上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Lecture 40: Robotics Foundation Models

Tao Huang

John Hopcroft Center, School of Computer Science, Shanghai Jiao Tong University

https://taohuang.info/cs3317

https://oc.sjtu.edu.cn/courses/89538

AI tools assisted in generating some figures in these slides. All such content has been reviewed, and the instructor is responsible for its accuracy.

Where We Are

Sim-to-real got one skill onto one robot. L40 asks whether one model can hold them all.

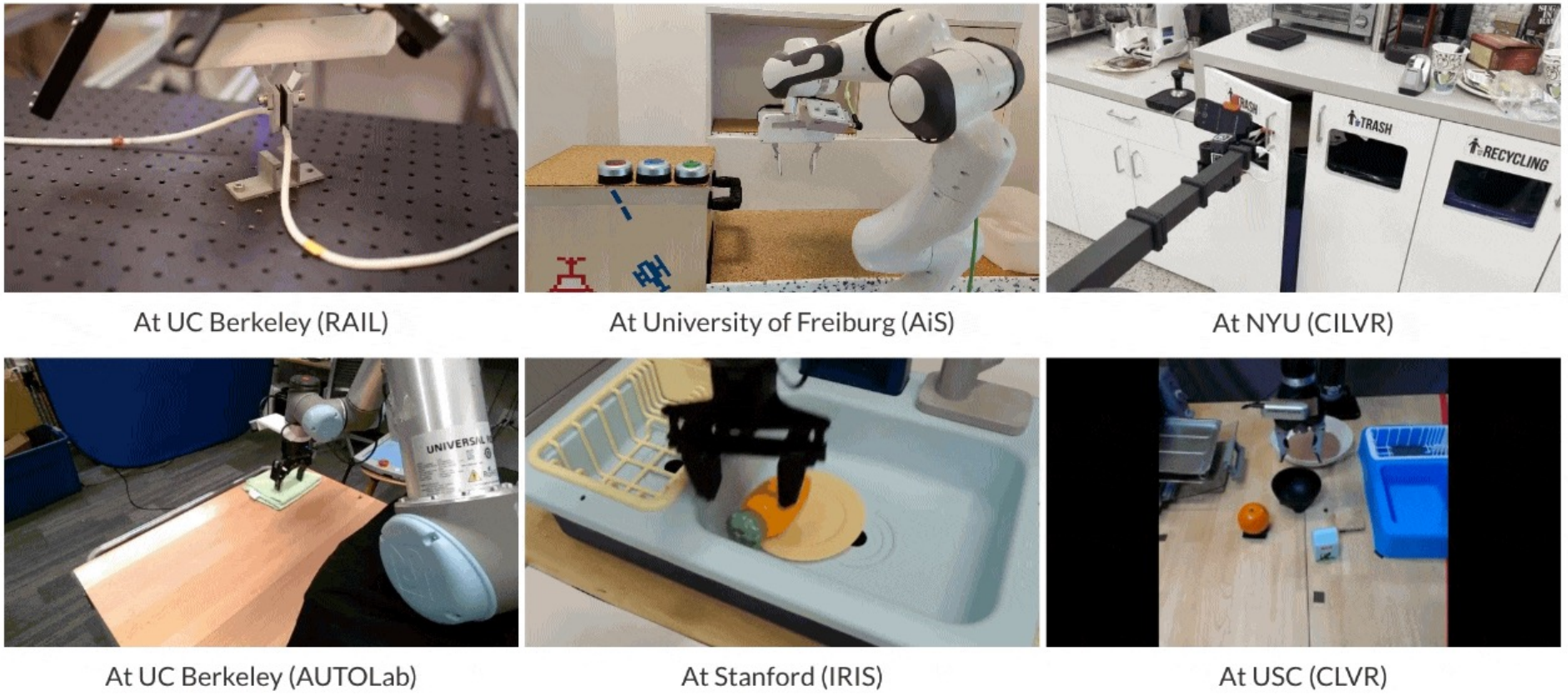
- L38 — mapped perception → action on one robot (a VLA policy).
- L39 — fed it cheaply with simulation, closing the reality gap.
- Both still give one policy, for one robot, on tasks it was trained on.

- **Today:** the leap to a foundation model — pretrain once, generalize broadly.

Same Brain, Different Bodies

One trained model — driving a dozen completely different robot bodies.

RT-1-X evaluation on in-distribution skills



At UC Berkeley (RAIL) | At University of Freiburg (AiS) | At NYU (CILVR)

At UC Berkeley (AUTOLab) | At Stanford (IRIS) | At USC (CLVR)

RT-1-X performing diverse tasks in 6 academic labs

GPT Had the Internet. Robots Don't.

GPT had the internet. ImageNet had 1.4M labeled photos.
Robots have neither.

Foundation models swallowed language and vision by pretraining once and adapting everywhere.
Why hasn't that happened for robots — and is it happening now?

Objectives

By the end of this lecture, you will be able to:

- Explain what makes a model a “foundation model,” and why robotics resisted one.
- Distinguish three axes of robot generalization: cross-task, cross-environment, cross-embodiment.
- Analyze the robot-data bottleneck and the three strategies attacking it.
- Evaluate where today's generalist policies actually generalize — and where claims outrun evidence.
- Sketch the data-flywheel argument for why the field bets that scaling will work.

