LaVA-Man: Learning Visual Action Representations for Robot Manipulation

Chaoran Zhu[1], Hengyi Wang[2], Yik Lung Pang[1], Changjae Oh[1]
[1]Queen Mary University of London [2]University Colleague London

CoRL Conference on Robot Learning, 2025

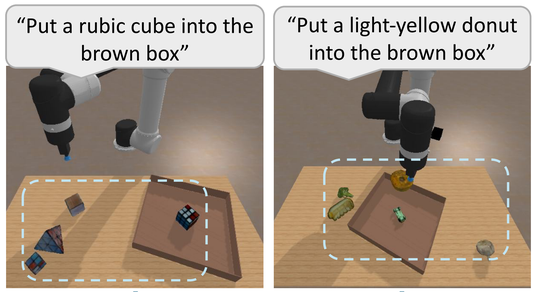
Project Website

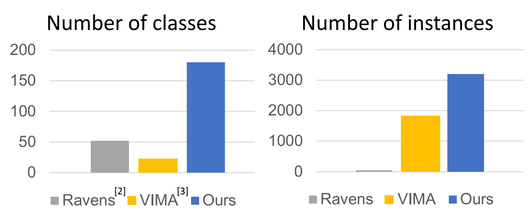
1. Core Contributions

- **Self-supervised pre-training** for learning visual-action representations, enabling few-shot robotic task adaptation.
- **New simulated dataset** (OOPP dataset) based on existing benchmarks [1] [2] with 3,200 real-scanned objects and 180 categories in full robot trajectories.

