Мультимодальные и поведенческие модели

Введение в Visual Language Action (VLA) модели

MIPT VLA 2026 | 7 сентября 2026

Алексей Староверов и Татьяна Земскова





МФТИ

Center for Cognitive Modeling

Система оценивания

Дифференцированный зачет - 10 баллов:

- 5 баллов - активность и выполнение лабораторных работ на семинарах
- 5 баллов - выполнение и сдача курсового проекта
- для получения зачета необходимо предоставить минимально рабочую версию курсового проекта

Алексей Староверов

Контакты:
Telegram - @Alstar8
Email - staroverov.av@phystech.edu

Татьяна Земскова

@wingrune
t.s.zemskova@gmail.com

Ссылка на
Telegram-чат курса

План курса

Module 1: Foundations of Embodied Intelligence

Модуль задаёт теоретическую основу курса: вводит парадигму VLA и ключевой математический аппарат генеративных моделей, необходимый для моделирования непрерывных пространств действий (continuous action spaces).

Module 2: Imitation Learning, Data, and Infrastructure

Модуль посвящён сбору данных и основным алгоритмам, используемым для обучения VLA стратегий.

Module 3: Advanced Perception and Environment Transfer

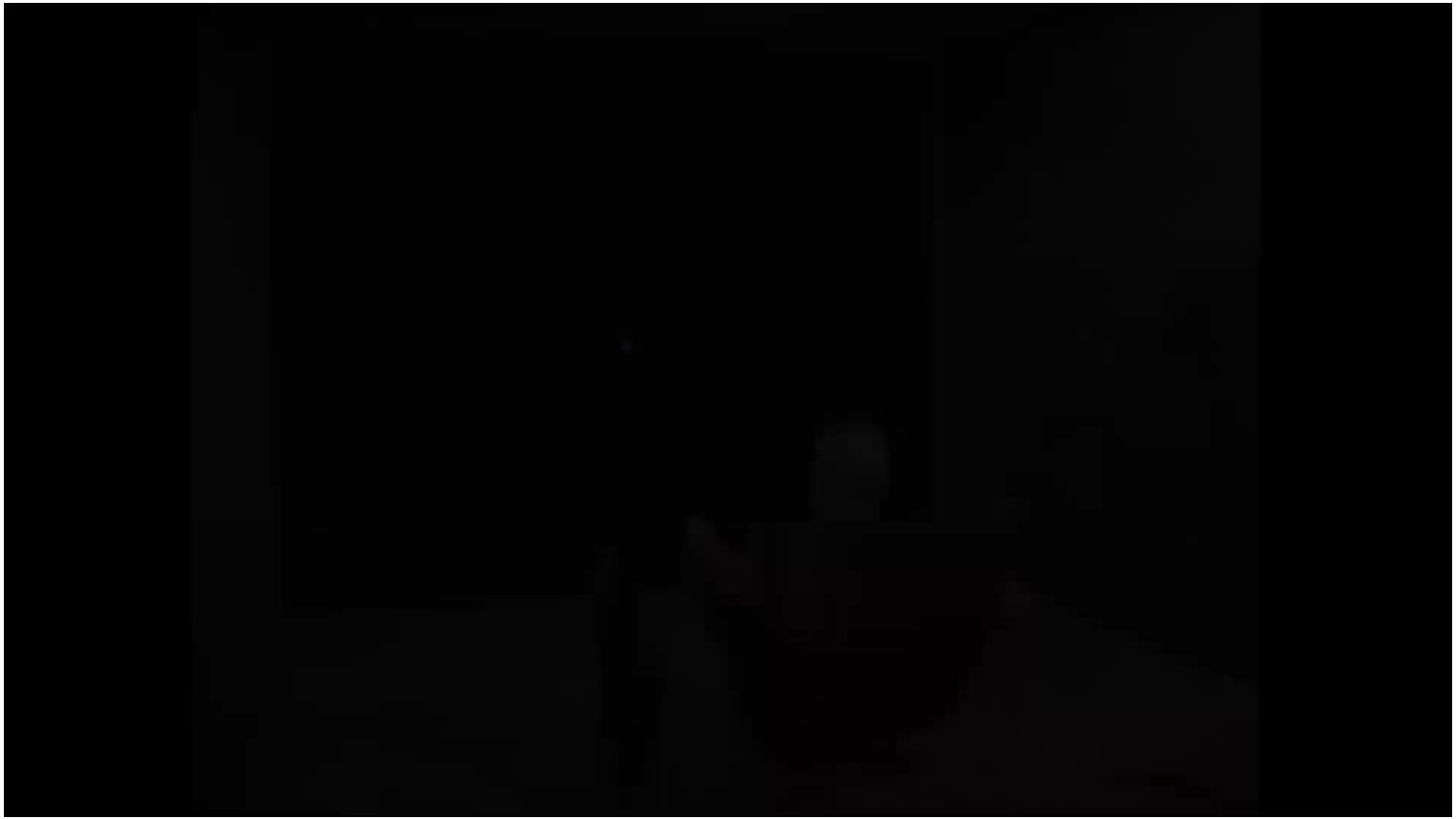
Модуль знакомит с методами работы со сложной 3D-геометрией, расхождениями аппаратного обеспечения и координацией двух манипуляторов.

Module 4: Complex Reasoning, Optimization, and World Models

Модуль охватывает продвинутую оптимизацию (RL) и когнитивные архитектуры (World Models) для long-horizon autonomy.

What is a VLA?

Multimodal and behavioural models



«Put that there» - Richard Bolt. 1980

Foundation models in the physical world

clean the house, fold the laundry, and make dinner

It sounds like you have a busy evening ahead! To help you tackle those tasks efficiently, here's a plan:

Cleaning the House

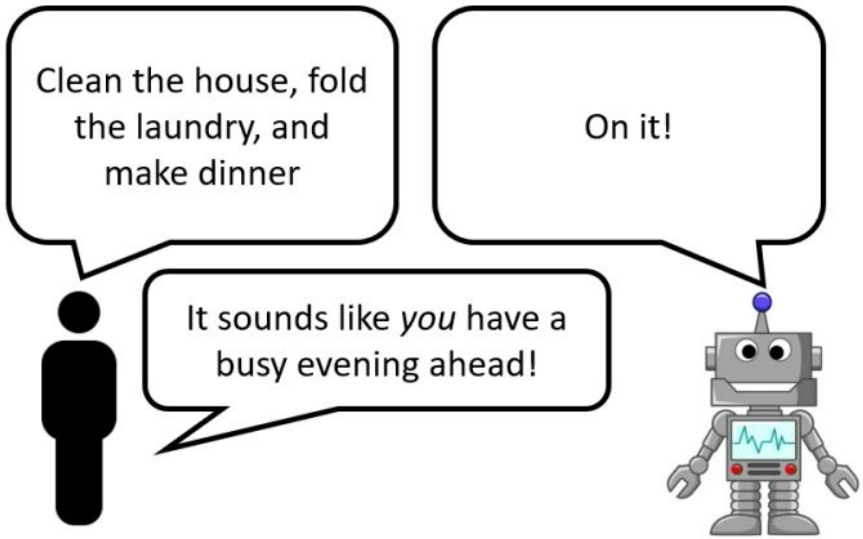
Let's break down the house cleaning into manageable steps:

- **Tidy Up:** Start by putting away any clutter. This makes the actual cleaning much easier.
- **Dusting:** Work from top to bottom. Dust shelves, furniture, and electronics.
- **Vacuum/Sweep:** Get rid of dirt and debris from floors.
- **Wipe Surfaces:** Clean countertops, tables, and other high-touch areas with an appropriate cleaner.
- **Bathrooms and Kitchen (if applicable):** Focus on these areas last, as they often require more targeted cleaning.

Folding the Laundry

To make laundry folding less daunting:

- **Sort by Type:** Fold similar items together (e.g., all shirts, then all pants).
- **Clear a Space:** Designate a clean, flat surface like a bed or table for folding.



Why Physical World is important?

The real world is very complicated!







Robotic learning is about more than just building useful robots

Lifelong learning is much more valuable when you live in a world that is complex enough to learn from!

This is a big opportunity!

Opportunity: the real physical world gives the AI system a lot to learn!

How do we get AI to do what we want?

Explaining a joke

Prompt

```
Explain this joke:

Joke: Did you see that Google just hired an eloquent whale for
their TPU team? It showed them how to communicate between two
different pods!
```

Model Response

```
Prediction: TPUs are a type of computer chip that Google uses
for deep learning. A "pod" is a group of TPUs. A "pod" is also a
group of whales. The joke is that the whale is able to
communicate between two groups of whales, but the speaker is
pretending that the whale is able to communicate between two
groups of TPUs.
```



The problem: our AI systems are trained to **model** the world, but we then demand that they **act** in this world.

$$\text{modeling} \neq \text{acting}$$

Solution: change the data to "trick" the AI system to behave the way we want

$$p_\theta(\mathbf{y}|\mathbf{x})$$

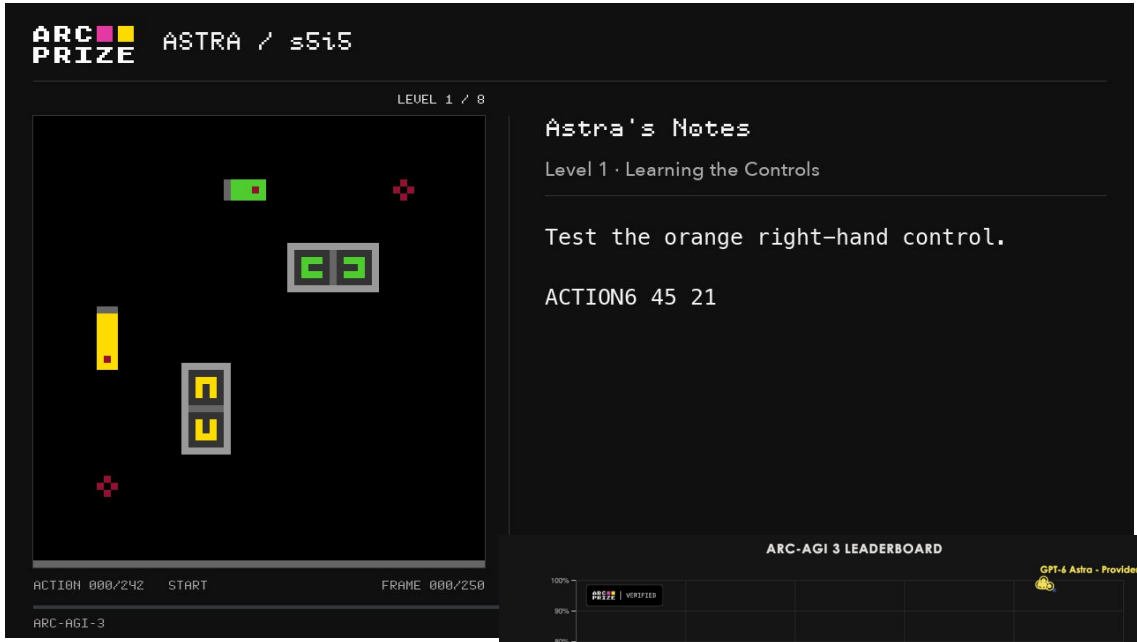
most *post-training* does this

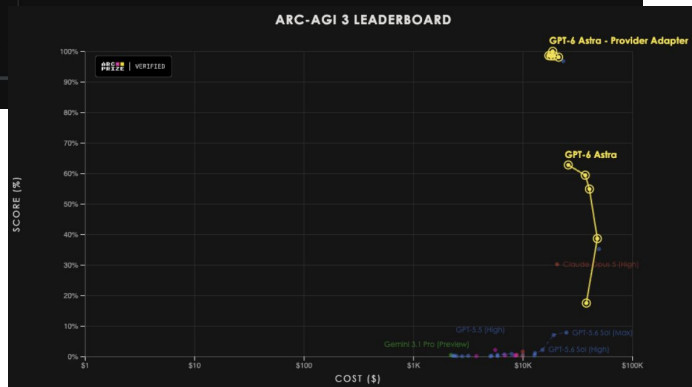




Recent breakthrough (GPT-6 Astra)

- GPT-6 Astra scores 62.7% for $26K on ARC-AGI-3 Semi-Private with our Standard harness, and 99.9% for $19K with a Provider Adapter harness.
- GPT-6 Astra surpasses the human baseline in action efficiency on ARC-AGI-3. It used fewer actions than the median tested human on 96% of levels.
- A key behavior observed in GPT-6 Astra was its ability to turn unfamiliar environments into compact symbolic world models. It represented game mechanics as logical rules and developed its own domain-specific language shorthand to track state and plan actions.





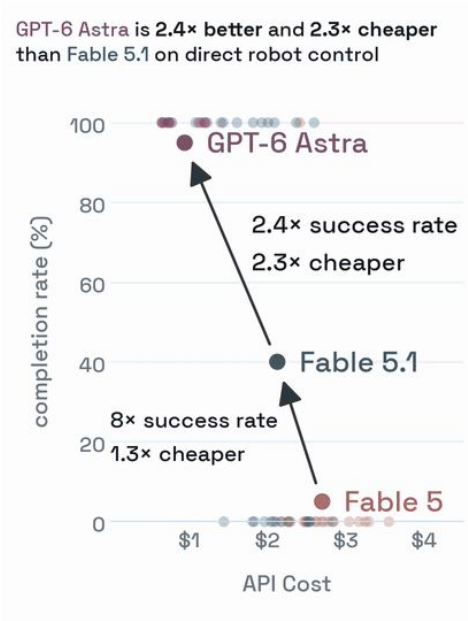
https://arcprize.org/blog/astra

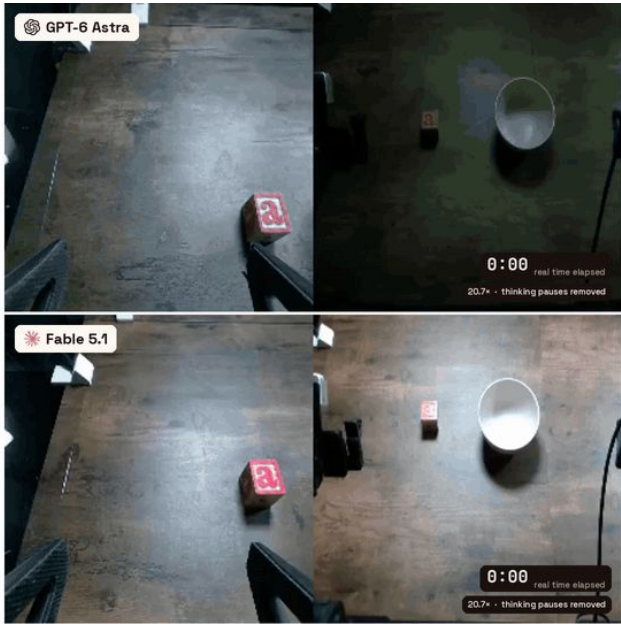
Recent breakthrough (GPT-6 Astra)

Embodiment - Bimanual I2RT YAM arms, 6-DoF per arm with parallel-jaw grippers

Control - Absolute end-effector poses (move_to): x, y, z, yaw, pitch, roll and gripper, per arm. The robot's IK converts poses to joint angles.

