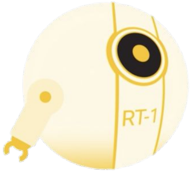
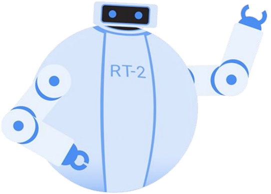
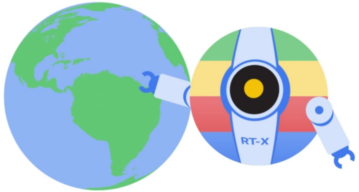
What's Missing for Robotics-First Foundation Models?

Ted Xiao

Google DeepMind

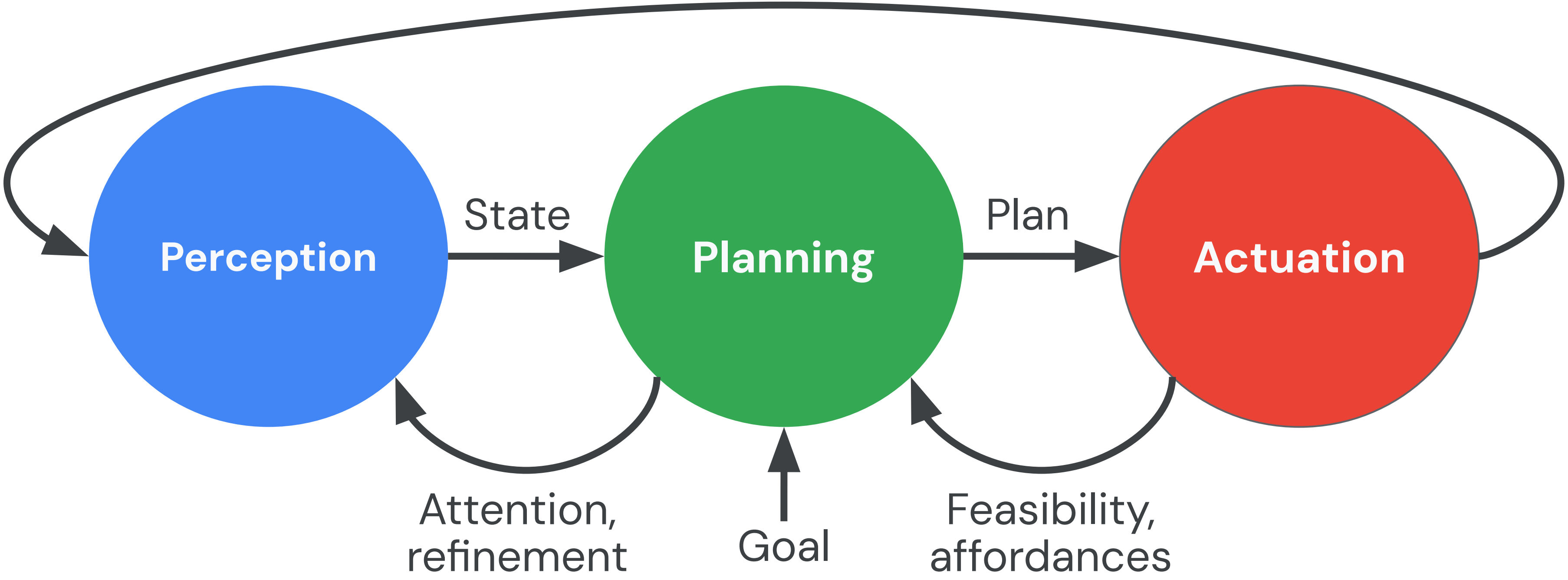






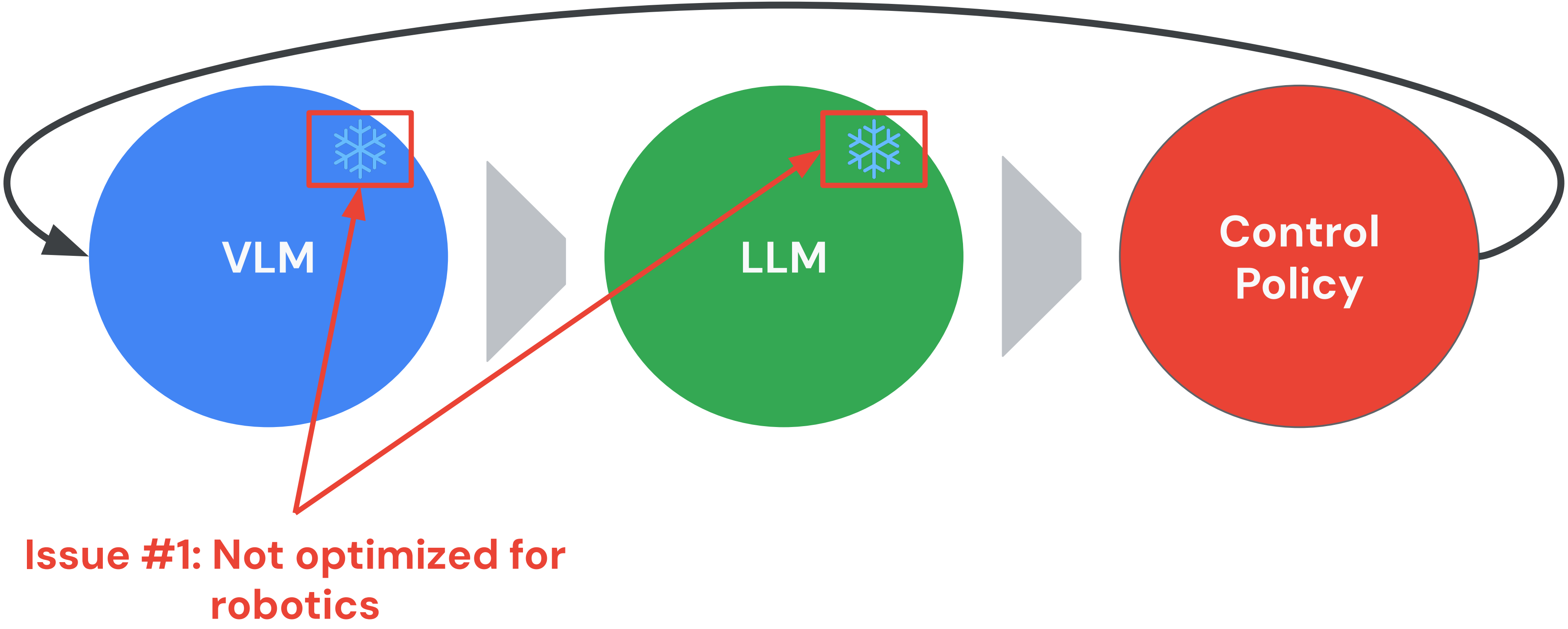
Agenda

The Robotics Information Flow

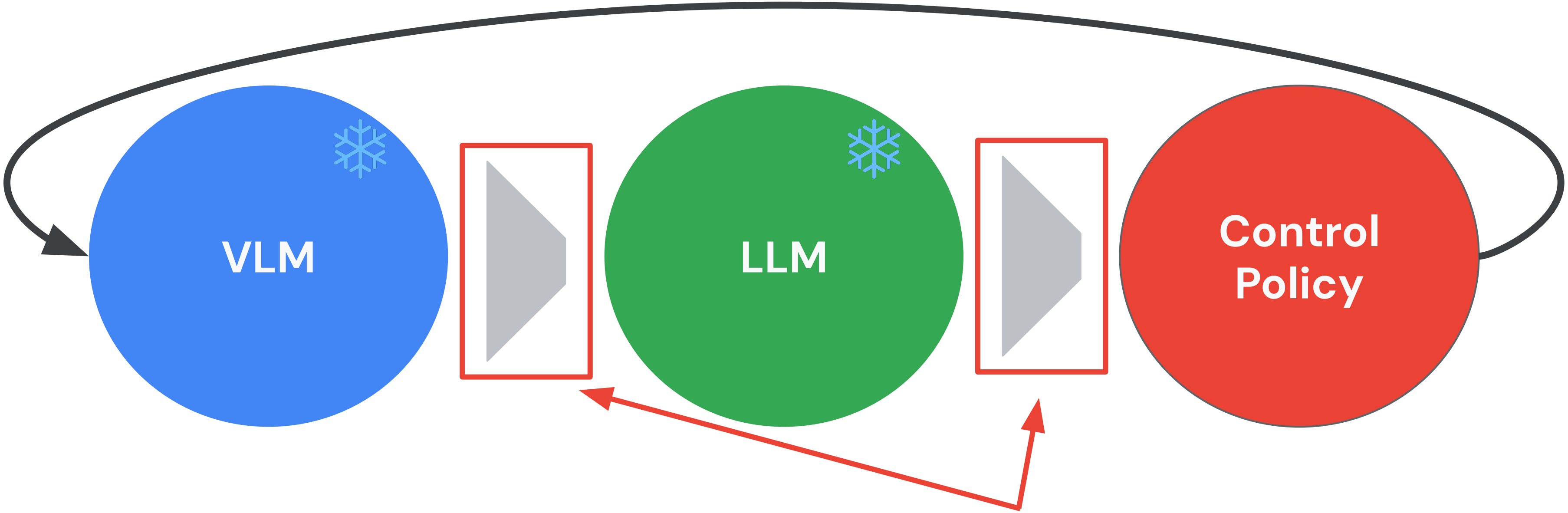


Foundation Models as Experts

Foundation Models as Experts



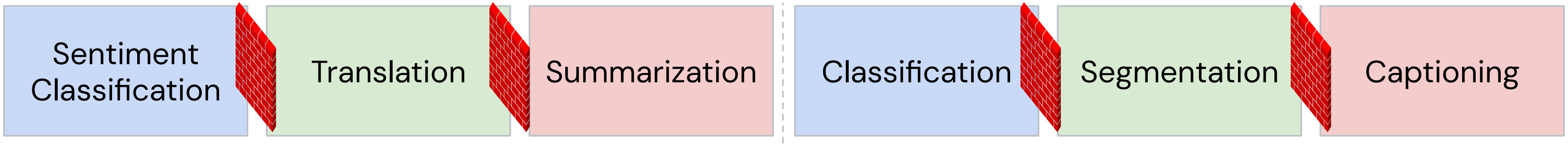
Foundation Models as Experts



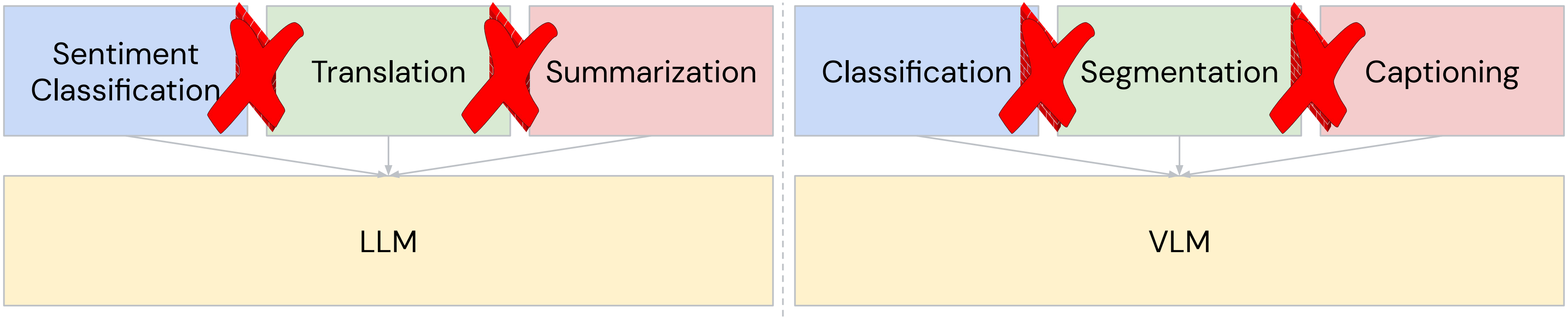
Issue #1: Not optimized for robotics

Issue #2: Narrow communication bandwidth between "intelligence modules"

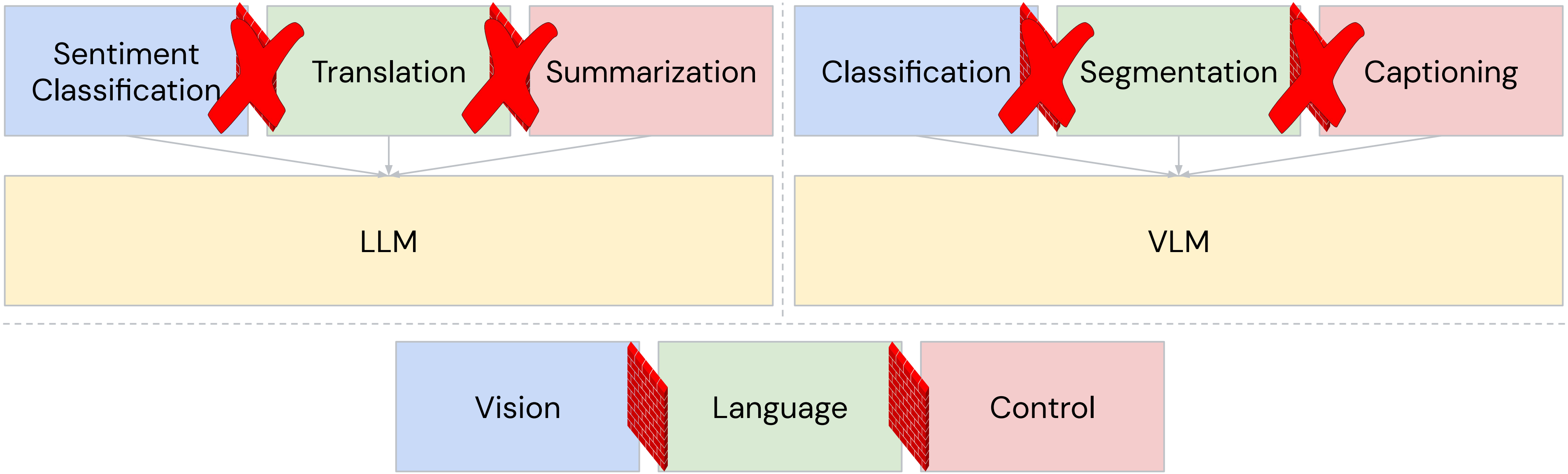
Foundation Model–fication of Robotics?



