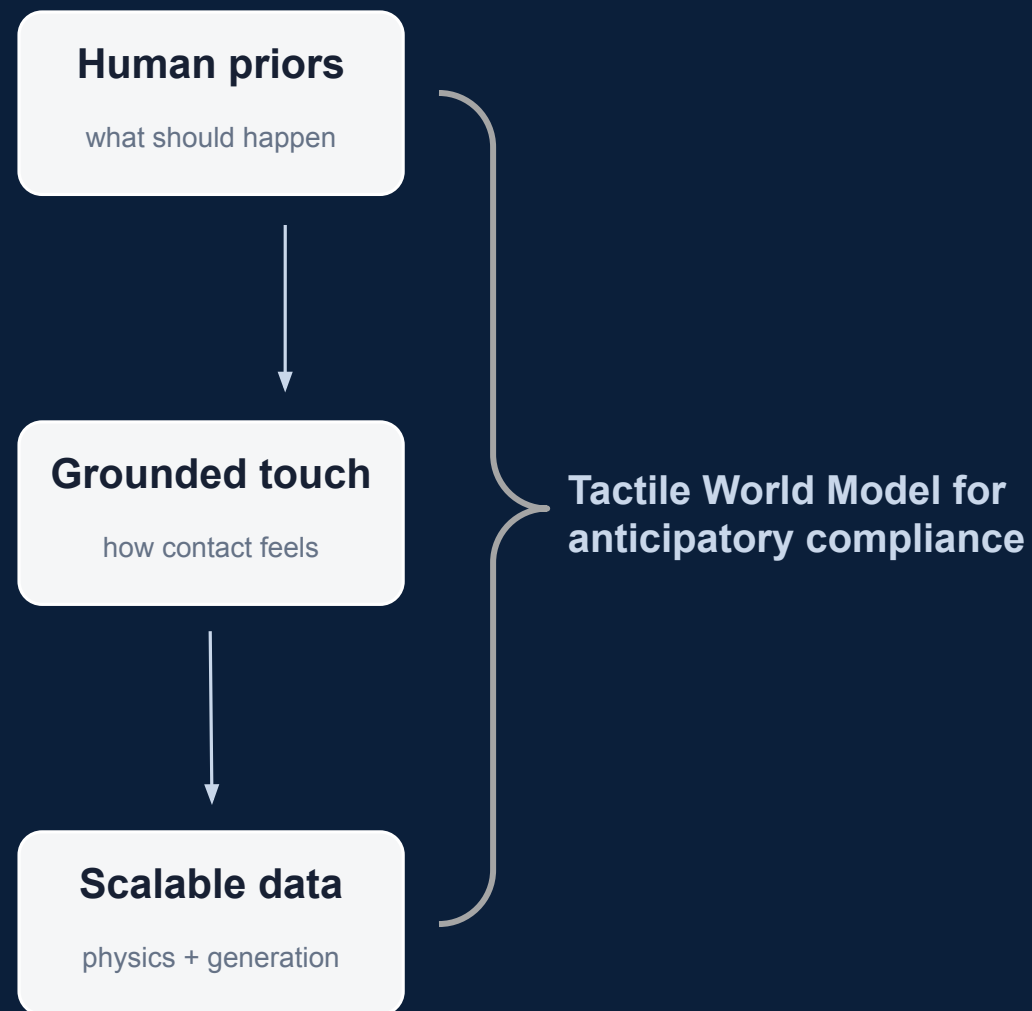


Contact-Rich Robot Learning from Human Videos & Tactile

Kelin Yu

3-rd Ph.D at the University of Maryland,
Advisor: Prof. Ruohan Gao

Research Intern at Amazon Industrial Robotics Group
Mentors: Dr. Alice Wu and Prof. Nima Fazeli



Imitation from Videos, Practice with ourselves

- Human needs to do practice after watching expert demos



Expert Demos



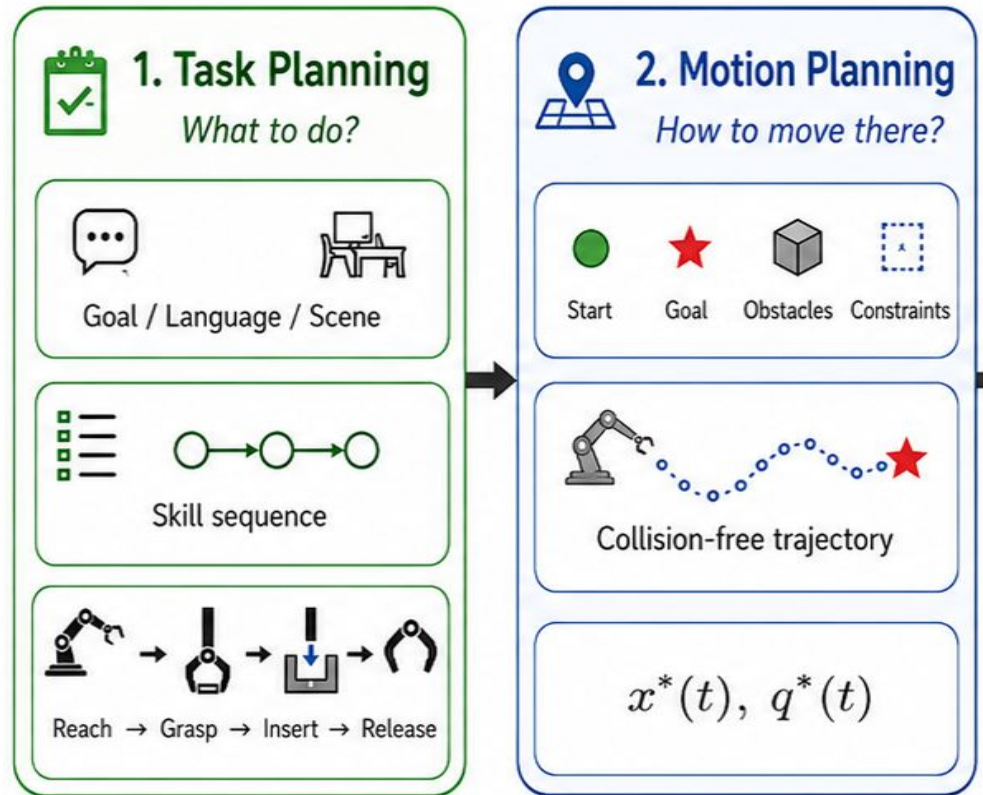
Practice

Imitation from Videos, Practice with ourselves

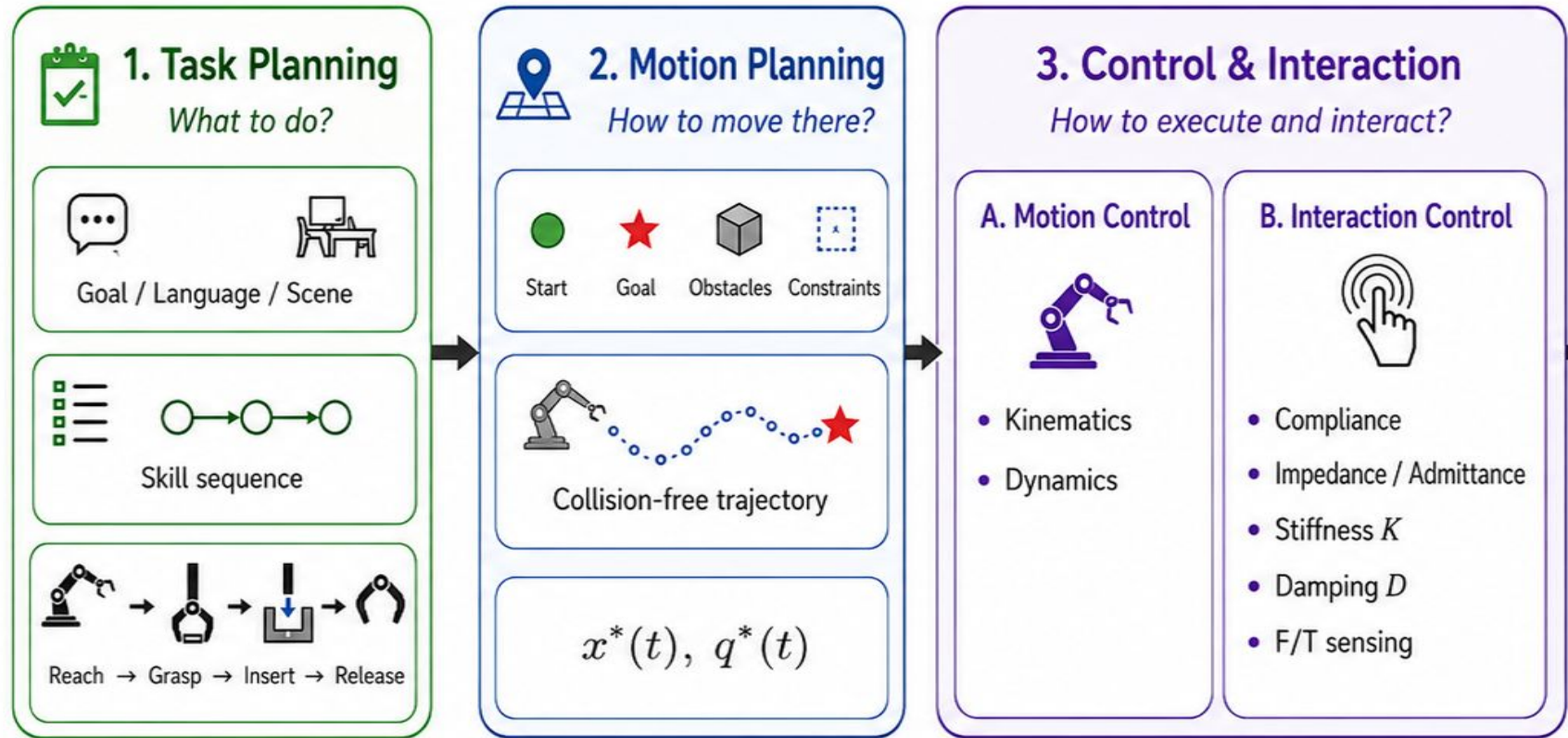


Imitate such behaviors from videos is impossible due to the embodiment gaps

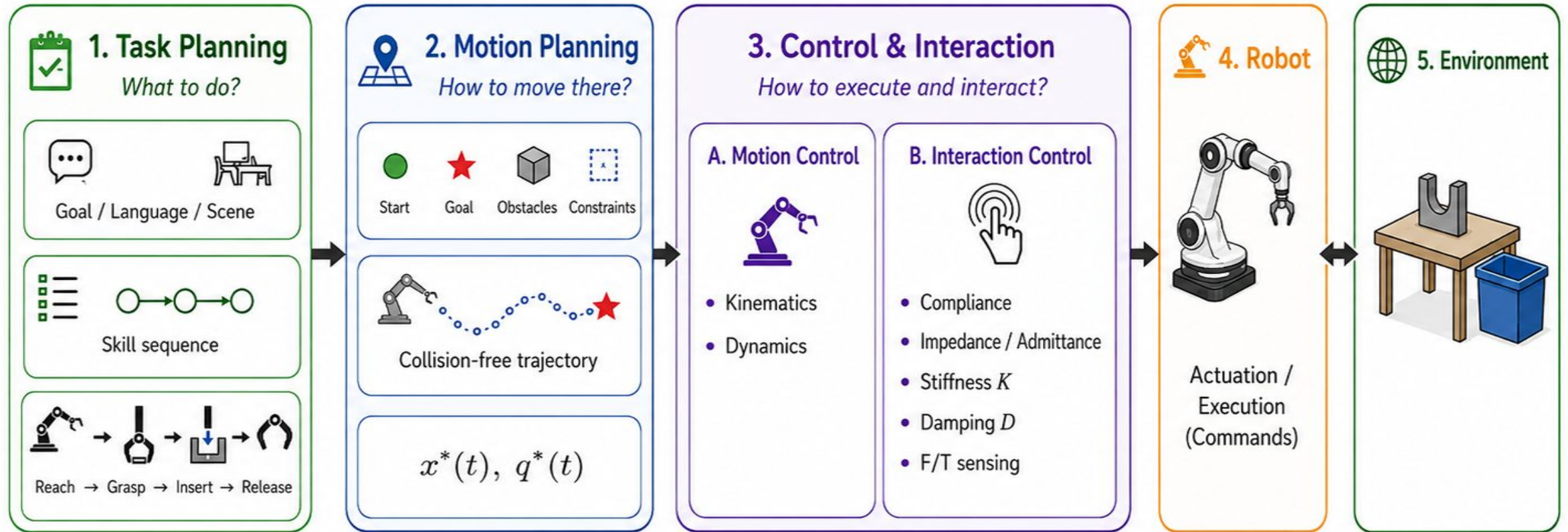
Robot Manipulation in My Mind



Robot Manipulation in My Mind



Robot Manipulation in My Mind



What can Robots Learn from Humans?

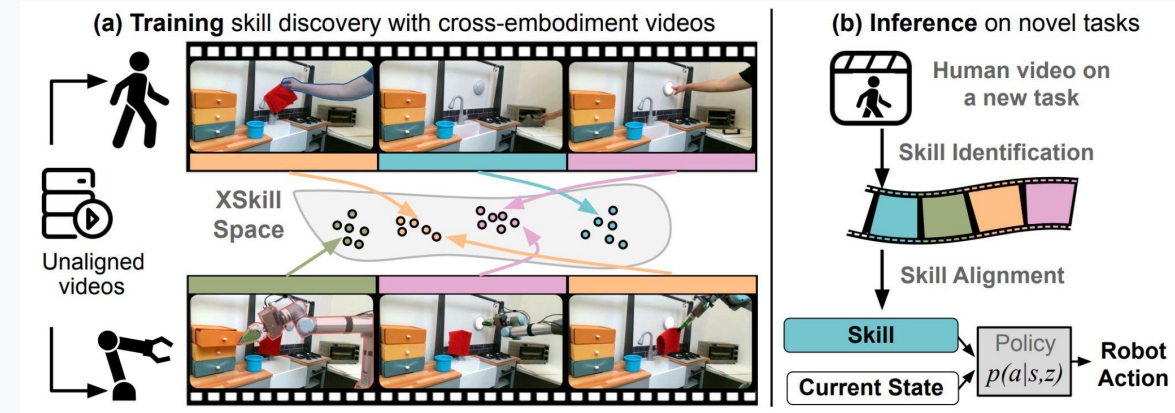
Task Planning

- Human's expert knowledge as task planner



SayCan, Zeng et al, CoRL 2022

Language guidance for Task Planning



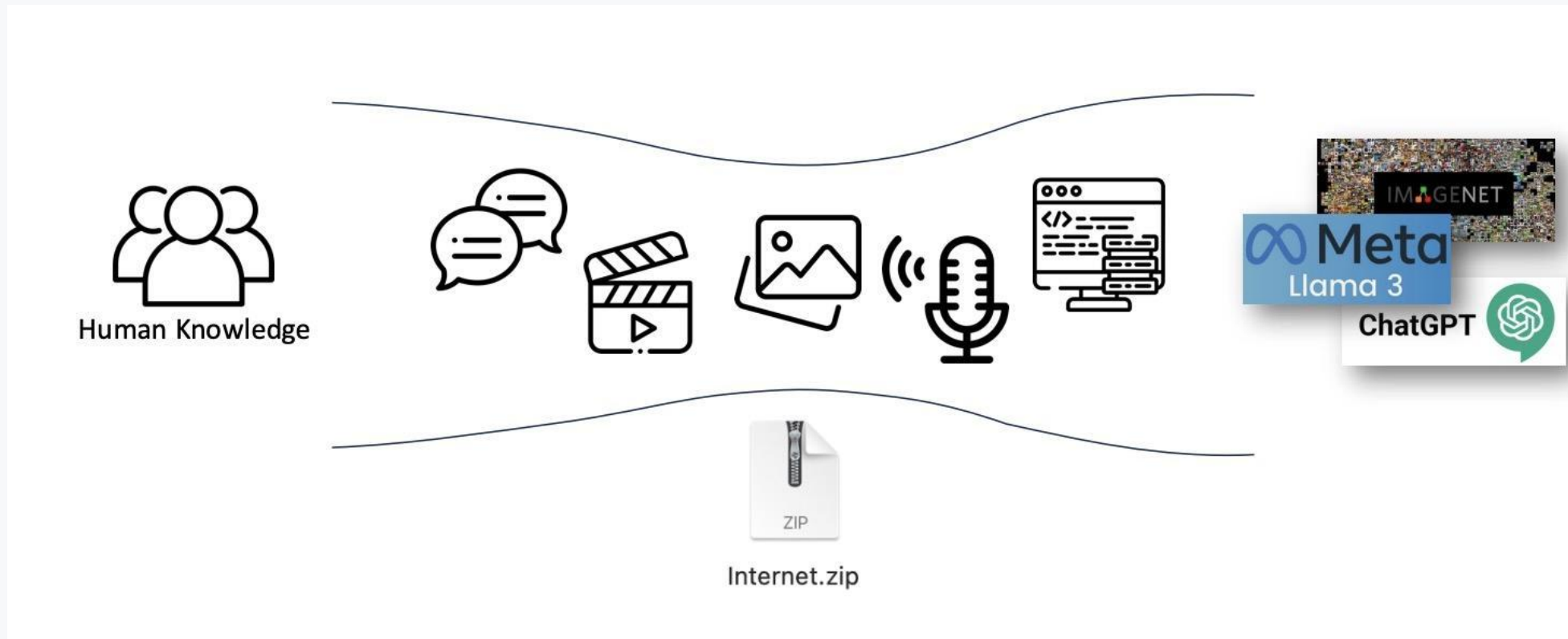
XSkill, Xu et al, CoRL 2024

Human video as the Task Planner

What can Robots Learn from Humans?

Motion Planning + Kinematics-based Control

- Imitation Learning from Human Data



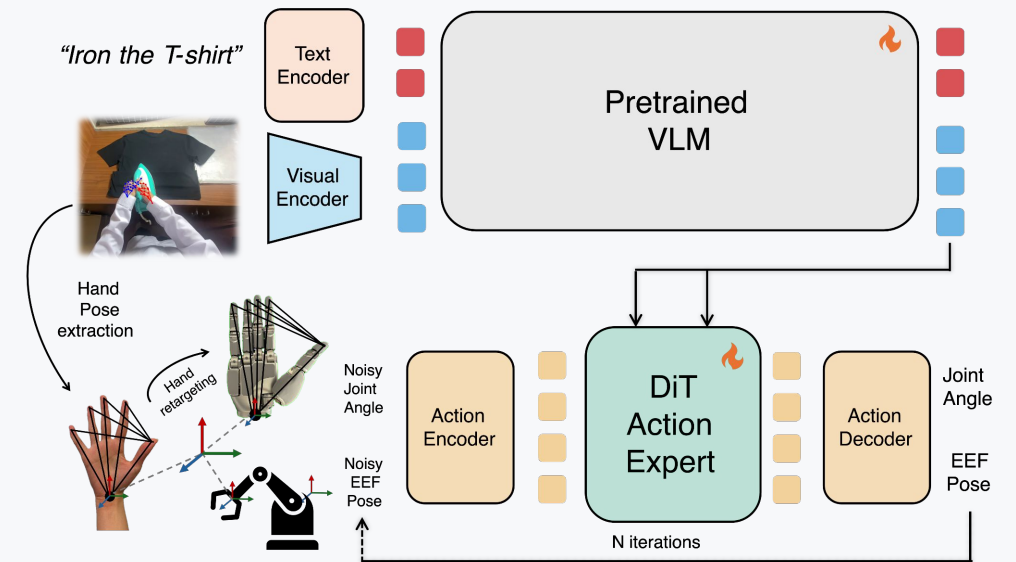
What can Robots Learn from Humans?

Motion Planning + Kinematics-based Control

- Imitation Learning from Human Data



EgoMimic, Kareer et al, ICRA 2025

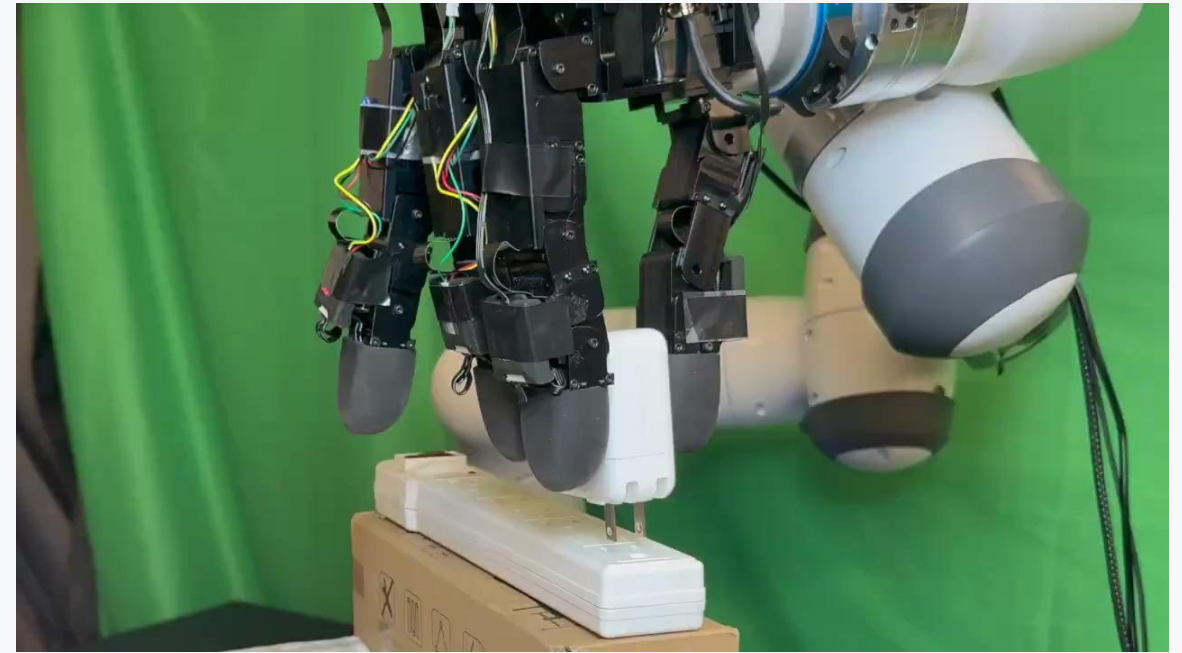
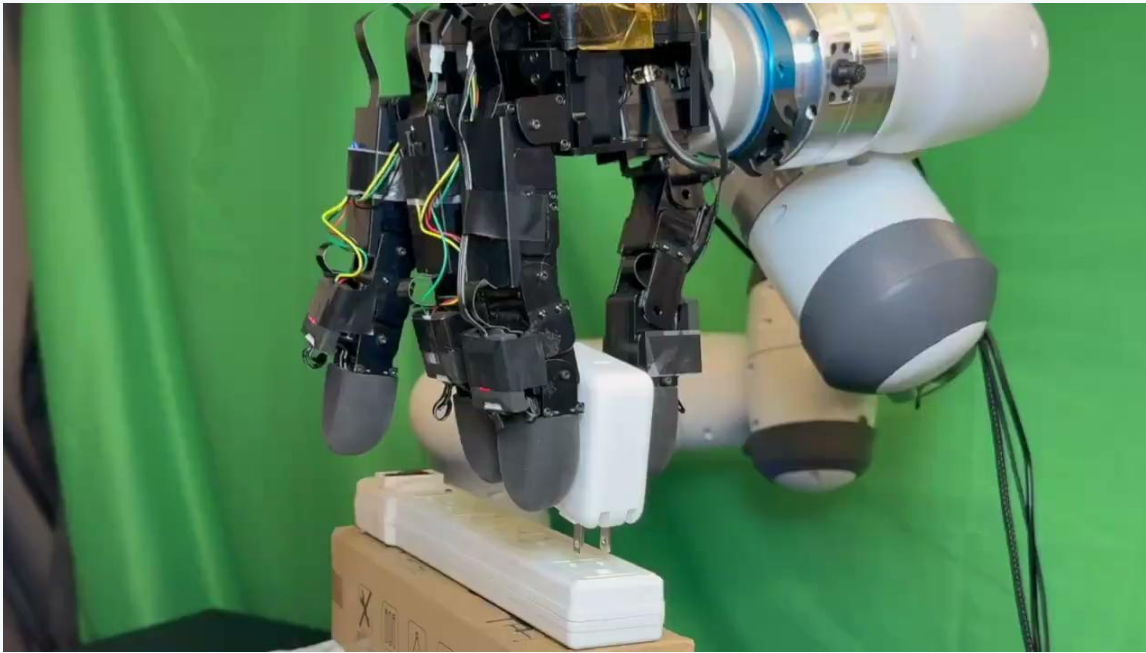


EgoScale, Zheng et al, Arxiv 2026

What can Robots Learn from Humans?

Motion Planning + Kinematics-based Control

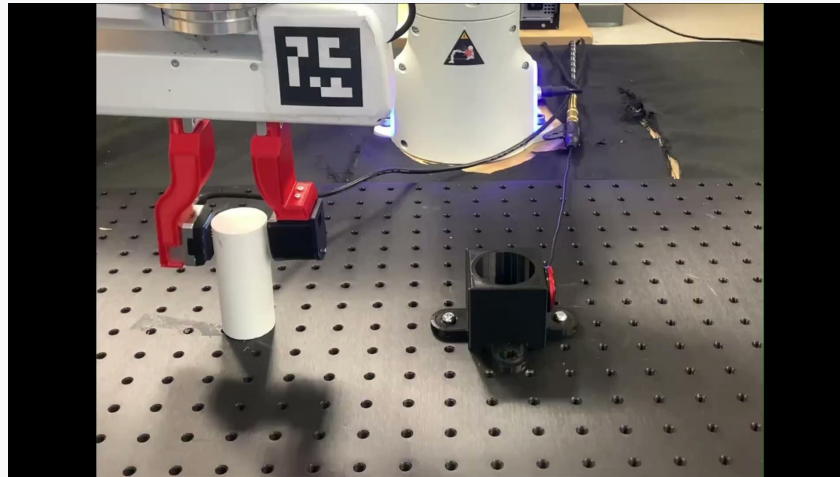
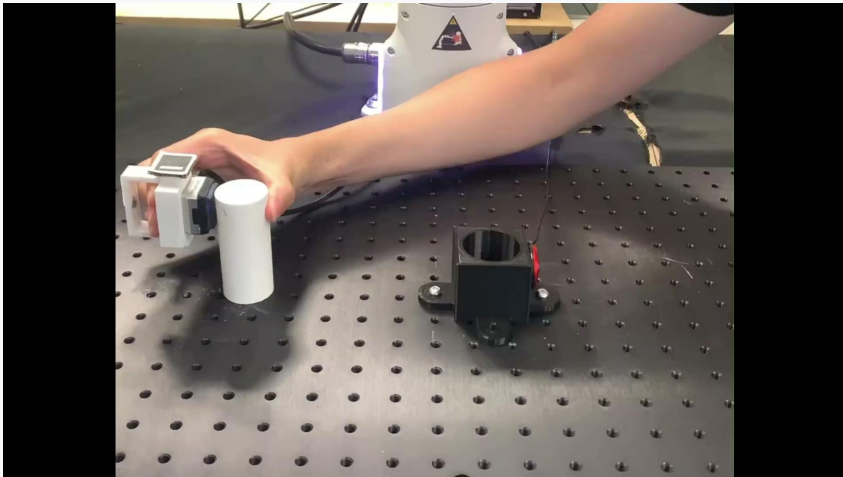
- Imitation Learning from Human Data
- Not enough for contact-rich manipulations



What can Robots Learn from Humans?

Contact-rich manipulations

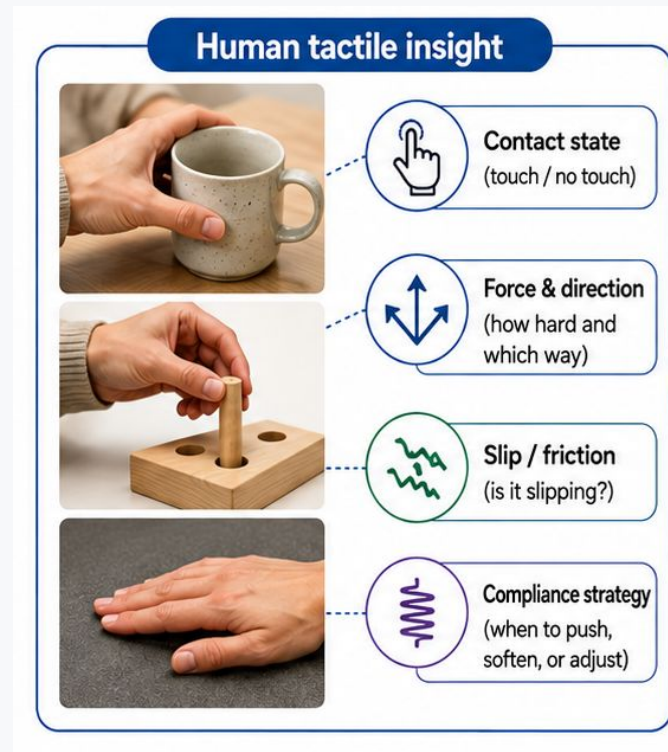
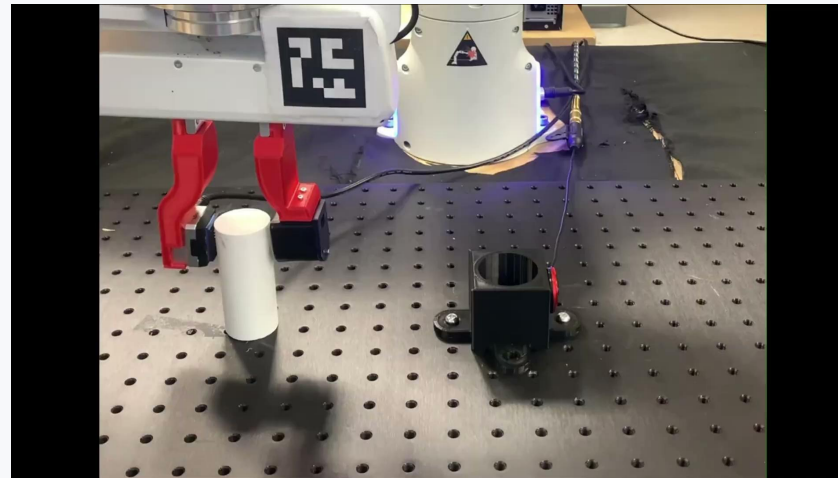
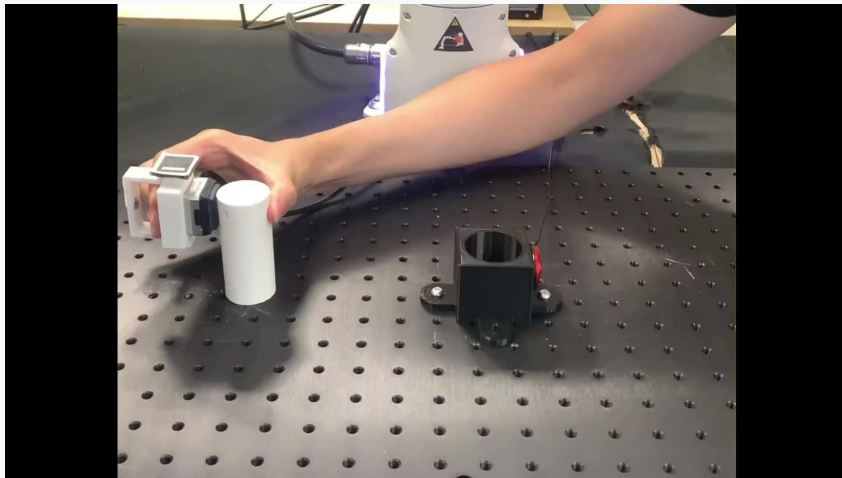
- Smooth, Diverse, Faster Motion for contact-rich manipulations



What can Robots Learn from Humans?

Contact-rich manipulations

- Smooth, Diverse, Faster Motion for contact-rich manipulations
- Tactile insight for contact reasoning and reactive behavior



Contact Rich Manipulation from Human Video and Tactile

Abundant human videos

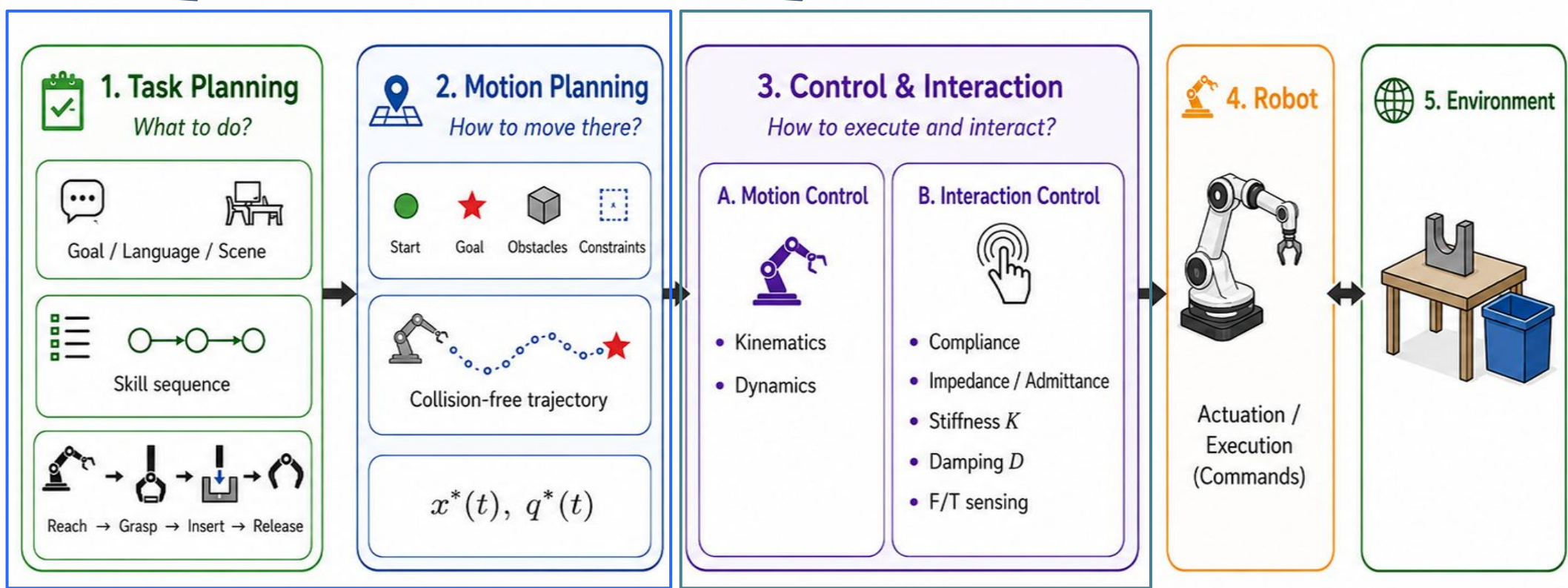
rich interaction priors
but no force or tactile signal

Scarce tactile data

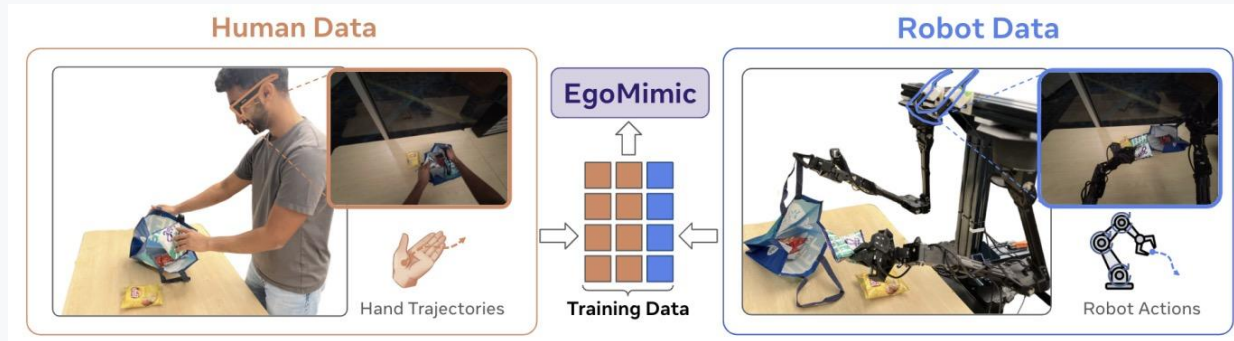
essential for contact
but sensor-specific and costly

My goal

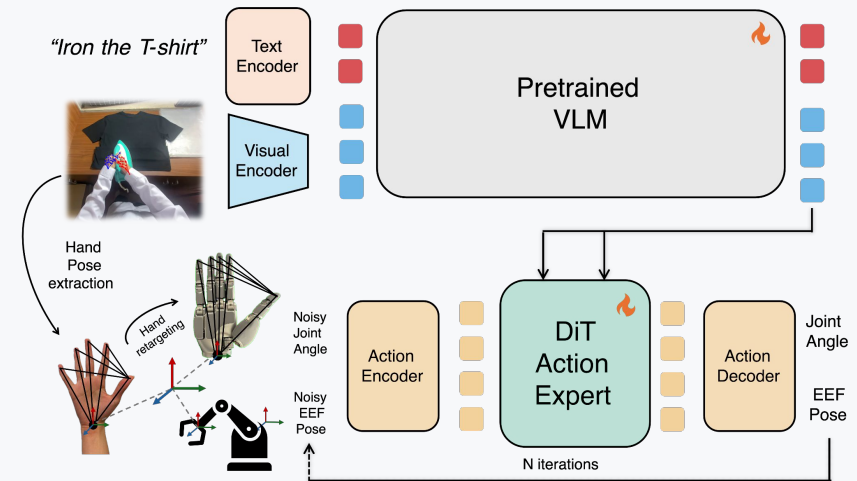
combine both to build robots
that anticipate contact