Observation - Three camera views (top, left wrist, right wrist) plus proprioceptive state, each turn





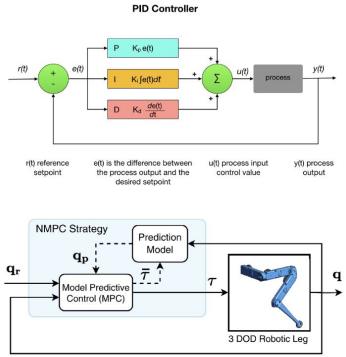
Inspect Robots

https://github.com/robocurve/inspect-robots

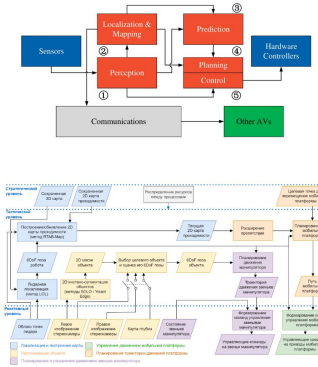
https://openai.robocurve.org/gpt-6-astra/

Evolution of control systems in robotics

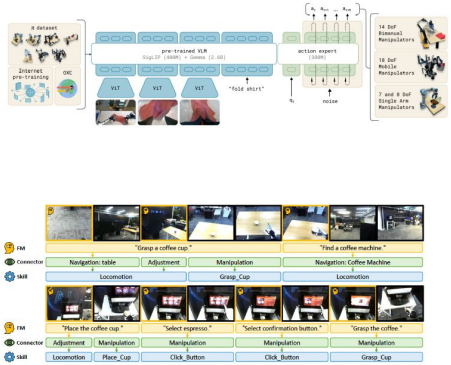
Классическое управление



Модульные архитектуры

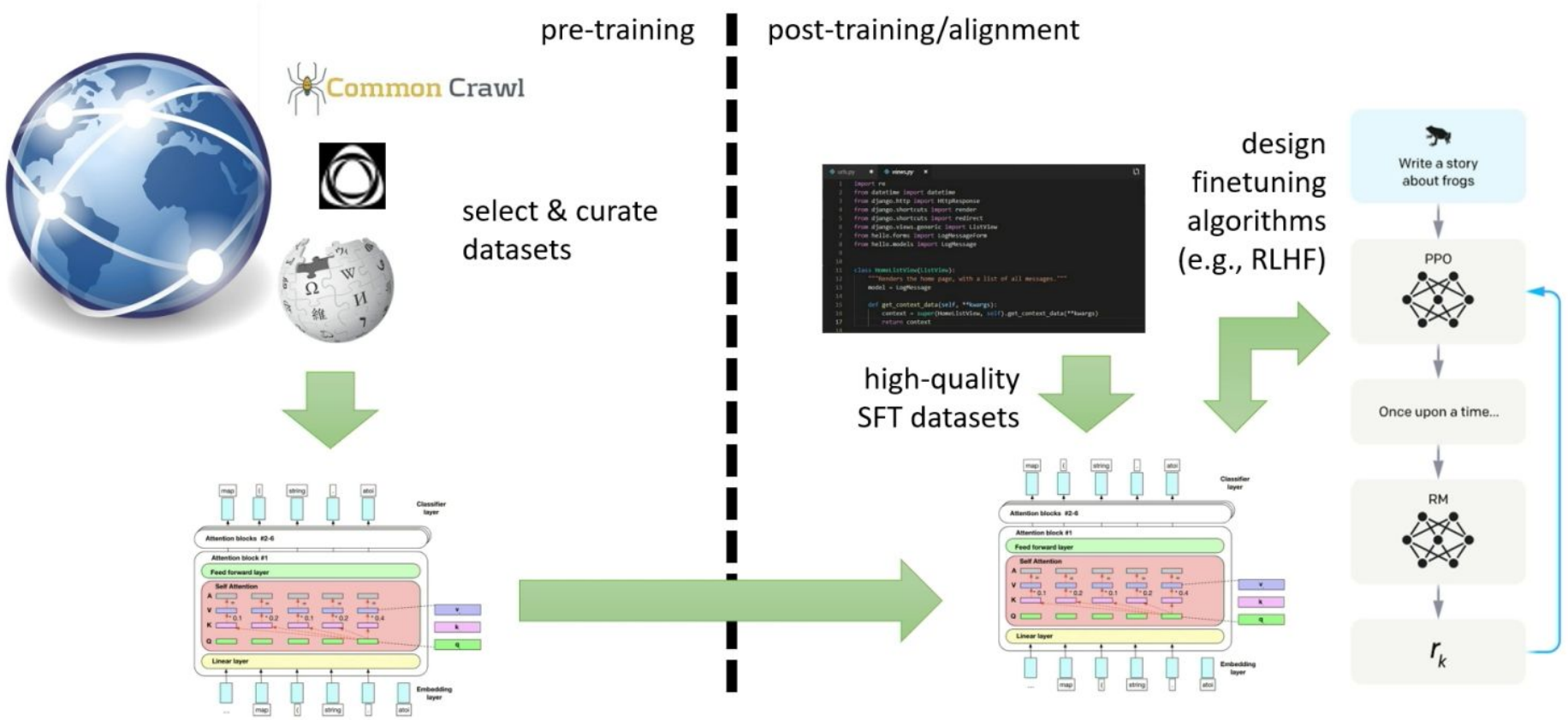


End-to-end VLA models





The anatomy of foundation model training



Why is RL training of robot foundation models hard?

We need to do the RL training in the *real world*

Why?





RL in the real world

+ we learn about the real world

- we need very sample-efficient algorithms to make this practical

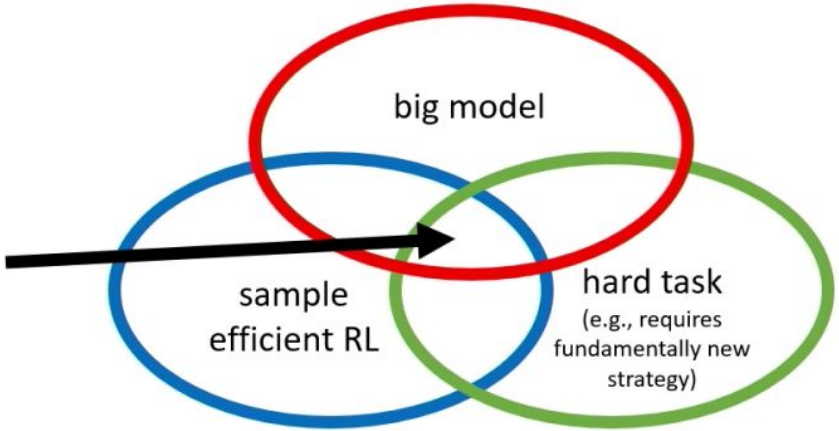
RL in a sim/sandbox

- we learn about the stuff we put into our sim

+ we can parallelize and use expensive but parallelizable algorithms

Lifelong learning is much more valuable when you live in a world that is complex enough to learn from!

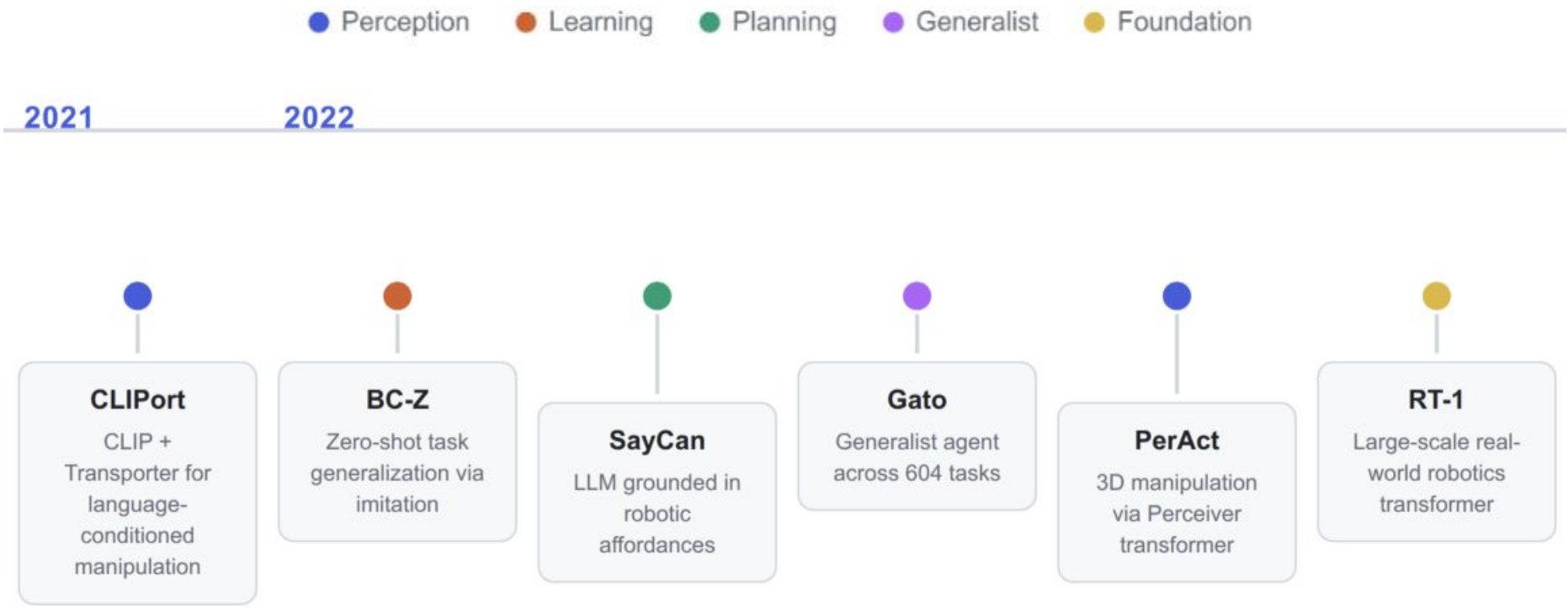
this combination is really hard



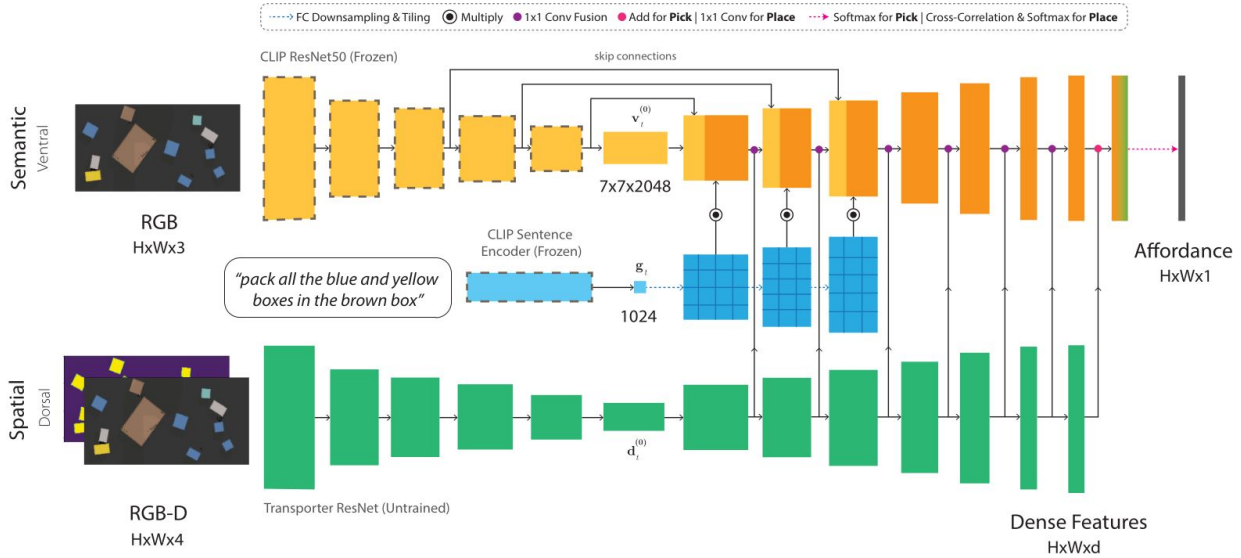
History of VLA models

Evolution of Robot Foundation Models

From sim-to-real transfer to generalist robot policies — the key milestones shaping embodied AI.



CLIPort (2021) — CLIP “what” + Transporter “where”



Novelty at the time · Shridhar, Manuelli, Fox. CLIPort: What and Where Pathways for Robotic Manipulation. CoRL 2021. arXiv:2109.12098

- Первыми объединили frozen internet vision-language model (CLIP) с spatially equivariant pick-and-place (Transporter).
- Two-stream: ventral semantic stream conditioned на language; dorsal stream держит metric spatial precision.
- Language goal («pack the blue and yellow boxes») становится pixel affordance map, а не parsed skill ID.
- Few-shot generalization на невиданные цвета, объекты и conjunctions — language grounding без гигантской policy.
- Задали шаблон, которым VLA пользуются до сих пор: frozen web semantics + spatial action head.

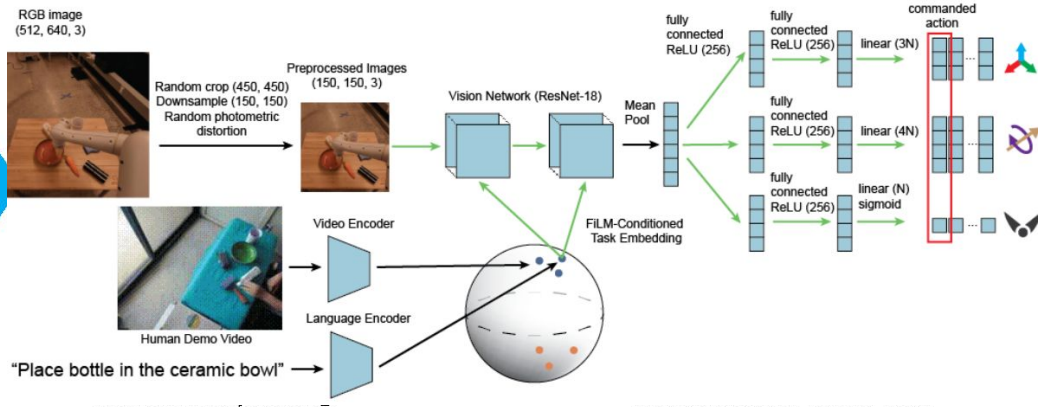
BC-Z (2022) — one imitator, many language tasks



Shared Autonomy Data Collection



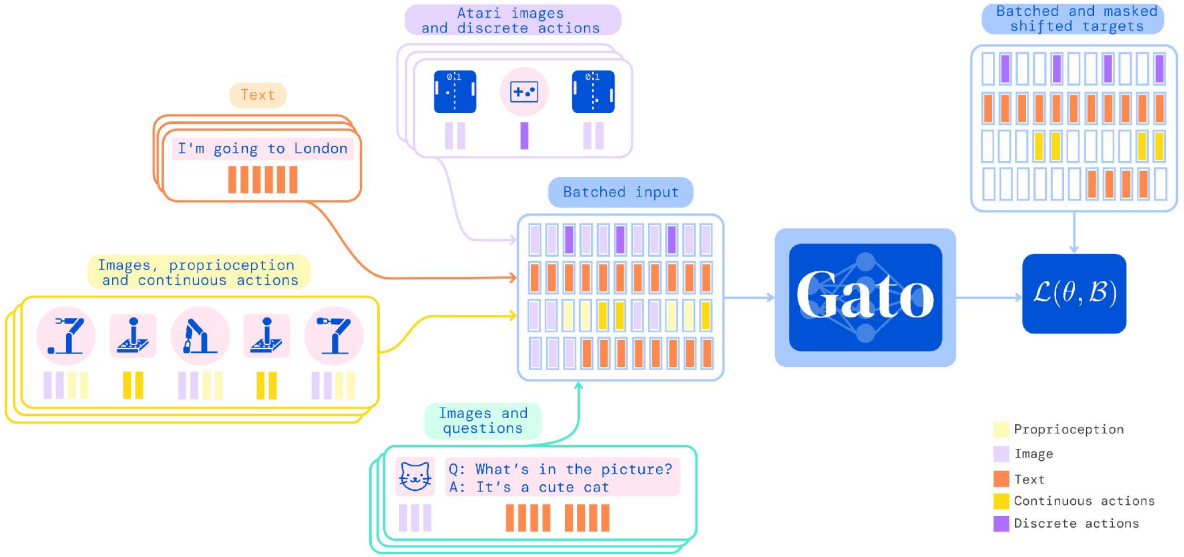
Diverse Multi-Task Dataset

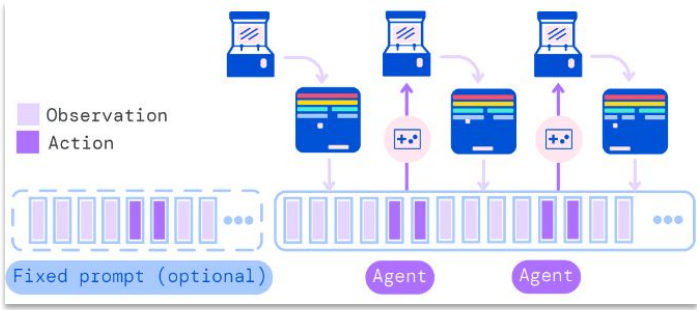


Novelty at the time · Jang, Irpan, Khansari et al. BC-Z: Zero-Shot Task Generalization with Robotic Imitation Learning. CoRL 2021/2022. arXiv:2202.02005

- Одна visuomotor policy на 100+ real tasks вместо specialist на каждый skill.
- Imitation conditioned на language (и video), поэтому новые формулировки можно пробовать zero-shot.
- Показали: robot fleet + language labels выгоднее, чем отдельный датасет на каждый глагол.
- Zero-shot здесь — новый language для уже обученных skills, ещё не web-scale semantics (это RT-2).
- Эмпирический аргумент: imitation + language — путь к multi-task real robots.
- Собрали датасет из 26к робо демонстраций, 125 часов, 12 роботов, 100 задач

Gato (2022) — one network, 604 tasks, including a robot

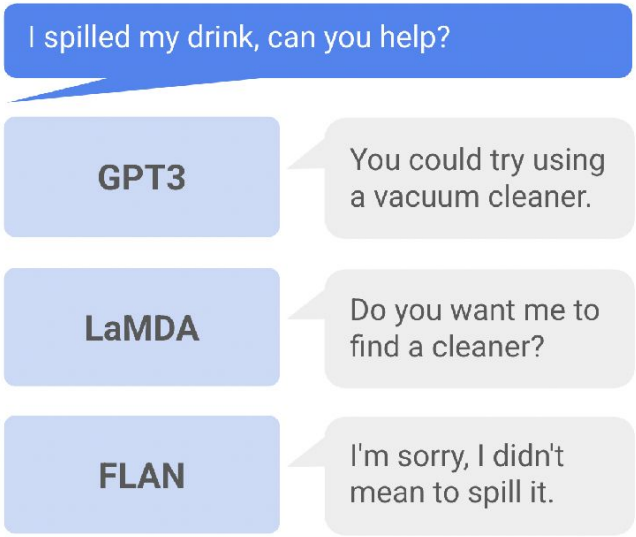


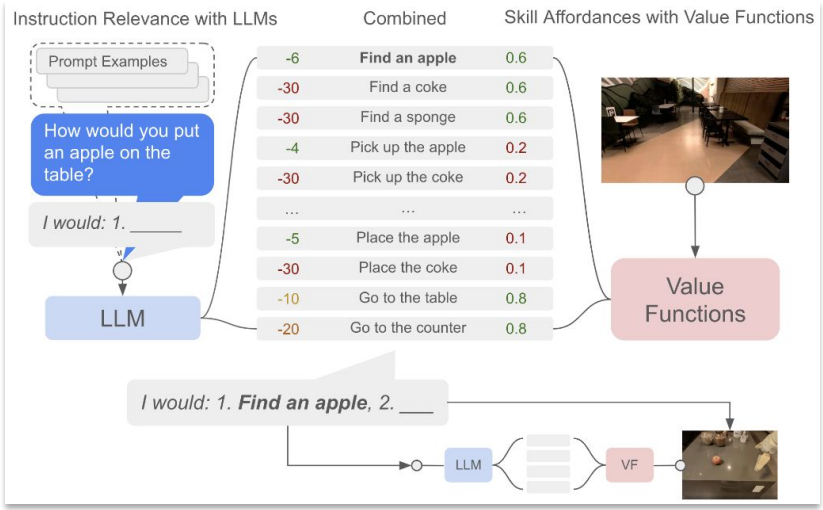


Novelty at the time · Reed, Zolna, Parisotto et al. A Generalist Agent. 2022. arXiv:2205.06175

- Один набор весов для Atari, captioning, dialogue и real-robot stacking — не mixture of experts.
- Всё — tokens: text, flattened images, proprioception и actions делят один transformer.
- Первый громкий тезис: generalist sequence model может включать physical control, не только текст и пиксели.
- По роботу результат скромный; новизна — shared tokenization, а не новый manipulation SOTA.
- Прямой предок RT-1 / RT-2 / OpenVLA: та же ставка «tokenize the robot and train a transformer».