1. The Recipe That Ate AI

The Foundation-Model Recipe

- In language and vision, the field **consolidated**: pretrain one big backbone on web-scale data (GPT, BERT, CLIP, ViT) — then adapt it cheaply to thousands of downstream tasks.

Foundation model = broad pretraining → broad transfer, often with emergent generalization.

Same scaling-law recipe you saw in the LLM module: scale + diversity → capability.

Why Robotics Resisted

- **Situation:** NLP & vision consolidated onto shared pretrained backbones.
- **Complication:** Robotics didn't — a separate model per robot, per task, per room.
- **Why:** there is no internet of robot actions. Data is physical — slow, costly, safety-bound, and welded to one specific body.

Every lab's dataset is tiny — and incompatible with the next lab's robot.

The Bet — and Three Axes

The bet: the same recipe works for robots — if we solve the data problem.

- Cross-task — “wipe the table” → “fold the towel” (a new skill)
- Cross-environment — a trained kitchen → a kitchen it's never seen (a new scene)
- Cross-embodiment — a Franka arm → a humanoid (a new body)

2. One Brain, Many Bodies

Trapped in Silos

- **Situation:** a policy trained on Robot A is useless on Robot B — different cameras, arms, grippers, action spaces.
- **Complication:** each lab's data stays siloed — and each silo is tiny.

ImageNet ≈ 1.4M images from one pipeline. The biggest robot datasets: a few thousand episodes on a single arm.

Pool Everything: Open X-Embodiment

Question: what if dozens of labs pooled their data and trained one model across many robot bodies?

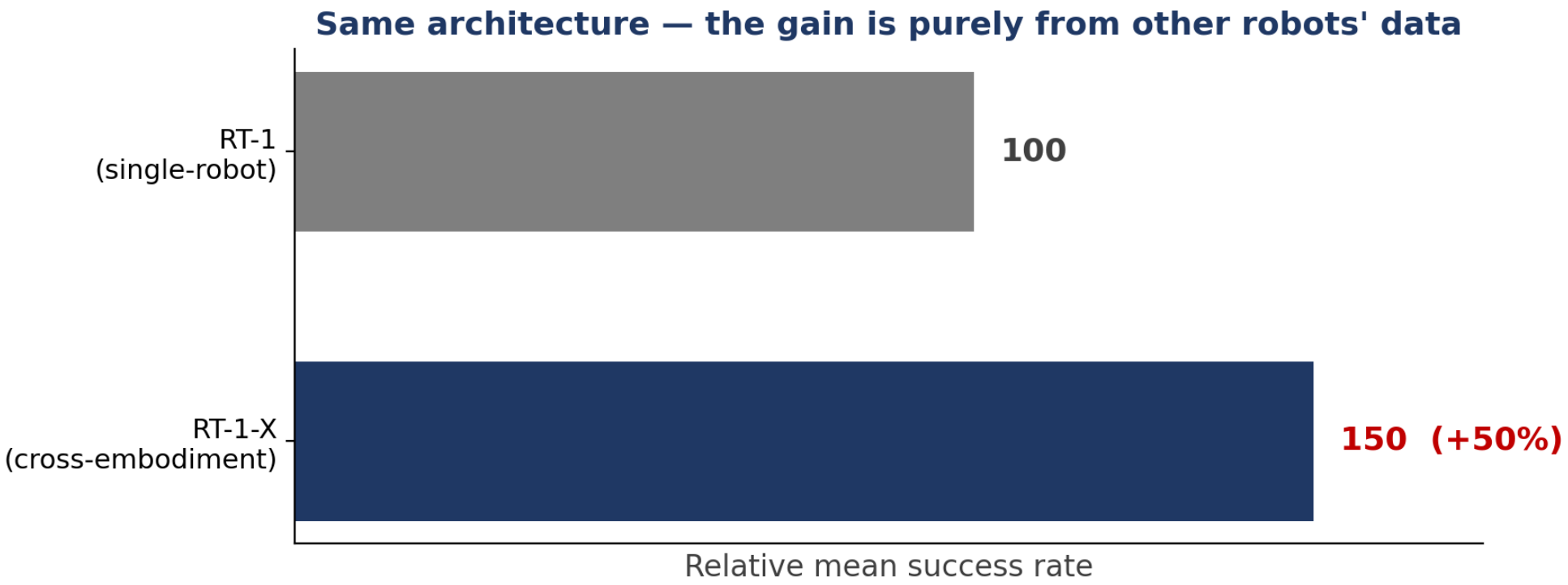
Answer: Open X-Embodiment / RT-X (Google DeepMind + the community).

1M+ trajectories · 22 robot embodiments · 60 datasets · 34 labs

The first serious "robotics ImageNet."

