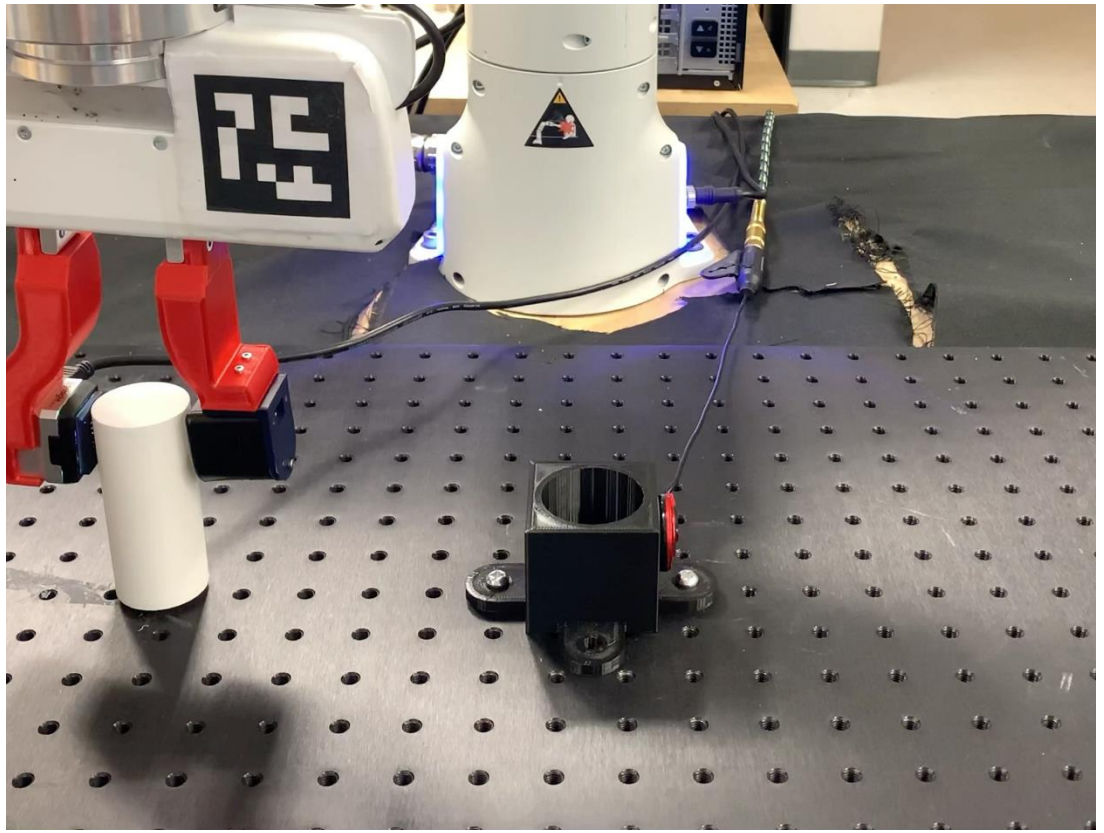


Yu and Han et al. “MimicTouch: Leveraging Multi-modal Human Tactile Demonstrations for Contact-rich Manipulation”, *CoRL 2024*

MimicTouch: Learning from Human Tactile Data



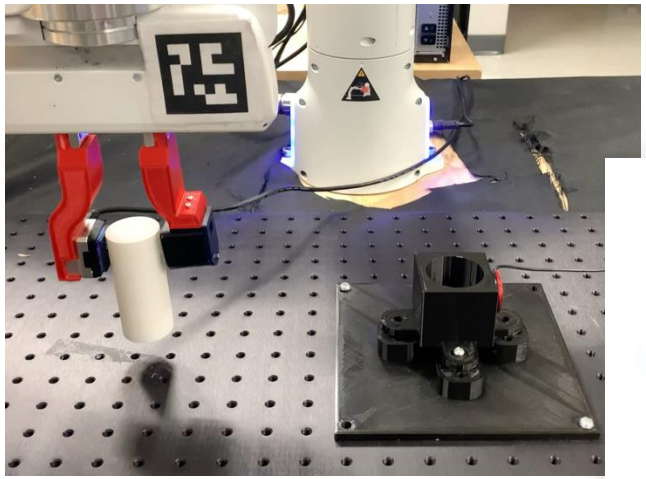
Hand-worn device to collect
tactile-audio demonstrations



