Design Decisions that Matter: Modality, State, and Action Horizon in Imitation Learning

Brendan Chharawala, Joshua Li, Stephie Liu, Shawn Yang, Colin Bellinger, David Liu, Chang Shu, Yue Hu, Pengcheng Xi

1. Problem definition

➢ Role of **teleoperation modality** in shaping *demonstration quality* and *downstream robot imitation learning performance* is still poorly understood

2. Contributions

➢ Comparative dataset of assistive task demonstrations (**VR controller and haptic pen**) paired with **NASA-TLX subjective workload measures**
➢ Analysis on the impact of teleoperation modality on demonstration quality and imitation learning model (Octo) finetuning performance
➢ Exploration on effects of data-related **fine-tuning design choices** (robot states and action horizon) on real world robot performance

3. Methodology

➢ Tasks on an UR10e robotic arm: (i) wipe table surface and (ii) turn desk lamp on/off
  ○ Teleoperation modalities per task: (i) VR controller and (ii) haptic 3D pen input
➢ 5 participants (20 episodes per modality per task, total 400 episodes)

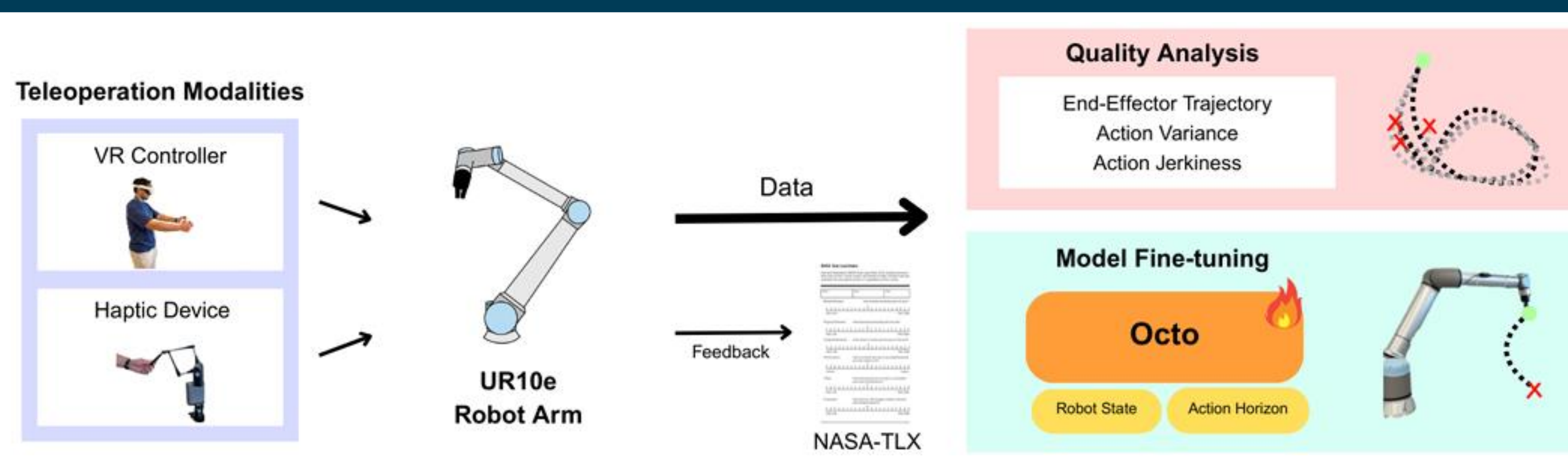
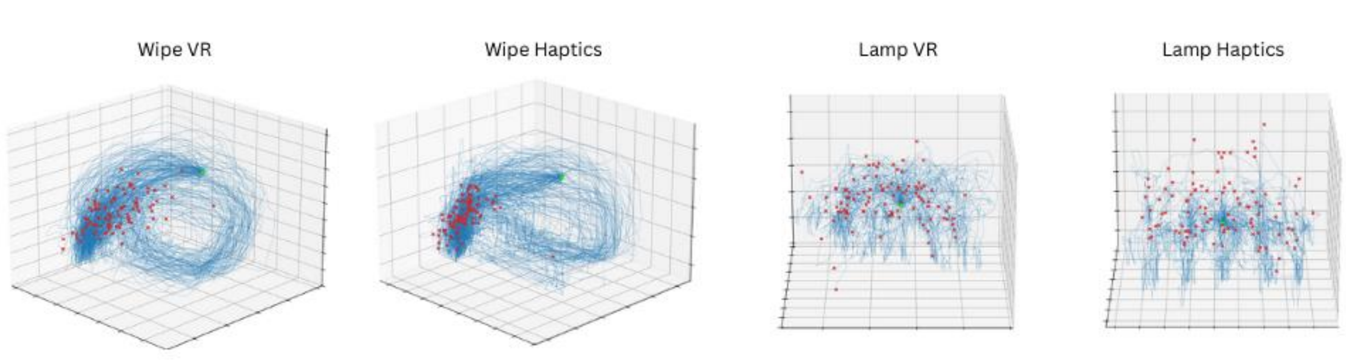