Foundation Model–fication of Robotics?

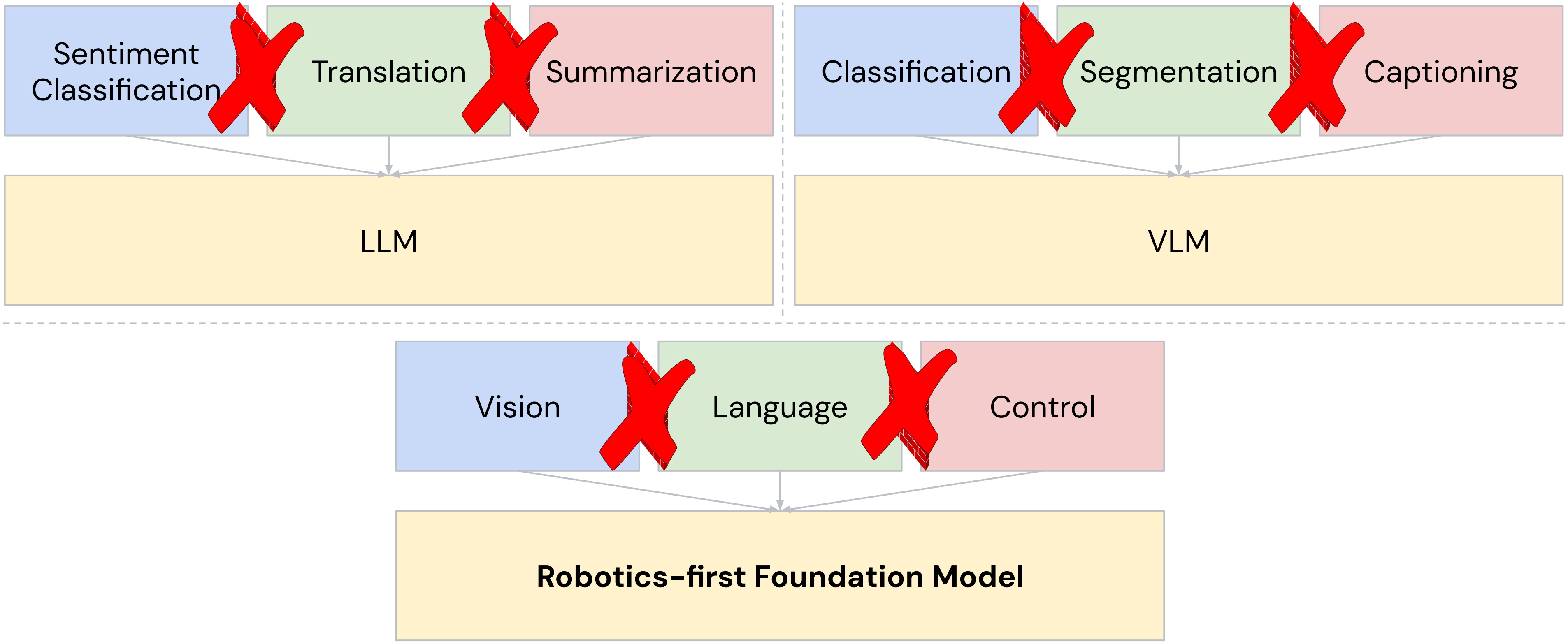


Foundation Model–fication of Robotics?



Google

Foundation Model–fication of Robotics?



Missing Foundation Model Pieces

	Non-robotics Foundation Models	**Robotics-first Foundation Model**
Positive Transfer from Scale	*Generalists beat specialists* *Scaling laws*	***???***

Missing Foundation Model Pieces

	Non-robotics Foundation Models	**Robotics-first Foundation Model**
Positive Transfer from Scale	*Generalists beat specialists* *Scaling laws*	***???***
Steerability and Promptability	*Prompt Engineering* *Few-shot Learning*	***???***

Missing Foundation Model Pieces

	Non-robotics Foundation Models	**Robotics-first Foundation Model**
Positive Transfer from Scale	*Generalists beat specialists* *Scaling laws*	???
Steerability and Promptability	*Prompt Engineering* *Few-shot Learning*	???
Scalable Evaluations	*Realistic Evals* *Predictive Benchmarks*	???

Google

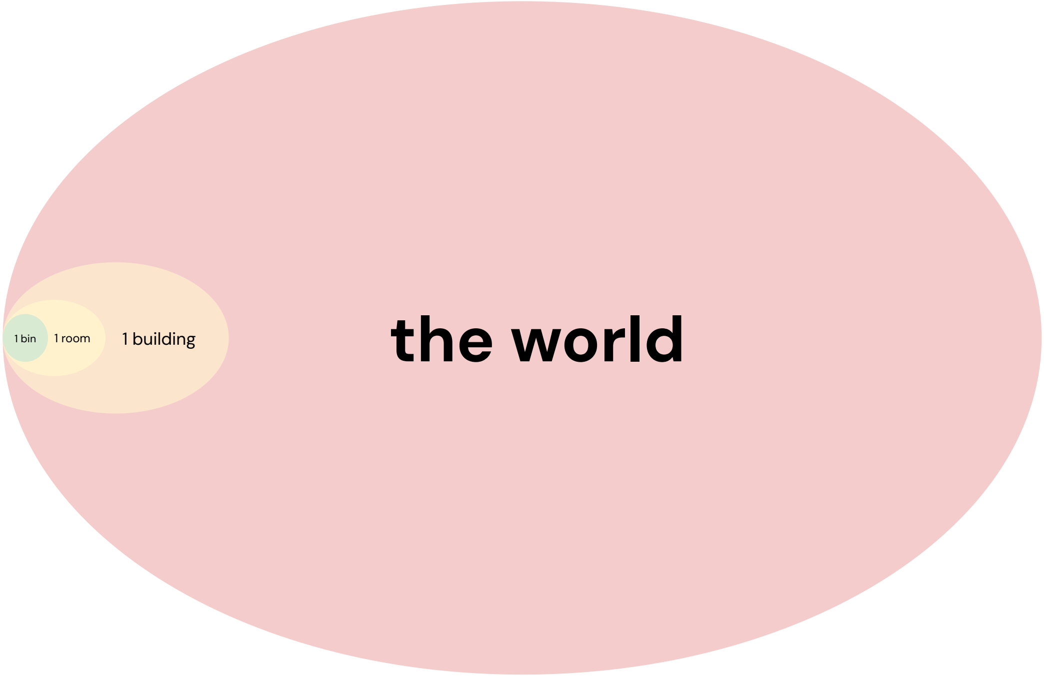
Missing Foundation Model Pieces

Claim: These missing properties are <u>necessary</u> for robotics to operate in the real world

Positive Transfer from Scale

Steerability and Promptability

Scalable Evaluations



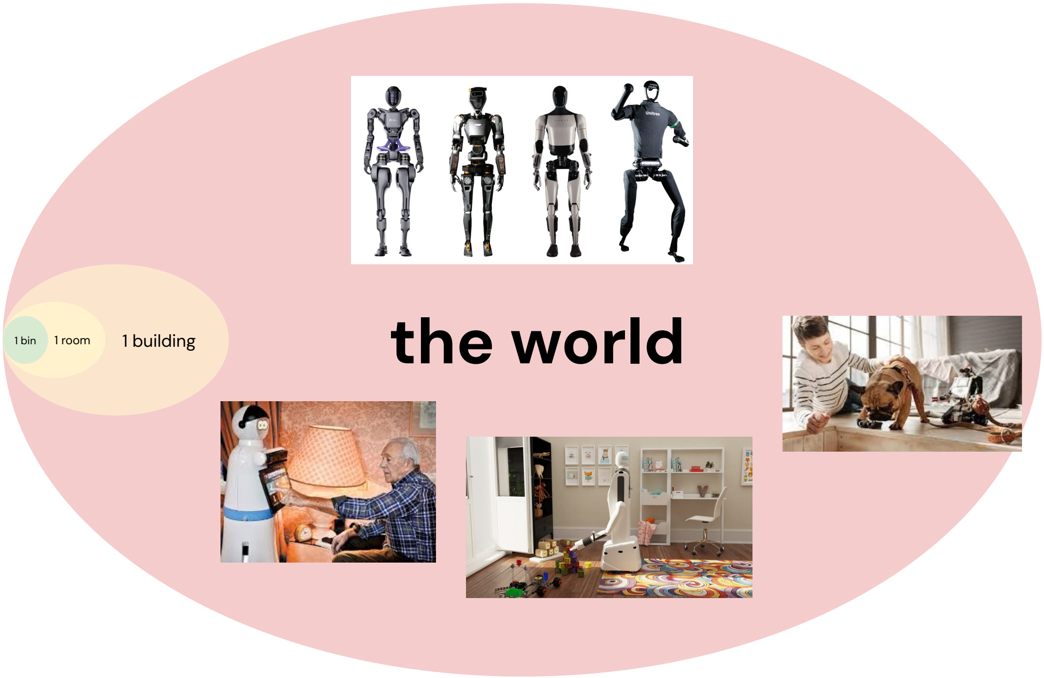
Missing Foundation Model Pieces

Claim: These missing properties are <u>necessary</u> for robotics to operate in the real world

Positive Transfer from Scale

Steerability and Promptability

Scalable Evaluations



Missing Foundation Model Pieces

Claim: These missing properties are <u>necessary</u> for robotics to operate in the real world

Positive Transfer from Scale

Steerability and Promptability

2024 level SoTA technology is not sufficient for general robotics.
At least one or two paradigm shifts (algorithms and data) required

Agenda

01 Why Robot Foundation Models?

02 Piece #1: Positive Transfer from Scaling

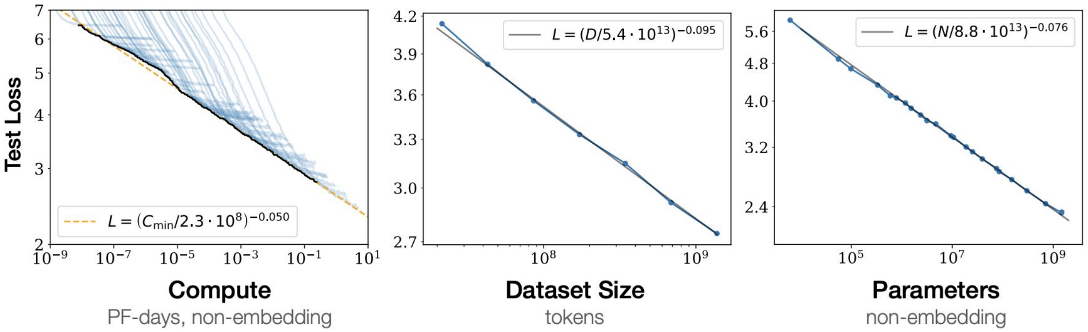
03 Piece #2: Steerability

04 Piece #3: Scalable Evaluation

05 Horizons

Lessons from Foundation Modeling: Data Scaling

- **Data scaling** a key ingredient in LLMs and VLMs
- ...but the internet already exists. No equivalent for robot data yet!



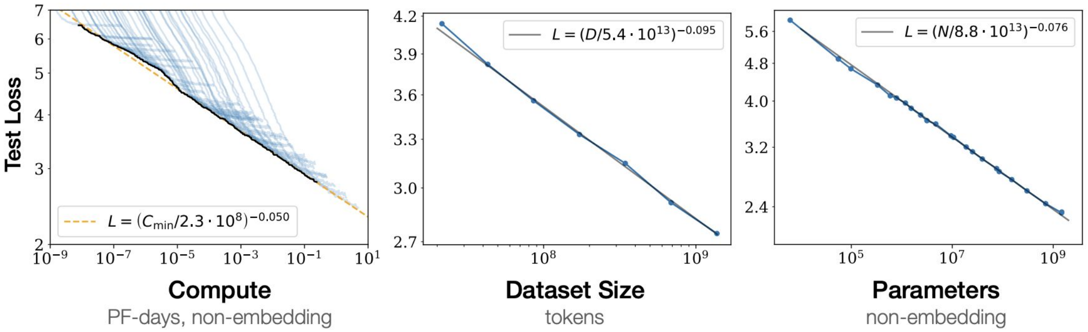
Source: Kaplan et al. 2020

Lessons from Foundation Modeling: Data Scaling

- **Data scaling** a key ingredient in LLMs and VLMs
- ...but the internet already exists. No equivalent for robot data yet!

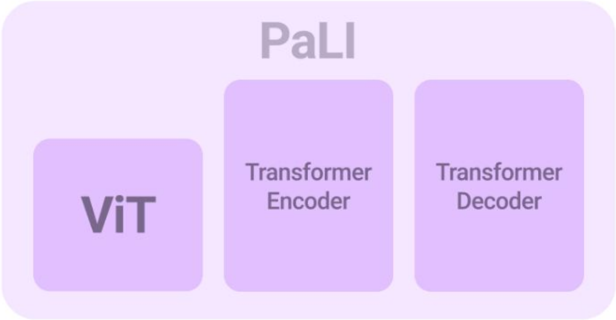
#1
Merge robot data with internet data?