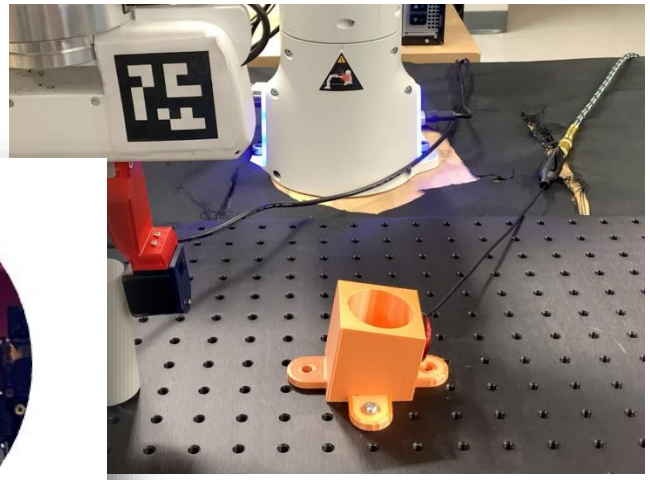
Learn tactile-audio policies
that can run on robots
(KNN-IL + RL finetuning)

Yu and Han et al. “MimicTouch: Leveraging Multi-modal Human Tactile Demonstrations for Contact-rich Manipulation”, *CoRL 2024*

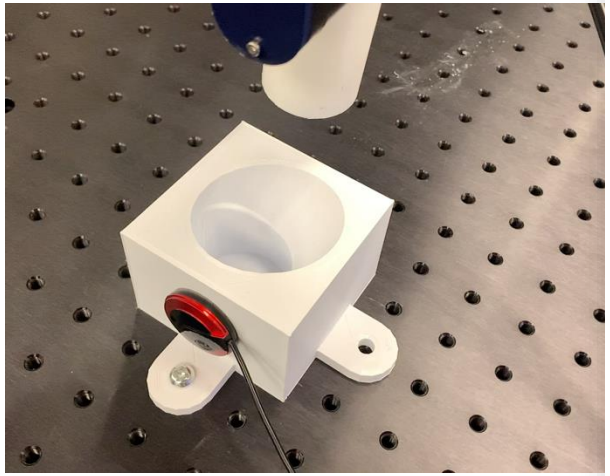
Trained policies zero-shot generalize to new settings



Base translation (10cm)



ilt right (-10 degrees)

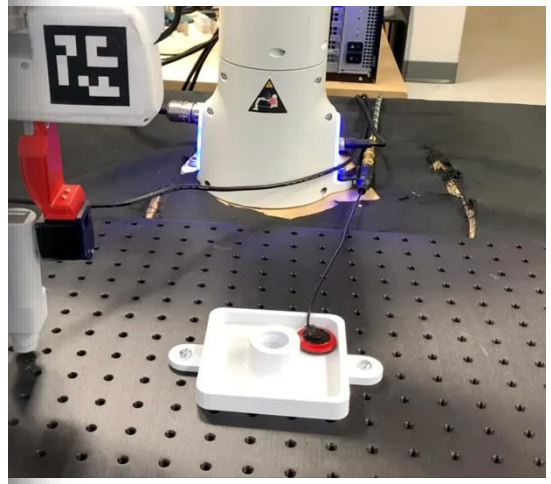


Two-stage insertion

Kelin Yu Yunhai Han

CoRL

Poster Session #4
Friday 4pm



Screw assembly

Yu and Han et al. “MimicTouch: Leveraging Multi-modal Human Tactile Demonstrations for Contact-rich Manipulation”, *CoRL 2024*

