Terrain Map-Conditioned VLAs for Autonomous Earthmoving: Preliminary Findings

Brian Blankenau[1*], Arturo Saucedo[1*], W. Jacob Wagner[1*], Isaac Blankenau[2], Dustin Nottage[1], Ahmet Soylemezoglu[1]

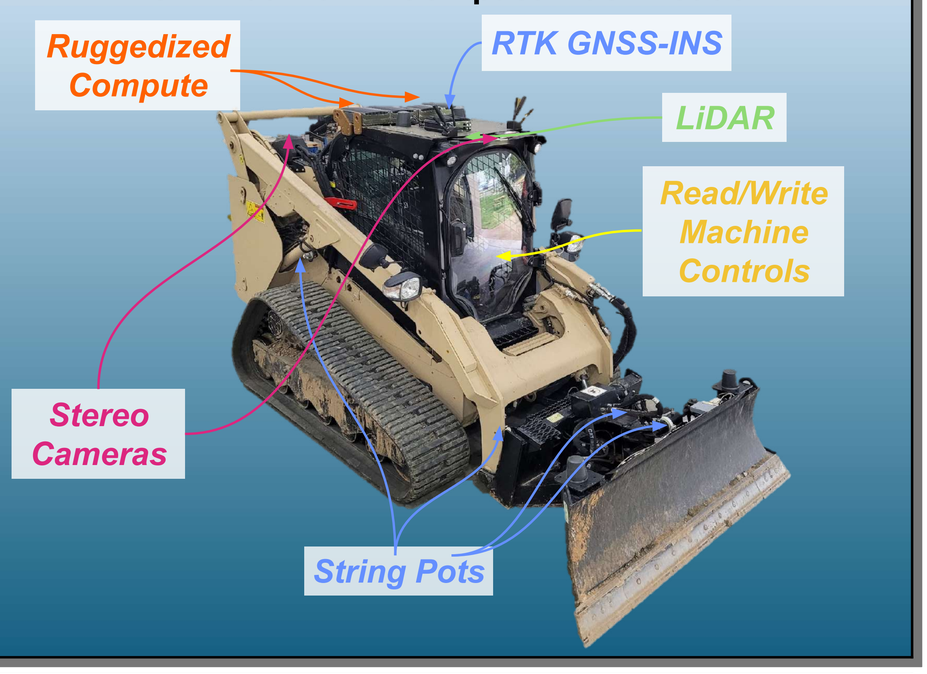
[1]CERL, [2]rerun.io (work performed at CERL; now at Rerun), [*]Equal Contribution.

Want to Chat @ ICRA? Reach Out: ***william.j.wagner@erdc.dren.mil +1-217-308-3852***

VLA Models for Earthmoving

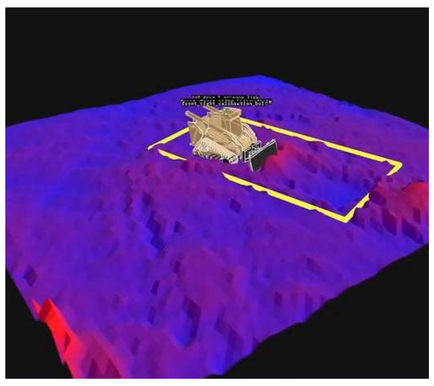
Vision-Language-Action models (VLAs) have demonstrated remarkable generalization in structured indoor environments. However, their efficacy in performing complex **field robotics** tasks remains largely unexplored. Earthmoving presents unique automation challenges as it includes mobile manipulation of **deformable terrain**, high contact forces, widely varying environmental conditions, etc. This study investigates how VLAs can be extended to **autonomous earthmoving** by utilizing **terrain map-conditioned policies** to achieve robust performance across diverse soil types.

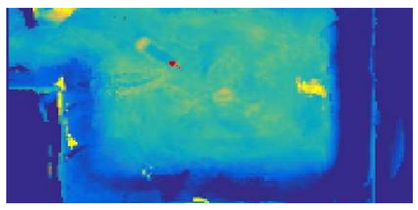
Robotics for Engineer Operations Robotic Compact Track Loader

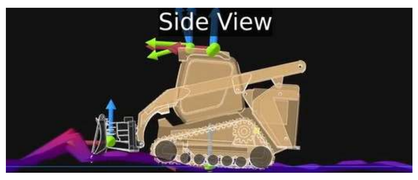


Earthmoving Aware Terrain Mapping

- Movement of the blade is tracked through the terrain to update heights with **swept volume displacements.**
- Soil failure physics is used to erode the map enabling realistic **pile formation** and **windrowing.**
- These proprioceptive mapping features ensure **continuous, occlusion-free terrain map updates** at the blade which is critical for policy learning.