#2
Merge all kinds of robot data?



Source: Kaplan et al. 2020

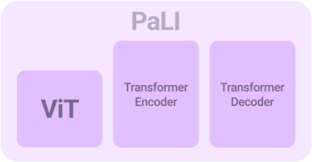
Vision-Language Models



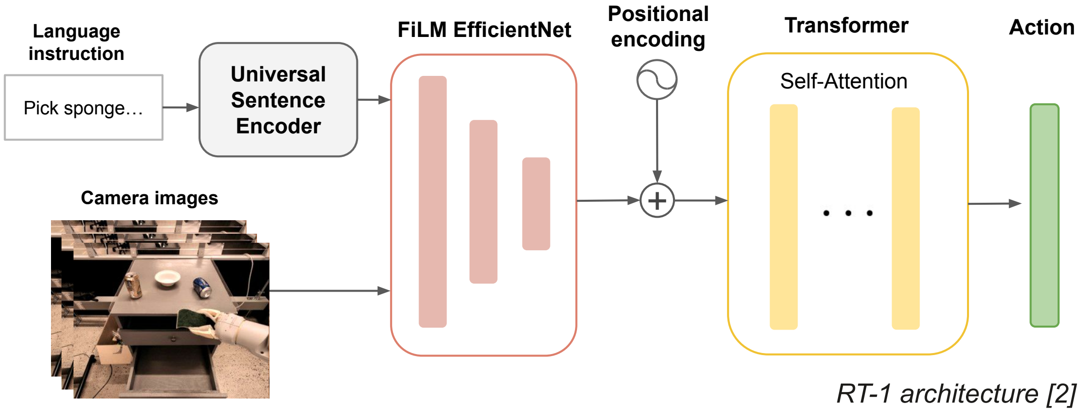
- VLMs encompass both **visual** and **semantic** understanding of the world

[1] PaLI: A Jointly-Scaled Multilingual Language-Image Model. Chen et al. 2022.

VLMs as Robot Policies



PaLI architecture [1]

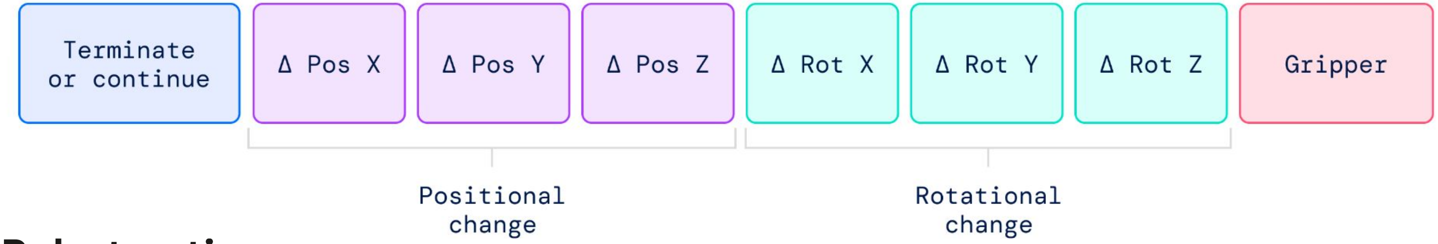


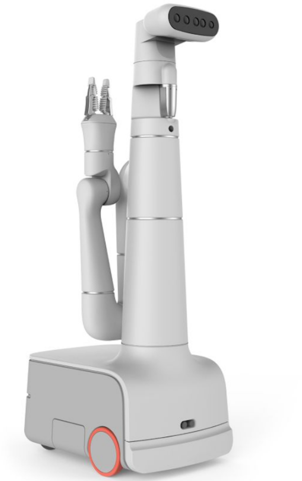
RT-1 architecture [2]

- **RT-1:** image + text → **discretized actions**
- Similar to a Visual-Language Model (VLM) with different **output tokens**
- Use large pre-trained VLMs directly as the **policy**!
- How do we **deal with actions** when using pre-trained VLMs?

[1] PaLI: A Jointly-Scaled Multilingual Language-Image Model. Chen et al. 2022.
[2] RT-1: Robotics Transformer for Real-World Control at Scale, Robotics at Google and Everyday Robots, 2022.

Representing Actions in VLMs





- **Robot actions:**
  - Moving the robot arm and gripper
  - Discretized into 256 bins
- **Actions in VLMs**
  - Convert to a string of numbers
  - Example: “1 127 115 218 101 56 90 255”
  - Alternatives:
    - *Float numbers* - more tokens needed
    - *Extra-IDs, least used* language tokens
    - *Human language (left, right etc.)* - can’t be directly executed on a robot

→ Vision-Language-Action (VLA) model!

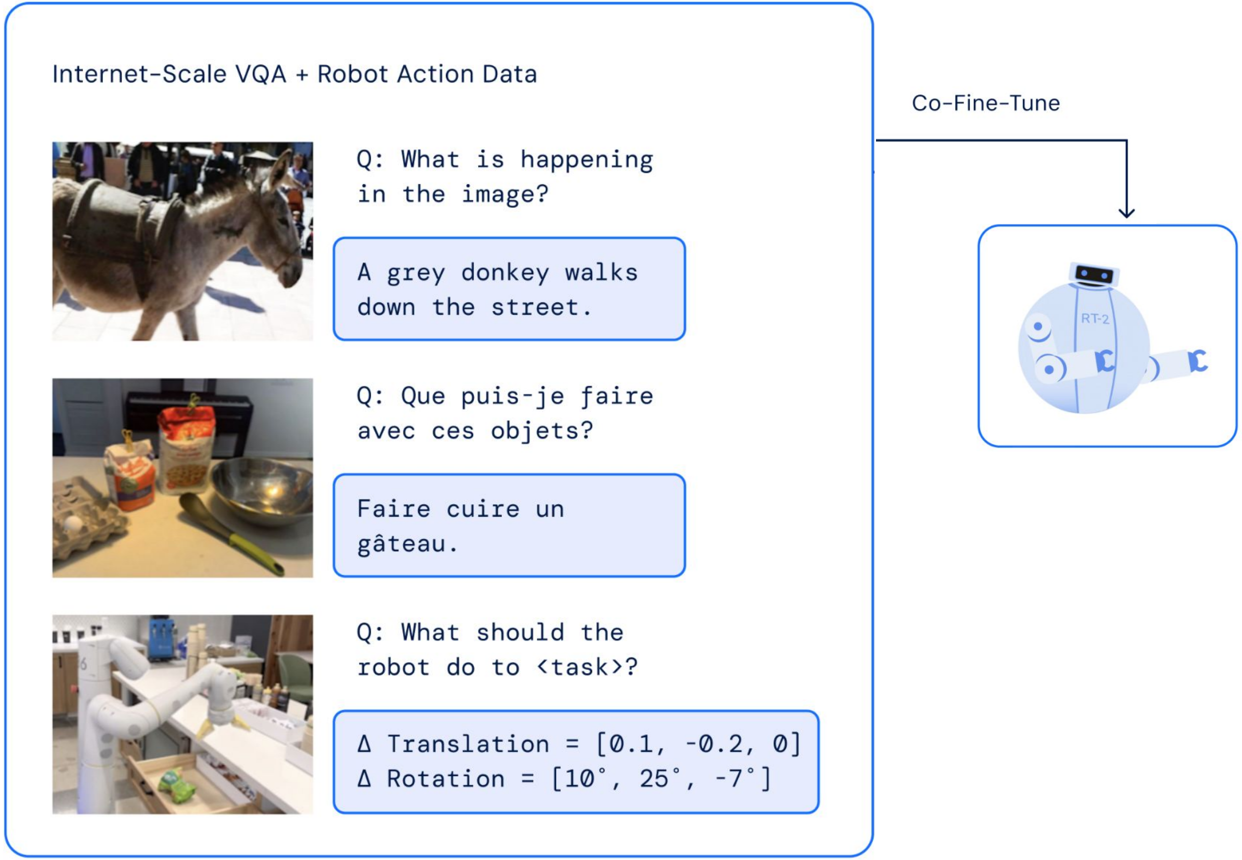
Training data and underlying models

Models

- PaLI-X (5B, 55B)
- PaLM-E (12B)

Data

- Pretraining: Web-data
- Robot data
  - RT-1 data
  - 13 robots
  - 17 months
  - 130k demos