Robot Learning from Embodied Human Data

***MimicPlay*:**
Learning from “Unembodied” Human Data

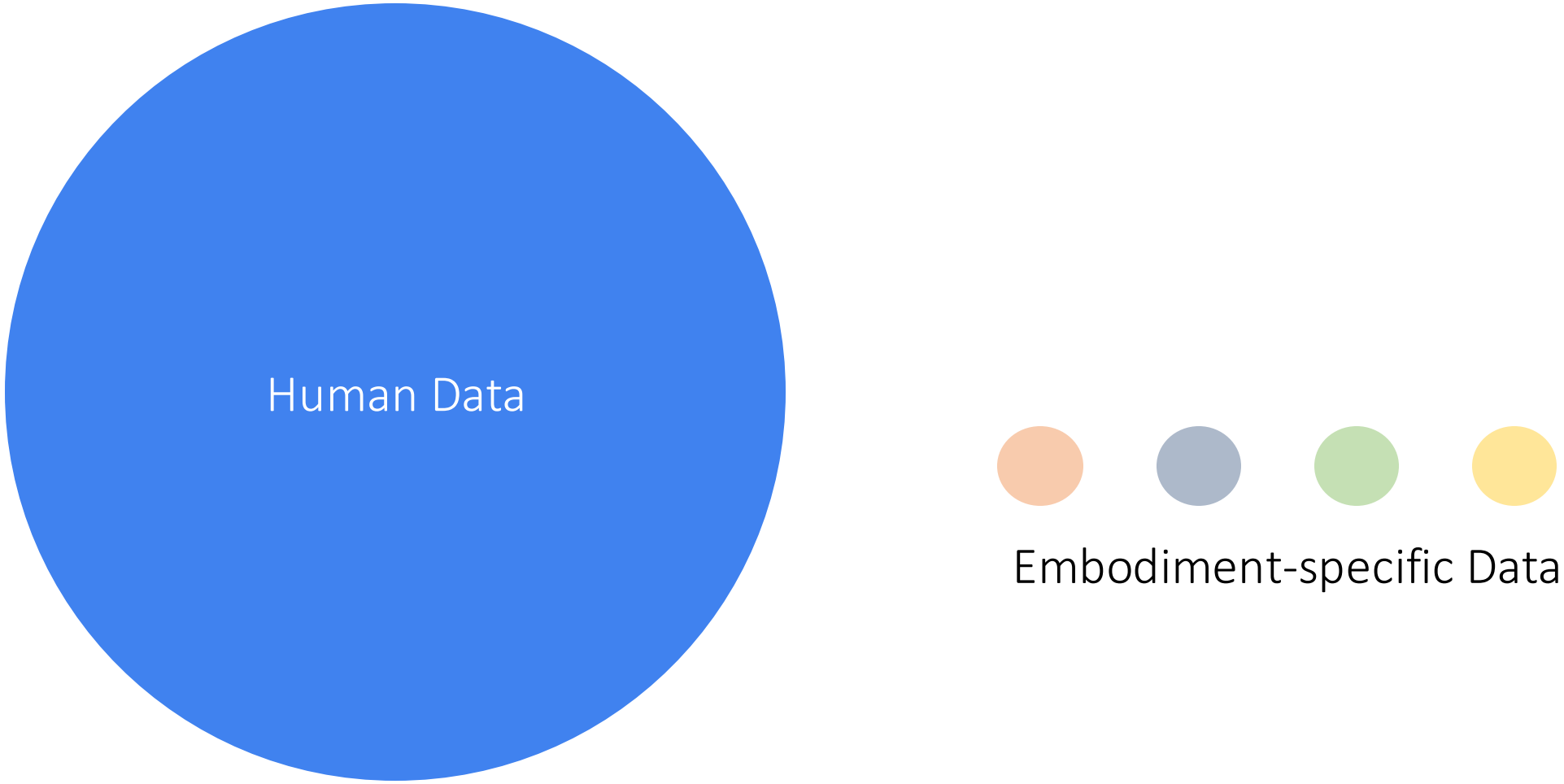
***EgoMimic*:**
Embodied Human Data as Robot Data

***MimicTouch*:**
Learning from Multimodal Embodied Experiences

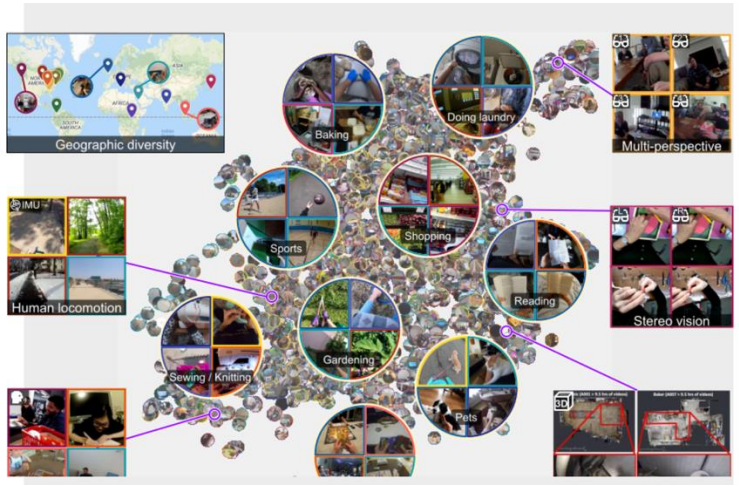
Beyond Scaling Up:
Robot Learning from a Human Perspective

Embodied Human Data as Foundation Dataset?