Learning from Human Videos



EgoMimic, Kareer et al, ICRA 2025



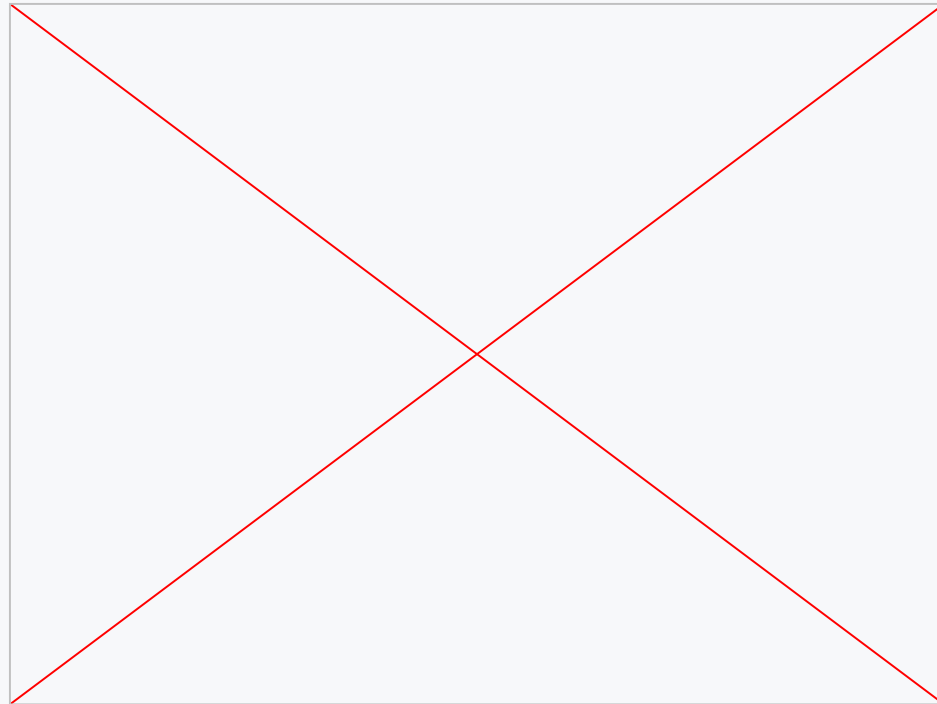
EgoScale, Zheng et al, Arxiv 2026

- Extract human hand pose, utilize human data as robot data
- Cotraining and Pretraining, with teleoperated robot data

Human and Robot have Embodiment Gaps

Huge embodiment gap:

- Dexterity/Action gaps between human and robot



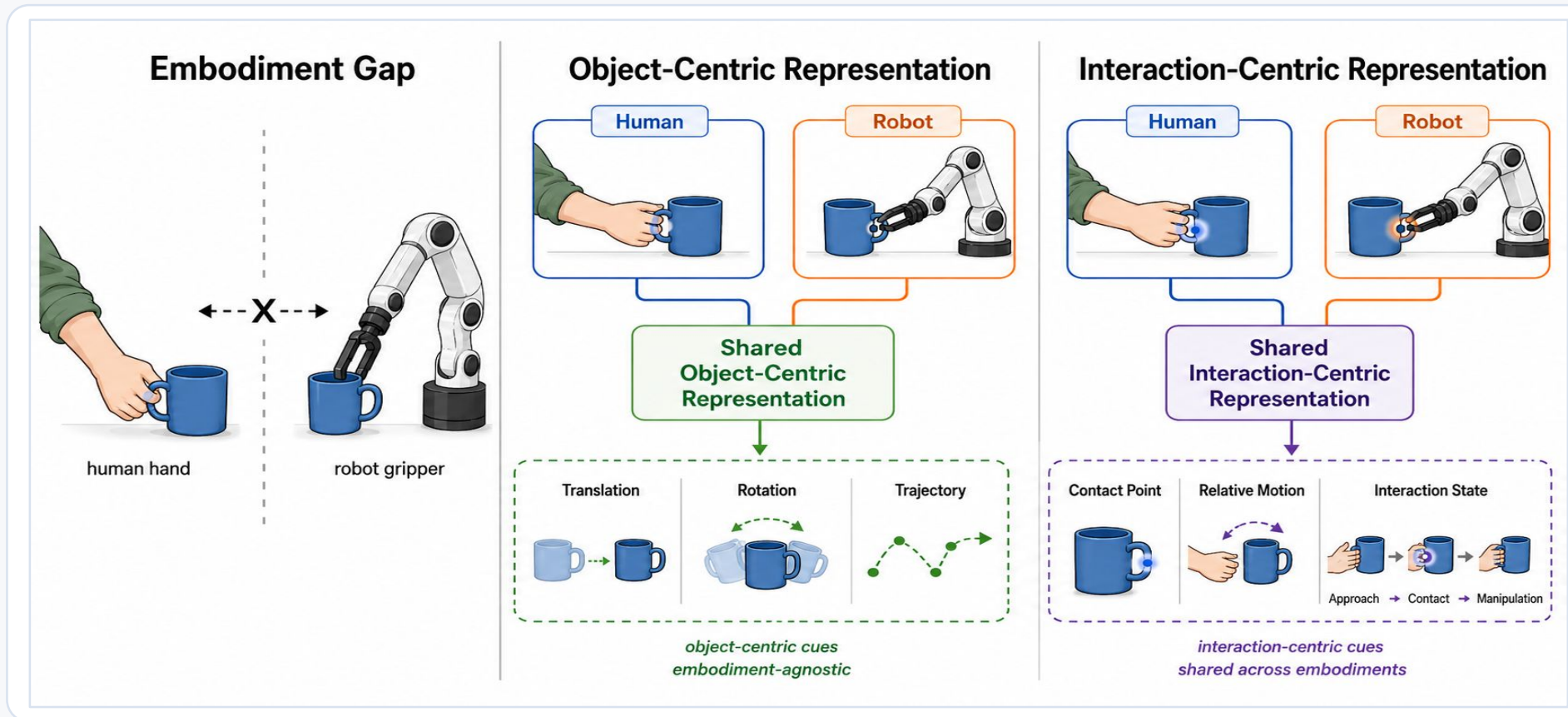
Human and Robot have Embodiment Gaps

Huge embodiment gap:

- Dexterity/Action gaps between human and robotx
- Sensory gaps (visual gap)



Human Videos Transfer through Objects and Interaction

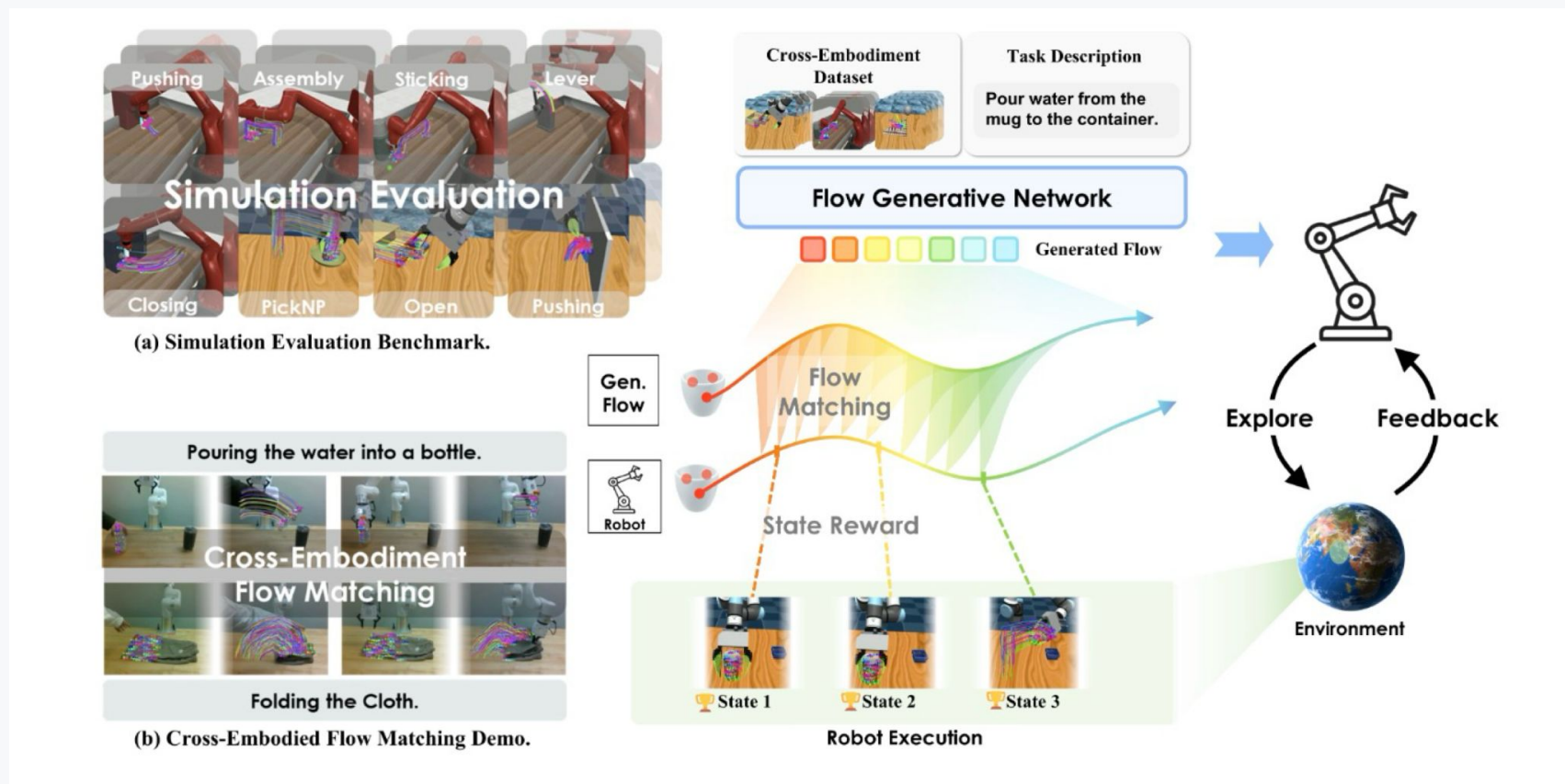


Human video is the largest manipulation dataset we have. The transferable part is not the demonstrator's embodiment — it is the task-relevant object motion.

What should happen?

Object-centric priors give policies a high-level physical intention before any robot-specific adaptation.

GenFlowRL: Shaping Rewards with Generative Object-Centric Flow in Visual Reinforcement Learning

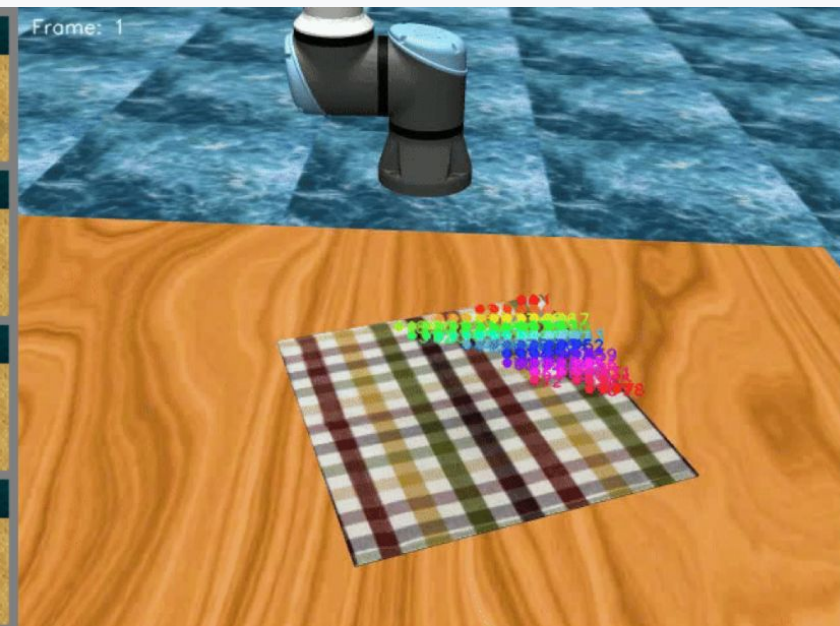


Kelin Yu*, Sheng Zhang*, Harshit Soora, Furong Huang, Heng Huang, Pratap Tokerr, Ruohan Gao
 ICCV 2025, **Spotlight and Best Paper Nomination** at ICRA FMNS Workshop

Generated Flow for Policy Design and Reward Shaping



A. Escontrela, et. al, NeurIPS 2023

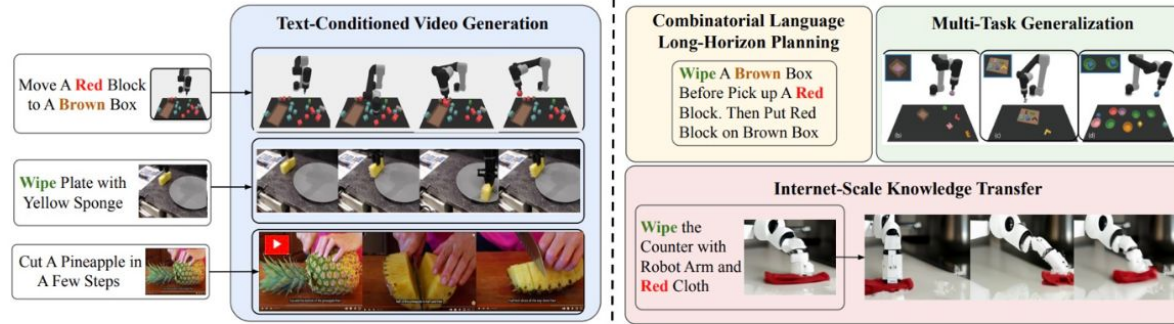


M. Xu, et. al, CoRL 2024

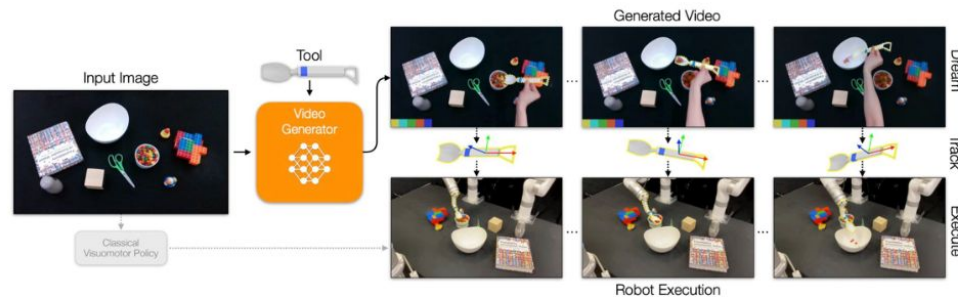
Motivation

- Generated video has been widely used for reward shaping, but hard to extract manipulation-centric features like motions
- Video is hard to be cross-embodiment
- We use shared object-centric flow for both policy design and reward shaping

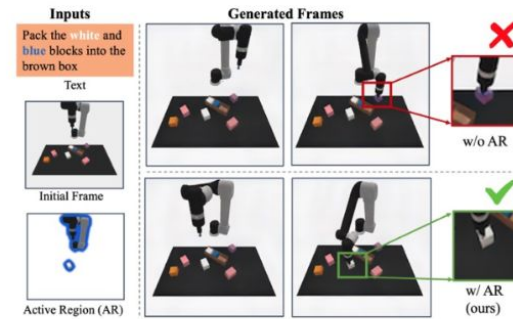
Generative Model For Robot Learning



Du. Y et al., 2023

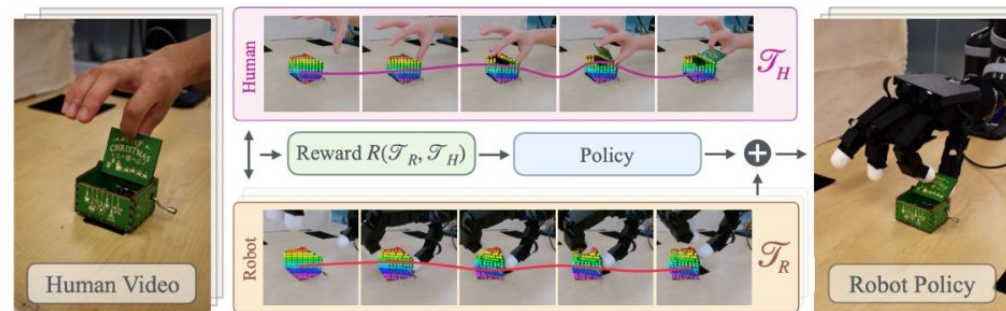
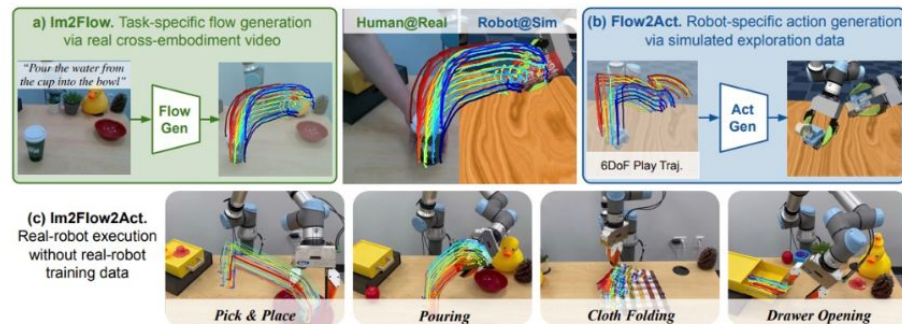
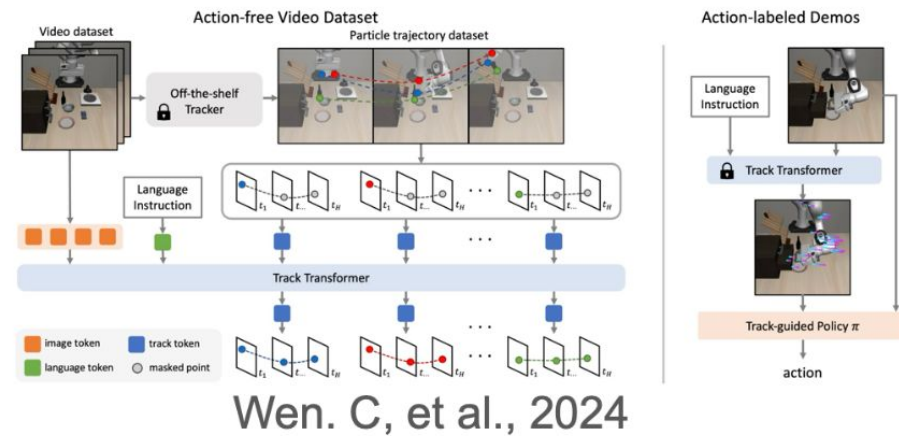


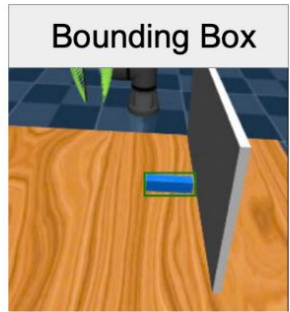
Liu. R et al., 2024



Huang. S et al., 2024

Flow-guided Robot Learning

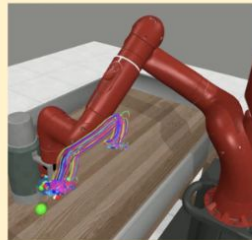




Sphere Robot



Sawyer Robot



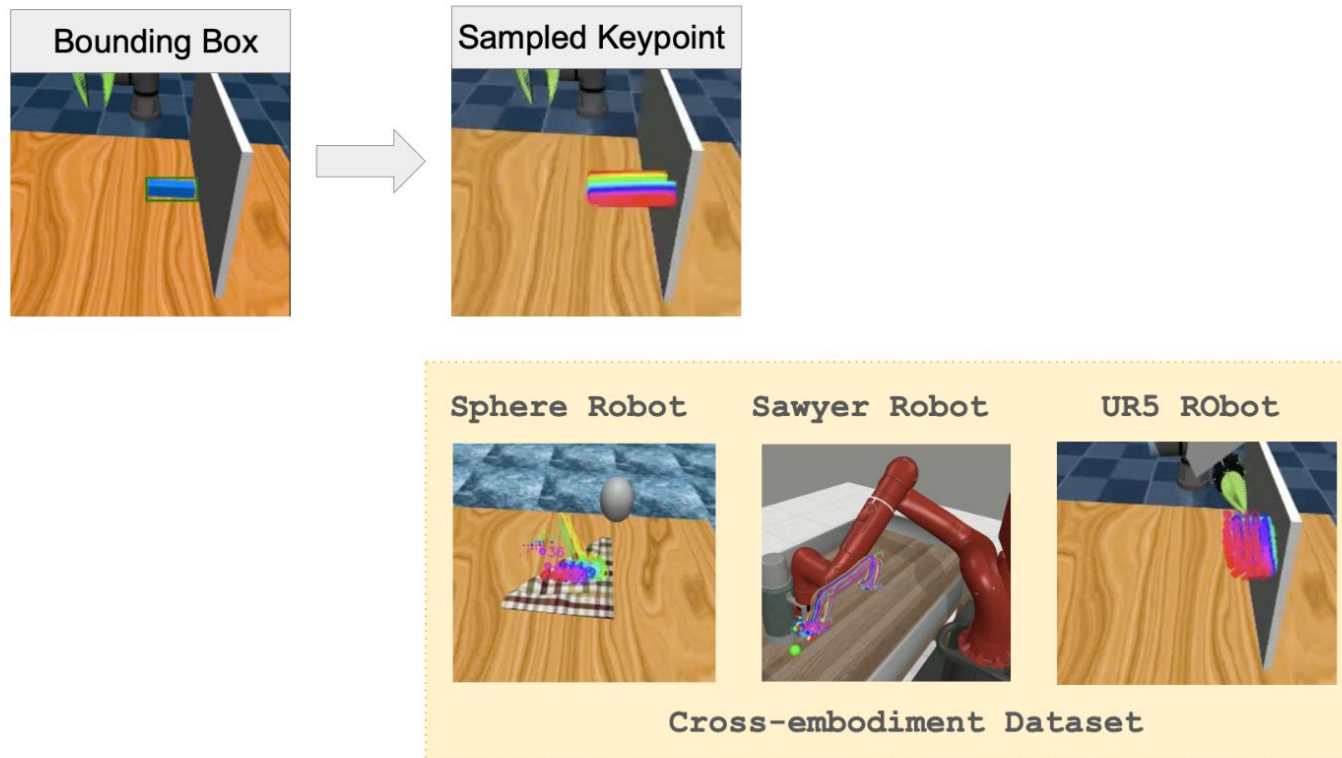
UR5 Robot



Cross-embodiment Dataset

Generate cross-embodiment object flow dataset:

- Detect object with a bounding box



Generate cross-embodiment object flow dataset:

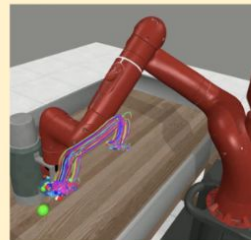
- Detect object with a bounding box
- Sampled Keypoints from the bounding box



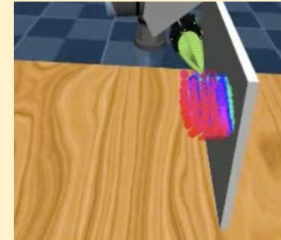
Sphere Robot



Sawyer Robot



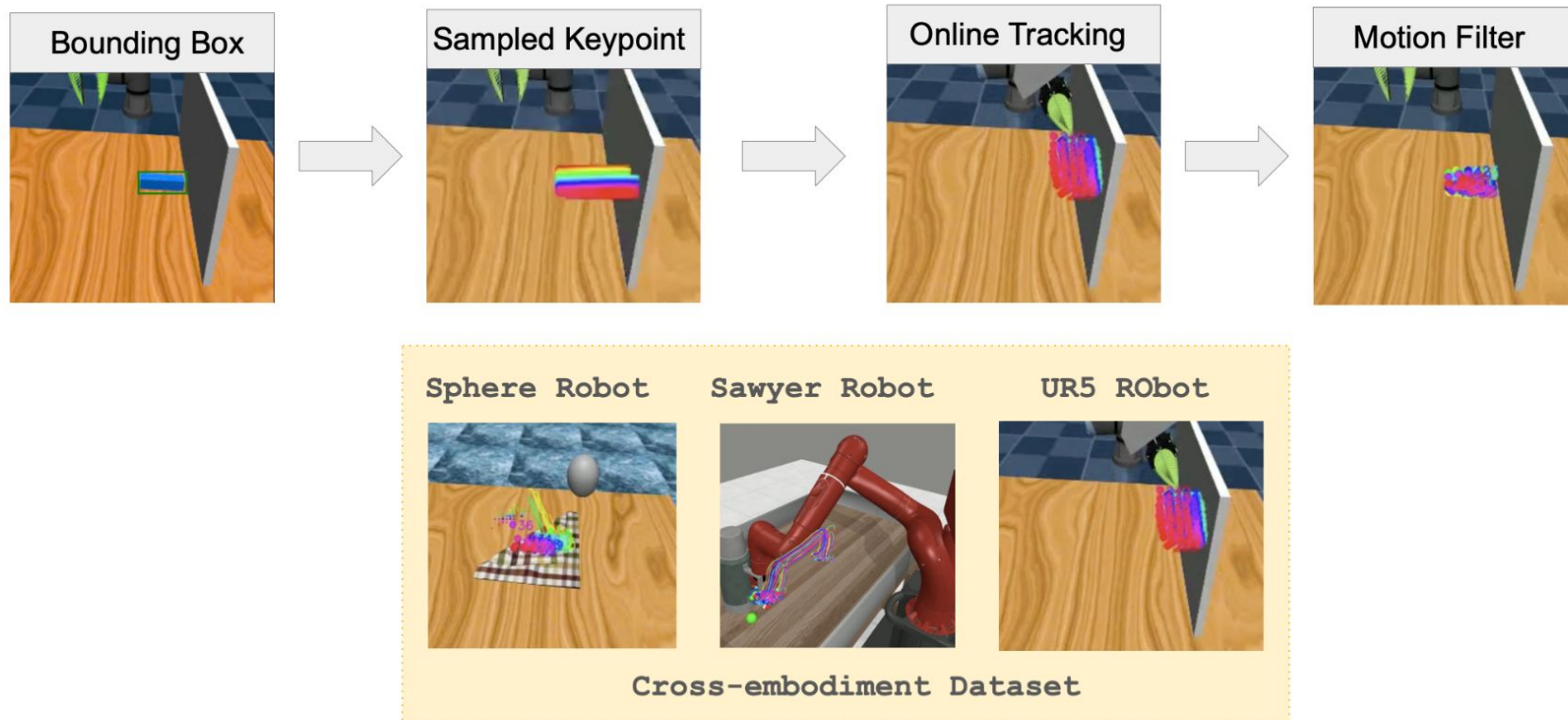
UR5 Robot



Cross-embodiment Dataset

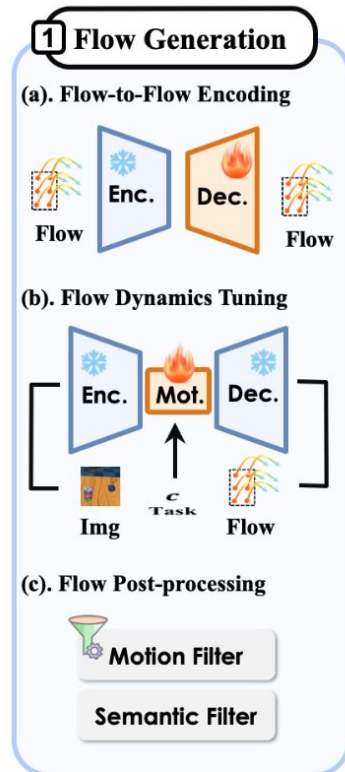
Generate cross-embodiment object flow dataset:

- Detect object with a bounding box
- Sampled Keypoints from the bounding box
- Doing Online Tracking with CoTracker



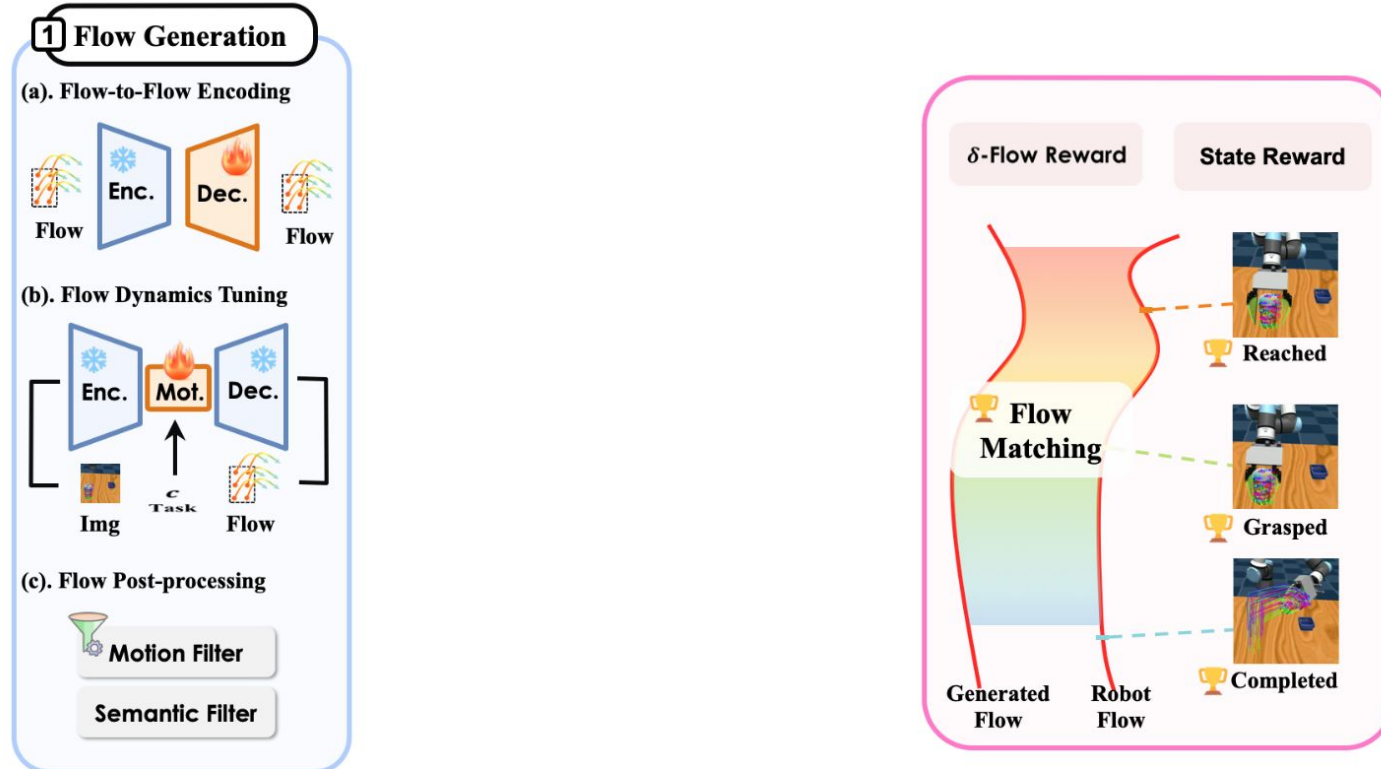
Generate cross-embodiment object flow dataset:

- Detect object with a bounding box
- Sampled Keypoints from the bounding box
- Doing Online Tracking with CoTracker
- Using Filter to remove keypoints out of the object



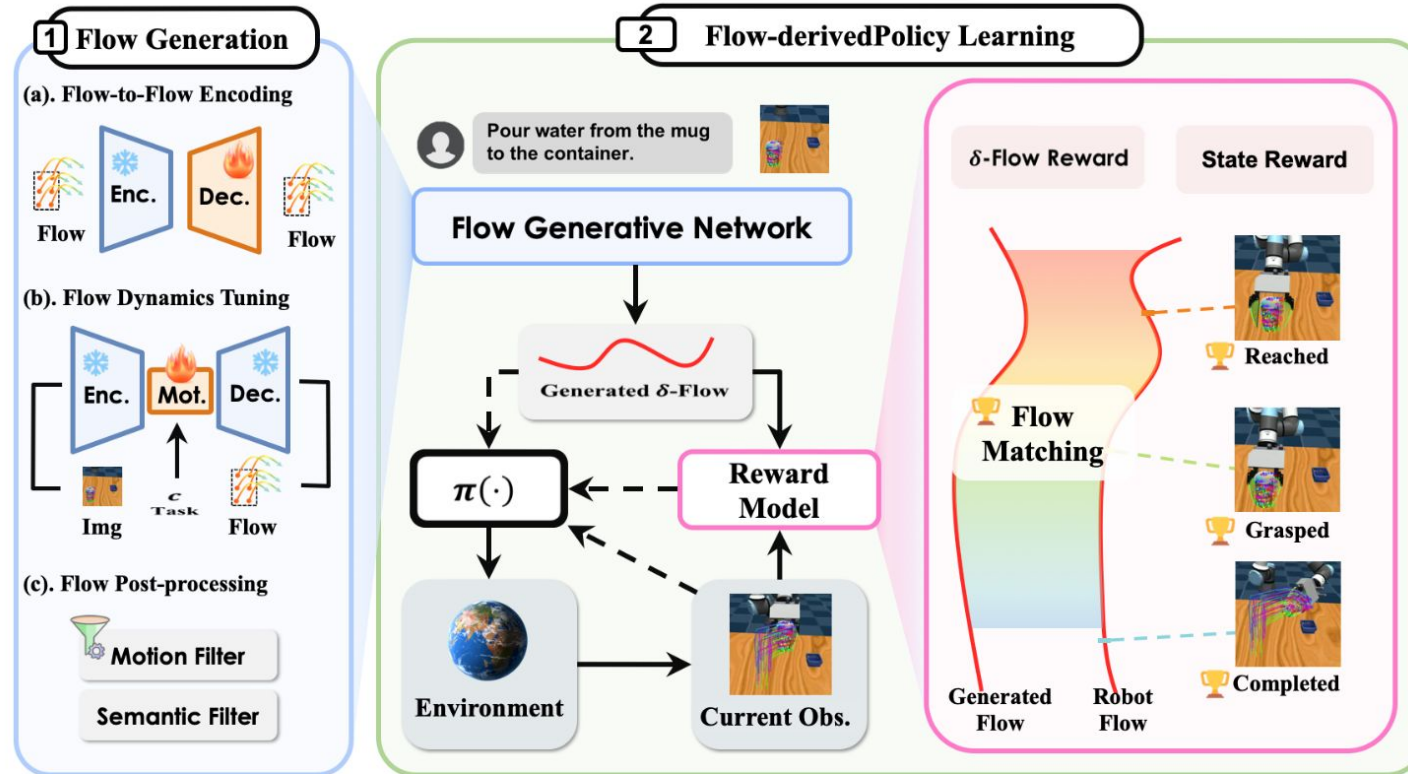
Formulate flow generation process with AnimateDiff

- Firstly, training the decoder of VQ-GAN in StableDiffusion
- Secondly, finetuning the UNet with LORA
- Before policy learning, process the generated flow with motion filter

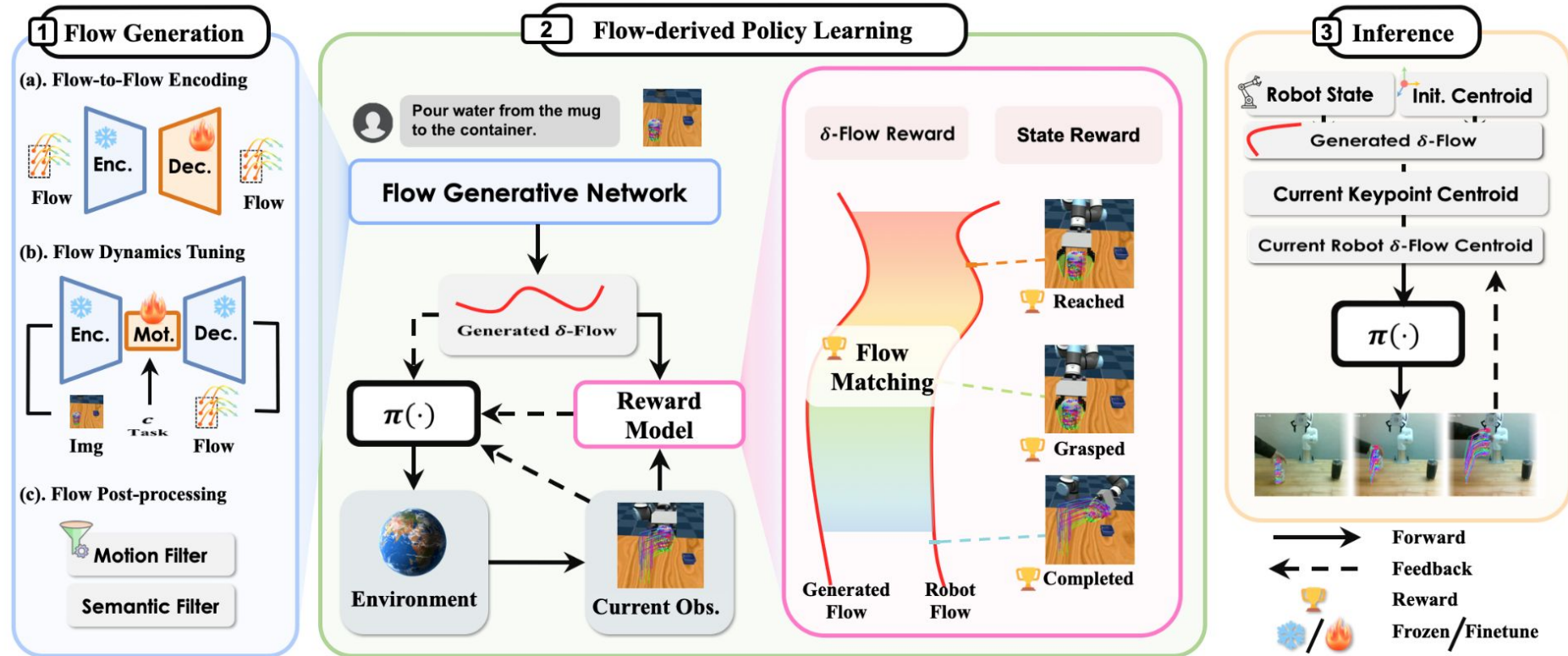


Adapt the expert flow into a flow-derived reward model

- Formulate the flow as δ -Flow and do flow matching between expert flow and execution flow for reward shaping
- Utilize task-specific state reward.