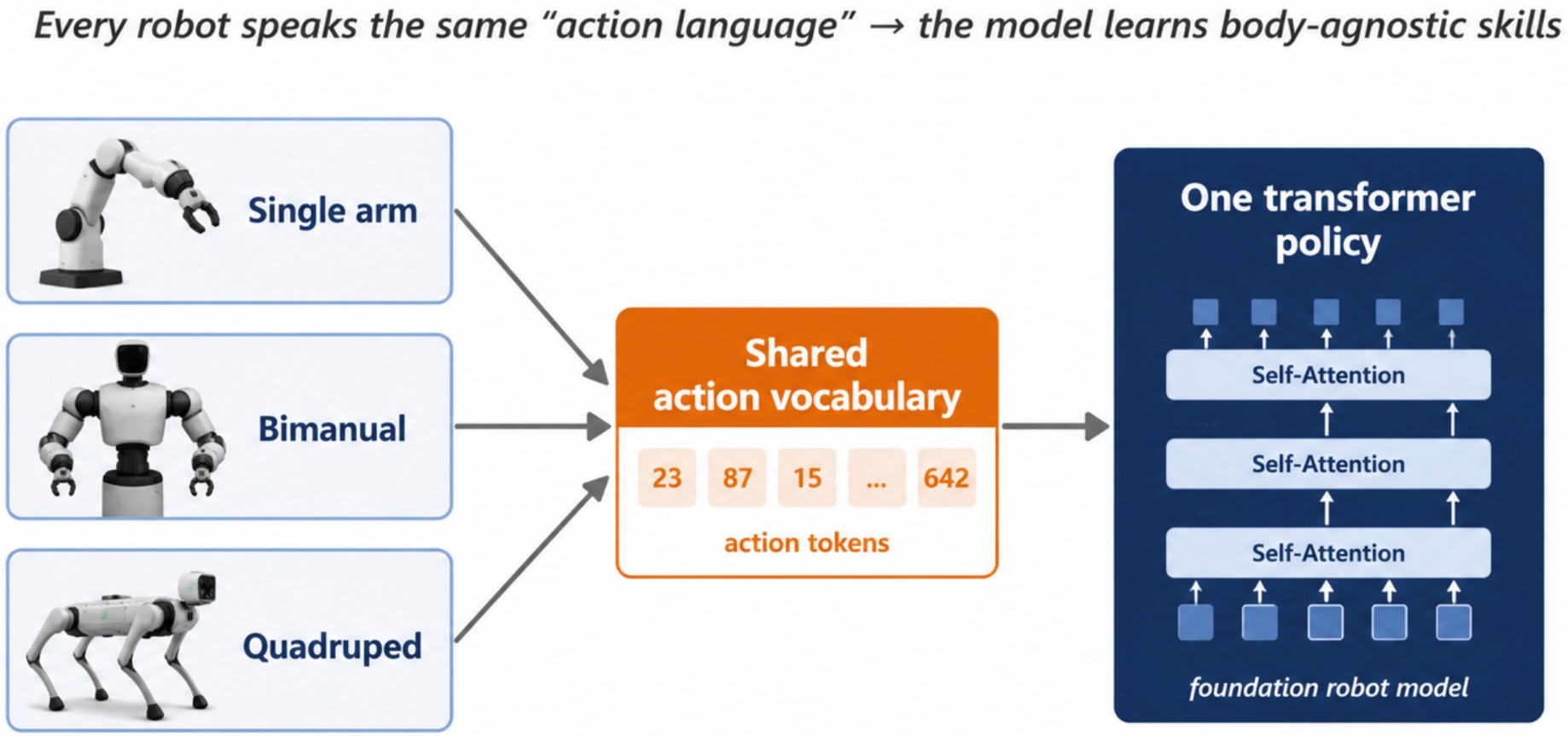
Positive Transfer

Skills learned on one body help another — pooling beats siloing.



A Shared Action Language

How can different robots share one model? Give them a common vocabulary.