Human embodiment is universal & has scalable data source

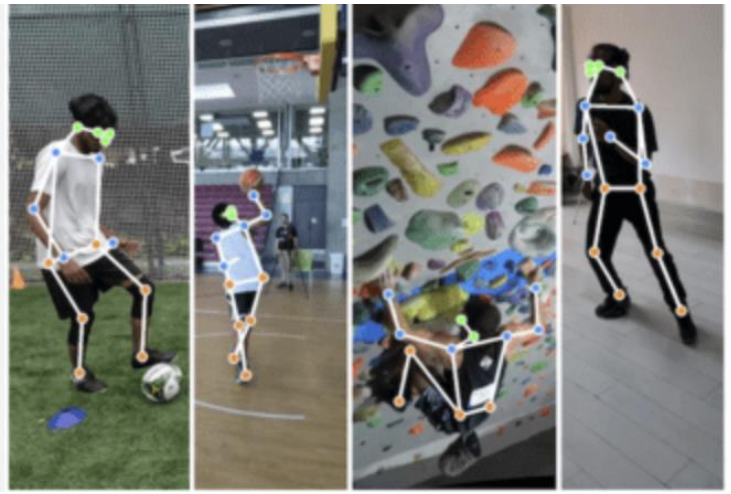


Can we leverage existing egocentric data?



Ego4D (Ego4D Consortium, 2021)

Epic Kitchens (Damen et al., 2020)

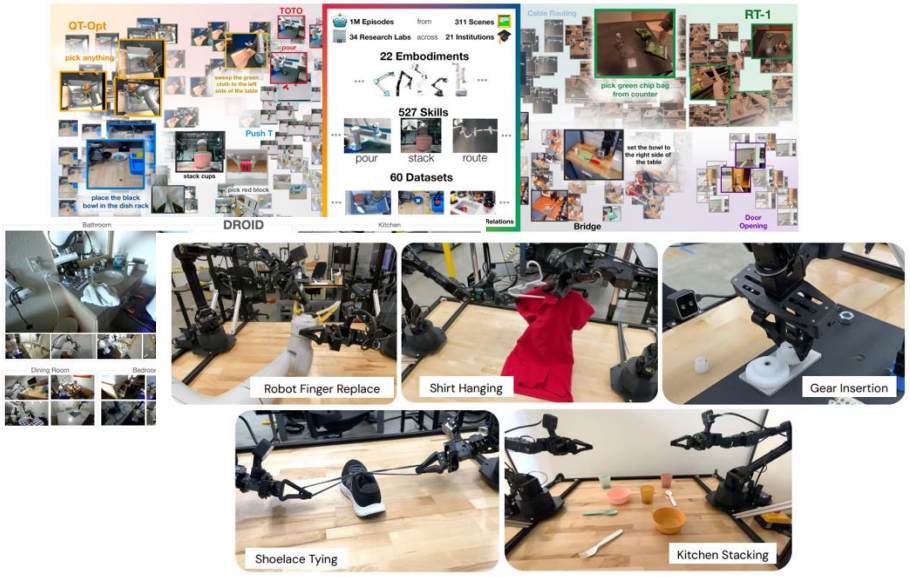


Ego-Exo4D (Ego4D Consortium, 2024)

Not easily … many problems:

- **No 3D hand tracks** (Ego4D, EpicKitchen)
- **No SLAM** (Ego4D, EpicKitchen)
- Most activities **unsuitable for robot** (Ego-Exo4D)

We need an egocentric dataset built for Robot Learning



?

Wishlist:

- Diverse tasks that are suitable for robots
- Split by task types: bimanual, mobile manipulation, navigation, etc.
- Learning-relevant annotations: SLAM, 3D Tracks, point cloud, etc.
- …

Robot Learning from Embodied Human Data

MimicPlay:
Learning from “Unembodied” Human Data

EgoMimic:
Embodied Human Data as Robot Data

MimicTouch:
Learning from Multimodal Embodied Experiences

Beyond Scaling Up:
Robot Learning from a Human Perspective

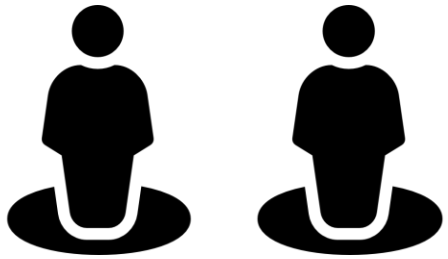
Beyond Scaling Up: Robot Learning from a Human Perspective

What is my human partner's intent / goal / next action?