2. Proposed Dataset

Omni-Object Pick-and-Place (OOPP) dataset

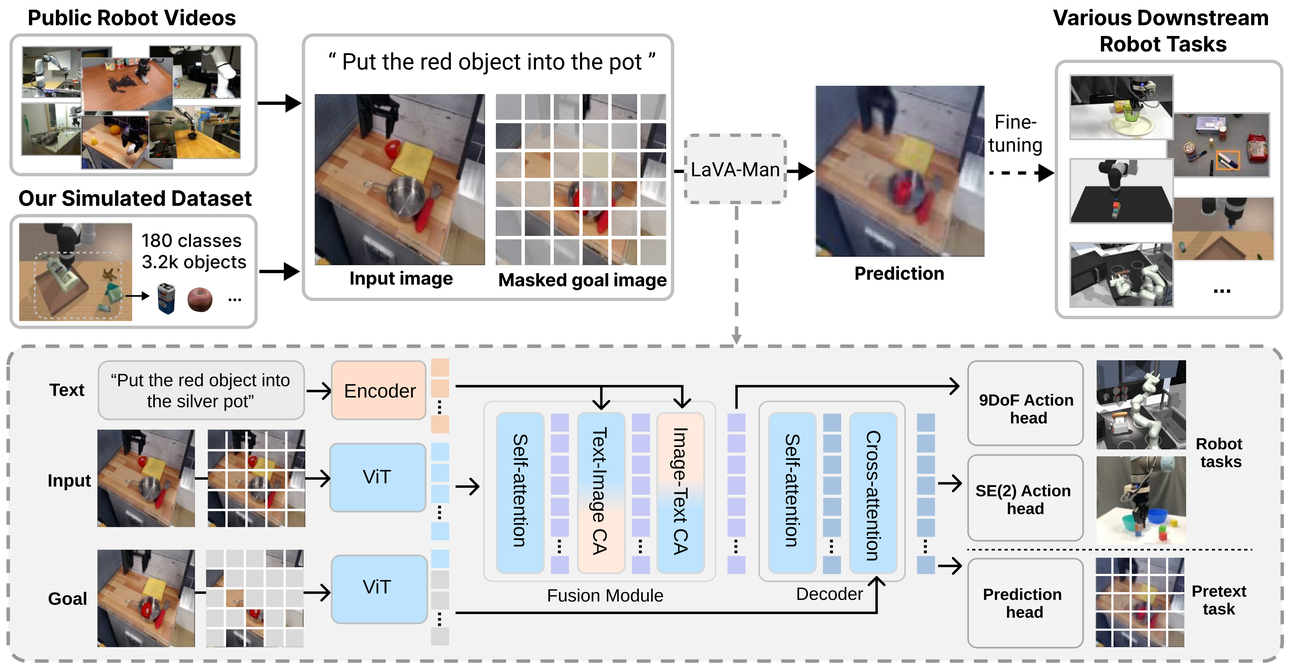




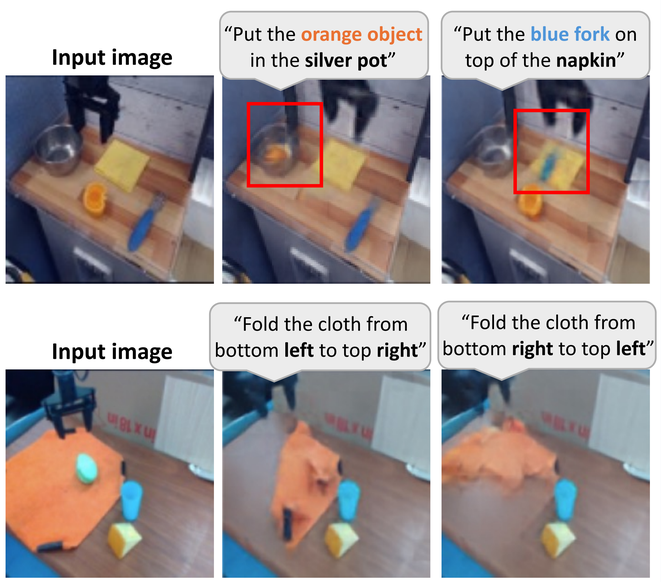
Reference

[1] T. Wu et al. OmniObject3D: Large-vocabulary 3D object dataset for realistic perception, reconstruction and generation. CVPR 2023.
[2] A. Zeng et al. Transporter networks: Rearranging the visual world for robotic manipulation. CoRL 2021.
[3] Y. Jiang et al. VIMA: Robot manipulation with multimodal prompts. ICML 2023.
[4] H. Walker et al. BridgeData V2: A dataset for robot learning at scale. CoRL 2023.
[5] A. Gupta et al. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning. CoRL 2020.

3. Self-supervised Pre-training

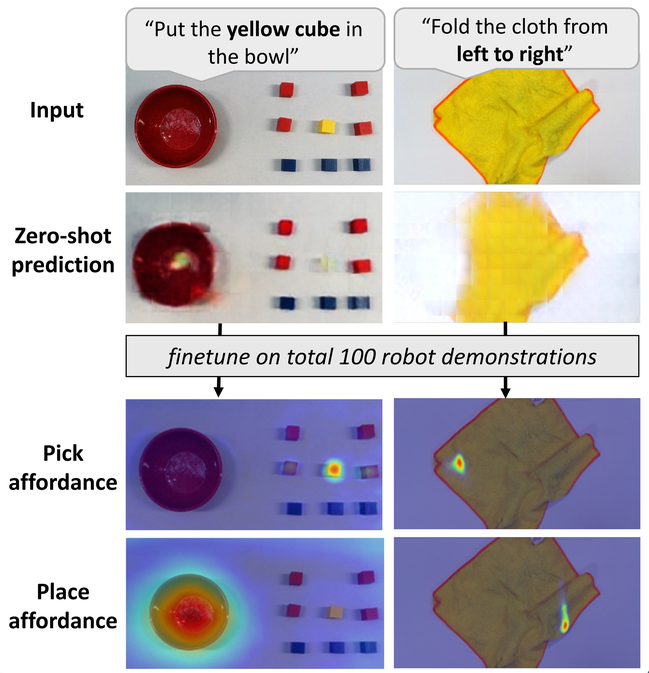


4. Predictions by Pre-trained model



- Handling fine-grained language commands
- Generalizing to unseen examples in the validation set [4]

5. From Predictions to Robot Affordance



6. Downstream Tasks