Scale Your Dataset Without Robot

Jiwon Kim* [1] Kyungzun Rim* [1] Ilkwon Hong [1] Suhyun Yoon [1]

* These authors contributed equally to this work.

[1]Robotics LAB, Hyundai Motor Company

Abstract

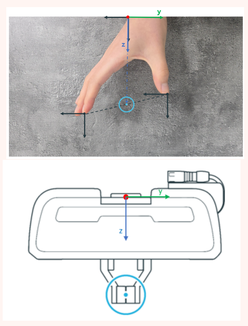
Imitation learning for robotic manipulation requires extensive demonstration data, yet traditional teleoperation methods are time-consuming, physically constrained, and produce biased datasets. We present a **novel VR-based data collection pipeline** that addresses these limitations by capturing natural hand demonstrations without robot control. Our approach transforms VR-tracked hand poses into robot-executable trajectories through automated post-processing. Our results demonstrate that **VR-based hand demonstrations** provide an accessible, efficient solution for scaling robot learning datasets while improving policy generalization and task performance.

Contribution

- **Bridging Gaps Between Human Demonstrations and Robot Movements:** We introduce a transformation pipeline that converts hand poses into gripper poses relative to the robot's base frame. This enables our approach to be applied across diverse robot types, from manipulators to humanoids.
- **Complementary Learning by Leveraging Cross-domain Data:** Human demonstrations and teleoperation datasets differ significantly in aspects such as motion speed, object placement, and camera viewpoint. By merging them into a unified dataset, we obtain policies with substantially improved performance. Notably, the weaknesses of each policy are compensated for when the datasets are combined.
- **Time-efficient and Space-free:** Our method shortens the episode length for data collection by more than 50%. Since only a VR device is required, no additional space or specialized hardware is needed.

Collection of Hand Pose Data

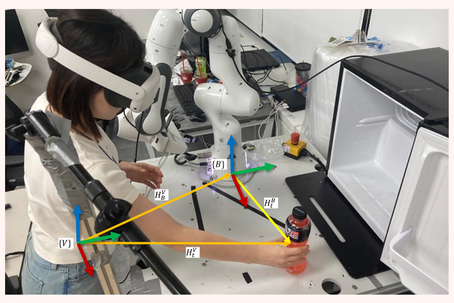
- Data Collection Process.
  1. **Robot base frame establishment** in the VR coordinate system
     - **Acquisition of specific H/W points** whose geometric relationship to the origin of the robot base frame is known
     - **Robot base frame pose estimation** in the VR coordinate system
  2. **Episode dataset collection**
     - Pinch gestures are used to signal the start and end of an episode
- Hand Pose Definition.



  - Tracked hand bones mapped to single hand frame H
  - **Position:** Origin set at wrist position (high tracking confidence)
  - **Rotation:** Axes defined by fingertip positions, emulating parallel-jaw gripper pose
  - **Task flexibility:** Frame definition adaptable to different robot/end-effector kinematics
  - See figure for an example definition of a hand frame.
    - Top: Raw hand-tracking outputs (dark gray) and defined hand frame (RGB)
    - Bottom: robot reference frame (Franka Hand)

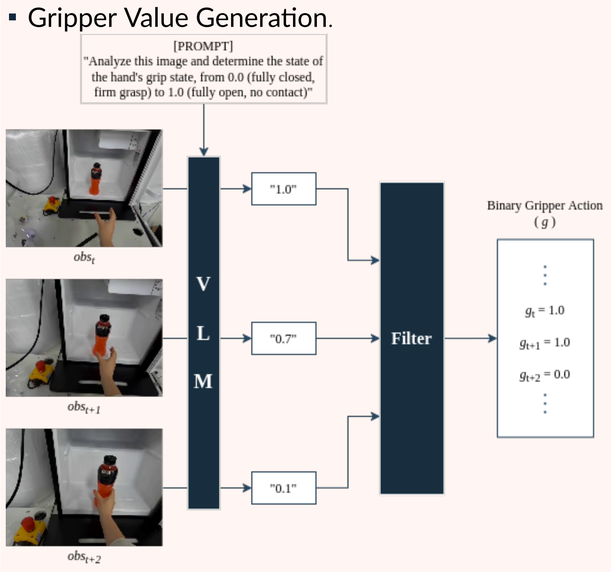
Post-Processing of Data

- Frame Transformation.