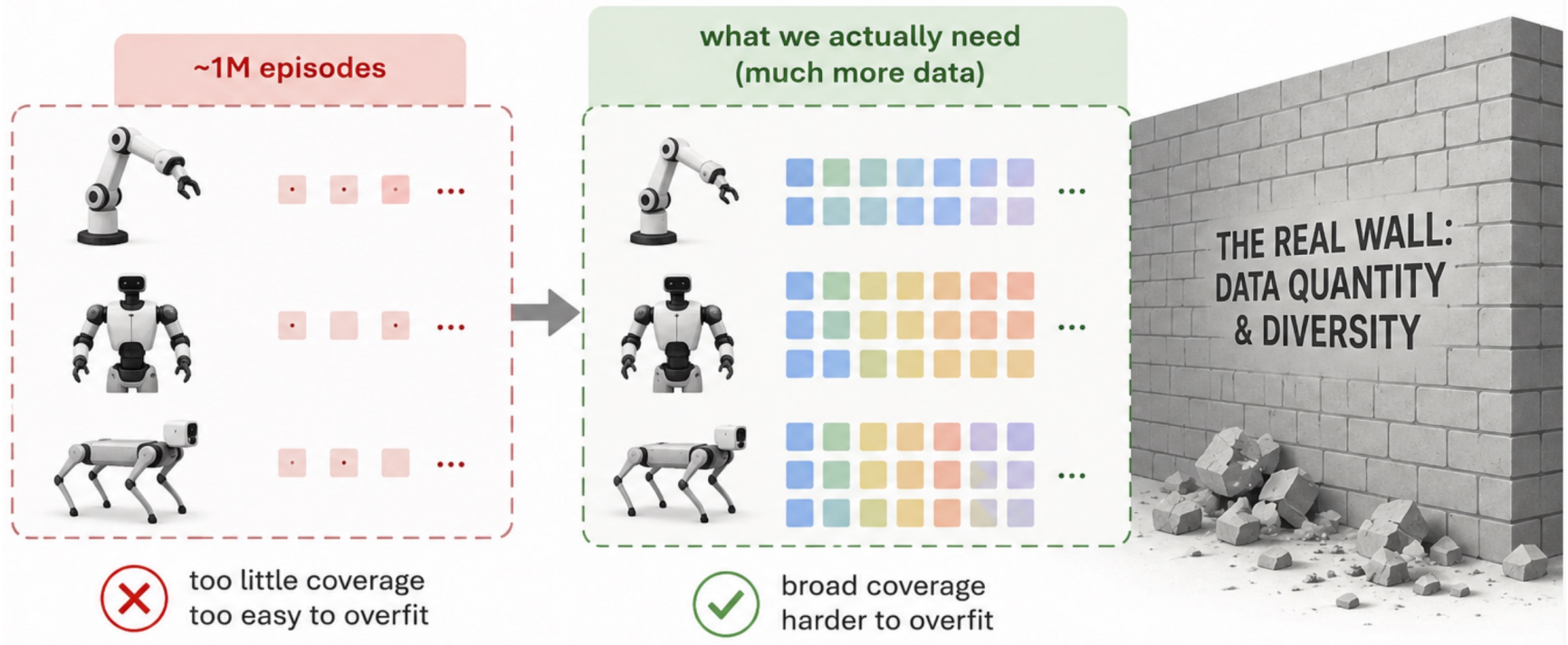


3. The Data Wall, Quantified

Still Too Small

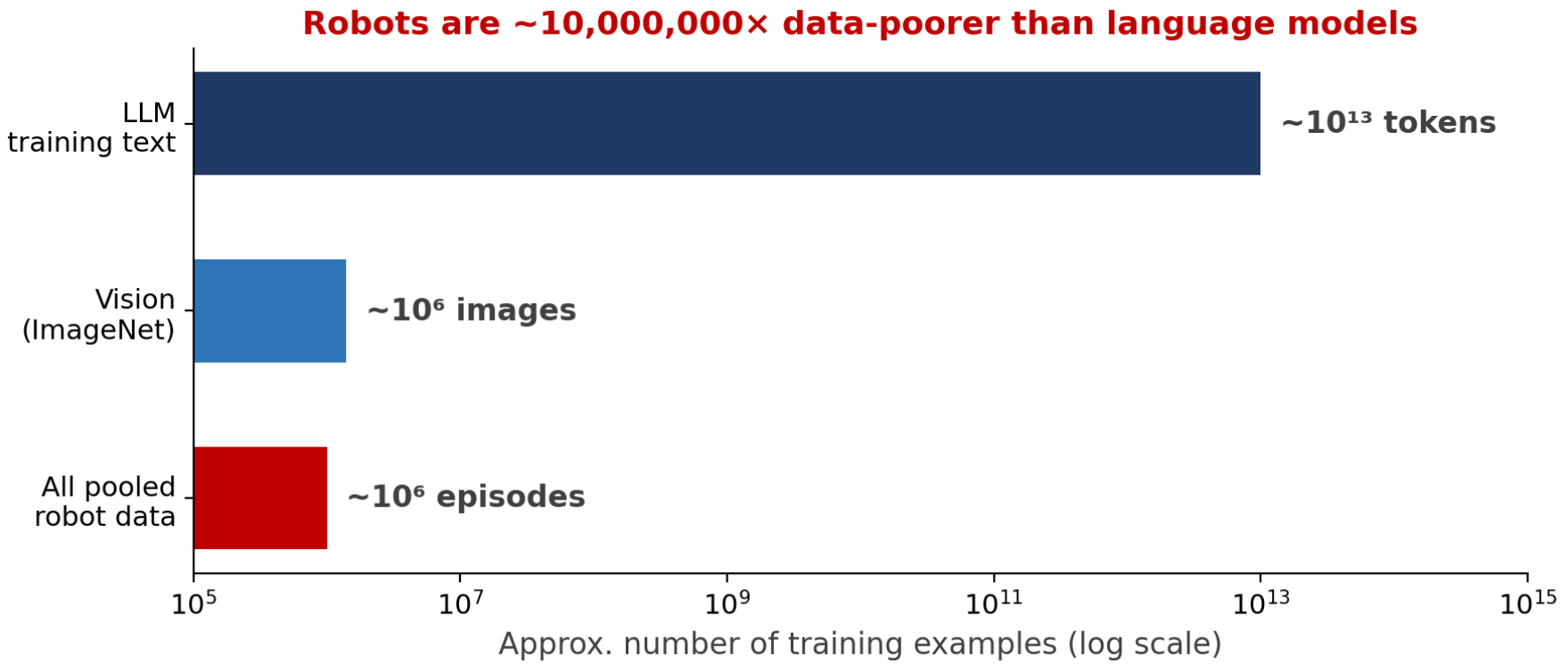
Situation: cross-embodiment pooling helped.

Complication: ~1M episodes total is still tiny — and you can't scrape robot actions off the web.