SayCan (2022) — an LLM that only says what the robot can do now

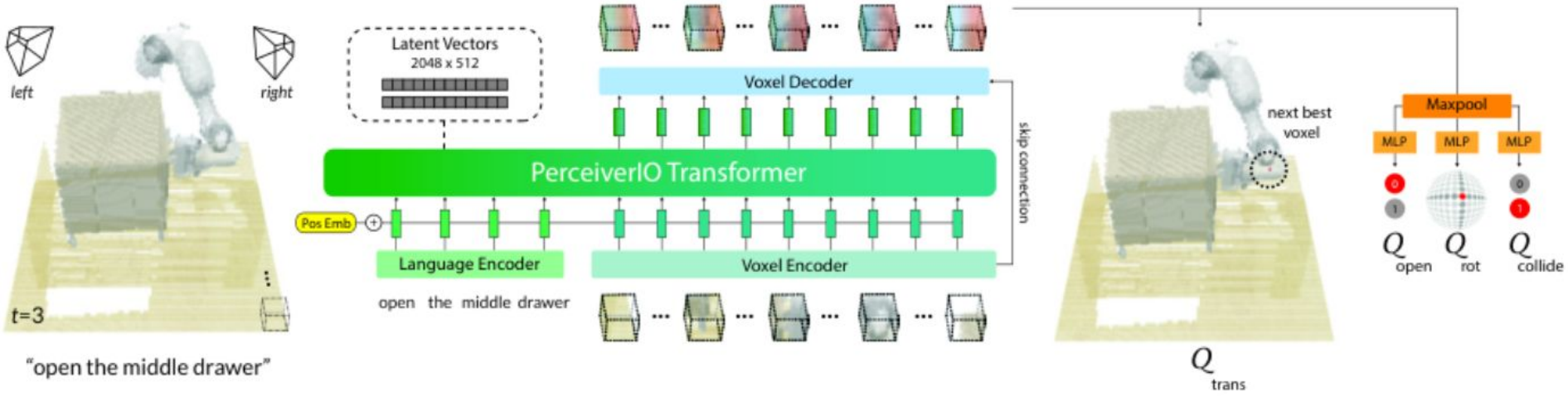




Novelty at the time · Ahn, Brohan, Brown et al. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. 2022. arXiv:2204.01691

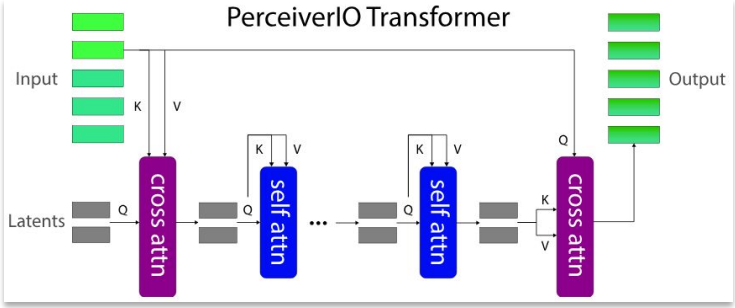
- Первая широко цитируемая система, где LLM — robot planner, а не captioner.
- LLM предлагает skills; выученная value / affordance model оценивает, что вообще возможно в сцене.
- Произведение «do as I can, not as I say» не давало free-form LLM планам требовать невозможное.
- Long-horizon kitchen tasks из одного language goal, skills — action vocabulary.
- Grounded affordances стали интерфейсом между LLM и реальным роботом (пока VLA не стали делать это end-to-end).

PerAct (2022) — language-conditioned actions in a voxel grid

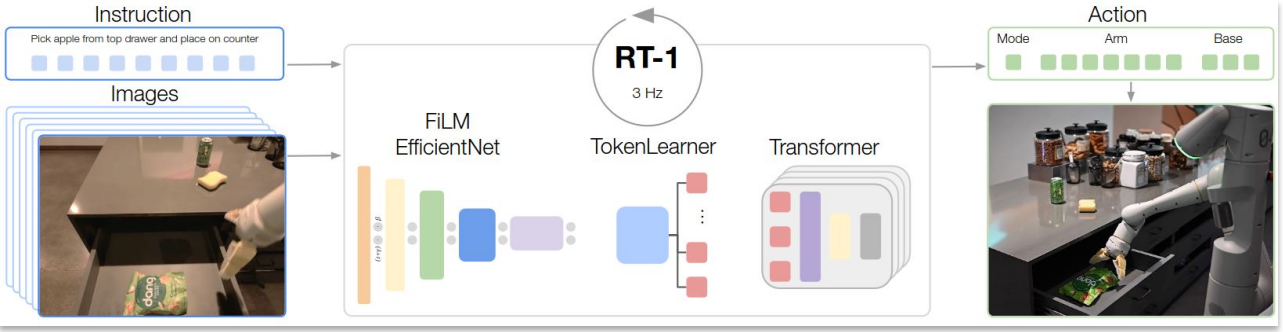


Novelty at the time · Shridhar, Manuelli, Fox. Perceiver-Actor: A Multi-Task Transformer for Robotic Manipulation. CoRL 2022. arXiv:2209.05451

- Language goals в стиле CLIPort подняли с 2D heatmaps до 6-DoF actions внутри 3D voxel grid.
- Multi-task RLBench с несколькими demos на task — 3D structure обошёл 2D baselines на spatial commands.
- Action — «кликни voxel, затем rotation»; translation-equivariant, как Transporter.
- Language-conditioned manipulation нужен metric 3D workspace, а не только плоскость камеры.

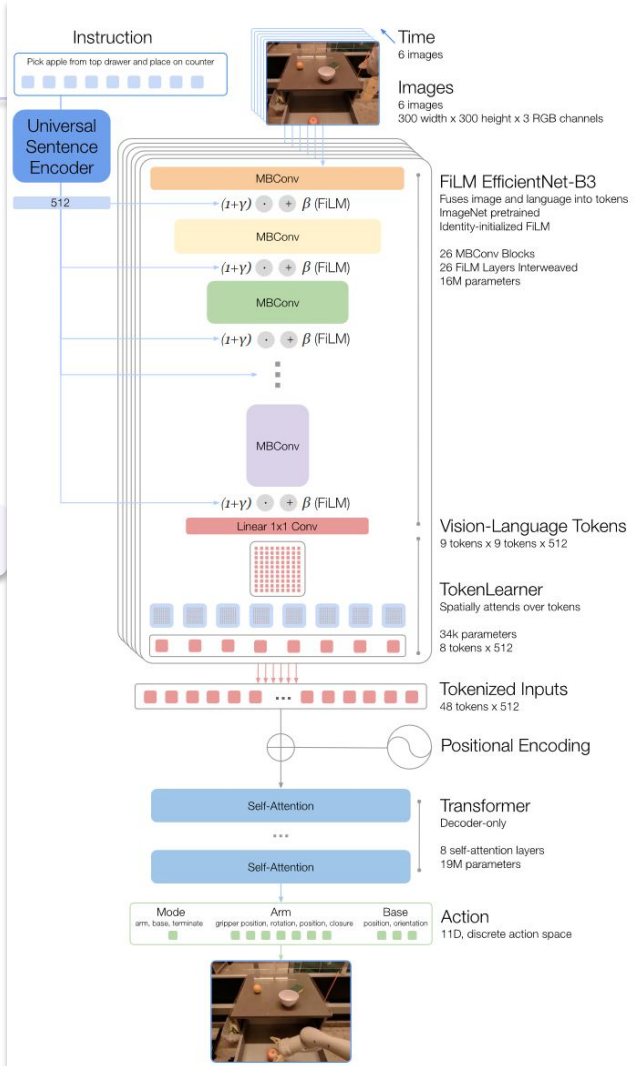


RT-1 (2022) — a transformer trained on a real robot



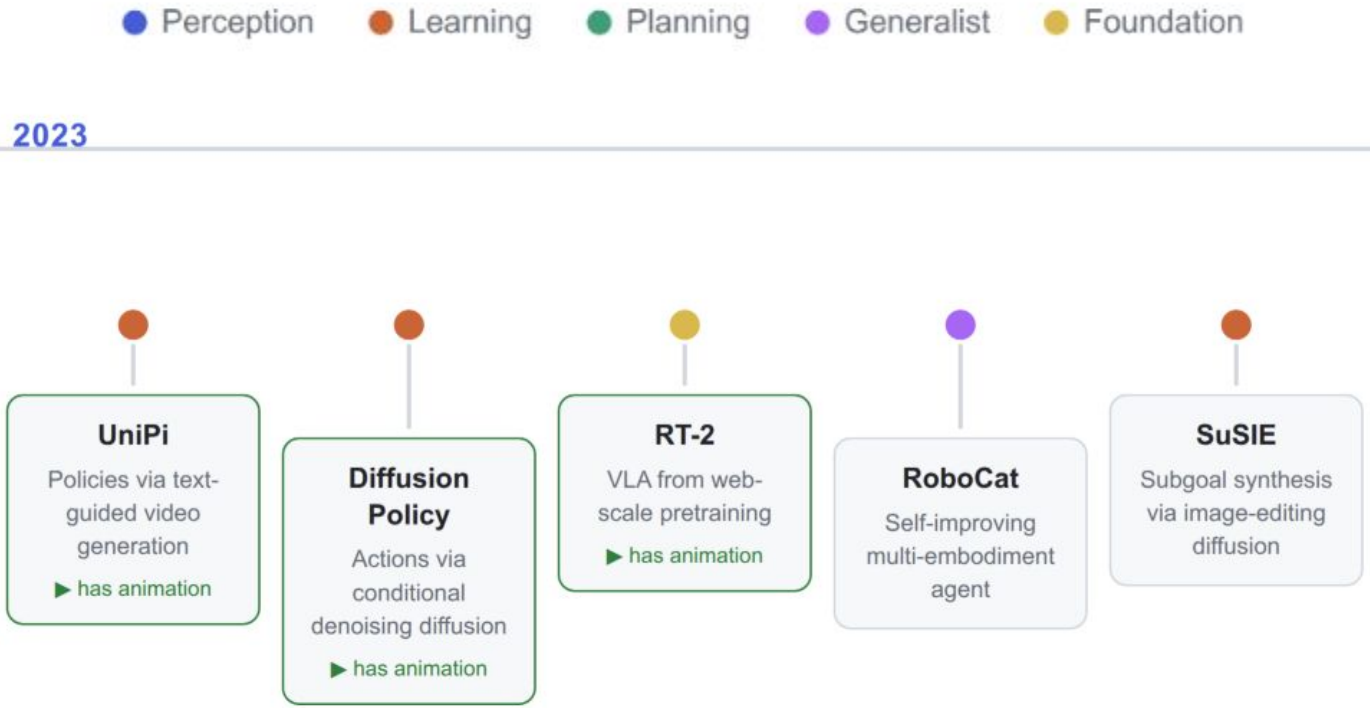
Novelty at the time · Brohan, Brown, Carbajal et al. RT-1: Robotics Transformer for Real-World Control at Scale. 2022. arXiv:2212.06817

- Первая масштабная transformer policy: ~130k real episodes, 700+ tasks, месяцы robot time.
- Images и actions — tokens; FiLM-conditioned EfficientNet + transformer decoder выдаёт action tokens.
- Data diversity (много tasks, много объектов) важнее, чем выучивать single-task demonstrator.
- Open-loop token actions ~3 Hz — медленно, но результатом была scaling curve, а не dexterity.
- Стал data-and-tokenization backbone, который RT-2 потом подключил к веб-VLM.

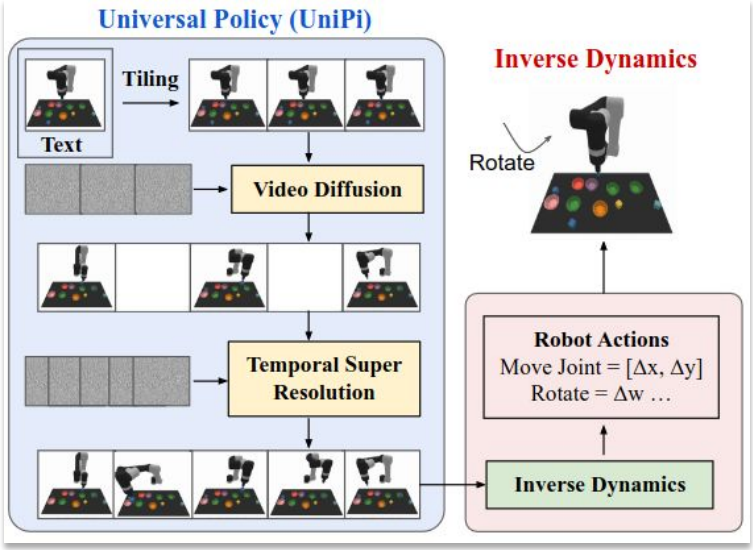


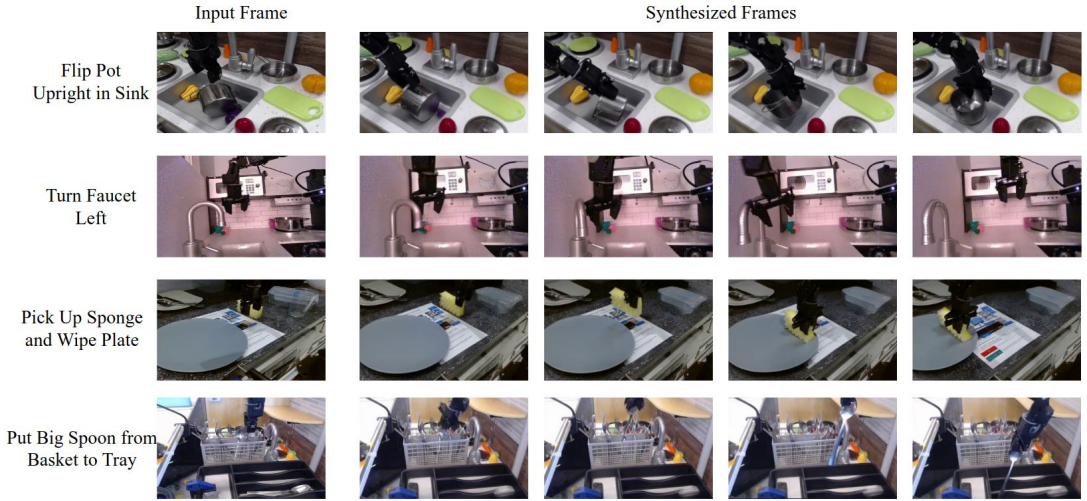
Evolution of Robot Foundation Models

From sim-to-real transfer to generalist robot policies — the key milestones shaping embodied AI.