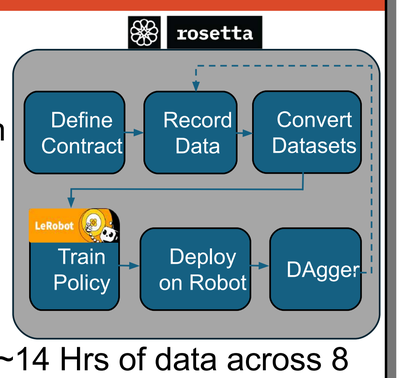






System Infrastructure

- **ROS2** software stack
- **LeRobot** model infrastructure
- **Rosetta** ROS2/LeRobot integration
  - Rosetta streamlines data recording, training, and model inference on ROS2 systems



Data Collection:

- A single human operator collected ~14 Hrs of data across 8 days on 3 earthmoving tasks.
- Data consisted of a range of environmental and soil conditions (overcast, rainy, sunny, windy, snowy, frozen soil, saturated soil, hard dry soil, etc.)

Model Architecture:

- Initial experimentation led to selection of SmolVLA base model. Future work will be based around $\pi_{0.5}$ policy.
- Weights are frozen for SIGLIP image encoders.

Robot State Features:

- Joint positions and velocities (lift, pitch, roll), track velocities, vehicle orientation (roll, pitch), vehicle twist, engine RPM.

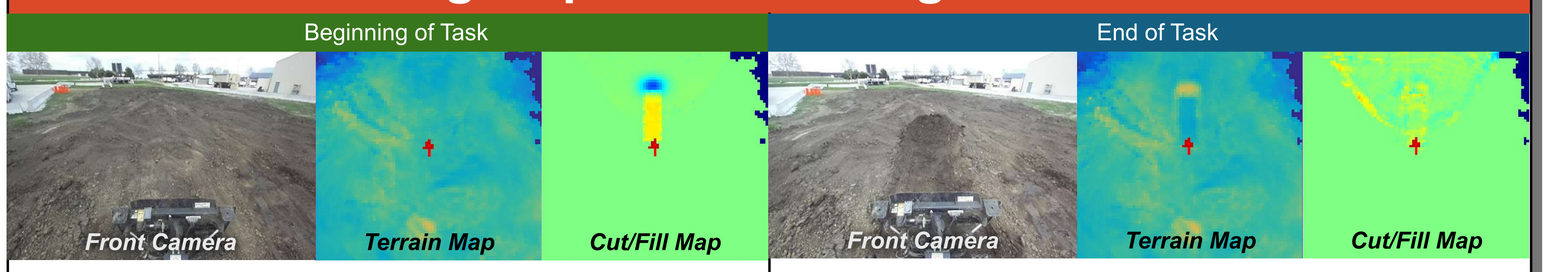
Robot Action Space:

- Joint efforts (Joystick commands similar to open loop velocity control).

Robot Control:

- 10 Hz- Base machine controls (slow vehicle dynamics)
- Simple **Anti-Stall Controller** reduces primary failure mode

Using Maps as VLA Image Features



Terrain Map

Need a means of providing the policy information of terrain state because:

- Depth is difficult to discern from camera image due to poor contrast of loose soil
- Blade obscures accumulated soil
- Tasks are long-horizon, require large work area, and ego-mounted cameras have limited FOV

Terrain maps are generated from an initial scan of the work site by the vehicle, and are continuously updated during operation.

Egocentric top-down terrain maps provide:

- Unobscured terrain heights as colored image
- Position of vehicle relative to terrain features
- Learnable representation (egocentric observation enables location agnostic task execution)
- Persistent total environment observaion- memory

Cut/Fill Map

For VLA earthmoving policies to be practical, models must:

- Build specific structures
- Allow users to specify location, shape, and size.