Results: Emergent skills

put strawberry into the correct bowl

pick up the bag about to fall off the table

move apple to Denver Nuggets

pick robot

place orange in the matching bowl

move redbull can to H

move soccer ball to basketball

move banana to Germany

move cup to the wine bottle

pick animal with different color

move coke can to Taylor Swift

move coke can to X

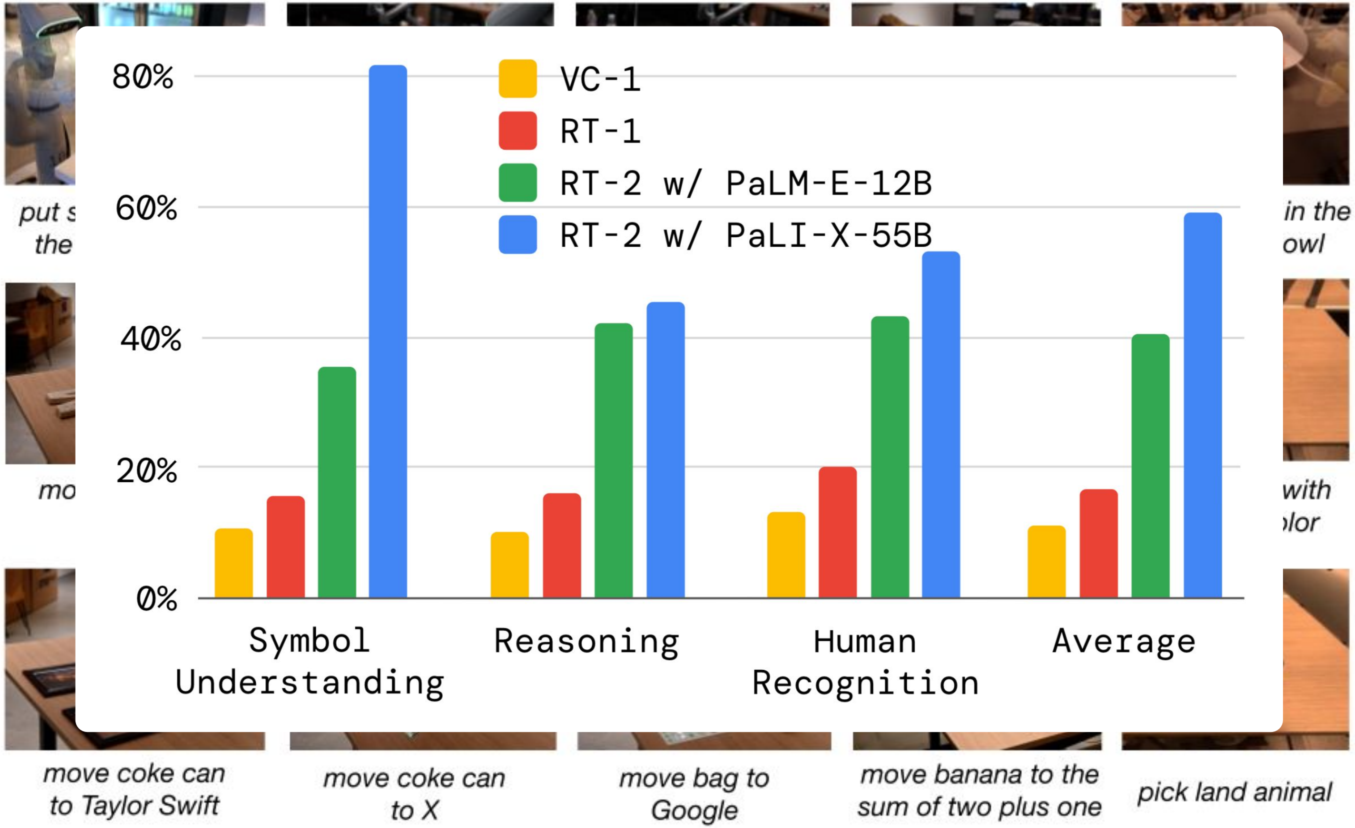
move bag to Google

move banana to the sum of two plus one

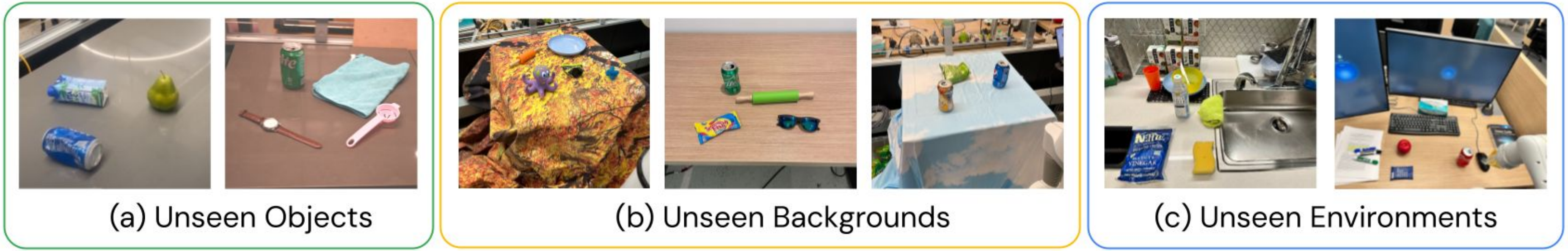
pick land animal

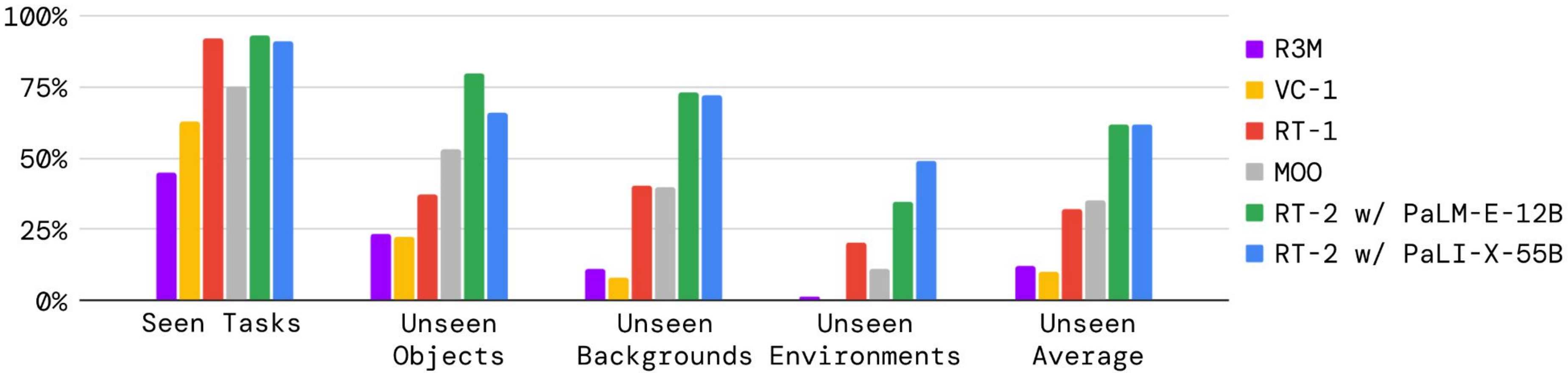
Results: Emergent skills

Google DeepMind



Results: Quantitative evals



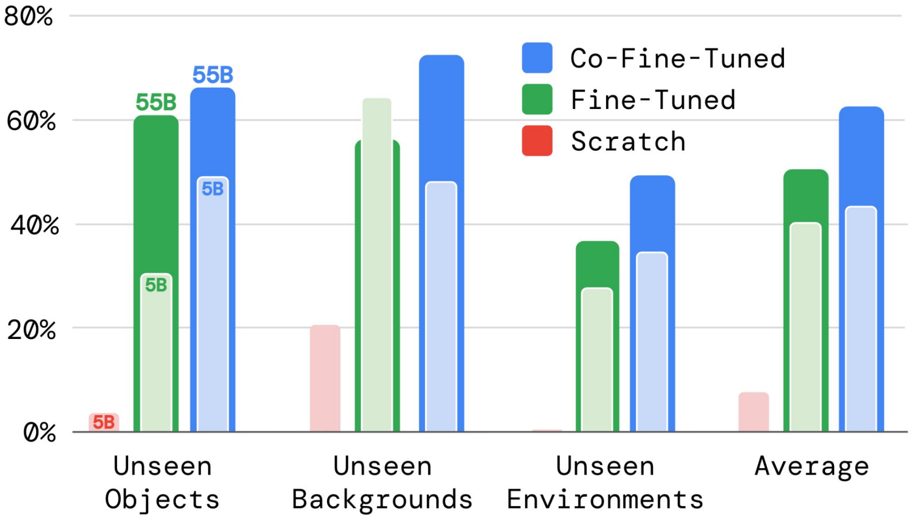


Results: Quantitative evals

(a) Unseen Objects (b) Unseen Backgrounds (c) Unseen Environments

RT2 w/ PaLI-X-55B ablations

- Co-Fine-Tuning with VQA data
- Fine-Tuning on robot data only
- Training on robot data from scratch



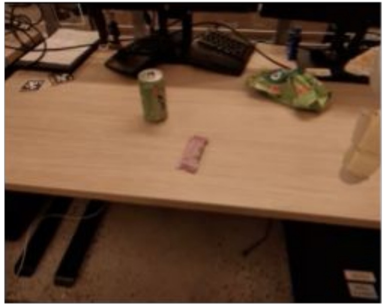
Positive Transfer: Chain-of-Thought with RT-2-PaLM-E

Prompt:
Given <img> Instruction:
Bring me a drink.
Prediction:
Plan: pick 7up can.
Action: 1 143 129 123 145
114 115 127

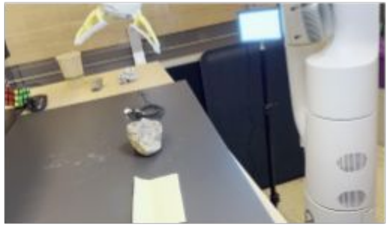
Prompt:
Given <img> Instruction:
Move all the objects
together.
Prediction:
Plan: move green can near
green rice chip bag.
Action: 1 128 126 127 135
123 119 127

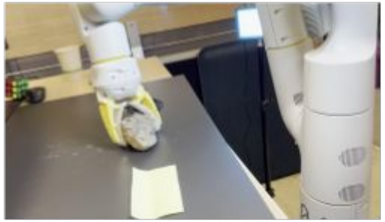
Prompt:
Given <img> Instruction:
Pick the object that is
different from all other
objects
Prediction:
Plan: pick rxbar
chocolate. Action: 1 128
129 125 131 125 128 127

Prompt:
Given <img> Instruction:
Move the green objects
together.
Prediction:
Plan: move green can near
green rice chip bag.
Action: 1 130 129 121 131
127 128 127



Prompt:
Given <img> I need to
hammer a nail, what
object from the scene
might be useful?
Prediction:
Rocks. Action: 1 129 138
122 132 135 106 127





The Open X-Embodiment Dataset



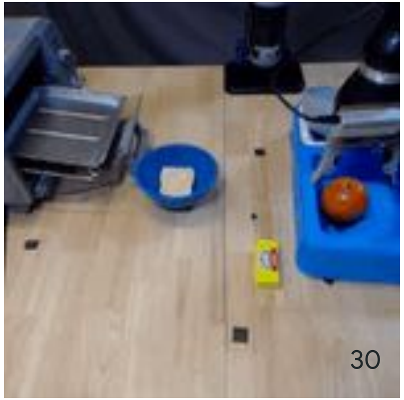
1M+ Real Robot Episodes

22 Robot Embodiments

34 Research Labs

300+ Scenes

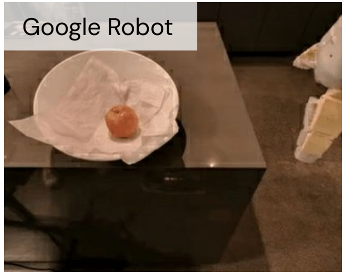


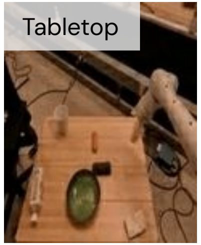


The Open X-Embodiment Dataset

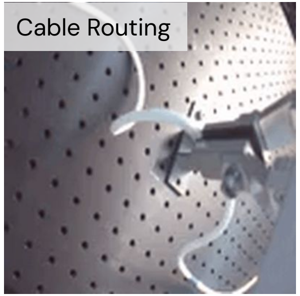
Many Embodiments



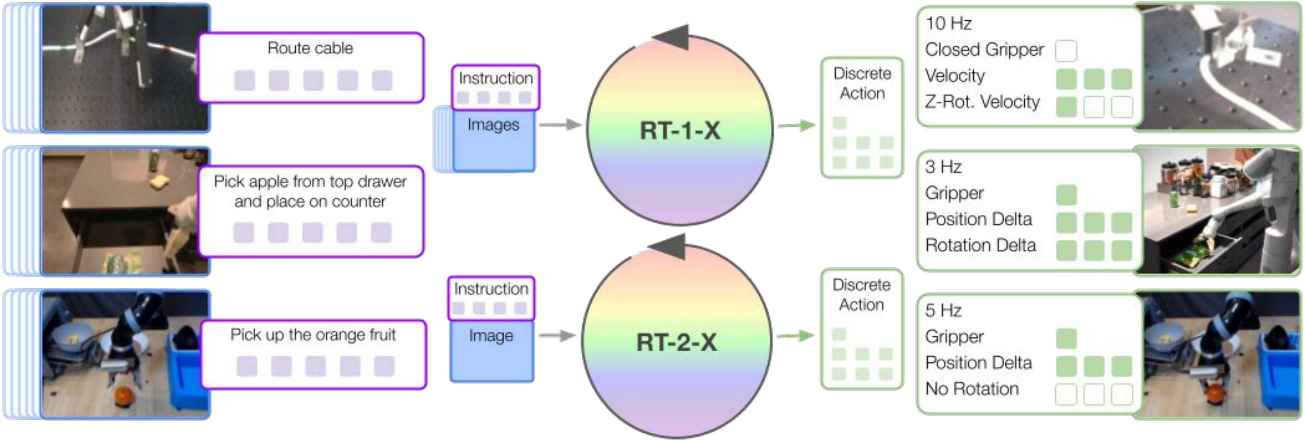
Many Scenes



Many Skills



Model Architectures



Inputs: RGB images and text instructions

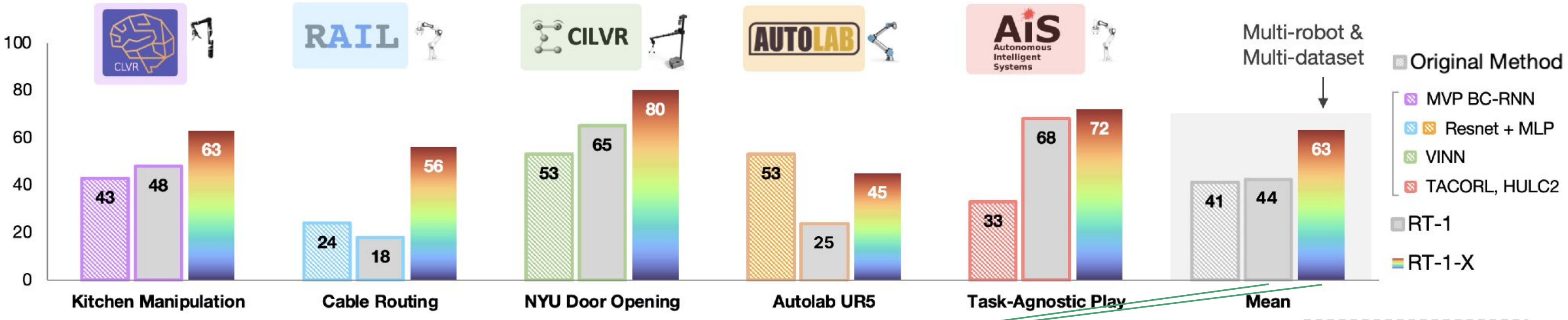
Outputs: discretized end-effector actions

Just RT-1 and RT-2 trained on X-Embodiment datasets

Velocity, delta position, absolute position

Different evaluations run at different frequencies

Results: Signs of Positive Transfer

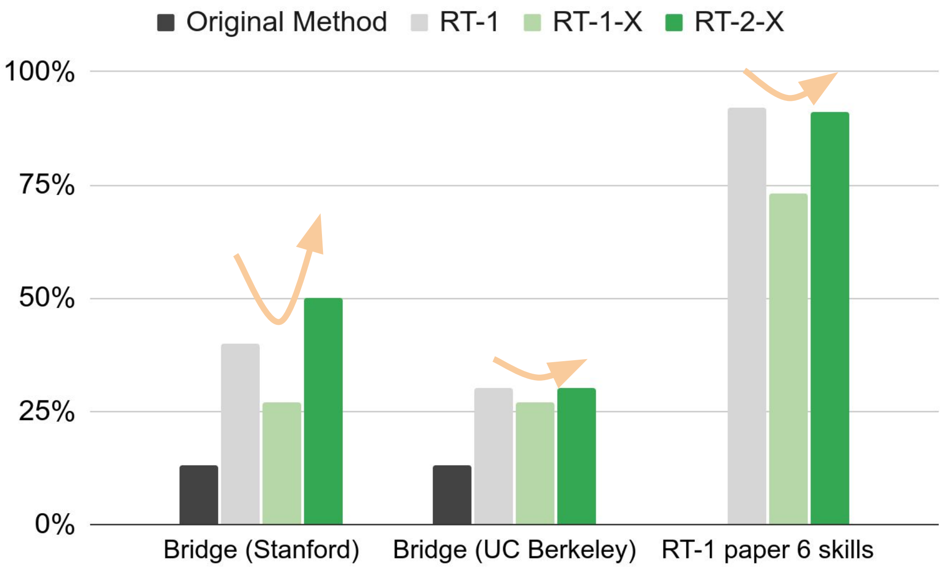


Generalist (RT-1-X) vs. Specialists (RT-1, Baselines)

50% improvement

- Training on data from **all robots** outperforms training on data from the particular evaluation robot

Results: Small Models Underfit

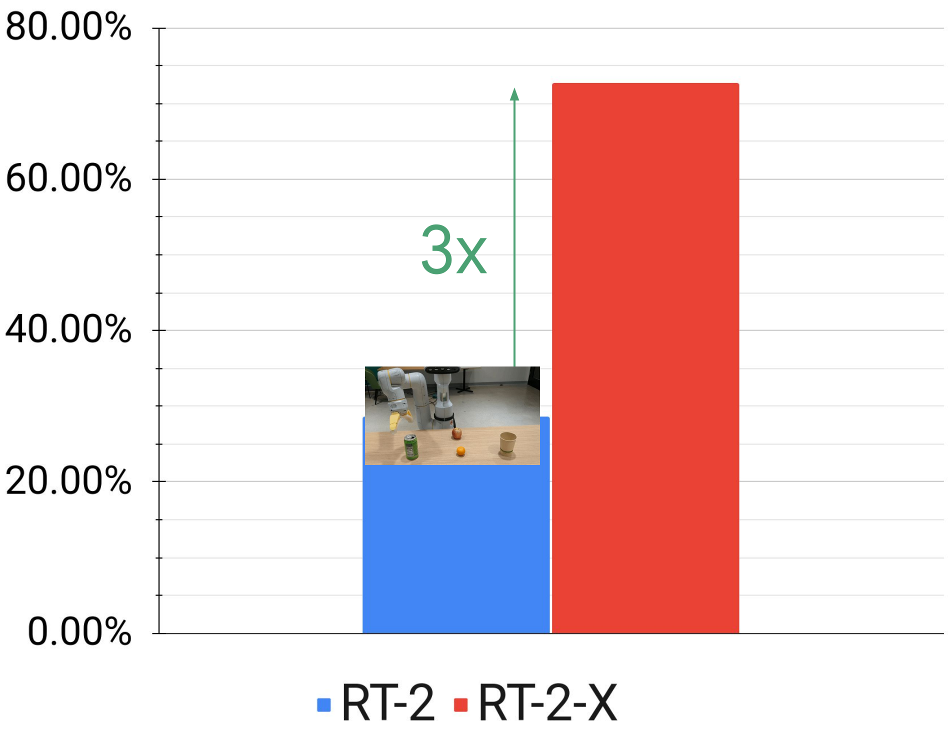


RT-1-X underfits for large datasets

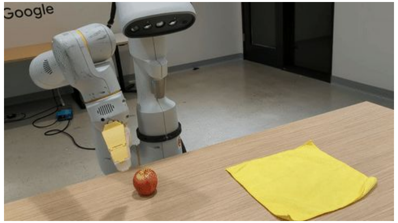
RT-2-X recovers performance

Is Web-scale Data Sufficient?

