UniDex-ViTac: Learning Unified Visuo-Tactile Dexterous Manipulation Policy from Human Video Data

Hyesung Lee[1,2], Si-Hwan Heo[1], Sungwook Yang[1†]
[1]KIST, [2]KAIST

Project Website

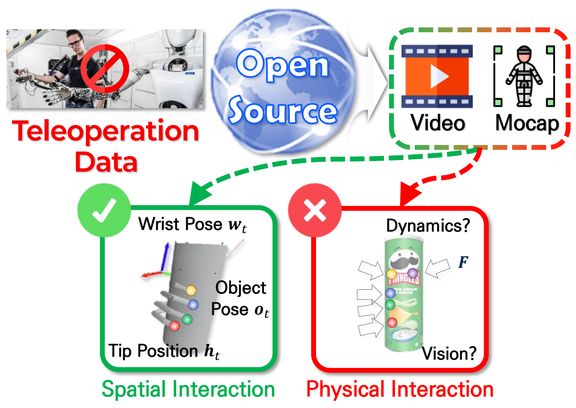
ICRA IEEE INTERNATIONAL CONFERENCE ON ROBOTICS AND AUTOMATION

MOTIVATIONS

Motivation 1

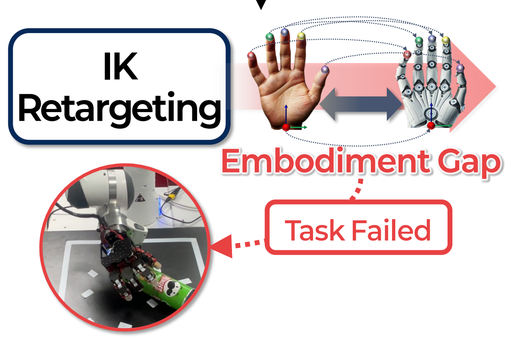
Human Videos fails to capture physical "Contact" signals.

Open Source Human Motion Data



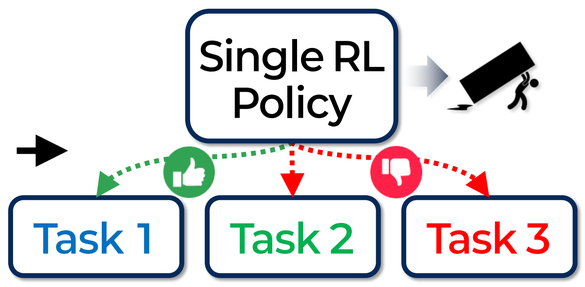
Motivation 2

Due to the "Embodiment Gap", IK retargeting fails to guarantee task success.