UniPi (2023) — a policy is a text-conditioned video

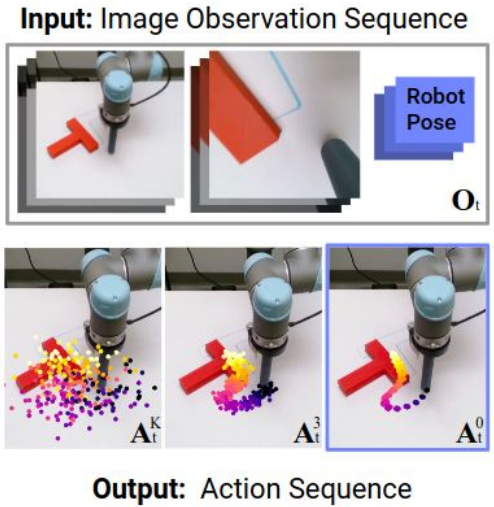


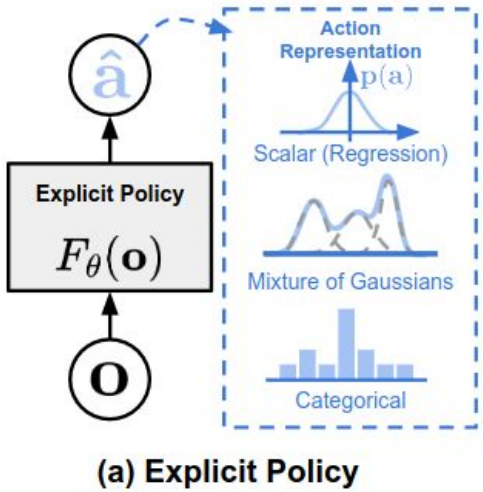


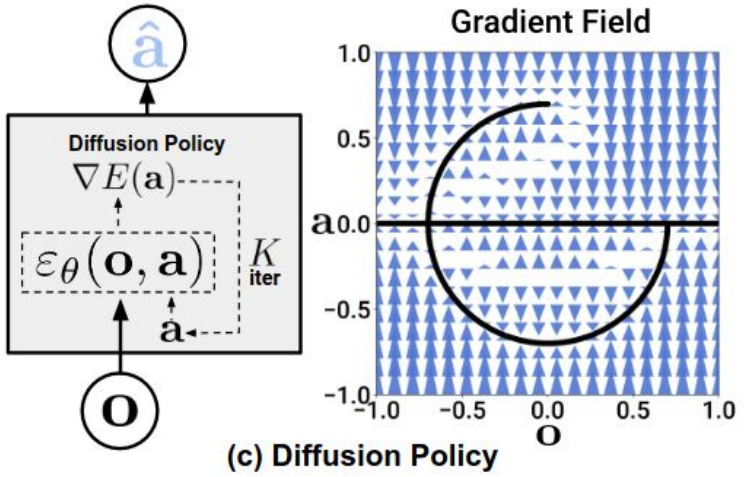
Novelty at the time · Du, Yang, Dai, Dudík, Salakhutdinov, Tenenbaum, Mordatch. Learning Universal Policies via Text-Guided Video Generation. 2023. arXiv:2302.00111

- Sequential decision-making как text-guided video generation — pixels как универсальный state и action interface.
- Planner синтезирует будущие frames; inverse dynamics позволяет получить (o_t, o_{t+1}) → motor commands.
- Internet video priors позволяют language собирать goals, которых не было как robot trajectories.
- Первый громкий аргумент: «physics уже в video», поэтому hand-built simulator может быть не нужен.
- Ограничение, которое сдвинуло поле: генерировать pixels слишком медленно для control loop — UVA / Cosmos уходят в latents.

Diffusion Policy (2023) — actions as conditional denoising

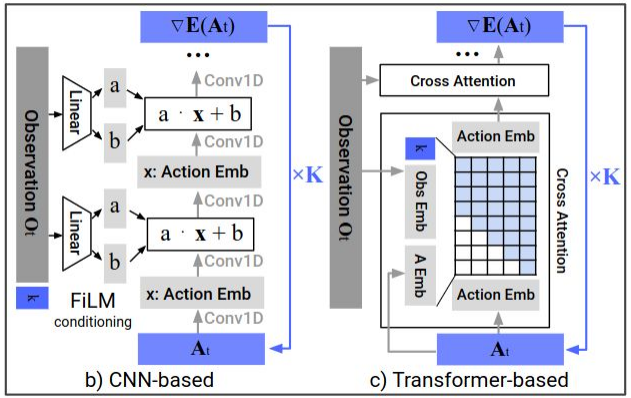




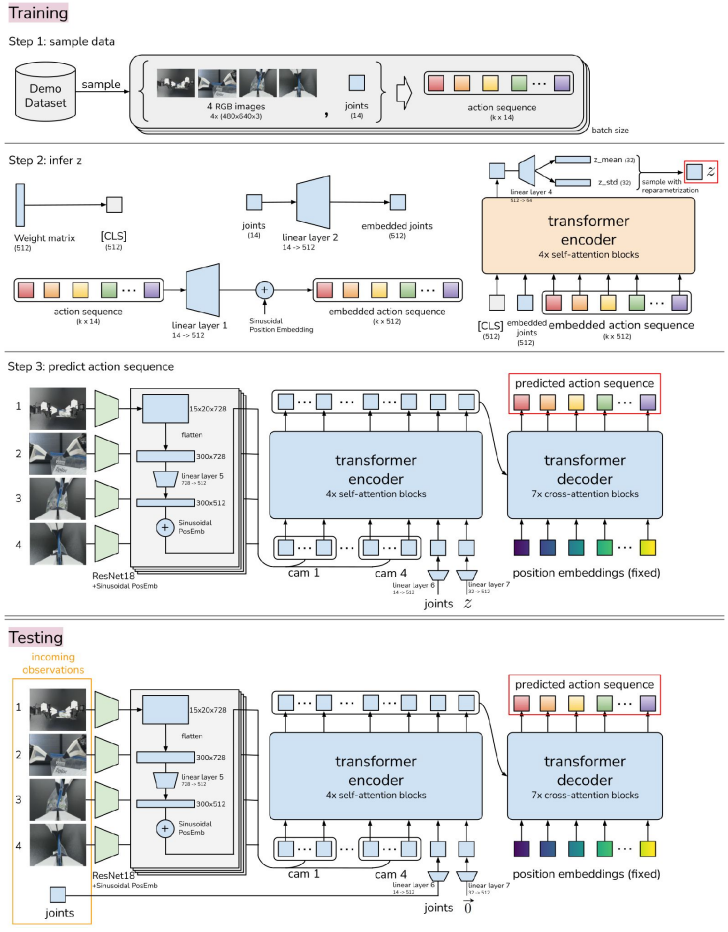


Novelty at the time · Chi, Feng, Du, Xu, Cousineau, Burchfiel, Song. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. RSS 2023. arXiv:2303.04137

- Closed-loop visuomotor control как denoising action chunk, а не regression к одному next pose.
- Решили multimodal-demo проблему: MSE behavioral cloning усредняет «слева от кружки» и «справа» в collision.
- Receding-horizon: предсказать T_p шагов, сделать несколько, replan — совместимо с real robots.
- Стал default continuous action head (CNN+FiLM или transformer) у Octo, GR00T, π-моделей.



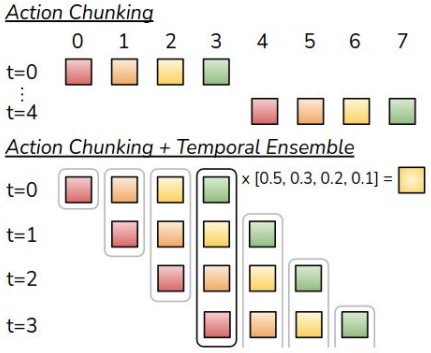
ACT (2023) — predict a motor program, not one joint tick



Zhao, Kumar, Levine, Finn. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware. CoRL 2023. arXiv:2304.13705

Novelty at the time

- Chunking K будущих joint targets (CVAE + transformer) бьёт one-step BC на тонких bimanual tasks.
- ALOHA: ~$20k железа + ~50 teleop demos / 10 мин, 80–90% на нескольких real tasks — не гигантский pretrained VLA.
- Temporal ensembling overlapping chunks сделал движение human-smooth на 50 Hz.
- CVAE latent z только на training; на test time z = 0 (prior mean).
- Action chunking стал default в Diffusion Policy, π0 и почти каждой современной VLA.



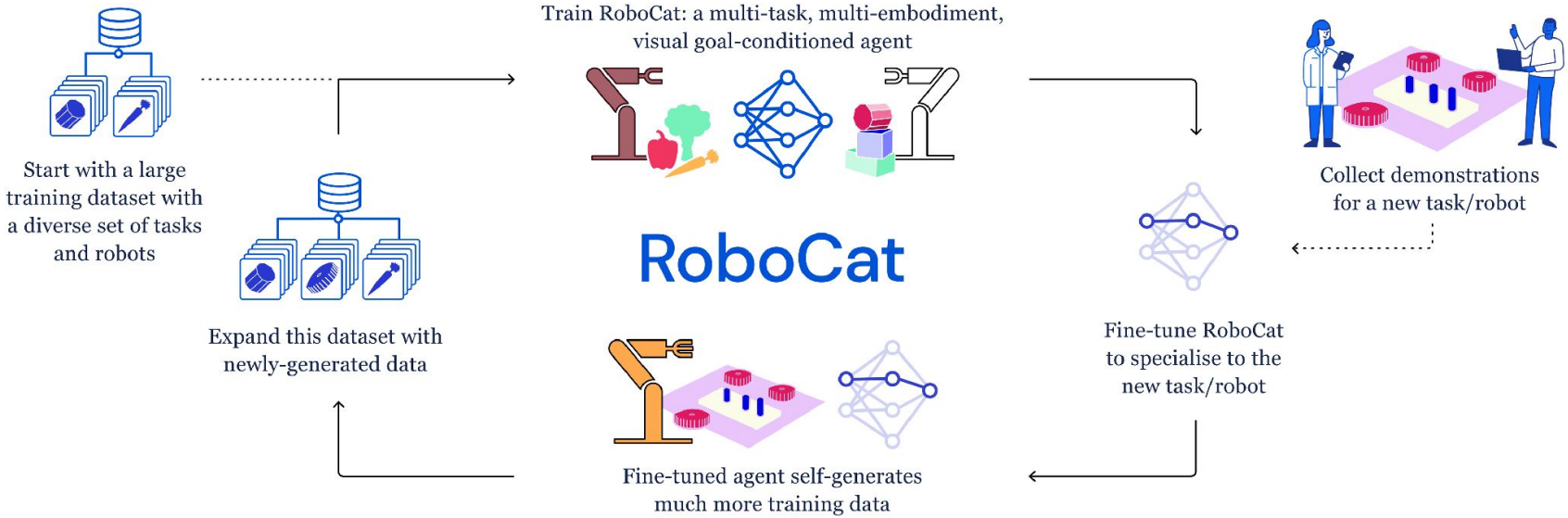
Algorithm 1 ACT Training

Given: Demo dataset $\mathcal{D}$, chunk size k, weight β.
Let a_t, o_t represent action and observation at timestep t, $\bar{o}_t$ represent o_t without image observations.
Initialize encoder $q_\phi(z|a_{t:t+k}, \bar{o}_t)$
Initialize decoder $\pi_\theta(\hat{a}_{t:t+k}|o_t, z)$
for iteration $n = 1, 2, ...$ **do**
  Sample o_t, $a_{t:t+k}$ from $\mathcal{D}$
  Sample z from $q_\phi(z|a_{t:t+k}, \bar{o}_t)$
  Predict $\hat{a}_{t:t+k}$ from $\pi_\theta(\hat{a}_{t:t+k}|o_t, z)$
  $\mathcal{L}_{reconst} = MSE(\hat{a}_{t:t+k}, a_{t:t+k})$
  $\mathcal{L}_{reg} = D_{KL}(q_\phi(z|a_{t:t+k}, \bar{o}_t) \parallel \mathcal{N}(0, I))$
  Update θ, ϕ with ADAM and $\mathcal{L} = \mathcal{L}_{reconst} + \beta\mathcal{L}_{reg}$

Algorithm 2 ACT Inference

Given: trained π_θ, episode length T, weight m.
Initialize FIFO buffers $\mathcal{B}[0 : T]$, where $\mathcal{B}[t]$ stores actions predicted *for* timestep t.
for timestep $t = 1, 2, ...T$ **do**
  Predict $\hat{a}_{t:t+k}$ with $\pi_\theta(\hat{a}_{t:t+k}|o_t, z)$ where $z = 0$
  Add $\hat{a}_{t:t+k}$ to buffers $\mathcal{B}[t : t + k]$ respectively
  Obtain current step actions $A_t = \mathcal{B}[t]$
  Apply $a_t = \sum_i w_i A_t[i] / \sum_i w_i$, with $w_i = \exp(-m * i)$

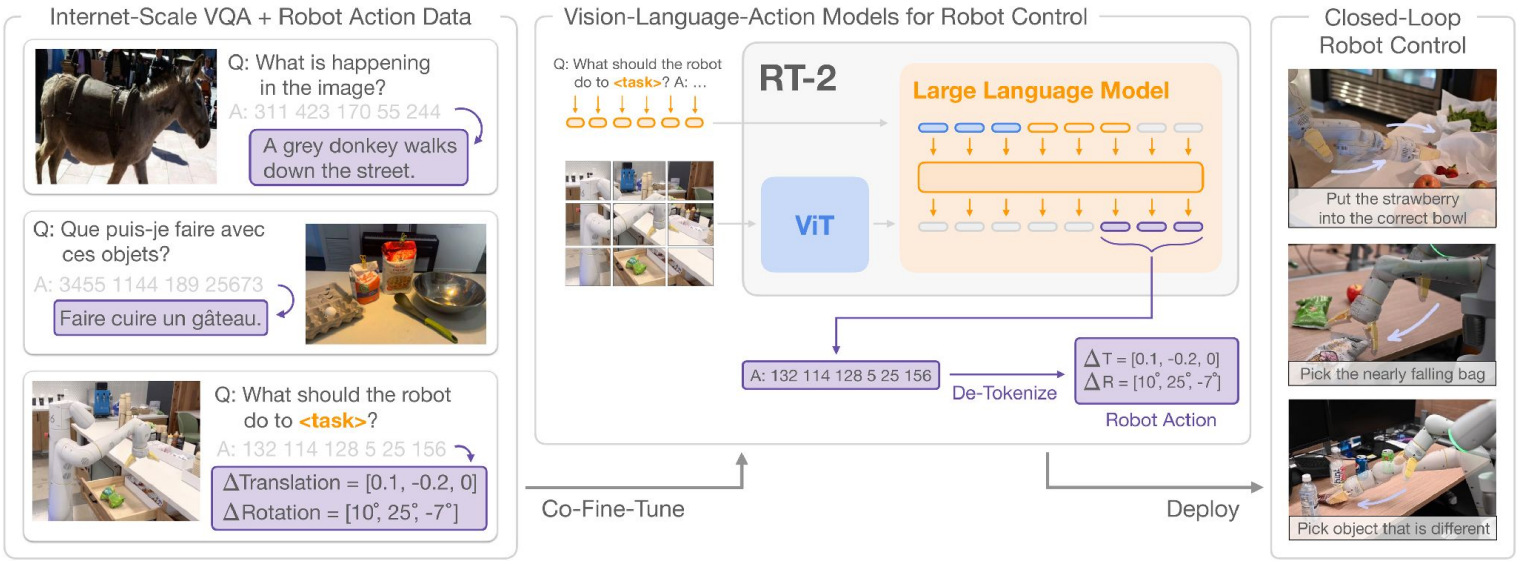
RoboCat (2023) — a generalist that grows its own data



Novelty at the time · Bousmalis, Vezzani, Rao et al. RoboCat: A Self-Improving Generalist Agent for Robotic Manipulation. 2023. arXiv:2306.11706

- Один agent на несколько arms и много tasks: fine-tune, deploy, потом учиться на своих rollouts.
- VQ-GAN tokens для images + sequence model для actions — generalist recipe до Open-X VLA.
- Self-improvement loop: policy одновременно skill и сборщик данных для следующей итерации.
- Few-shot adaptation к новому embodiment или набору объектов, не только к новой language command.
- «Generalist» может значить растущий датасет и семью роботов, а не только больший frozen backbone.

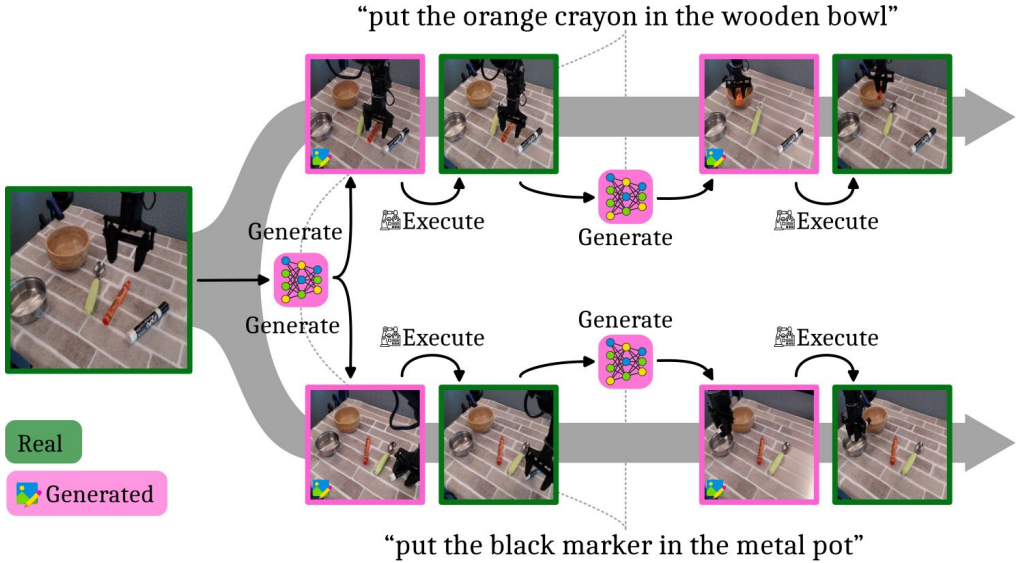
RT-2 (2023) — the first vision-language-action model

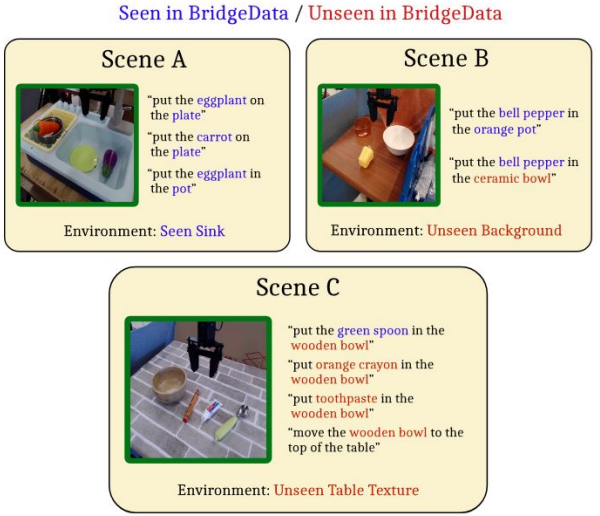


Novelty at the time · Brohan, Brown, Carbajal et al. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. CoRL 2023. arXiv:2307.15818

- Co-fine-tune веб-VLM на robot trajectories и VQA: motor commands — обычные language tokens.
- Первый явный transfer веб-знаний в control: icons, semantic categories, unseen objects, chain-of-thought.
- Замкнули петлю, которую открыл RT-1: backbone больше не robotics-only transformer, а предобученный VLM.
- Зафиксировали VLA recipe, которому OpenVLA, π0 и Gemini Robotics до сих пор следуют.
- Частота всё ещё ~3–10 Hz discrete tokens — dexterity осталась у diffusion / flow, а не у самого RT-2.

