ViTacGen: Robotic Pushing with Vision-to-Touch Generation

Zhiyuan Wu[1], Yijiong Lin[2], Yongqiang Zhao[1], Xuyang Zhang[1], Zhuo Chen[1], Nathan Lepora[2], Shan Luo[1]

[1]King's College London, [2]University of Bristol

Background & Motivation

Robotic Pushing requires precise control and perception to handle complex object dynamics and frictional interactions.

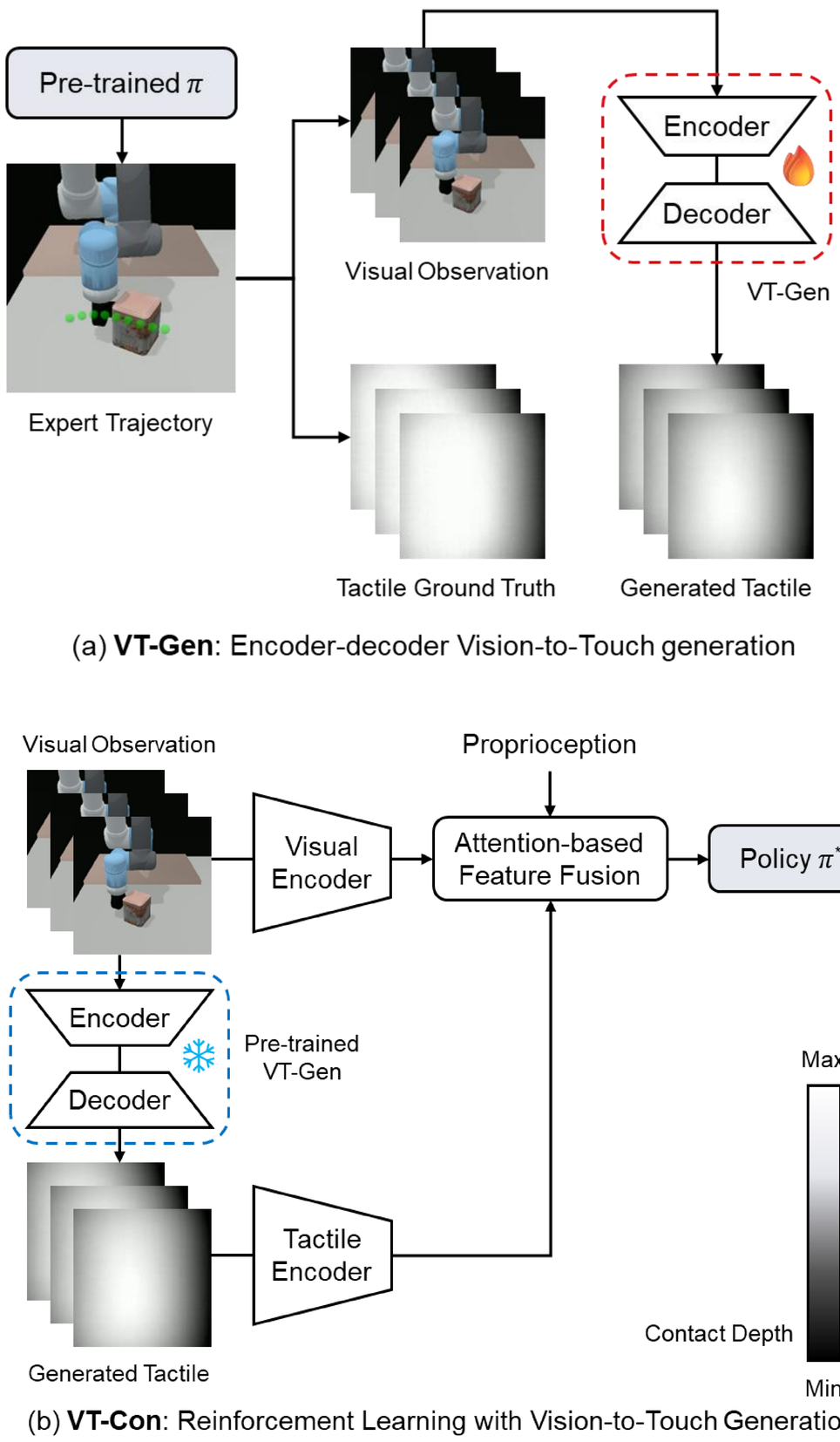
1.**Tactile sensing** in robotic pushing