The real wall isn't architecture. It's data — quantity and diversity. (callback: L38)

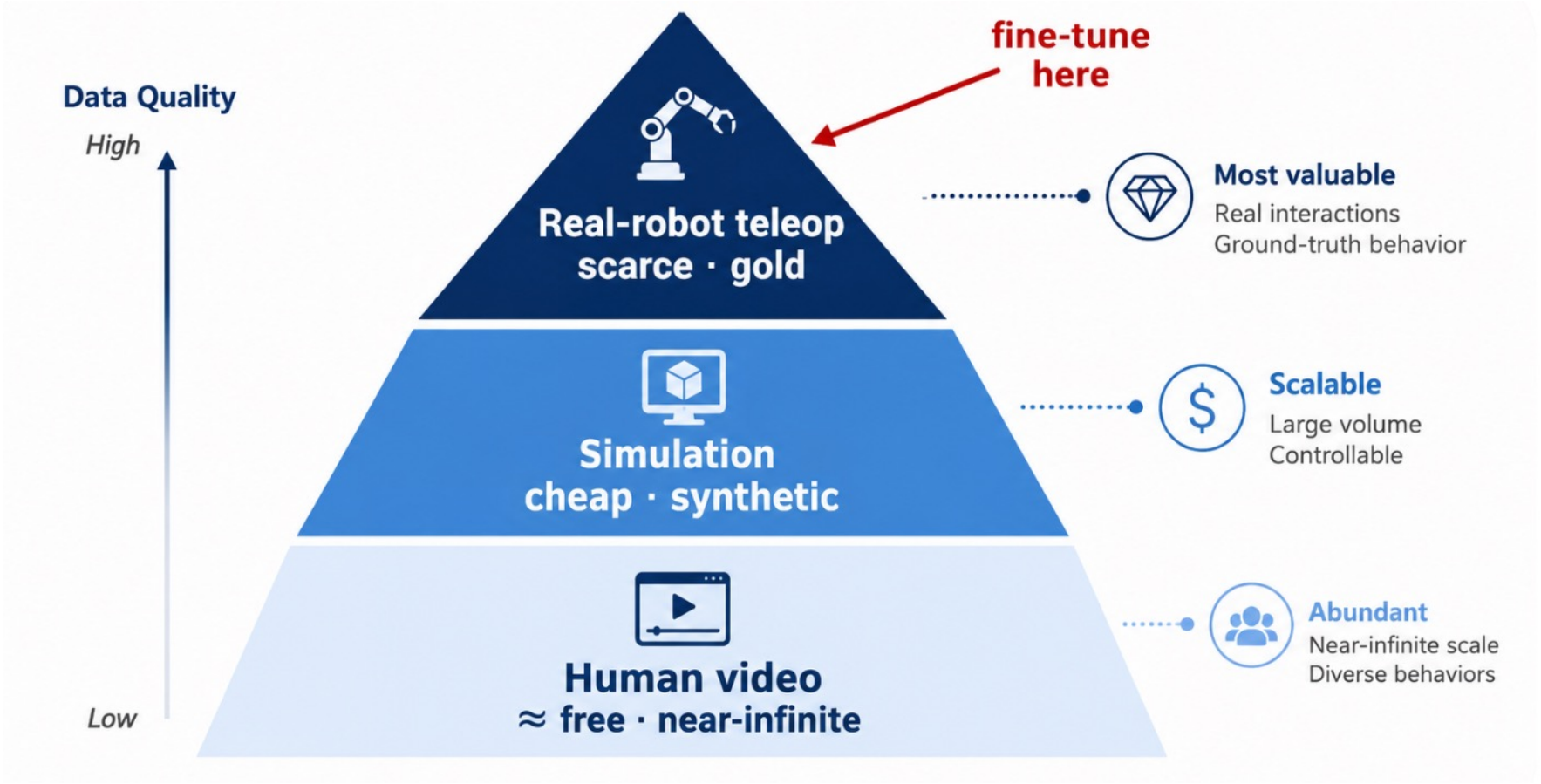
The Scale Gap, in Numbers



Robots are roughly ten million times data-poorer than language models.