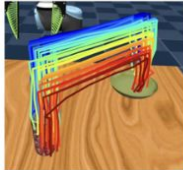
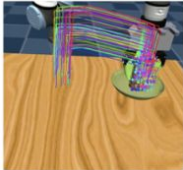
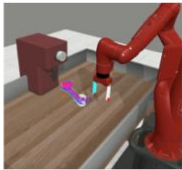
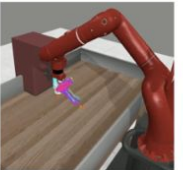
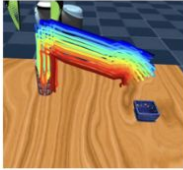
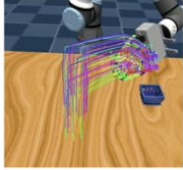
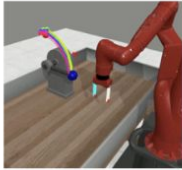
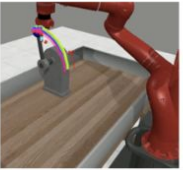
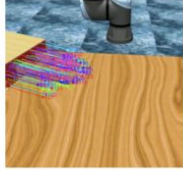
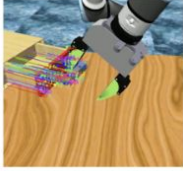
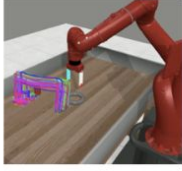
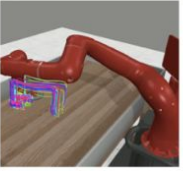
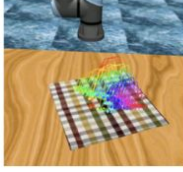
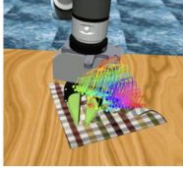
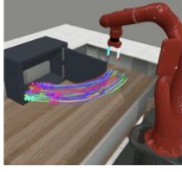
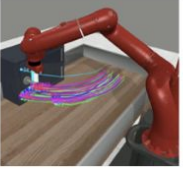
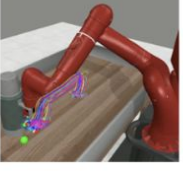


Then, we can utilize the flow-derived reward model in various task with our flow generative network.

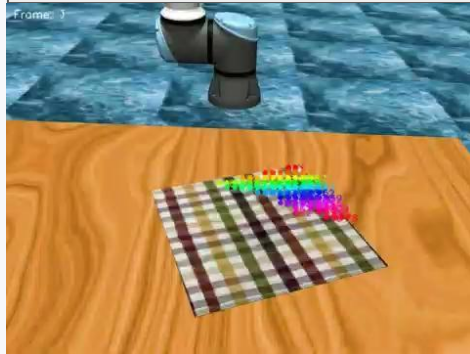
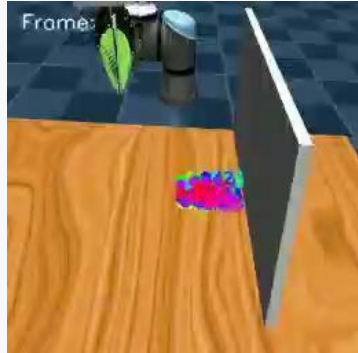


Finally, we train a flow-derived policy, which maps robot observation and generated δ -flow as the input into the robot 6Dof actions.

The task setup and execution for ten different manipulation tasks

Prompt	Generated Flows	Robot Execution	Prompt	Generated Flows	Robot Execution
PickNPlace Put the mug into the plate			Pusing Push the coffee mug to the goal		
Pouring Pour water from the mug to the bowl			Levering Push the lever to the top		
Opening Open the drawer			Assembly Assemble the peg with circle to the cylinder		
Pouring Fold the cloth to the diagonal corner			Closing Close the cabinet door		
Pouring Pivot the peg perpendicular to the wall			Stick Pulling Pick a stack and use it to pull the kettle		

Experiments: Tasks

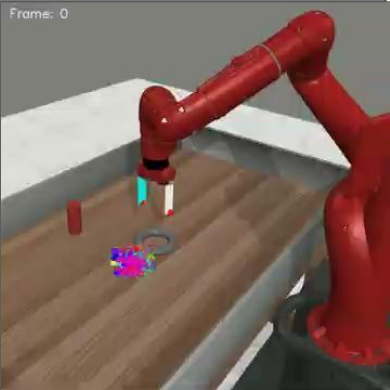
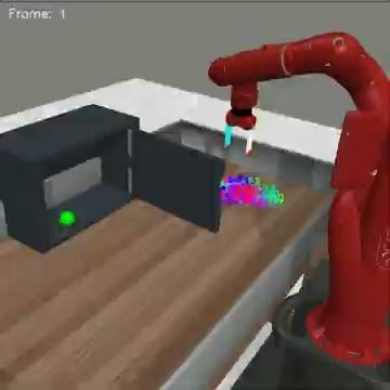
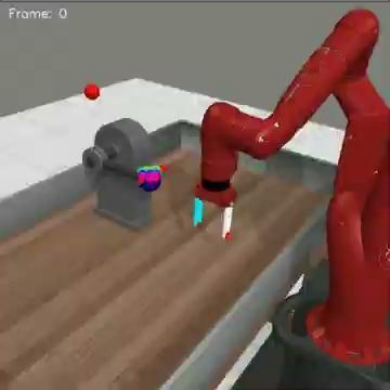
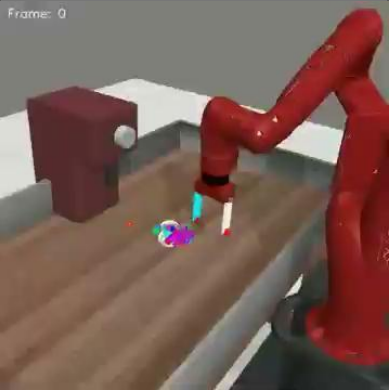
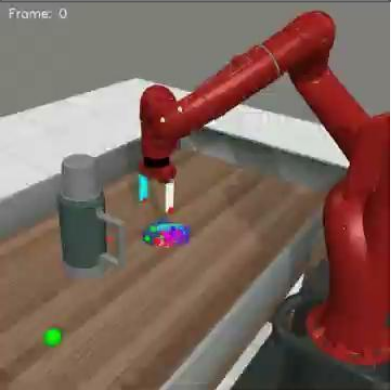
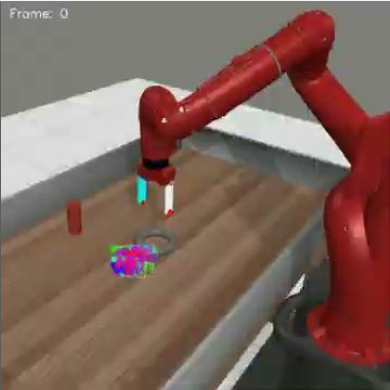
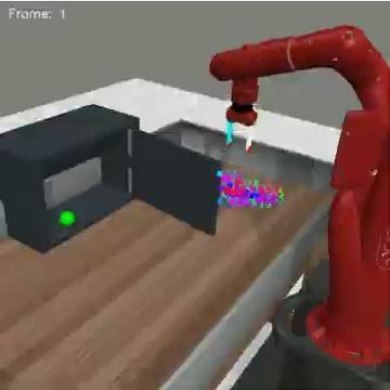
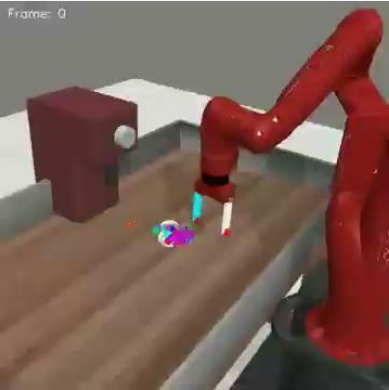
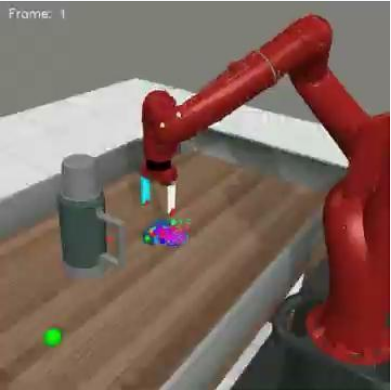
	PickNPlace	Pouring	Opening	Folding	Pivoing
Generated					
Rollout					

Experimental Results: Object-centric vs Video

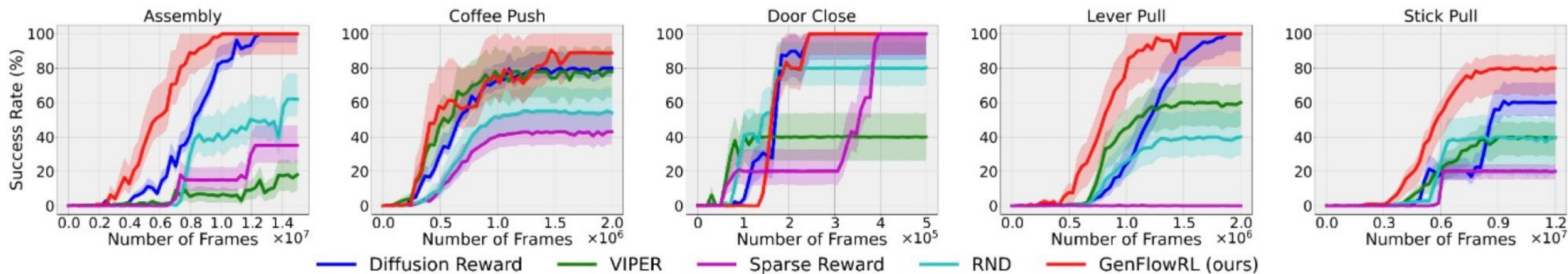
	Demonstration-Conditioned					Language-Conditioned				
	PickNP.	Pour	Open	Fold	Pivot	PickNP.	Pour	Fold	Fold	Pivot
Heuristic [49]	70	50	30	0	0	/	/	/	/	/
Im2Flow2Act [42]	100	95	95	90	60	90	85	90	35	45
GENFLOWRL (Ours)	100	100	100	95	90	95	95	95	80	85

- **Im2Flow2Act**: Flow Generation + Flow-based Imitation Learning
- **Heuristic**: 3D Flow as the Ground Truth + Trajectory Optimization

Experiments: Tasks

	Assembly	Close	Lever	Coffee	Stick
Generated	Frame: 0 	Frame: 1 	Frame: 0 	Frame: 0 	Frame: 0 
Rollout	Frame: 0 	Frame: 1 	Frame: 1 	Frame: 0 	Frame: 1 

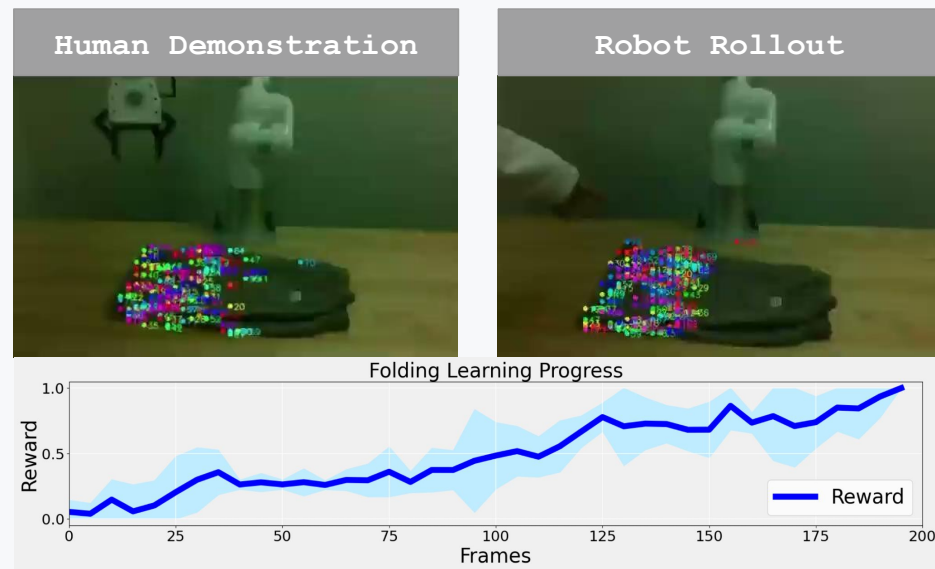
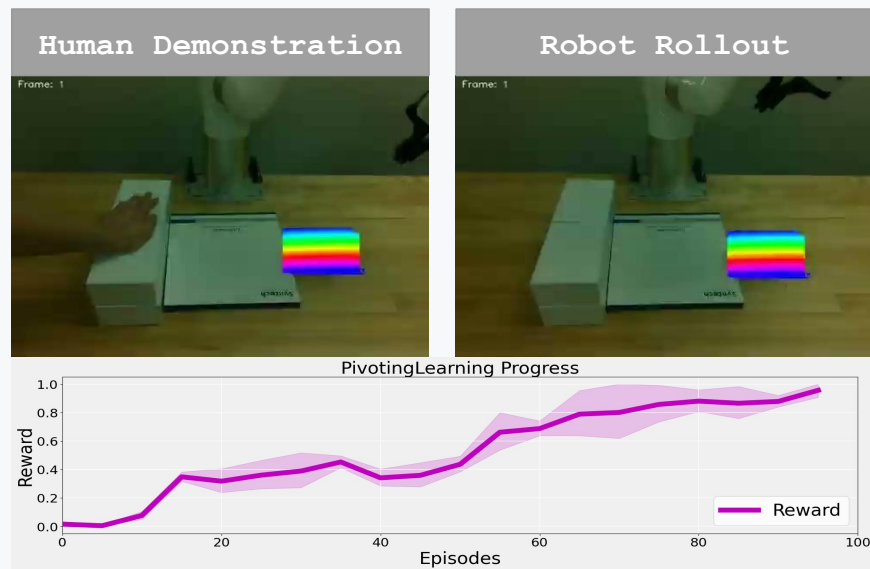
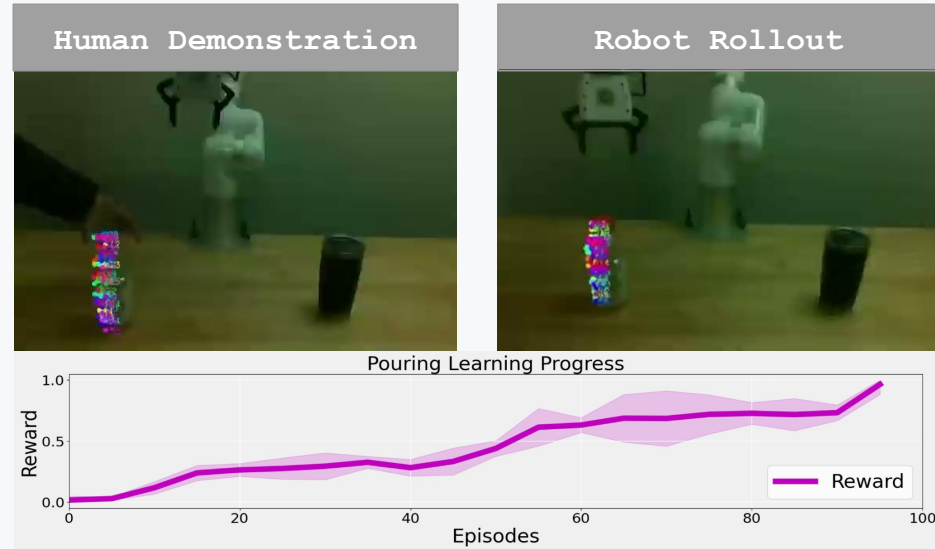
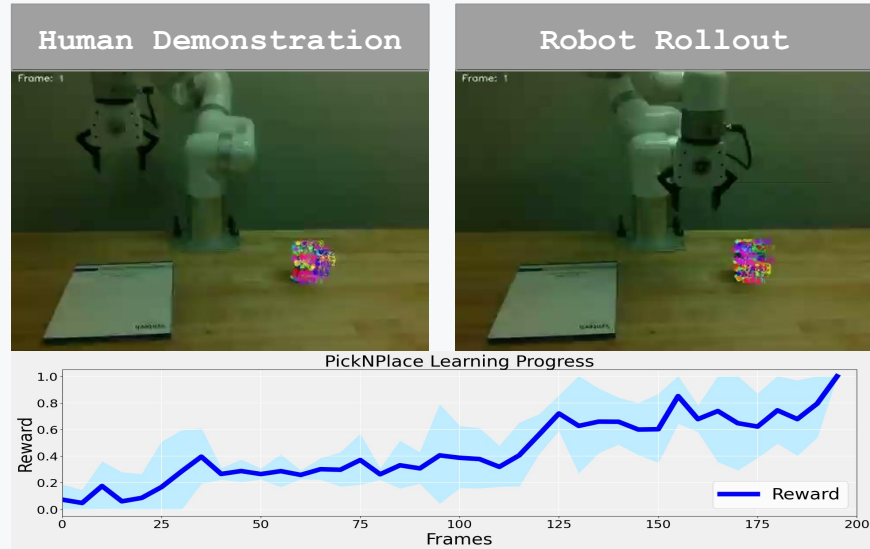
Experimental Results: Object-centric vs Video



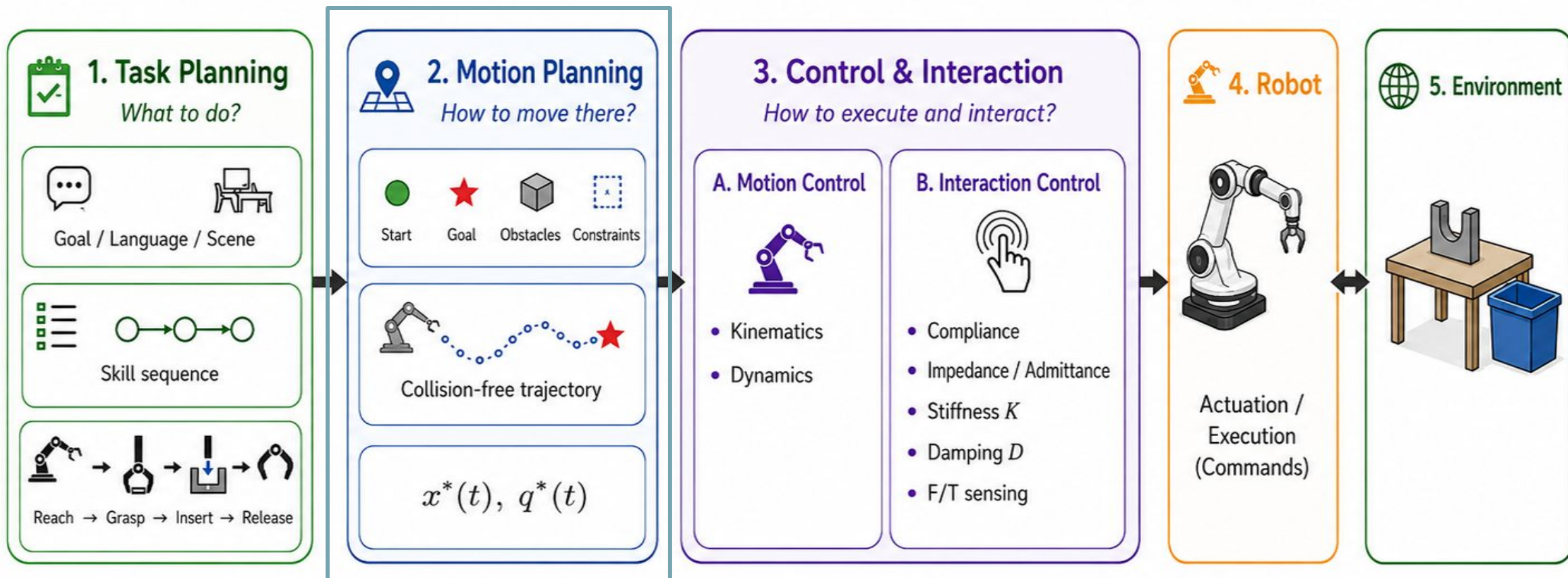
Compare our method with other video-guided RL baselines on the MetaWorld Benchmark:

- Diffusion Reward
- Video Prediction Model
- Random Network Distillation
- Sparse Reward

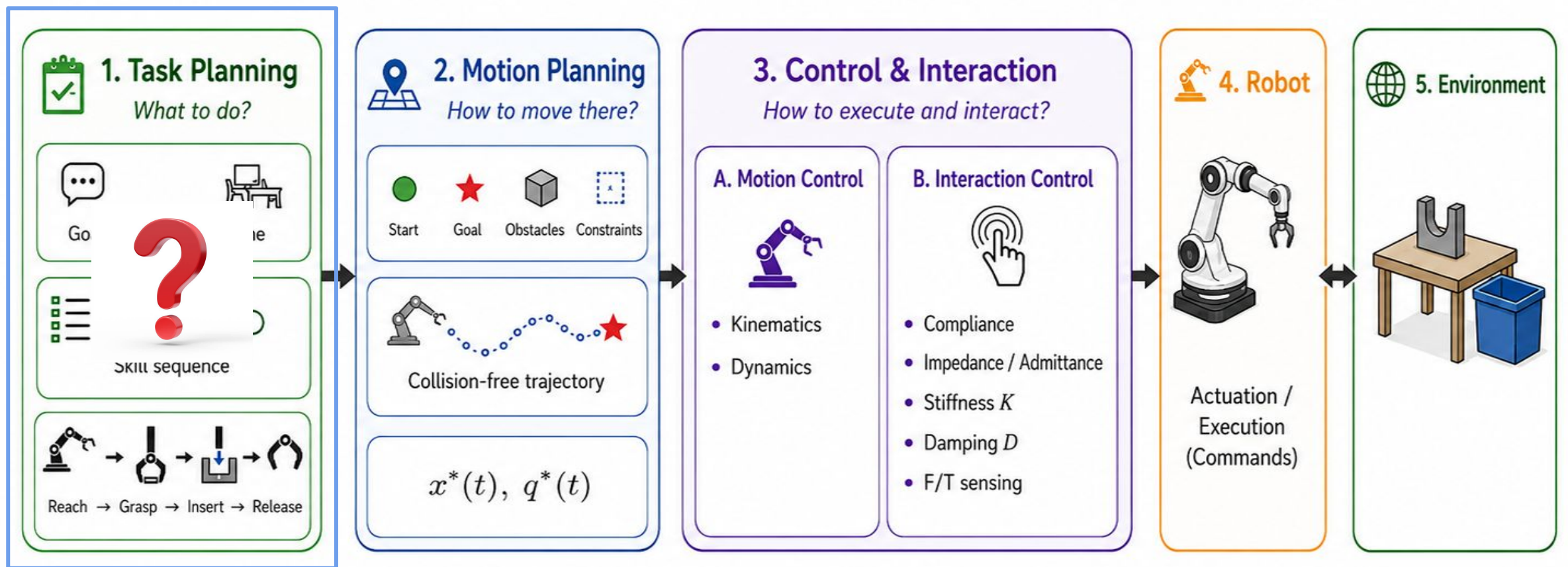
Case Study: Aligning Human Video with Robot Demos



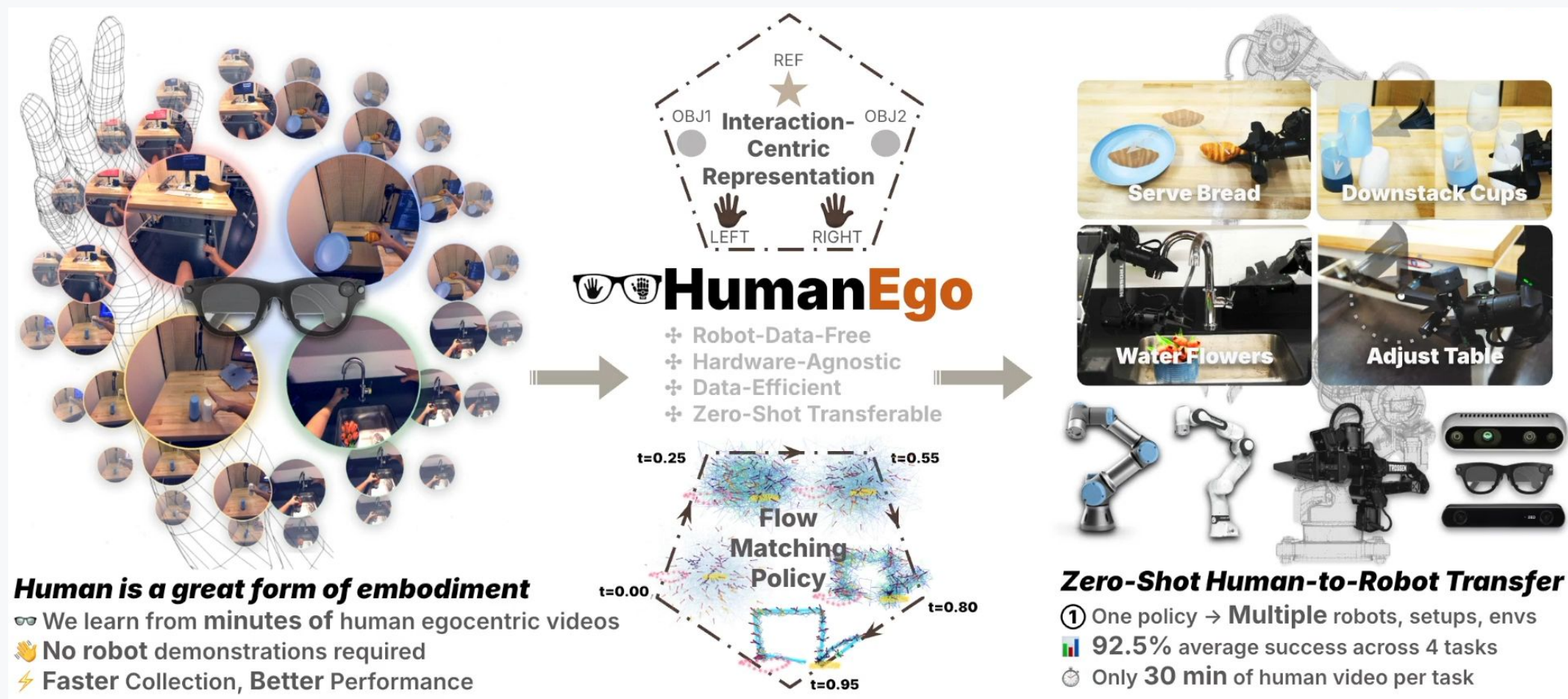
We learned motion priors from human videos via object-centric flow



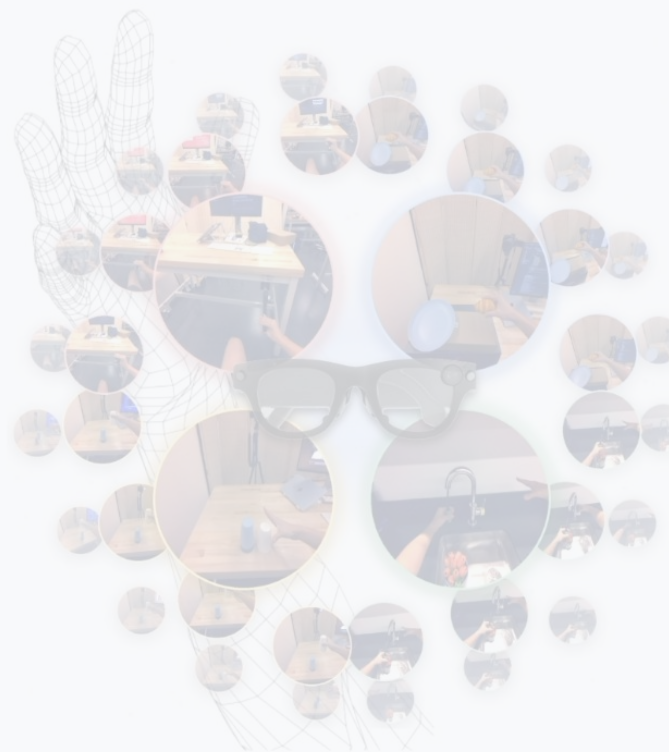
We learned motion priors from human videos via object-centric flow



HumanEgo: Zero-shot Human2Robot via Interaction-centric Representations



Interactive Centric Representations

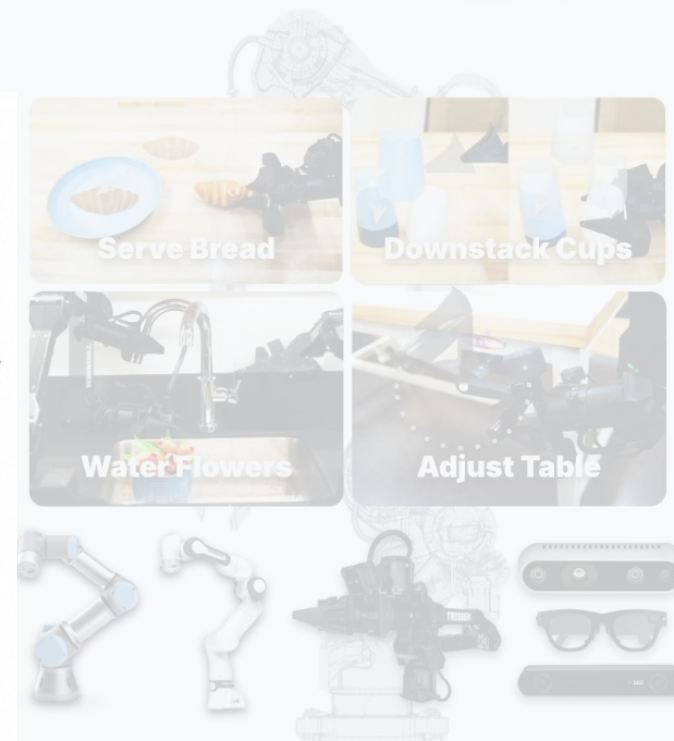
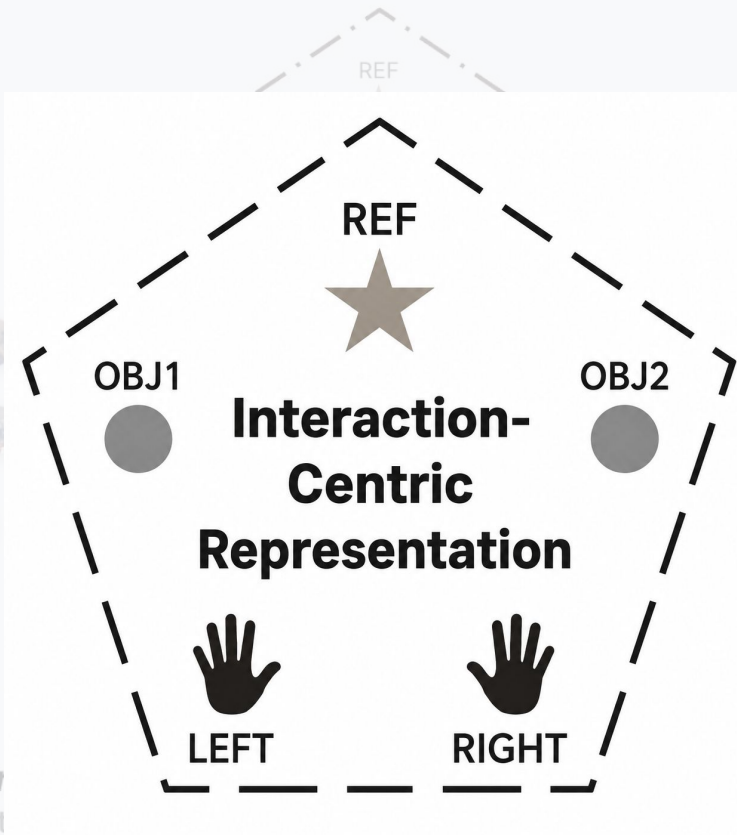


Human is a great form of embodiment

👁️ We learn from minutes of human egocentric video

👉 No robot demonstrations required

⚡ Faster Collection, Better Performance



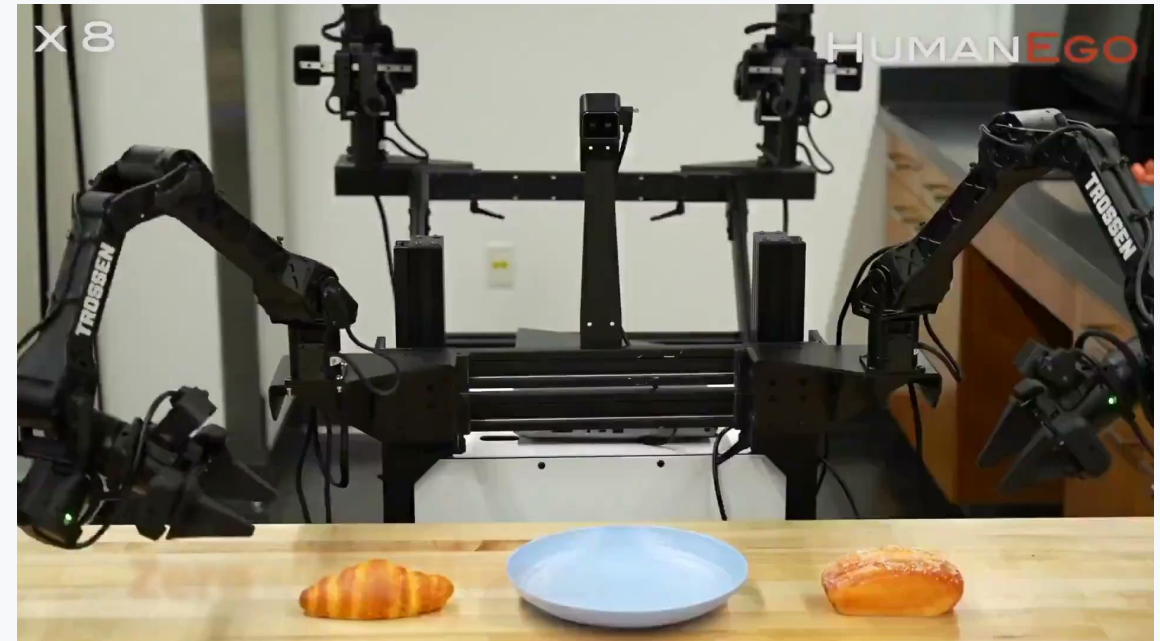
Zero-Shot Human-to-Robot Transfer

① One policy → Multiple robots, setups, envs

📊 92.5% average success across 4 tasks

🕒 Only 30 min of human video per task

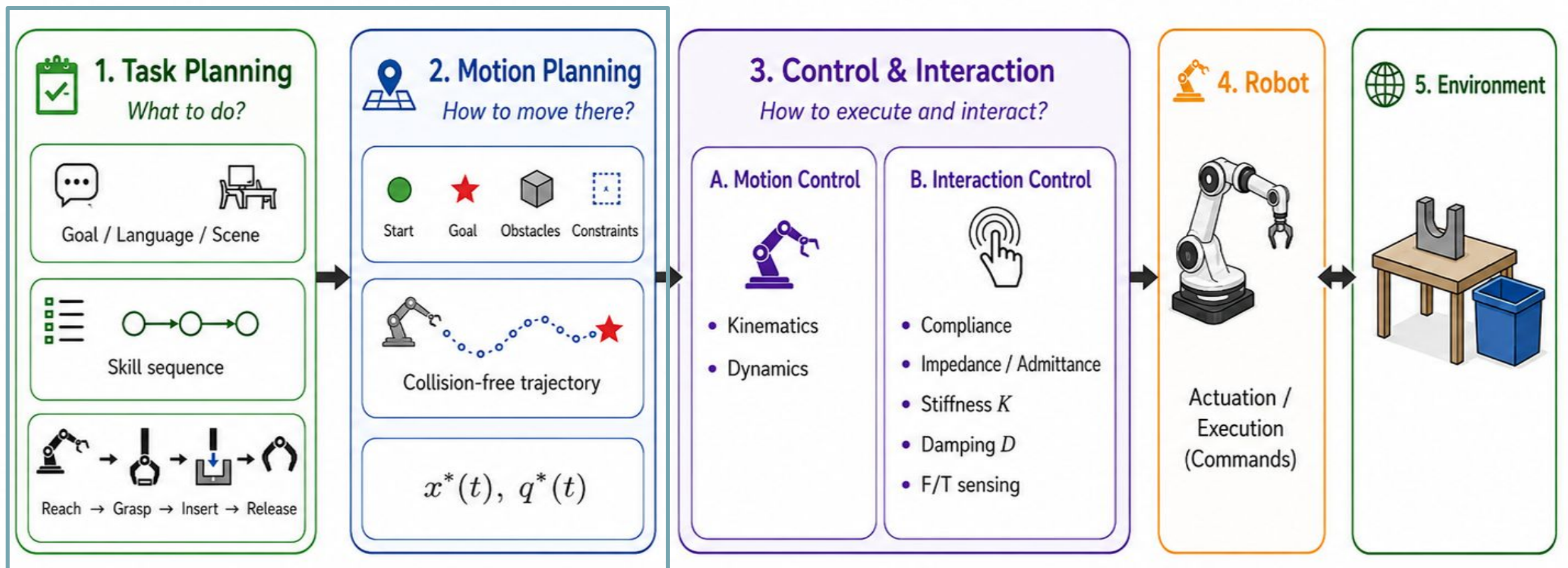
HumanEgo: Zero-shot Human2Robot via Interaction-centric Representations



Ego-centric Human Videos

Zero-shot Human2Robot Policy

We learned motion priors from human videos via object-centric flow



Why Robot need MultiSensory Perception



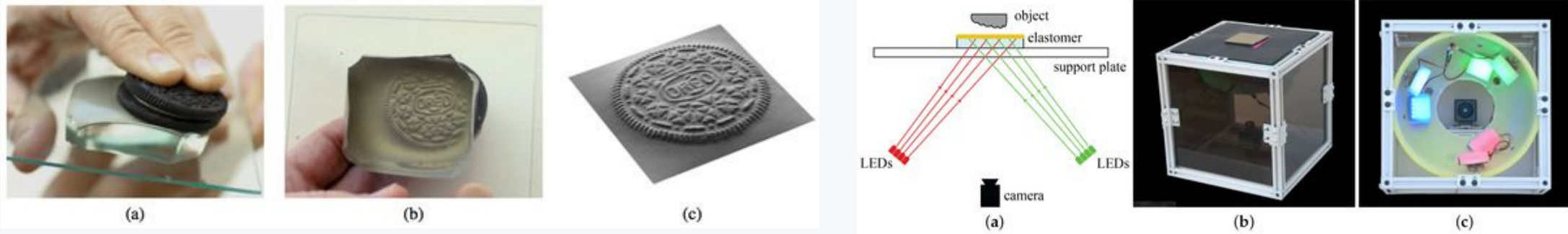
See and catch the ball

Feel the deformable of earphone

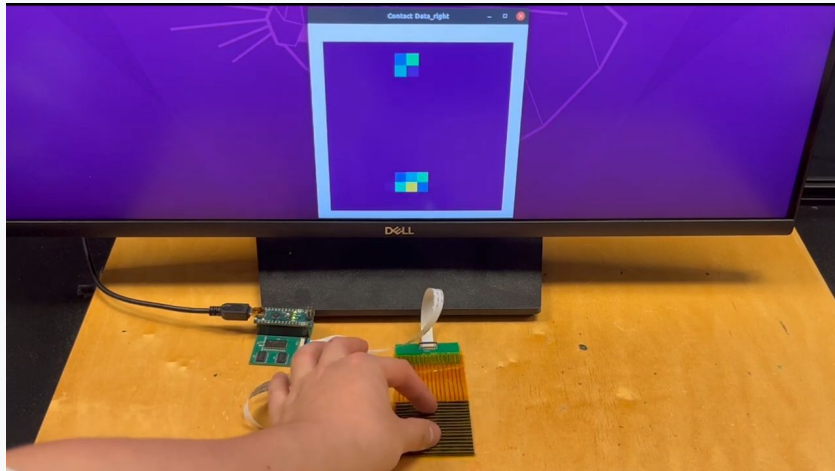
Hear the sound from Vacuum

H. Li,et al. "See, Hear, Feel: Smart Sensory Fusion for Robotic Manipulation". In CoRL, 2022.

Tactile Sensing



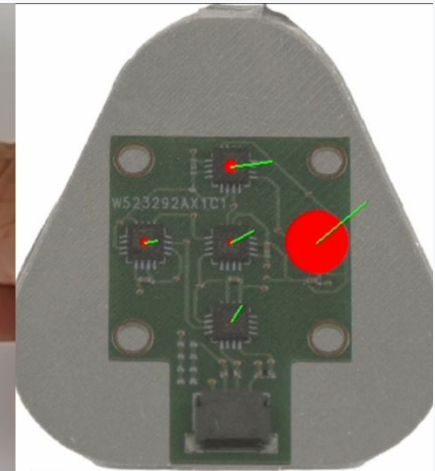
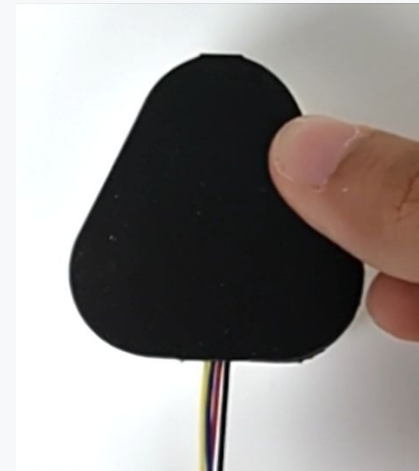
W. Yuan, et al. "GelSight: High-Resolution Robot Tactile Sensors for Estimating Geometry and Force". In Sensors, 2017.