SuSIE (2023) — subgoals as image edits, not as language

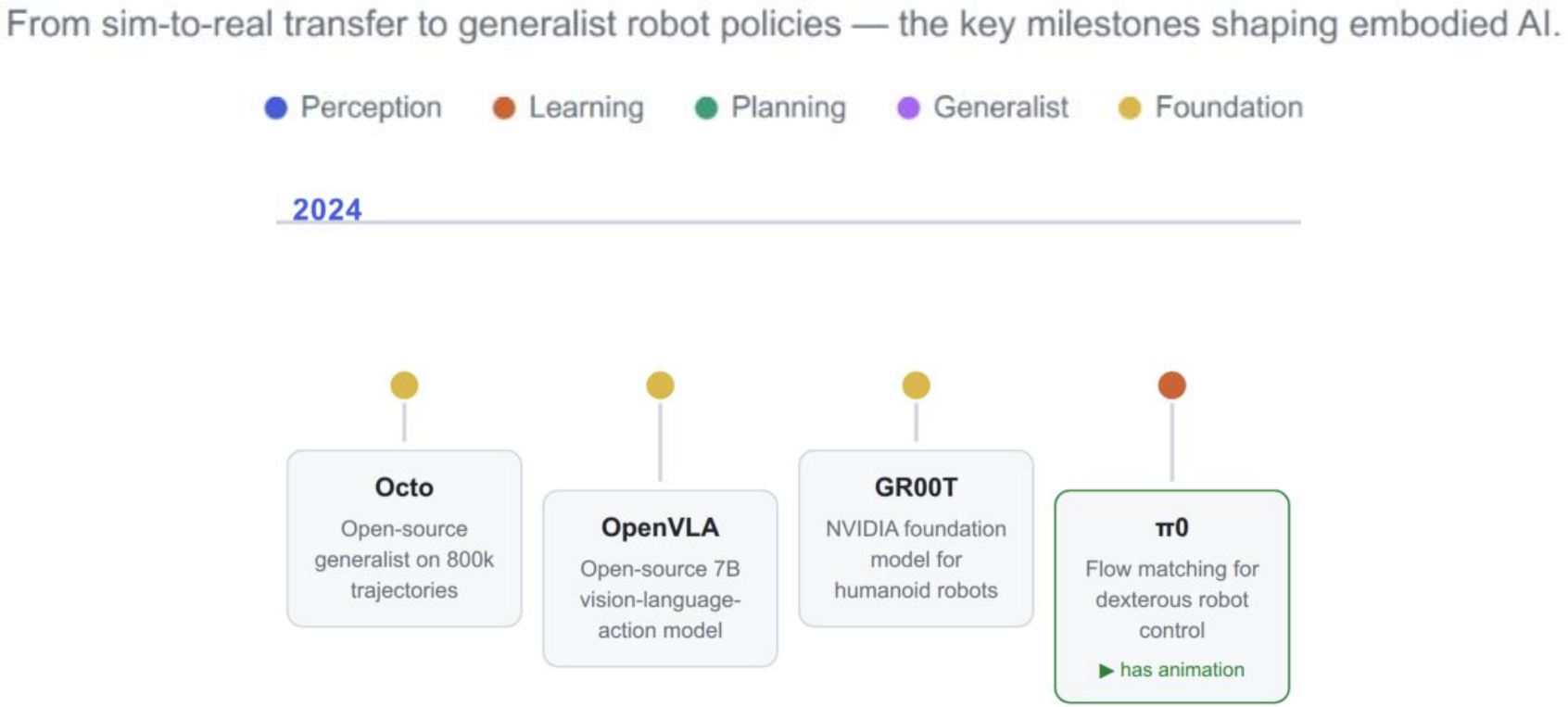




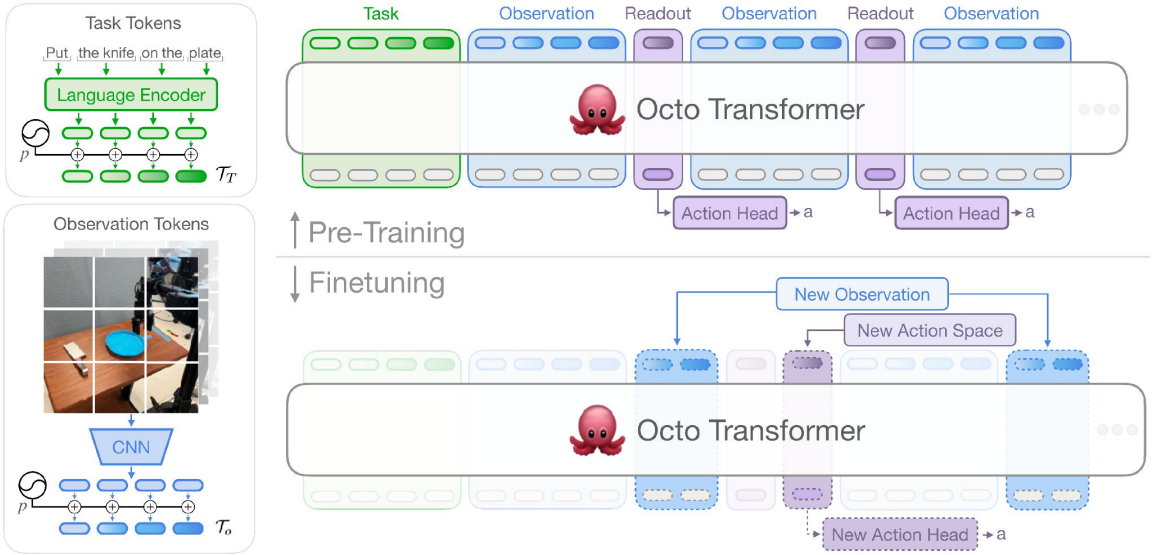
Novelty at the time · Black, Nakamoto, Atreya, Walke, Finn, Kumar, Levine. Zero-Shot Robotic Manipulation with Pretrained Image-Editing Diffusion Models. 2023. arXiv:2310.10639

- Image-editing diffusion model (InstructPix2Pix-style) рисует следующий subgoal image из language.
- Отдельная low-level goal-conditioned policy (diffusion policy) доезжает до отредактированной картинки — иерархия без skill library.
- Zero-shot long-horizon из human video / image priors, а не из полной task demonstration.
- Subgoals живут в pixel space и композируются проще, чем discrete skill name.
- Предок стеков «video / image as the plan» (SuSIE → UniPi-подобные planners → world-model policies).

Эволюция Robot Foundation Models



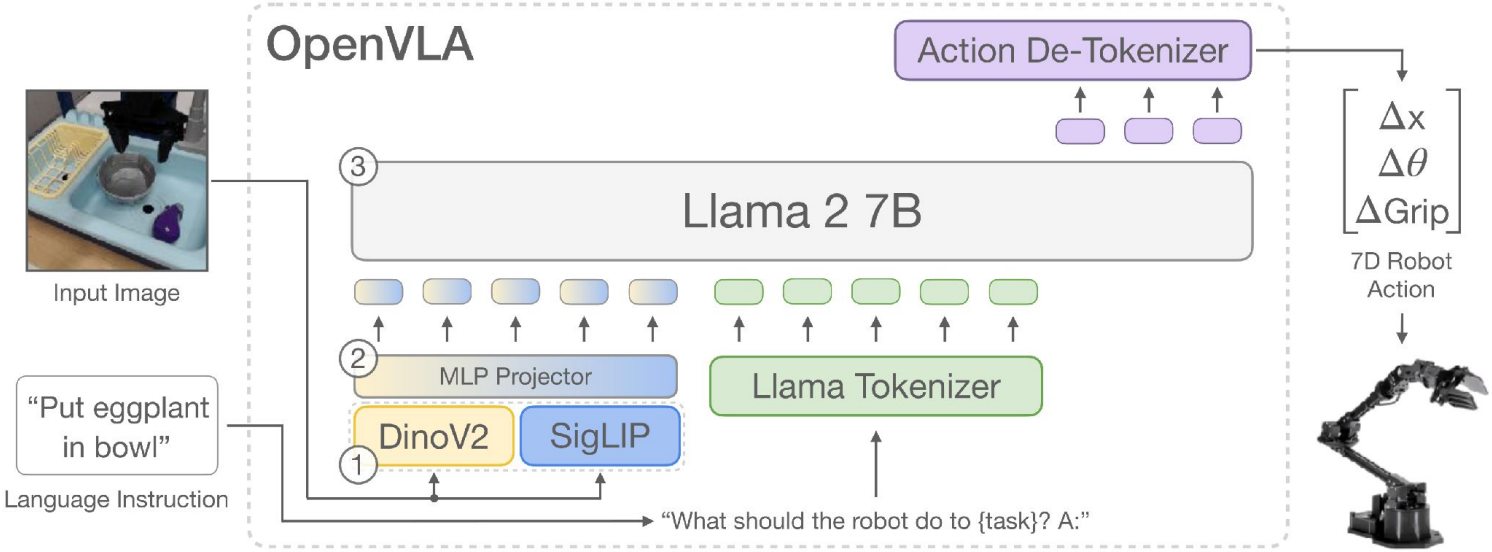
Octo (2024) — the open generalist anyone could fine-tune



Novelty at the time · Octo Model Team. Octo: An Open-Source Generalist Robot Policy. RSS 2024. arXiv:2405.12213

- Первый открытый transformer-generalist на ~800k траекториях Open-X-Embodiment, с public weights.
- Modular blockwise attention по language, images и proprioception — не closed checkpoint RT-2.
- Diffusion action head вместо discrete tokens, поэтому fine-tuning на новом роботе реально работает.
- Стал community baseline, который можно ablate: OpenVLA и следующие VLA должны были обойти Octo, а не только RT-2.
- Показали: Open-X достаточно для generalist, если architecture модульная, а actions непрерывные.

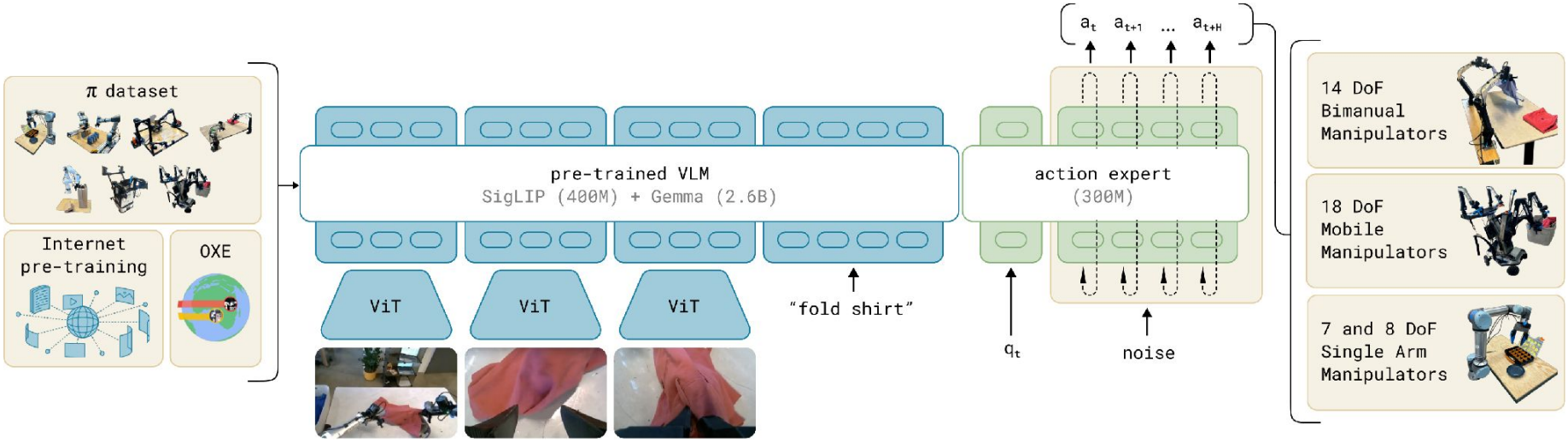
OpenVLA (2024) — an open 7B VLA that matched closed models



Novelty at the time · Kim, Pertsch, Karamcheti et al. OpenVLA: An Open-Source Vision-Language-Action Model. CoRL 2024. arXiv:2406.09246

- Open 7B VLA: SigLIP + DinoV2 → MLP projector → Llama 2, обучение на ~970k эпизодах Open-X.
- Первая воспроизводимая open-weight модель класса RT-2, включая LoRA-рецепт под нового робота за ночь.
- Dual vision encoders (semantics + spatial) вместо одной CLIP-башни — дизайн, который потом копируют.
- Выигрывает на language-grounded multi-object задачах; на одном точном skill всё ещё проигрывает Diffusion Policy.
- Превратили VLA из Google-only демо в то, что лаборатория может скачать, fine-tune и цитировать.

π0 (2024) — a VLM with a flow-matching action expert



Novelty at the time · Black, Brown, Driess et al. π0: A Vision-Language-Action Flow Model for General Robot Control. 2024. arXiv:2410.24164 · Lipman et al., Flow Matching, ICLR 2023, arXiv:2210.02747

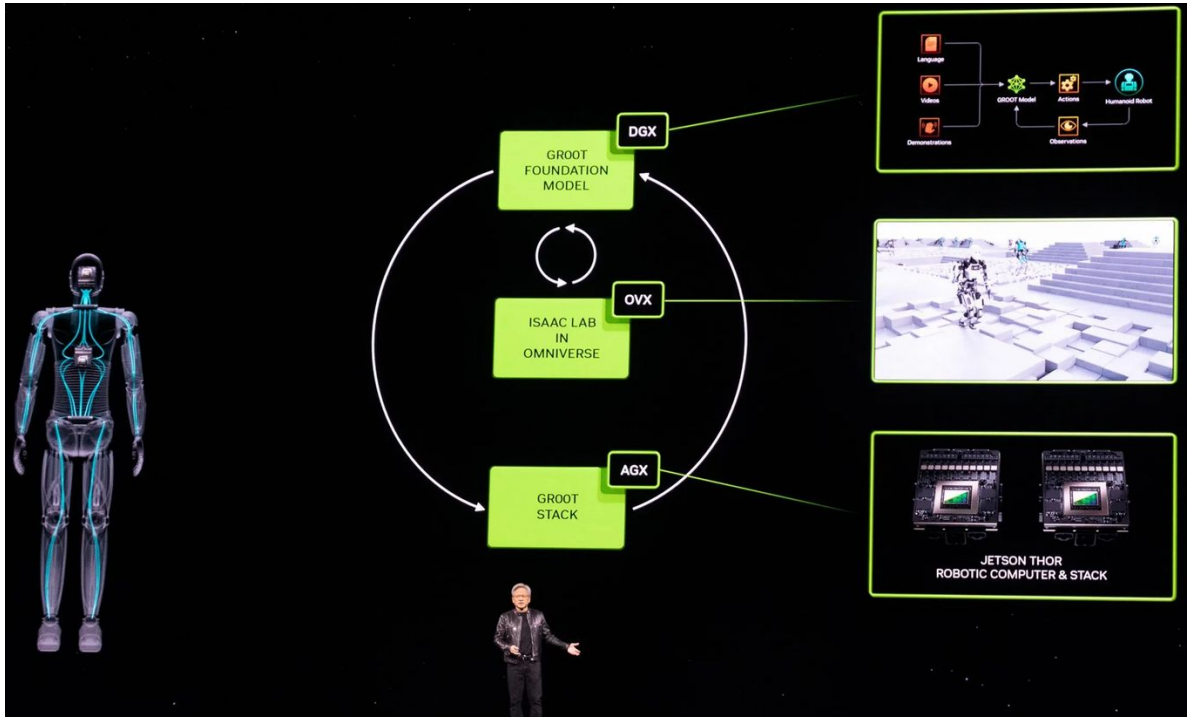
- Discrete action tokens заменены на flow matching: vector field из шума в continuous action chunk.
- Предобученный VLM для semantics + отдельный action expert — dexterity без отказа от language.
- Pretrain на 7 конфигурациях роботов / 68 задачах, затем prompt или fine-tune (laundry, bussing, assembly).
- Почти прямые OT (Optimal Transport) -пути flow matching требуют гораздо меньше sampling steps, чем изогнутый diffusion SDE.
- Стал default VLA recipe 2025–26: tokens для контекста, flow / diffusion для motor control (π0.5, π0.6).