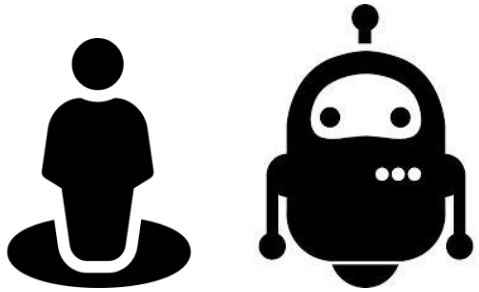
There is tremendous value in machines understanding us based on our embodied experience.
Now we have a data-centric path to achieve this.

Co-GAIL: Learning for Human-Robot Collaboration

Human-human
collaboration data



Human-robot
collaboration

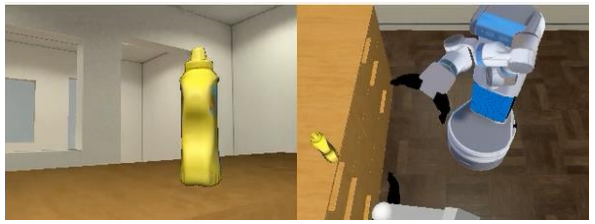


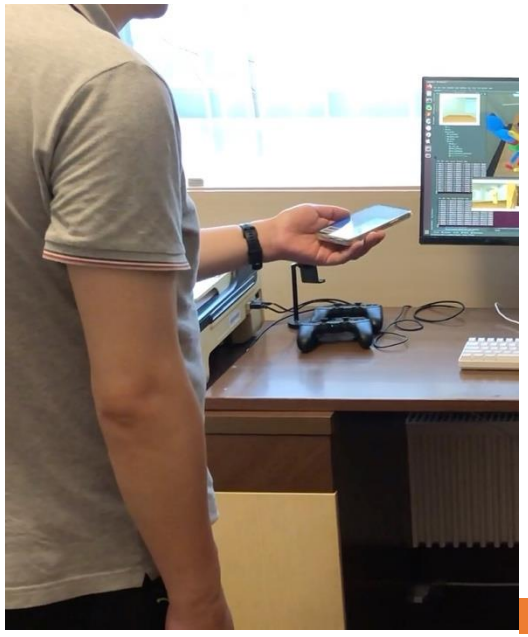
Chen Wang

Wang et al., Co-GAIL: Learning Diverse Strategies for Human-Robot Collaboration (CoRL 2021)

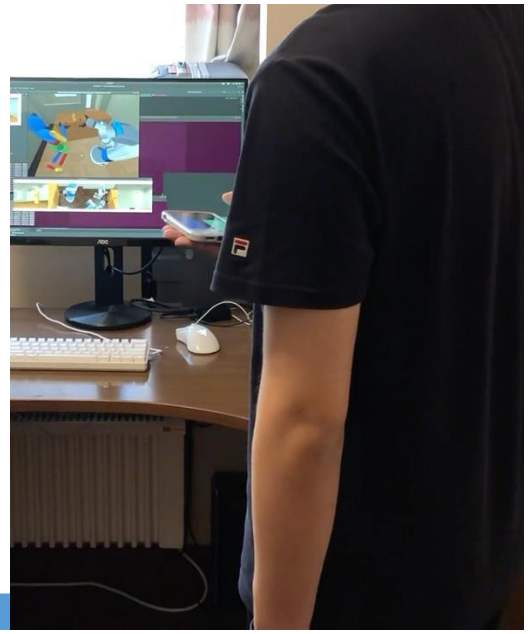
Collecting Human-Human Collaboration Data

