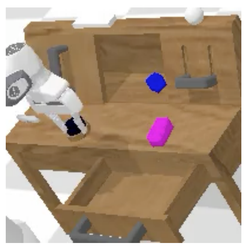
Red Teaming Language-Conditioned Robot Models via Vision Language Models

Sathwik Karnik*, Zhang-Wei Hong*, Nishant Abhangi*, Yen-Chen Lin, Tsun-Hsun Wang, Pulkit Agrawal

Different language instructions can cause robots to fail

Success

Before

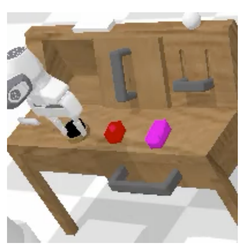


After

Prompt: Slide the pink block to the right

Failure

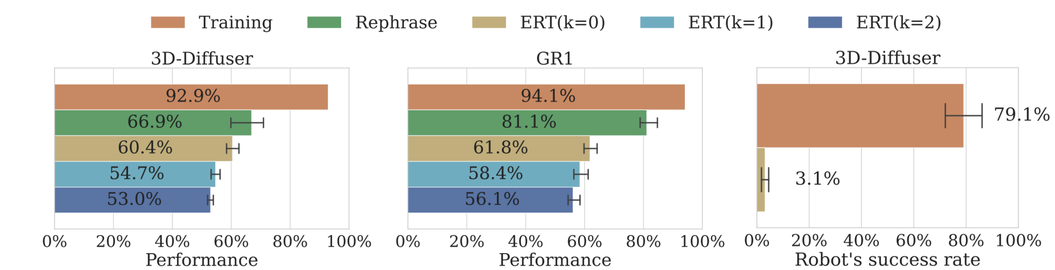
Before



After

Prompt: Sweep the pink block to the right

ERT discovers many instructions that cause language-conditioned robots to fail

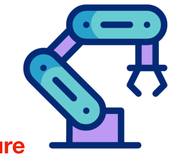


(a) CALVIN (b) RLBench

Embodied Red Teaming (ERT)

Task: close_drawer

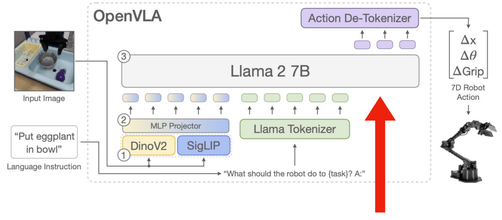
ERT: Move your arm towards the open drawer and push it until it closes completely.



success or failure

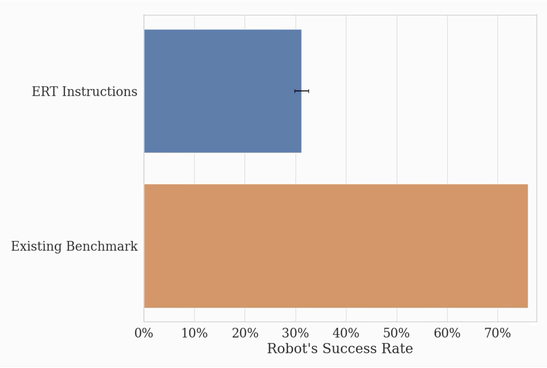
Iterative refinement to generate new instructions

ERT discovers many instructions that cause robotic foundation models to fail



OpenVLA is based on an LLM

Task: pick coke can



Sample ERT Instructions

Task	CALVIN Training Instruction	ERT Instruction
push_pink_block_left	slide the pink block to the left	Move the pink block to the left side of the table.
rotate_red_block_right	grasp the red block and turn it right	Rotate the position of the red block toward the right.
turn_on_lightbulb	turn on the yellow light	Move the robot arm to the switch and toggle it to turn on the lightbulb.

ERT instructions are diverse