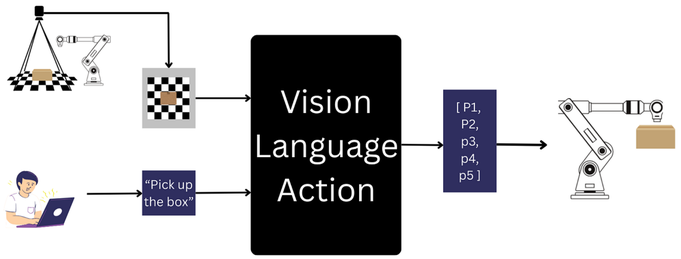
Human-Hand-Aware Vision-Language-Action Models for Safer Construction HRC

Cynthia Zhang[1], Zhaofeng Hu[2], Ci-Jyun Polar Liang[2]
[1]The Wheeler School; [2]Dept. of Civil Engineering, Stony Brook University

CROSS LAB

Stony Brook University

BACKGROUND & MOTIVATION



VLA Architecture (source: Medium)

Vision-Language-Action (VLA) models are a promising approach for **manipulation and human-robot collaboration (HRC)** in **construction,** because they can perform diverse tasks from visual observations and language instructions. However, during HRC, **workers' hands** may enter the robot workspace, creating potential collision risks. Existing safety approaches often rely on **separate hand detectors or additional safety modules**, which leads to our key question:

Does a pretrained VLA already possess human-hand awareness, and can this capability be directly leveraged for safer control?

Research Goal: Enable **Stop → Hold → Resume** behavior while preserving the VLA's original task capability.

VLA for HRC in construction (source: CROSSLAB)

HandVerse-3D SAFETY DATASET

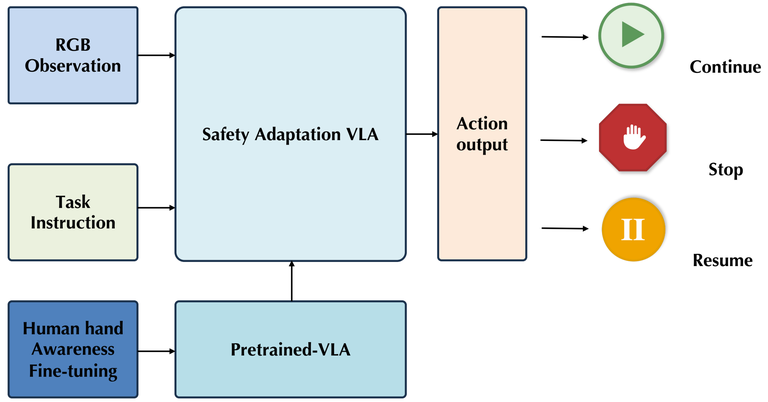
We introduce **HandVerse-3D**, a scalable pipeline for generating **7 categories**, **100+ diverse 3D hand models** across broad appearances and use cases.

Categories: Basic human hands, complex appearances, construction gloves, construction equipment, construction tools, poses, special

Dataset generation pipeline: Language Prompt → Nano Banana 2 → Hand Image → Hunyuan3D → 3D Hand Model

Method

- Inspired by methods that use targeted interventions to alter robot behavior, we **reverse the idea and repurpose intervention-based adaptation for safety**.
- **Observation + Instruction → Pretrained VLA → Safety Adaptation → Action**
- Desired behavior: **Continue / Stop / Resume**

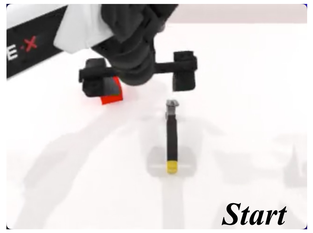


The Framework of Safety

RESULTS

Experiments

Trained a Pi0.5 with video dataset collection of 500 videos by Full Fine-tuning in RoboTwin. Then Model evaluation on the seen/unseen per 20 episodes:



Scenarios	Stop Rate	Task Success Rate
Seen	74%	42%
Unseen	68%	34%

CONCLUSION

- Pretrained VLA models already exhibit human-hand awareness, enabling meaningful stop-and-resume behavior without a separate hand detector.
- HandVerse-3D provides a scalable way to introduce diverse hand appearances and improve robustness to unseen hands.

FUTURE WORK

- Improve **stop reliability, resume behavior, and task-success preservation**.
- Expand to **more construction tasks, worker appearances, and real-robot HRC scenarios**.

References:

Kim, Moo Jin, et al. "Openvla: An open-source vision-language-action model." *arXiv preprint arXiv:2406.09246* (2024).
Black, Kevin, et al. "$\pi_0 $: A Vision-Language-Action Flow Model for General Robot Control." *arXiv preprint arXiv:2410.24164* (2024).
Intelligence, Physical, et al. "$\pi_ {0.5} $: a Vision-Language-Action Model with Open-World Generalization." *arXiv preprint arXiv:2504.16054* (2025).
Yang, Xianghui, et al. "Hunyuan3d 1.0: A unified framework for text-to-3d and image-to-3d generation." arXiv preprint arXiv:2411.02293 (2024).
Chen, Tianxing, et al. "Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation."-
- arXiv preprint arXiv:2506.18088 (2025).
Wang, Xin, et al. "Freezevla: Action-freezing attacks against vision-language-action models." arXiv preprint arXiv:2509.19870 (2025).