Embodied AI: Perception, Representation and Action

Imitation Learning and Behavior Cloning

From Expert Demonstrations to Generalist Policies

Module 3 — Learning Methods

Today's Lecture: Five Objectives in One Causal Chain

Where RL Left Us

Reinforcement learning requires millions of environment interactions, reward specification that is fragile and prone to hacking, exploration cost that makes physical deployment unsafe, and long-horizon credit assignment that remains brittle.

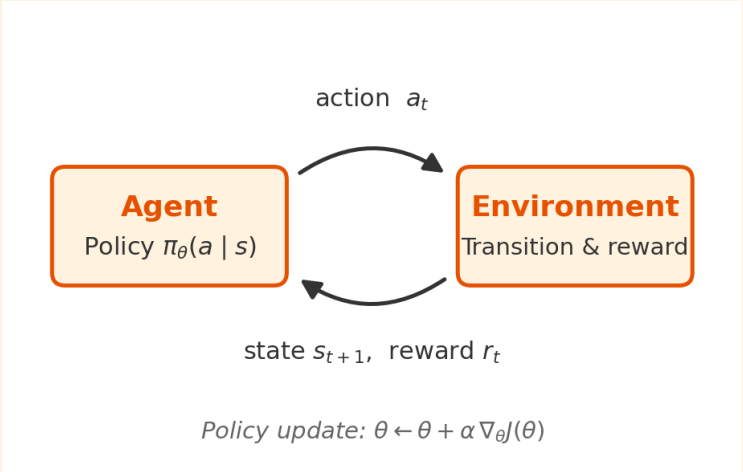
Today: Imitation Learning

1. Behavioral cloning and distribution shift
2. DAgger and interactive data collection
3. Teleoperation systems
4. Scaling and the primacy of diversity
5. Human video learning

These five objectives form a single causal chain, where each solves the problem left unresolved by the previous.

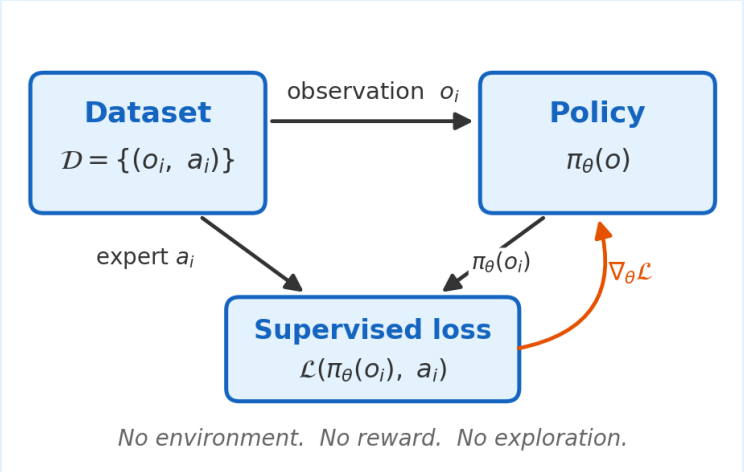
Behavioral Cloning: Supervised Learning on Demonstrations

Reinforcement Learning



Learn by trial and error guided by a scalar reward signal. Requires millions of interactions, careful reward design, and exploration.

Behavioral Cloning



Learn by mimicking an expert. Dataset of (observation, action) pairs. Train with a standard supervised loss. No exploration, no reward, no credit assignment.

The Behavioral Cloning Objective

A supervised learning problem in every respect. The training signal comes from expert demonstrations; the architecture, loss, and optimization are standard.

Symbol Definitions

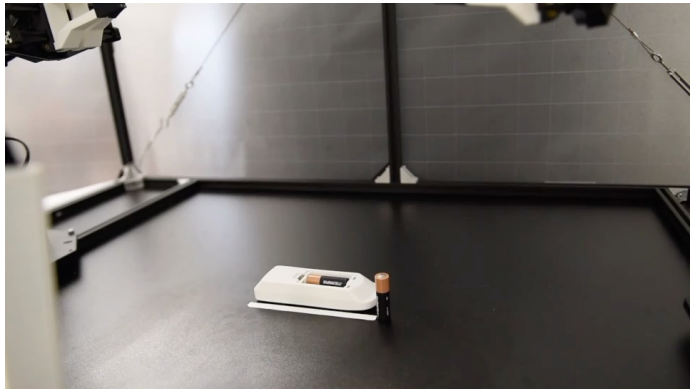
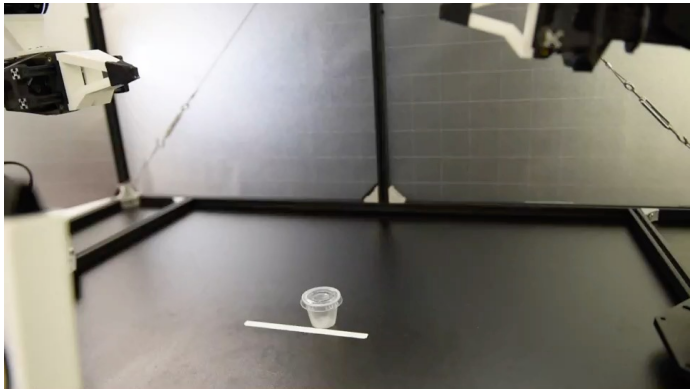
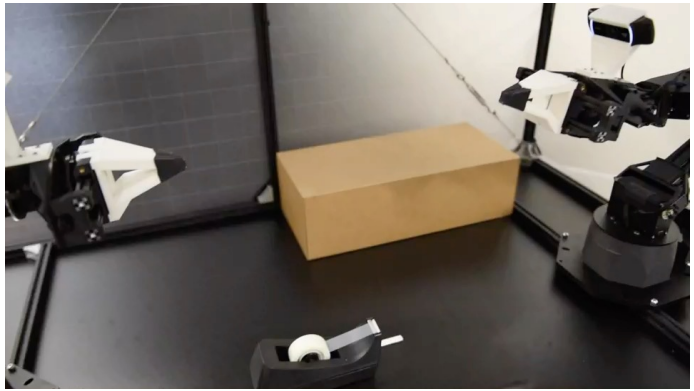
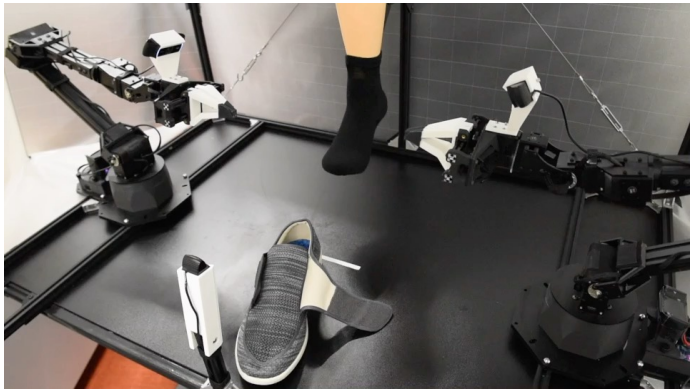
Symbol	Domain	Meaning
π_θ	function	Learned policy parameterized by θ
o_t	observation space	Observation at timestep t
a_t	action space	Expert action at timestep t
$\mathcal{D}$	dataset	Demonstration dataset of observation-action pairs
$\mathcal{L}$	loss	Per-sample loss (MSE or negative log-likelihood)

BC Training Objective (EQ1)

$$\theta^* = \arg\min_\theta \mathbb{E}_{(o_t, a_t) \sim \mathcal{D}} \mathcal{L}\big(\pi_\theta(o_t),\ a_t\big)$$

Standard supervised learning. The objective is identical to training a classifier or a regressor. The only thing that distinguishes BC from any other supervised learning problem is what the labels mean.

Behavioral Cloning in Action: Precision Manipulation

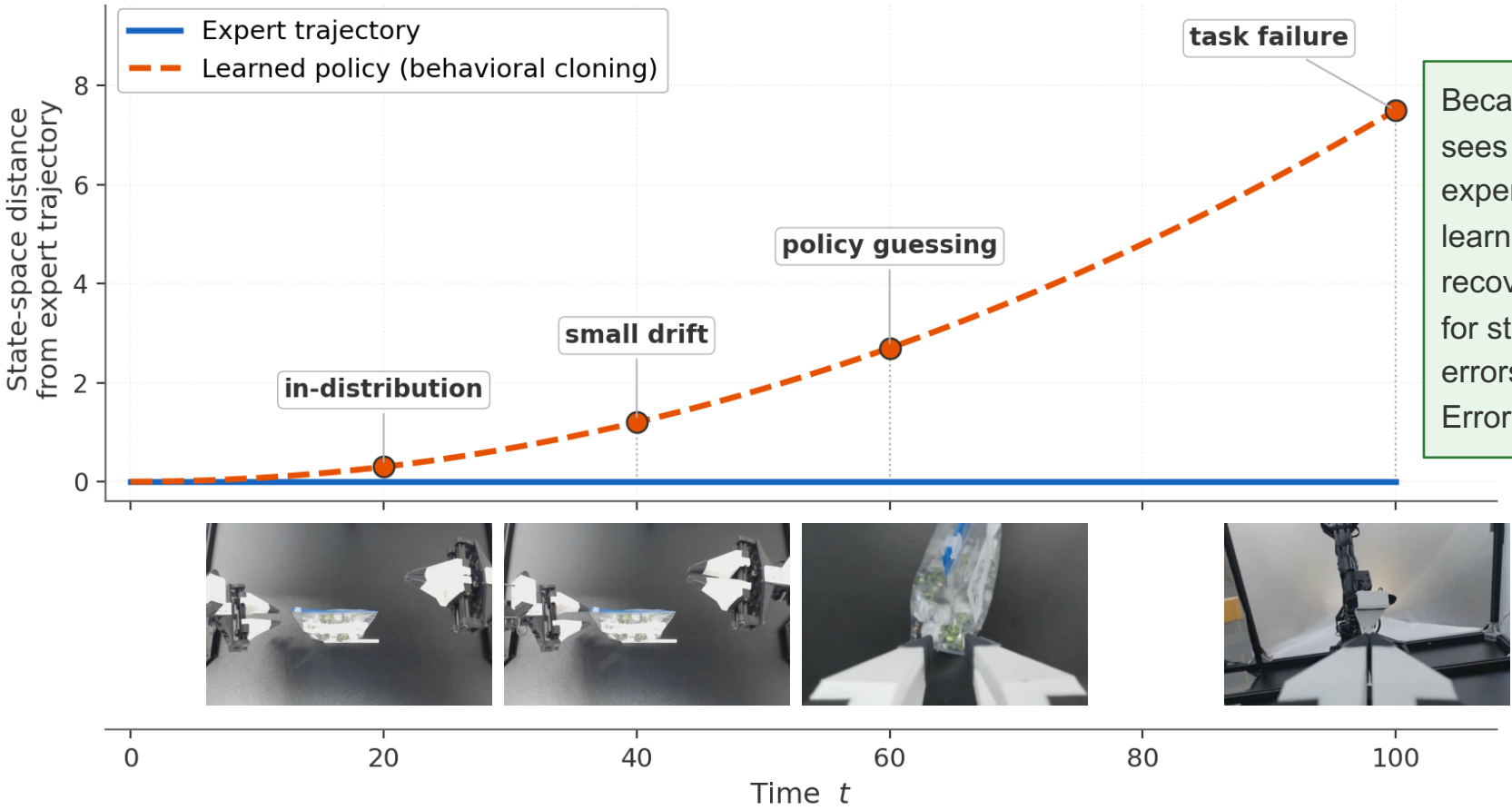
ALOHA : Bimanual manipulation learned end-to-end with action chunking on approximately 50 demonstrations per task.

