



Vision Language Models in Robotics

Vladimír Petřík

vladimir.petrík@cvut.cz

15.12.2025

Motivation: Talking to Robots

- ▶ Traditional Pipeline:
 - ▶ Object Detector (bounding boxes)
 - ▶ Planner (geometry)
 - ▶ Controller (inverse dynamics)



Motivation: Talking to Robots

- ▶ Traditional Pipeline:
 - ▶ Object Detector (bounding boxes)
 - ▶ Planner (geometry)
 - ▶ Controller (inverse dynamics)
- ▶ Problem: "Pick up the empty cup"
 - ▶ Detector needs the class "empty_cup"
 - ▶ Hard to generalize to new objects



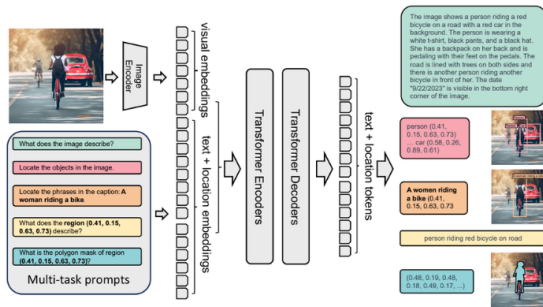
Motivation: Talking to Robots

- ▶ Traditional Pipeline:
 - ▶ Object Detector (bounding boxes)
 - ▶ Planner (geometry)
 - ▶ Controller (inverse dynamics)
- ▶ Problem: "Pick up the empty cup"
 - ▶ Detector needs the class "empty_cup"
 - ▶ Hard to generalize to new objects
- ▶ Solution: Vision Language Models (VLM)
 - ▶ Understand semantic concepts
 - ▶ Connect pixels to language directly



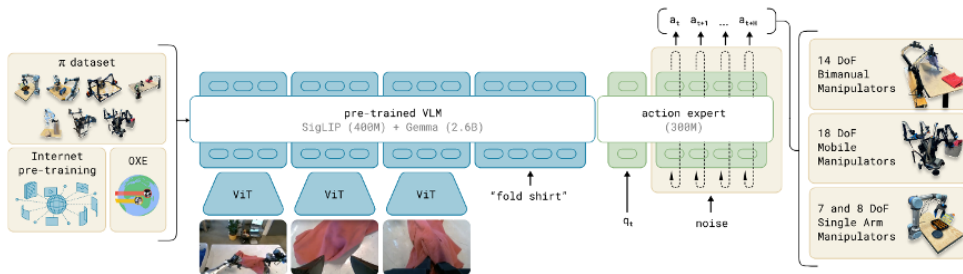
Vision-Language Models (VLM) - High Level

- ▶ **The Core Idea:** Giving Large Language Models "eyes".
 - ▶ **Input:** Image + Text - **Output:** Text reasoning
 - ▶ Trained on internet-scale data
- ▶ Why is this better than standard detectors?
 - ▶ Standard Detector: Output is "Box: Cup".
 - ▶ VLM: Output is "The cup is empty and close to the edge."
 - ▶ They understand **context** and **relationships** - crucial for robotics.



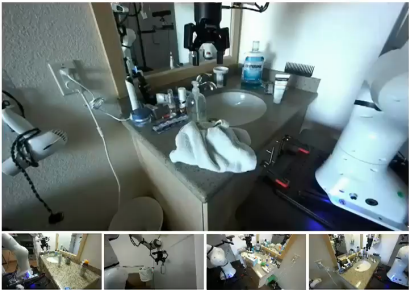
From VLM to VLA (Vision-Language-Action) model

- ▶ Robots need to output **actions**, not just text
- ▶ π_0 architecture
 - ▶ 3 input images (from three cameras)
 - ▶ query text (task description)
 - ▶ robot configuration q
 - ▶ outputs a horizon of robot actions








We need data for training - DROID

Bathroom

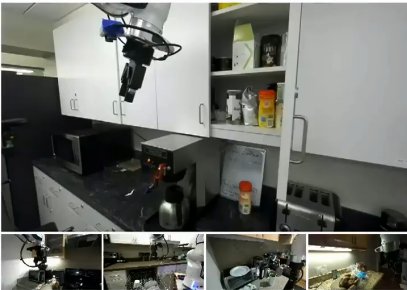


DROID

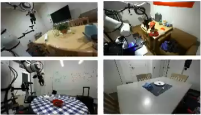
Distributed Robot Interaction Dataset

-  76k Episodes
-  564 Scenes
-  52 Buildings
-  13 Institutions
-  86 Tasks / Verbs