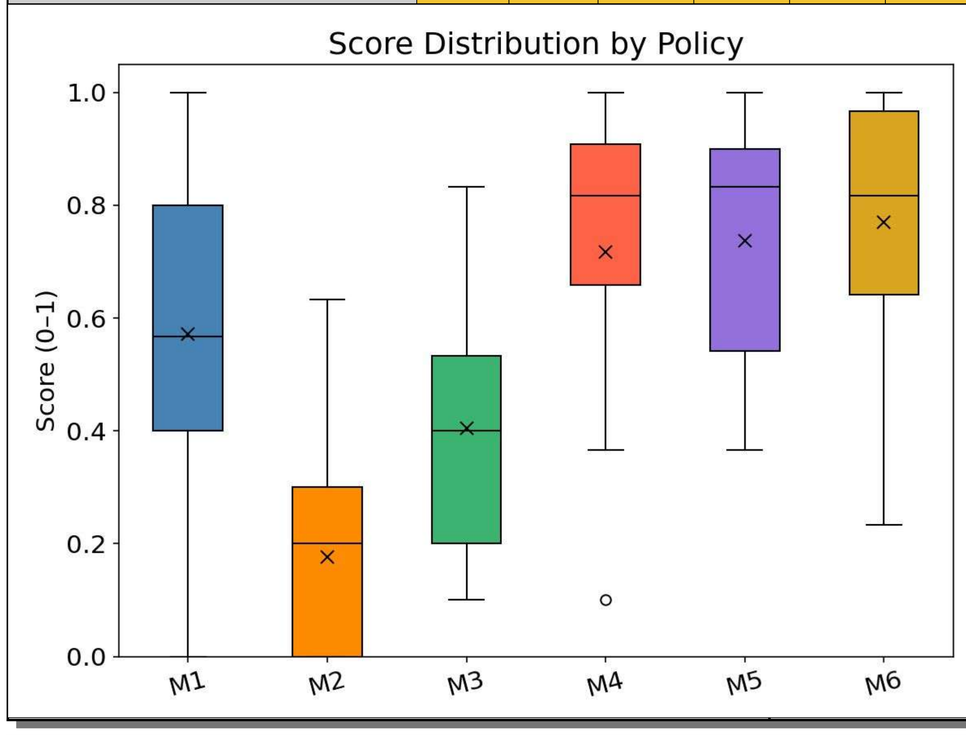
This information is contained in the goal-generated Cut/Fill map.

- Cut/Fill map shows the difference between the current terrain map and an initial goal
- Dynamically updated throughout task execution

Cut/Fill maps can be generated at train-time with a pure self-supervised approach:

- Record the terrain map during an arbitrary earthmoving task
- Provide the episodes final terrain map as the desired state/goal and reprocess the episode

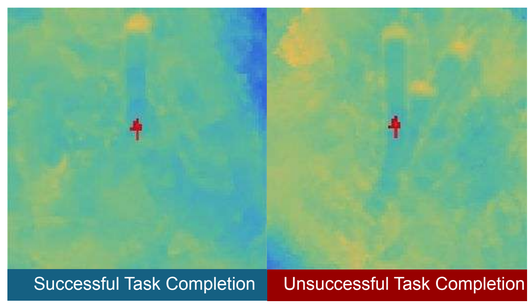
Dataset Breakdown	M1	M2	M3	M4	M5	M6
Front Image	X		X	X		X
Map Image		X	X		X	X
Cut/Fill Map				X	X	X
State Features	X	X	X	X	X	X



Feature Ablation Study

How does the inclusion of terrain and cut/fill maps as VLA inputs impact policy performance?

We trained six policies with different mixtures of input features and evaluated their performance on the prompt “Build a pile”.



Dataset details:

- 615 demonstrations across 5 tasks
- Observation consists of 19 vehicle states + images as specified in Dataset Breakdown

Early Conclusions

- Addition of Cut/Fill map with self-supervision significantly improves policy performance
  - Reduces key failure modes: vehicle run-away, re-starting pile creation, out-of-slot cuts
  - Supplies a mechanism to specify location information for desired operations
- Near parity between policies trained on map images alone (M5) and map and camera images (M6)
  - Excellent candidate for simulation-based pretraining using simple renderless simulations
- Frozen SIGLIP image encoders can be used with map images, though fine tuning may be preferred
- Addition of anti-stall controller improves task execution and eliminates a key failure mode

Future work

Dataset

- Increase task variety to cover broader range of goal earthmoving profiles
- Add simulation data for pre-training

Policy Architecture and Training

- Move to π series base models
- Expand model history (KV caching)
- Offline RL with advantage conditioning
- Stage aware reward modeling

Long Horizon Planning

- High-level diffusion stage planner for goal maps

Model Features

- Include cut-depth and loose soil depth
- Incorporate soil property knowledge from online soil property estimation

Platform:

- Expand to more capable machines - D3 Bulldozer
- Improve localization

References

- Shukor, Mustafa, et al. "Smolvla: A vision-language-action model for affordable and efficient robotics." arXiv preprint arXiv:2506.01844 (2025).
- Miki, Takahiro, et al. "Elevation mapping for locomotion and navigation using gpu." 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2022.
- Wagner, William Jacob. “Towards soil-aware autonomous terrain shaping for bulldozing”. Diss. University of Illinois Urbana-Champaign, 2025.
- LeRobot. “Unfolding Robotics: Open-Source Shirt Folding from Data to Deployment”. Hugging Face, https://huggingface.co/spaces/lerobot/robot-folding.
- Rosetta - https://github.com/iblnkn/rosetta