B. Huang et al, 3D-VTac, CoRL 2024

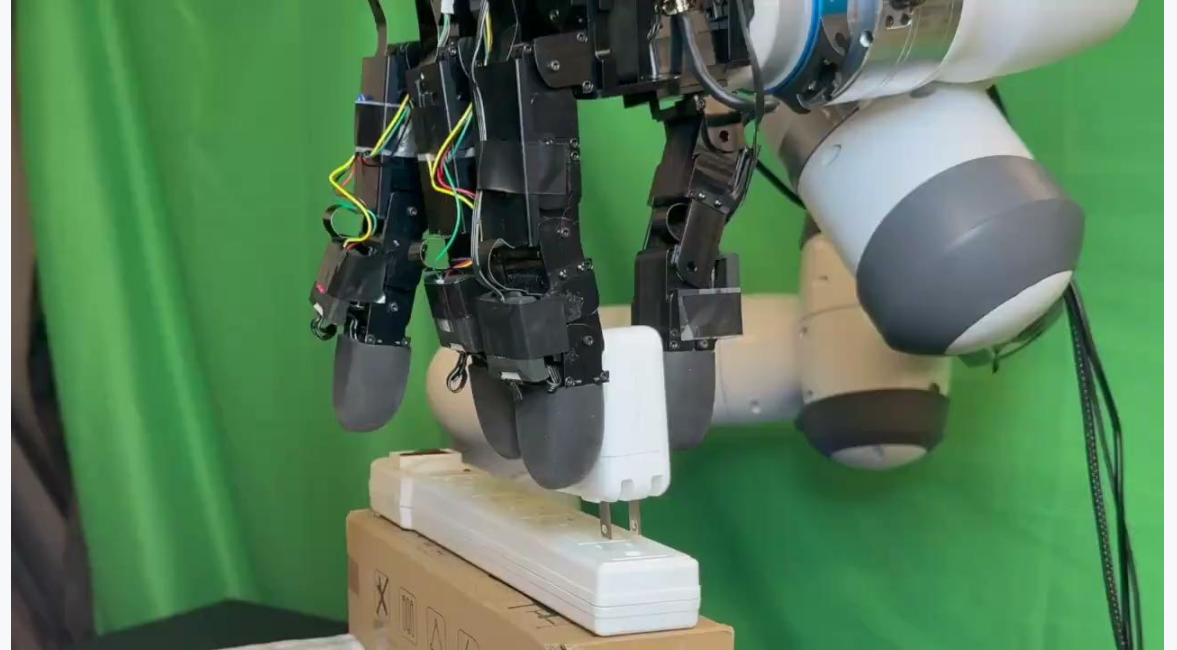


Capacitive Tactile Sensor



Bhirangi et al, AnySkin, ICRA 2025

With v.s Without Tactile



W, Wi,et al. TactAlign. In RSS, 2026.

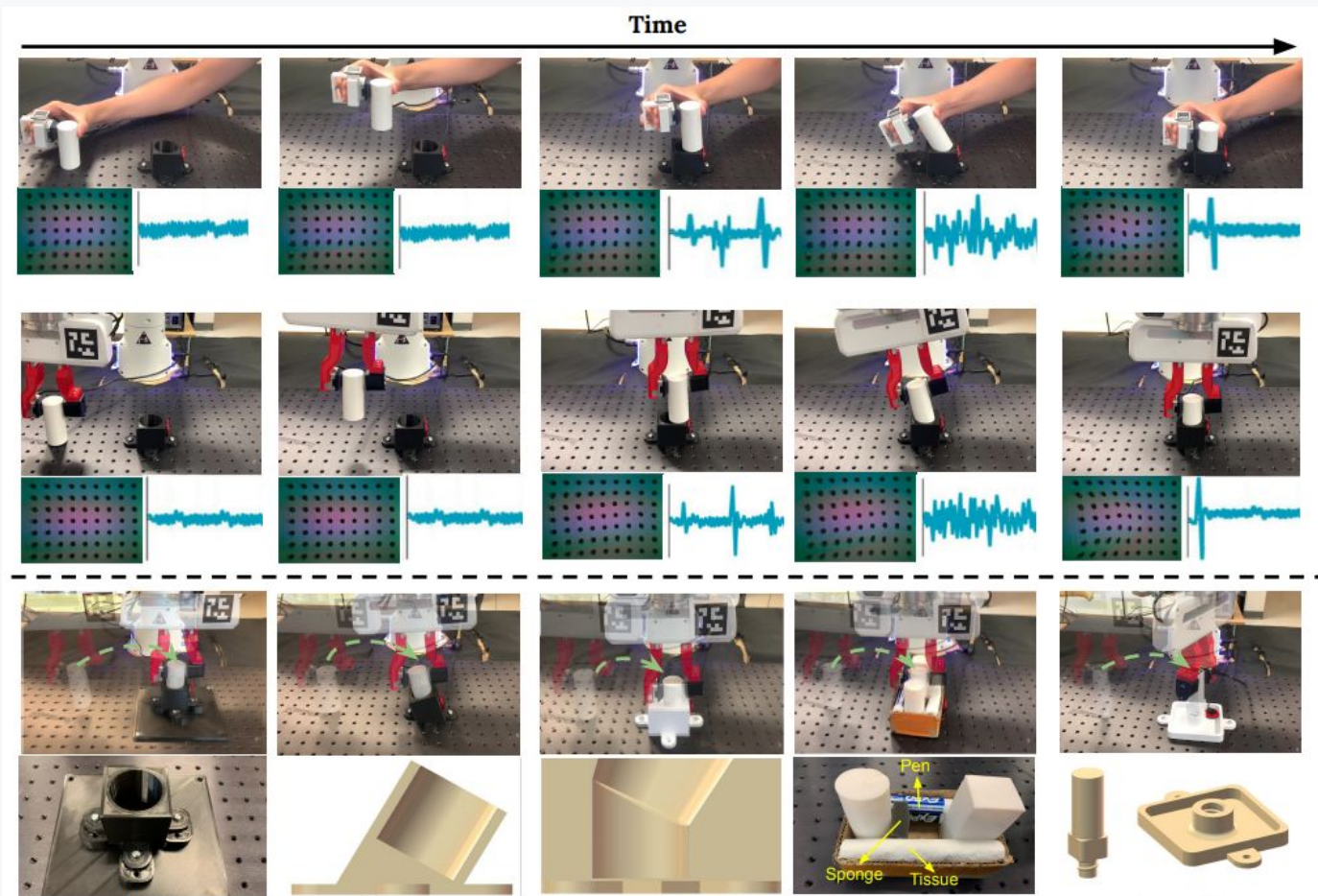
Learning from Human Tactile Insights

How about learning multi-modal policy from human?

Can embodied sensorimotor experiences helps?

Out of scaling up, are human experience helpful for the performance of contact-rich manipulation tasks?

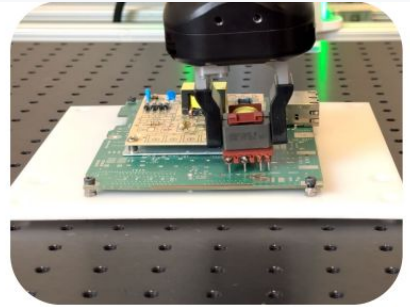
MimicTouch: Leveraging Multi-modal Human Tactile Demonstrations for Contact-rich Manipulation



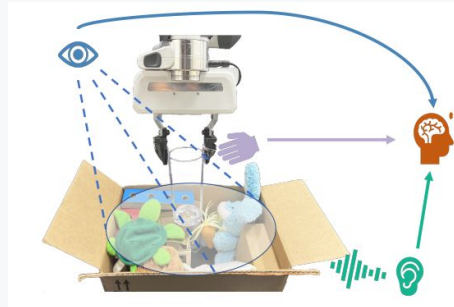
Kelin Yu*, Yunhai Han*, Qixian Wang, Vaibhav Saxena, Danfei Xu, Ye Zhao,
MimicTouch: Leveraging Multi-modal Human Tactile Demonstrations for Contact-rich Manipulation
CoRL 2024, **Best Paper Award** at NeurIPS TouchProcessing Workshop

First contribution: Why tactile and audio feedback only for insertion tasks

Current learning-based methods for insertion tasks assume



Luo. J , et al., ICRA 2024



Hao. L , et al., CoRL 2022

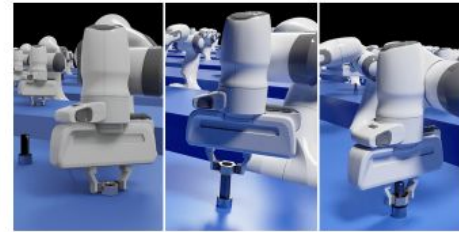
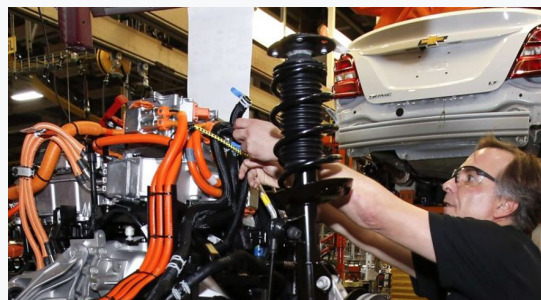


Fig. 8. Rendering of achieved goal states of our trained subpolicies for FrankaNutBoltEnv. Left: Pick. Middle: Place. Right: Screw.

Yashraj. N , et al., RSS 2022

Good lighting conditions & Open space

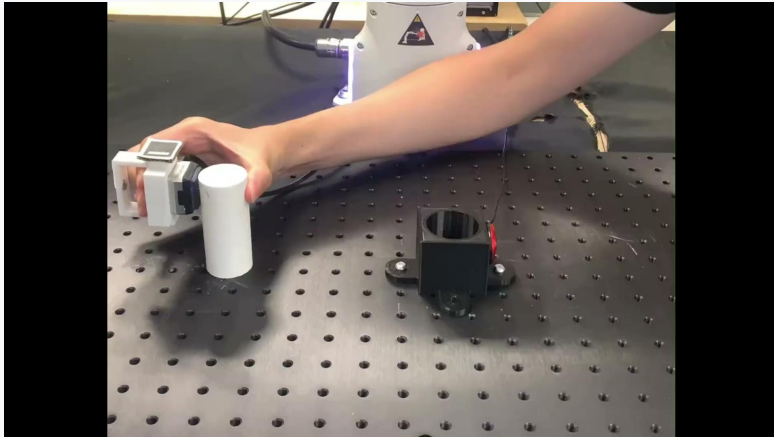
What are the common industrial setting (car assembly)



Poor lighting conditions & Constrained Space
Visual policy fails & even no place to put cameras!

Second contribution: Why Human Play Data?

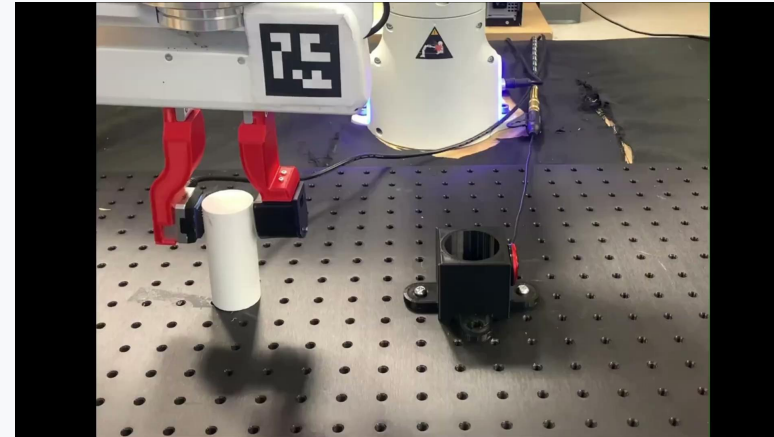
Human Play Data



2X

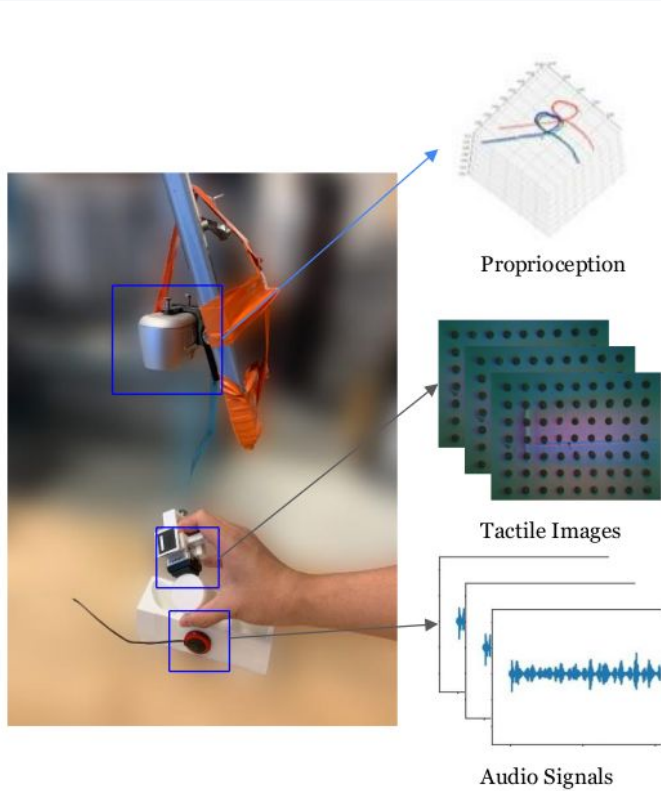
VS

Teleoperation



- Utilize **Human** Tactile Feedback
 - Sense contact dynamics during data collection
- 10 times **faster**
 - Human execution speed
- **Diverse** motions
 - Smooth human motions
 - No constraints on action space (6D)
- **Easy** for data collection
 - No expertise required

Human Tactile Demonstrations



- Retarget poses of human finger tips to the **robot proprioception**.
- **Tactile images** reflect the change of the object poses with different contacts
- **Audio Signals** reflect the vibrations with slippery and collisions between contacts.

Multi-modal Representation Learning

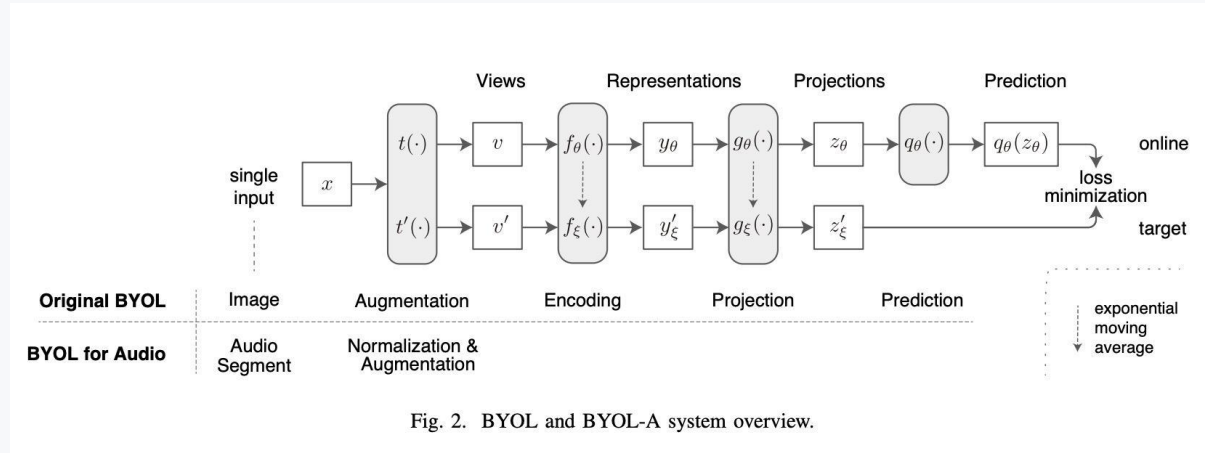
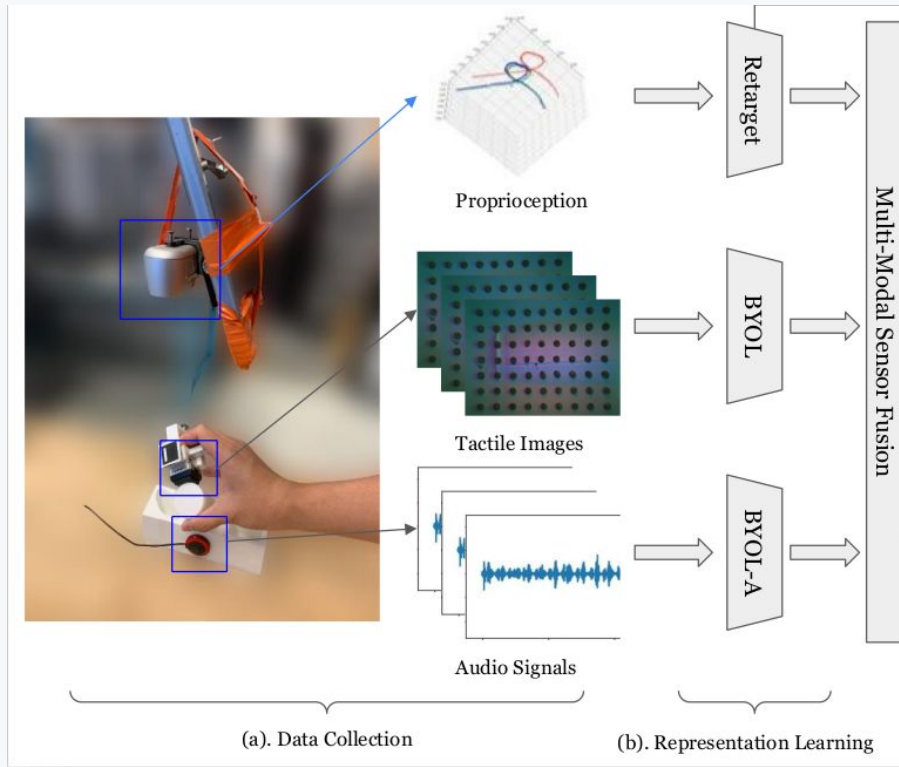


Fig. 2. BYOL and BYOL-A system overview.

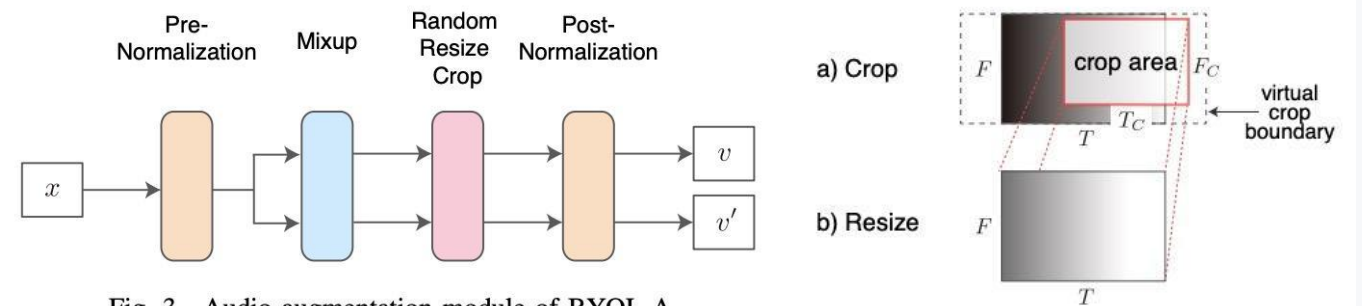
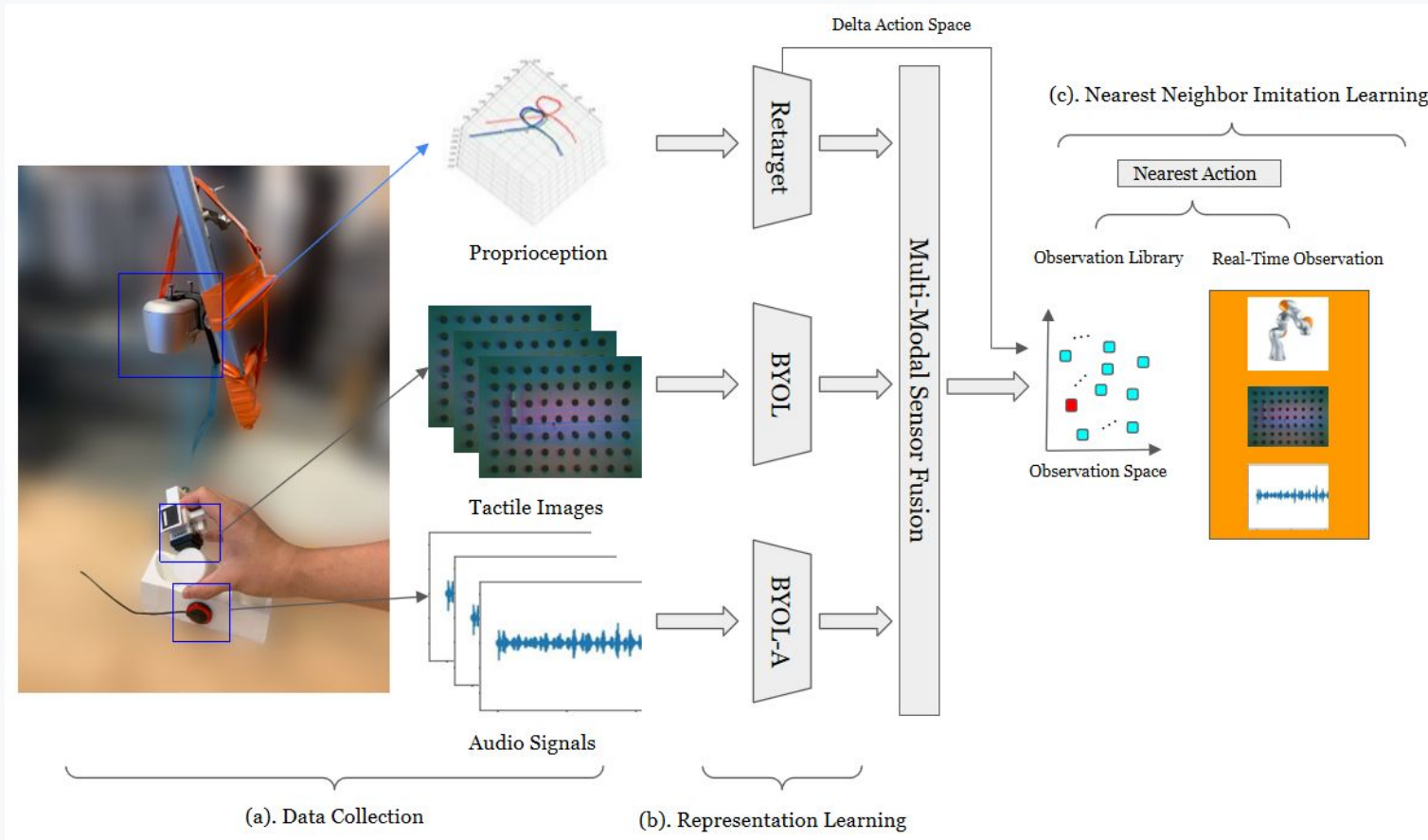


Fig. 3. Audio augmentation module of BYOL-A.

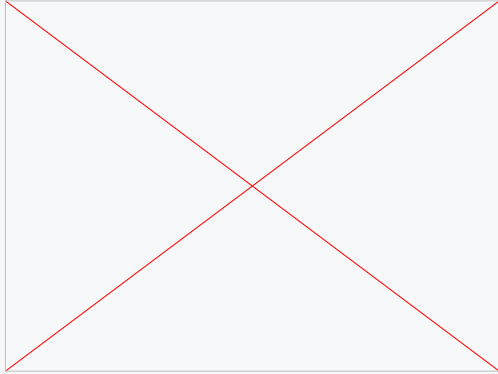
VINN-based Imitations



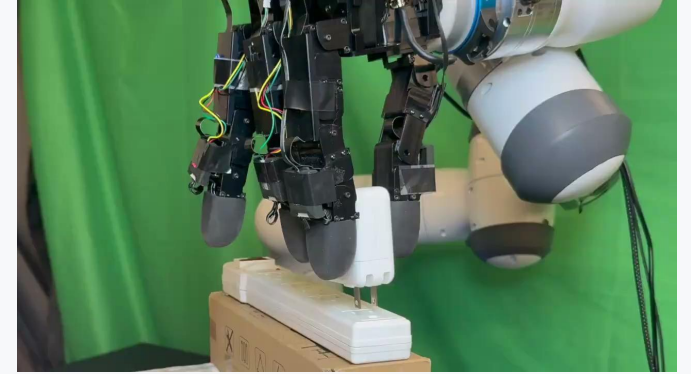
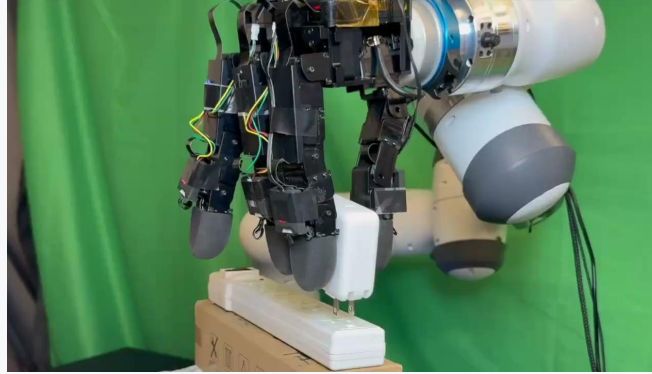
Why VINN [1]?

- **Less corivance shift**, can reduce dangerous behaviors caused by embodiment gaps and 6-dimensional actions space, to ensure the **safety** for human-to-robot transferring learning.
- Can be used for demos with **sparse actions**. (Since human's motions are much faster)

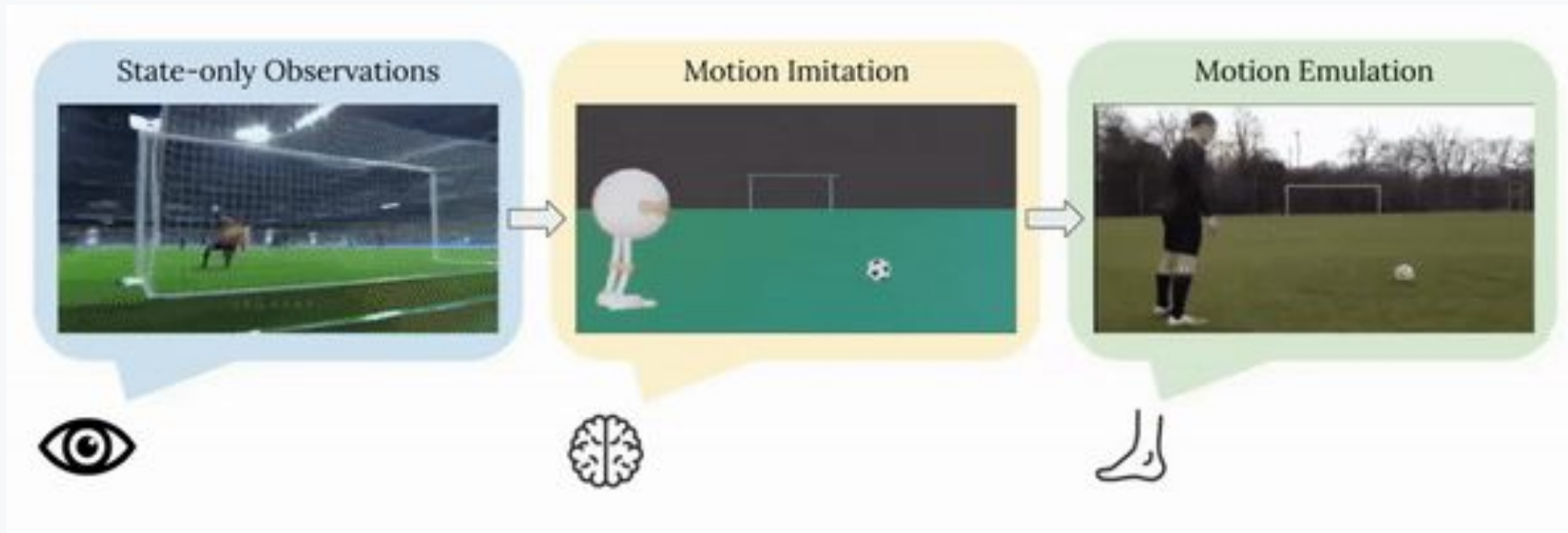
Online Residual Reinforcement Learning



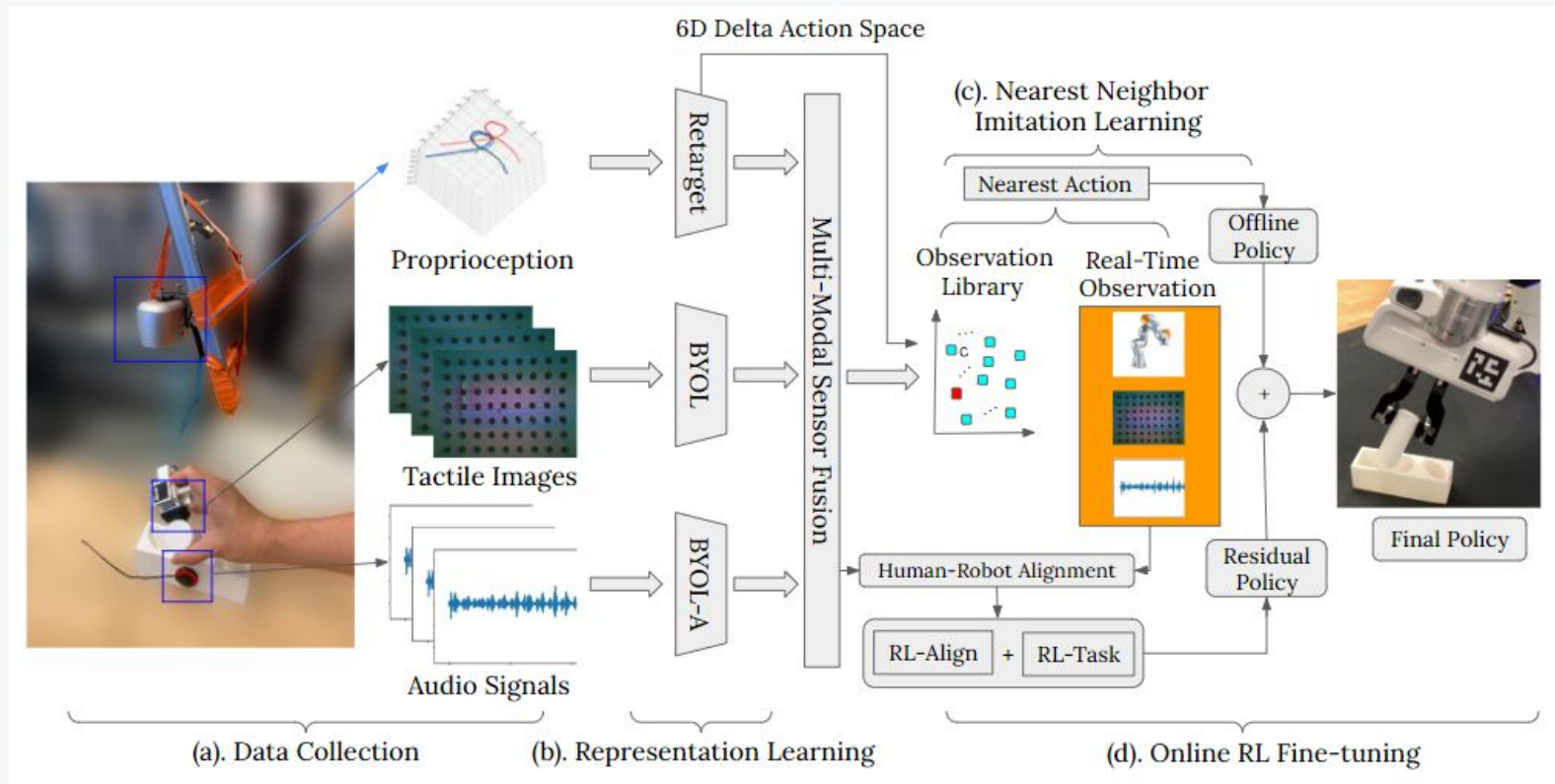
Motion Gaps



Dynamic Gaps



Online Residual Reinforcement Learning



- Using Real World Residual RL to solve the **embodiment gaps** between human demonstrations and robot executions. (motion difference)
- Finetuning the offline policy through interactive learning through **dense local contact** (inferior reasoning of the contacts)

Human tactile Data Collection

Two metrics:

- The number of usable demonstrations collected per hour.
- The success rate for collecting usable demonstrations.

How to define usable?

- The robot successfully inserts the object into the hole without any slipping or falling.
- The robot completes the task within 100 actions.

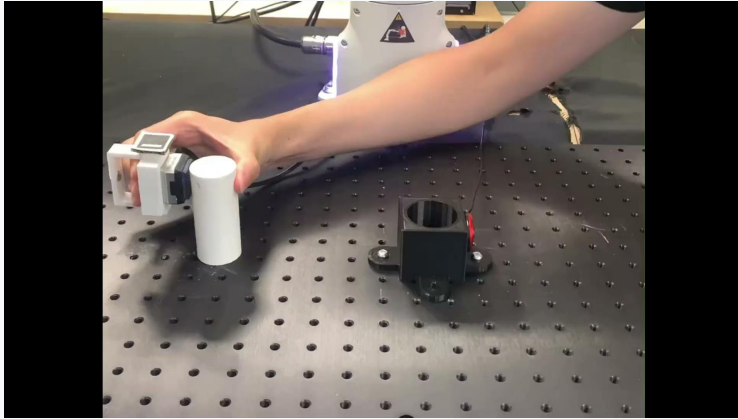
Methods	Frequency	Success Rate
Spacemouse Teleoperation	19 traj/hr	38.5% (20/52)
Hand Teleoperation	44 traj/hr	58.8% (20/34)
Human Tactile Demonstrations	104 traj/hr	83.3% (20/24)

Table 1: Data collection throughput for Human Tactile Demonstrations and Teleoperation systems.