Kitchen




Dining Room




Bedroom



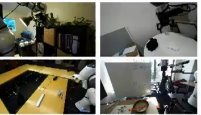
Laboratory



Laundry Room

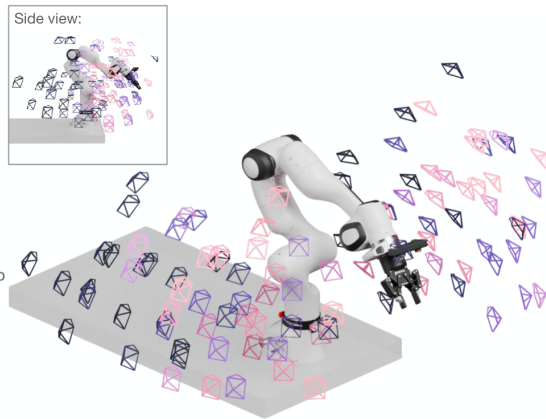
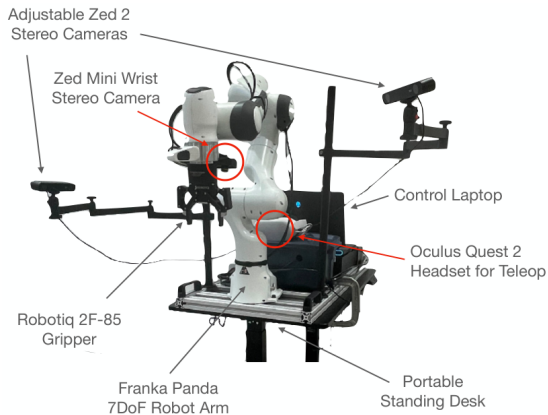


Office





DROID



How it works after training



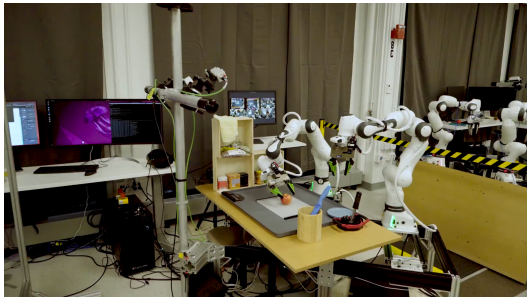
Put the apple into the pot and close the lid.



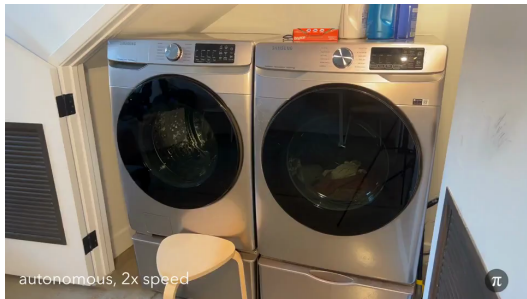
Put chips into the bowl.



Examples from other platforms - impressive results



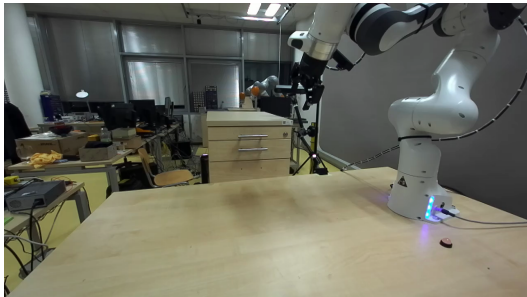
Slice the apple.



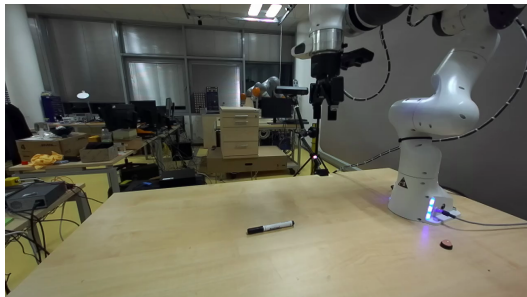
Take the laundry out of the washing machine.