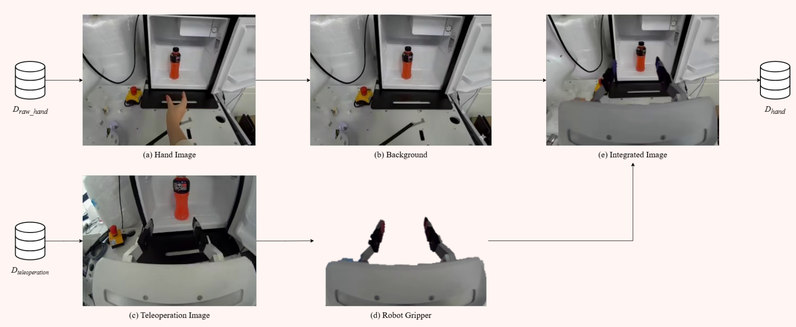


$$ {}^{B}\mathbf{H}_{t}{}' = {}^{V}\mathbf{H}^{B^{-1}} \times {}^{V}\mathbf{H}_{t} \quad (1) $$

  - ${}^{B}\mathbf{H}_{t}{}'$: Pose of the end-effector at time t
  - ${}^{V}\mathbf{H}^{B}$: Homogeneous transform to the robot base frame B, both from the VR frame V
  - ${}^{V}\mathbf{H}_{t}$: Homogeneous transform to the hand at time t (wrist pose)
- Gripper Value Generation.

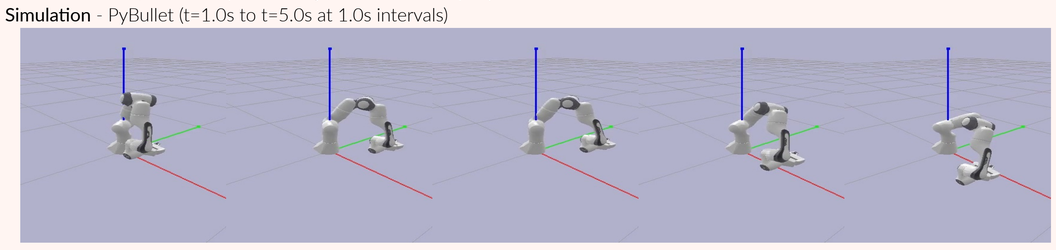


  - **Vision Language Model** [1] for hand state classification
    - Open / Closed Grip Detection
  - **Robustness improvement** against VLM prediction errors
    - Moving average over N frames to decide the state transition
    - Filter based on the velocity of hand movement under the assumption that holding or releasing of an object can only happen when it is sufficiently slow
  - Result: **95 % accuracy** in gripper states estimation
- Image Modification

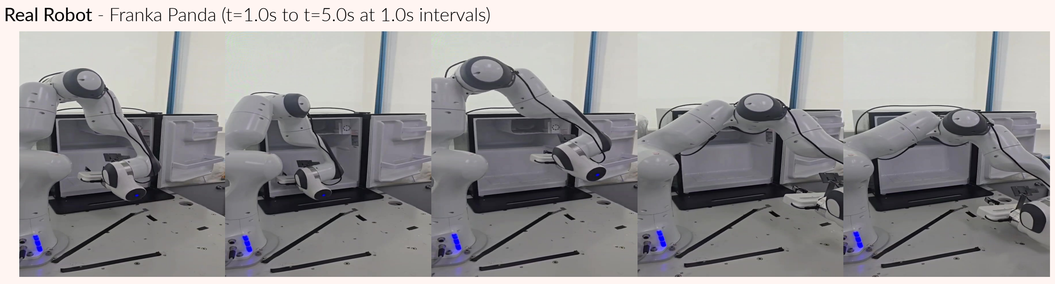


  - **Bridging the visual gap** through utilization of existing teleoperation dataset.
  - **Modification of collected human demonstration images** via a three-stage process:.
    1. Human hand removal from demonstration images using generative inpainting.
    2. Segmentation of the robot gripper from teleoperation images.
    3. Composition of the extracted gripper onto the inpainted background.
- Joint Pose Generation
  - **Inverse Kinematics**: Without actual robot execution, we compute joint configurations that minimize trajectory smoothness while avoiding singularities through null-space optimization.