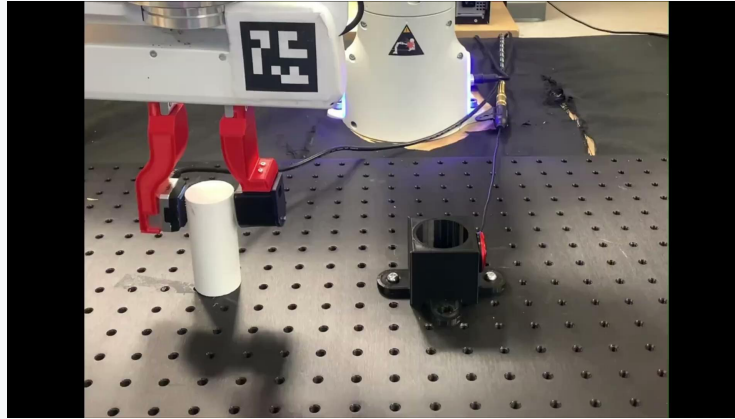
Human tactile Data Collection

2X

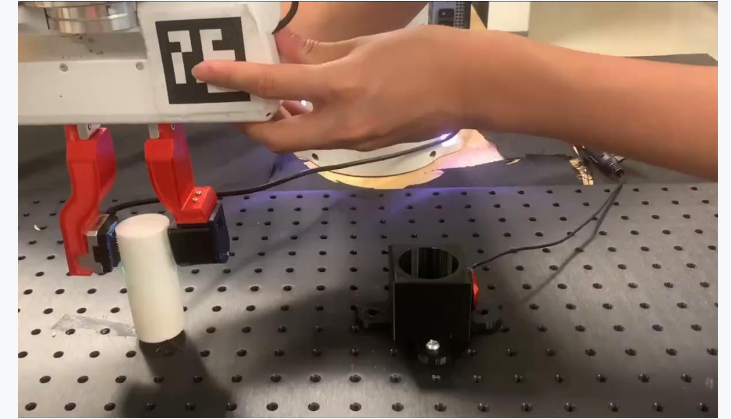
Human Play Data



Teleoperation



Hand Operation



- Human tactile demonstrations can be collected significantly more efficiently than teleoperation systems for contact-rich tasks.
- Human tactile demonstrations can seamlessly integrate human's tactile feedback and motion capability, whereas teleoperation systems struggle to capture such dynamic tactile-guided motions.

Online RL Training

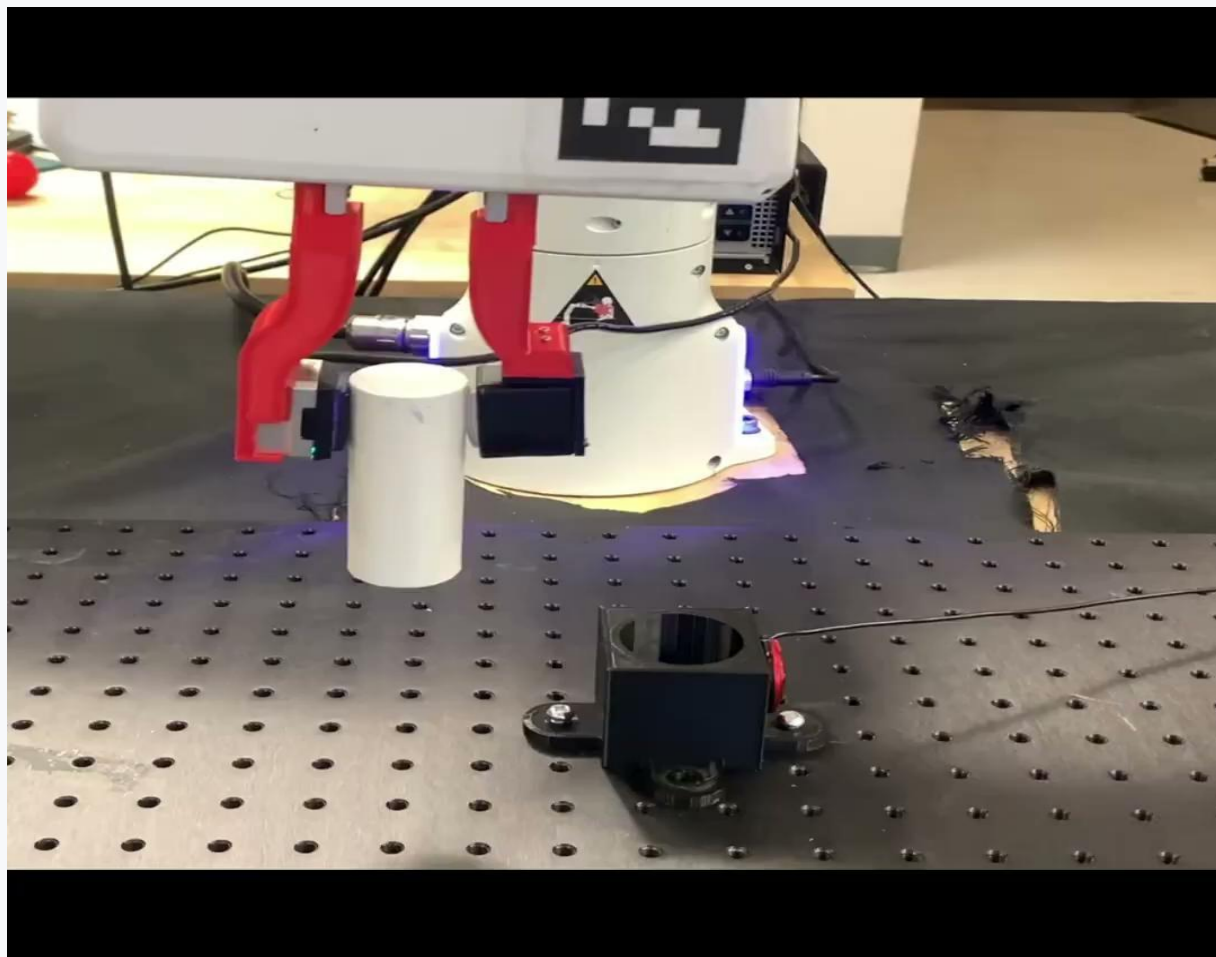
Real-world RL Training for **Three hours**

Success Rate (During Training)

- After each hour, we compute the task success rates over 25 policy rollouts, in alignment with the offline policy evaluation.
- For other time instances, we only compute the task success rates over 10 policy rollouts to minimize sensor wear

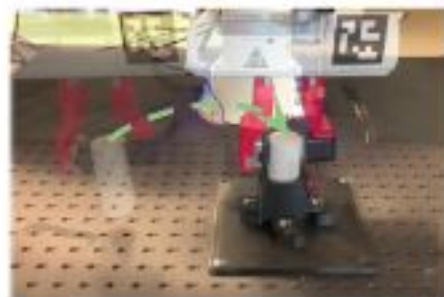


Online RL Training



4X

Zero-shot Generalization



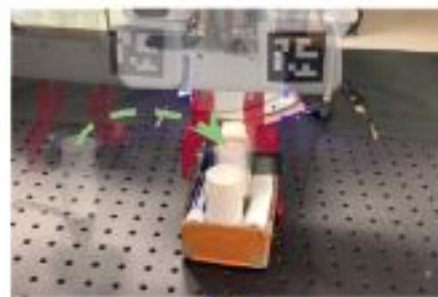
Shifting



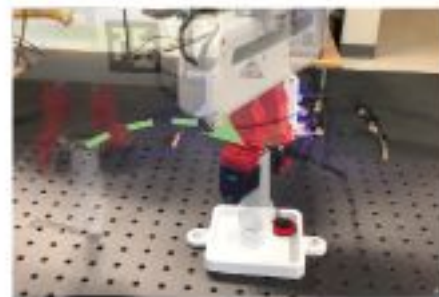
Tilting



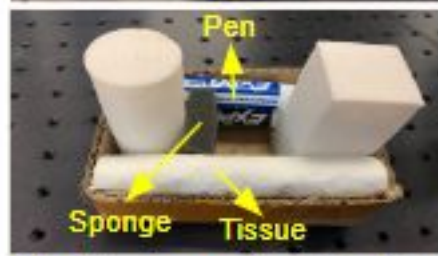
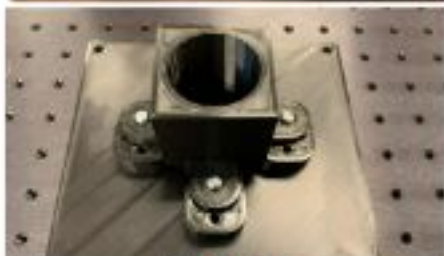
Two-stage Packing



Multi-material Packing



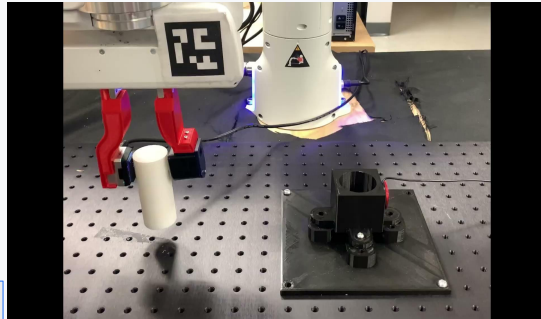
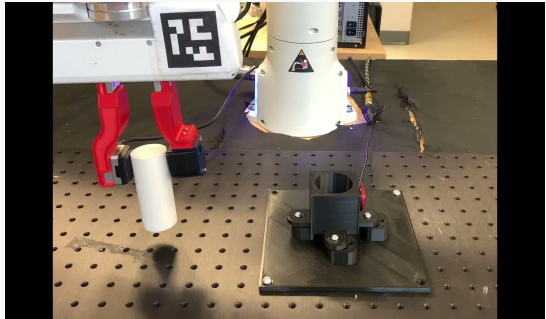
Assembly



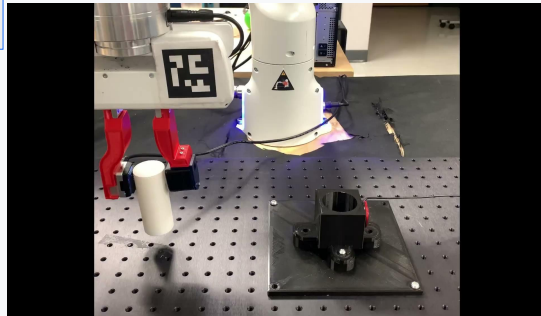
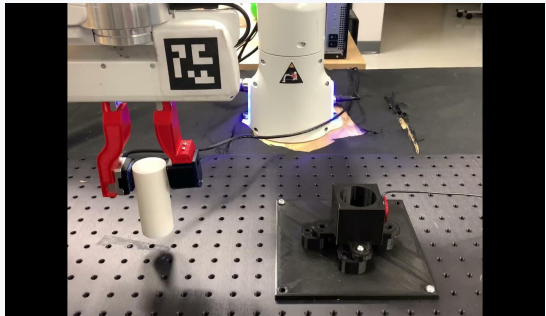
Policy	Shift	10°	20°	Two-stage	Rigid	Soft	Assem (I)	Assem (A)	Assem
<i>Openloop</i>	60%	56%	40%	52%	52%	36%	32%	37.5%	12%
MimicTouch	92.5%	92%	80%	88%	80%	64%	76%	68.4%	52%

Zero-shot Generalization

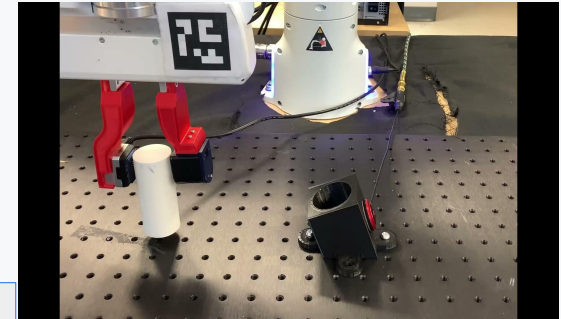
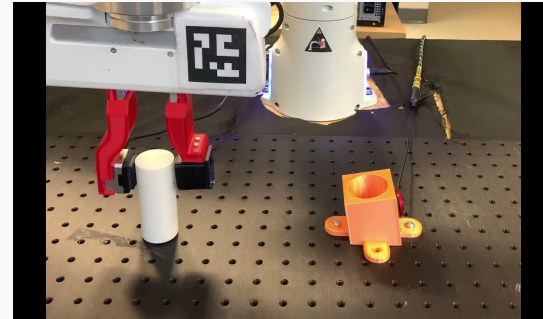
Shift 0.8cm



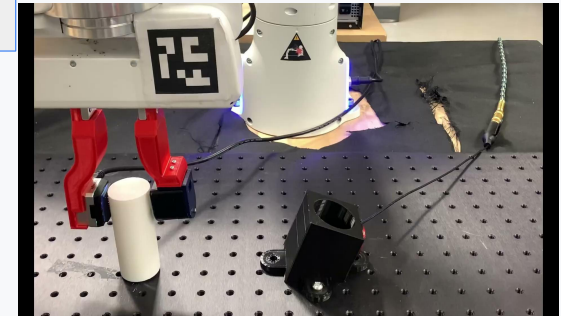
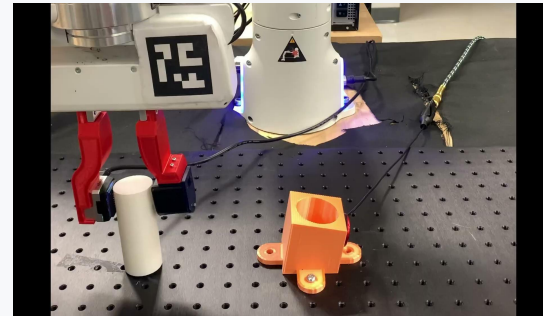
2X



Tilt 10-20 Degrees



2X

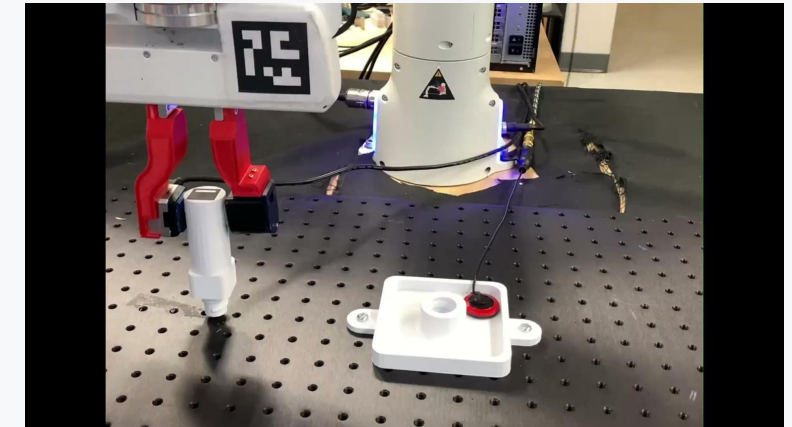
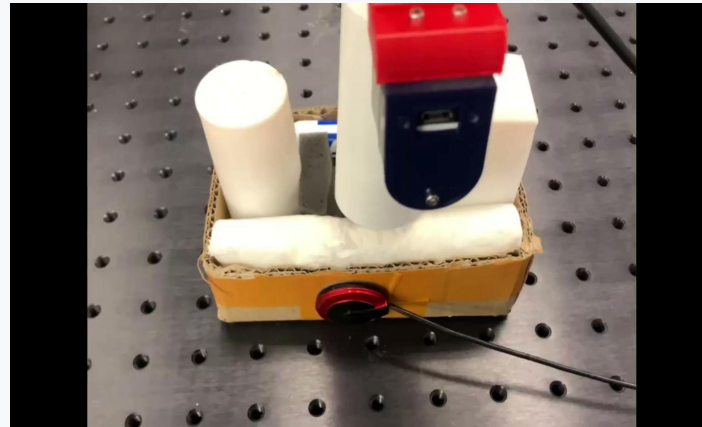
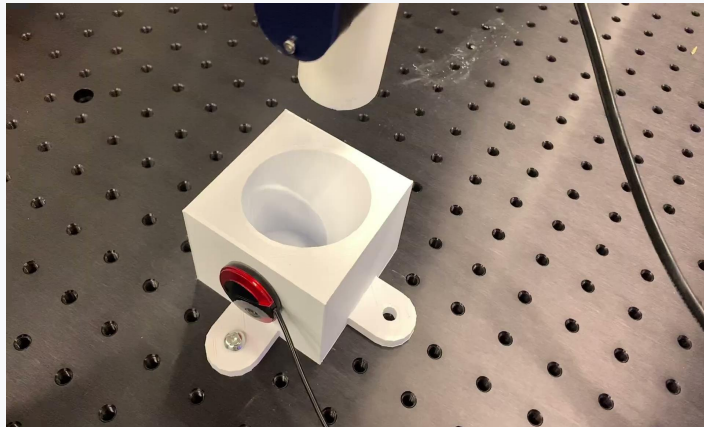
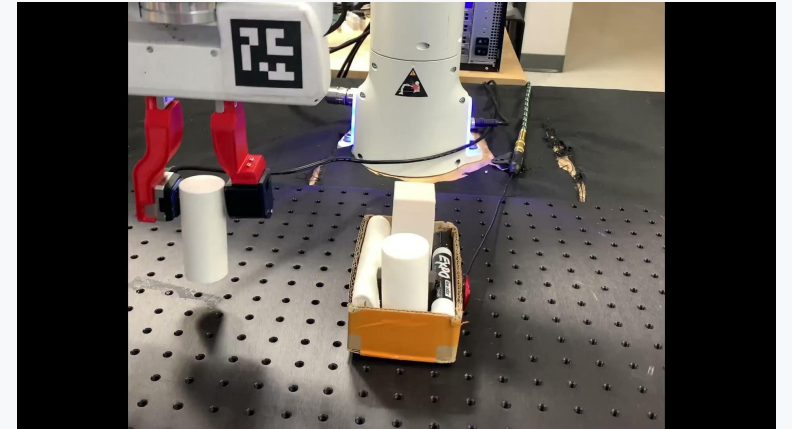
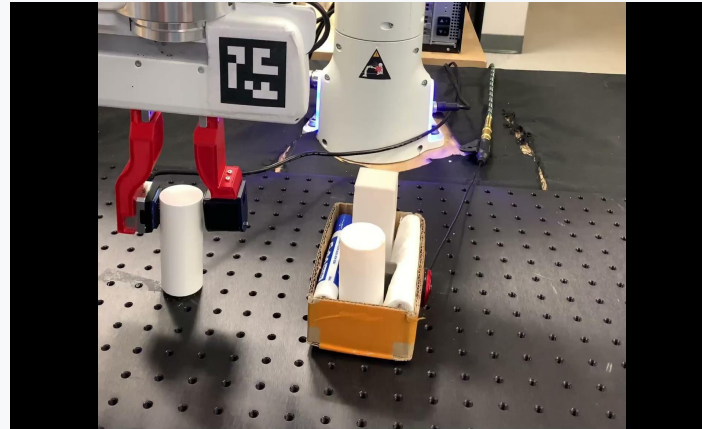
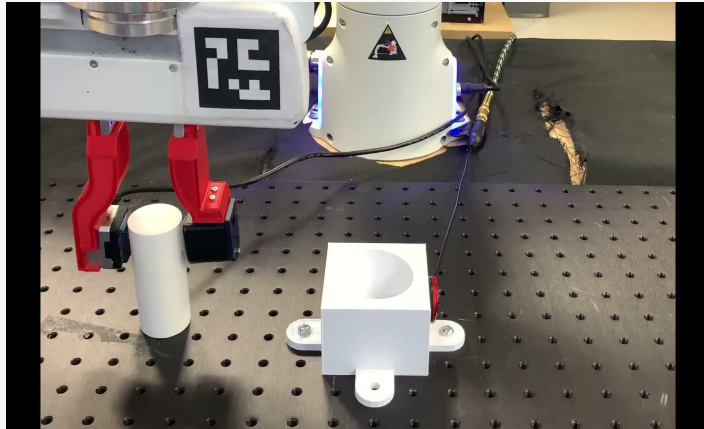


10 Degrees

20 Degrees

Zero-shot Generalization

Multi-material Dense Packing (Pen)



Two-stage Dense Packing

Multi-material Dense Packing (Soft)

Furniture Assembly

2X

Limitations

- No Visual Feedbacks
- Need to embed Tactile Sensors
- Cannot be adapt to more general policies
- No Reward Shaping

Watching Human Demos is not enough for Contact-rich Manipulations

- Human needs to do practice after watching expert demos
- Using visual policy as motion prior and tactile for practicing and residual correction



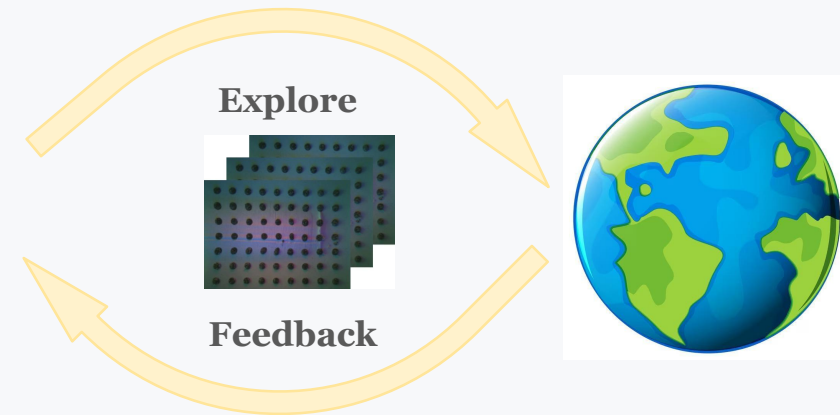
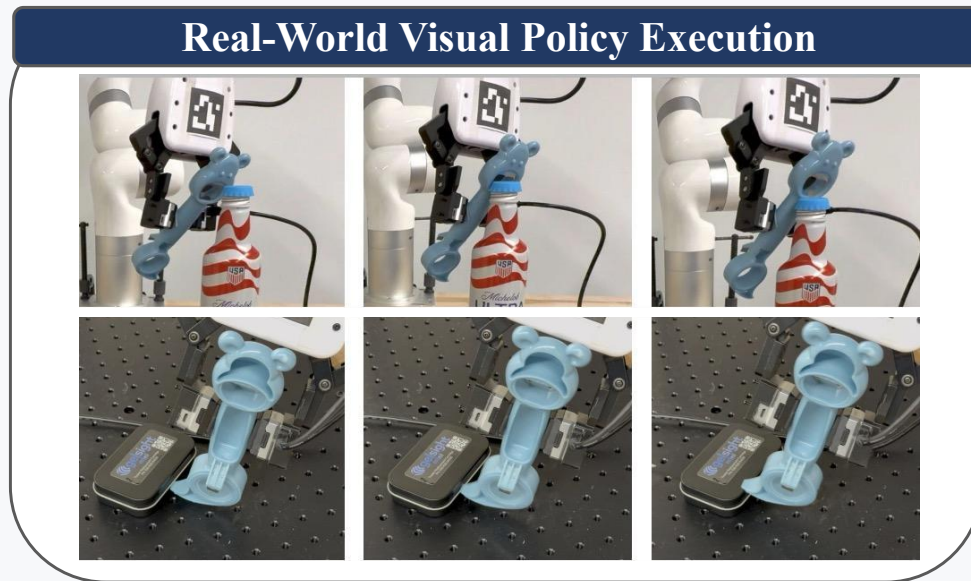
Expert Demos



Practice

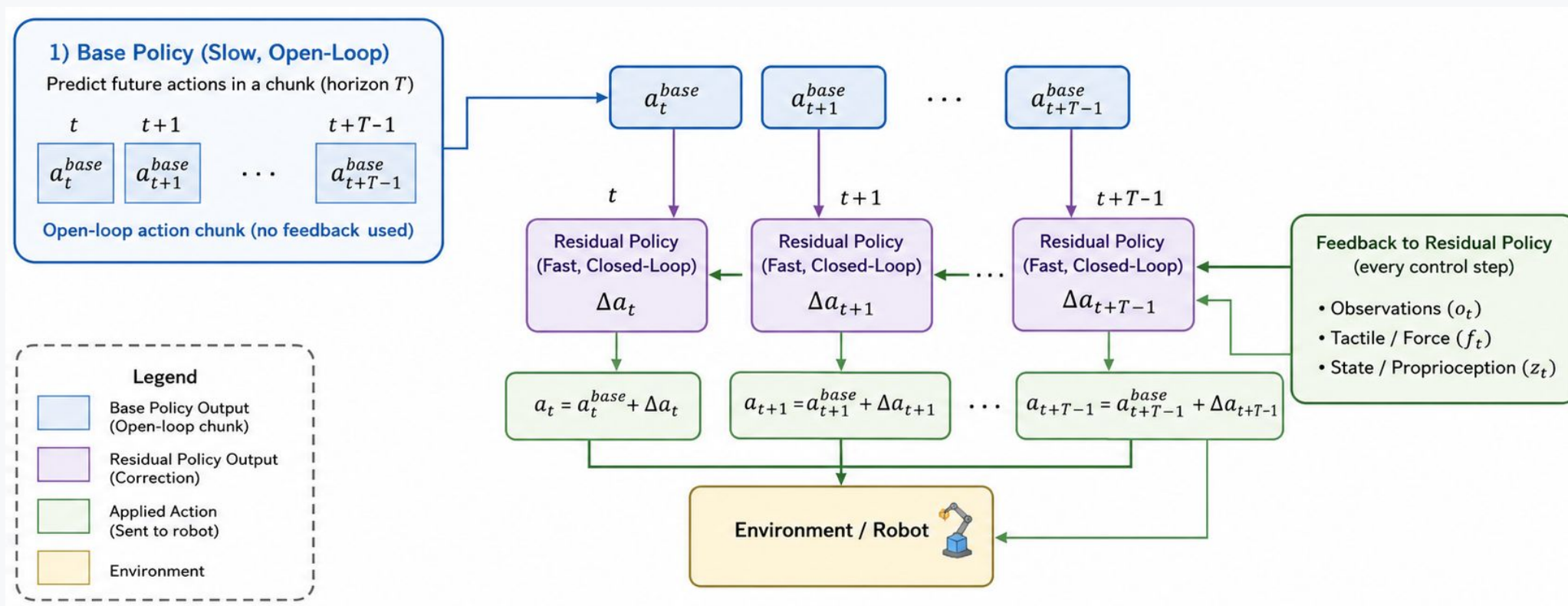
Watching Human Demos is not enough for Contact-rich Manipulations

- A new multisensory learning pipeline, bridging the gap between visual data and tactile data



Watching Human Demos is not enough for Contact-rich Manipulations

- A new multisensory learning pipeline, bridging the gap between visual data and tactile data
- Using visual policy as motion prior and tactile for practicing and residual correction



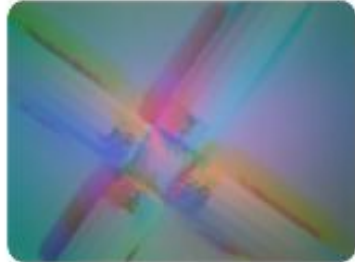
Watching Human Demos is not enough for Contact-rich Manipulations

- A new multisensory learning pipeline, bridging the gap between visual data and tactile data
- Using visual policy as motion prior and tactile for practicing and residual correction
 - Why not choosing simulator?

G.T.



Simulator



Large Sim2Real Gap for Visuo-Tactile

Box Opening



3X

Hard to simulate all objects and dynamics

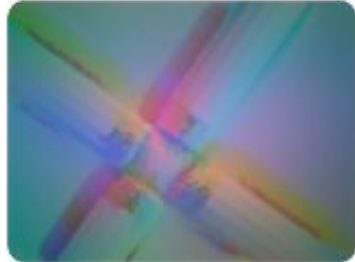
Watching Human Demos is not enough for Contact-rich Manipulations

- A new multisensory learning pipeline, bridging the gap between visual data and tactile data
- Using visual policy as motion prior and tactile for practicing and residual correction
 - Why not choosing simulator?
- One take: Real-World RL is the final step for the last millimeter in all directions

G.T.



Simulator



Large Sim2Real Gap for Visuo-Tactile

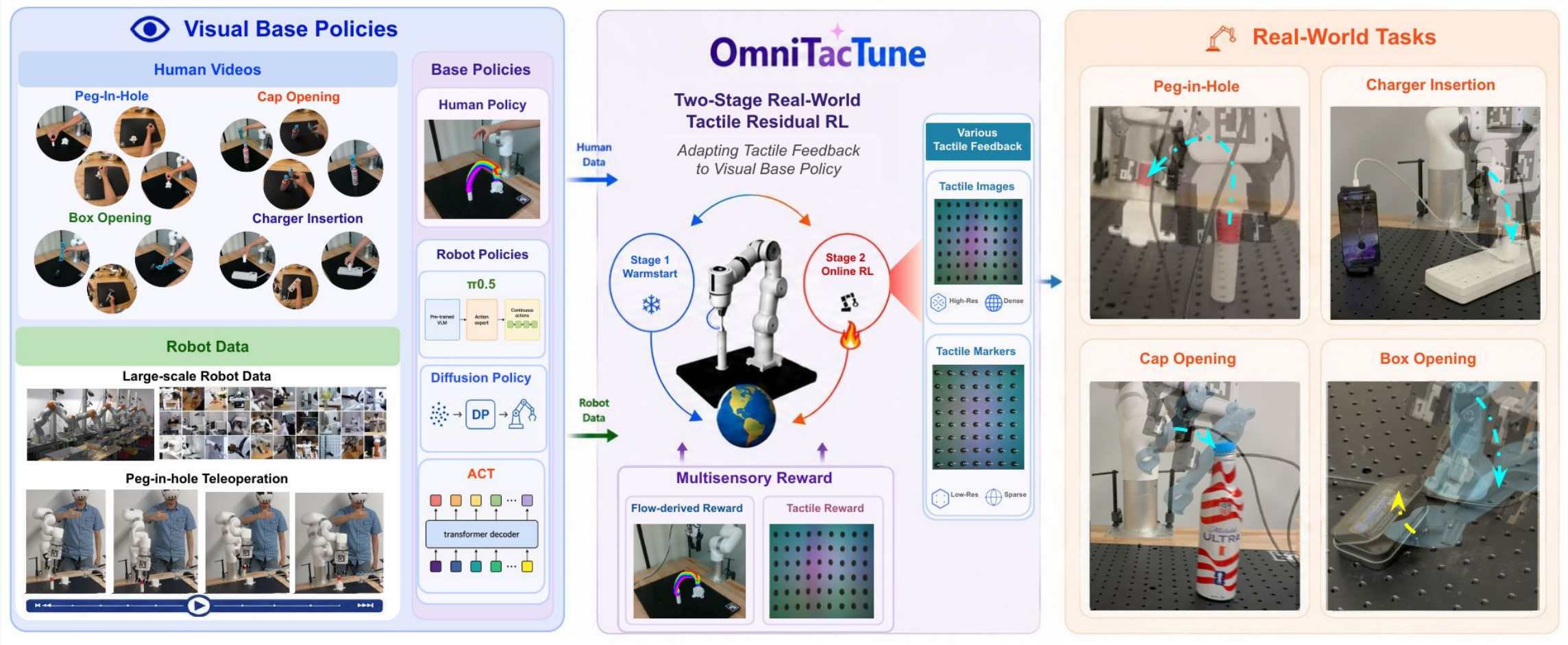
Box Opening



3X

Hard to simulate all objects and dynamics

OmniTacTune : Policy-Agnostic Real-World RL for Tactile Adaptation

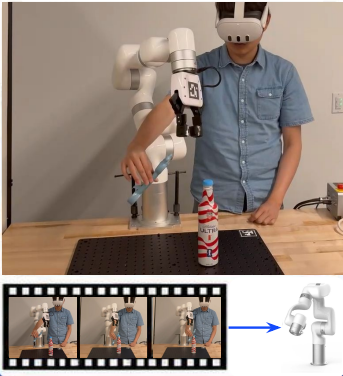


OmniTacTune : Policy-Agnostic Real-World RL for Tactile Adaptation

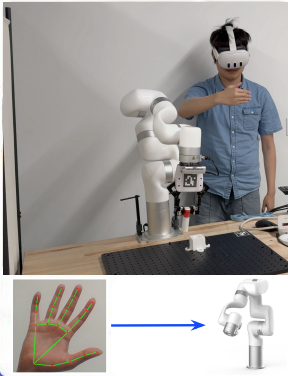
1. Data Collection and Base Policies

(a) Data Collection Sources

Human Video + Retargeting

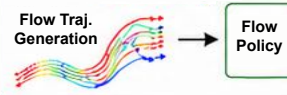


VR Teleoperation



(b) Base Policies

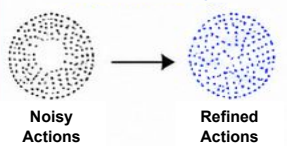
Human Policy



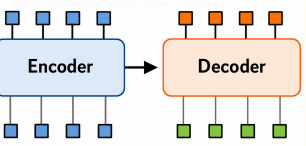
$\pi_{0.5}$



Diffusion Policy



ACT

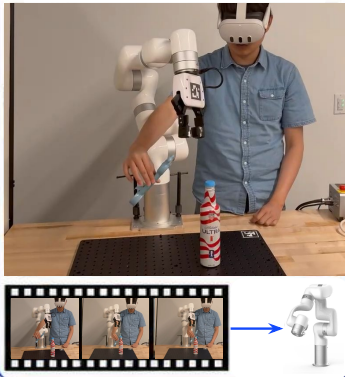


OmniTacTune : Policy-Agnostic Real-World RL for Tactile Adaptation

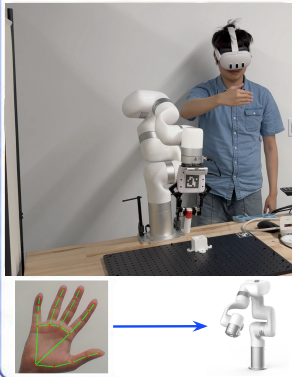
1. Data Collection and Base Policies

(a) Data Collection Sources

Human Video + Retargeting

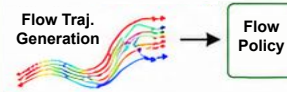


VR Teleoperation



(b) Base Policies

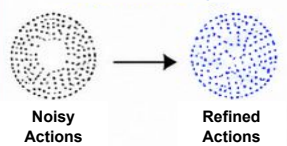
Human Policy



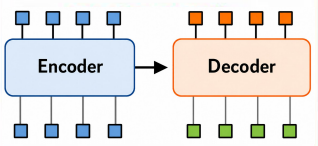
$\pi_{0.5}$



Diffusion Policy

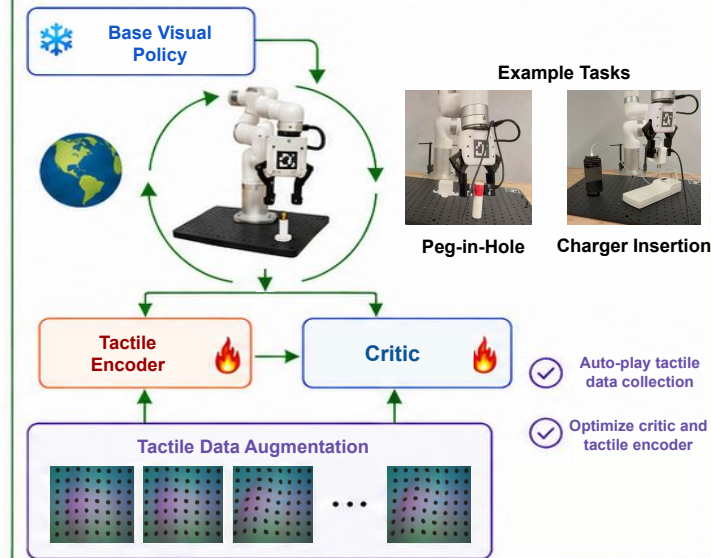


ACT

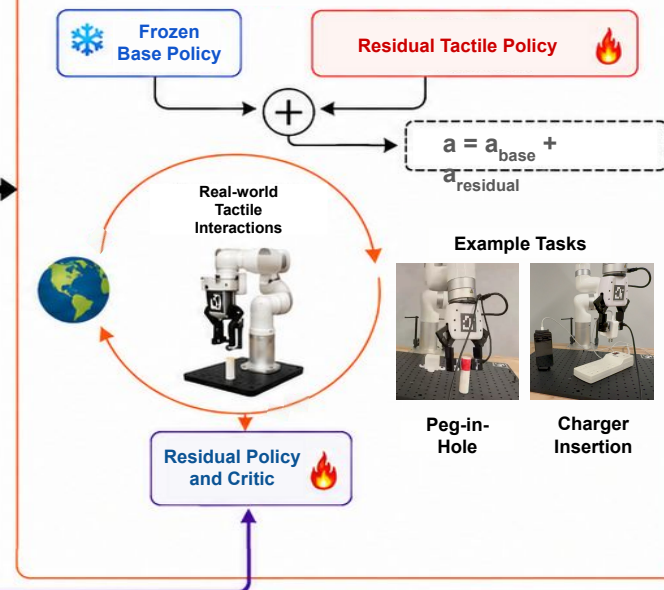


2. Two-stage Real-world Tactile Residual RL

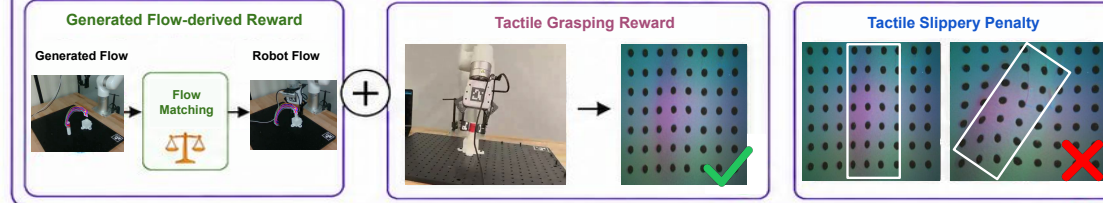
Stage 1: Warmstart



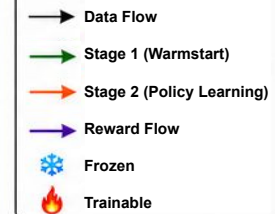
Stage 2: Policy Learning



Dense Reward Shaping Pipeline



Legend



Policy-Agnostic Flow Interface

How Generated Flow Makes Residual RL Policy-Agnostic

Any Frozen Visual Policy

Human-flow policy, ACT, DP, or $\pi_{0.5}$
outputs a base action chunk



Shared Object-Centric Interface

generated keypoint goals + base-policy action chunk
plus tactile feedback, no hidden states