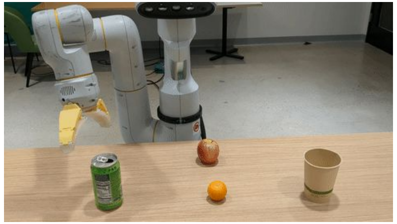
RT-2-X outperforms RT-2 by 3x in emergent skill evaluations



put apple on cloth



move apple near cloth



move apple between cup and apple

Data Scaling and Positive Transfer Recap

Real-world robot demonstration dataset → Co-train on robot data alongside internet data → Add robot data from different embodiments

[RT-1] *[RT-2]* *[RT-X]*

Increasing data interoperability by treating robot actions as just another data modality

Data Scaling and Positive Transfer Recap

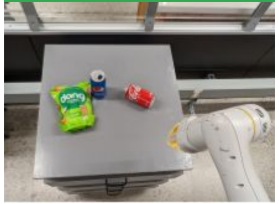
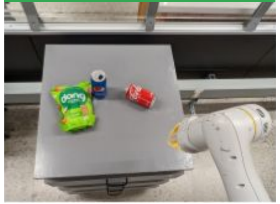
Real-world robot demonstration dataset → Co-train on robot data alongside internet data → Add robot data from different embodiments

[RT-1] *[RT-2]* *[RT-X]*

Better **in-distribution** performance → Generalization to internet **semantics** → Generalization to **spatial** concepts

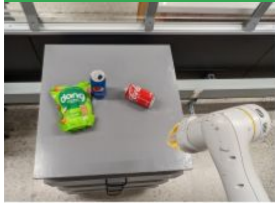
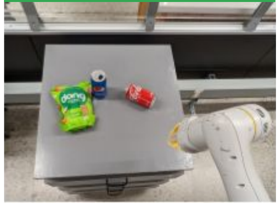
...But Many Open Challenges!

VLAs overfit to robotics data distributions

VQA Image	Robot Image	Robot Image
		
VQA Prompt	**Action Prompt**	**VQA Prompt**
Q: What is happening in the image?	Q: What action should the robot take to pick coke can?	Q: What is happening in the image?
A: A grey donkey walks down the street. ✓	A: 1 130 129 121 131 127 128 127 ✓	A: 1 127 127 127 127 127 127 127 ✗

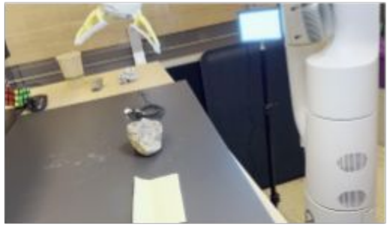
…But Many Open Challenges!

VLAs overfit to robotics data distributions

VQA Image	Robot Image	Robot Image
		
VQA Prompt	**Action Prompt**	**VQA Prompt**
Q: What is happening in the image?	Q: What action should the robot take to pick coke can?	Q: What is happening in the image?
A: A grey donkey walks down the street. ✓	A: 1 130 129 121 131 127 128 127 ✓	A: 1 127 127 127 127 127 127 127 ✗

Reasoning mixes unpredictably with low-level robot action control

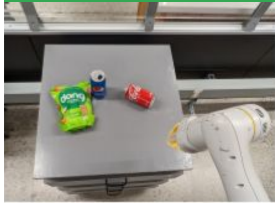
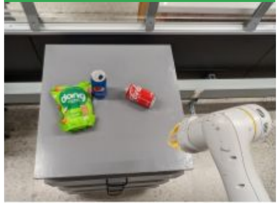
Prompt:
Given <img> I need to hammer a nail, what object from the scene might be useful?
Prediction:
Rocks. Action: 1 129 138 122 132 135 106 127



If coke can added to scene, planning works but actions break!

...But Many Open Challenges!

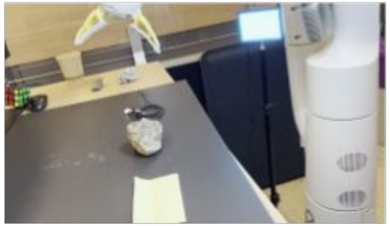
VLAs overfit to robotics data distributions

VQA Image	Robot Image	Robot Image
		
VQA Prompt	**Action Prompt**	**VQA Prompt**
Q: What is happening in the image?	Q: What action should the robot take to pick coke can?	Q: What is happening in the image?
A: A grey donkey walks down the street. ✓	A: 1 130 129 121 131 127 128 127 ✓	A: 1 127 127 127 127 127 127 127 ✗

[3] Grounding Multimodal Large Language Models in Actions, Szot et al., 2024.

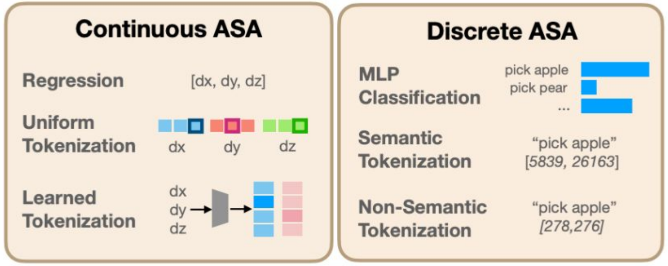
Reasoning mixes unpredictably with low-level robot action control

Prompt:
Given <img> I need to hammer a nail, what object from the scene might be useful?
Prediction:
Rocks. Action: 1 129 138 122 132 135 106 127



If coke can added to scene, planning works but actions break!

Action representations and tokenization decision choices are underexplored



Agenda

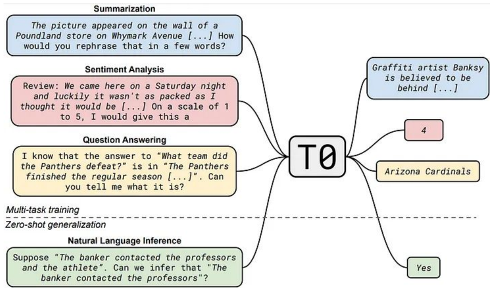
Google

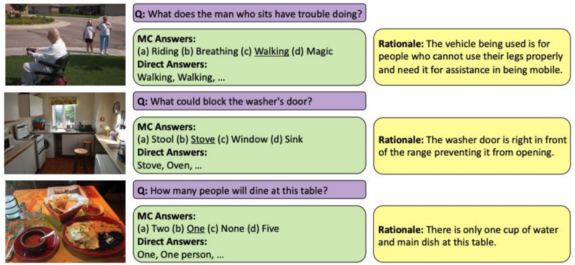
We convey intent to robot policies via very <u>constrained</u> interfaces…

…but LLM reasoning is enabled by large context <u>bandwidths</u>.

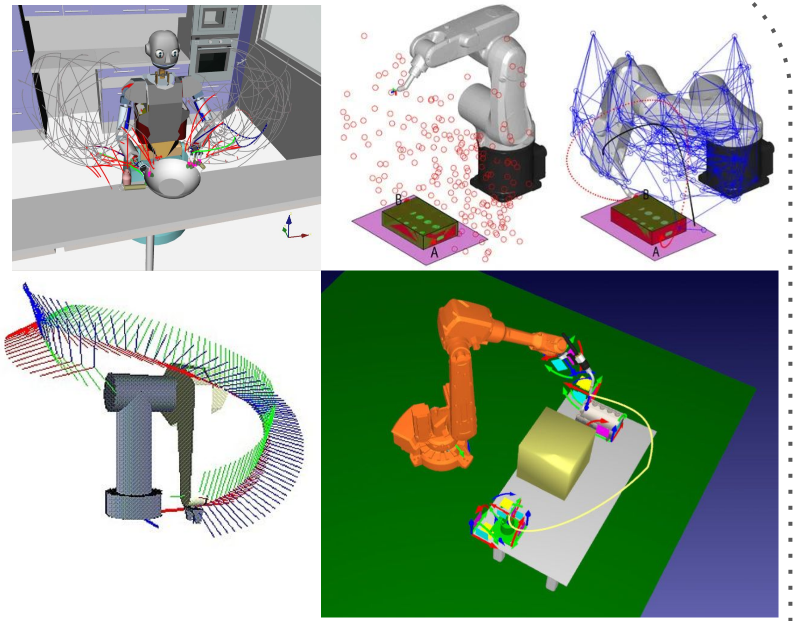
Where is my promptable generalist robot??