Behavioral Cloning in Action: Mobile and Whole-Body Manipulation

Mobile ALOHA: mobile bimanual manipulation trained via behavioral cloning with co-training on static data plus approximately 50 task-specific demonstrations.

If BC works this well, why do we need the rest of this lecture?

Why Does BC Fail? The Compounding-Error Picture



Because BC only sees states the expert visited, the learner has no recovery behavior for states its own errors produce. Errors compound.

The Compounding-Error Bound

Symbol Definitions

Symbol	Domain	Meaning
T	integer	Trajectory horizon (number of timesteps)
ϵ	real	Per-timestep policy error rate
$J(\pi)$	real	Total expected cost of rolling out policy π
π^*	policy	Expert (reference) policy

Compounding-Error Bounds (EQ2)

$$J(\pi_{\mathrm{BC}}) - J(\pi^*) \;\leq\; \mathcal{O}\big(\epsilon T^2\big) \qquad \text{(Behavioral Cloning)}$$

A policy making errors 1% of the time can fail almost deterministically over a 100-step trajectory under BC.

Learning Connection

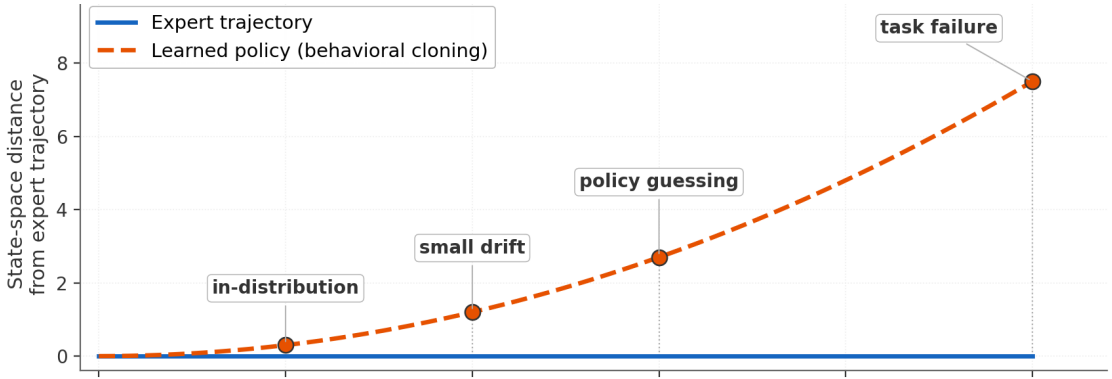
Action chunking, which recent theory (Zhang et al., 2025) has shown converts **exponential** compounding errors to **polynomial** ones in continuous control. Diffusion Policy and ACT are most powerful not because they are bigger networks, but because they combine expressive action distributions with chunking.

Think-Pair-Share

Imagine you are teaching a robot to pour coffee. Partner up and discuss for two minutes:

1. What would it take to train this policy with reinforcement learning?
2. What would it take to train it with behavioral cloning?
3. Which would you choose, and what is the trade-off?

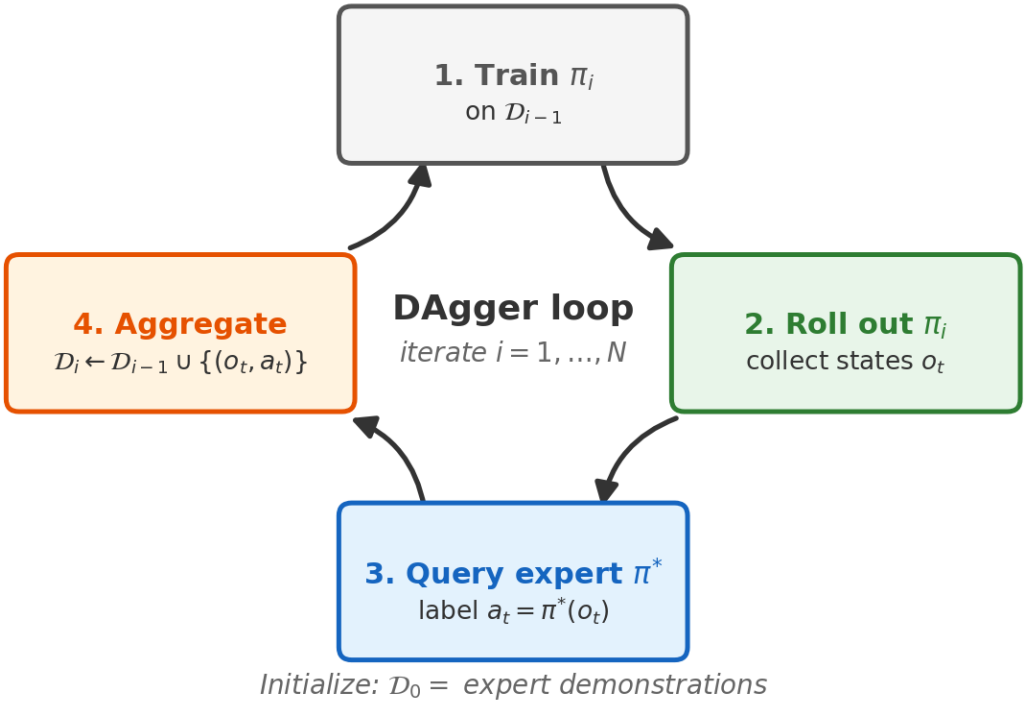
Can We Fix Covariate Shift?



The problem: the policy visits states it never saw in training, and errors compound.

If the problem is that training and deployment distributions differ, what would you change?

DAgger: Formal Definition and Algorithmic Structure



$$J(\pi_{\mathrm{BC}}) - J(\pi^*) \leq \mathcal{O}(\epsilon T^2) \qquad \text{(Behavioral Cloning)}$$

$$J(\pi_{\mathrm{DAgger}}) - J(\pi^*) \leq \mathcal{O}(\epsilon T) \qquad \text{(DAgger)}$$

DAgger: Formal Definition and Algorithmic Structure

Symbol Definitions

Symbol	Domain	Meaning
π^*	function	Expert policy mapping observations to actions
π_i	policy	Learner policy at iteration i, parameterized by θ_i
$\mathcal{D}_i$	dataset	Aggregated observation-action dataset at iteration i
d_{π_i}	distribution	Distribution of states visited when following π_i
β_i	real in $[0, 1]$	Mixing coefficient (expert-to-learner blend)
$\mathcal{L}$	loss	Per-sample loss (MSE or negative log-likelihood)

DAgger: Dataset Aggregation (Ross, Gordon, Bagnell, 2011)

Initialize: $\mathcal{D}_0 \leftarrow \{(o_t, \pi^*(o_t))\}$ from expert demonstrations.

For $i = 1, 2, \ldots, N$ **do:**

1. Train on aggregated data *(EQ3)*

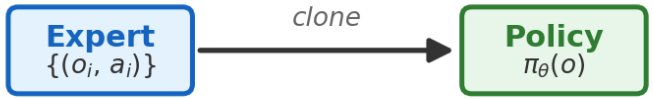
$$\pi_i \;=\; \arg\min_{\pi} \; \mathbb{E}_{(o,\,a)\,\sim\,\mathcal{D}_{i-1}} \; \mathcal{L}\big(\pi(o),\; a\big)$$

2. Roll out mixture policy: sample states $o_t \sim d_{\beta_i \pi^* + (1-\beta_i)\pi_i}$
3. Query expert on visited states: $a_t = \pi^*(o_t)$
4. Aggregate: $\mathcal{D}_i \leftarrow \mathcal{D}_{i-1} \cup \{(o_t, a_t)\}$

end for

Reward Inference: An Alternative Framing

Behavioral Cloning / DAgger: clone behavior

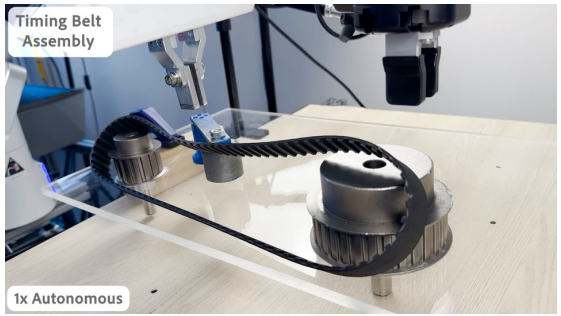


IRL family: infer reward, then learn policy

Method	Mechanism	Status in 2026
MaxEnt IRL (Ziebart et al., 2008)	Infer the reward that makes the expert's behavior maximally likely under a maximum-entropy distribution	Foundational; largely superseded by deep IRL variants
GAIL (Ho and Ermon, 2016)	Adversarial training: learner tries to match expert state-action distribution; discriminator distinguishes them	Foundational; inspired DPO-style preference learning for robotics (e.g., GRAPE, 2024)

The reward-inference family asks: rather than cloning the behavior, can we infer the reward that explains it? Survey papers cover this family in depth; for today we note it exists as an alternative framing.