Flow-Tactile Residual Policy

predicts bounded contact correction
 Δa_{res}
only when correction is needed

Policy-Independent Reward

reaching + grasp/contact + flow matching + tactile stability – safety penalty

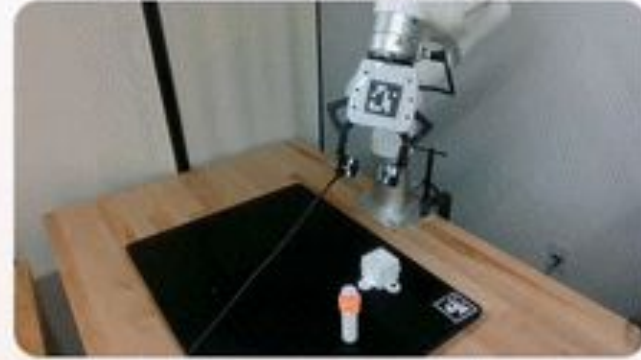


Real-World Residual RL

updates residual policy, critic, and tactile encoder
while keeping the base policy frozen

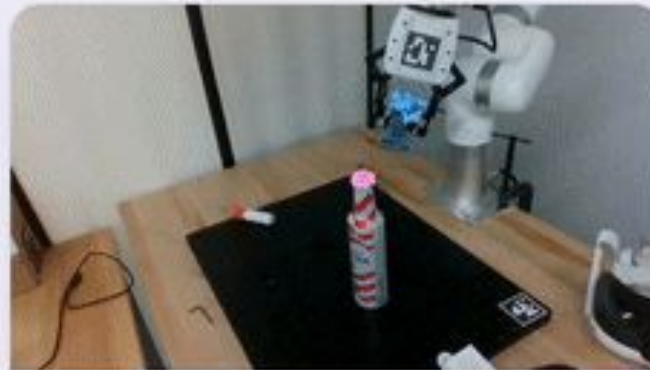
Policy-Agnostic Flow Interface

PEG-IN-HOLE Spatial Generalization



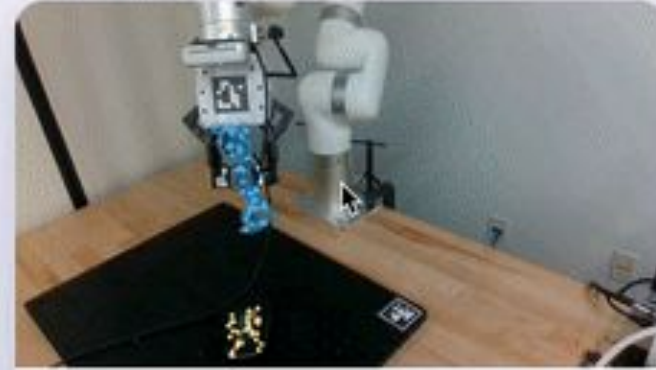
Charger Insertion

Guidance + reward shaping



Cap Opening

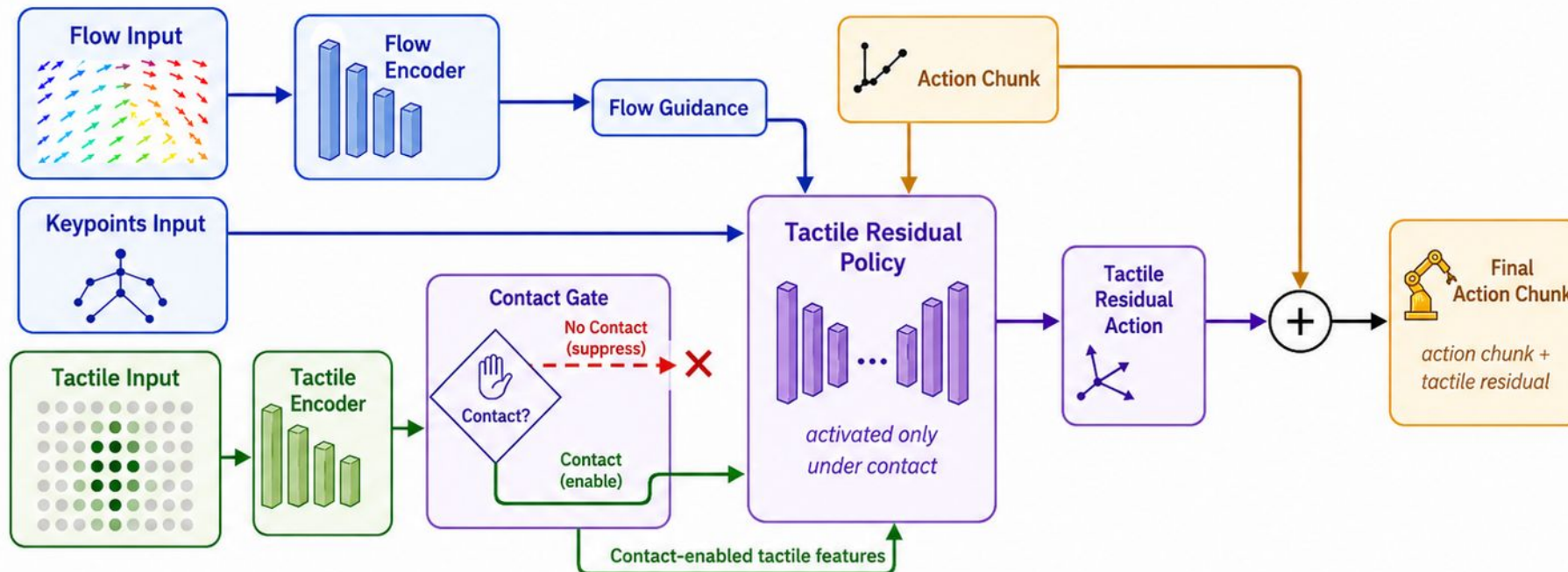
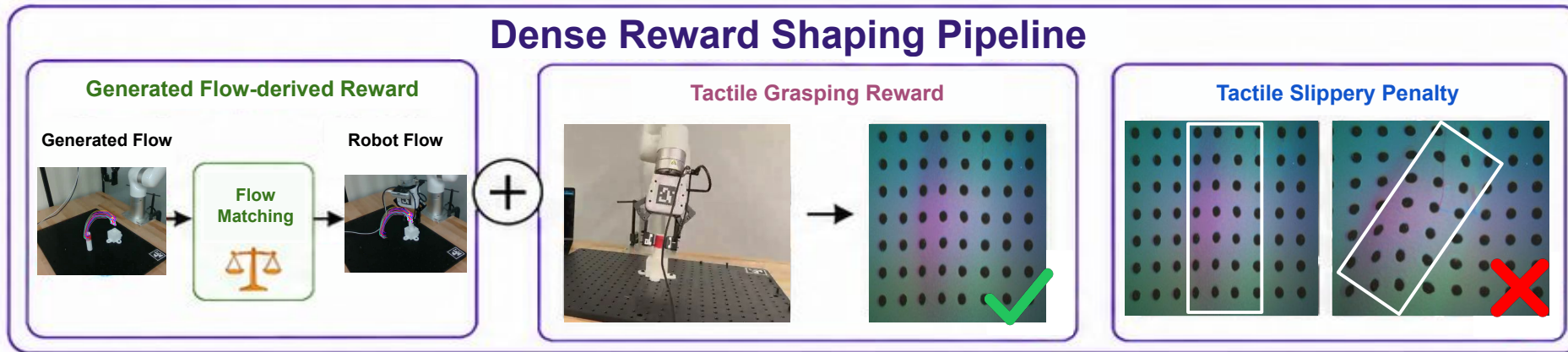
Guidance + reward shaping



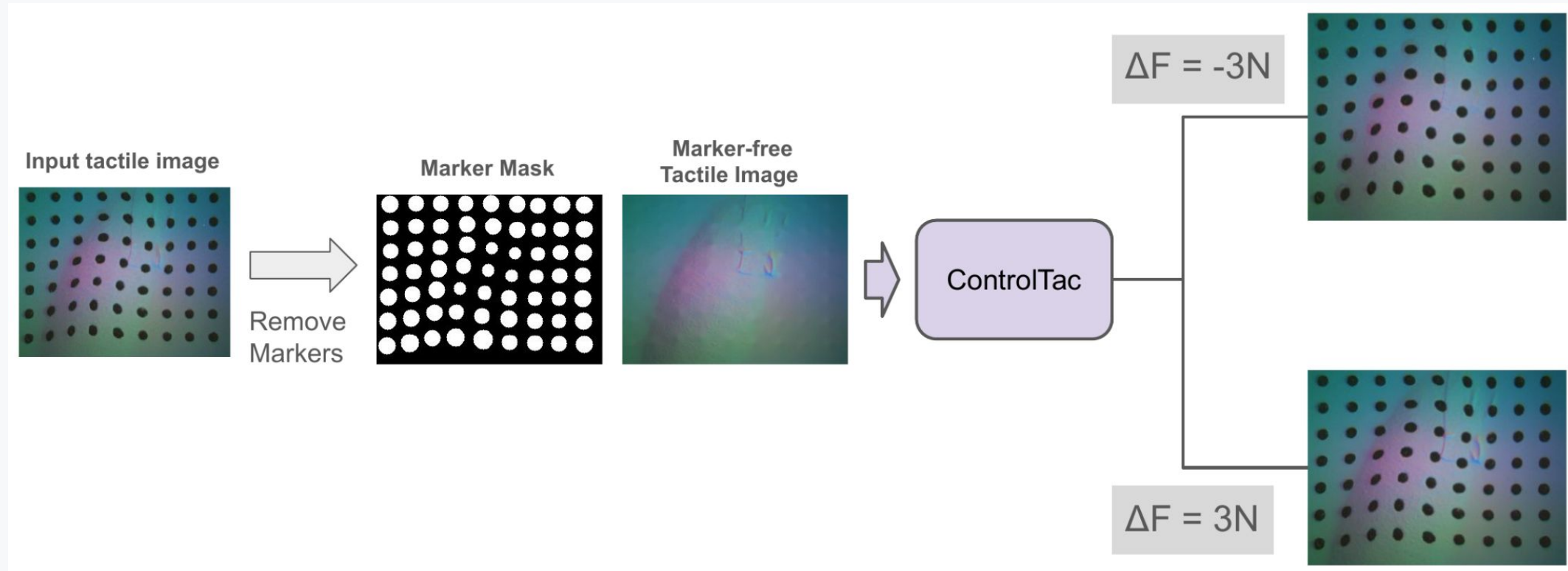
Box Opening

Guidance + reward shaping

MultiSensory Reward Shaping and Policy Design



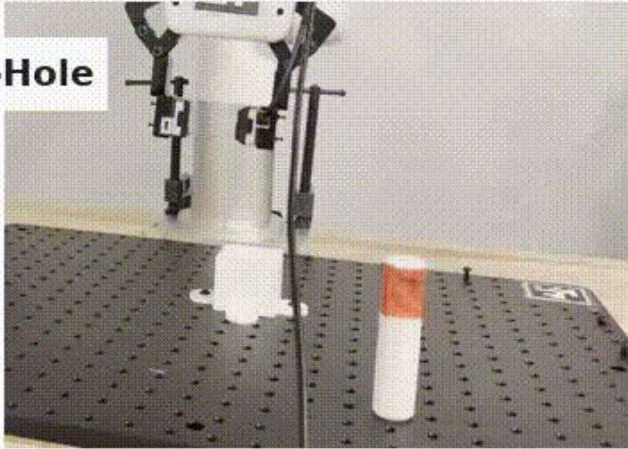
Tactile Data Generation for Sample Efficiency and Bootstrapping



ControlTac trajectory-level augmentation. A real tactile image is converted to a marker-free reference, perturbed to new contact forces (e.g. $\Delta F = -3N$ and $\Delta F = +3N$) via ControlTac, and recomposed with the original marker mask — producing **physically plausible tactile variants** for tactile-aware bootstrapping.

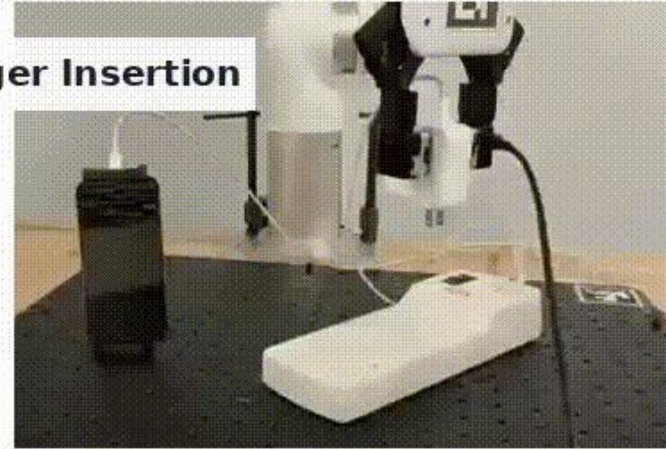
Contact-rich Manipulation Tasks

Peg-in-Hole



3X

Charger Insertion



Cap Opening

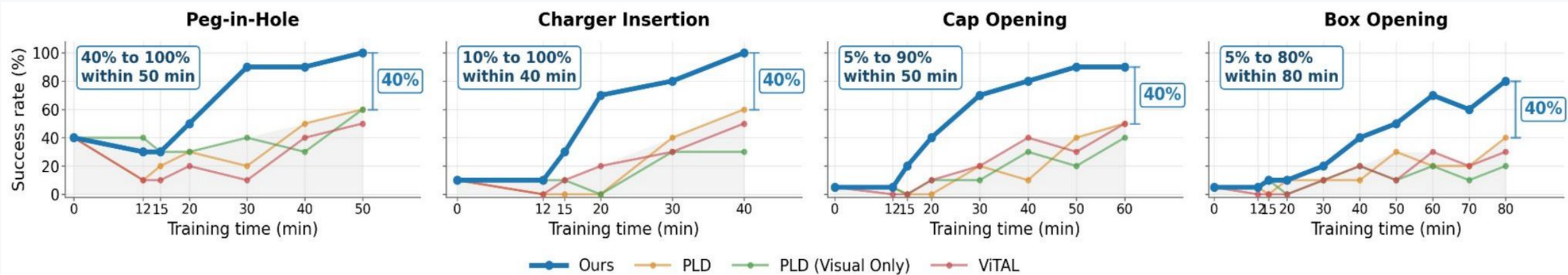


3X

Box Opening



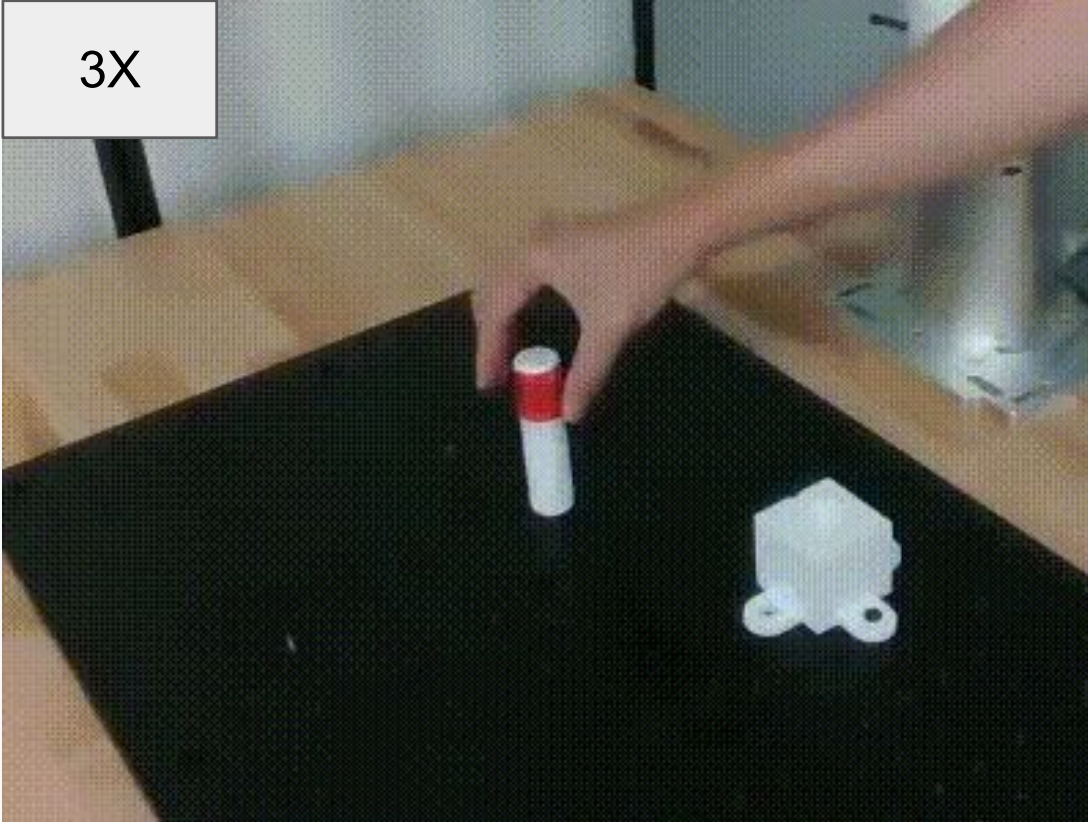
Comparing with Baselines



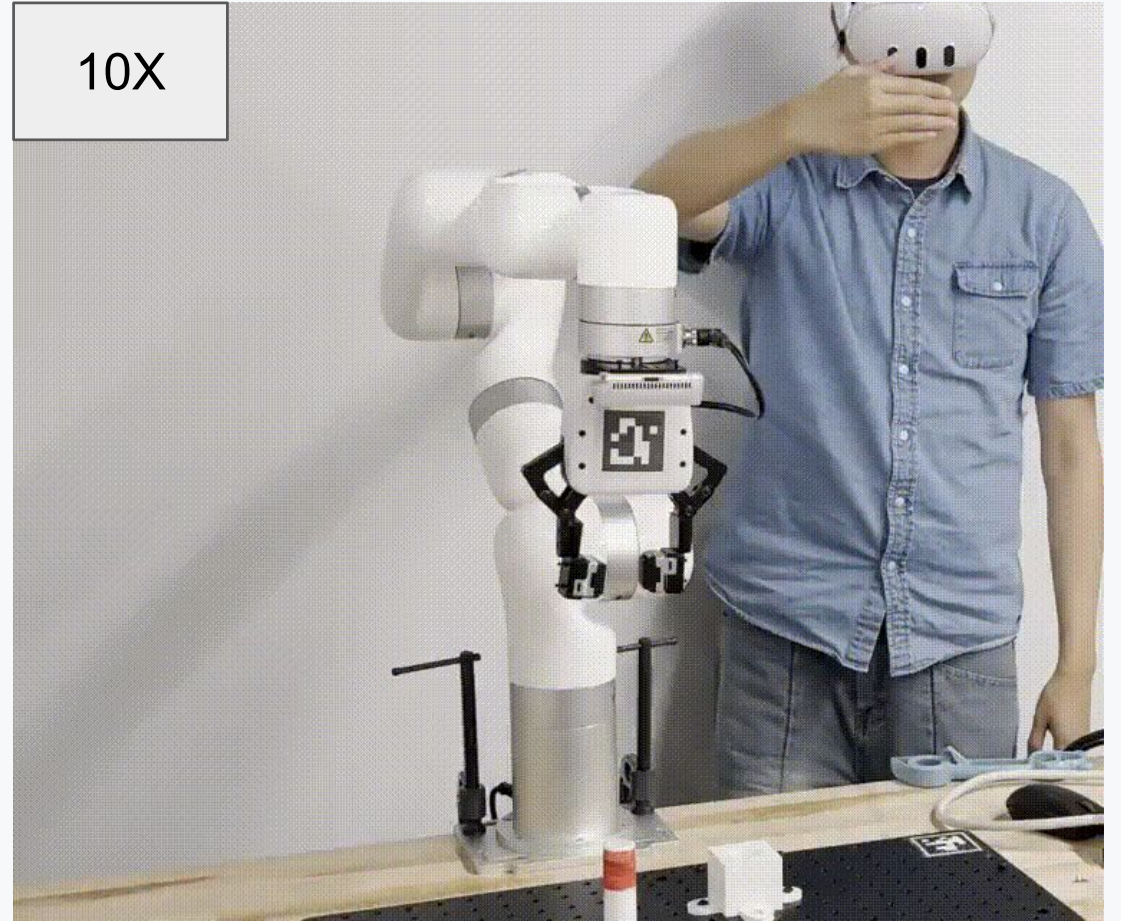
Method	Peg-in-Hole	Charger Insertion	Cap Opening	Box Opening	Average
PLD* [22]	65%	60%	50%	35%	52.5%
PLD [22] (Visual Only)	60%	30%	40%	20%	37.5%
ViTAL [63]	50%	50%	45%	30%	43.75%
Ours	100%	100%	90%	85%	93.75%

General Pipeline: Across different Data Source

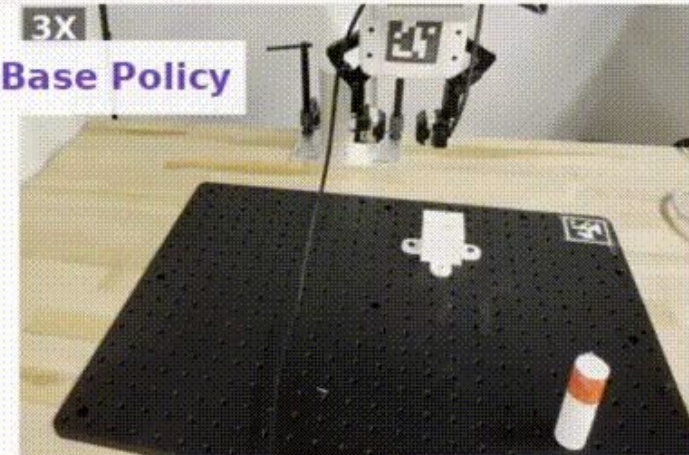
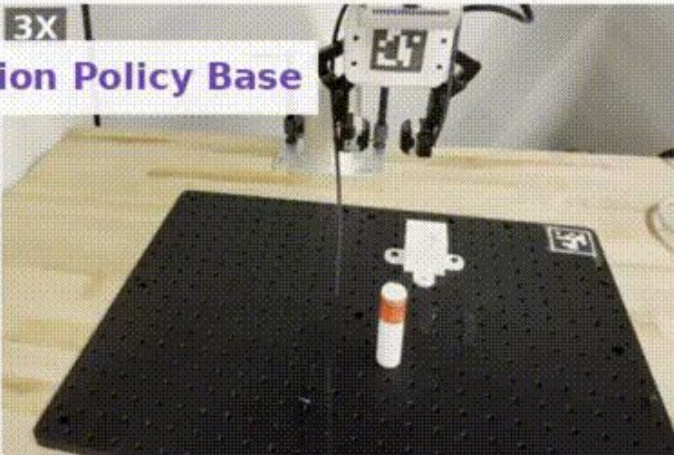
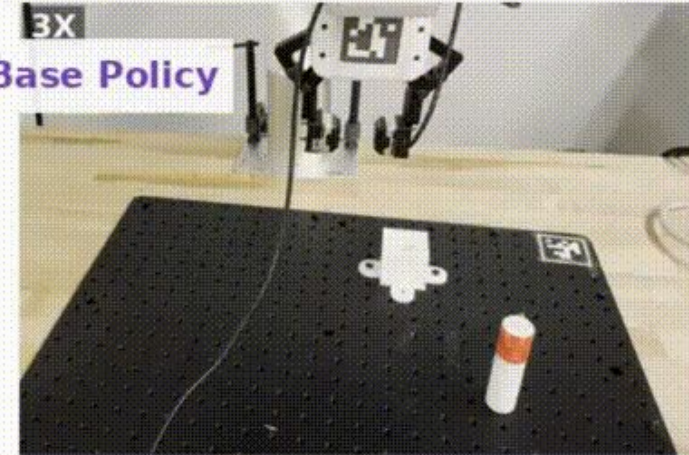
3X



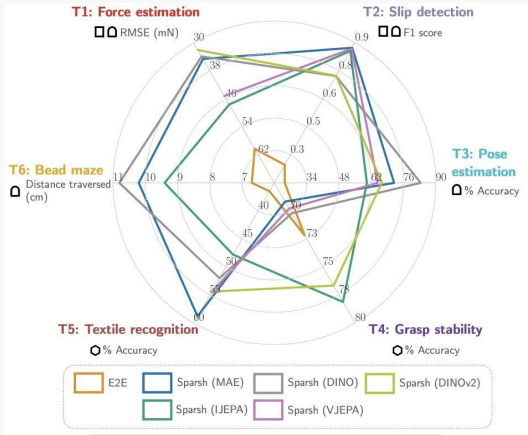
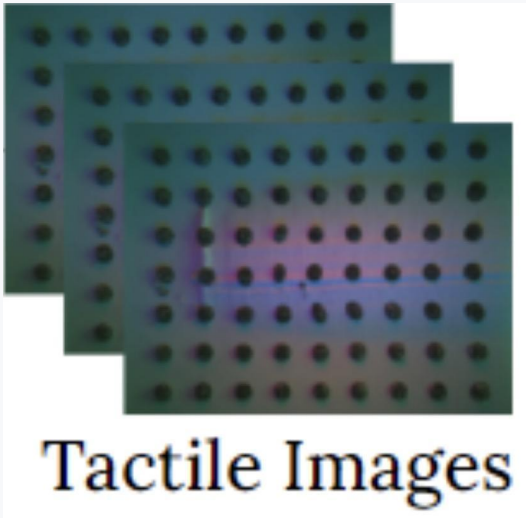
10X



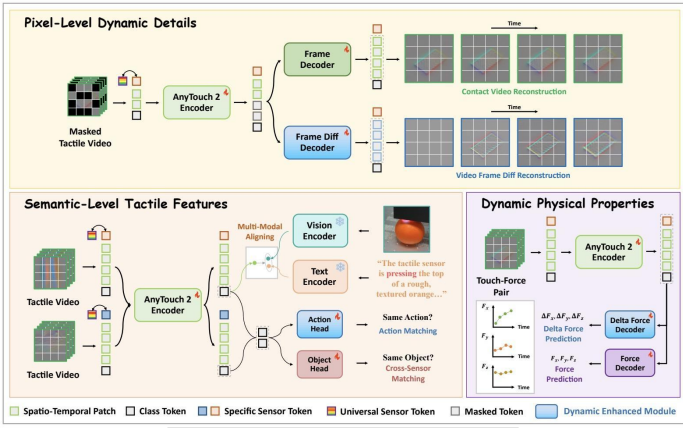
General Pipeline: Across Different Base Policies



General Pipeline: Across Different Tactile Representations

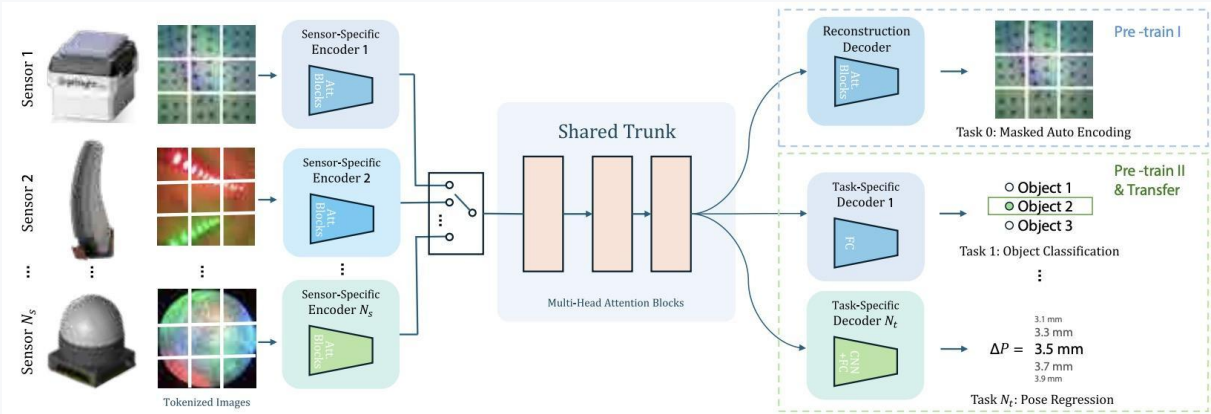
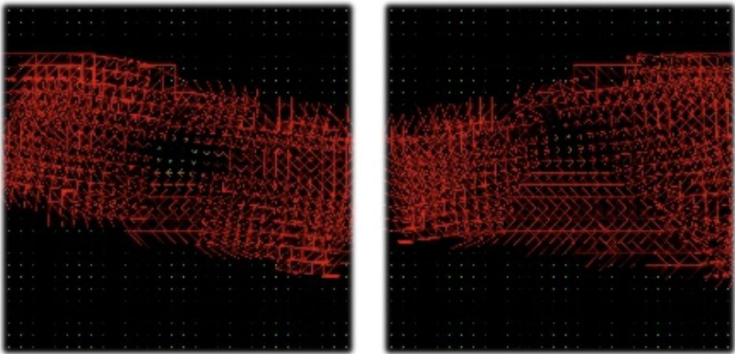


Higuera, CoRL 2024



Feng, ICLR 2026

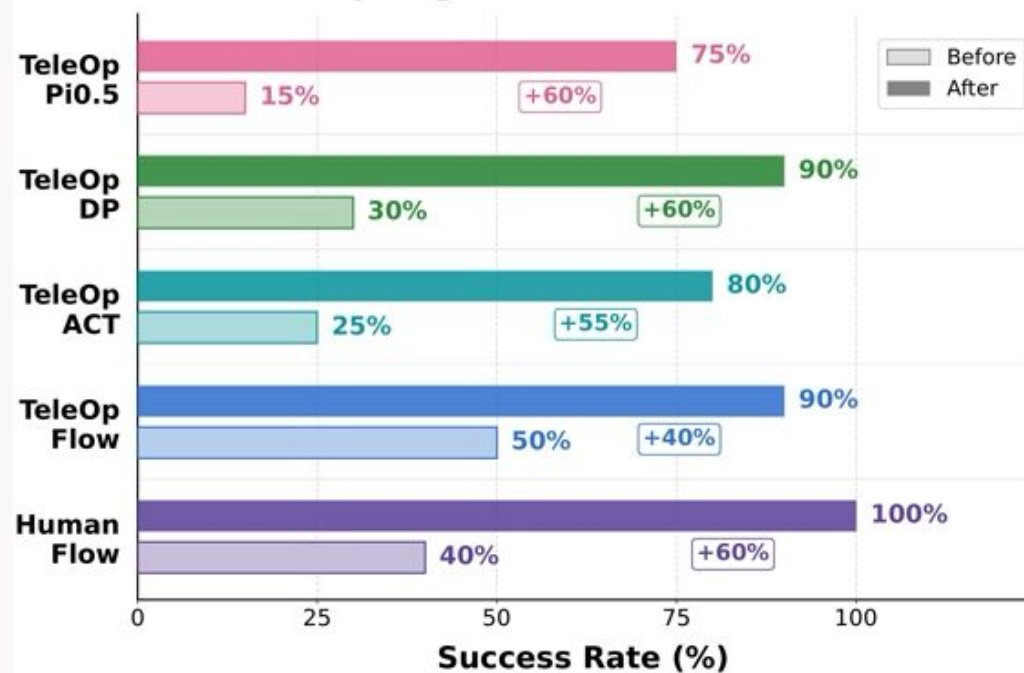
Tactile Force Maps



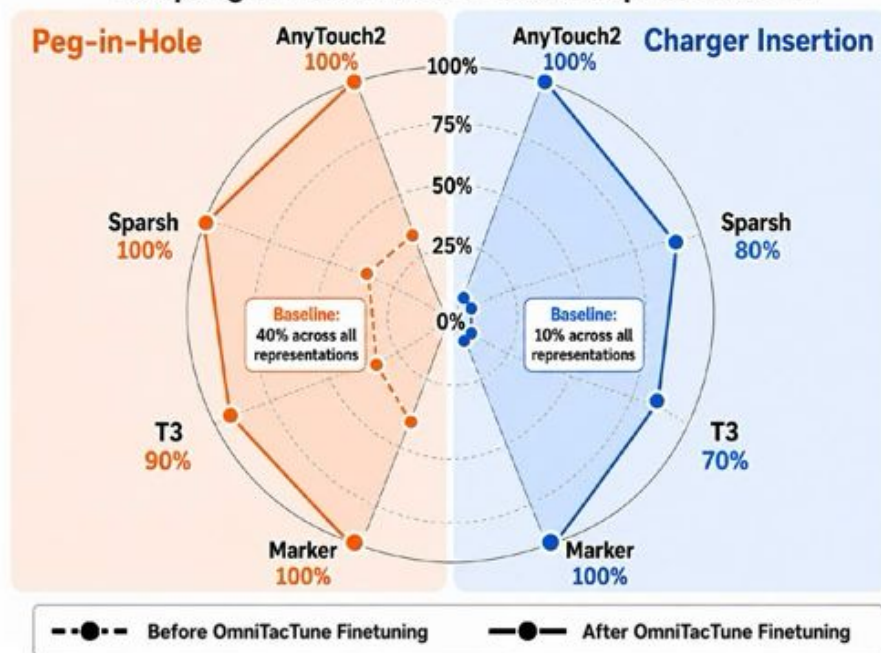
Zhao, CoRL 2024

Results

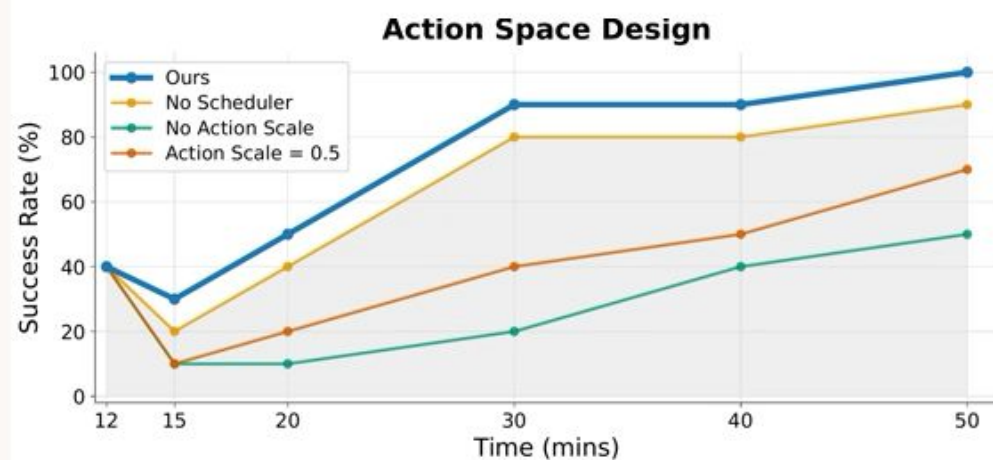
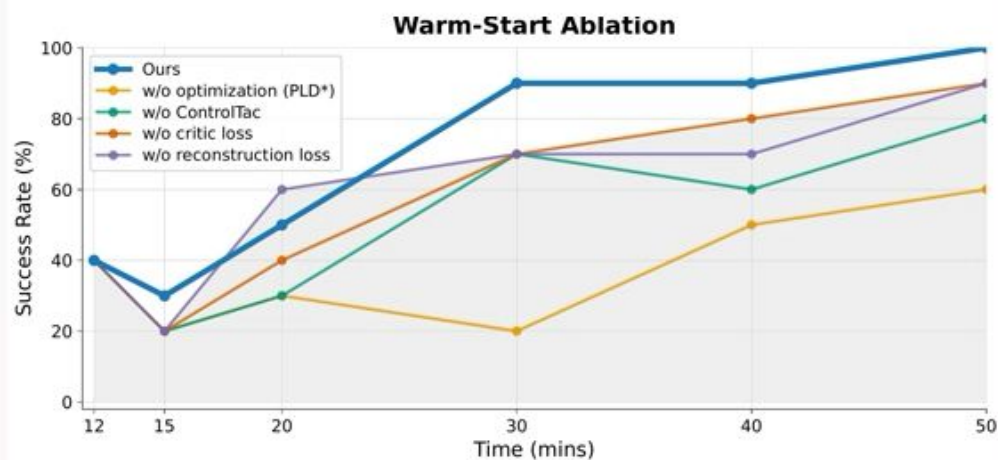
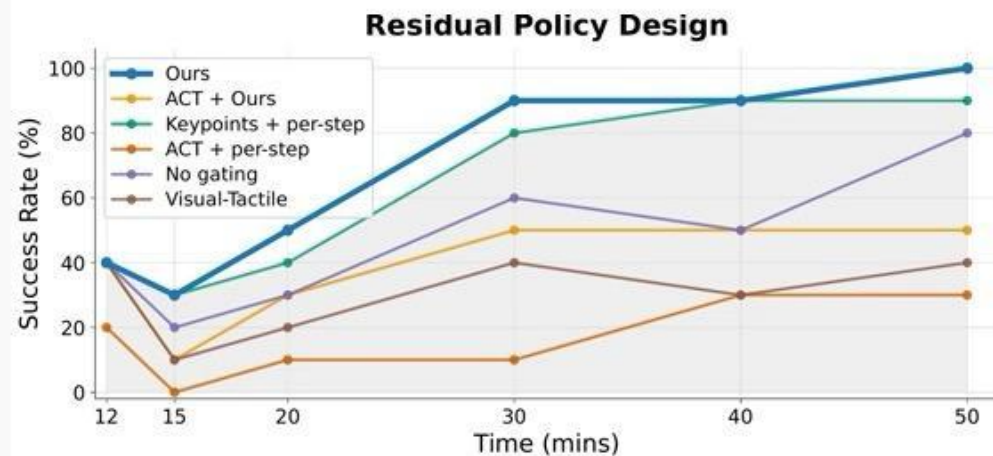
Adapting Diverse Base Policies