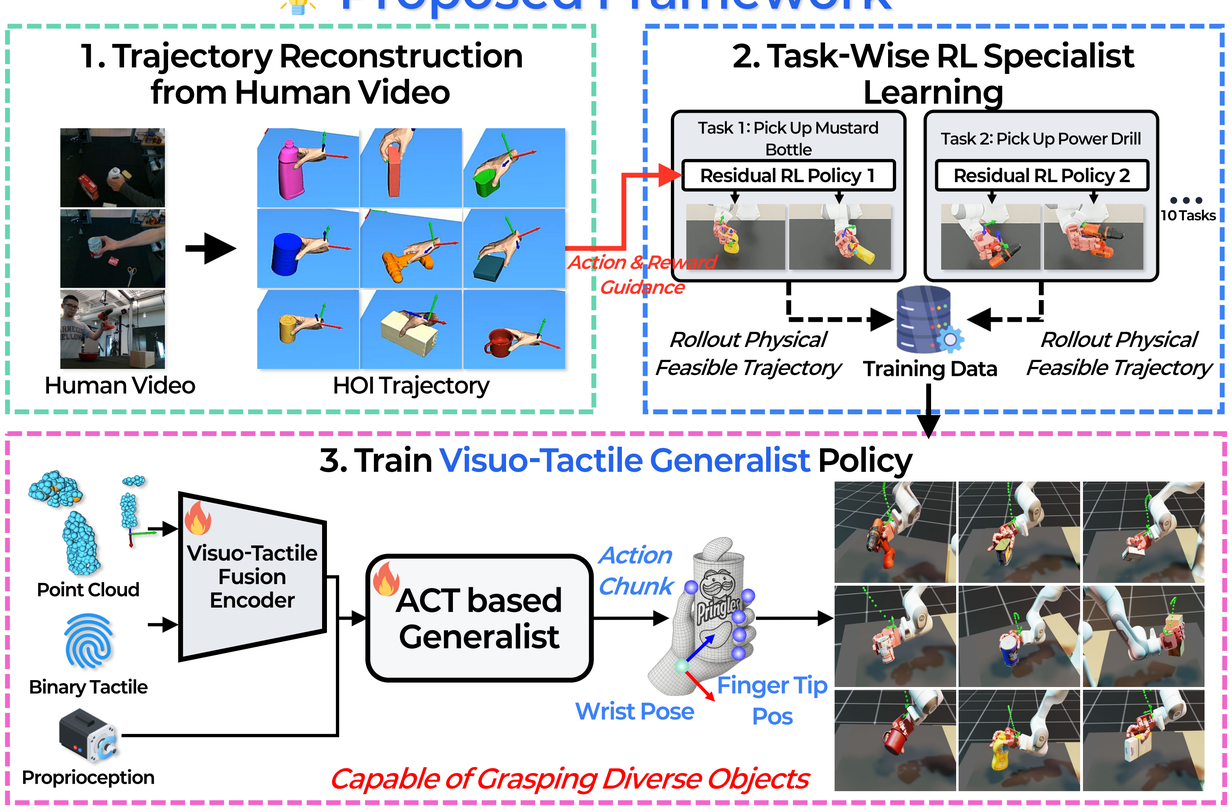
Motivation 3

RL struggles to master "diverse, complex dexterous manipulation tasks" with a single policy