Reality check

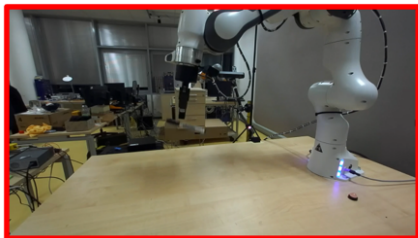


Open the top drawer.

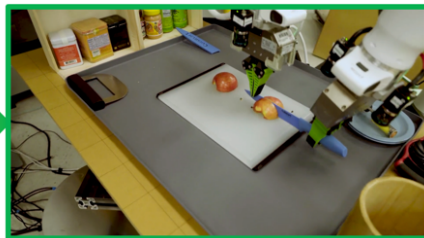


Rotate the marker.

So where are we actually?



?



Evaluating Vision-Language-Action Models

- ▶ VLAs can handle deformable objects or sliding apples
- ▶ They still fail on some simple tasks



Evaluating Vision-Language-Action Models

- ▶ VLAs can handle deformable objects or sliding apples
- ▶ They still fail on some simple tasks
- ▶ Key questions about their performance remain:
 - ▶ What are the failure modes?
 - ▶ How well do VLAs understand human language instructions?
 - ▶ Can VLAs generalize across diverse objects, scenes and tasks?

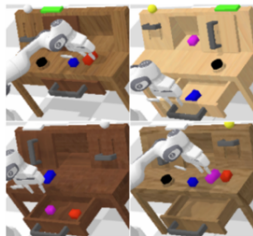


Evaluating Vision-Language-Action Models

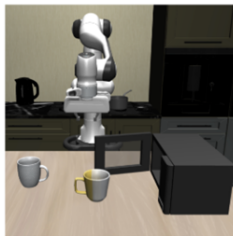
- ▶ VLAs can handle deformable objects or sliding apples
- ▶ They still fail on some simple tasks
- ▶ Key questions about their performance remain:
 - ▶ What are the failure modes?
 - ▶ How well do VLAs understand human language instructions?
 - ▶ Can VLAs generalize across diverse objects, scenes and tasks?
- ▶ Problem: real-world evaluation is very expensive. . .



Solution: Simulated benchmarks



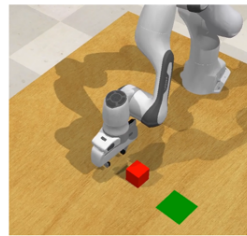
CALVIN [1]



LIBERO [2]

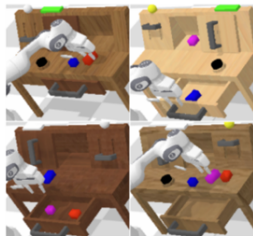


VLABench [3]

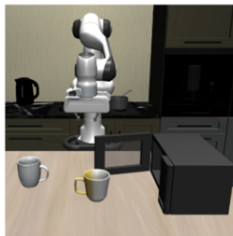


COLOSSEUM [4]

Solution: Simulated benchmarks



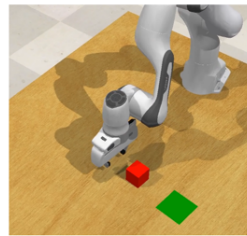
CALVIN [1]



LIBERO [2]



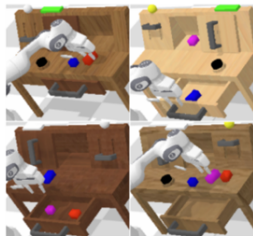
VLABench [3]



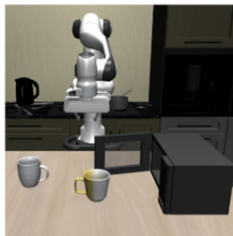
COLOSSEUM [4]

► Can we trust the results?