2.Reliance on **high-quality** tactile sensors

3.**Manufacture inconsistency** across sensors

Using **Contact Depth** to Standardly Represent Tactile Images

Methodology



(a) **VT-Gen**: Encoder-decoder Vision-to-Touch generation

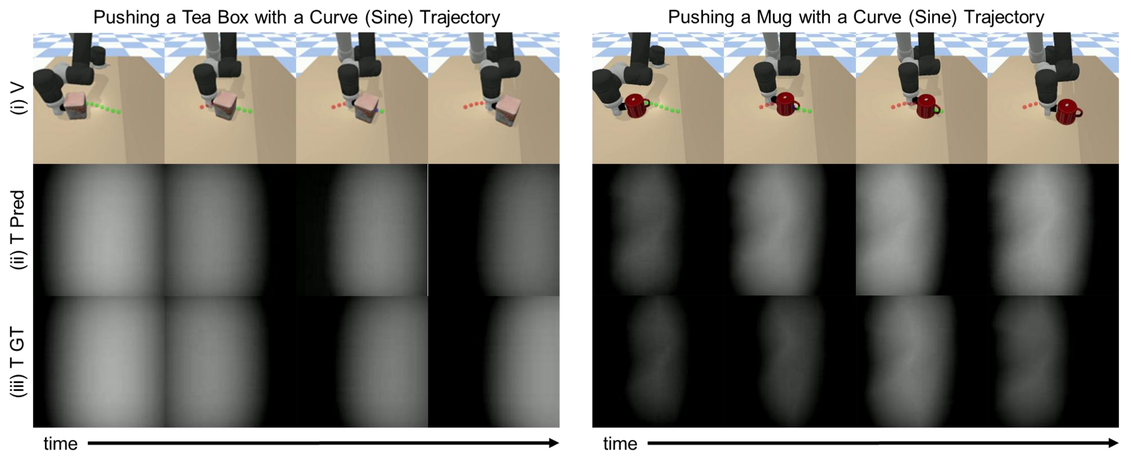
(b) **VT-Con**: Reinforcement Learning with Vision-to-Touch Generation

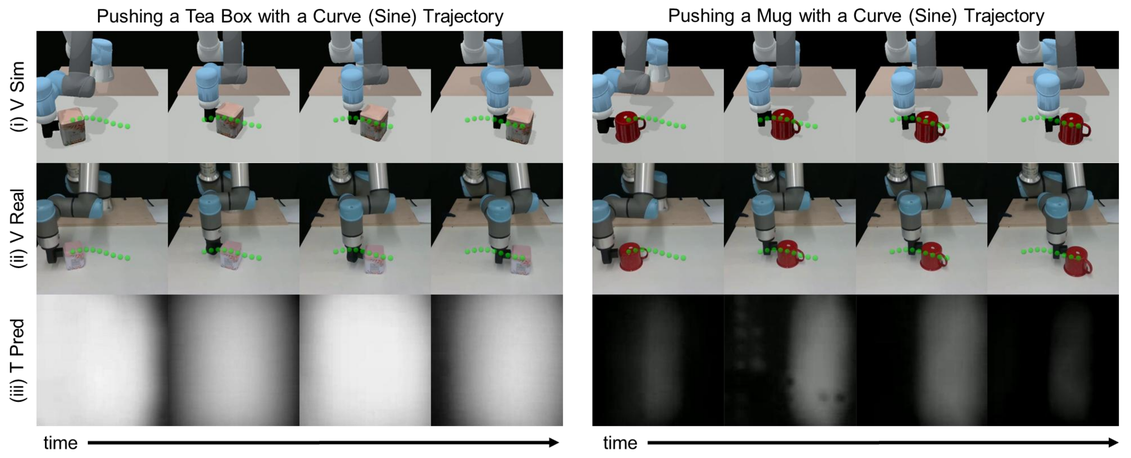
Used Objects

Training Objects

Unseen Objects

Experiments





More qualitative results are available in our website

TABLE II

QUANTITATIVE RESULTS FOR ROBOTIC PUSHING IN SIMULATION. WE COMPARE TWO VERSIONS OF OUR VITACGEN FRAMEWORK: ONE WITH VISUAL-ONLY INPUTS AND THE OTHER WITH VISUAL & ACTUAL TACTILE DATA FOR REFERENCE. THESE ARE EVALUATED AGAINST BASELINES INCLUDING VISUAL-ONLY, TACTILE-ONLY, AND COMBINED VISUAL-TACTILE APPROACHES [9]. WE CONDUCT 100 TRIALS FOR EACH EXPERIMENTAL CONDITION. WE REPORT CUMULATIVE REWARDS, EPISODE LENGTH, DISTANCE ERROR, AND SUCCESS RATE WITH A THRESHOLD OF 2.5 CM. THE BEST RESULTS ARE HIGHLIGHTED IN RED, WHILE THE SECOND-BEST RESULTS ARE HIGHLIGHTED IN BLUE.