Strengths and Limitations of Language





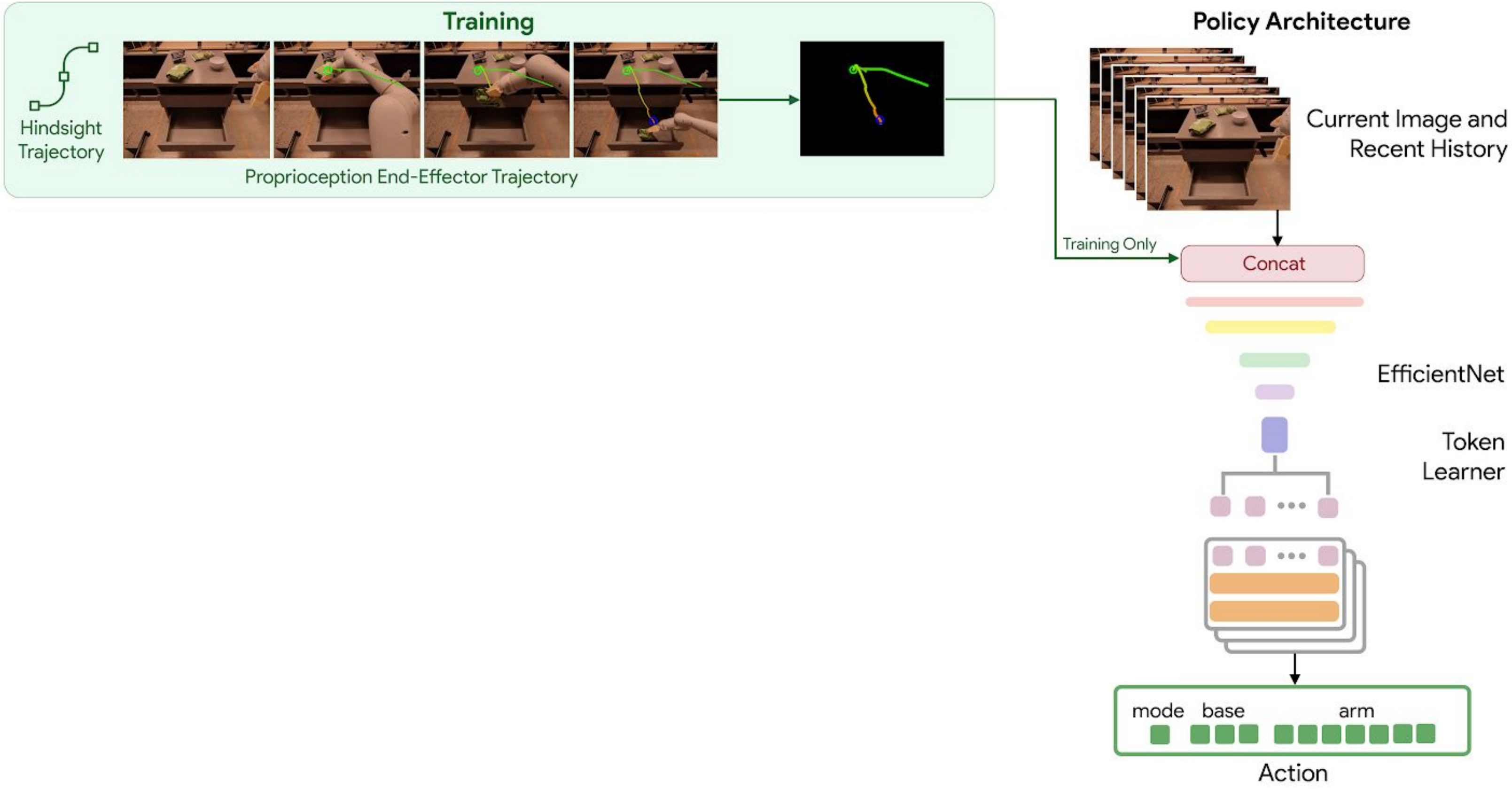
High-level Language Knowledge



Low-level Robotics Knowledge

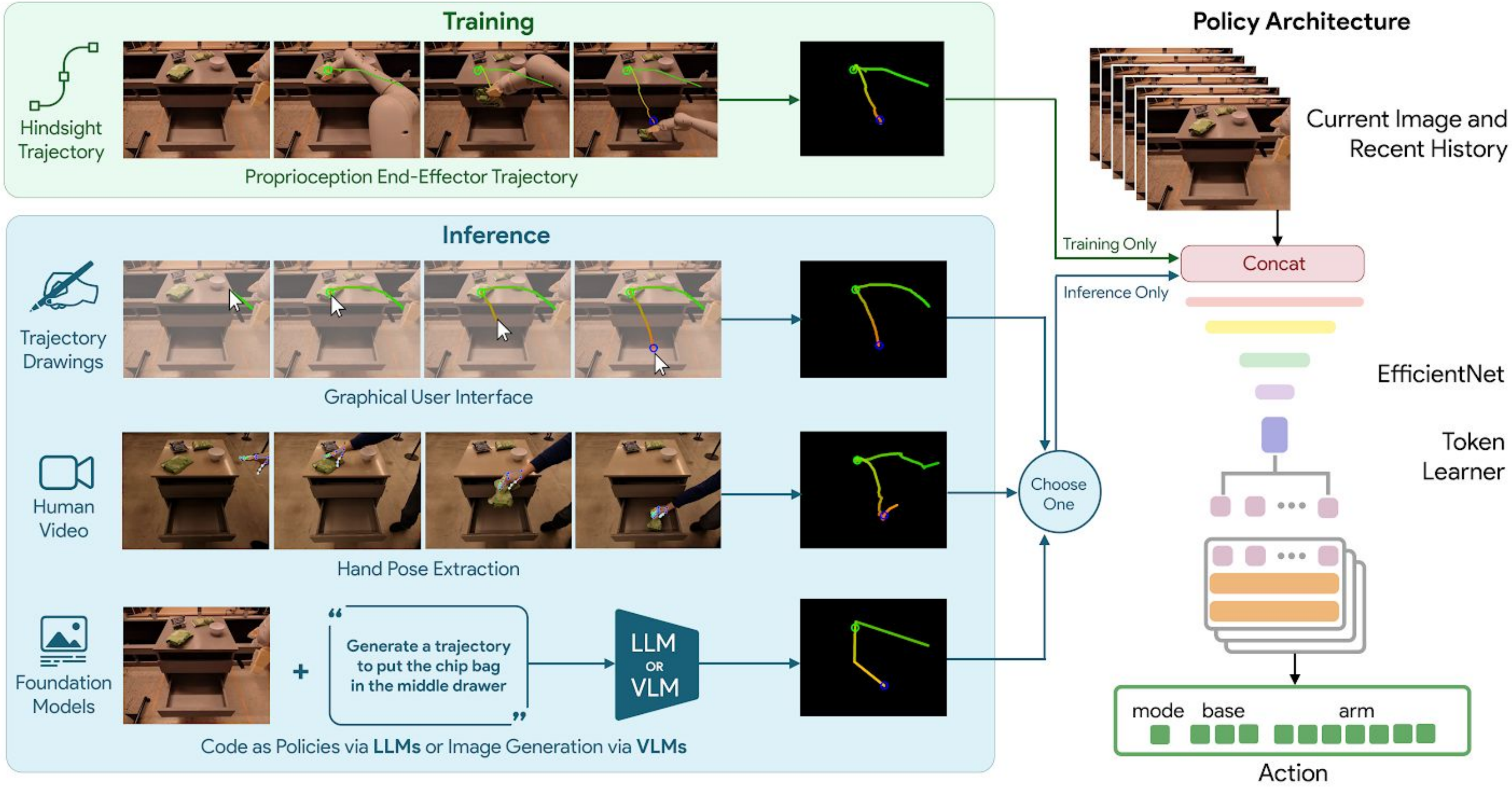
Motion-centric Representations: Hindsight Trajectories

RT-Trajectory

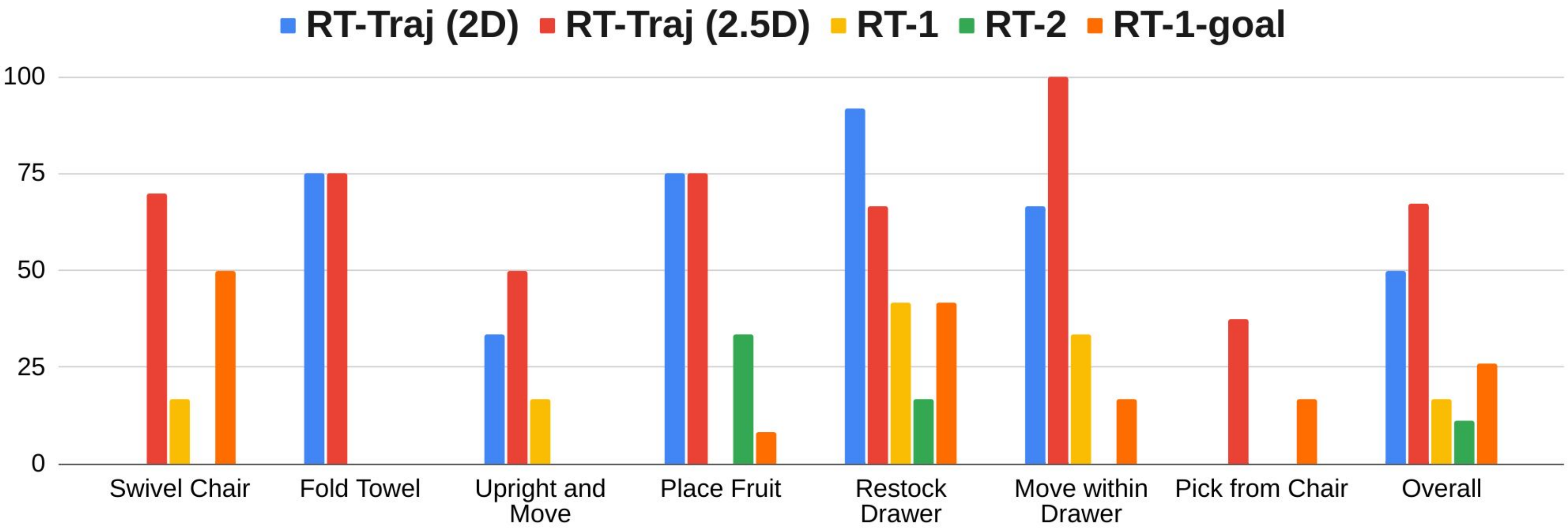


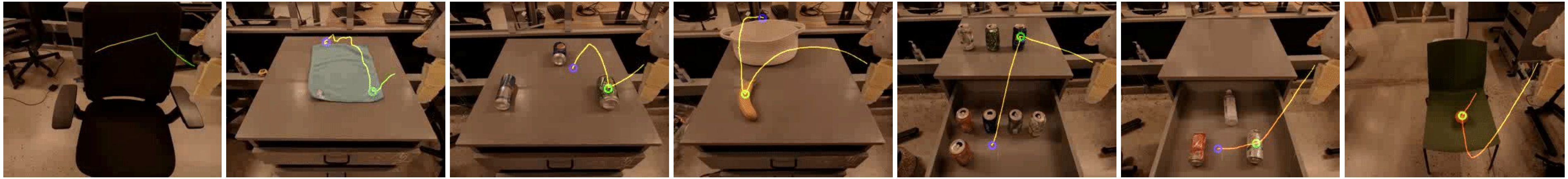
Motion-centric Representations: Hindsight Trajectories

RT-Trajectory



Results: Quantitative Evaluations





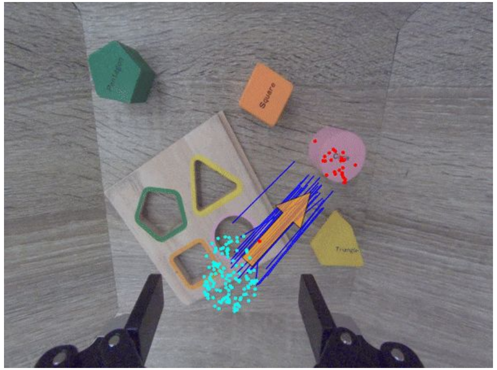
Results: Prompt Engineering via Trajectories

Ego-centric trajectory representations enable broad generalization:

- Novel motions (new heights, new shapes, new curvatures)
- Visual distribution shifts (new furniture, new rooms, new objects, new lighting)
- Behavior modulation within skills (specify exactly *how* to accomplish the task)

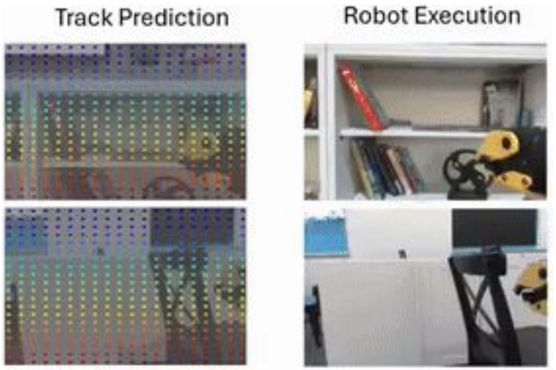
Concurrent Work: Tracks, Flow, Motion

Motions and trajectories are a powerful representation which capture the unique properties of robotics: actions, dynamics, physics, change



RoboTAP

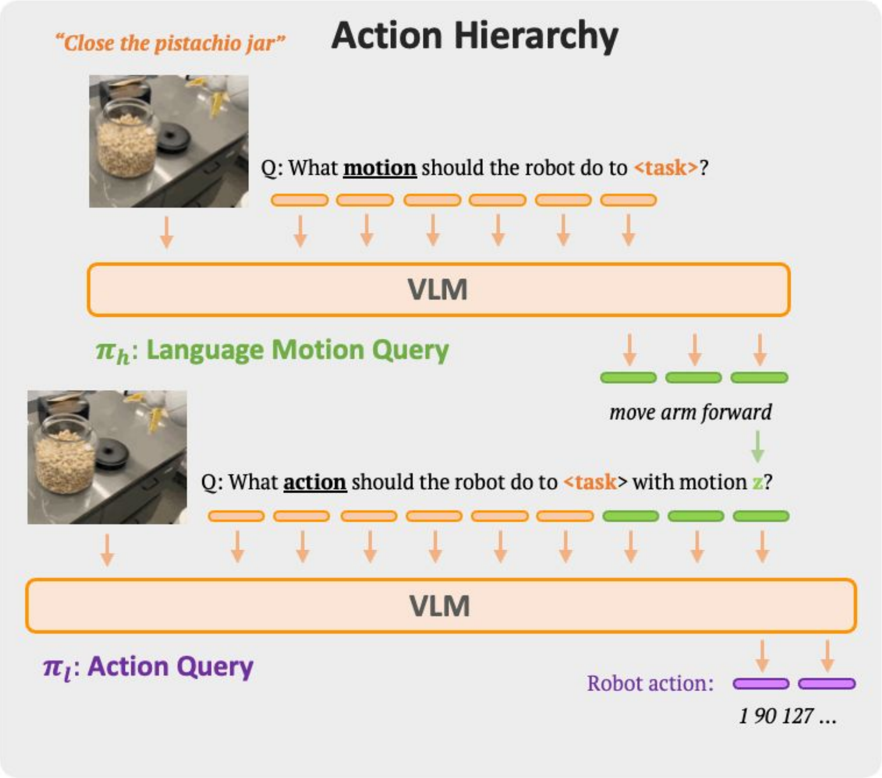
Any-point Trajectory Modeling



Track2Act

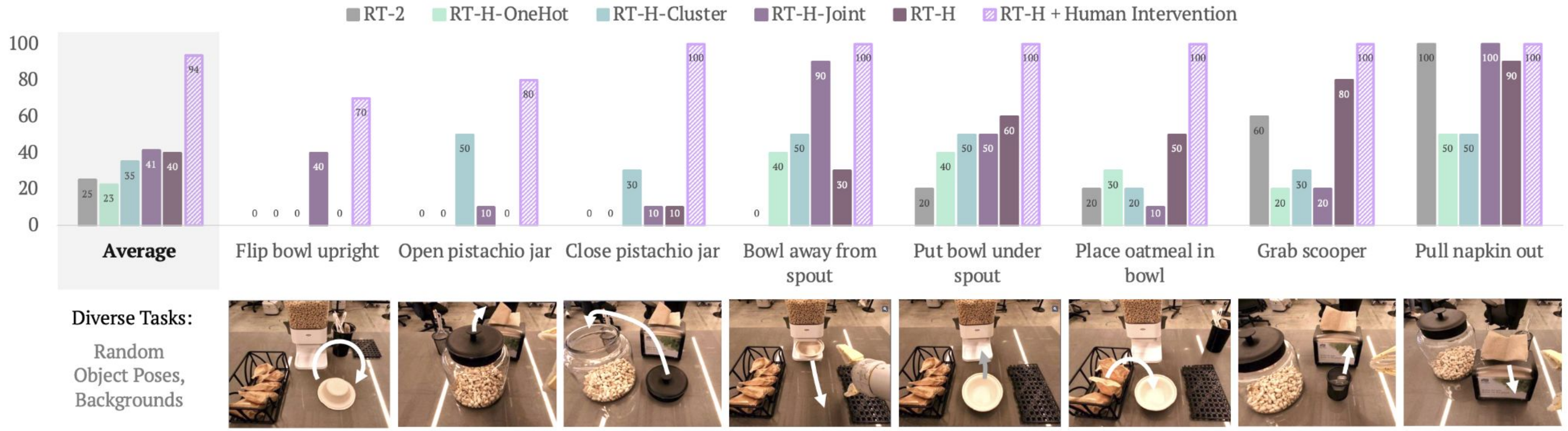
[4] RoboTAP: Tracking Arbitrary Points for Few-Shot Visual Imitation, Vecerik et al., 2023.
[5] Any-point Trajectory Modeling for Policy Learning, Wen et al., 2024.
[6] Track2Act: Predicting Point Tracks from Internet Videos enables Diverse Zero-shot Robot Manipulation, Bharadhwaj et al. 2024.