Solution: Simulated benchmarks



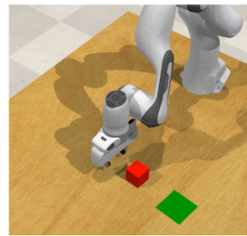
CALVIN [1]



LIBERO [2]



VLABench [3]

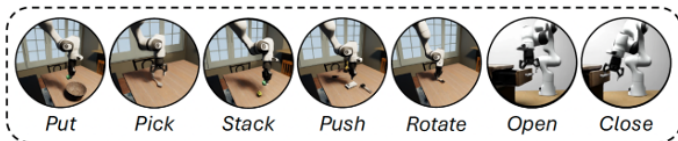


COLOSSEUM [4]

- ▶ Can we trust the results?
- ▶ Real-world performance of VLAs does not correspond to simulated performance
 - ▶ Visual gap
 - ▶ Control gap



REALM: a real-to-sim validated generalization benchmark



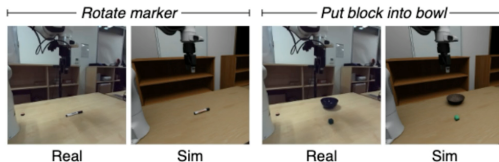
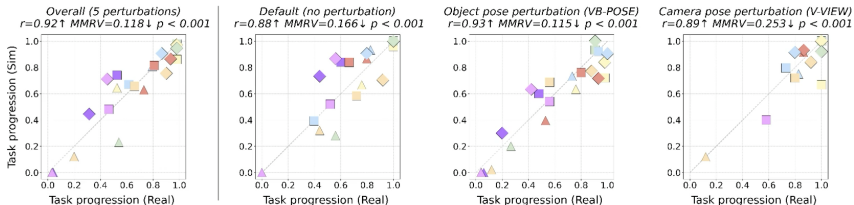
7 manipulation skills



15 perturbations | 3 categories



Real-to-Sim Validation



Closing the gap: *aligned robot control*



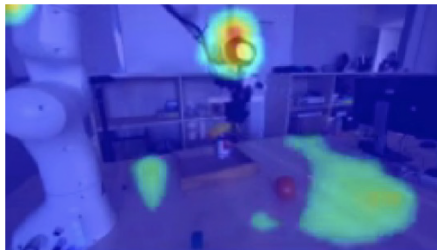
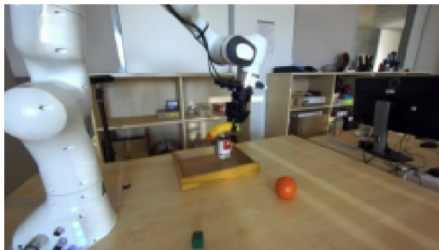
2x speed

Visual validation

Input image

π_0 attention maps

Real



Sim



REALM benchmark

- ▶ We have a trusted simulator - 800 rollouts in the real world compared to simulation.
- ▶ We can build generalization benchmark via perturbations.
 - ▶ Visual perturbations
 - ▶ Semantic perturbations
 - ▶ Behavioral perturbations

Perturbation	Description & Implementation
Default	Testing a skill under no specific perturbations.
<i>Visual</i>	
V-AUG	Randomize <i>blur</i> and <i>contrast</i> .
V-SC	Randomly spawn <i>new distractors</i> in the scene.
V-VIEW	Random shifts to external <i>camera pose</i> .
V-LIGHT	Randomize illumination <i>color</i> and <i>intensity</i> .
<i>Semantic</i>	
S-PROP	Reference objects based on their properties.
S-LANG	Reference similar verbs and remove articles.
S-MO	Reference spatial relationships in the scene.
S-AFF	Reference human needs and use cases.
S-INT	Reference facts about the world that typically require knowledge from Internet-scale text data.
<i>Behavioral</i>	
B-HOBJ	Randomize manipulated object <i>mass</i> .
<i>Visual+Behavioral</i>	
VB-POSE	Randomize manipulated <i>object pose</i> .
VB-MOBJ	Randomize object <i>size</i> and <i>shape</i> .
<i>Semantic+Behavioral</i>	
SB-NOUN	Reference <i>another known object in the scene</i> .
SB-VRB	Change the <i>tested skill</i> for another compatible one.
<i>Visual+Semantic+Behavioral</i>	
VSb-NOBJ	Sample a <i>new unseen manipulated object</i> .