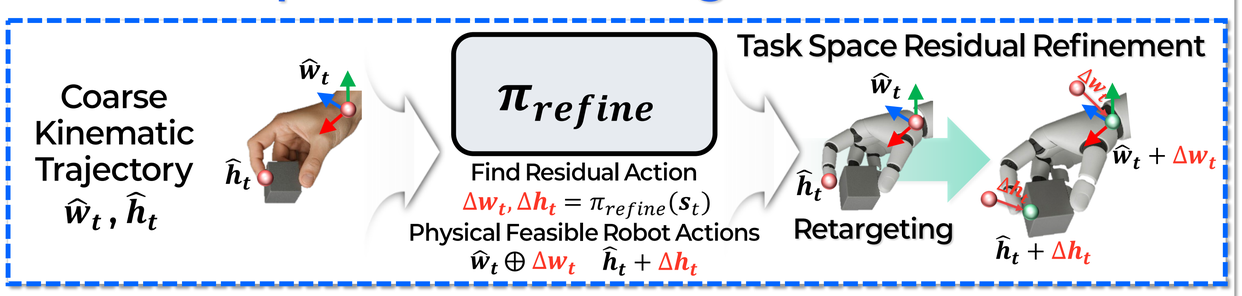


METHODS

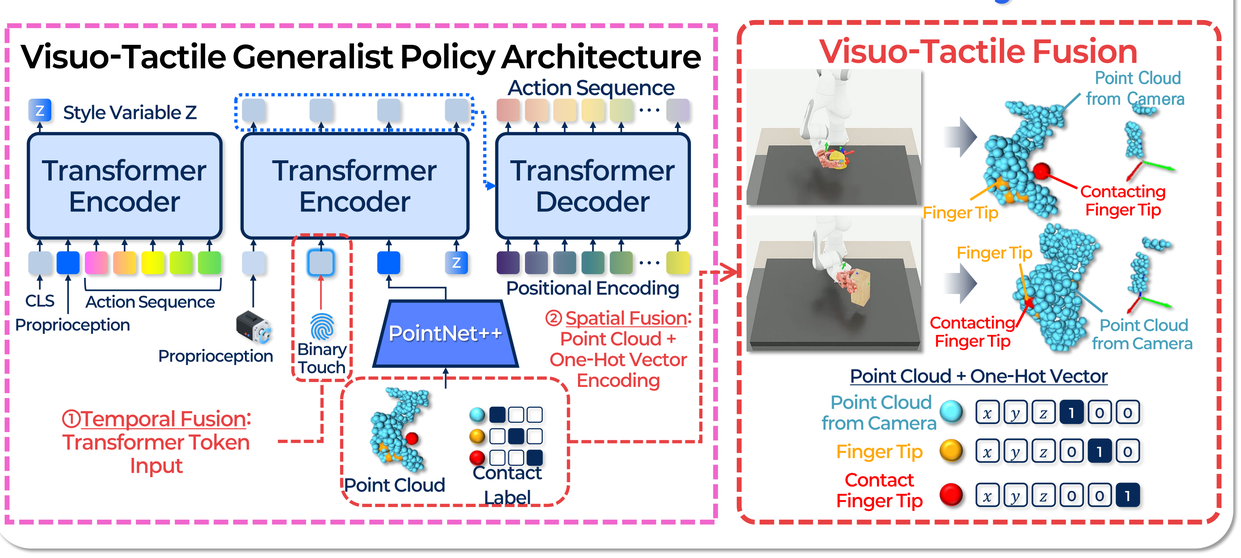
Proposed Framework



1. RL Specialist Learning via Residual RL



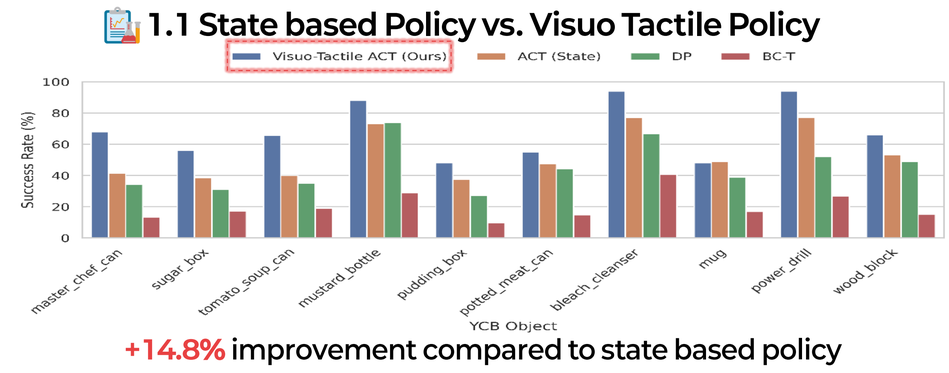
2. Visuo-Tactile Generalist Policy



EXPERIMENTS & RESULTS

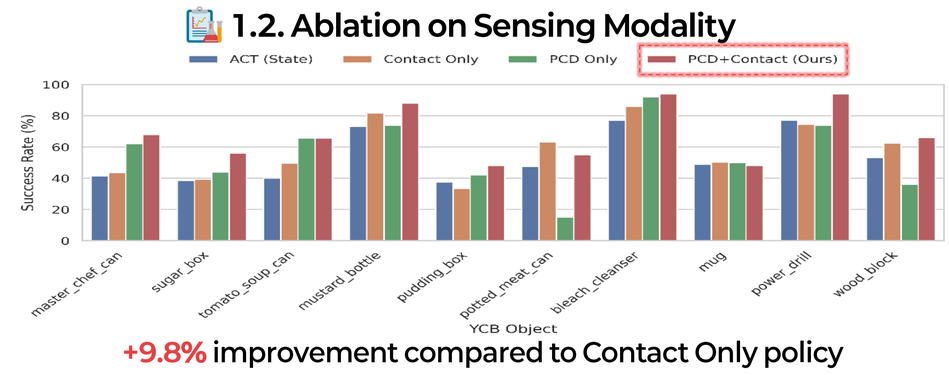
1. Ablation Study (Simulation Validation)