HIL-SERL: Human-in-the-Loop Reinforcement Learning

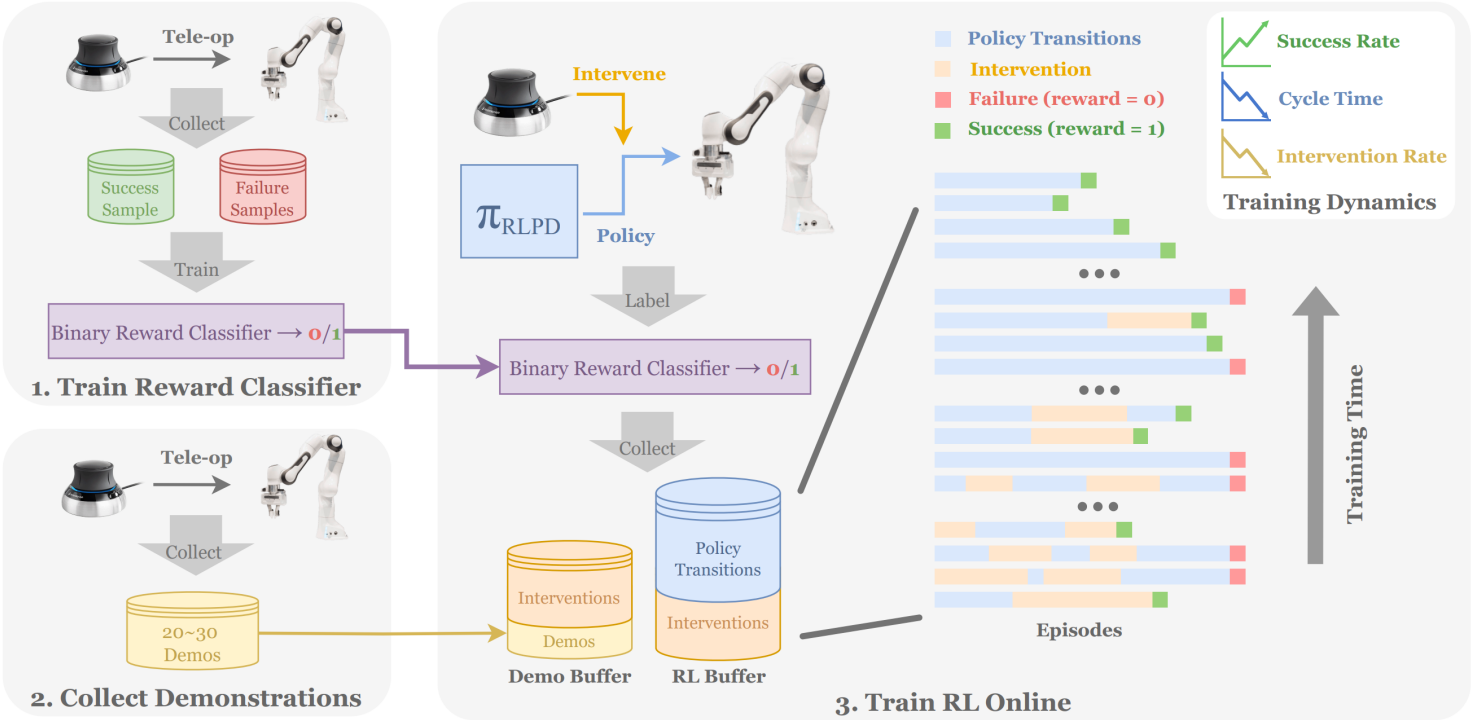


Off-policy human-in-the-loop reinforcement learning. The human operator intervenes when the policy fails, and those interventions are used as a corrective signal. HIL-SERL reaches near-perfect success rates on precise real-world tasks in one to two and a half hours of real-world training.

Key Insight

HIL-SERL validates the DAgger insight in the foundation-model era: closing the expert-learner loop remains the most sample-efficient way to drive success rates to near-perfect on precise tasks, even when the underlying policy is a modern neural network.

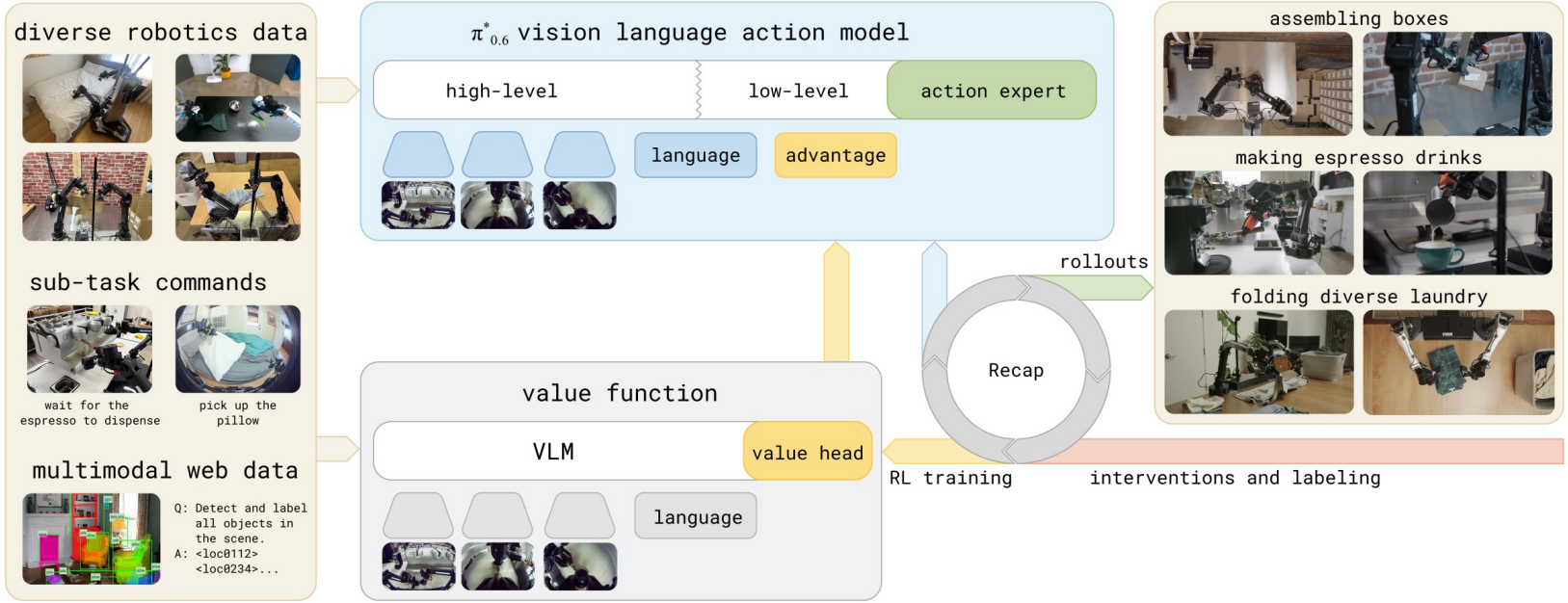
HIL-SERL: Human-in-the-Loop Reinforcement Learning



$$\mathcal{L}_Q(\phi) = E_{\mathbf{s},\mathbf{a},\mathbf{s}'}\left[\left(Q_\phi(\mathbf{s},\mathbf{a}) - \left(r(\mathbf{s},\mathbf{a}) + \gamma E_{\mathbf{a}'\sim\pi_\theta}[Q_{\bar{\phi}}(\mathbf{s}',\mathbf{a}')]\right)\right)^2\right] \quad (1)$$

$$\mathcal{L}_\pi(\theta) = -E_{\mathbf{s}}\left[E_{\mathbf{a}\sim\pi_\theta(\mathbf{a})}[Q_\phi(\mathbf{s},\mathbf{a})] + \alpha\mathcal{H}(\pi_\theta(\cdot|\mathbf{s}))\right], \quad (2)$$

π*0.6 and RECAP: The November 2025 Frontier



Learning Connection

Interactive data collection, which seemed historical when DAgger was proposed in 2011, has become the research frontier again for foundation-model-era robot learning. The trajectory from DAgger in 2011 to HIL-SERL in 2025 to RECAP in late 2025 is a single idea scaling with the models it is applied to.

π*0.6 and RECAP: The November 2025 Frontier

Today's Lecture: Five Objectives in One Causal Chain

Where RL Left Us

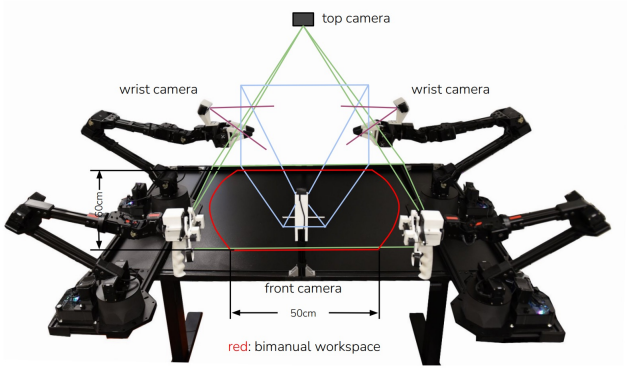
Reinforcement learning requires millions of environment interactions, reward specification that is fragile and prone to hacking, exploration cost that makes physical deployment unsafe, and long-horizon credit assignment that remains brittle.

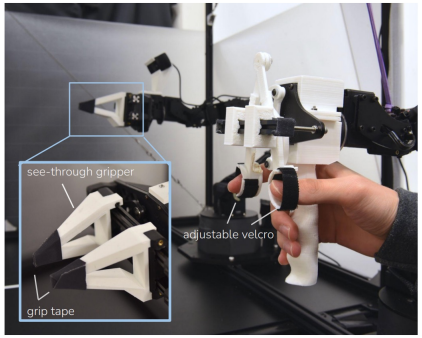
Today: Imitation Learning

1. Behavioral cloning and distribution shift
2. DAgger and interactive data collection
3. Teleoperation systems
4. Scaling and the primacy of diversity
5. Human video learning

These five objectives form a single causal chain, where each solves the problem left unresolved by the previous.