Human operator view



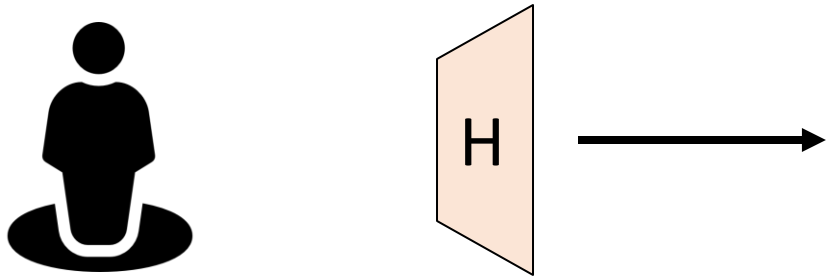


Robot operator view



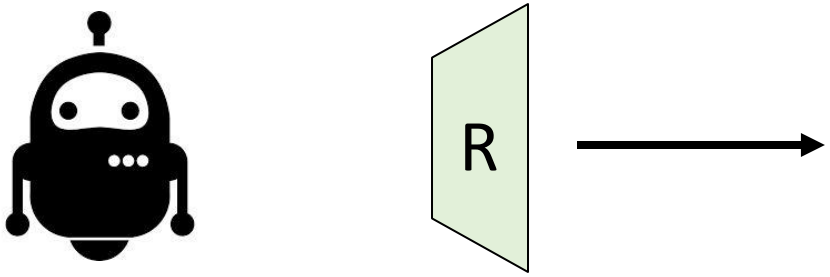
Wang et al., Co-GAIL: Learning Diverse Strategies for Human-Robot Collaboration (CoRL 2021)

Robot Learning from a Generative Human Model



Generative human model

generate diverse and plausible human collaborative behaviors by learning from human-human interaction data



Robot policy model

Learn to assist humans effectively by practicing with a generative human model

Wang et al., Co-GAIL: Learning Diverse Strategies for Human-Robot Collaboration (CoRL 2021)

Collaborating with a Real (teleoperated) Human

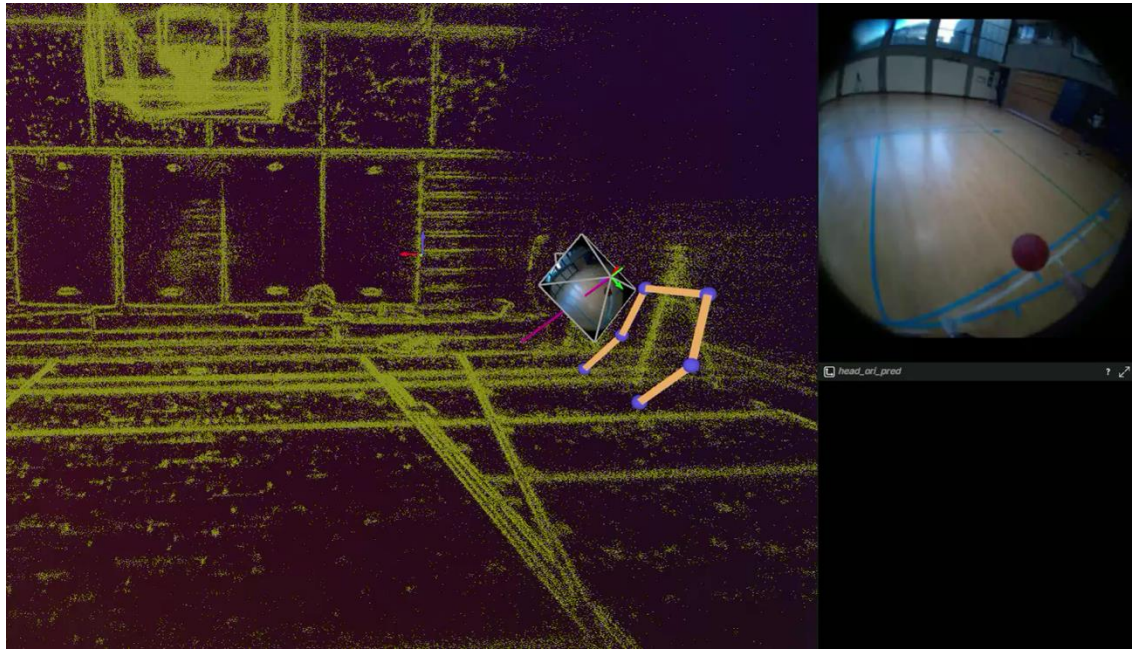
Challenge: Getting human-human collaboration data useful for robot learning.

Wang et al., Co-GAIL: Learning Diverse Strategies for Human-Robot Collaboration (CoRL 2021)