Fig. 1. Overview of data collection and learning pipeline.

➢ NASA-TLX survey: subjective metrics affecting teleoperation usability
➢ Data quality analysis metrics: measure smoothness and control precision with *end-effector trajectories, action variance, and jerkiness*
➢ Finetuned Octo policy to assess how different input modalities influence learning performance

4. Results



	VR	Haptics
Wipe	0.00057	0.00016
Lamp	0.00015	0.00006

	VR	Haptics
Wipe	0.23 ± 0.33	0.12 ± 0.14
Lamp	0.12 ± 0.17	0.08 ± 0.08

Fig. 2. Top: end-effector trajectories. Middle left: action variance. Middle right: jerkiness (mean ± std).

➢ **Data quality:** VR provides higher quality data for **broad task actions**, haptic is better for **precise tasks**
  ○ Generally, VR has higher action variance and jerkiness (less stability compared to haptic)
➢ **NASA-TLX:** VR supports **scalable data collection and ease of use**, haptics for **high-fidelity data and better performance**

Task	Mental Dem.	Physical Dem.	Temporal Dem.	Perf.	Effort	Frustr.
Wipe	10.2	8.4	8.6	10.0	9.6	6.3
Lamp	8.7	8.2	8.1	7.4	8.7	7.5

Col. Method	Mental Dem.	Physical Dem.	Temporal Dem.	Perf.	Effort	Frustr.
Haptics	10.0	10.0	9.0	7.8	9.2	6.2
VR	8.9	6.6	7.7	9.6	9.1	7.6

Fig. 3. Top: NASA-TLX results per task (lower = better). Bottom: NASA-TLX results per modality (lower = better)

5. Finetuning Design Discussion

Configuration	Wipe Success (%)	Wipe Pose Err (cm)	Lamp Success (%)	Lamp Pose Err (cm)
Mixed, AH 10	**73**	**3.4**	**80**	**2.0**
VR, AH 10	47	4.7	53	3.4
Haptic, AH 10	40	4.7	53	3.2
Mixed, AH 15	27	4.6	47	4.6
Mixed, AH 5	20	10.5	33	5.3
Mixed, AH 10, P	20	9.7	13	$3.7^{\dagger}$

Fig. 4. Success rate (%) and pose alignment error (cm). AH = action horizon, P = proprioception included.

➢ **Finetuned highest success rate: mixed, no robot proprioception, action horizon = 10**
➢ Excluding robot state input may enhance performance
➢ Action horizon has an optimal value (e.g. 10)
➢ Camera setup and lighting affect performance significantly

Ackowledgement

This work is supported by the Aging in Place Challenge Program at the National Research Council of Canada.

ARC

National Research Council Canada

Conseil national de recherches Canada

Canada