ALOHA: Teleoperation as a First-Class Research Problem





Features

Leader arms: Low-cost kinematic replicas driven by human hands

Joints: Passive gravity compensation, no high-end servos

Followers: Paired followers execute identical joint trajectories in real time

Cost: Approximately $20,000 total system

ALOHA enables bimanual manipulation tasks requiring millimeter precision that were previously uncollectable at academic budgets. The hardware design is released open-source with complete bills of materials.

Key Insight

Teleoperation hardware design is a first-class research contribution because it determines what data is collectable. No algorithmic advance can compensate for data that cannot be gathered. ALOHA's impact on the field traces directly to design choices that made bimanual demonstrations practical at academic scale.

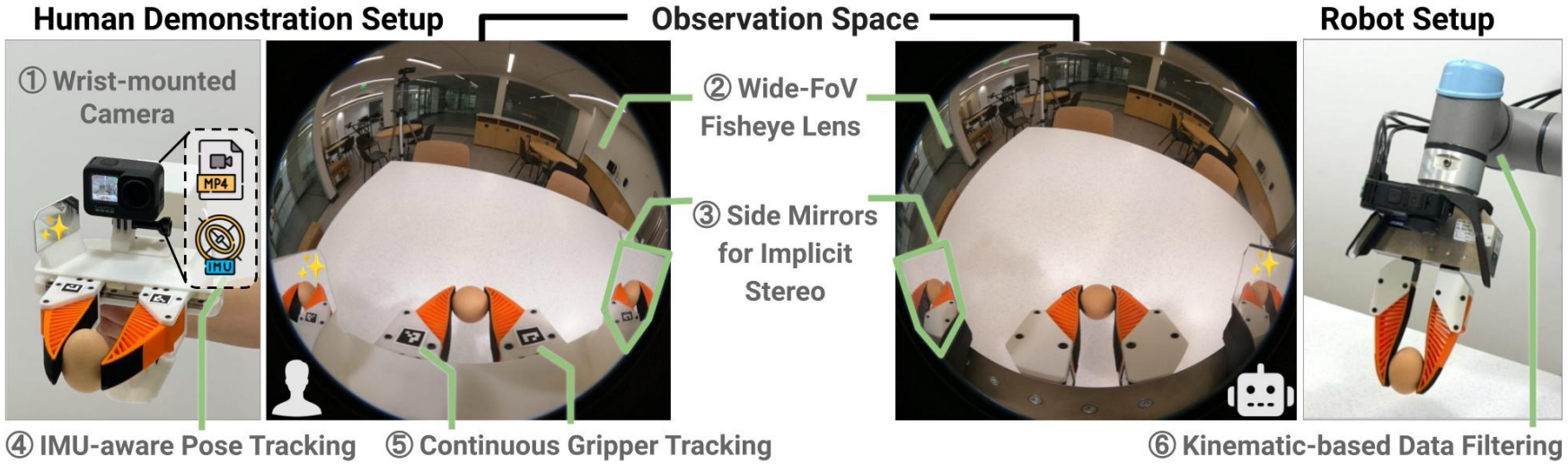
Mobile ALOHA: Extending the Paradigm

Wheeled base + raised arm mount: appropriate height for whole-body human demonstrations.

Collection method: operator rides the base while teleoperating the arms through the leader-follower interface.

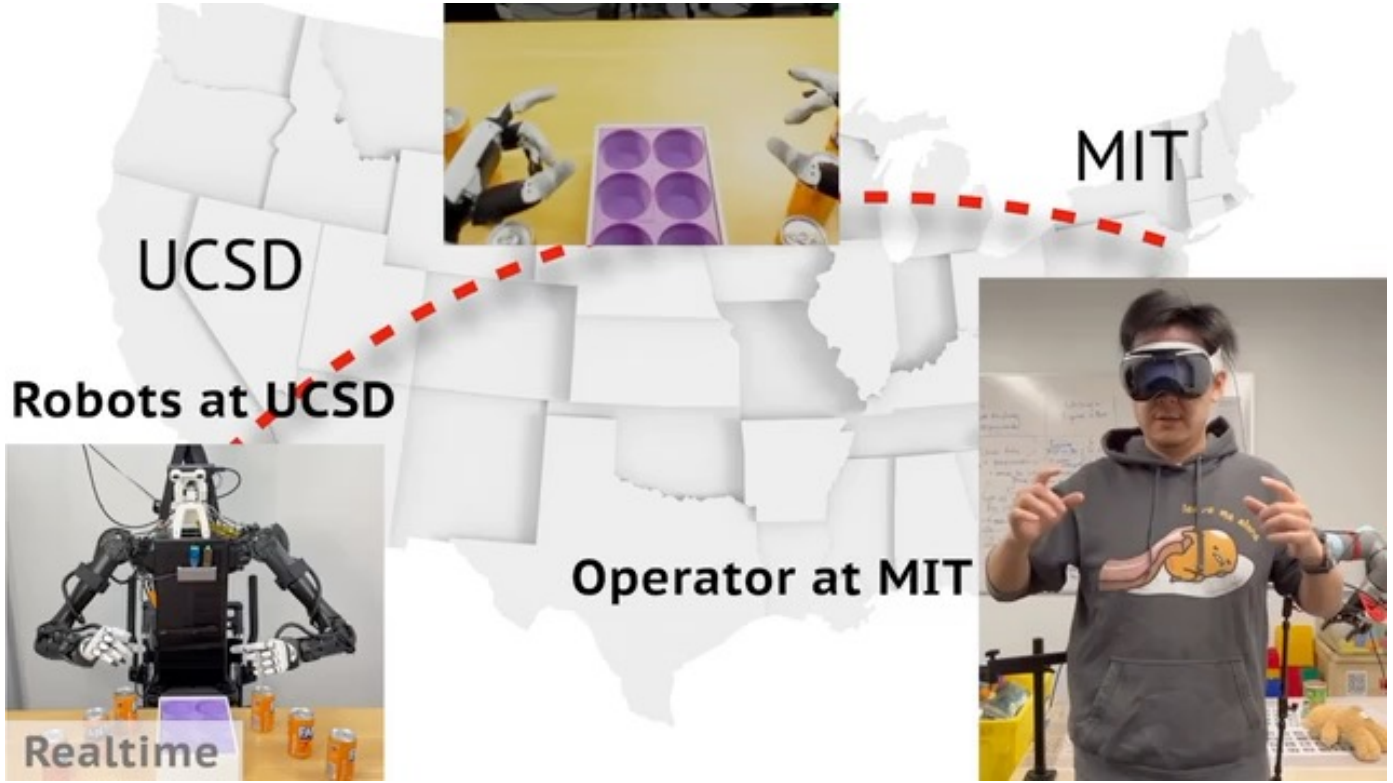
Co-training with static ALOHA: reduces per-task demonstration requirement to approximately 50 for complex tasks such as cooking shrimp or riding an elevator.

UMI: Universal Manipulation Interface



https://umi-gripper.github.io/

Open-TeleVision: Teleoperation with Immersive Active Visual Feedback



The Data Collection Design Space

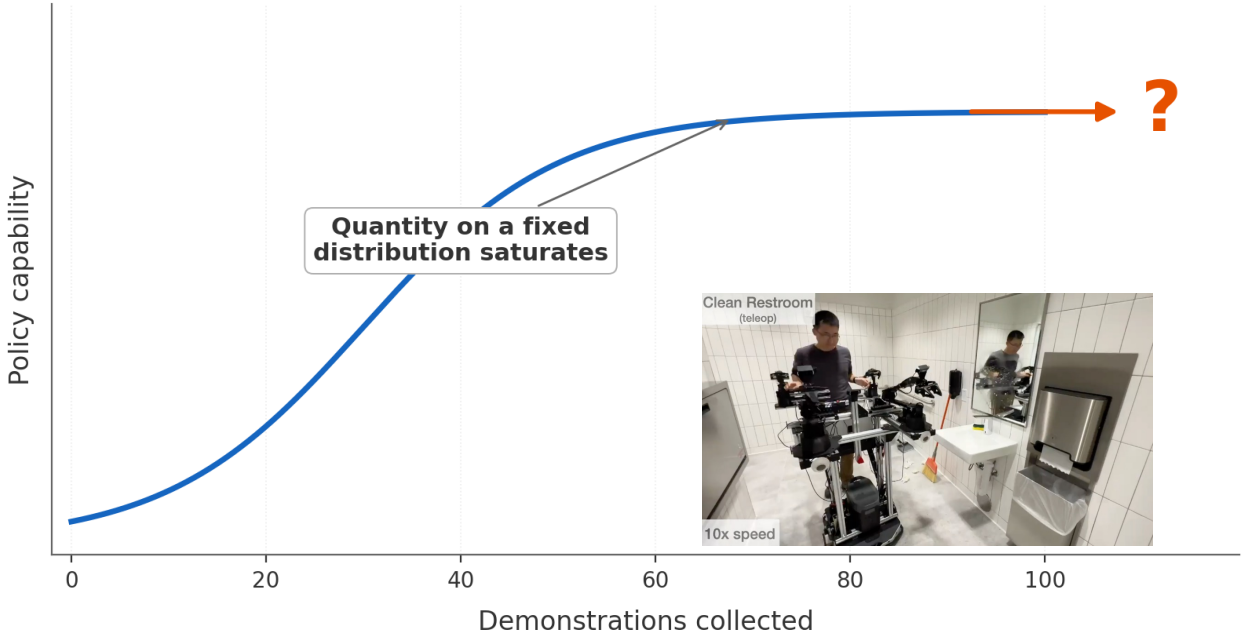
Design Dimension	ALOHA (leader-follower)	UMI (handheld gripper)	Open-TeleVision (immersive VR)
Teleoperation Mechanism	Physical leader arms kinematically paired with followers	Handheld gripper with GoPro + fisheye lens; no robot required during collection	VR headset streams stereoscopic robot view; operator commands through tracked controllers
Dexterity Profile	Bimanual, moderate dexterity, no fingers	Single-handed, parallel-jaw contact	Full arm dexterity, humanoid whole-body
Cost and Accessibility	~$20K; requires robot at every collection session	~$1K; usable anywhere	Several $K for VR + robot
Task Coverage	Bimanual tabletop, contact-rich	Any task a handheld gripper can perform, including in-the-wild	Humanoid and dexterous manipulation, large workspaces

Think-Pair-Share

You are collecting data for a policy that needs to handle diverse kitchen scenes across 50 different homes.