Is language enough, if it's *hierarchical* and *granular*?
RT-Hierarchy



- Idea: predict granular language motions before predicting low-level robot actions
  - "move arm forward", "rotate arm clockwise", "close gripper"
- Can be viewed as chain-of-thought / planning for language-based skills

Results: RT-H Outperforms RT-2



No other policy class (RT-1, RT-2) was able to learn from challenging new data

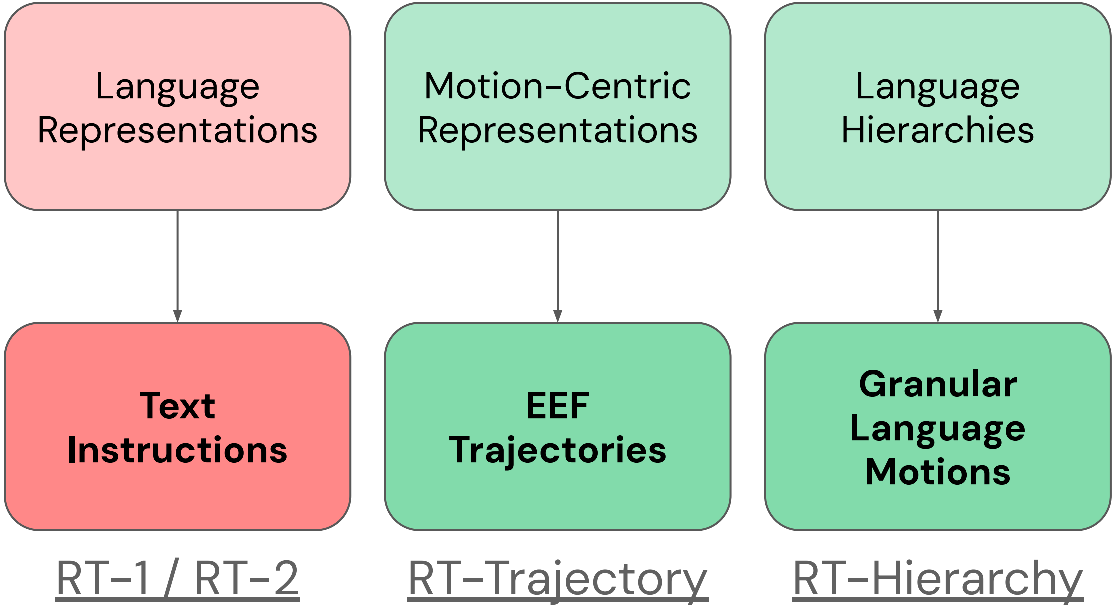
Results: Language Interventions

Task: "Close the pistachio jar"

Action Hierarchies Improve Performance and Enable Intervention

RT-H bottleneck often was language motion prediction rather than low-level action prediction: language motions easier to collect interventions for!

Steerability Recap



We have proofs of concept for promptable robots…

…but do we have enough robot data to support these algorithms?

Google

We have proofs of concept for promptable robots…

…but do we have enough robot data to support these algorithms?

Robot data is not guaranteed to be a bottleneck because we don't yet know what kind of robot data we need

Agenda

01 Why Robot Foundation Models?

02 Piece #1: Positive Transfer from Scaling

03 Piece #2: Steerability

04 Piece #3: Scalable Evaluation

05 Horizons

Google

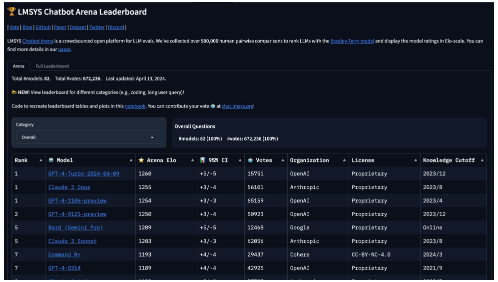
AI has an Evaluation Problem

- All roads lead to generalist models, but generalist models that can "do anything" need to be evaluated on "everything"!
- How do you <u>scalably</u> evaluate a broad set of capabilities?

LLMs Target Human Data Distribution

↓

Evaluate on Humans Directly



🏆 LMSYS Chatbot Arena Leaderboard

| Vote | Blog | GitHub | Paper | Dataset | Twitter | Discord |

LMSYS Chatbot Arena is a crowdsourced open platform for LLM evals. We've collected over **500,000** human pairwise comparisons to rank LLMs with the Bradley-Terry model and display the model ratings in Elo-scale. You can find more details in our paper.

Arena | Full Leaderboard

Total #models: **82**. Total #votes: **672,236**. Last updated: April 13, 2024.

NEW! View leaderboard for different categories (e.g., coding, long user query)!

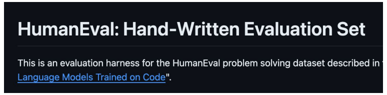
Code to recreate leaderboard tables and plots in this notebook. You can contribute your vote at chat.lmsys.org!

Category: Overall

Overall Questions

#models: 82 (100%) #votes: 672,236 (100%)

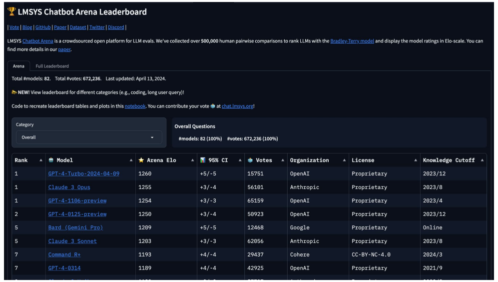
Rank	Model	Arena Elo	95% CI	Votes	Organization	License	Knowledge Cutoff
1	GPT-4-Turbo-2024-04-09	1260	+5/-5	15751	OpenAI	Proprietary	2023/12
1	Claude 3 Opus	1255	+3/-4	56101	Anthropic	Proprietary	2023/8
1	GPT-4-1106-preview	1254	+3/-3	65159	OpenAI	Proprietary	2023/4
2	GPT-4-0125-preview	1250	+3/-4	50923	OpenAI	Proprietary	2023/12
5	Bard (Gemini Pro)	1209	+5/-5	12468	Google	Proprietary	Online
5	Claude 3 Sonnet	1203	+3/-3	62056	Anthropic	Proprietary	2023/8
7	Command R+	1193	+4/-4	29437	Cohere	CC-BY-NC-4.0	2024/3
7	GPT-4-0314	1189	+4/-4	42925	OpenAI	Proprietary	2021/9



AI has an Evaluation Problem

- All roads lead to generalist models, but generalist models that can "do anything" need to be evaluated on "everything"!
- How do you scalably evaluate a broad set of capabilities?

LLMs Target Human Data Distribution	Robots Target Physical Data Distribution
↓	↓
Evaluate on Humans Directly	Evaluate on ???


🏆 LMSYS Chatbot Arena Leaderboard

| Vote | Blog | GitHub | Paper | Dataset | Twitter | Discord |

LMSYS Chatbot Arena is a crowdsourced open platform for LLM evals. We've collected over **500,000** human pairwise comparisons to rank LLMs with the Bradley-Terry model and display the model ratings in Elo-scale. You can find more details in our paper.

Arena | Full Leaderboard

Total #models: **82**. Total #votes: **672,236**. Last updated: April 13, 2024.

NEW! View leaderboard for different categories (e.g., coding, long user query)!

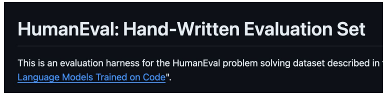
Code to recreate leaderboard tables and plots in this notebook. You can contribute your vote at chat.lmsys.org!

Category: Overall

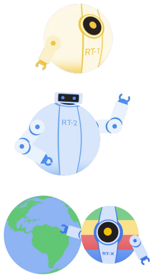
Overall Questions

#models: 82 (100%) #votes: 672,236 (100%)

Rank	Model	Arena Elo	95% CI	Votes	Organization	License	Knowledge Cutoff
1	GPT-4-Turbo-2024-04-09	1260	+5/-5	15751	OpenAI	Proprietary	2023/12
1	Claude 3 Opus	1255	+3/-4	56101	Anthropic	Proprietary	2023/8
1	GPT-4-1106-preview	1254	+3/-3	65159	OpenAI	Proprietary	2023/4
2	GPT-4-0125-preview	1250	+3/-4	50923	OpenAI	Proprietary	2023/12
5	Bard (Gemini Pro)	1209	+5/-5	12468	Google	Proprietary	Online
5	Claude 3 Sonnet	1203	+3/-3	62056	Anthropic	Proprietary	2023/8
7	Command R+	1193	+4/-4	29437	Cohere	CC-BY-NC-4.0	2024/3
7	GPT-4-0314	1189	+4/-4	42925	OpenAI	Proprietary	2021/9


HumanEval: Hand-Written Evaluation Set

This is an evaluation harness for the HumanEval problem solving dataset described in t Language Models Trained on Code".



RT-1: 3,000 Trials

RT-2: 6,000 Trials

RT-X: 3,600 Trials

Measuring Axes of Generalization

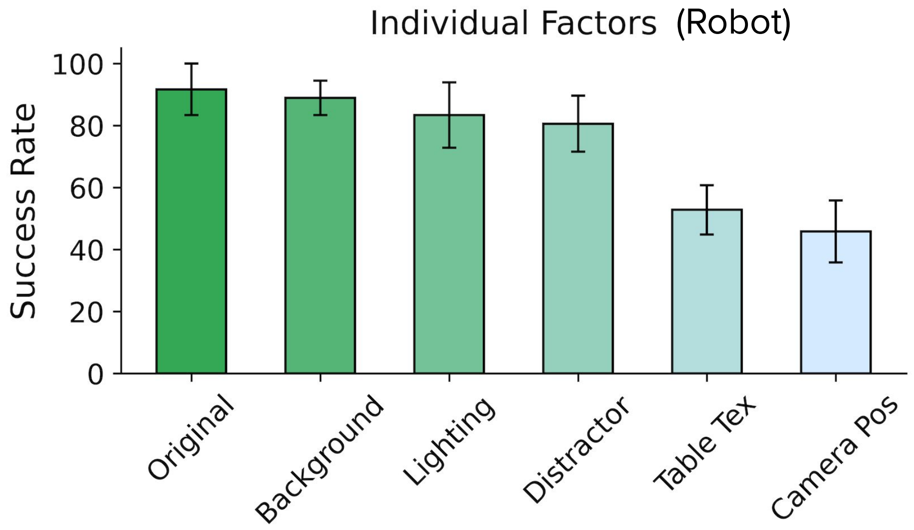
Can we *systematically* measure policy generalization?

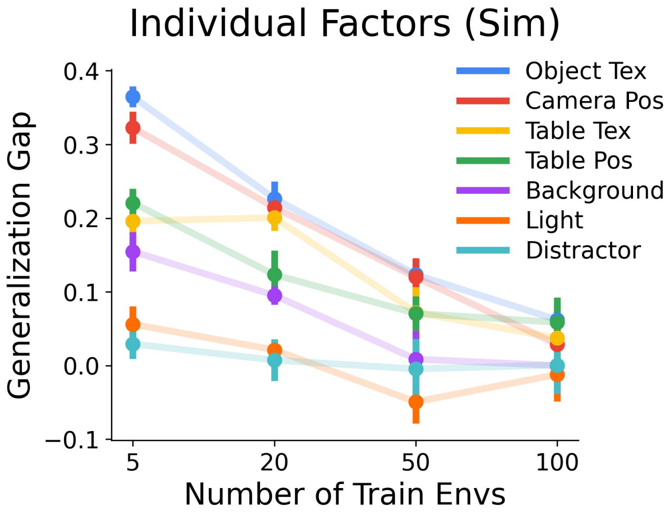
Real Robot

Table (x3) Background (x3) Distractors (x3) Lighting (x2) Camera Pose (x3)

Evaluation Metrics: success rate, generalization gap (train – test success rate)