- For verification, the computed trajectories are replayed on sim/real robots.
  - **Simulation** - PyBullet (t=1.0s to t=5.0s at 1.0s intervals)

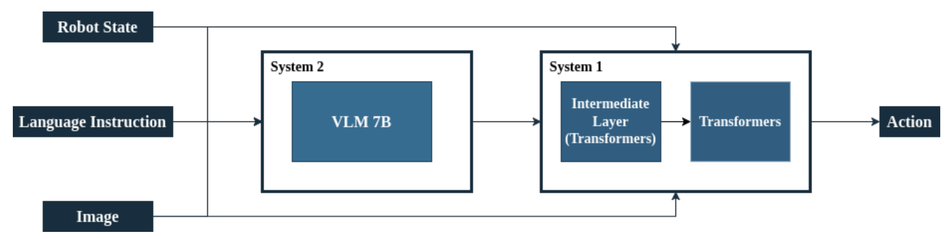


  - **Real Robot** - Franka Panda (t=1.0s to t=5.0s at 1.0s intervals)



Policy Training

- **Dataset Scale:** 46 episodes per task—insufficient for common VLA baselines (SmolVLM, π_0)
- **Custom VLA:** Developed a sample-efficient dual-system model inspired by [2]



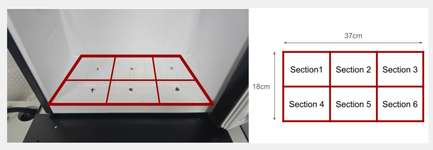
(a) Custom VLA Architecture

Experiments

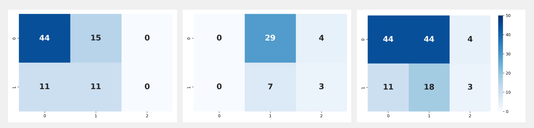
- Experimental Setup
  - Meta Quest 3
  - Deployed on a Franka Panda, a 7-DoF manipulator mounted on a tabletop setup.
  - NVIDIA H100 GPU for training, A6000 for inference.
- Dataset Analysis
  - 3 types of Datasets
    - Dataset collected via teleoperation
    - Dataset using hand poses
    - Merged datasets (dataset collected via teleoperation + dataset using hand poses)
  - Dataset Episode Length

Dataset	# Episodes	Avg. Episode Length (s)
Teleoperation	46	17.13
Hand	83	5.88
Merged	129	9.89

    - Hand demonstrations are **3× faster** to collect than teleoperation (5.88s vs 17.13s)
  - Dataset Distribution Over Sections



(a) Definition of Target Object Section



(b) Hand (c) Teleoperation (d) Merged

- Policy Performance
  - Improved policy performance with hand-collected data attributed to two main factors:
    1. **Expanded spatial coverage:** The hand dataset, collected without robot constraints, covered underrepresented areas.
    2. **Enhanced motion dynamics:** The hand dataset provides larger, more decisive movements that are difficult to achieve through teleoperation.

Table 1. Success Rate of Policies Trained on Different Datasets

	1 Image Input			2 Image Input
	Teleoperation	Hand	Teleoperation + Hand (expanded dataset)	Teleoperation
Section 1(5)	0/5(0%)	0/5(0%)	**3/5(60%)**	2/5(40%)
Section 2(5)	0/5(0%)	0/5(0%)	**4/5(80%)**	5/5(100%)
Section 3(5)	0/5(0%)	0/5(0%)	**2/5(40%)**	3/5(60%)
Section 4(5)	0/5(0%)	0/5(0%)	**3/5(60%)**	0/5(0%)
Section 5(5)	3/5(60%)	0/5(0%)	**4/5(80%)**	4/5(80%)
Section 6(5)	3/5(60%)	0/5(0%)	**3/5(60%)**	5/5(100%)
Success Rate	6/30(20%)	0/30 (0 %)	**19/30(63%)**	19/30(63%)

Note: A policy trained with an additional image input (third-eye view) provided as a reference.

Future Work

- **Robust Hand Tracking:** Current VR tracking fails under occlusion and low-light conditions. Future work will generate smooth trajectories from sparse high-confidence poses.
- **Visual Domain Adaptation:** Our gripper replacement method uses fixed views from teleoperation data, limiting visual realism. Future work will dynamically render 3D gripper meshes based on hand pose and address camera perspective differences between VR and robot-mounted camera.
- **Improved Pose Mapping:** Replace naive fingertip-based rotation with more reliable skeletal features for accurate hand-to-gripper pose conversion.

References

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Figure AI. Helix: A vision-language-action model for generalist humanoid control, 2025.

[3] A. Iyer, Z. Peng, Y. Dai, I. Guzey, S. Haldar, S. Chintala, and L. Pinto. Open teach: A versatile teleoperation system for robotic manipulation. *arXiv preprint arXiv:2403.07870*, 2024.

[4] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.