Closing the gap: *aligned robot control*



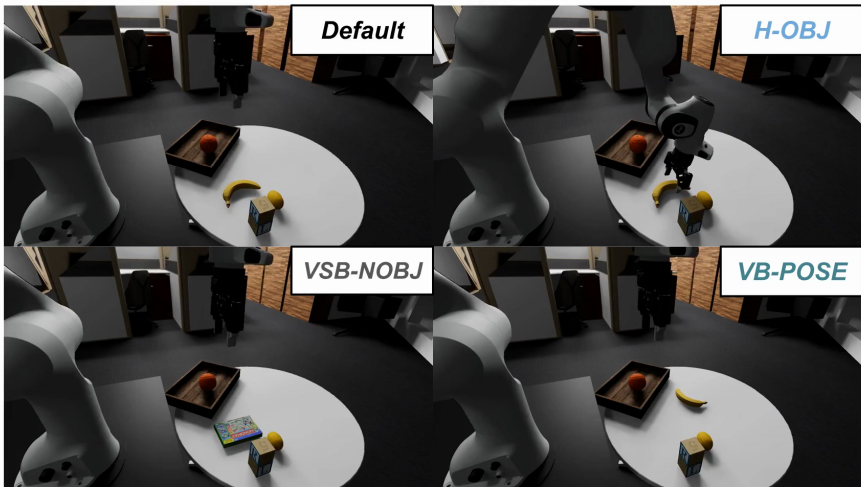
2x speed

Semantic perturbations

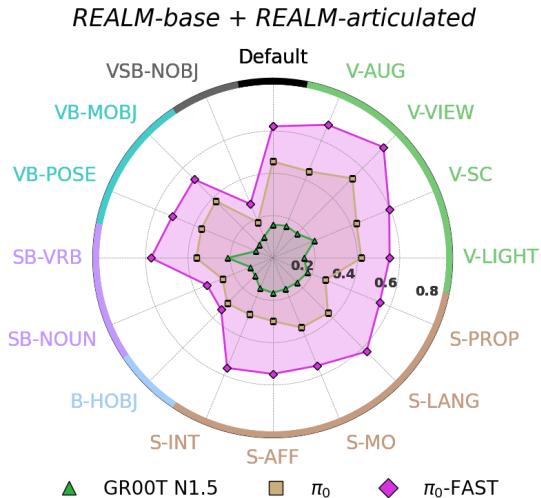


"put the banana into the box"

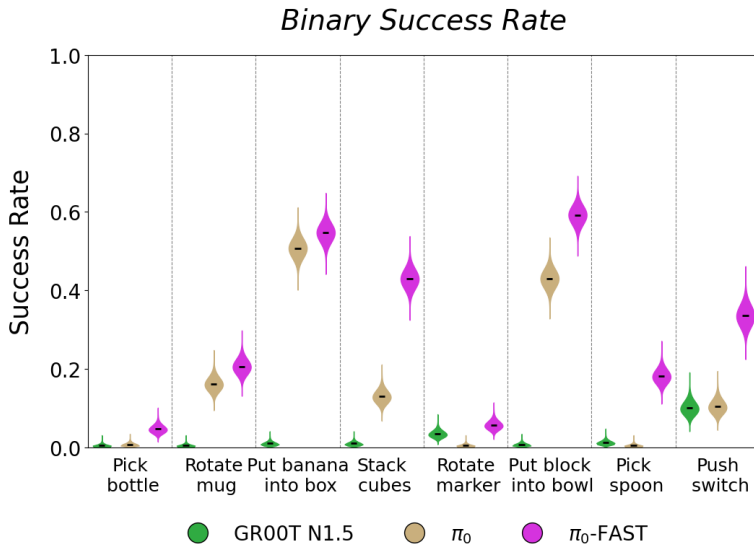
Behavioral perturbations



Performance of VLAs



Performance of VLAs

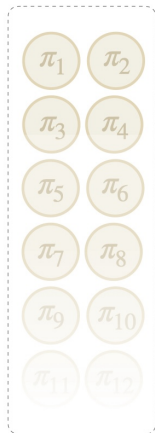


Takeaways for VLAs

- ▶ High-fidelity simulation with aligned robot control can serve as a valuable **proxy** for real-world performance
- ▶ Noticeable drop in performance under semantic perturbations
- ▶ High sensitivity to camera view
- ▶ All skills are short-horizon
 - ▶ Put something into something
 - ▶ Wipe a board
- ▶ Reliability and robustness have not yet been achieved - we need more data

