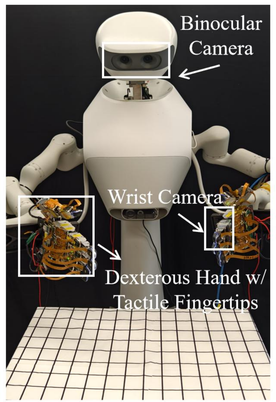
ViTacFormer: Learning Cross-Modal Representation for Visuo-Tactile Dexterous Manipulation

Liang Heng[1,2,3*], Haoran Geng[1*†], Kaifeng Zhang[3], Pieter Abbeel[1], Jitendra Malik[1]

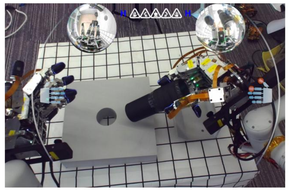
[1]University of California, Berkeley; [2]Peking University; [3]Sharpa

Hardware Setup



(a) Hardware System

(b) Human Teleoperator



(c) First-person VR View

➢ Overview of our system hardware and teleoperation setup

Motivation

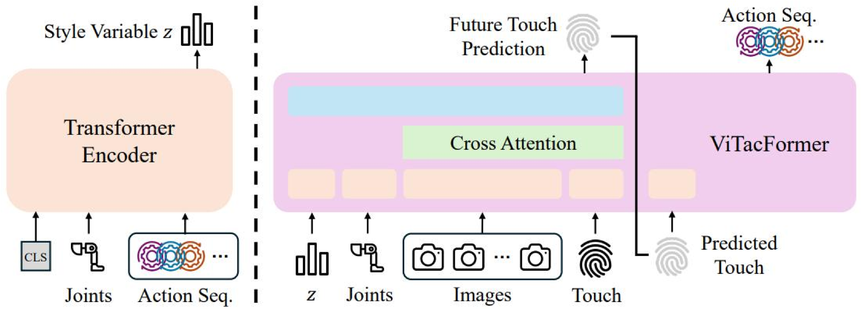
➢ *"How can we effectively incorporate **tactile sensing** into **vision-based** imitation learning policies for **dexterous manipulation**?"*

Project Page

➢ We propose **ViTacFormer**, which **fuses high-resolution vision and touch via cross-attention, and forecasts future tactile states** to stabilize cross-modal representation learning.

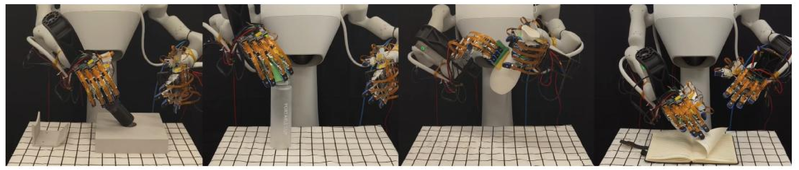
Methods

➢ **Cross-Attention Fusion:** Fuse vision and touch -> shared representation
➢ **Auto-regressive tactile prediction:** Predicts the next-step tactile signal and feeds it back
➢ **Two-stage training:** GT tactile (75%) → Predicted tactile (25%)



Short-horizon Task:

➢ **4 tasks**: *1) Peg Insertion, 2) Cap Twist, 3) Vase Wipe, 4) Book Flip*



Metrics:
① Success Rates
② Human Normalized Score

$$\text{HNS} = \frac{\sum_{i=1}^{N} w_i \cdot s_i}{3 * \sum_{i=1}^{N} w_i},$$

Task	Peg Insertion	Cap Twist	Vase Wipe	Book Flip
DP	2/10	0/10	3/10	1/10
ACT	4/10	4/10	3/10	2/10
HATO	4/10	1/10	4/10	3/10
ACTwT	6/10	6/10	4/10	4/10
Ours	**10**/10	**10**/10	**9**/10	**9**/10

Experiments

Long-horizon Task:

➢ **Make Hamburger**

Stage	1	2	3	4	5	6	7	8	9	10	11	Overall
ACT	2.4	2.5	**1.9**	**2**	0.7	2.2	1.6	**2.8**	2.2	2.2	0.7	0.61
Ours	**2.9**	**3**	**1.9**	1.8	**2.7**	**2.9**	**2**	**2.8**	**2.4**	**2.5**	**3**	**0.88**