1.1 State based Policy vs. Visuo Tactile Policy



+14.8% improvement compared to state based policy

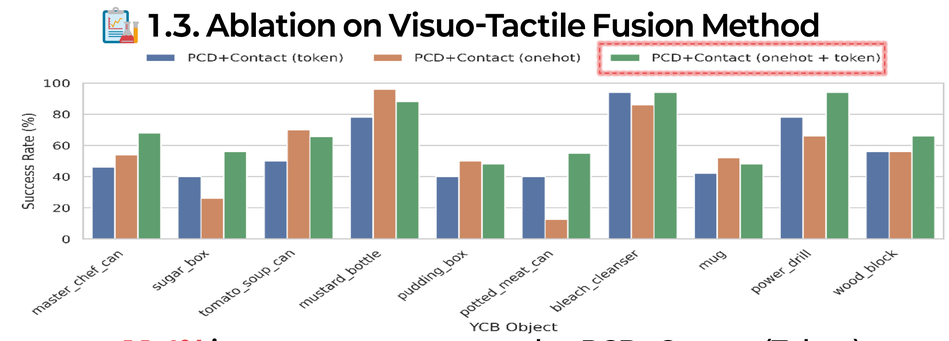
1.2. Ablation on Sensing Modality



+9.8% improvement compared to Contact Only policy

+12.8% improvement compared to Point Cloud Only policy

1.3. Ablation on Visuo-Tactile Fusion Method



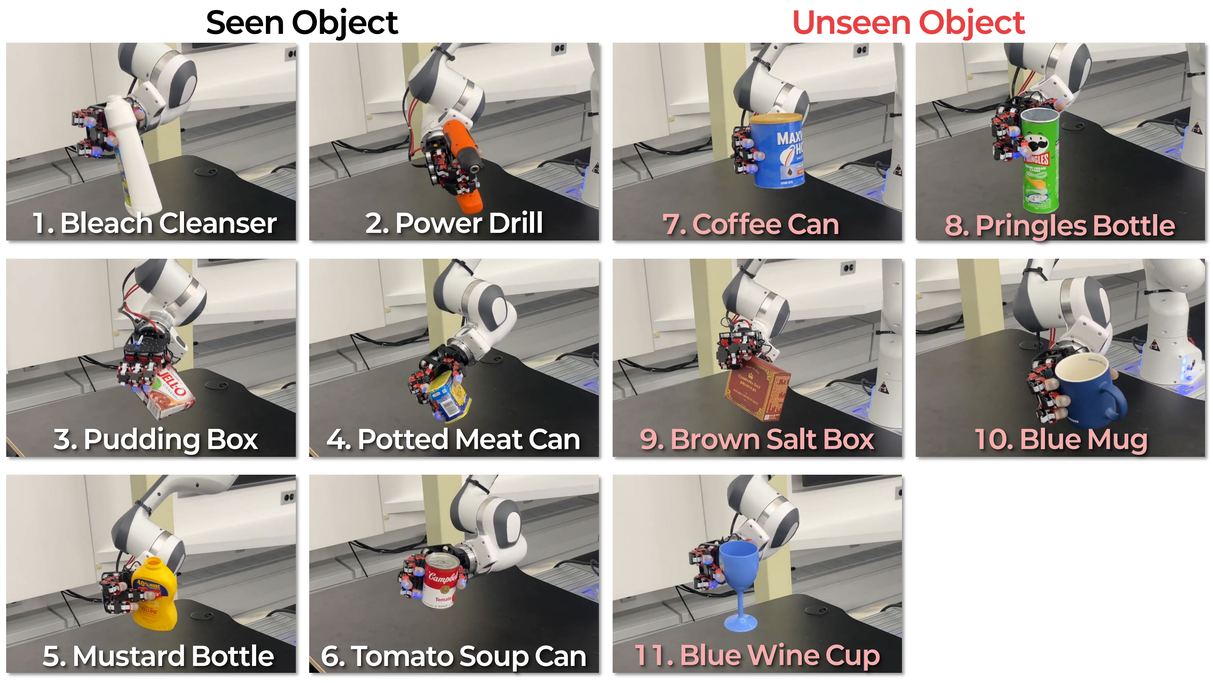
+11.4% improvement compared to PCD+Contact(Token)

+11.9% improvement compared to PCD+Contact(one-hot)

2. Real-World Validation

Our visuo-tactile policy proved particularly effective for small objects. Tactile feedback successfully mitigated visual occlusion during grasping.

Seen Object | Unseen Object



Seen Object

	Point Cloud + Binary Tactile	Point Cloud Only
Bleach Cleanser	9/10	10/10
Mustard Bottle	7/10	7/10
Power Drill	8/10	7/10
Potted Meat Can	7/10	4/10
Tomato Soup Can	5/10	3/10
Pudding Box	6/10	2/10

Small Objects

Unseen Object

	Point Cloud + Binary Tactile	Point Cloud Only
Coffee Can	9/10	9/10
Brown Salt Box	4/10	2/10
Plastic Wine Cup	6/10	7/10
Blue Mug Cup	2/10	0/10
Pringles Bottle	10/10	9/10

CONCLUSIONS

- Our framework enables task generalization across diverse objects by training a visuo-tactile policy solely on fully open-source human video data, without any teleoperated robot demonstrations.
- By fusing global 3D point clouds with local binary tactile signals, our generalist policy achieves 68.3% in simulation and 66.4% on a physical 16-DoF hand across seen and unseen objects