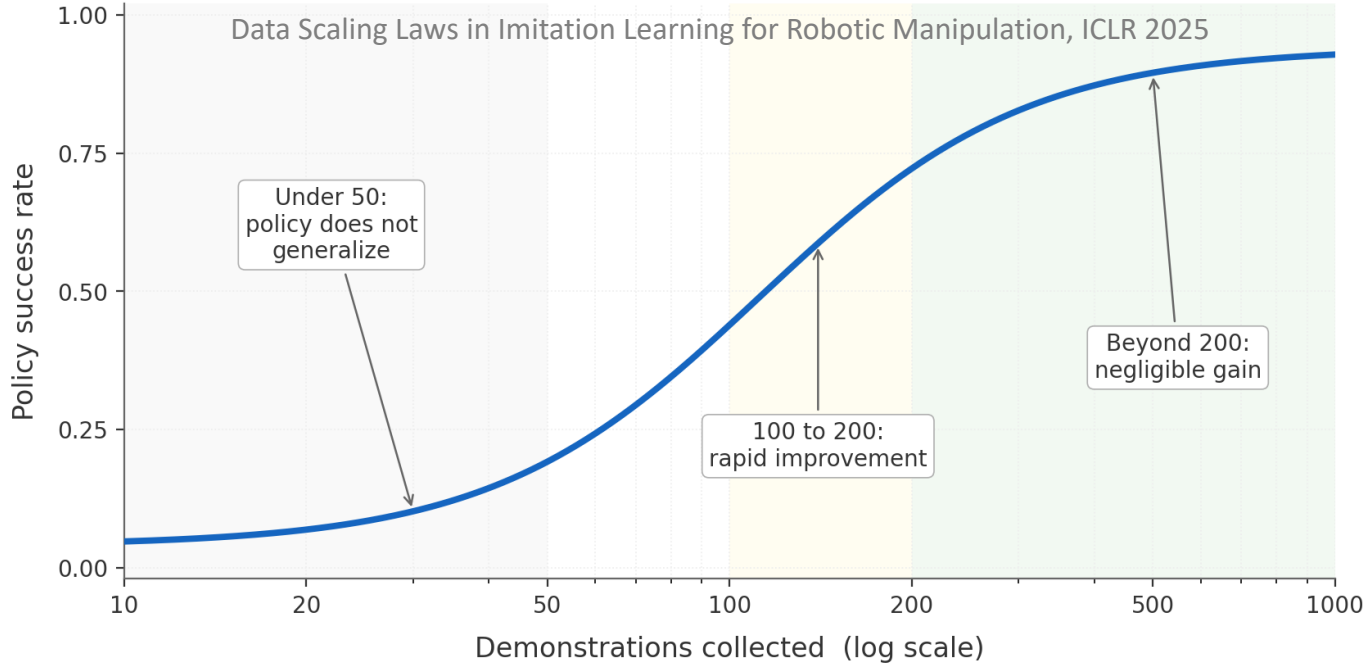
Which data collection system would you choose, and what are the trade-offs?

Teleoperation Hits a Ceiling



Even excellent teleoperation saturates. Collecting more demonstrations of the same task produces more data but not more generalization.

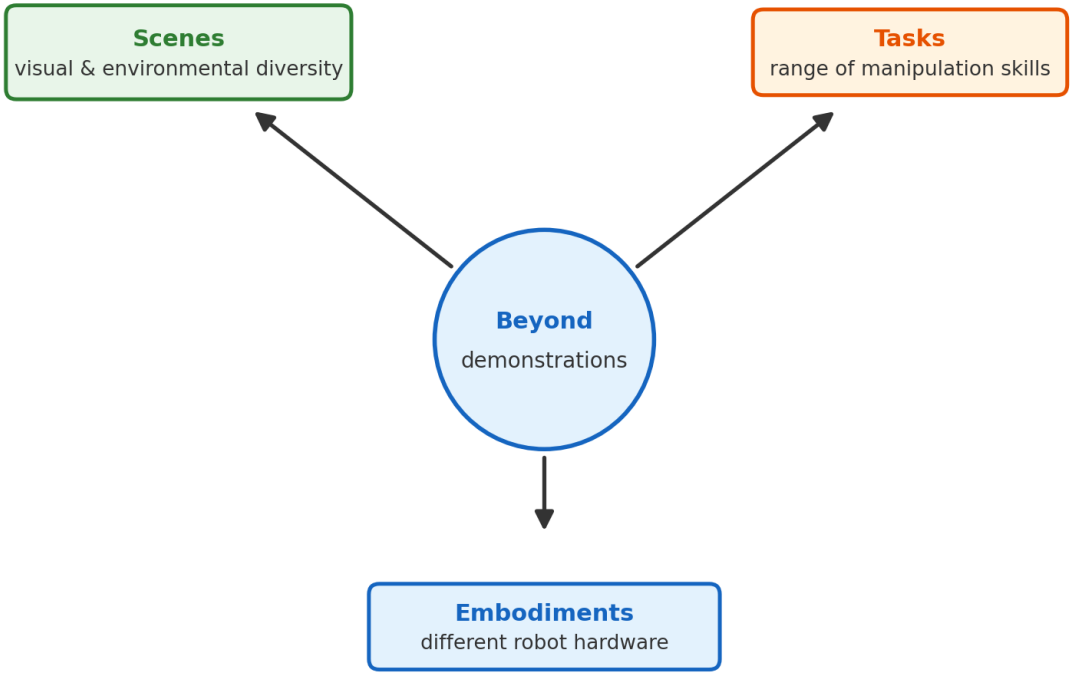
The Scaling Curve: Quantity Saturates



Key Insight

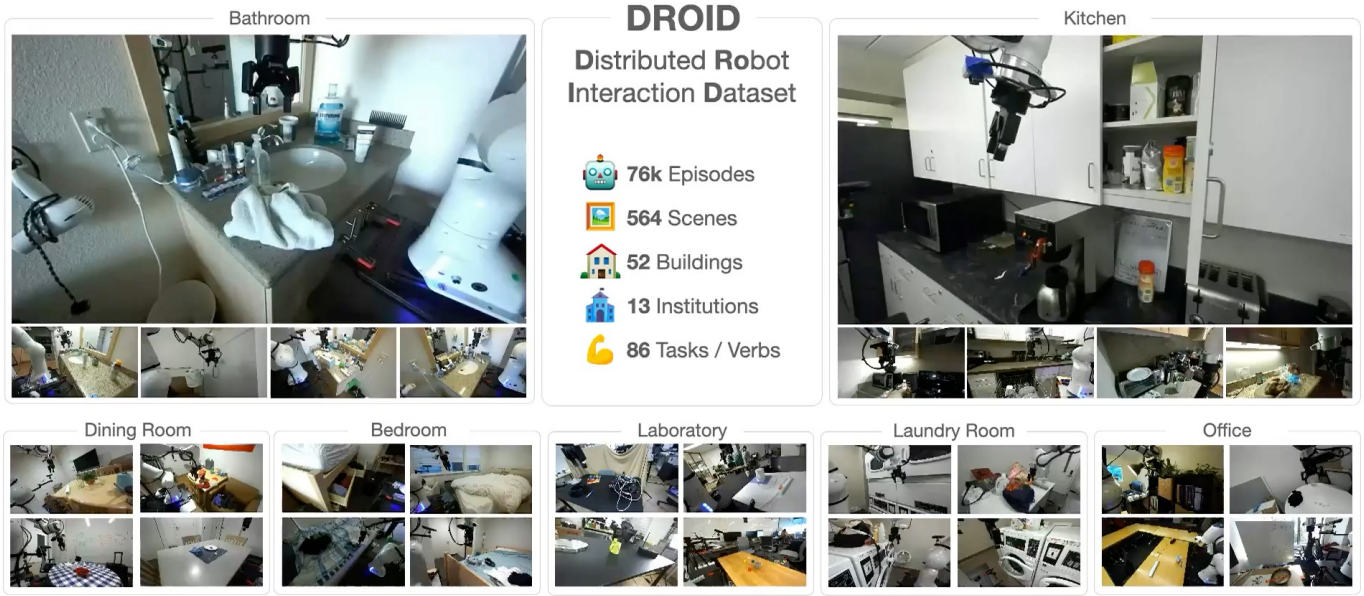
Imitation learning does not scale the way language models do. On a fixed task, policy performance plateaus after roughly 50 to 200 demonstrations. If demonstrations are what you have, scaling them by another order of magnitude will not meaningfully improve performance.

What Else Could We Scale?



Three candidate answers.

DROID: Betting on Scene and Object Diversity

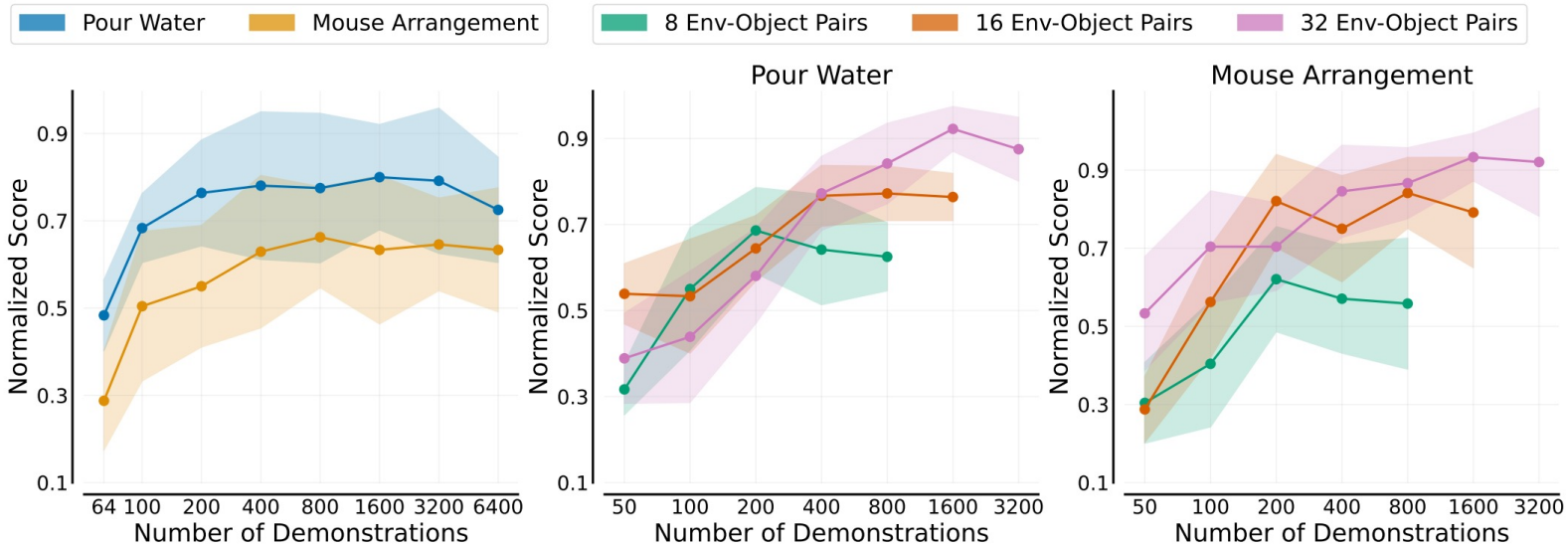


DROID's Design Bet

Diversity lives in scenes, objects, and tasks. Fix the embodiment to reduce variability; vary everything else.