Object Type	Method	Rewards (mean±std) ↑	Epi. Len. (mean±std) ↓	Dist. Err. (mm) ↓	Succ. Rate (%) ↑
Tea Box	Visual only	-155.73±63.06	335.93±38.15	47.54	12.0
	Tactile only	-150.10±62.03	337.10±36.72	50.62	11.0
	Visual & Tactile	-147.01±72.21	334.80±39.34	50.02	13.0
	Ours (Visual only)	-84.35±79.03	267.24±51.90	27.81	84.0
	Ours (Visual & Tactile)	-44.83±22.64	268.57±30.20	23.08	92.0
Meat Can	Visual only	-169.27±80.65	330.02±45.84	62.23	16.0
	Tactile only	-149.75±17.06	324.44±45.39	53.77	25.0
	Visual & Tactile	-125.72±79.06	313.98±55.18	53.75	30.0
	Ours (Visual only)	-88.19±79.86	250.90±48.50	29.81	81.0
	Ours (Visual & Tactile)	-44.49±20.96	251.66±41.86	30.38	86.0
Mug	Visual only	-112.65±52.30	328.78±43.09	54.85	20.0
	Tactile only	-106.35±50.17	317.51±49.73	47.19	31.0
	Visual & Tactile	-106.01±50.06	311.20±51.95	42.83	38.0
	Ours (Visual only)	-41.53±19.72	266.03±43.91	27.39	86.0
	Ours (Visual & Tactile)	-34.92±13.86	270.99±38.21	27.07	83.0

TABLE I

QUANTITATIVE RESULTS FOR VISION-TO-TOUCH GENERATION IN RL SETTING. WE REPORT PSNR, SSIM [36], AND LPIPS [37].

Object Type	PSNR ↑	SSIM ↑	LPIPS ↓
Tea Box	30.75	0.9482	0.0101
Meat Box	20.50	0.8657	0.0327
Mug	20.25	0.8222	0.0417

TABLE III

QUANTITATIVE RESULTS FOR ROBOTIC PUSHING IN REAL WORLD. SINCE OUR VITACGEN IN REAL WORLD SETTING IS DEPLOYED ON A VISUAL-ONLY ROBOTIC SYSTEM, WE COMPARE IT AGAINST VISUAL-ONLY INPUT BASELINE IN [9]. WE CONDUCT 50 TRIALS FOR EACH EXPERIMENTAL CONDITION AND MANUALLY MEASURE AVERAGE DISTANCE ERROR AND SUCCESS RATE WITH A THRESHOLD OF 2.5 CM.

Object Type	Method	Dist. Err. (cm) ↓	Succ. Rate (%) ↑
Tea Box	baseline	6.5±1.9	14.0
	Ours	**2.6±0.8**	**76.0**
Meat Can	baseline	8.2±1.9	8.0
	Ours	**1.9±0.5**	**82.0**
Mug	baseline	7.2±2.1	10.0
	Ours	**1.8±0.7**	**86.0**

TABLE IV

QUANTITATIVE RESULTS FOR PUSHING UNSEEN OBJECTS IN REAL WORLD. WE CONDUCT 50 TRIALS FOR EACH EXPERIMENTAL CONDITION AND MANUALLY MEASURE AVERAGE DISTANCE ERROR AND SUCCESS RATE WITH A THRESHOLD OF 4.0 CM.

Object Type	Dist. Err. (cm) ↓	Succ. Rate (%) ↑
Apple	3.3±1.1	82.0
Coffee Box	4.1±2.0	72.0
Ceramic Cup	3.9±1.8	74.0
Olive Jar	3.7±1.7	78.0
Soup Can	4.1±1.7	70.0

Contributions

- We propose ViTacGen, a novel robot manipulation framework designed for visual robotic pushing with vision-to-touch generation in RL to **eliminate the reliance on high-resolution real tactile sensors**, enabling effective zero-shot deployment on visual-only robotic systems.
- We introduce an encoder-decoder vision-to-touch generation network VT-Gen that generates **contact depth images as a standardized tactile representation** directly from visual image sequence, with an RL network VT-Con that learns robust policies using feature fusion and contrastive learning on visual and generated tactile observations.
- We demonstrate the effectiveness of our proposed methods in **both simulation and real world**, through extensive qualitative and quantitative experiments, supported by comprehensive ablation studies.