Future: Learning from Egocentric Collaboration Data

Source: Project Aria & University of Bristol

Generative Human Models from Egocentric Data

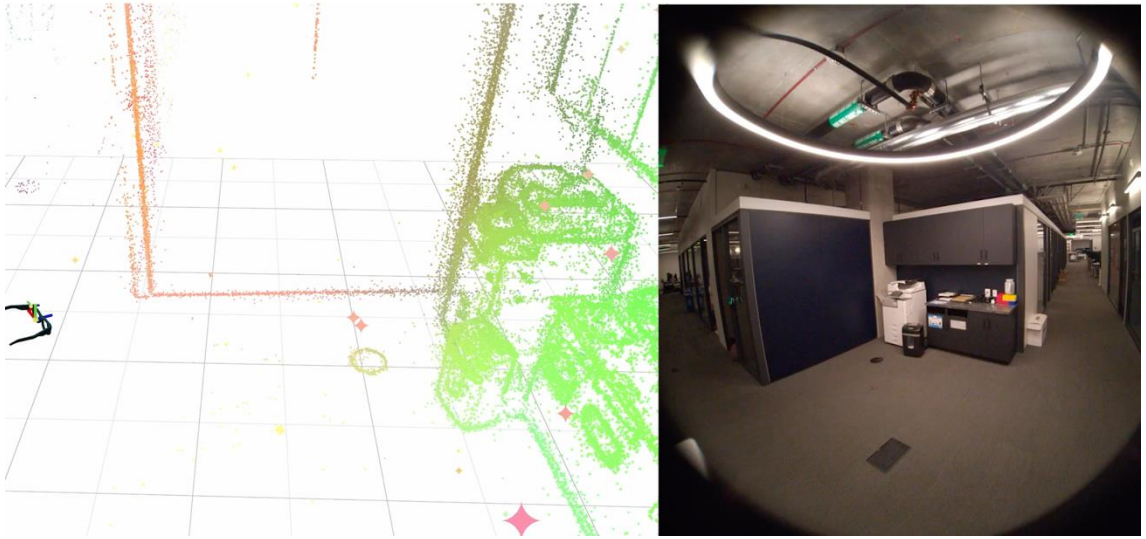


Wenqi Jia

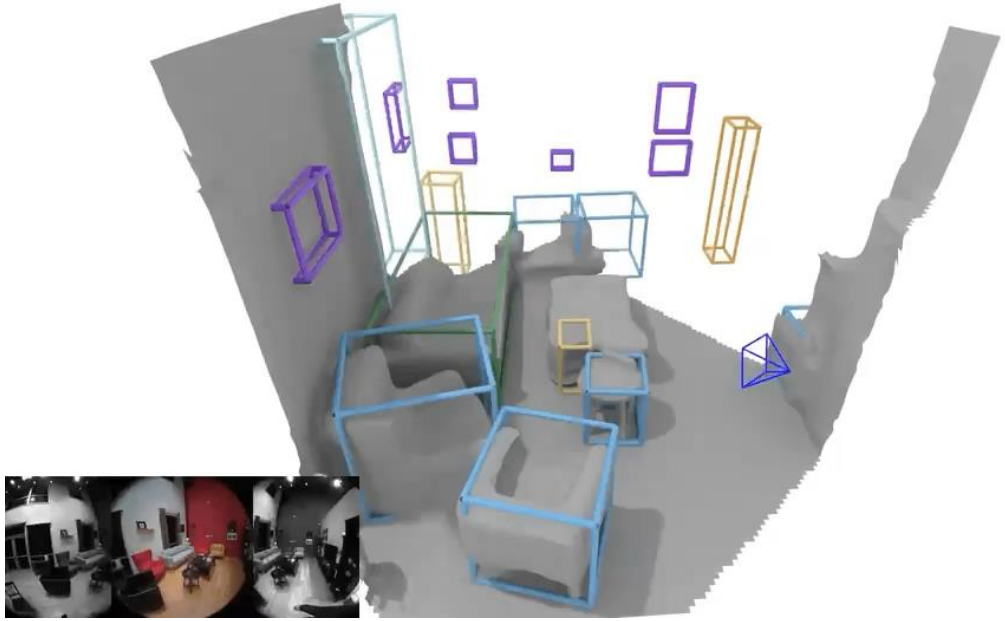
Generate plausible future behaviors (head, hand, gaze predictions) given egocentric observations (images, poses).