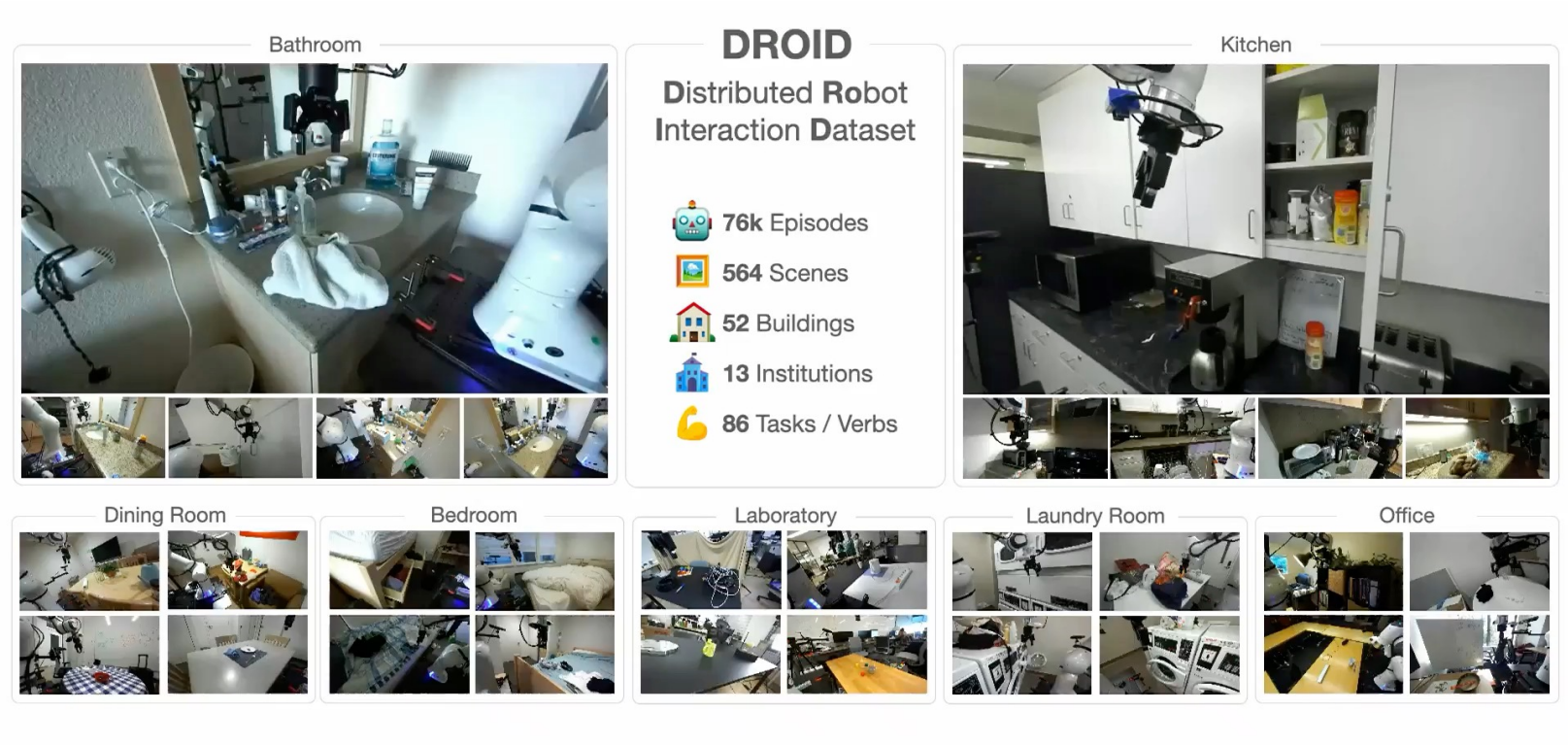
Where Robot Data Comes From

So where does foundation-scale robot data come from? The data pyramid.



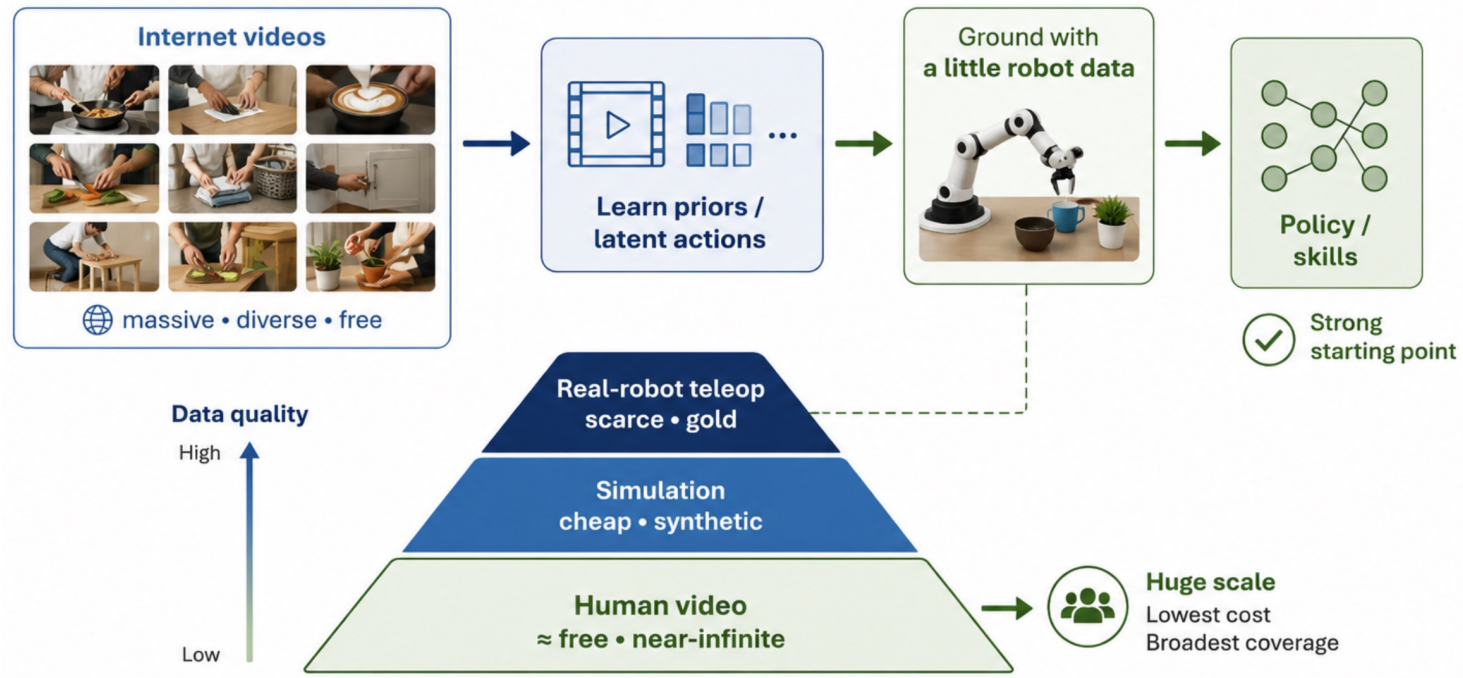
Strategy 1 — Pool Real Teleop

- **DROID:** 76k trajectories · 350 hours · 564 scenes · 50 collectors · 3 continents, all on one standardized arm. Plus AgiBot World (1M+ demos).
- **The data flywheel:** more robots deployed → more data → better models → more robots.



