VTDexManip: A Dataset and Benchmark for Visual-tactile Pretraining and Dexterous Manipulation with Reinforcement Learning

Qingtao Liu[1], Yu Cui[1], Zhengnan Sun[1], Gaofeng Li[1], Jiming Chen[1], Qi Ye[1*]

[1]Department of Control Science and Engineering, Zhejiang University Hangzhou, Zhejiang, China

Project page

ICLR

Motivation and Challenges

- Tactile feedback provides crucial information when vision is obstructed, **fusion of visual and tactile signals** can significantly improve robotic manipulation capabilities. While some visual-tactile datasets exist for simpler tasks, they lack data for complex multi-fingered manipulation tasks. These datasets do **not fully address the need for multi-task, complex manipulation scenarios.**
- The collection of robotic visual-tactile datasets is **costly and challenging for complex task with dexterous hands.** Leveraging human manipulation videos for robotic task pretraining has shown promise in prior works.

Solution

✓ The first visual-tactile manipulation dataset

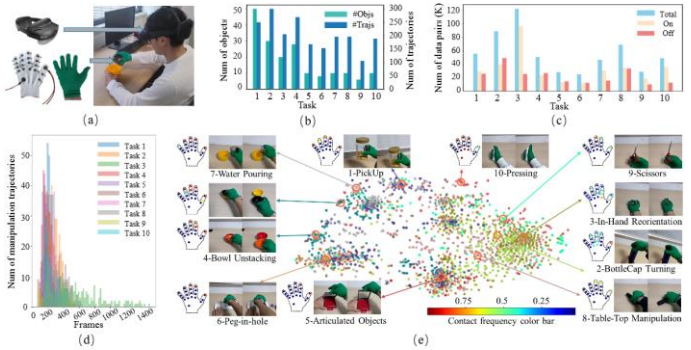
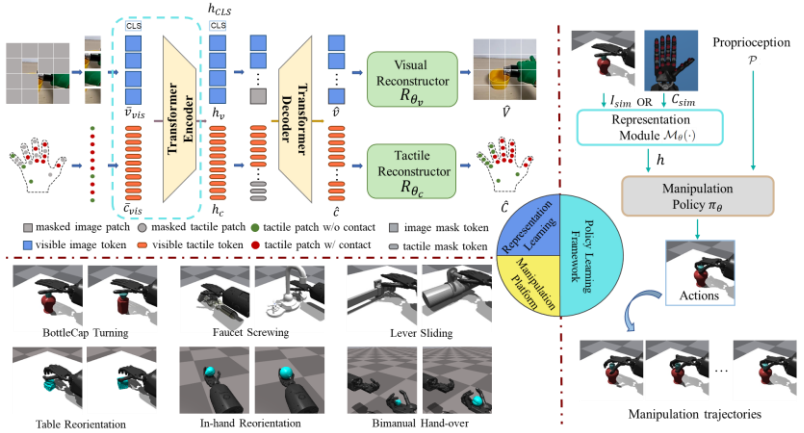


Figure 1: Visualization of our dataset. (a) Our collection system. (b) The number of trajectories and objects. (c): The number of total frames (On: frames w/ contact; Off: frames w/o contact). (d) The distribution of the number of frames. (e) t-SNE of the tactile signals.

✓ A visual-tactile pretrain network, a manipulation platform and RL learning framework



Physical System

Notes:

(a) Hardware:
1-Shadow Hand; 2-Azure Kinect camera;
3-Tactile collection board; 4-Bottle.
(b-f) Tasks: BottleCap Turning, Faucet Screwing,Lever Sliding, Table Reorientation, In-hand Reorientation.

Experiment Results

➢ Comparison with different modalities

Tasks	Split	Base	T-Pretrain	V-Pretrain	VT-JointPretrain
BottleCap Turning	Seen	55.9± 5.6	75.4± 2.9	70.8± 7.2	**83.7± 0.9**
	Unseen	36.8± 9.4	68.6± 5.6	58.5±14.2	**81.3± 0.5**
Faucet Screwing	Seen	49.0±12.0	60.0±12.3	57.9± 7.0	**80.1± 1.8**
	Unseen	43.9±10.5	51.9±12.1	51.8± 6.5	**73.6± 2.1**
Lever Sliding	Seen	5.8 ± 4.4	53.1±23.1	27.9±14.9	**89.3± 3.6**
	Unseen	2.2 ± 1.9	48.3±20.7	20.5±10.9	**79.6± 6.1**
Table Reorientation	Seen	51.8± 6.3	68.8± 1.8	74.2± 9.4	**85.0± 1.4**
	Unseen	46.7± 7.3	69.8± 2.3	69.2±10.0	**84.6± 1.1**
In-hand Reorientation	Seen	38.1± 2.4	42.1± 2.7	55.7± 1.5	**62.2± 5.0**
	Unseen	33.7± 1.6	35.8± 2.6	53.5± 1.7	**55.1± 2.7**
Bimanual Hand-over	Seen	8.0 ± 4.4	35.0±10.2	37.7±10.9	**45.5± 1.5**
	Unseen	3.3 ± 1.4	20.7± 6.0	23.1± 7.0	**26.6± 1.9**
Task Mean	Seen	34.8± 5.8	55.7± 8.8	54.0± 8.5	**72.2± 2.4**
	Unseen	27.8± 5.0	49.2± 9.4	46.1± 8.1	**66.8± 2.7**