Shared Progress with the XR & Wearable AI Community

EgoAllo uses egocentric ( ) SLAM poses and images

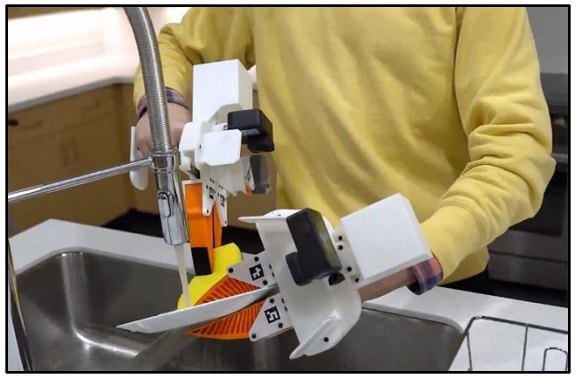


EgoAllo, Yi et al., 2024



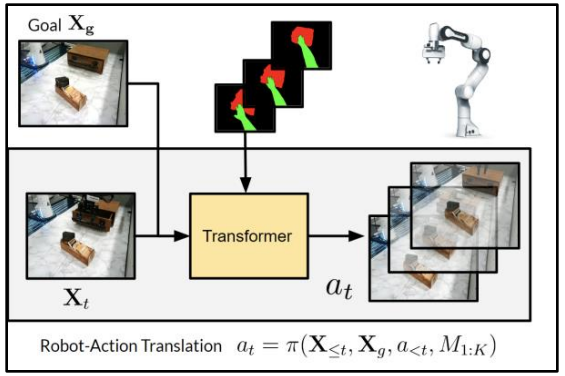
EFM3D Benchmark, Straub et al., 2024

Many exciting works from the community!

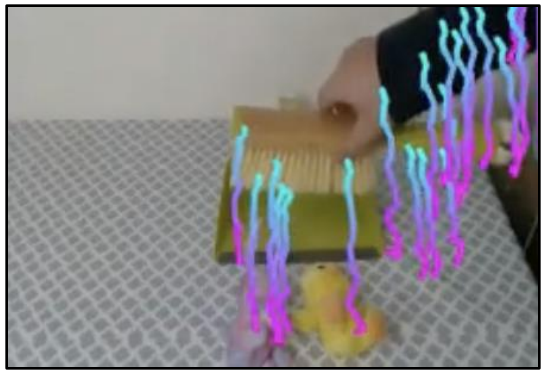


UMI
(Chi *et al.*)

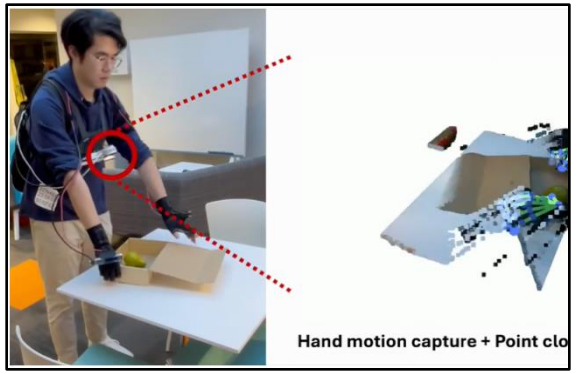
Robot Utility Model
(Etukuru *et al.*)



HOPMAN
(Bharadhwaj *et al.*)

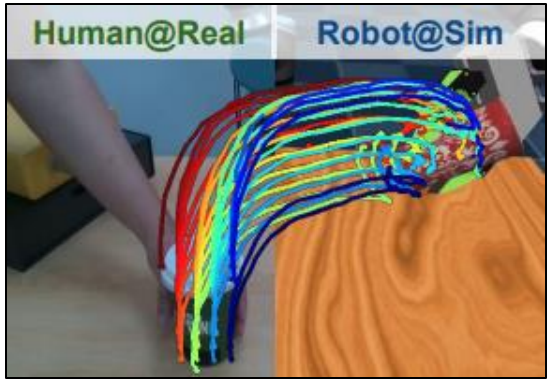


ATM
(Wen, Lin, So, *et al.*)



DexCap
(Wang *et al.*)

AR2-D2
(Duan *et al.*)



Im2Flow2Act
(Xu *et al.*)

...

Why Should *You* be Excited About Embodied Human Data?

- Embodied human data is robot data in disguise
- The data capturing technology is ready & advancing
- Robots are becoming more human-like and accessible
- Next-generation foundation models will understand humans at a sensorimotor level

Robot Learning from Embodied Human Data

1. *MimicPlay*: Long-Horizon Imitation Learning by Watching Human Play (CoRL 2023)
2. *EgoMimic*: Scaling Imitation Learning through Egocentric Video *(*CoRL 2024-WS*)*
3. *MimicTouch*: Leveraging Multi-modal Human Tactile Demonstrations for Contact-rich Manipulation (CoRL 2024)
4. *Co-GAIL*: Learning Diverse Strategies for Human-Robot Collaboration (CoRL 2021)

Thanks!

danfei@gatech.edu

School of Interactive Computing at Georgia Tech is Hiring!

https://www.ic.gatech.edu/faculty-hiring