GR00T (2024) — NVIDIA’s humanoid foundation-model bet

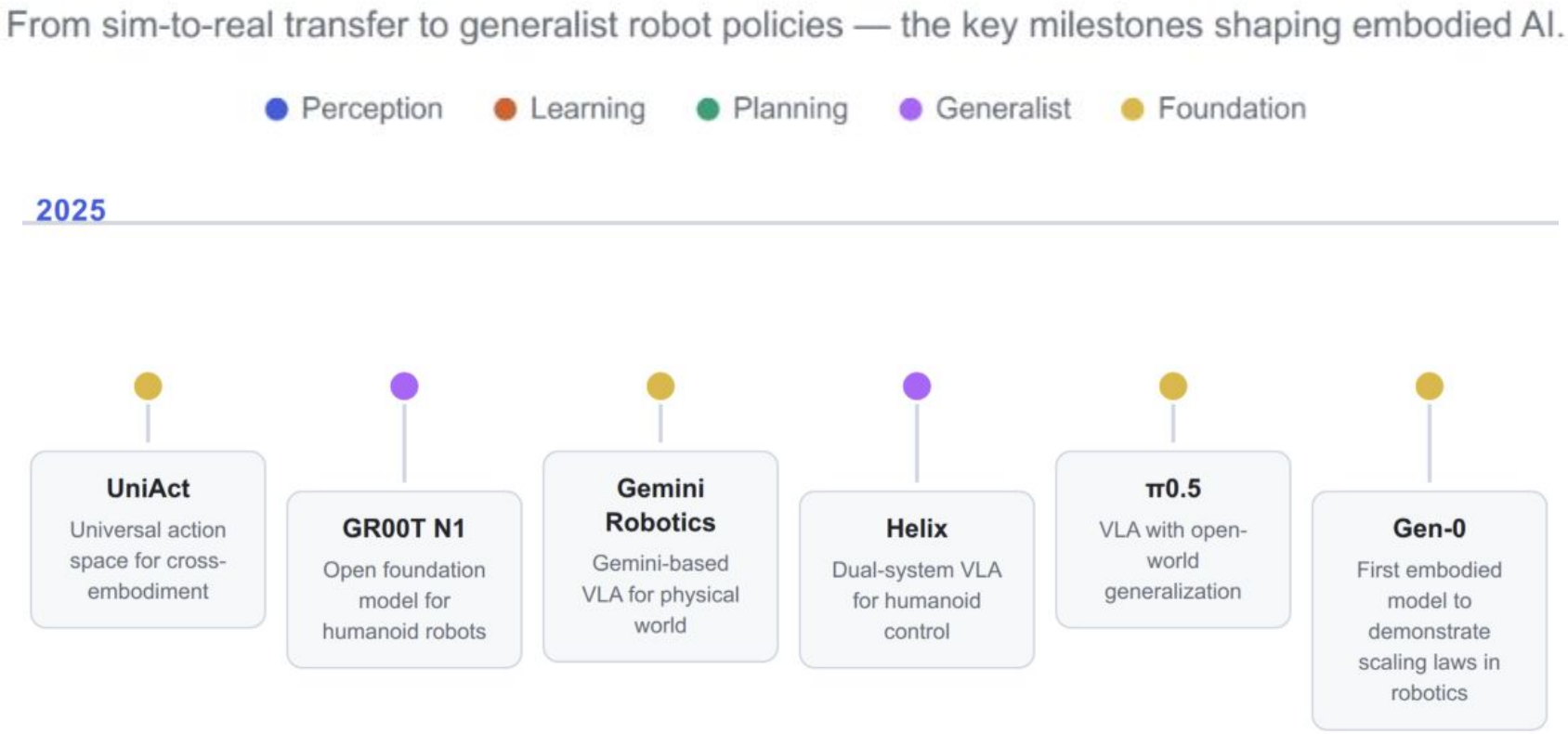
NVIDIA. Project GR00T announcement, 2024. (weights and the N1 paper follow in 2025, arXiv:2503.14734)

Novelty at the time

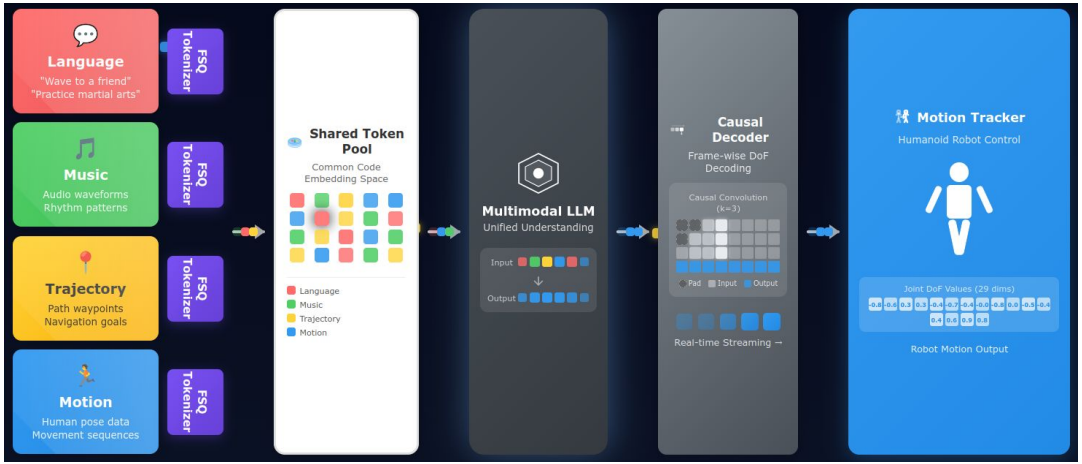
- Первый крупный industry анонс foundation model для humanoids, а не tabletop arms.
- Новостью был data claim: robot trajectories + human video + synthetic data в одном pretraining mix.
- Cross-embodiment как цель первого класса — один мозг, много тел
- Задали повестку 2025 (dual-system humanoid VLA, EgoScale, Cosmos), хотя 2024 был ещё teaser.
- Новизна тогда — product category, а не публичная architecture или число на benchmark.

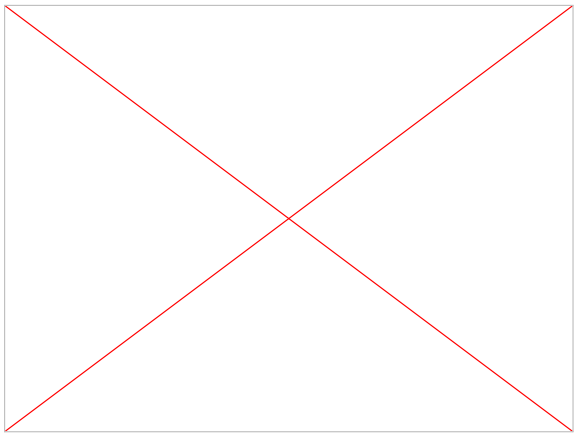


Эволюция Robot Foundation Models



UniAct (2025) — one action space across robot bodies

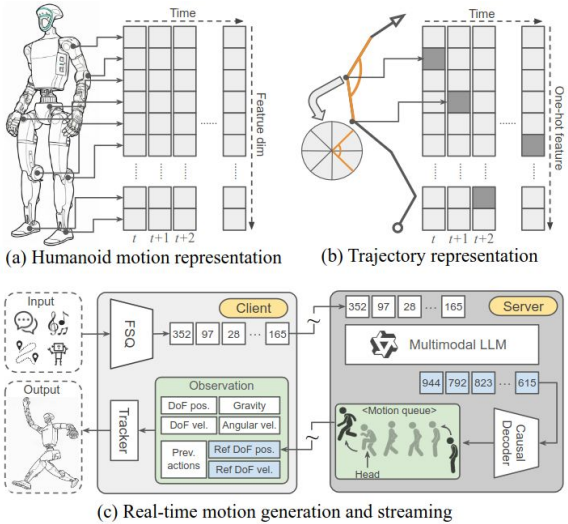




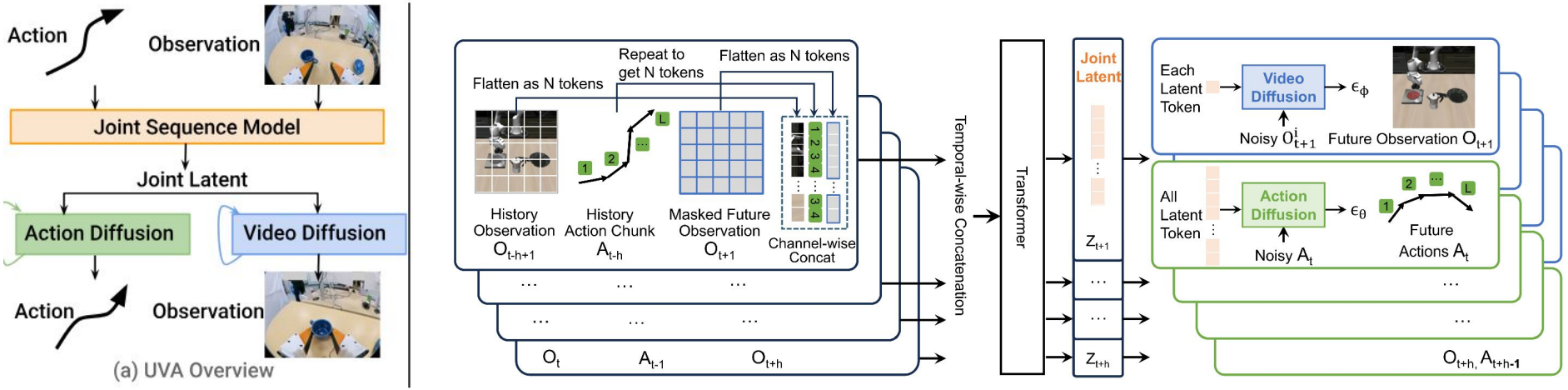
Yu, Xu, Zhang et al. UniAct: Universal Action Space for Embodied Cross-Embodiment Policy Learning. 2025. arXiv:2501.10105

Novelty at the time

- Cross-embodiment сдвинули от «shared backbone, per-robot head» к universal action vocabulary.
- Heterogeneous robots мапятся в одно action space, поэтому данные WidowX могут учить Airbot.
- Embodiment gap частично проблема action interface, а не только vision encoder.



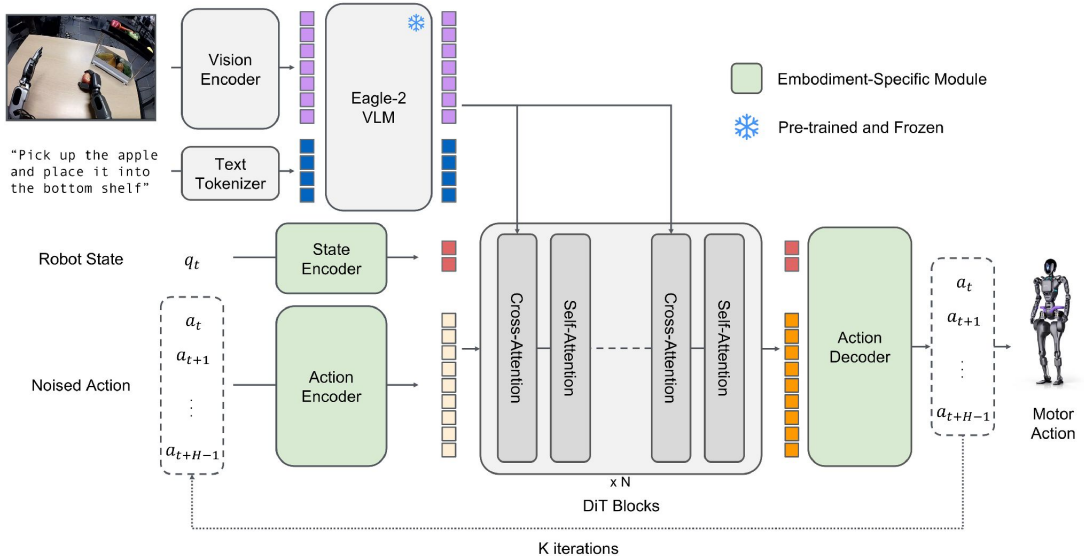
UVA (2025) — video and actions from one latent, pixels optional



Novelty at the time · Li, Gao, Sadigh, Song. Unified Video Action Model. 2025. arXiv:2503.00200

- Общий transformer latent Z на историю images и action chunks; две маленькие heads декодируют video и actions.
- На deploy video head отключают — inference на control-rate, знание video model остаётся в weights.
- Masked training: одна сеть — и policy, и video model, и forward / inverse dynamics.
- Выигрывал «generate video then read actions» по latency и action-only diffusion на части multi-task setups.
- Урок UniPi в инженерии: video pretraining оставить, pixel generation из control loop убрать.

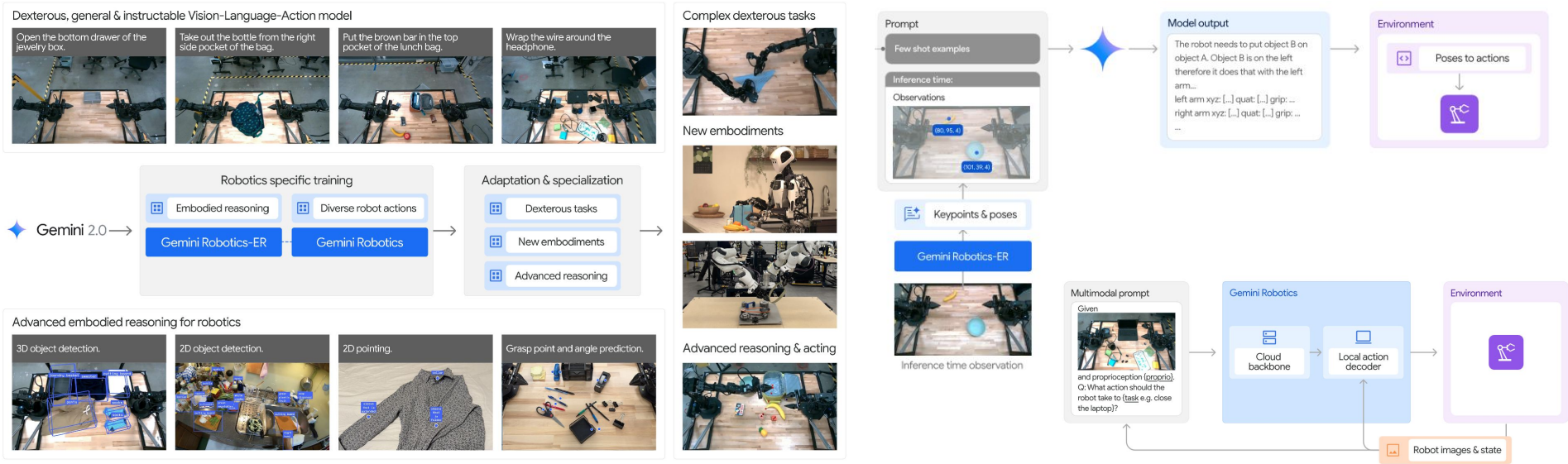
GR00T N1 (2025) — open dual-system humanoid VLA



Novelty at the time · NVIDIA, Bjorck, Castañeda et al. GR00T N1: An Open Foundation Model for Generalist Humanoid Robots. 2025. arXiv:2503.14734

- Первая открытая NVIDIA humanoid VLA: System 2 (VLM, slow) + System 1 (diffusion transformer, fast motor commands).
- Heterogeneous pretraining mix: robot trajectories, human video и synthetic data в одной модели.
- Dual-system split — главный architectural claim: reasoning и reflexes — не одна сеть на одном Hz.
- Лаборатории получили humanoid checkpoint, который можно запустить, а не только keynote 2024.
- Линия: Helix-style System 1/2, Octo-style open weights, π-style continuous actions.