➢ Tactile threshold setting for pertaining

	Force threshold (N) for RL training					
	0.01N		0.5N		1.0N	
	Seen	Unseen	Seen	Unseen	Seen	Unseen
0.2V (0.05N)	83.7± 0.9	81.3± 0.5	82.9± 1.2	80.6± 0.2	74.4± 4.8	65.0± 8.8
0.55V (0.5N)	80.5± 6.7	77.8± 6.3	85.3± 5.1	82.3± 2.1	82.5± 4.3	81.6± 4.2
0.75V (1.0N)	81.6± 3.5	79.3± 5.1	86.4± 3.8	84.2± 3.0	80.8± 4.4	77.8± 7.2

➢ benchmarking different methods

Method	Modality	Pretrain	Joint pretrain	Seen	Unseen
T	t	✘	-	50.8± 2.5	47.0± 2.1
V	v	✘	-	24.0± 3.0	22.2± 2.9
V+T	v+t	✘	-	23.6± 2.6	19.3± 2.9
V-MVP	v	✔	-	35.2± 2.7	29.4± 2.4
V-Voltron	v	✔	-	40.0± 1.9	31.7± 1.5
V-R3M	v	✔	-	37.0± 0.7	26.2± 2.1
V-CLIP	v	✔	-	61.3± 1.5	49.4± 1.8
V-ResNet	v	✔	-	54.1± 0.5	46.8± 0.6
V-MVP+T	v+t	✔	✘	38.5± 2.5	35.3± 2.3
V-Voltron+T	v+t	✔	✘	39.8± 2.1	34.7± 2.0
V-R3M+T	v+t	✔	✘	38.9± 2.1	31.0± 1.5
V-CLIP+T	v+t	✔	✘	65.4± 1.7	55.9± 1.7
V-ResNet+T	v+t	✔	✘	55.4± 1.9	44.1± 1.8
V-Pretrain+T-Pretrain	v+t	✔	✘	62.6± 6.3	53.3± 7.3
VT-JointPretrain	v+t	✔	✔	**74.3± 0.6**	**65.7± 0.7**

➢ Tactile noise setting for RL

	σ=0.01N		σ=0.1N		σ=1N	
	Seen	Unseen	Seen	Unseen	Seen	Unseen
v1	**60.3± 7.8**	**62.7± 6.0**	37.0± 10.9	33.7± 6.4	34.4± 10.2	33.5± 8.0
v2	**83.8± 1.4**	**81.2± 0.8**	79.1± 1.4	79.9± 2.2	41.2± 10.1	40.7± 4.8
v3	84.6± 4.6	81.7± 9.6	**87.1± 4.2**	**84.6± 7.8**	86.8± 5.5	83.8± 8.1

➢ Evaluation on different visual and tactile modalities

Method	Vision viewpoints	Tactile thresholds	Seen	Unseen
V-ego (V-Pretrain)	ego-centric	-	70.8± 7.2	58.5±14.2
V-arm	on the arm	-	58.2±16.9	58.5±17.0
V-3rd	third view	-	46.2±17.6	37.4±18.8
VT-arm	on the arm	0.01N	78.0± 4.9	73.3± 7.0
VT-3rd	third view	0.01N	82.4± 2.3	79.4± 4.2
T-1 (T-Pretrain)	-	0.01N	75.4± 2.9	68.6± 5.6
T-50	-	0.5N	60.8± 9.1	48.8±14.4
T-100	-	1.0N	64.3± 5.9	46.6± 9.7
VT-50	ego-centric	0.5N	82.9± 1.2	80.6± 0.2
VT-100	ego-centric	1.0N	74.4± 4.8	65.0± 8.8
VT-JointPretrain	ego-centric	0.01N	**83.7± 0.9**	**81.3± 0.5**

➢ Deploying single modality after joint pretraining to RL

Method	V-Pretrain	VT-JointPretrain-MaskT	T-Pretrain	VT-JointPretrain-MaskV	VT-JointPretrain
Seen	70.8± 7.2	73.3± 2.9	75.4± 2.9	72.6± 2.7	83.7± 0.9
Unseen	58.5±14.2	65.7± 5.0	68.6± 5.6	66.1± 7.0	81.3± 0.5