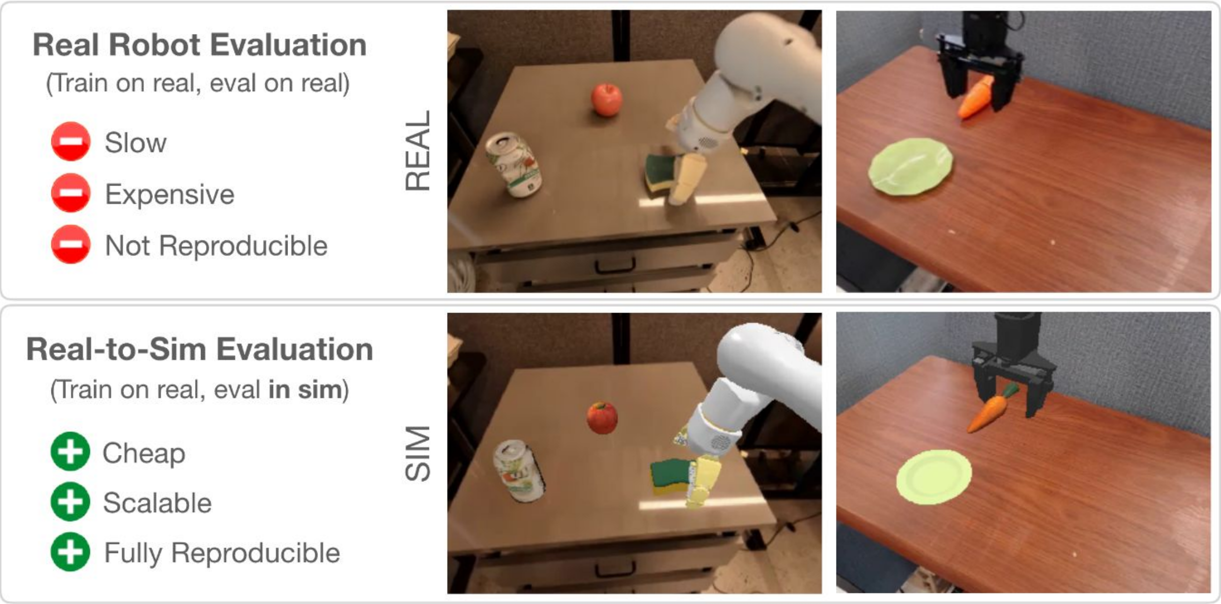
Impact of Individual Factors

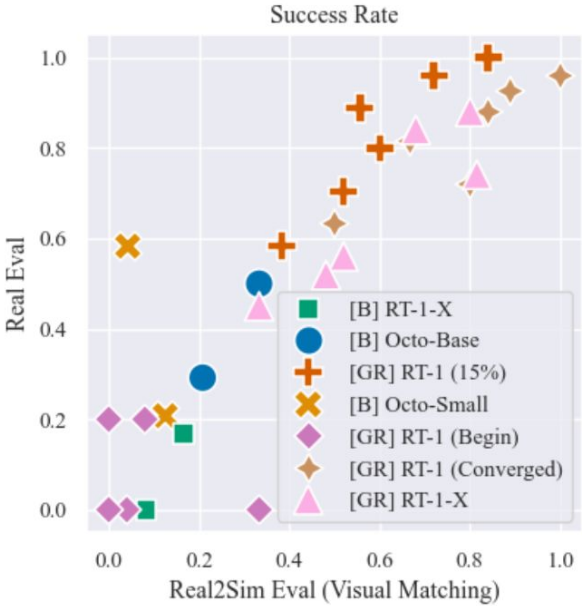




"Easier" factors: background, lighting, distractor
"Harder" factors: table position, table texture, camera position, object texture

Real-to-Sim Evaluation for Real-world Robot Policies





Key Insight: A simulation "good enough" for useful evaluation signal may be much easier to build than a full digital clone for training

World Models for Evaluation

PRISM-1

UniSim

Genie

[4] PRISM-1, Wayve, 2024
[5] UniSim: Learning Interactive Real-World Simulators, Yang et al., 2024
[6] Genie: Generative Interactive Environments, Bruce et al., 2024

World Models for Evaluation

PRISM-1

UniSim

Genie

Real world evaluations will always be the gold standard.
Scaled evaluations will be solved by unit economics and products.

[4] PRISM-1, Wayve, 2024
[5] UniSim: Learning Interactive Real-World Simulators, Yang et al., 2024
[6] Genie: Generative Interactive Environments, Bruce et al., 2024

Agenda

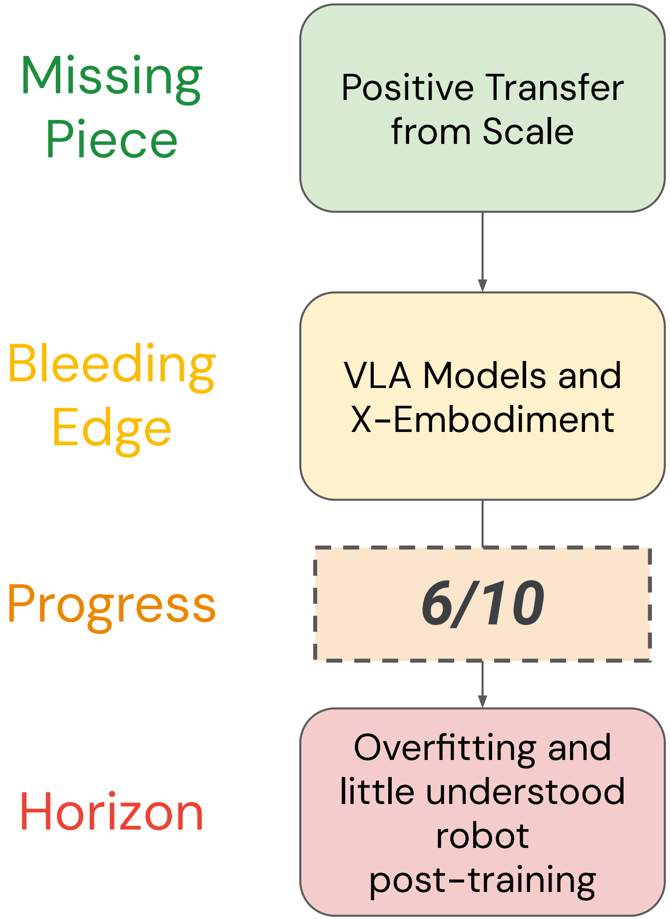
01 Why Robot Foundation Models?

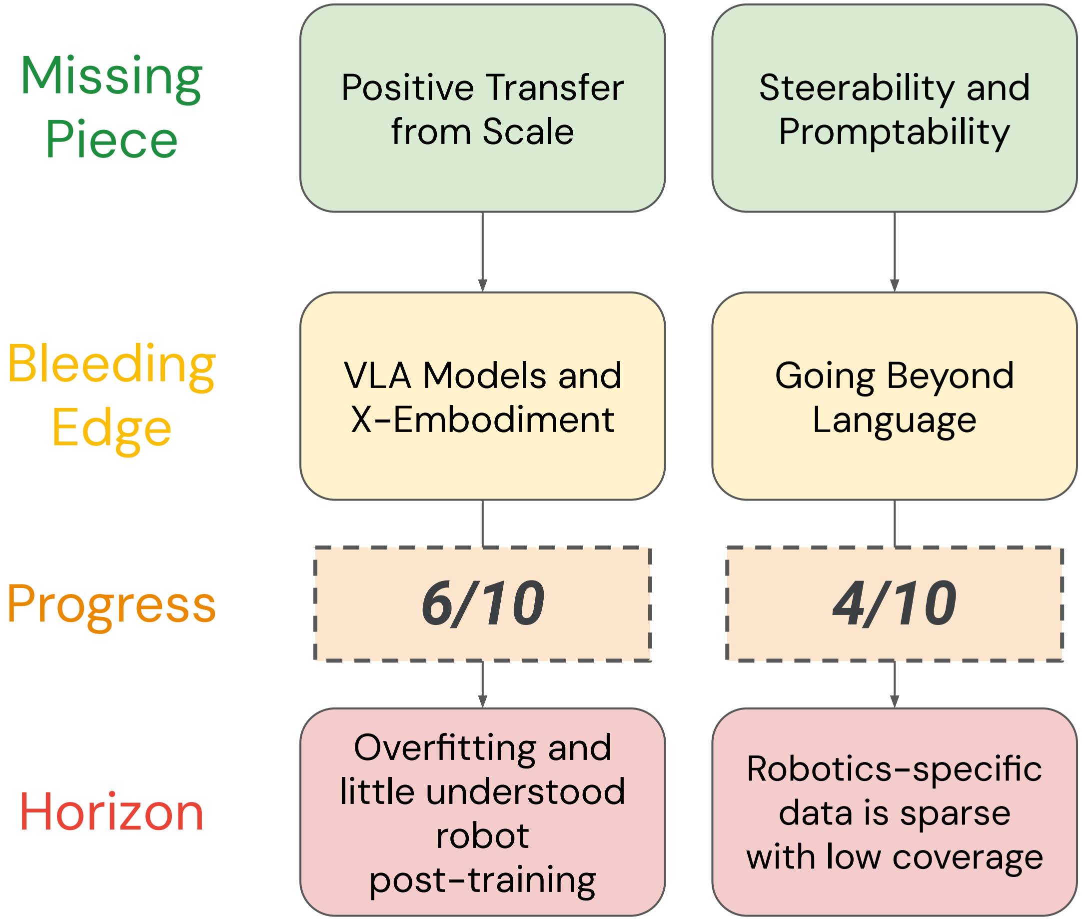
02 Piece #1: Positive Transfer from Scaling

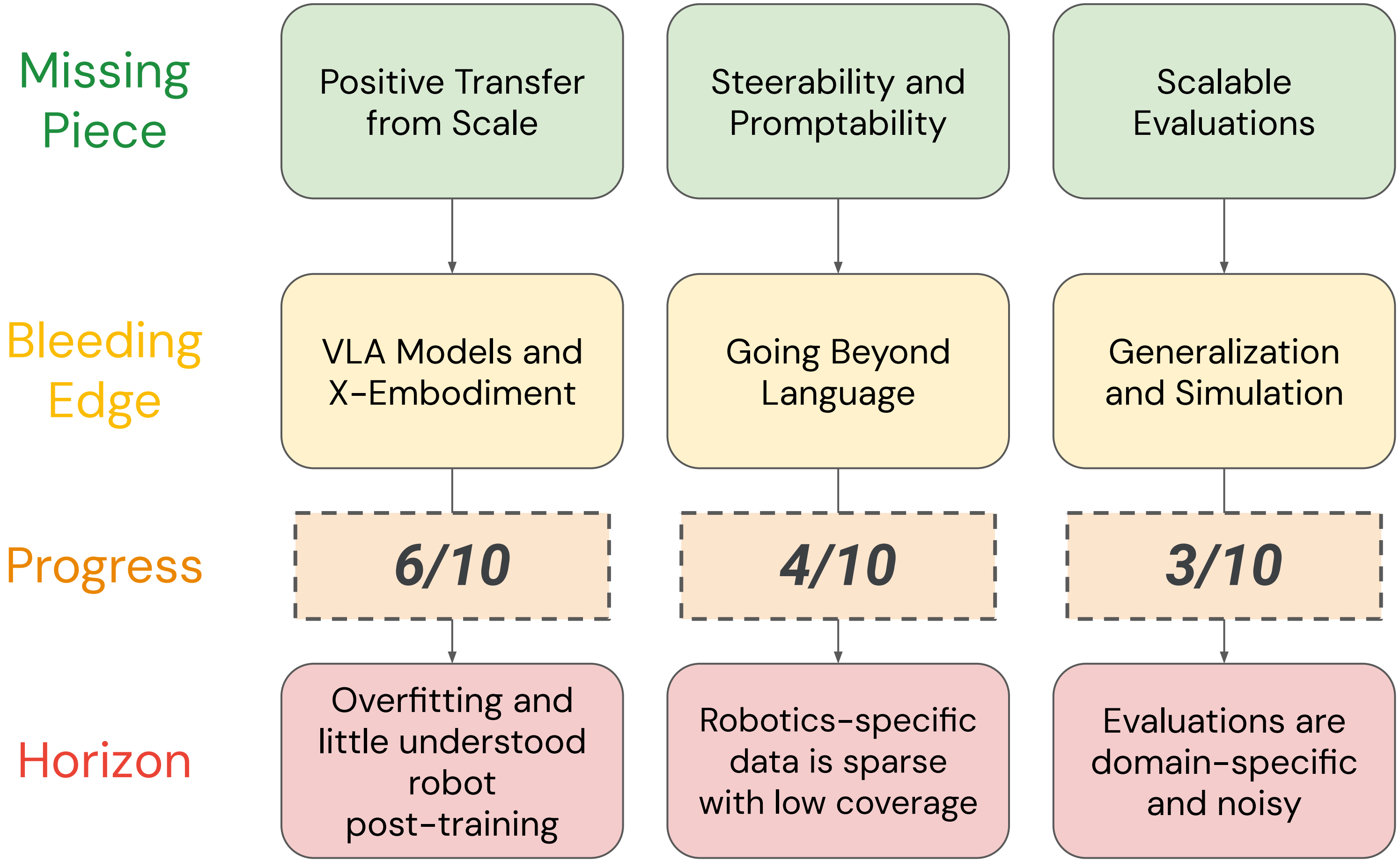
03 Piece #2: Steerability

04 Piece #3: Scalable Evaluation

05 Horizons


Missing Piece
Positive Transfer from Scale
Bleeding Edge
VLA Models and X-Embodiment
Progress
6/10
Horizon
Overfitting and little understood robot post-training


Missing Piece
Positive Transfer from Scale
Steerability and Promptability
Bleeding Edge
VLA Models and X-Embodiment
Going Beyond Language
Progress
6/10
4/10
Horizon
Overfitting and little understood robot post-training
Robotics-specific data is sparse with low coverage


Missing Piece
Positive Transfer from Scale
Steerability and Promptability
Scalable Evaluations
Bleeding Edge
VLA Models and X-Embodiment
Going Beyond Language
Generalization and Simulation
Progress
6/10
4/10
3/10
Horizon
Overfitting and little understood robot post-training
Robotics-specific data is sparse with low coverage
Evaluations are domain-specific and noisy

Google

Predictions

Overfitting and little understood robot post-training

Robotics research splits into pre-training and post-training

Robotics-specific data is sparse with low coverage

Robot data engines accelerated by industry and startups

Evaluations are domain-specific and noisy

Evaluations via simulators/world models vs. product deployments

Thank you!

tedxiao@google.com

Google DeepMind