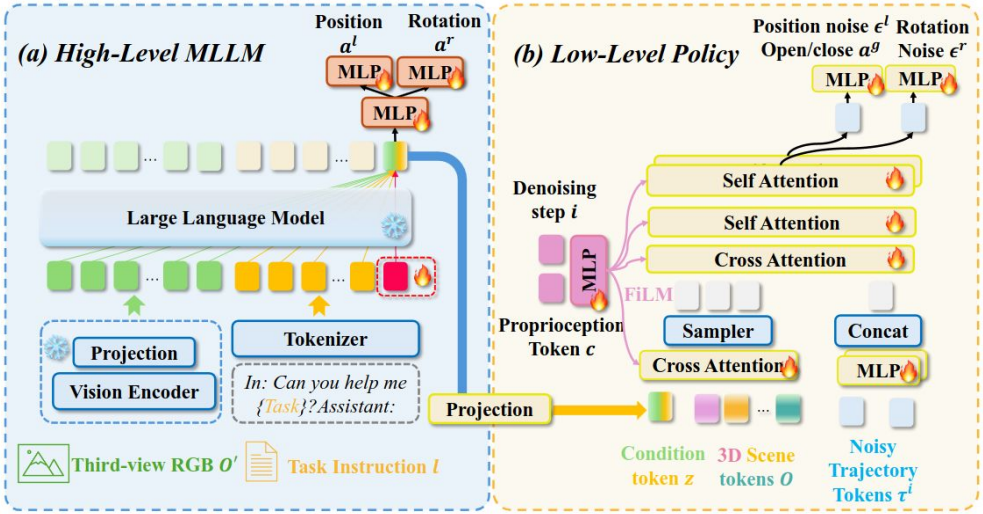
Gemini Robotics (2025) — a frontier VLM that thinks, then acts

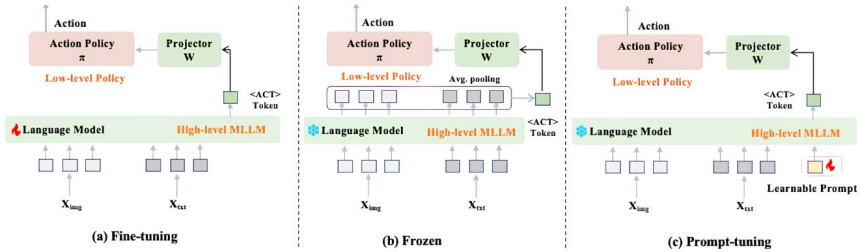


Novelty at the time · Google DeepMind. Gemini Robotics: Bringing AI into the Physical World. 2025. arXiv:2503.20020

- В robot loop поставили frontier Gemini VLM, а не 7B open Llama fine-tune.
- Embodied reasoning: spatial relations, affordances и multi-step plans на language до motor tokens.
- Веб-scale multimodal pretraining уже содержит много physical common sense — если модели дать говорить.
- Think-then-act — противоположность pure reactive VLA: chain-of-thought часть control.
- Closed-model counterpart к OpenVLA / π0: та же VLA idea, гораздо больший pretrained brain.

Helix (2025) — one VLA, whole humanoid upper body, 200 Hz

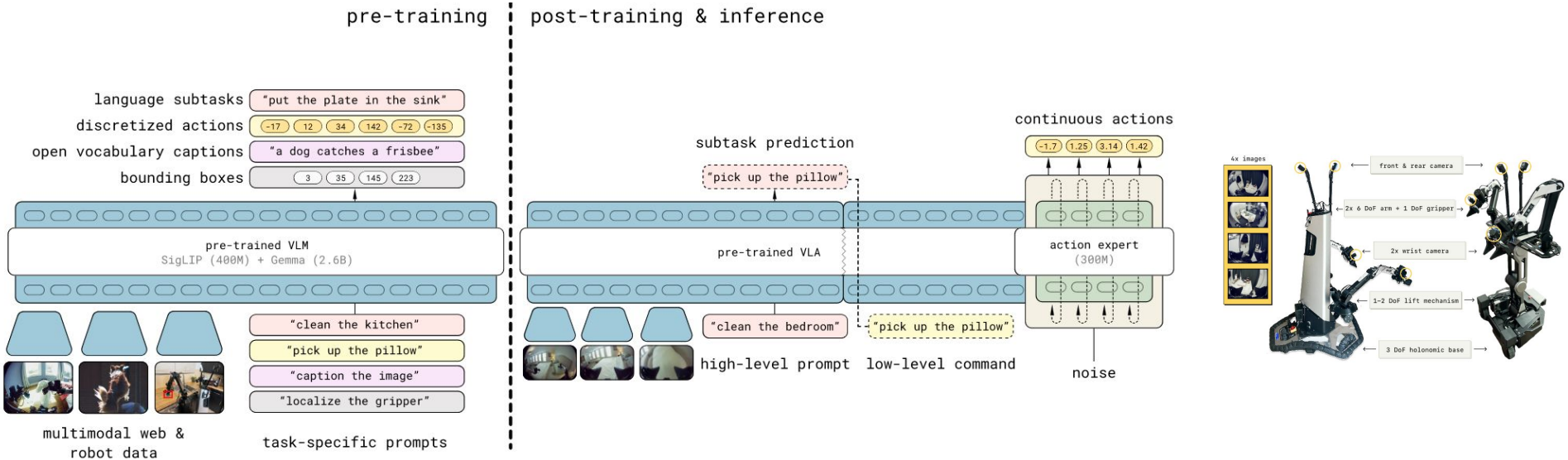




Novelty at the time · Figure AI. Helix: A Vision-Language-Action Model for Generalist Humanoid Control. 2025. figure.ai/news/helix

- Первая VLA, которая командует всей humanoid upper body — arms, wrists, torso, head и fingers — из language.
- Dual system на embedded GPUs: S2 — 7B VLM на 7–9 Hz; S1 — 80M visuomotor transformer на 200 Hz.
- Один набор weights, без task-specific fine-tune; ~500 часов multi-operator teleop с hindsight language labels.
- Onboard, low-power, closed-loop на dexterous rates — deployment claim, а не лабораторная 3 Hz token policy.
- Сделали «System 1 / System 2 humanoid VLA» default story, которую потом открыл GR00T N1.

π0.5 (2025) — open-world VLA, not only a tabletop



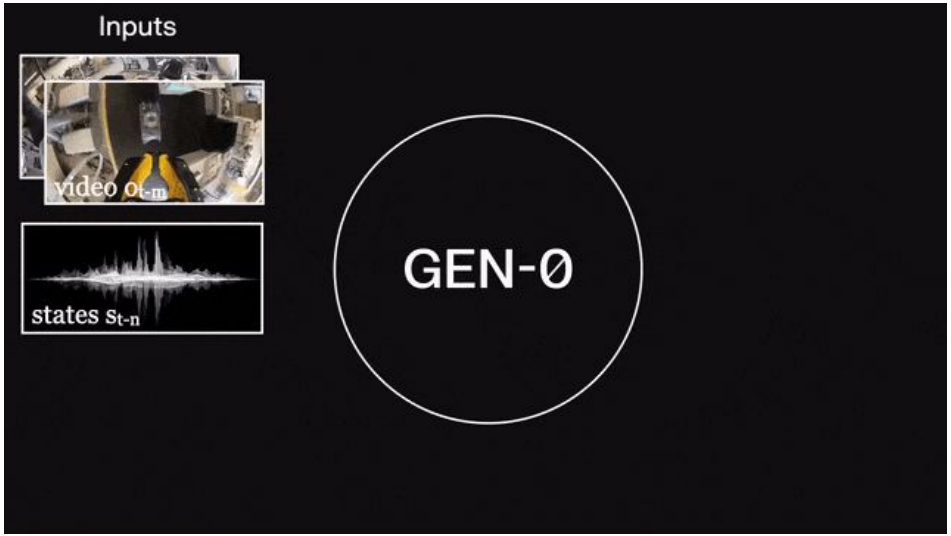
Novelty at the time · Physical Intelligence. π0.5: a VLA with Open-World Generalization. 2025. arXiv:2504.16054

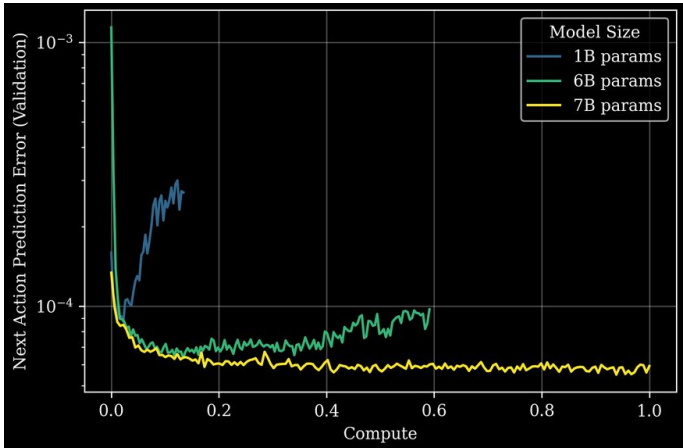
- Сдвинули π0 от scripted tabletop skills к целым комнатам: kitchens, bedrooms, unseen layouts.
- Hierarchy: high-level inference по language / subgoals, low-level flow-matching motor expert.
- Обучение на open-world generalization — новые объекты и новые комнаты на test time, не фиксированный object set.
- Сделали «VLA in the house», не только demo video.
- Прямой шаг к π0.6: тот же stack, больше open-world data и более плотный high-level loop.

Gen-0 (2025) — scaling laws claimed for physical AI

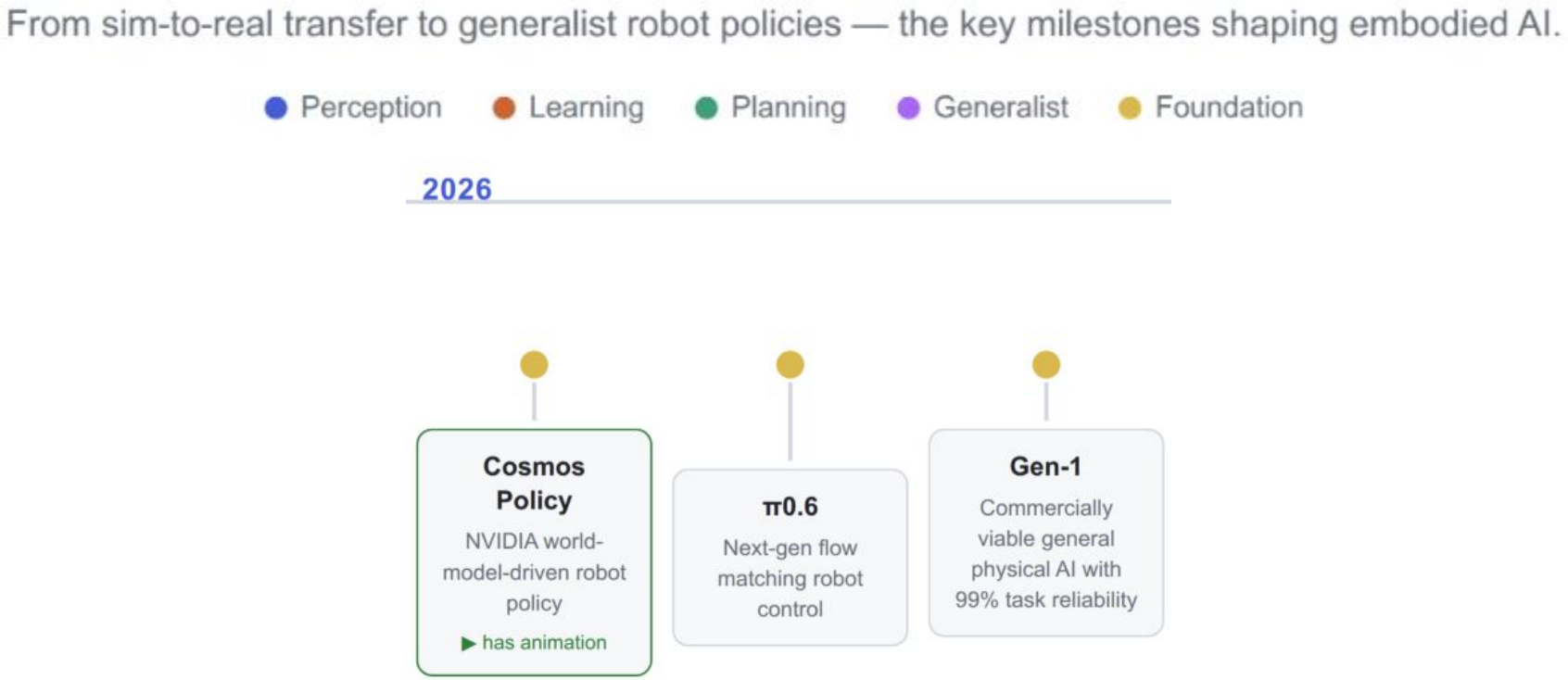
Generalist AI. Gen-0: the first embodied model to demonstrate scaling laws in robotics. Nov 2025. generalistai.com

- Industry claim: model, чья performance масштабируется предсказуемо с data и compute, как у LLM.
- Новизна тогда — framing scaling laws для роботов, которого у большинства academic VLA ещё не было.
- General physical AI как foundation-model product, а не per-task imitation paper.
- Closed weights и blog eval — цифры как vendor claim, не как публичный benchmark.
- Подготовили reliability pitch Gen-1: сначала scale, потом commercial robustness.

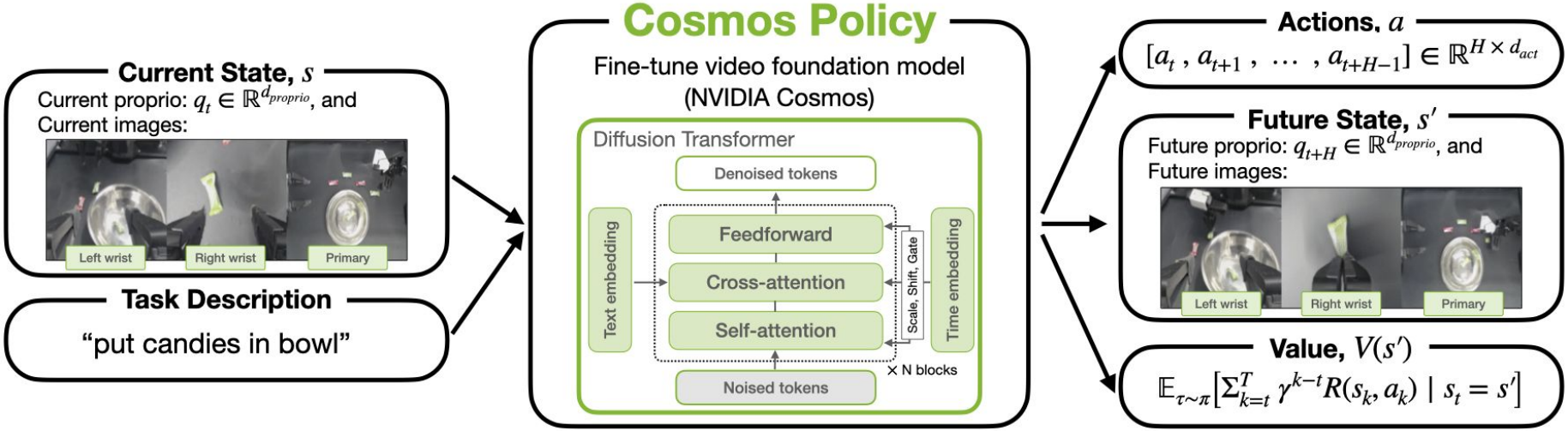




Эволюция Robot Foundation Models



Cosmos Policy (2026) — a world-model-driven robot policy



Novelty at the time · NVIDIA. Cosmos Policy. 2026. arXiv:2601.16163

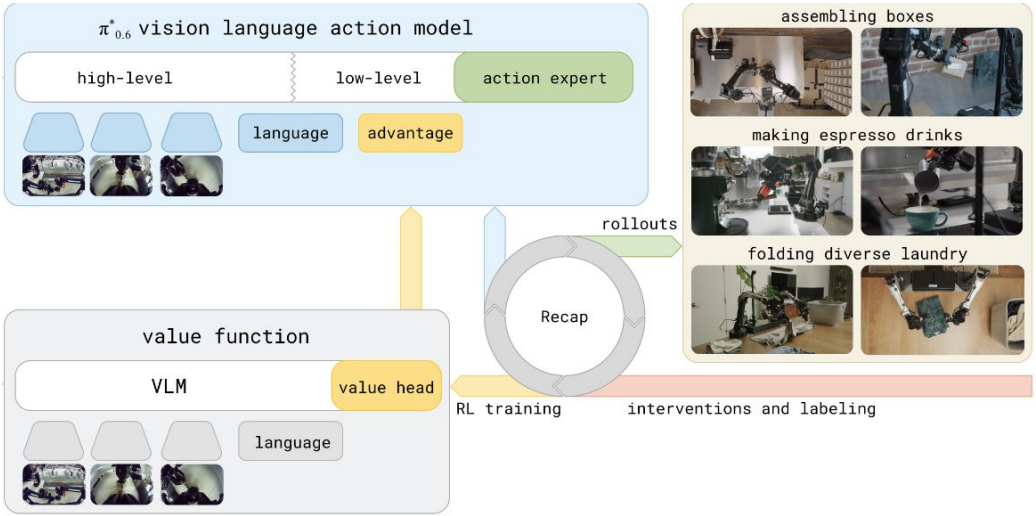
- Robot policy планирует внутри video world model NVIDIA Cosmos, а не только имитирует action chunks.
- Отделяет «как будет выглядеть мир» от «какую motor command послать» — идея UniPi, в latents, на foundation scale.
- Diffusion / generative rollout как imagination, затем policy читается из imagined future.
- Первая громкая industrial world-model policy после десятилетия исследований в духе Dreamer.
- Замыкает петлю 2018–2026 на timeline: world models больше не игрушка, а product stack.