Adapting Across Different Tactile Representations



Ablations



One-take 40 mins Training Demo

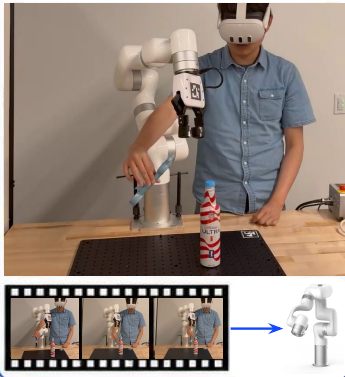


Tactile Data Augmentation for Sample Efficiency

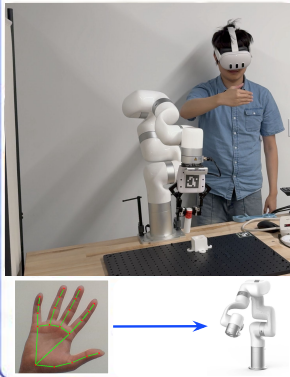
1. Data Collection and Base Policies

(a) Data Collection Sources

Human Video + Retargeting

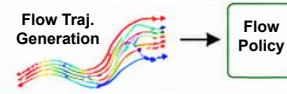


VR Teleoperation



(b) Base Policies

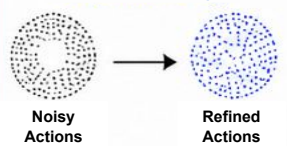
Human Policy



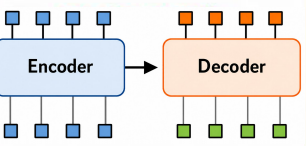
$\pi_{0.5}$



Diffusion Policy

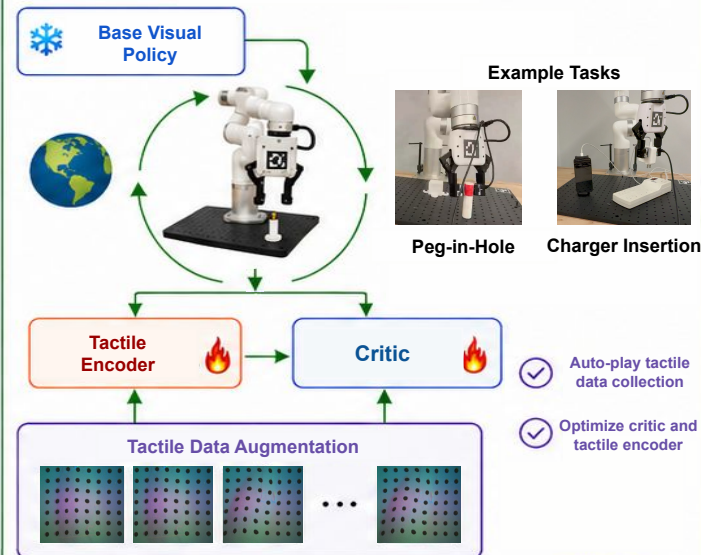


ACT

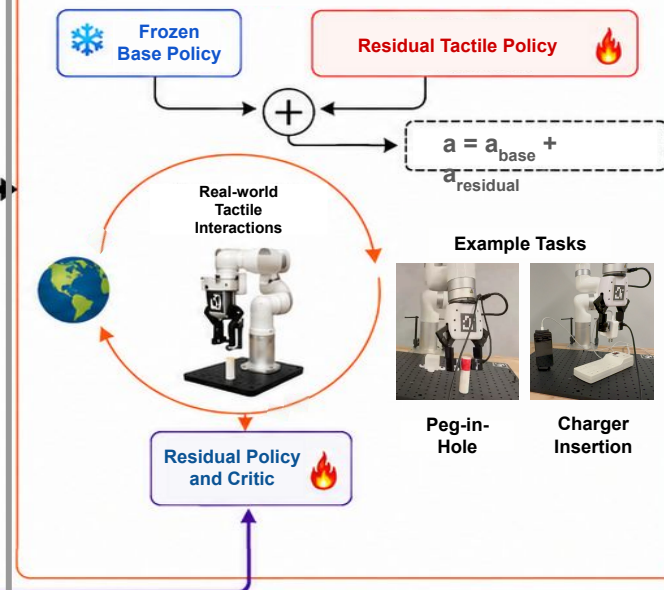


2. Two-stage Real-world Tactile Residual RL

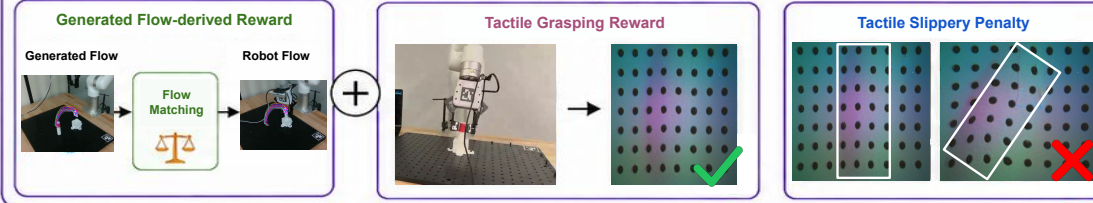
Stage 1: Warmstart



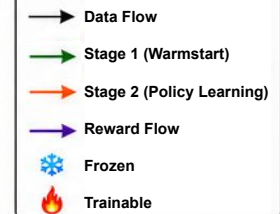
Stage 2: Policy Learning



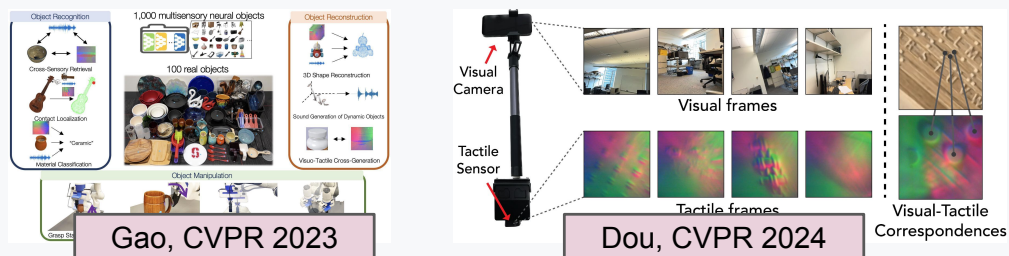
Dense Reward Shaping Pipeline



Legend

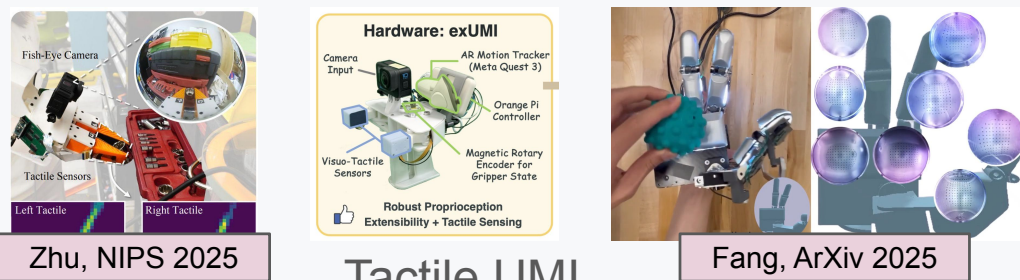


Recent Work in Tactile Data Scaling

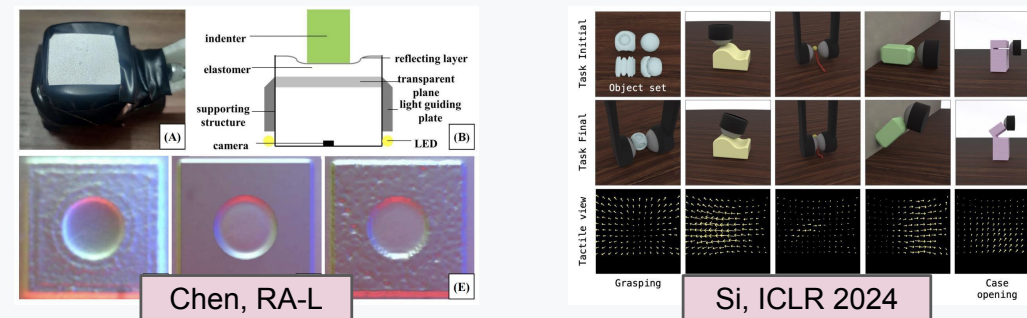


Calibrated Visual-Tactile Data

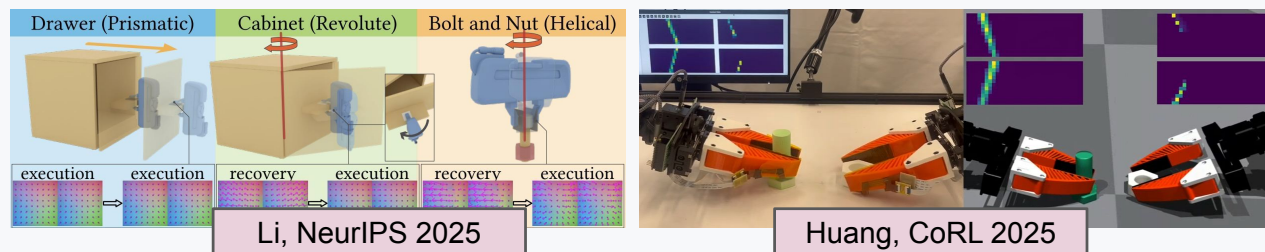
- Require huge efforts for calibrations



- Specific hardware for tactile scaling



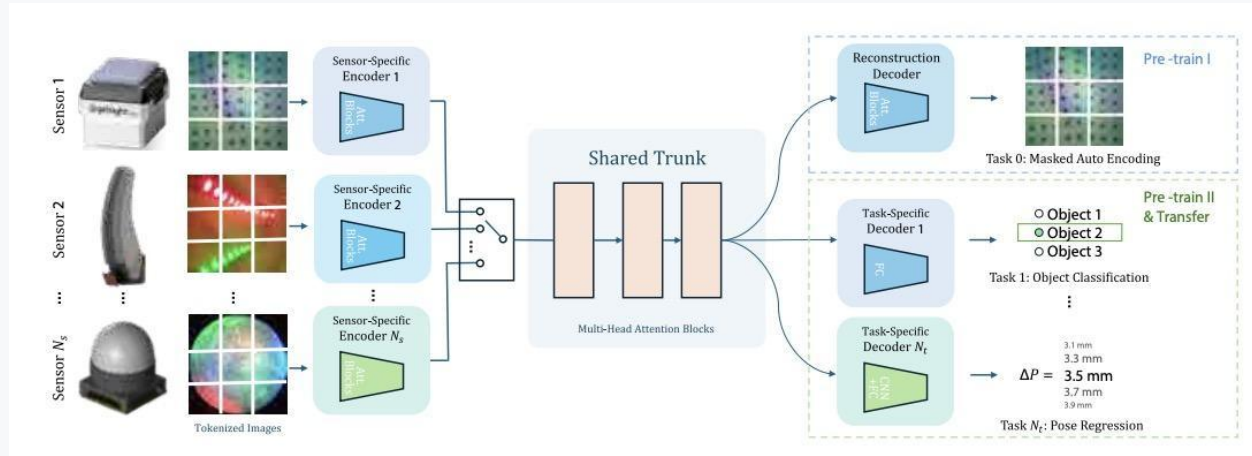
Differentiable Tactile Simulator



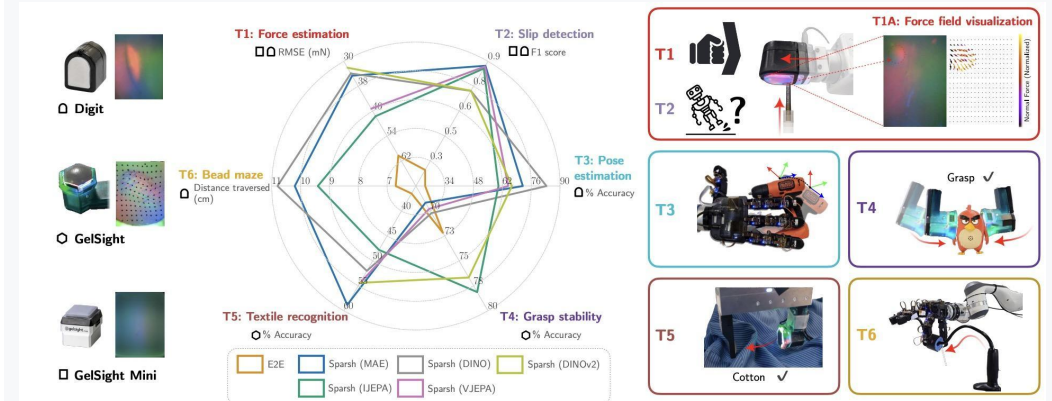
GPU-accelerated Tactile Simulator

- Scaling tactile data via physics-based simulator

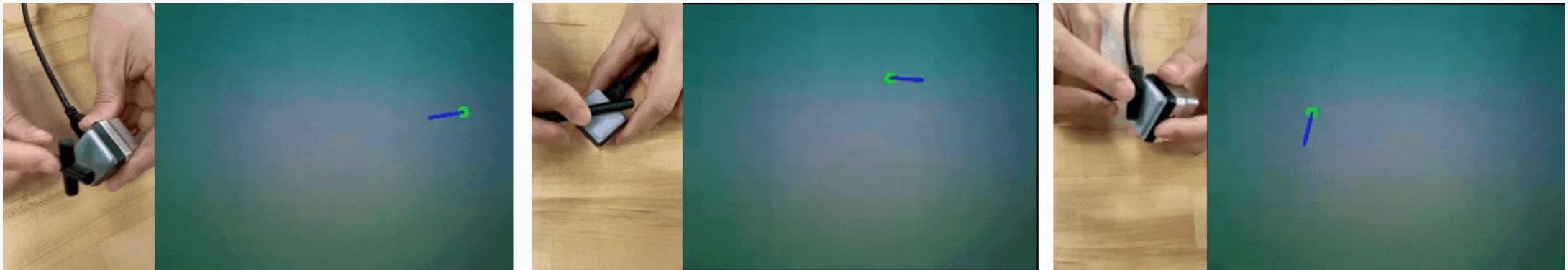
Tactile Data Usage



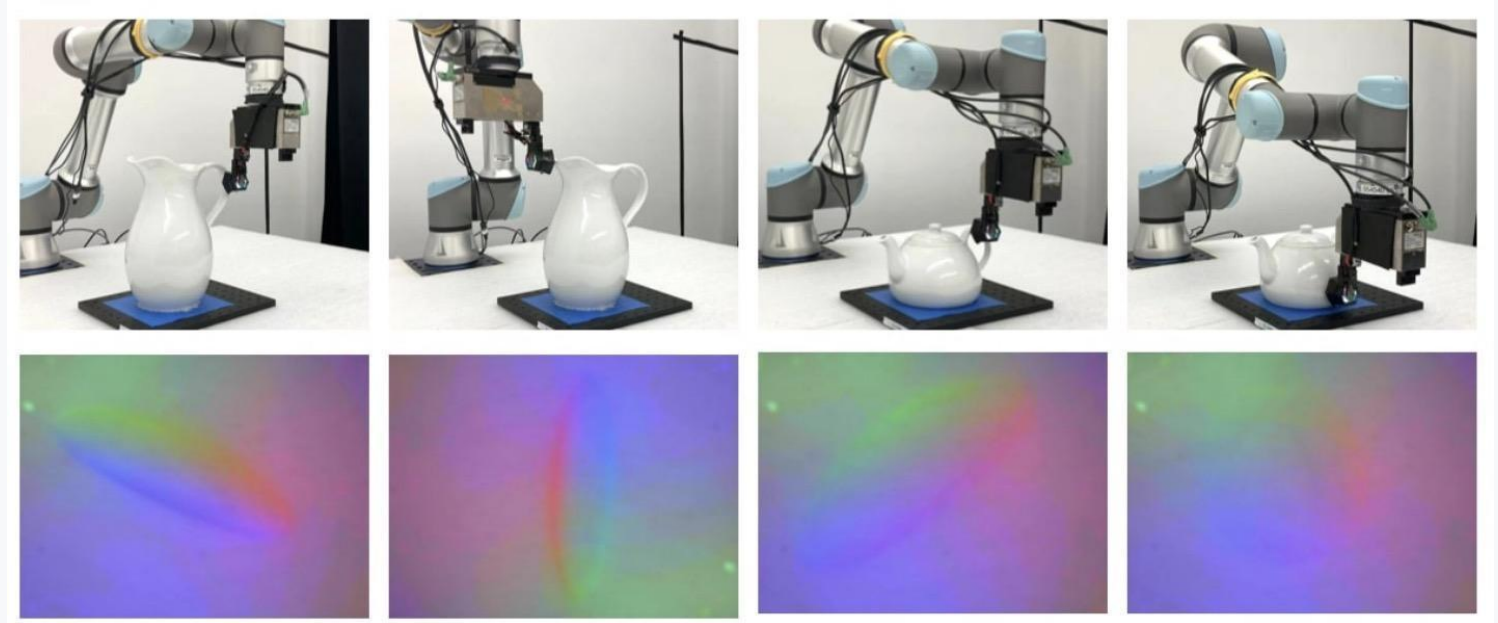
Zhao et al., T3, Corl 2024



Higuera et al., Spash, Corl 2024

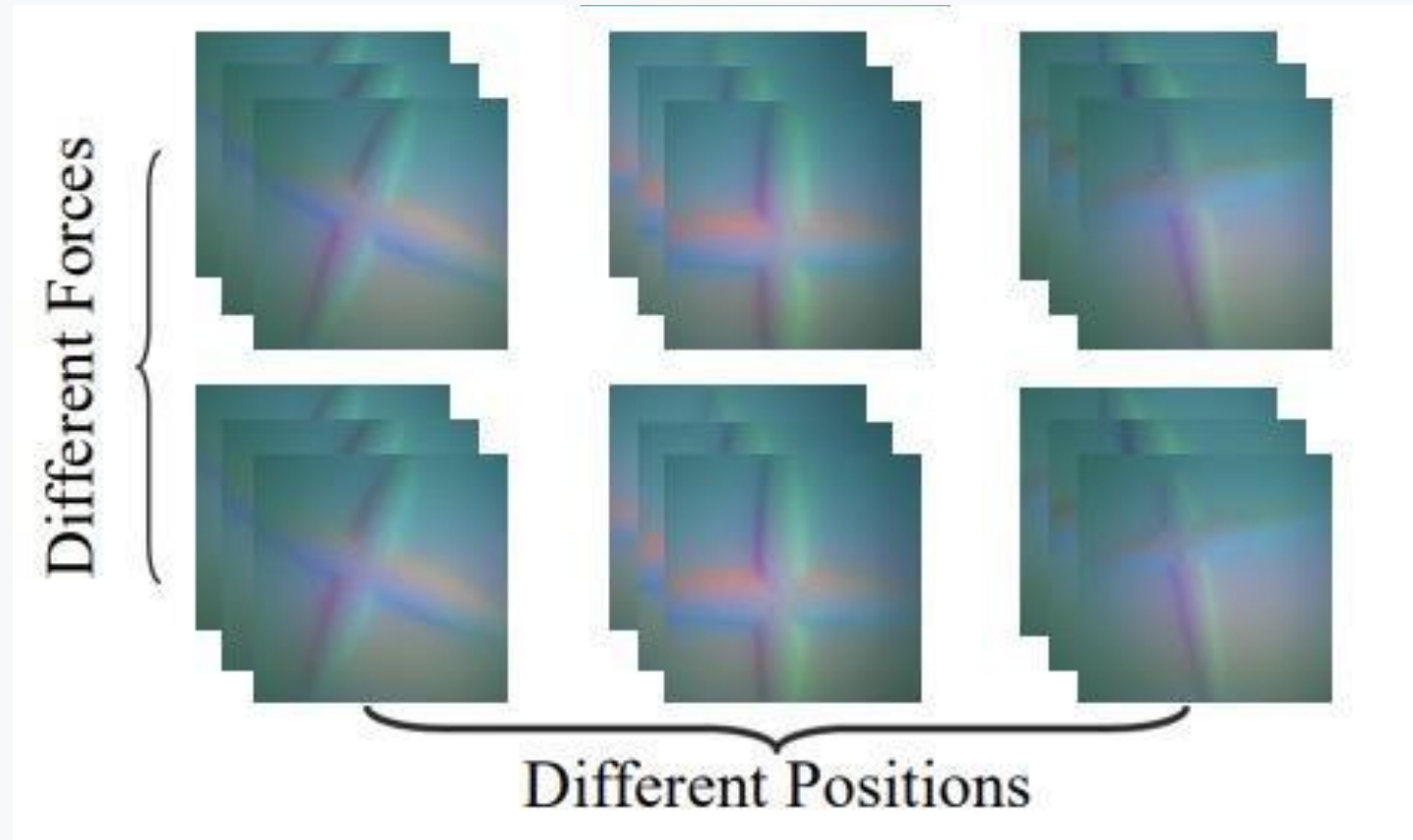
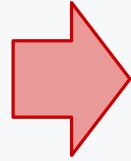


Touch is Essential, but Tactile data is Fragmented

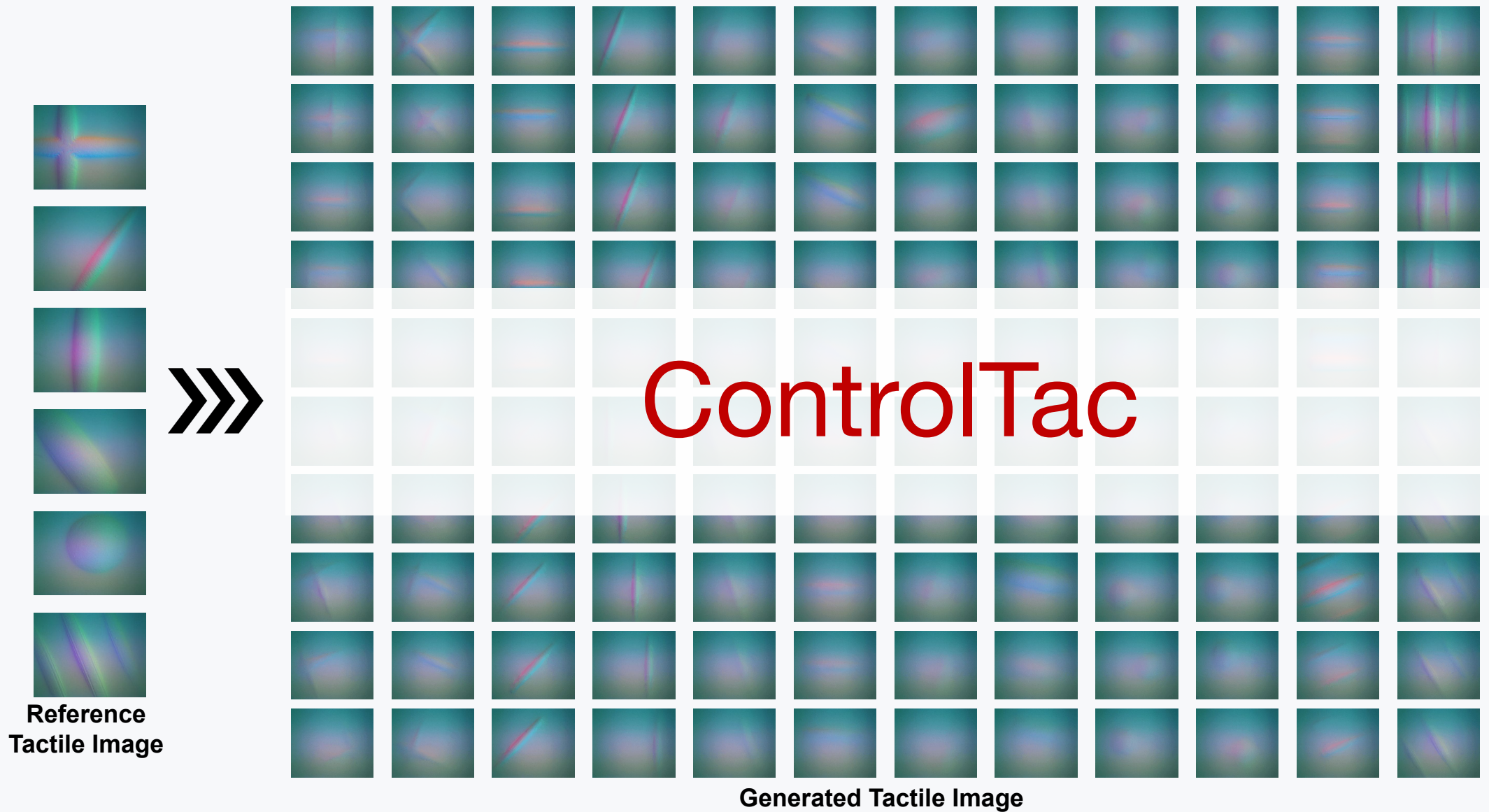


Si et al., RA-L

Touch is Essential, but Tactile Data is Fragmented

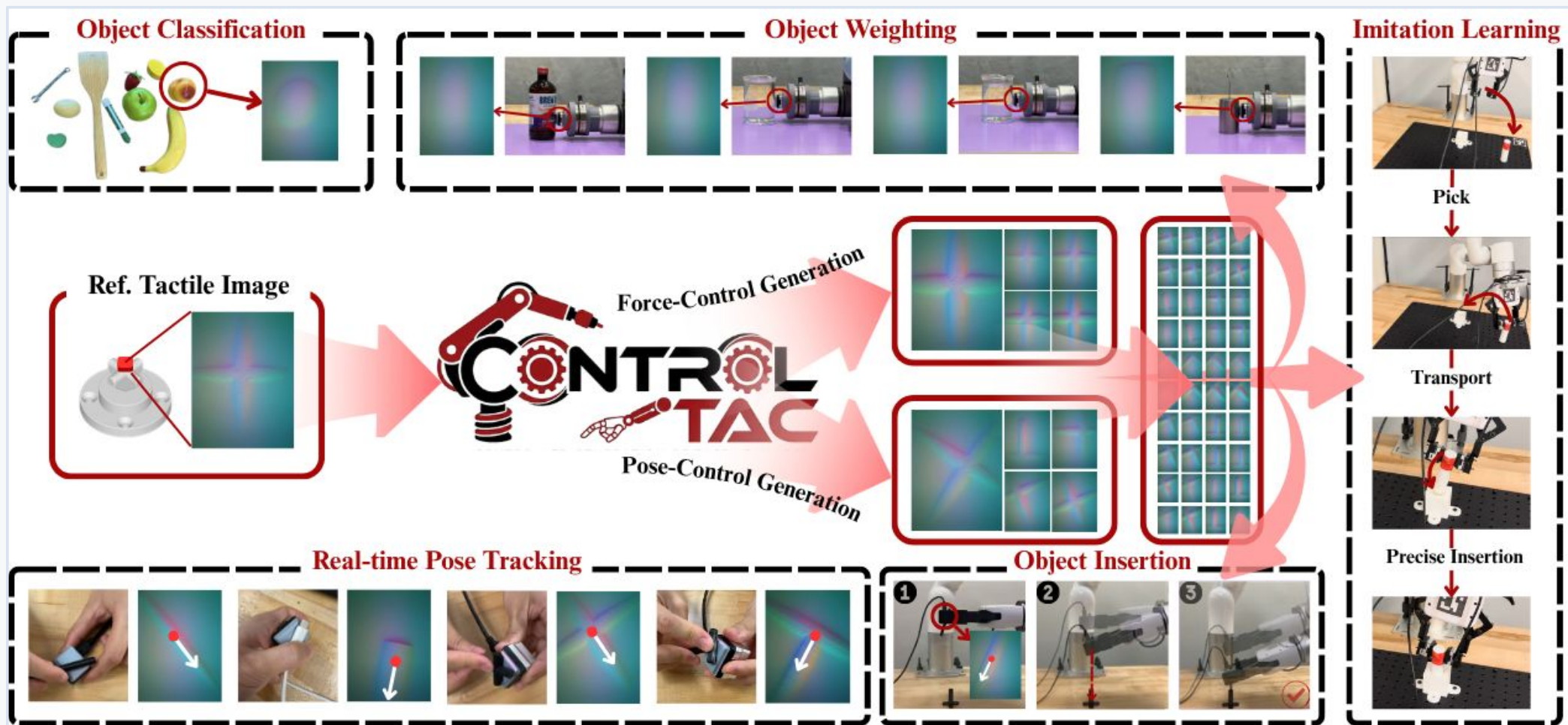


Each contact corresponds to various contact positions and forces

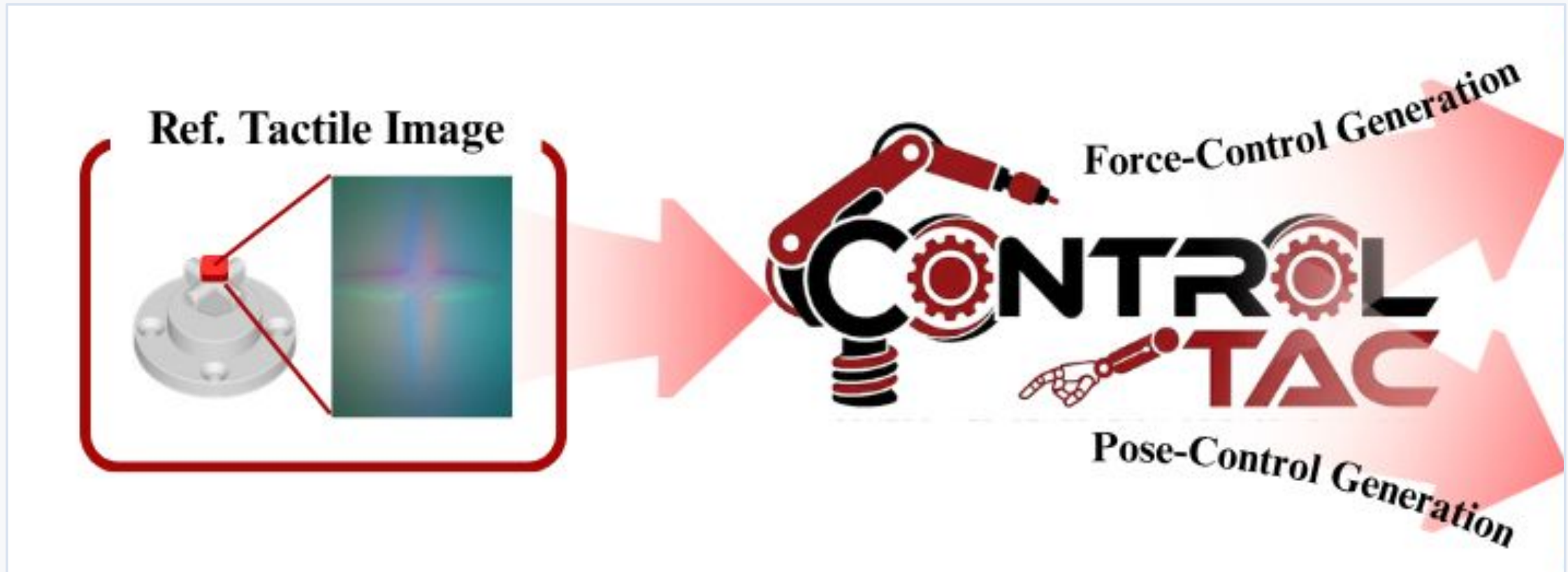


Dongyu Luo*, Kelin Yu*, Amir-Hossein Shahidzadeh, Cornelia Fermuler, Yiannis Aloimonos, Ruohan Gao
ControlTac: Force- and Position-Controlled Tactile Data Augmentation with a Single Reference Image
ICCV CDEL Workshop **Oral**, In Submission

ControlTac: Physically Controlled Tactile Image Generation

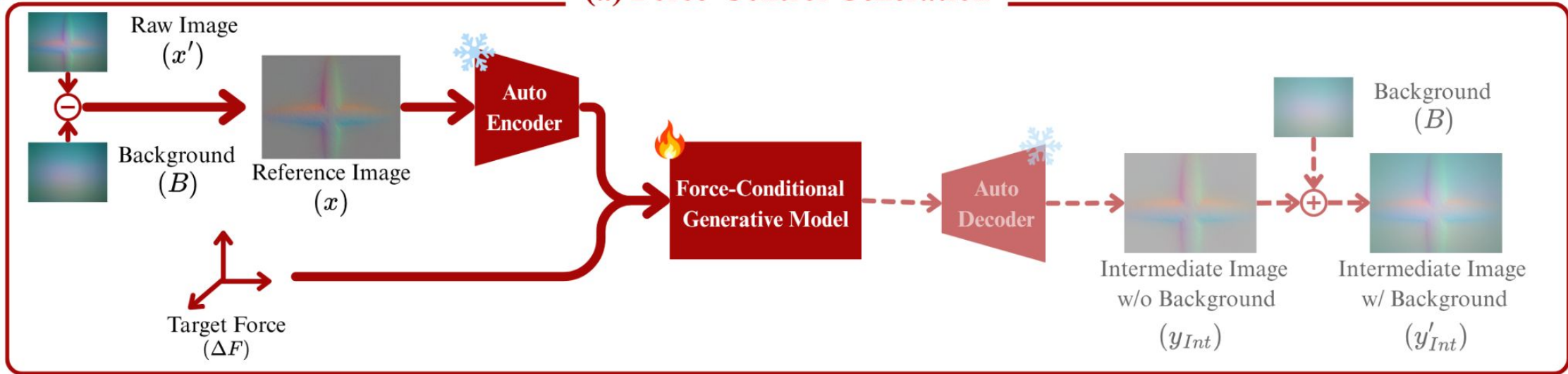


ControlTac: Physically Controlled Tactile Image Generation



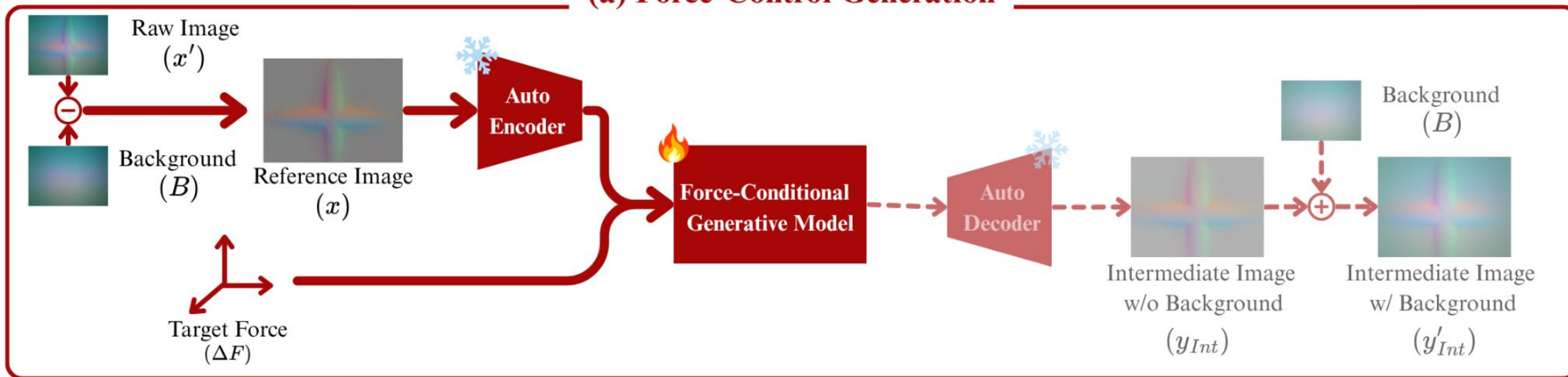
ControlTac: Physically Controlled Tactile Image Generation

(a) Force-Control Generation

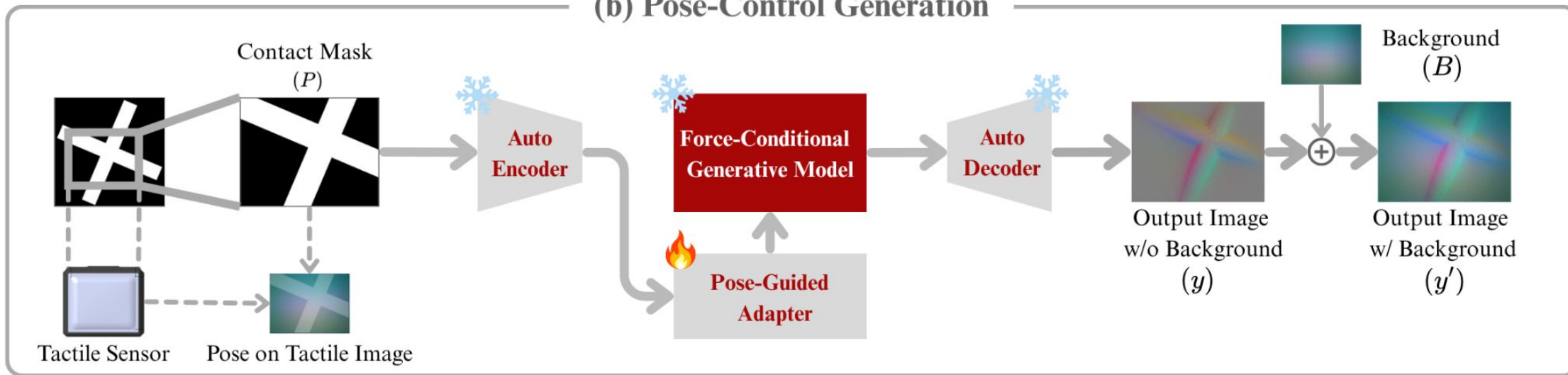


ControlTac: Physically Controlled Tactile Image Generation

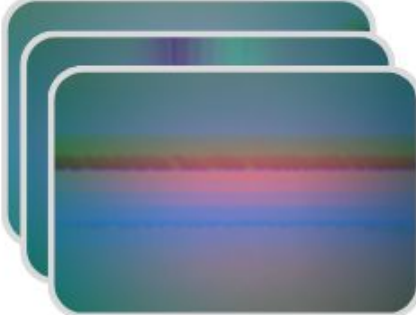
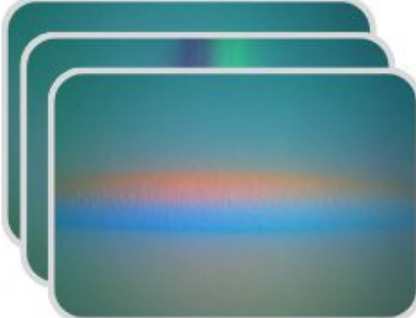
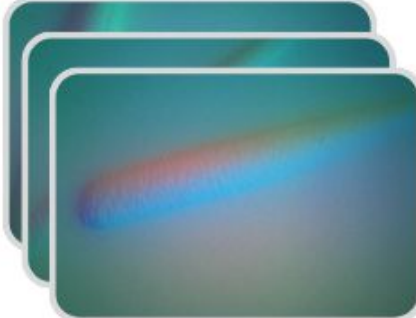
(a) Force-Control Generation