DROID aggregates 76,000 trajectories collected across 564 scenes on a single Franka Panda embodiment with a Robotiq gripper. 50 operators contributed demonstrations spanning 86 tasks across 3 continents. Total data volume is approximately 350 hours.

DROID: Betting on Scene and Object Diversity

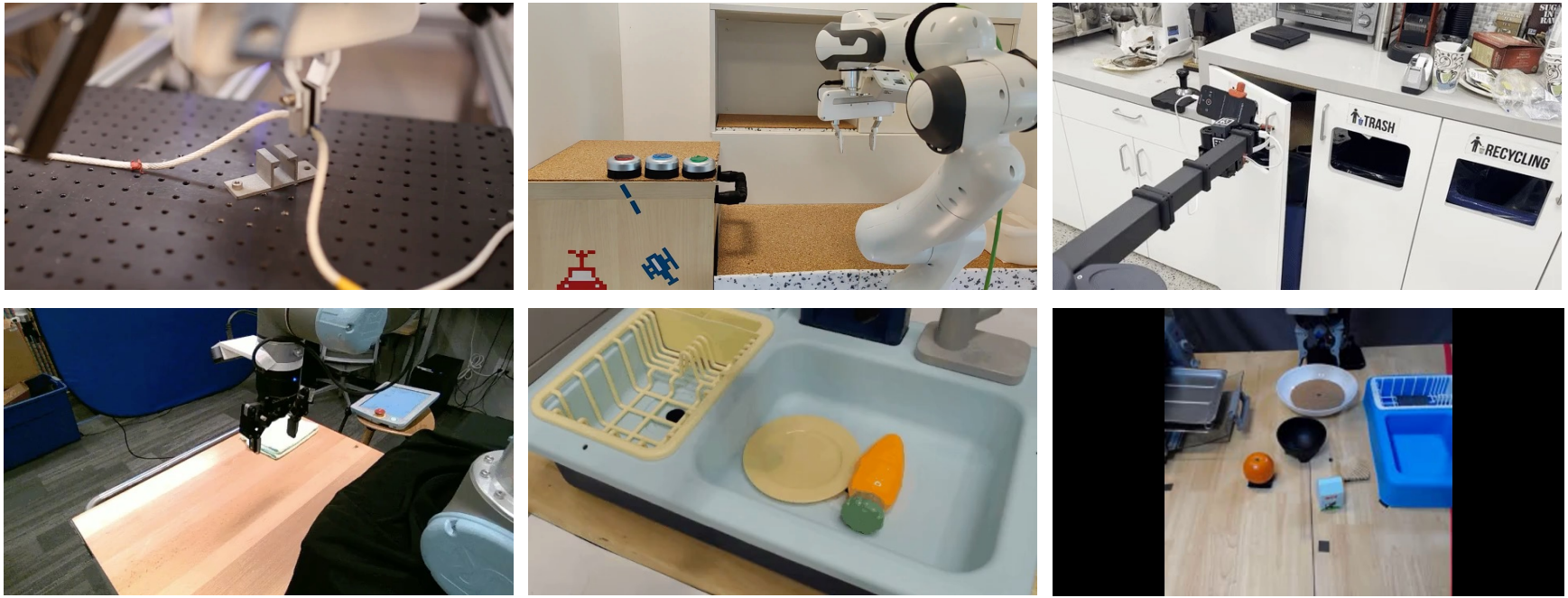


DROID's Design Bet

Diversity lives in scenes, objects, and tasks. Fix the embodiment to reduce variability; vary everything else.

DROID aggregates 76,000 trajectories collected across 564 scenes on a single Franka Panda embodiment with a Robotiq gripper. 50 operators contributed demonstrations spanning 86 tasks across 3 continents. Total data volume is approximately 350 hours.

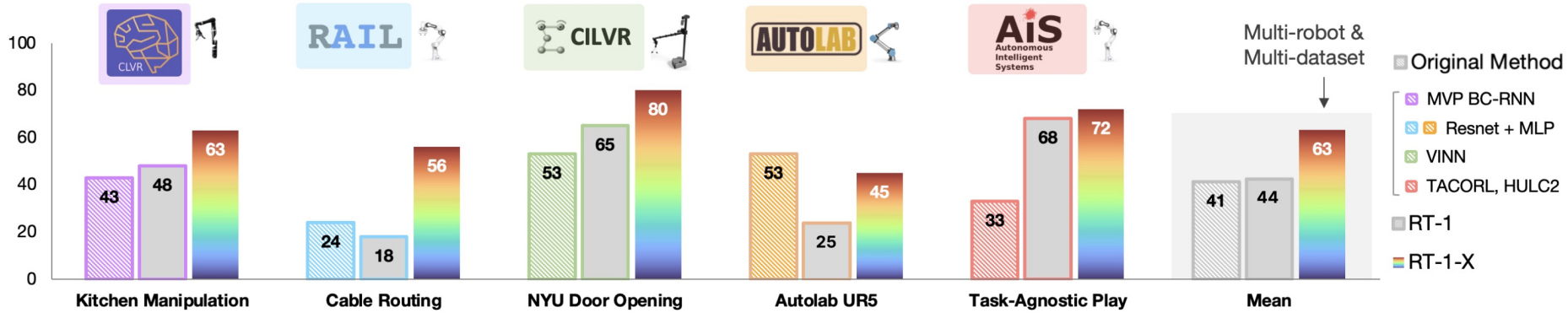
Open X-Embodiment and RT-X: Betting on Embodiment Diversity



Dataset and Design Bet

Over 1 million trajectories pooled from 60 existing datasets, spanning 22 robot embodiments, contributed by 34 labs at 21 institutions. The bet: diversity lives across embodiments — accept heterogeneous action spaces as the price of scale.

Open X-Embodiment and RT-X: Betting on Embodiment Diversity



RT-1 and RT-1-X share the same network structure.

Today's Lecture: Five Objectives in One Causal Chain

Where RL Left Us

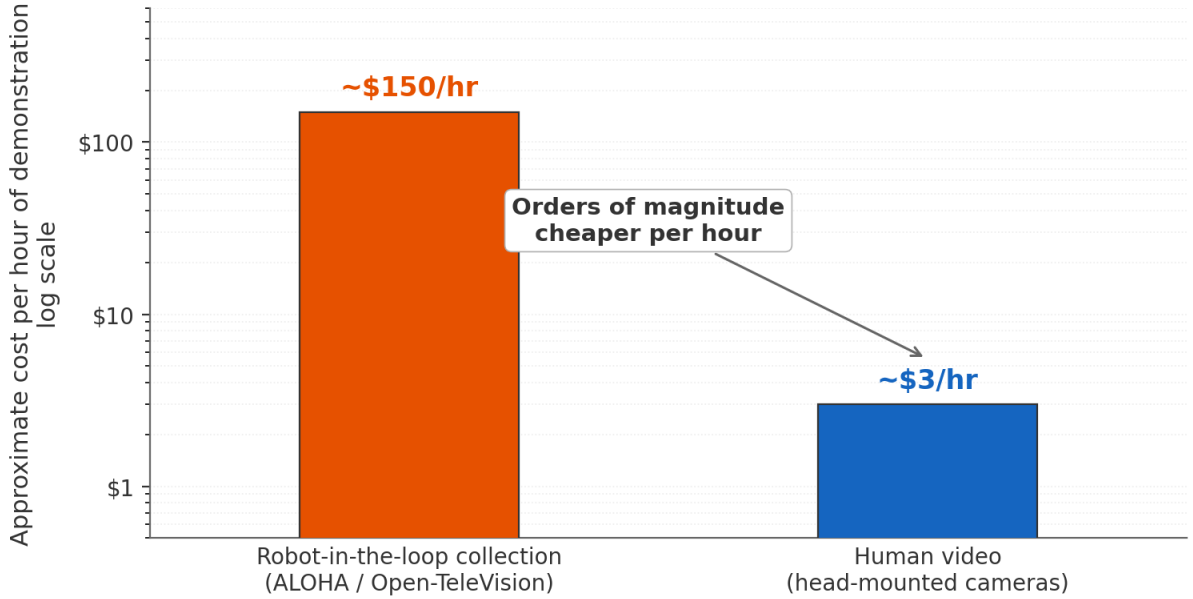
Reinforcement learning requires millions of environment interactions, reward specification that is fragile and prone to hacking, exploration cost that makes physical deployment unsafe, and long-horizon credit assignment that remains brittle.