Strategy 2 — Learn from Human Video

- The internet is full of humans doing tasks
  — **effectively free and near-infinite visual data.**
- Extract priors / latent actions from human video, then ground them with a little robot data.

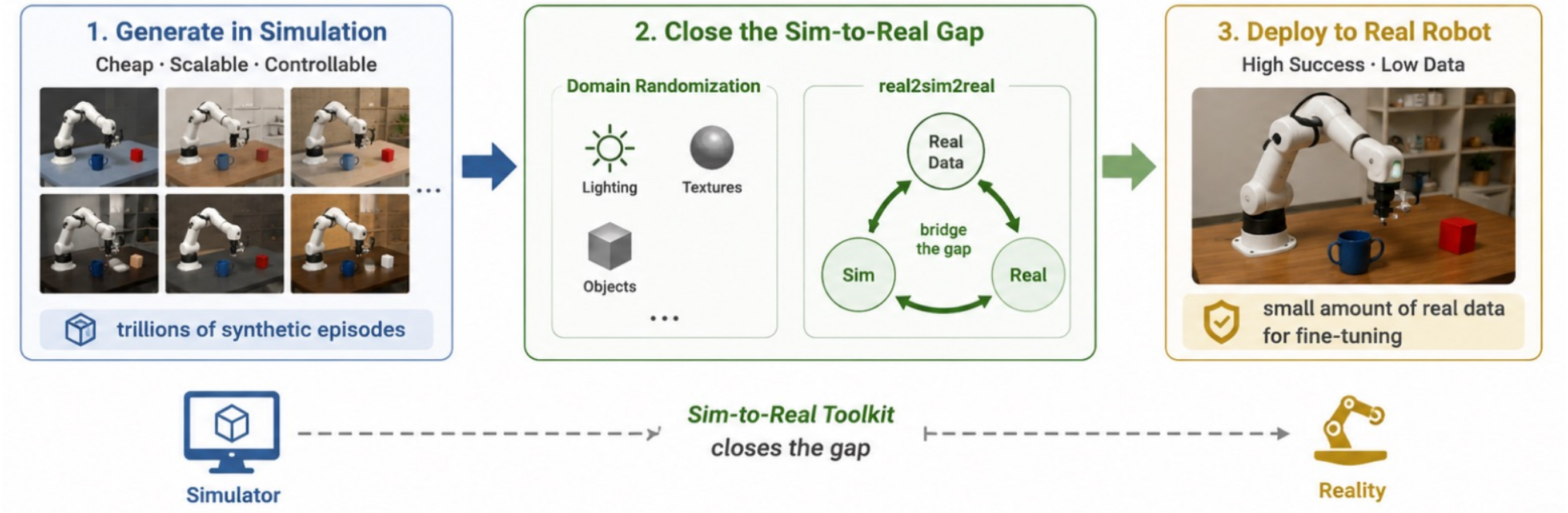


The cheapest, largest slice of the pyramid.

Strategy 3 — Simulation

Generate trillions of cheap synthetic episodes. (callback: L39)

The sim-to-real toolkit — domain randomization, real2sim2real — closes the gap between simulator and reality.



Human video + simulation form the base; real teleop is the precious top.

4. Does It Actually Generalize?

The Generalists Have Arrived

The payoff of pooled + human-video + sim data: generalist policies now exist.

- π0 / π0.5 / π0.7 — Physical Intelligence
- GR00T — NVIDIA (humanoid)
- Gemini Robotics — Google DeepMind
- AgiBot GO-1 · 1X NEO