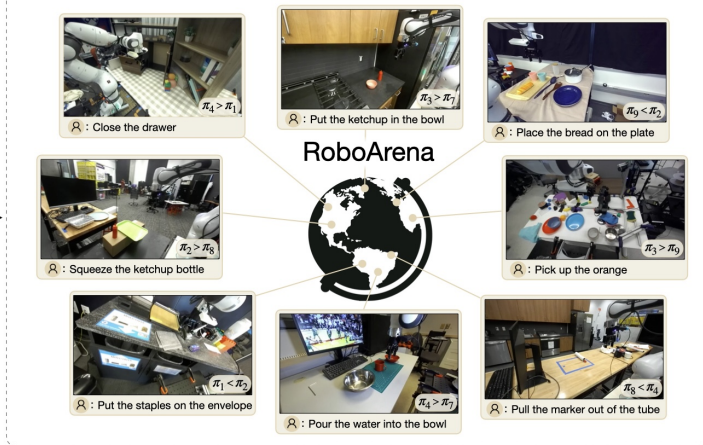


Generalist Policy Pool



Distributed Network of Evaluators

(each runs A/B comparisons on *whatever* scene & task they choose)



Aggregate pairwise policy preferences

Policy Ranking

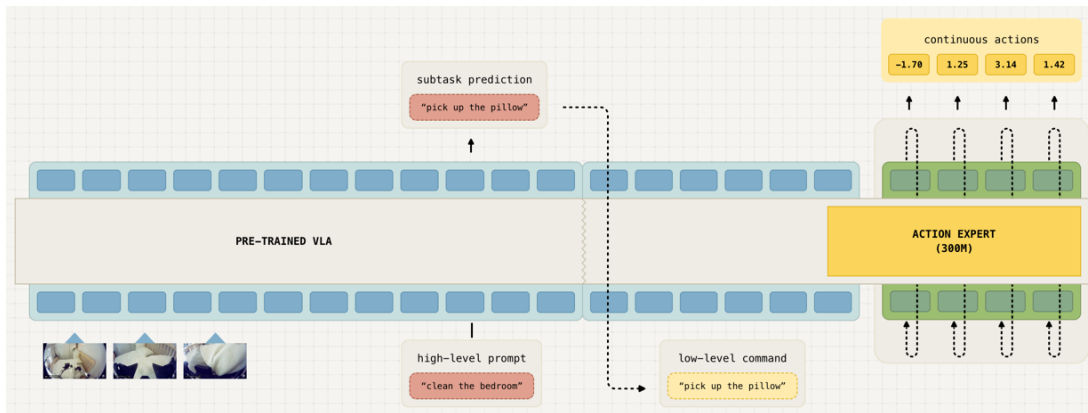
Policy	Score
π_4	1750
π_2	1321
π_1	1109
π_9	965
π_3	855

RoboArena Leaderboard

Rank	Policy	Score	SD	# A/B Evals	Open Source
1	pi05_droid	1867	29.1	490	
2	pi0_fast_droid	1819	27.9	654	
3	paligemma_fast_specialist_droid	1806	28.8	873	✓
4	paligemma_vq_droid	1780	30.3	680	✓
5	paligemma_fast_droid	1744	31.5	882	✓
6	paligemma_diffusion_droid	1652	43.8	678	✓
7	dam	1238	219.3	60	✓
8	pi0_droid	887	32	939	
9	paligemma_binning_droid	707	28.7	566	✓



Hierarchical reasoning - π 0.5



Conclusion

- ▶ VLAs can revolutionize robotics
 - ▶ Active research area
 - ▶ Needs data
- ▶ To remember
 - ▶ what is VLA
 - ▶ difference between simulated and real-world evaluations
- ▶ Topics not covered
 - ▶ Other modalities for VLA
 - ▶ Chain of thought reasoning for VLA
 - ▶ Safety issues
- ▶ Do you want to contribute to VLM/VLA research?



Exam

- ▶ CIIRC: B670
- ▶ Written (2h):
 - ▶ Theoretical questions
 - ▶ Computation with transformations
 - ▶ Kinematics of open kinematic chains
- ▶ Oral exam (10 minutes)