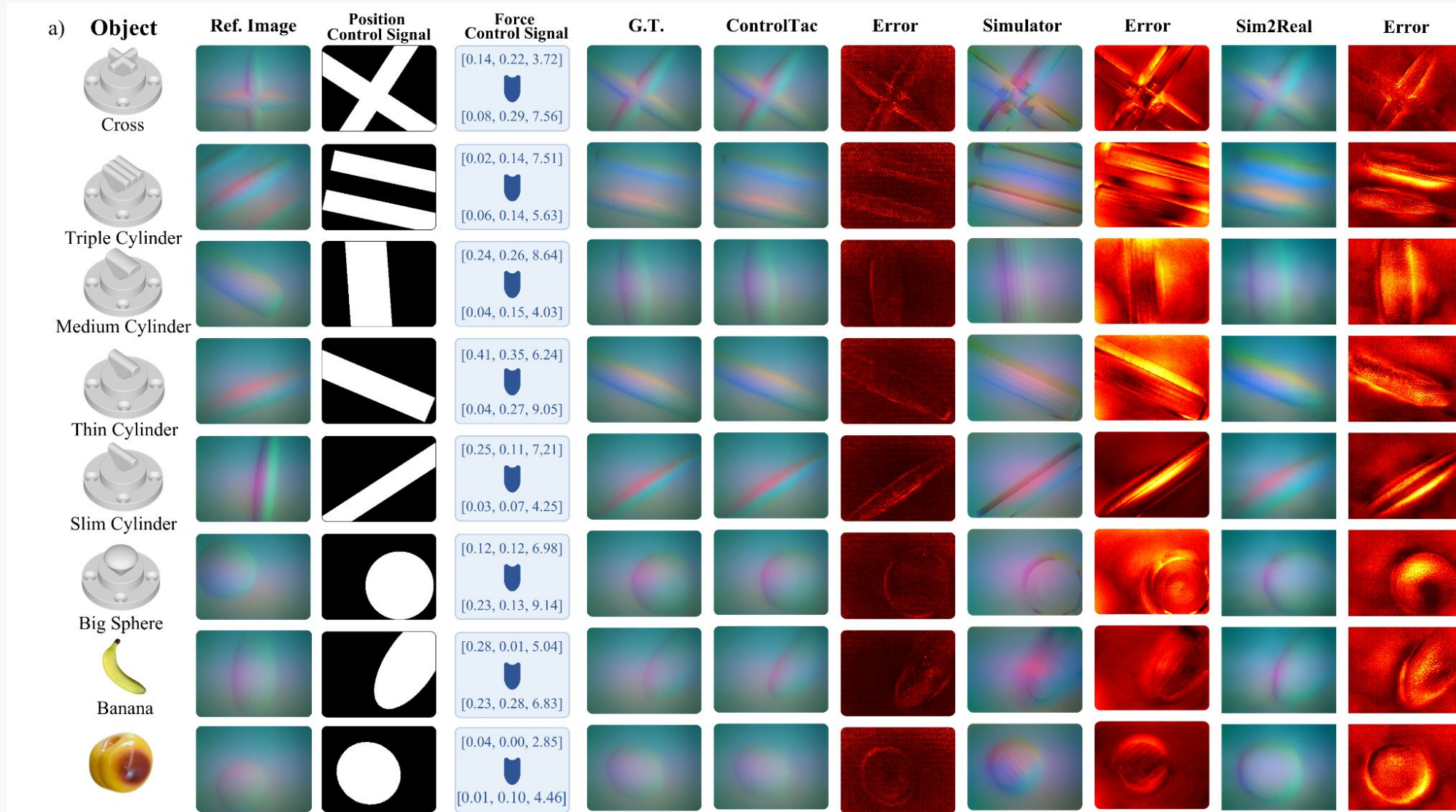
(b) Pose-Control Generation



ControlTac: Physically Controlled Tactile Image Generation

	Text2Tac	Vis2Tac	Simulation	ControlTac
Input	The touch of braided rope is smooth, polished, firm, cool, grainy.			
Generated				
Real				

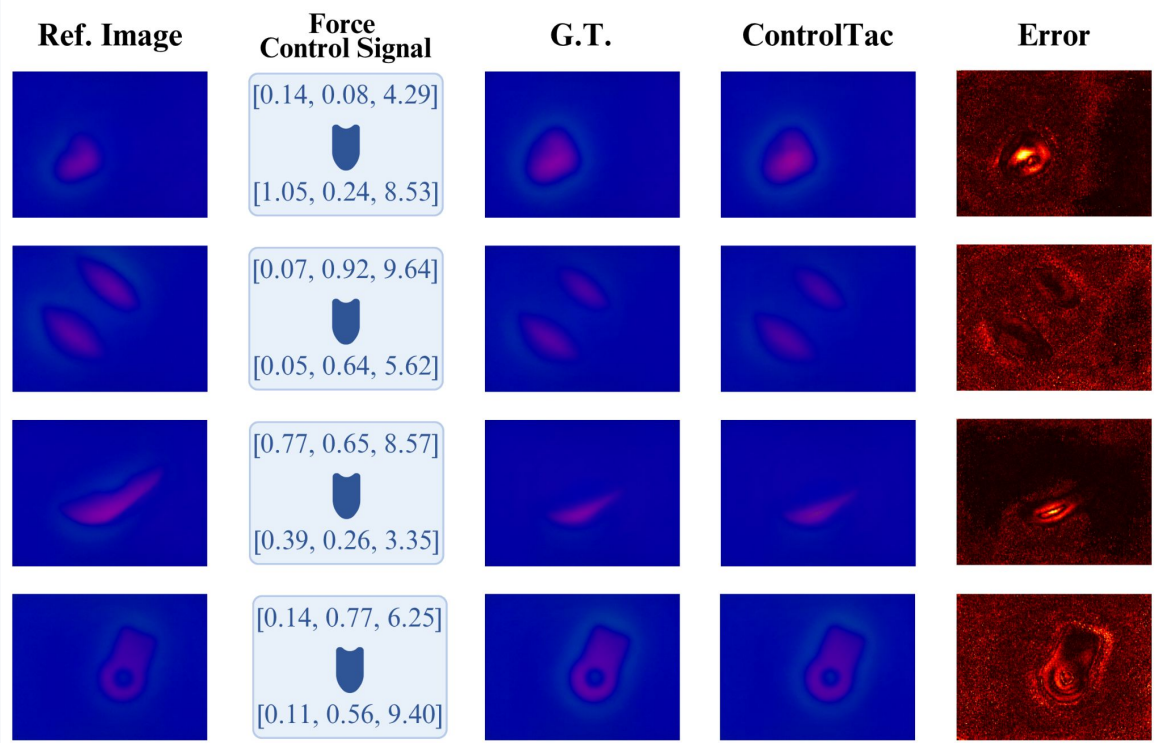
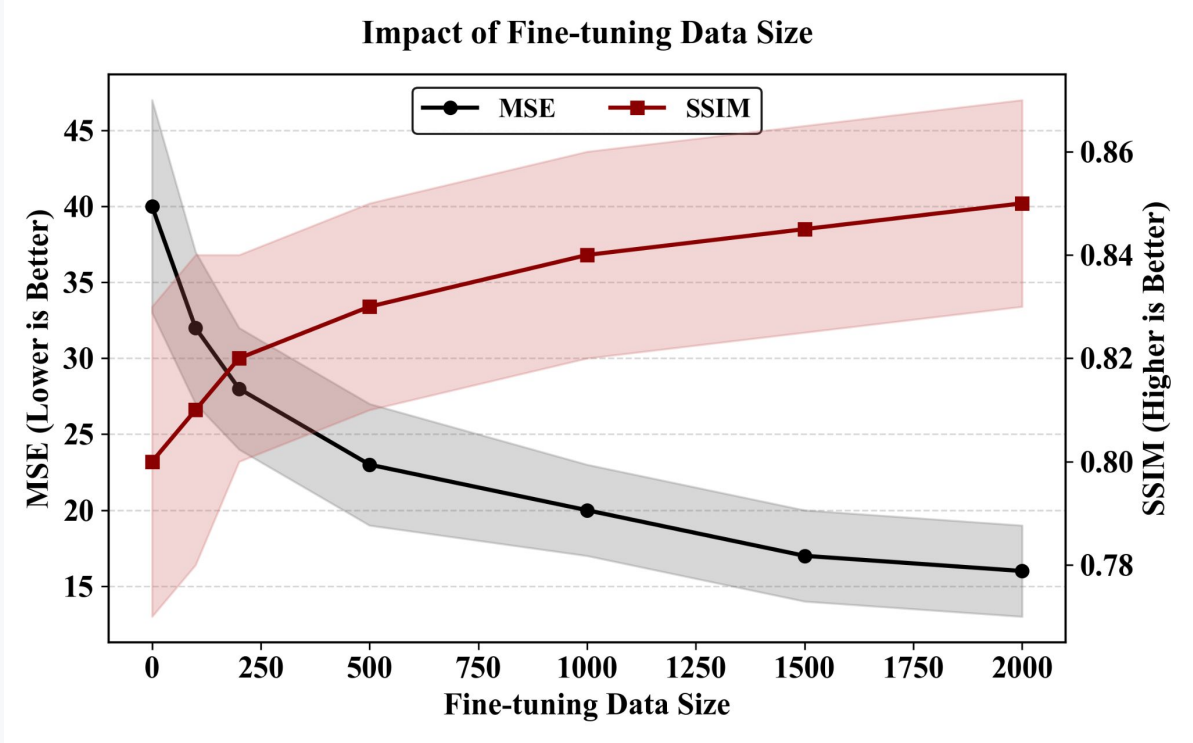
ControlTac: Physically Controlled Tactile Image Generation



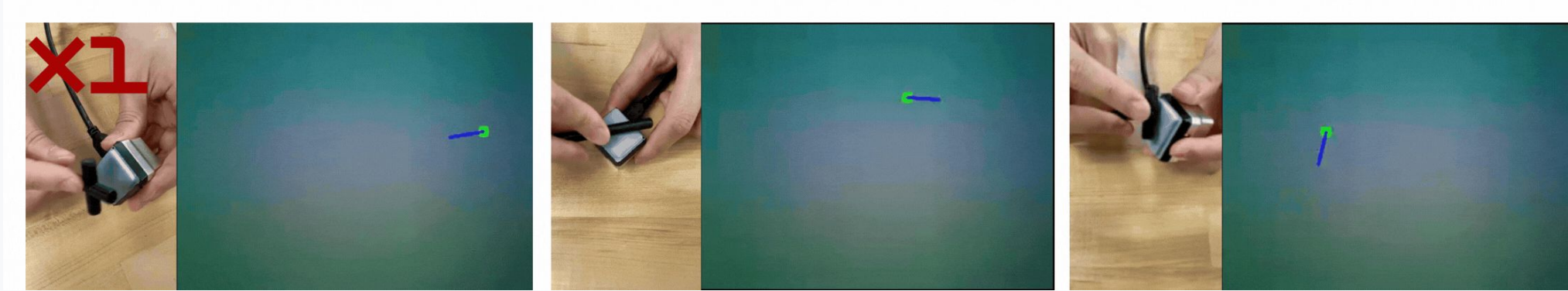
ControlTac: Physically Controlled Tactile Image Generation

METHOD	SEEN OBJECTS		UNSEEN: FEELANYFORCE		UNSEEN: ANYTOUCH2	
	MSE↓	SSIM↑	MSE↓	SSIM↑	MSE↓	SSIM↑
Simulator	1054±19	0.68±0.03	1065±23	0.69±0.03	2157±14	0.61±0.04
Sim2Real	239±17	0.74±0.03	253±25	0.73±0.03	545±31	0.70±0.05
Hybrid	31±5	0.81±0.04	37±6	0.75±0.04	—	—
Separate	157±8	0.79±0.04	199±11	0.72±0.05	—	—
CONTROLTAC (Ours)	23±2	0.83±0.03	26±3	0.79±0.04	29±2	0.81±0.02

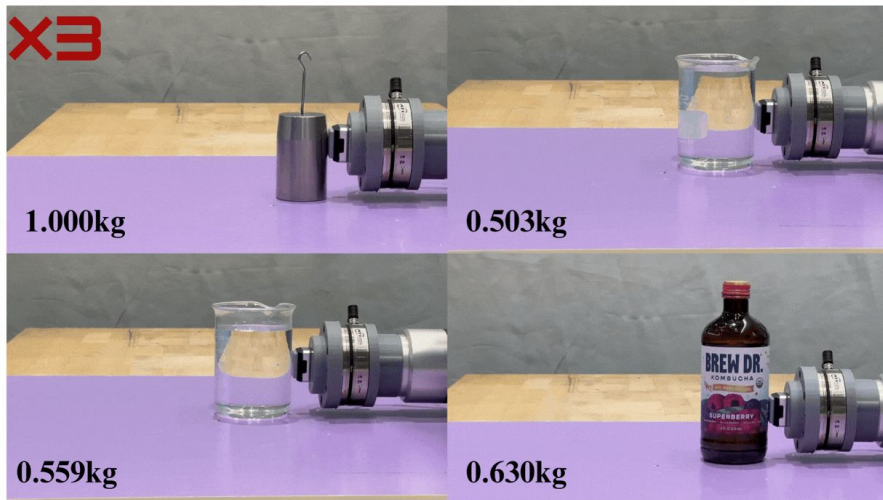
Cross-sensor Evaluation



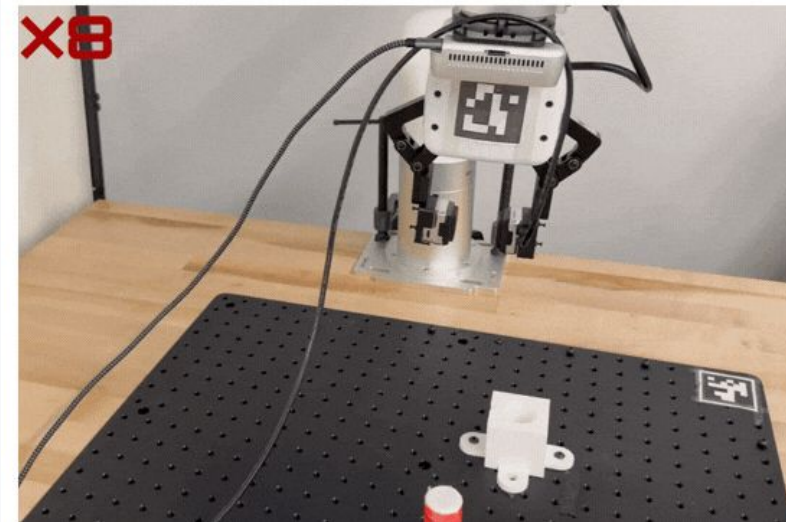
ControlTac: Downstream Tasks with Generated Data



Pose Tracking



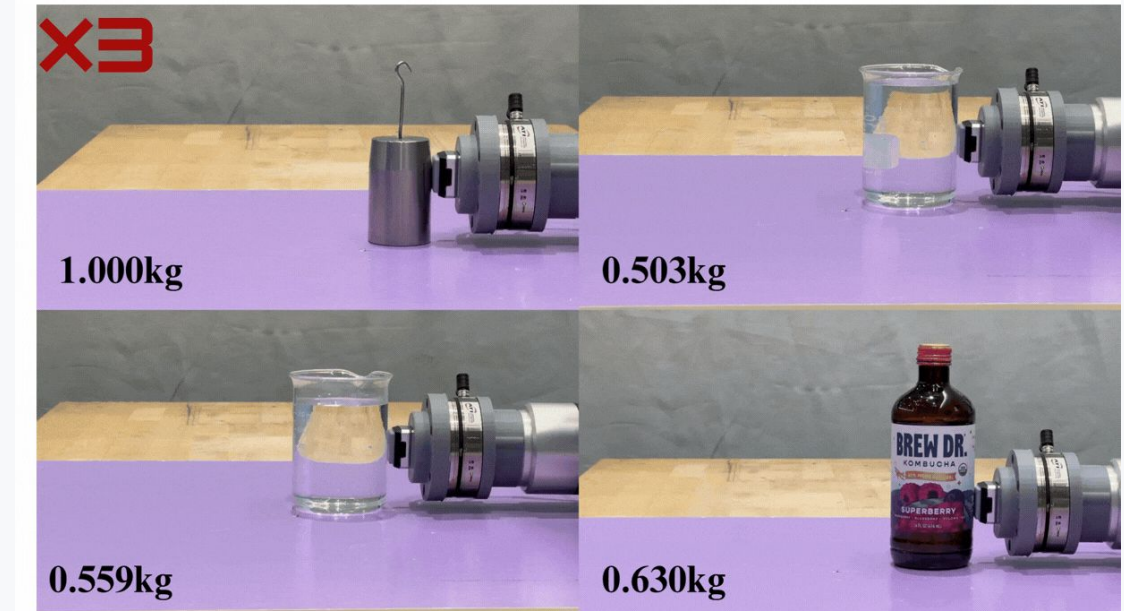
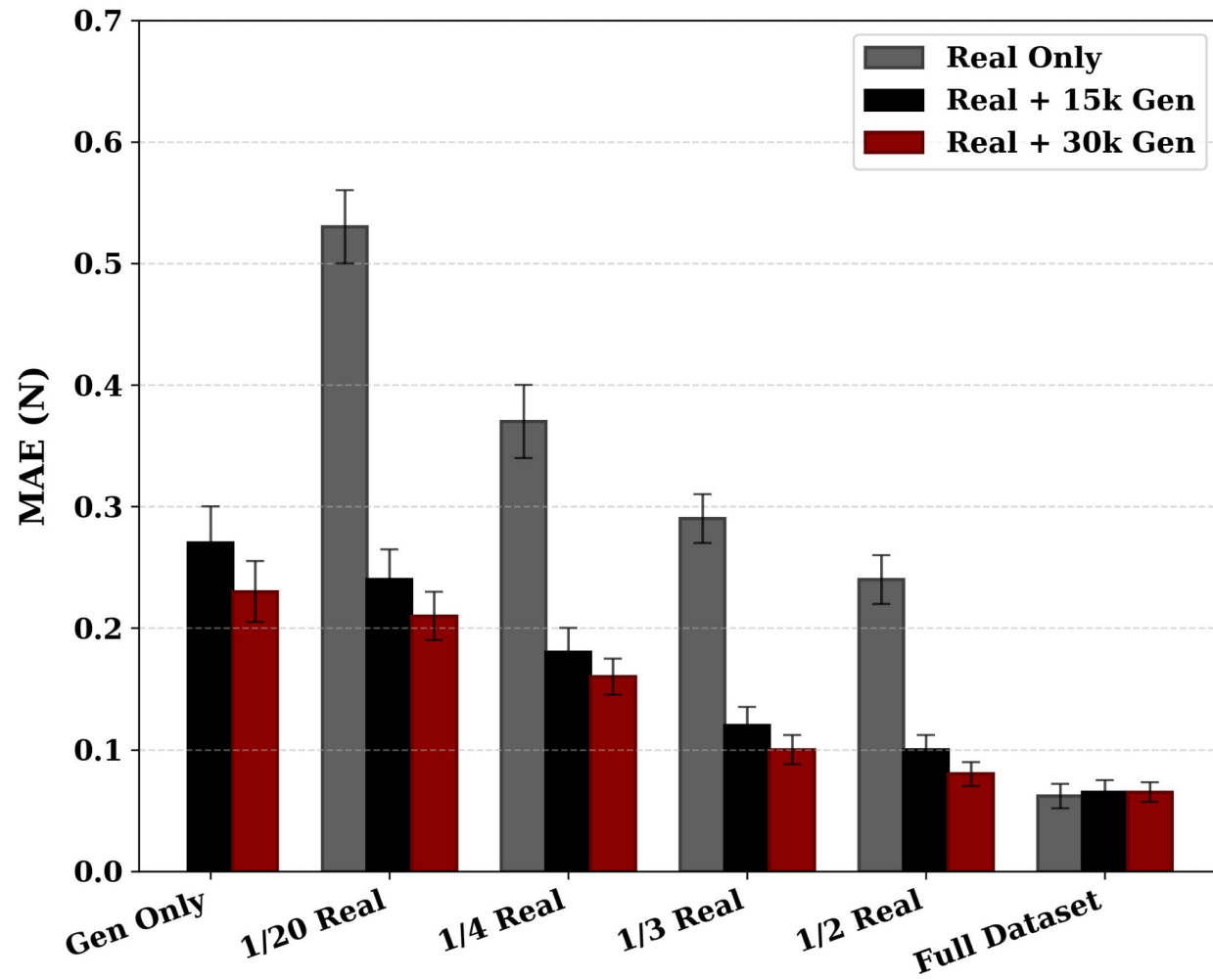
Object Weighting



Imitation Learning

Force Estimation

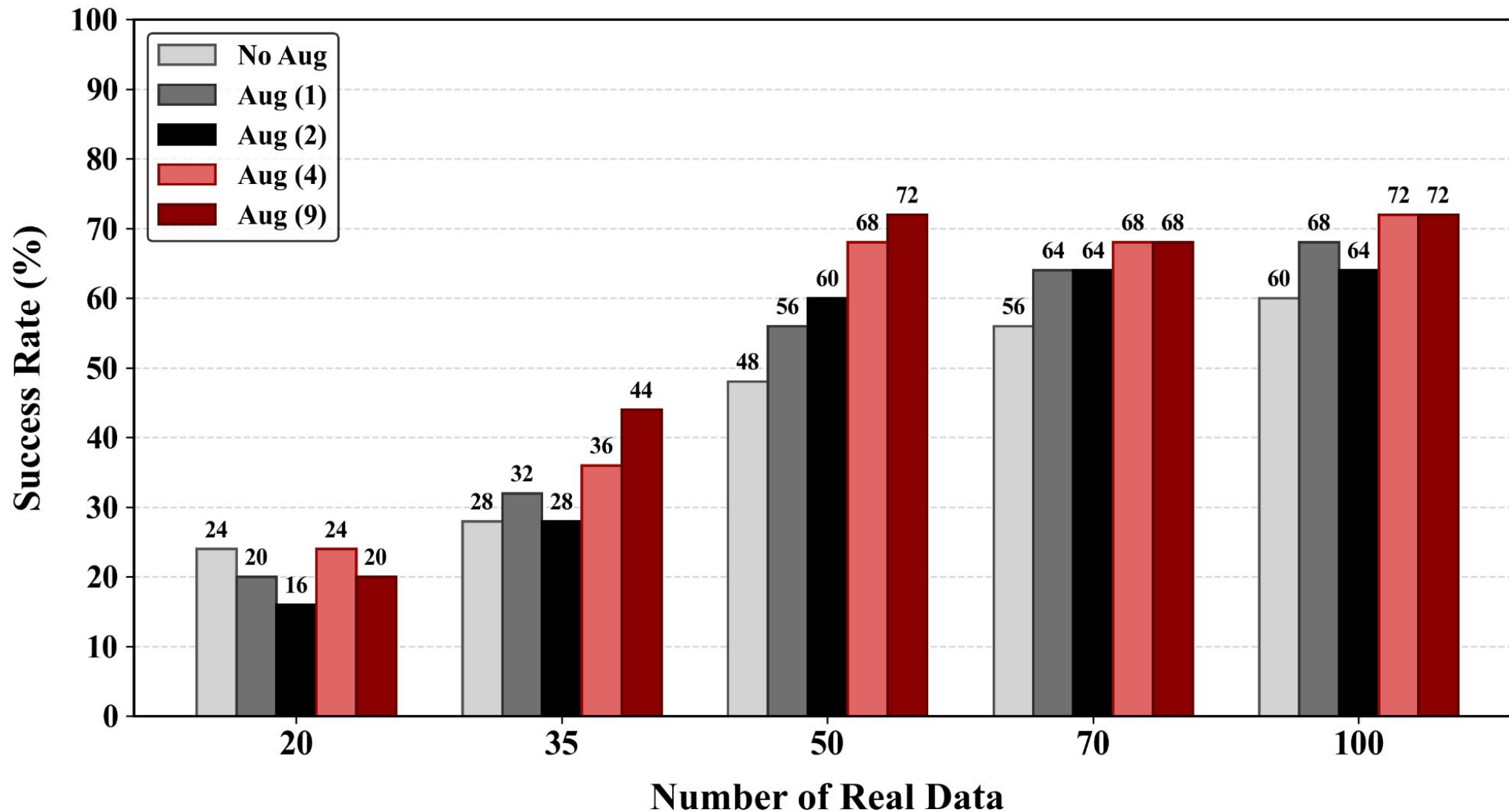
MAE Comparison: Real vs Generated Data



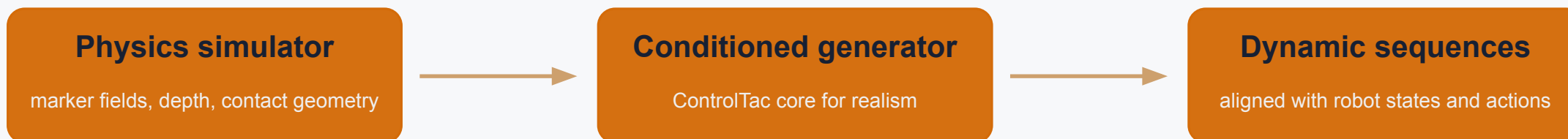
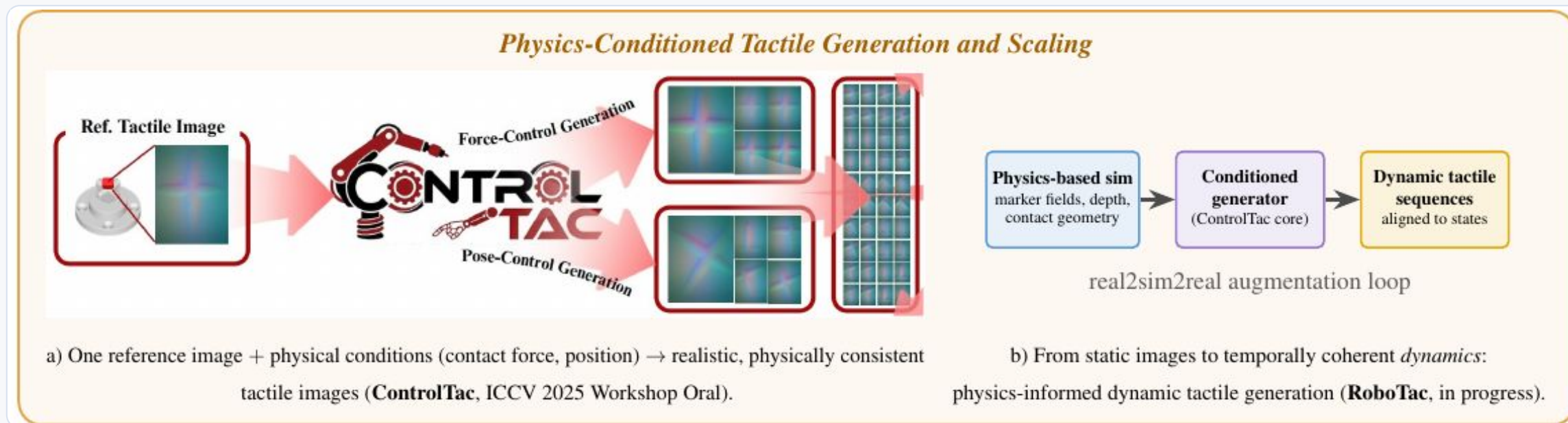
ControlTac: Downstream Tasks with Generated Data

METHOD	CYLINDER (3 TYPES)			CROSS		
	X	Y	ANG	X	Y	ANG
PCA	15	13	22	56	19	18
Real	8	8	4	6	6	2
Simulator	17	15	6	19	18	5
Sim + Real	12	13	6	17	16	4
Ours (fixed force)	9	8	5	7	9	4
Ours (unfixed force)	4	5	3	3	4	1

ControlTac: Downstream Tasks with Generated Data

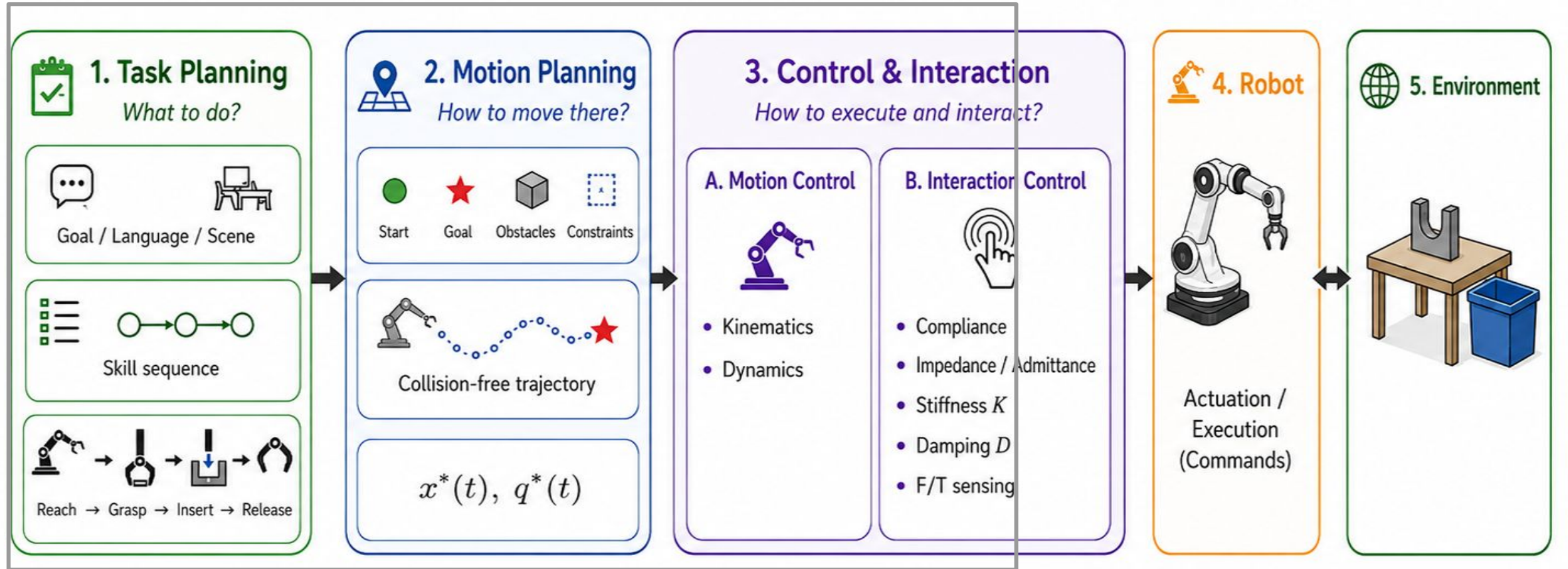


From Static Images to Tactile-action Dynamics

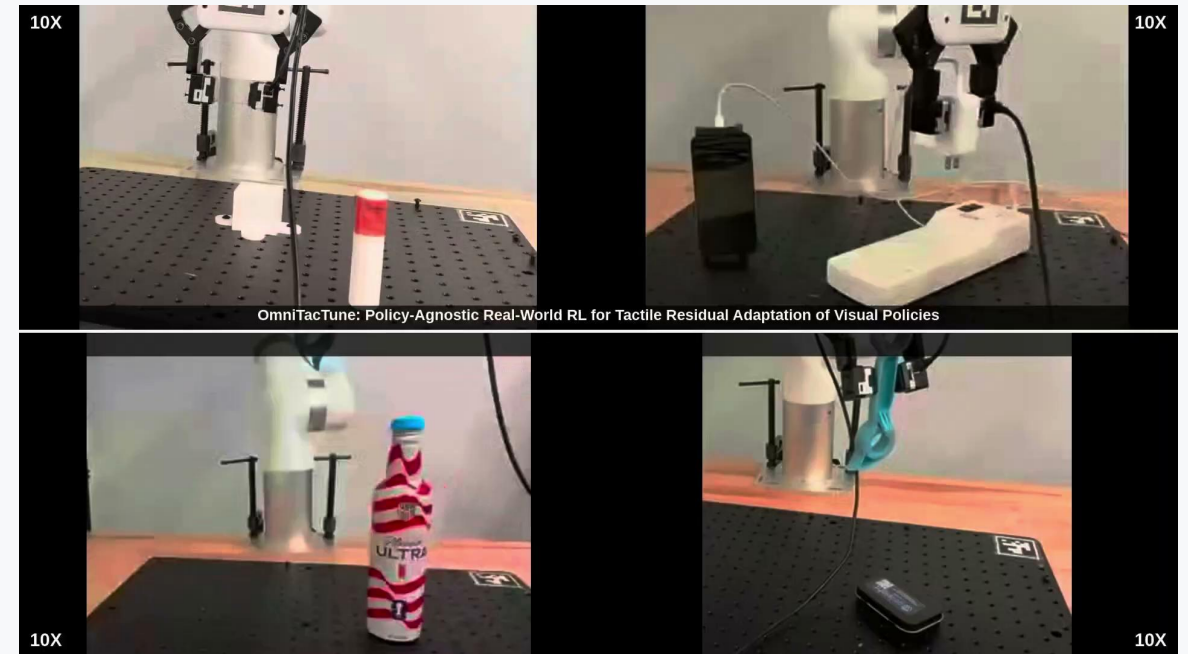


RoboTac direction: a real-to-sim-to-real augmentation loop for tactile world models and policy tuning.

Robot Manipulation in My Mind



Robot Manipulation in My Mind

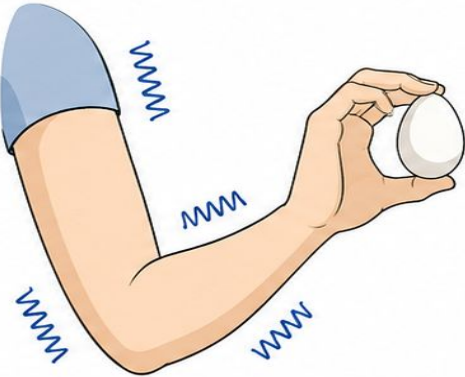


- Cannot reach 100% success rates for those tasks which are easy for human
- One gap comes from controller & hardware sides, which cannot be bridged

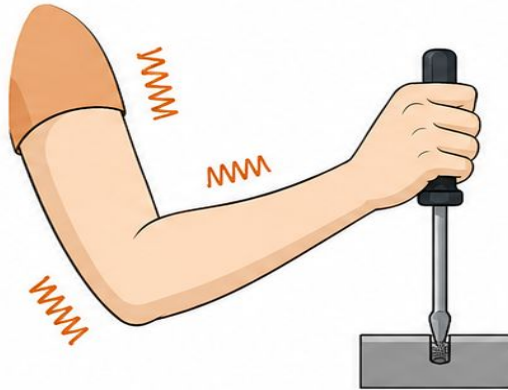
Future Idea: Controller / Stiffness Gaps between Human and Robot

Human Arm: Task/Object-Dependent Compliance

1 Handle fragile object
soft

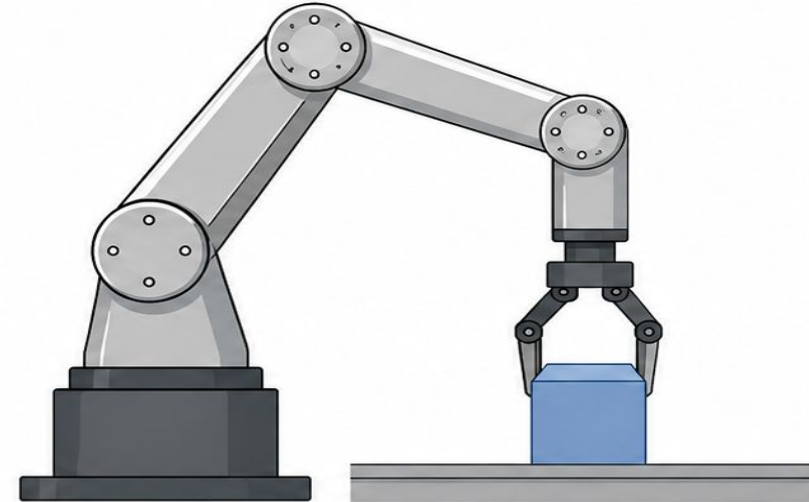


2 Use tool / apply force
stiff



can be soft or stiff depending on task and object

Robot Arm: Rigid Body



mostly rigid / fixed mechanical structure

Human: adaptable compliance

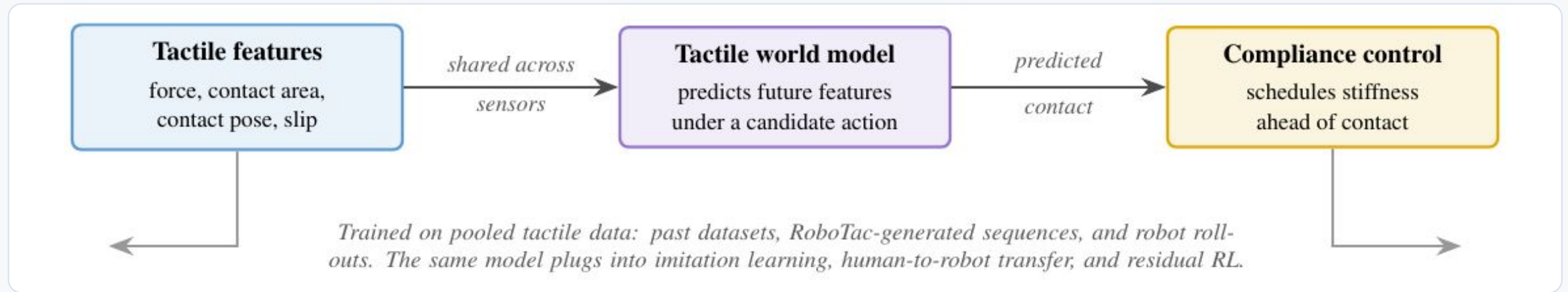


versus



Robot: rigid embodiment

Future Idea: A Tactile World Model over Shared Physical Features



The representation should be physical, not sensor-specific pixels.

State: normal/shear force, contact area, contact pose, slip
 Data: prior tactile datasets, RoboTac sequences, OmniTacTune rollouts, public tactile data
 Use: contact grounding, imagined rollouts, and compliance control

Hypothesis

Many small sensor-specific datasets become one trainable corpus in a common physical space.

Thanks for your listening!

Q & A