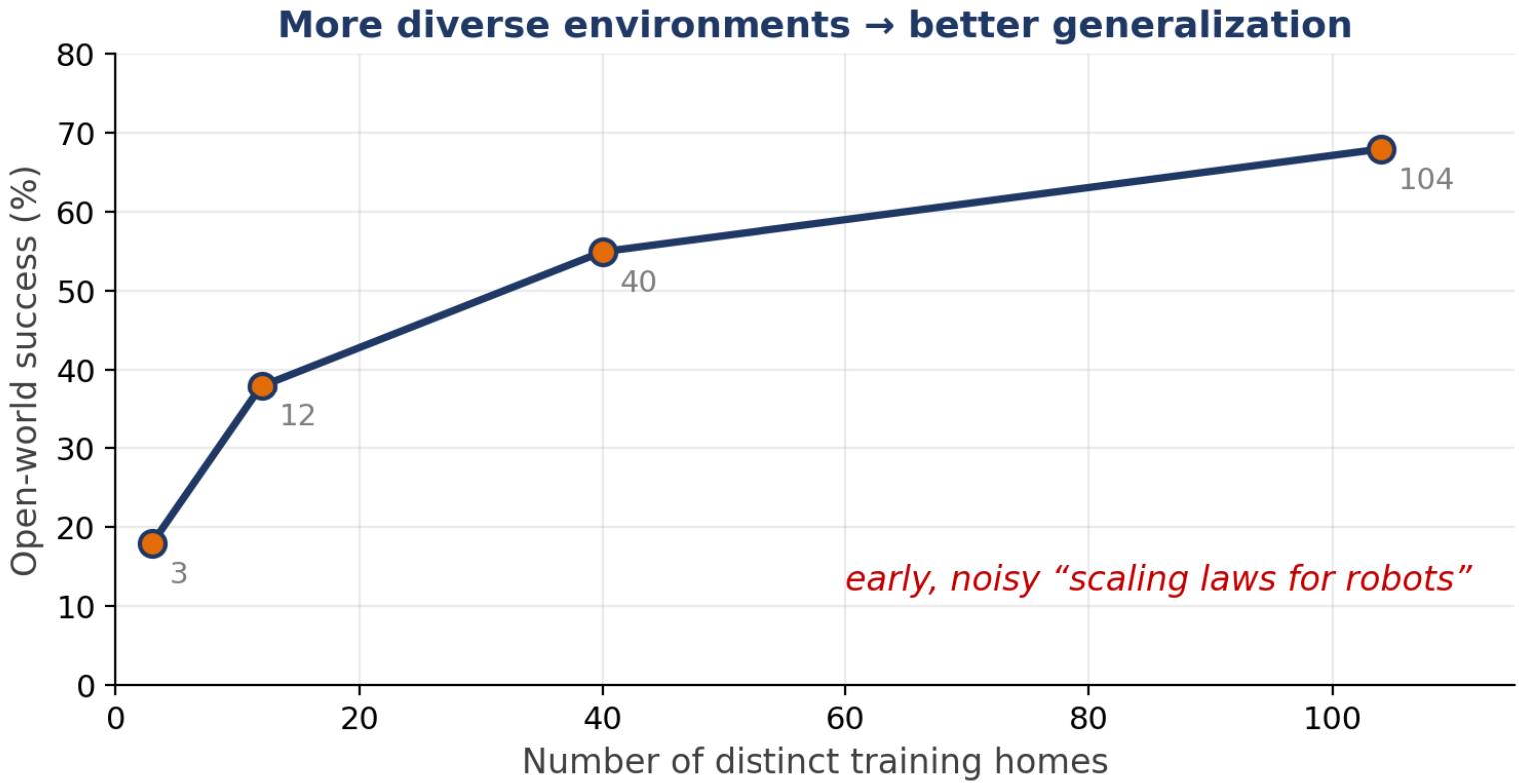
All pitched as robotic foundation models — one model, broad tasks.

The Real Test: A New Home

- **Complication:** the true test isn‘t a trained task — it’s a new task in a home the robot has never seen.
- **Answer:** π0.5 cleans kitchens & bedrooms in entirely unseen homes.

Scaling Laws for Robots?

π0.5's generalization improved as training scaled from 3 → 104 homes.



π0.7 pushes further: steerable, with emergent capabilities.

Think: Believe the Demo?

A viral clip shows a humanoid making coffee in a "new" kitchen.
Name three things you'd need to know before believing it "generalizes."

Discuss — then we'll compare your checklist with mine.

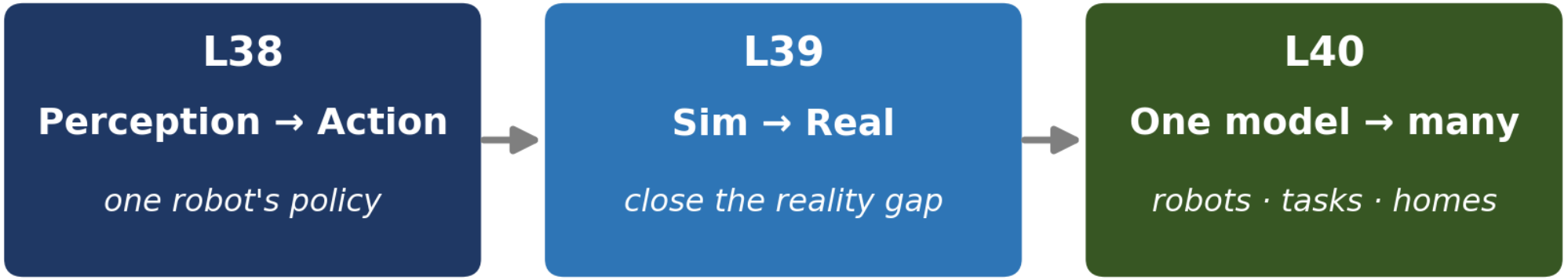
An Honest Reckoning

- Success rates are still well below human.
- No internet-scale robot data yet — the pyramid is thin.
- Embodiment heterogeneity is only partly tamed.
- Evaluation is slow, expensive, non-standardized.
- Safety and reliability for the home are open problems.

The "ImageNet moment for robotics" is a bet on a trend line — not a finished result.

5. Three Lectures, One Thread

The Module, End to End



The through-line: the data-and-generalization problem

L38 → L39 → L40: it was always the data-and-generalization problem.

Summary

- Robotics resisted foundation models — there's no internet of robot actions.
- Pooling across bodies (Open X-Embodiment) gives positive cross-embodiment transfer.
- The binding constraint is data — robots are ~7 orders of magnitude poorer than LLMs.
- The fix is a pyramid: human video + simulation at the base, scarce real teleop on top.
- Generalist policies now reach unseen homes — early, imperfect, a bet on scaling.

The Frontier

- The data-flywheel race — every deployed robot is a data collector.
- Humanoid foundation models — betting on one general-purpose body.
- World models as a data engine — robots that imagine their own training data.