Today: Imitation Learning

1. Behavioral cloning and distribution shift
2. DAgger and interactive data collection
3. Teleoperation systems
4. Scaling and the primacy of diversity
5. Human video learning

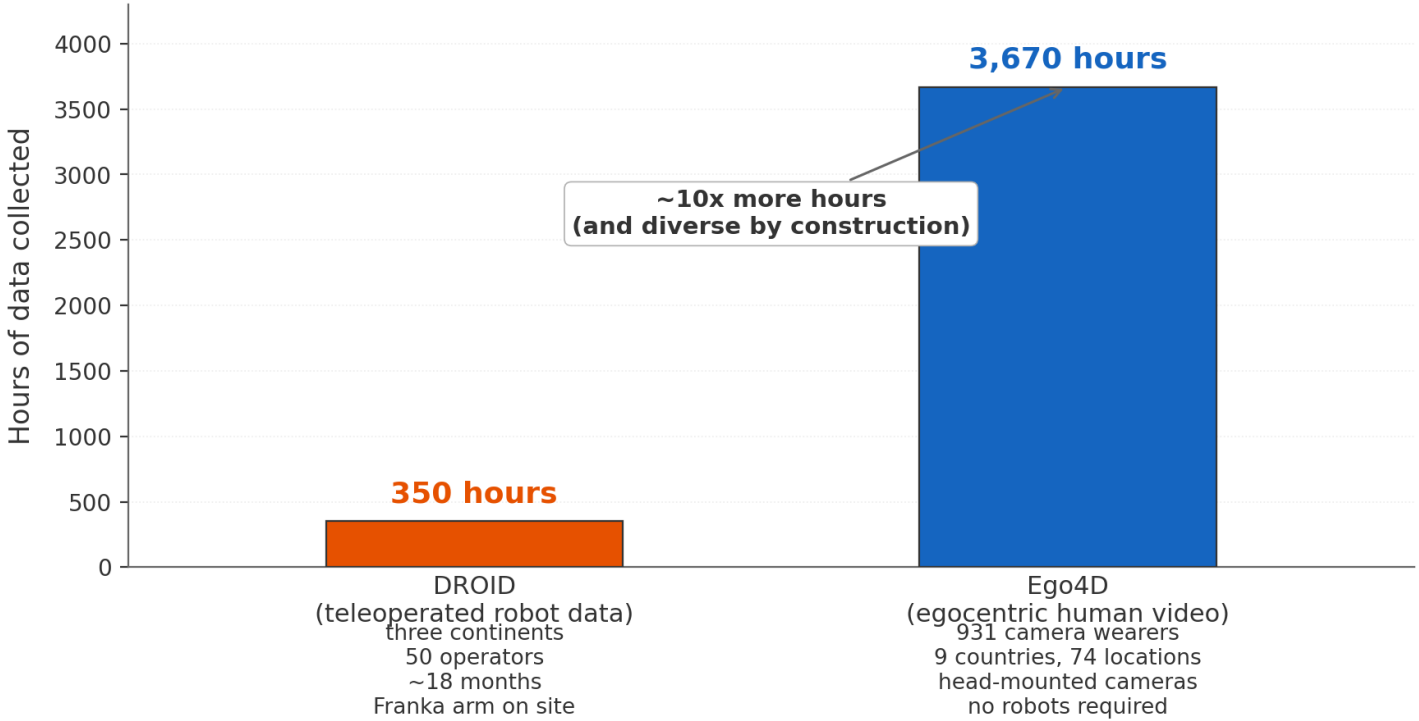
These five objectives form a single causal chain, where each solves the problem left unresolved by the previous.

Teleoperation Cannot Deliver the Diversity We Need



If diversity is what scales, and teleoperation is this expensive per hour, we need a qualitatively different data source.

The Diversity Asymmetry: Human Video versus Teleoperation



If diversity is the axis that scales, human video sits on the axis.

The Embodiment Gap

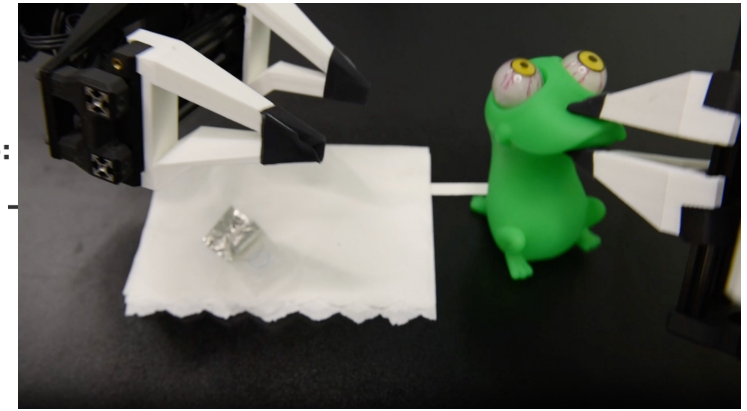
Human hand

- ~21 degrees of freedom
- diverse contact geometry
- distributed force, multiple digits

Parallel-jaw gripper

- 1 degree of freedom
- simple opposing-surface contact
- force concentrated between two plates

The embodiment gap:
not a coordinate change



Kinematic

reachable workspaces differ;
no simple transform equates them

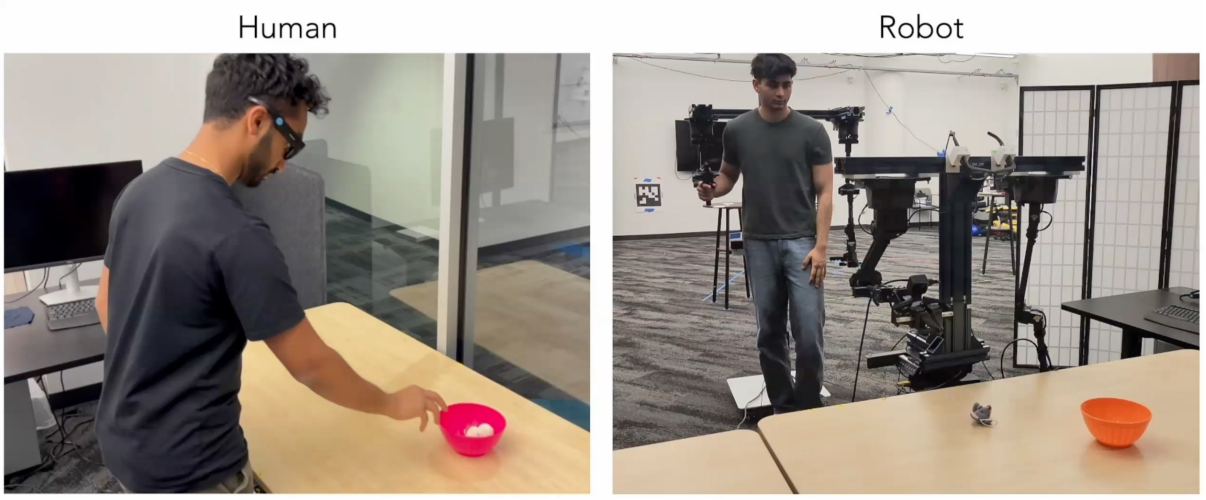
Geometric

human fingers wrap and pinch;
parallel-jaw grippers only oppose

Dynamic

human force distributes across digits;
grippers apply a single opposing force

Strategy One: Explicit Alignment with Co-Training

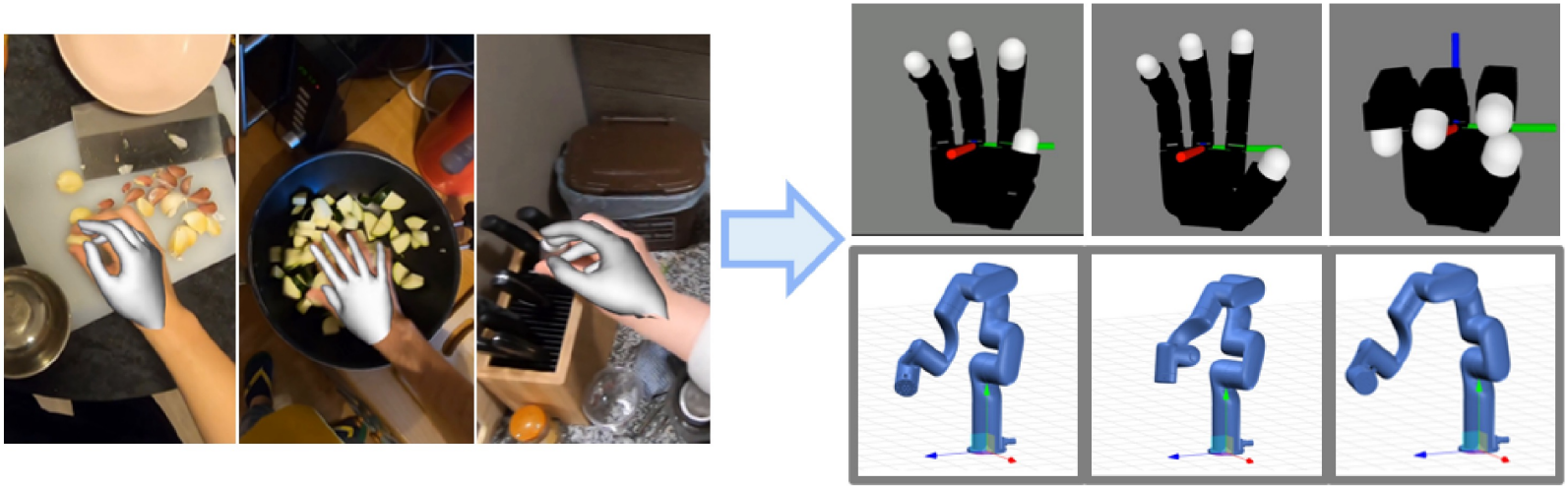


EgoMimic collects egocentric human video using Project Aria glasses, capturing both first-person visual observations and 3D hand poses. The human data is co-trained with teleoperated robot data, with hand poses explicitly aligned to the robot's end-effector frame. The approach yields 34 to 228 percent improvement over robot-only baselines.

Design Bet — Engineer the Bridge

Map human motion to robot action through alignment, then co-train to let the policy learn what alignment cannot capture. Accept that the human-to-robot mapping is engineered rather than learned.

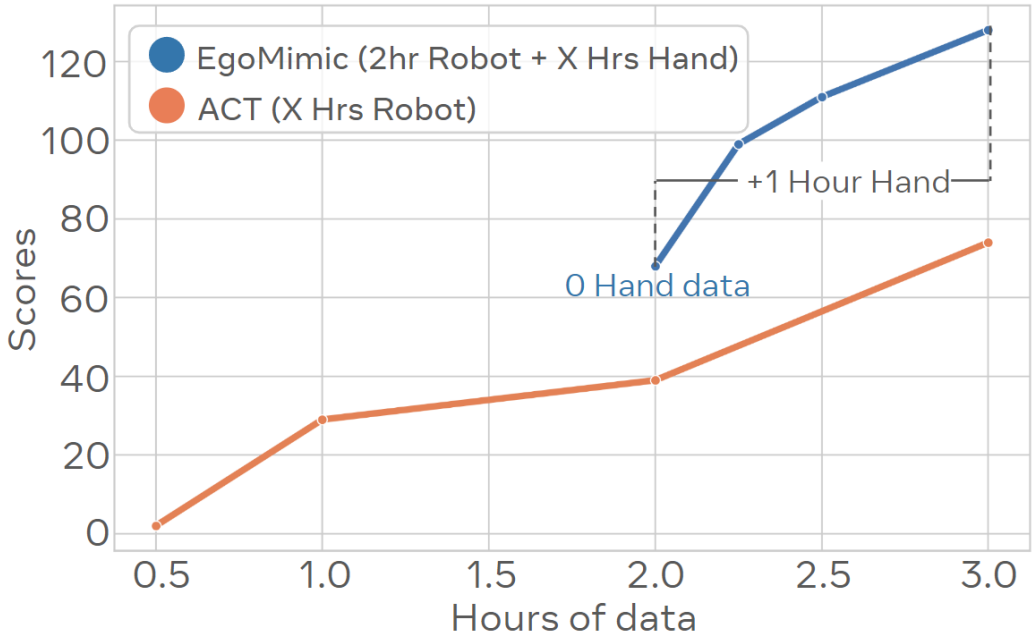
Strategy One: Explicit Alignment with Co-Training



Design Bet — Engineer the Bridge

Map human motion to robot action through alignment, then co-train to let the policy learn what alignment cannot capture. Accept that the human-to-robot mapping is engineered rather than learned.

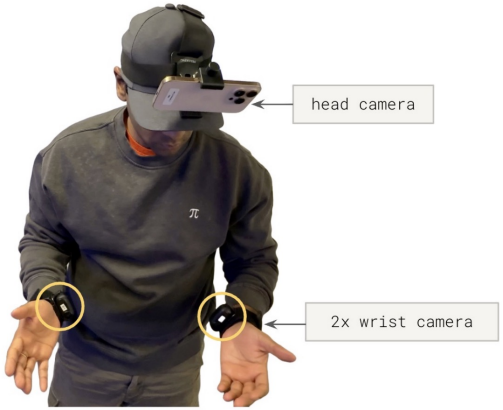
Strategy One: Explicit Alignment with Co-Training

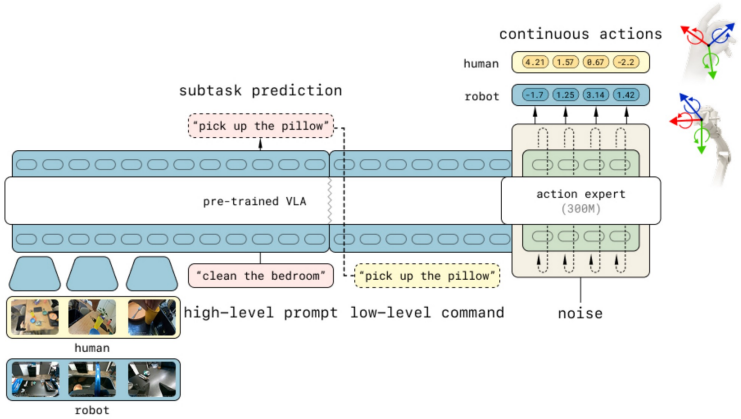


Design Bet — Engineer the Bridge

Map human motion to robot action through alignment, then co-train to let the policy learn what alignment cannot capture. Accept that the human-to-robot mapping is engineered rather than learned.

Strategy Two: Emergent Alignment Through Scale





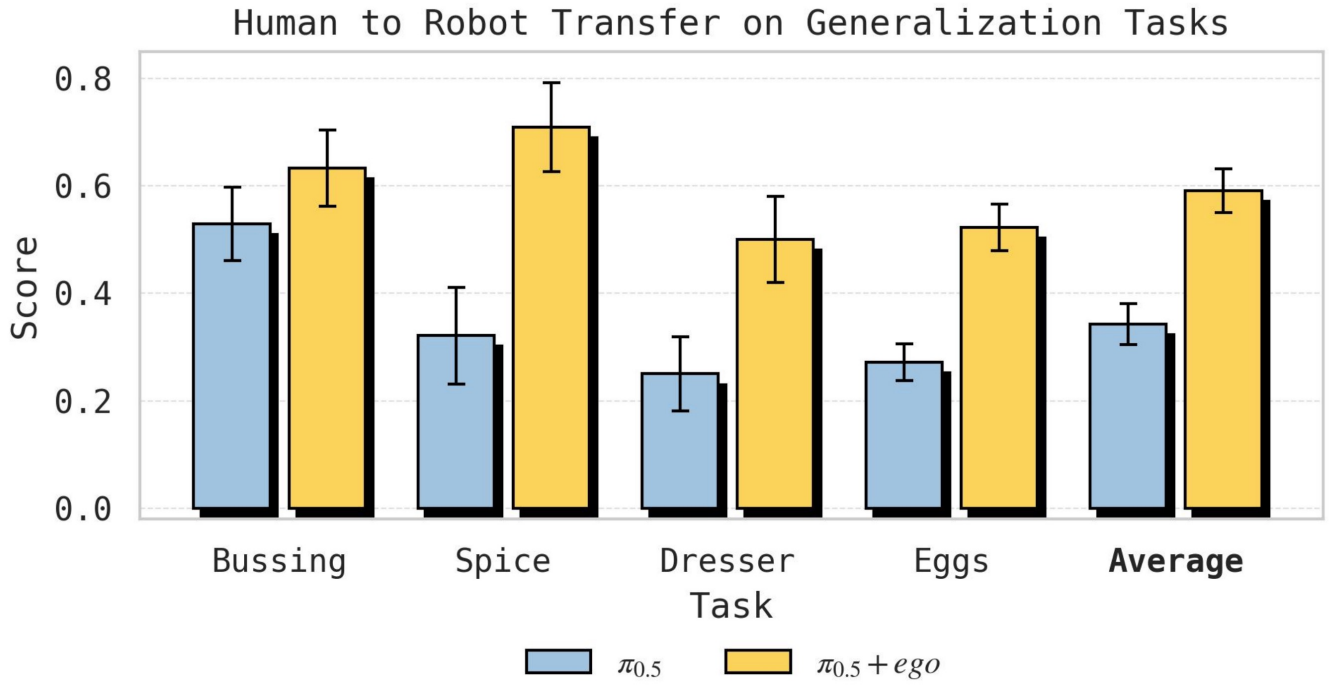
Physical Intelligence co-trained a vision-language-action model on heterogeneous data including teleoperated robot demonstrations and egocentric human video, without explicit alignment.

Design Bet — Scale Past the Gap

At sufficient model and data scale, human and robot representations converge in latent space. The gap dissolves rather than requires engineering.

Emergence of Human to Robot Transfer in Vision-Language-Action Models (https://arxiv.org/pdf/2512.22414)

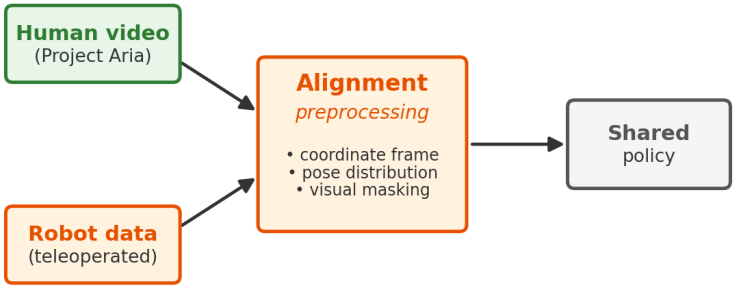
Strategy Two: Emergent Alignment Through Scale



Emergence of Human to Robot Transfer in Vision-Language-Action Models (https://arxiv.org/pdf/2512.22414)

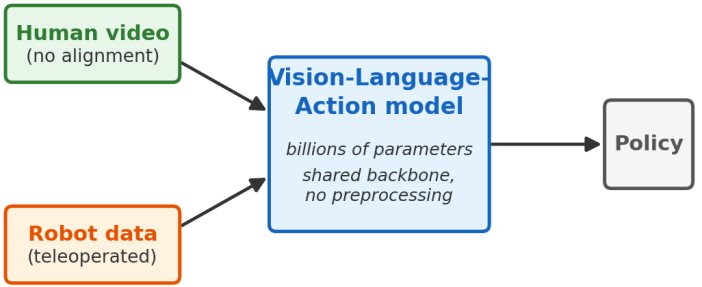
Two Strategies, One Open Research Question

Strategy 1: Align the data before co-training



Bet: engineer the alignment

Strategy 2: Emergent alignment at scale



Bet: scale past the gap

What counts as a demonstration when no robot action labels are present? Will the field close the embodiment gap by engineering the alignment, or by scaling past it?

Key Insight

Whether the embodiment gap closes by engineering or by scale is an open research question as of this lecture. The answer will shape imitation learning research over the next several years.

The Ten Messages of Today's Lecture

Objective 1: BC and Distribution Shift

1. BC converts RL's hard problems into supervised learning, which is why you should care.
2. The cost is covariate shift, which is structural rather than incidental.

Objective 2: DAgger and Interactive Data Collection

1. DAgger fixes covariate shift by closing the expert-learner loop.
2. Interactive data collection has returned as a research frontier with HIL-SERL and π*0.6.

Objective 3: Demonstration Collection Systems

1. ALOHA establishes teleoperation hardware as a first-class research problem.
2. ALOHA is one point in a design space; different points enable different data.

Objective 4: Scaling and Diversity

1. Imitation learning saturates when scaled by quantity on a fixed task.
2. Diversity across scenes, tasks, and embodiments is what actually scales.

Objective 5: Human Video Learning

1. Human video solves diversity at low cost but introduces the embodiment gap.
2. Two strategies address the gap; the choice is an open research question.