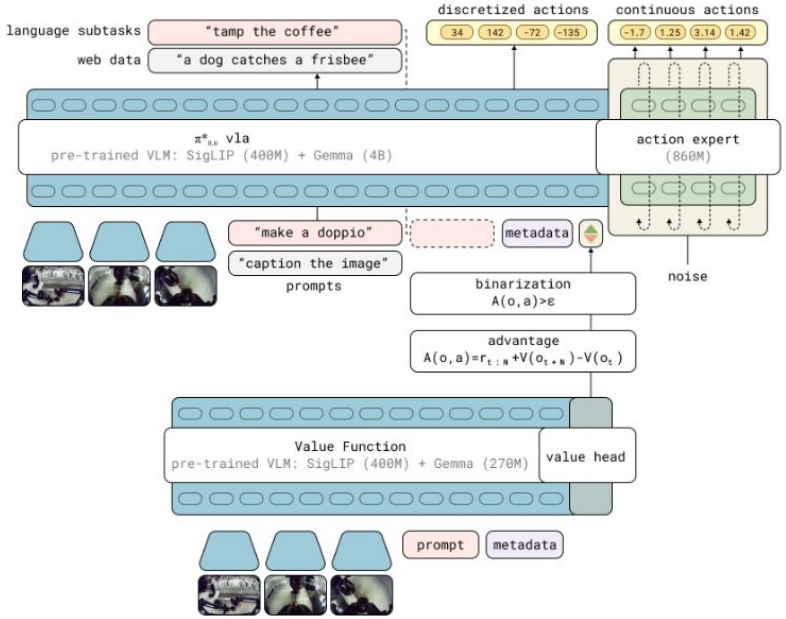
π0.6 (2026) — next-generation flow-matching control

Physical Intelligence. π0.6. 2026. pi.website/blog/pistar06

Novelty at the time

- Добавили Value head к VLA и показали как делать дообучение с RL
- Blog numbers — vendor-reported; architectural story — продолжение π0.5, нет открытого кода
- flow matching всё ещё winning action expert в 2026.



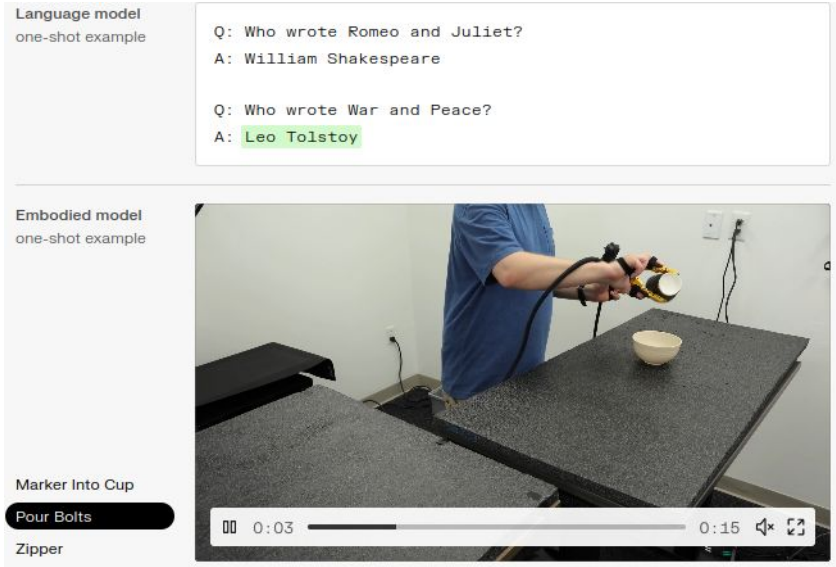


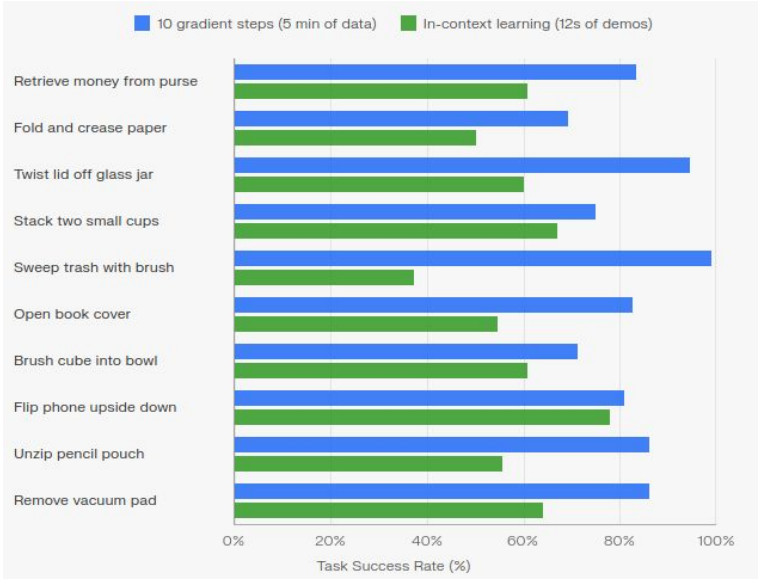
GEN 1.5 (2026) — zero-shot physical generalization

Generalist AI. GEN-1.5 2026. https://generalistai.com/blog/gen-1.5

Novelty at the time

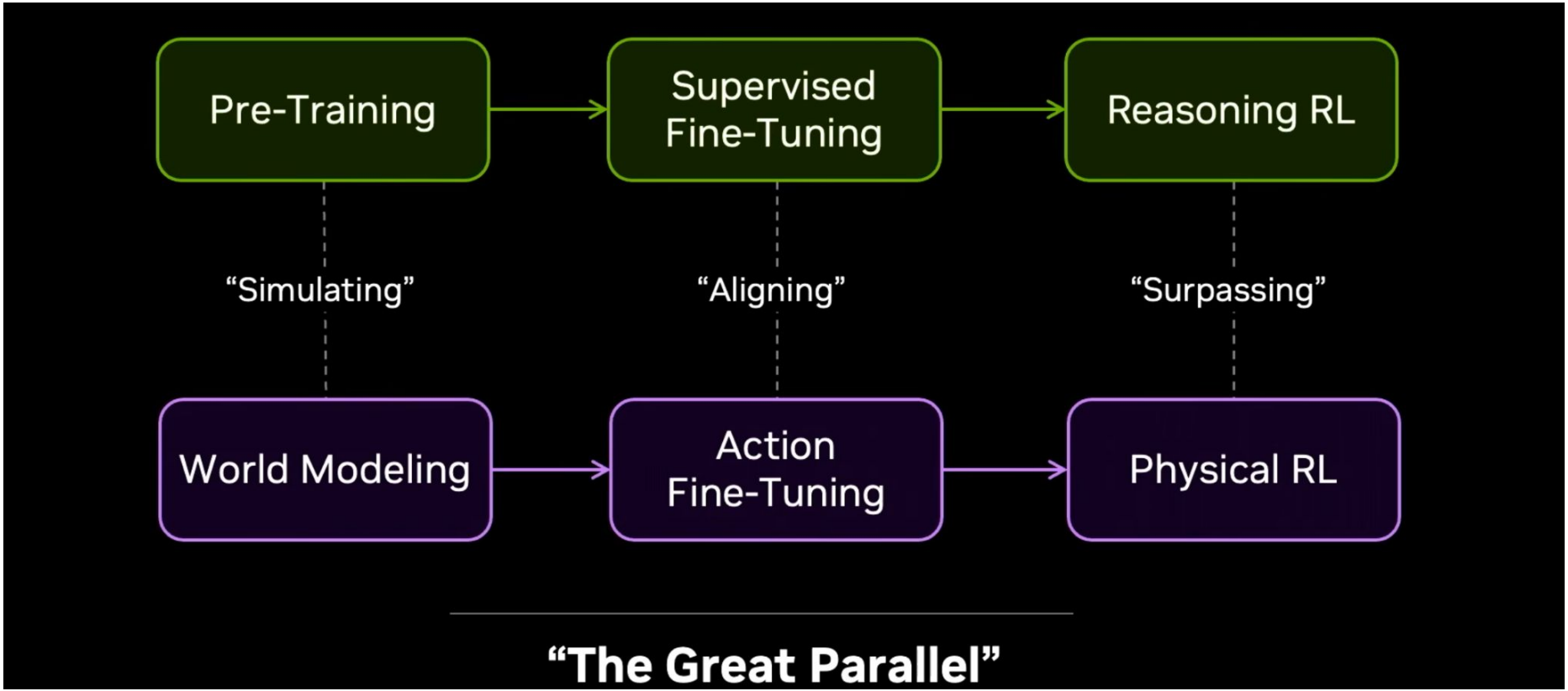
- Пример выполнения задачи как цель для робота (One-shot Learning)
- Обработка входных данных: Модель анализирует видеопоток (с объемом контекстной памяти до 30 секунд), текстовые (языковые) инструкции, проприоцептивные данные и сигналы от других сенсоров.



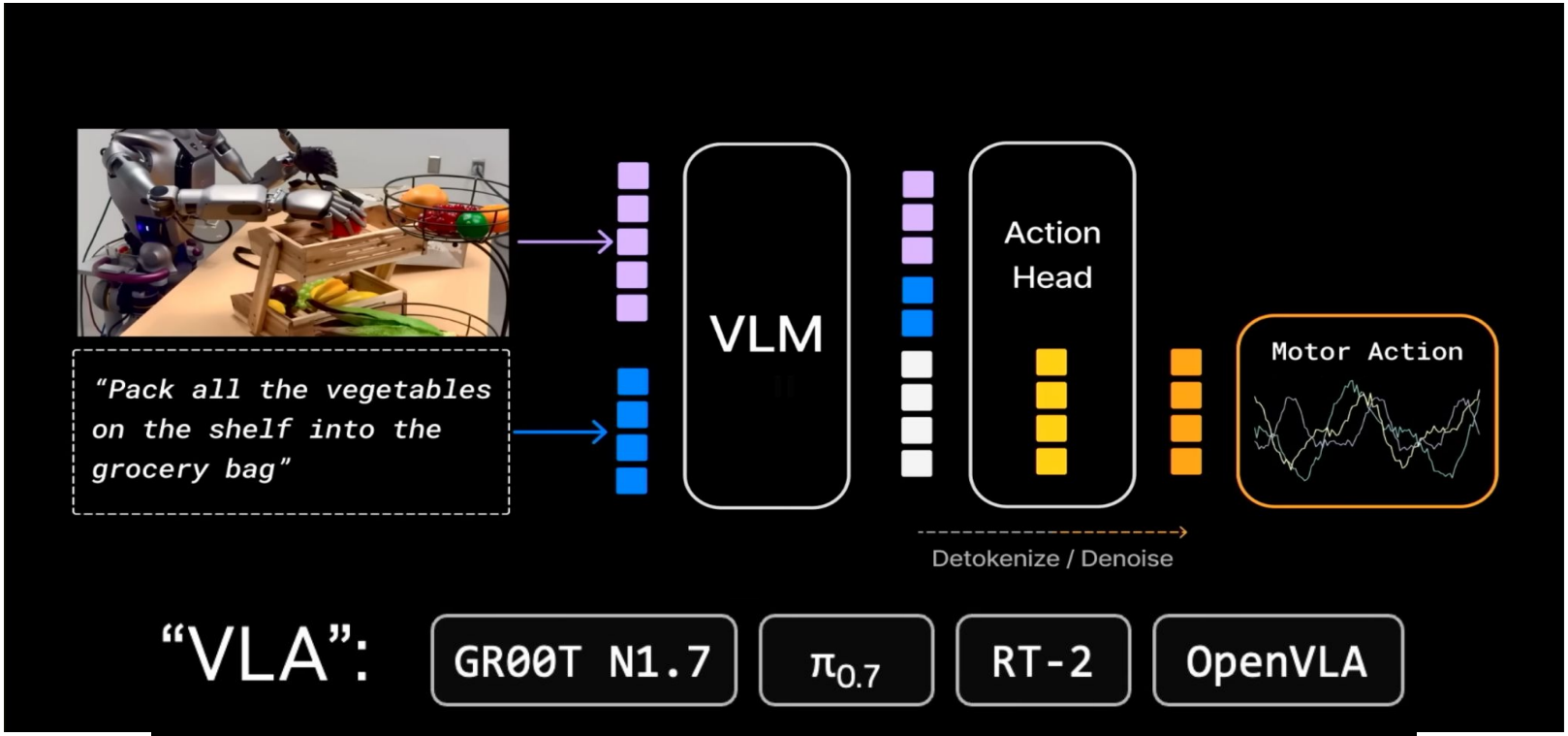


Future of VLA models

Robotics' End Game



Where the most amount of parameters sits?

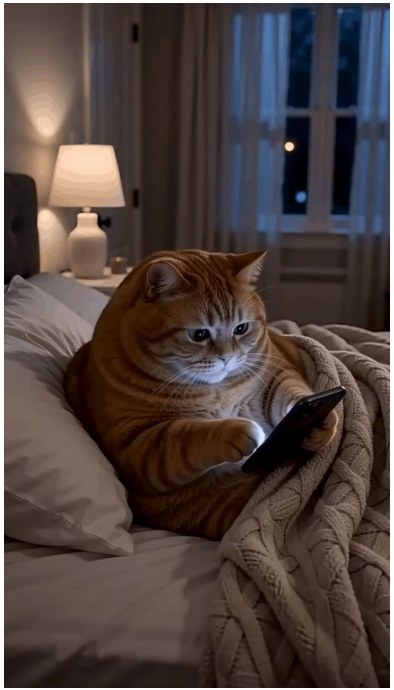


What pre-training abilities we are looking for?

Text generalization
(GPT, LLAMA, etc.)



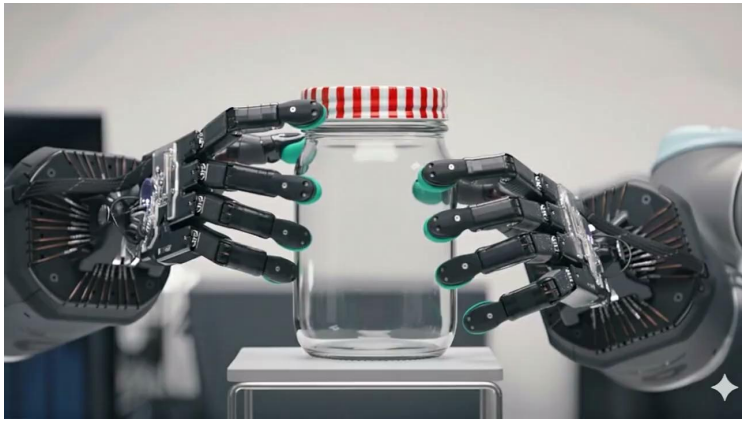
Video generalization
(Video World models)



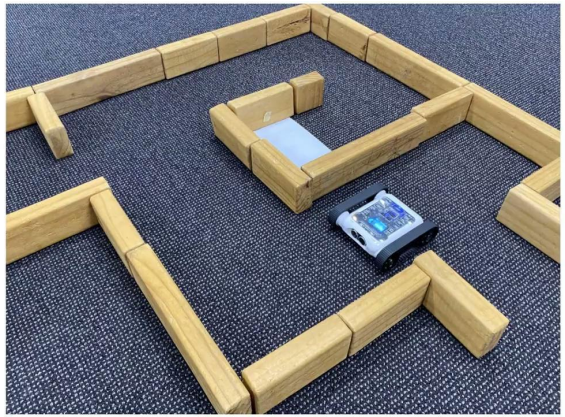


From AI video slop to physics and reasoning

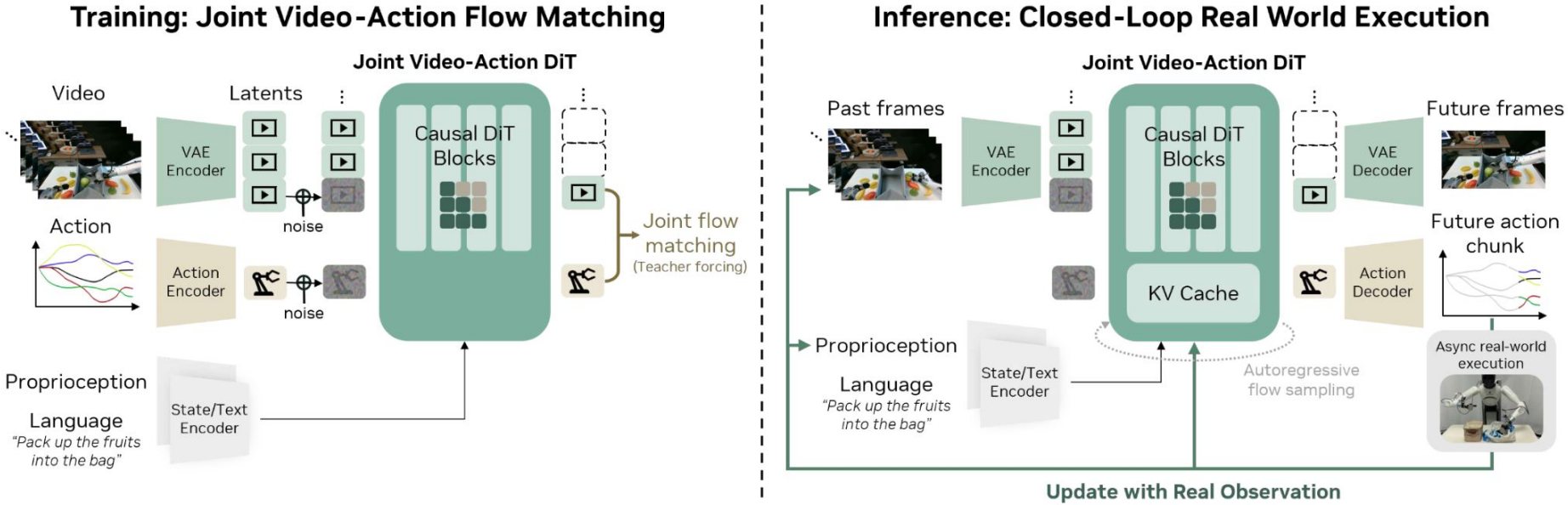






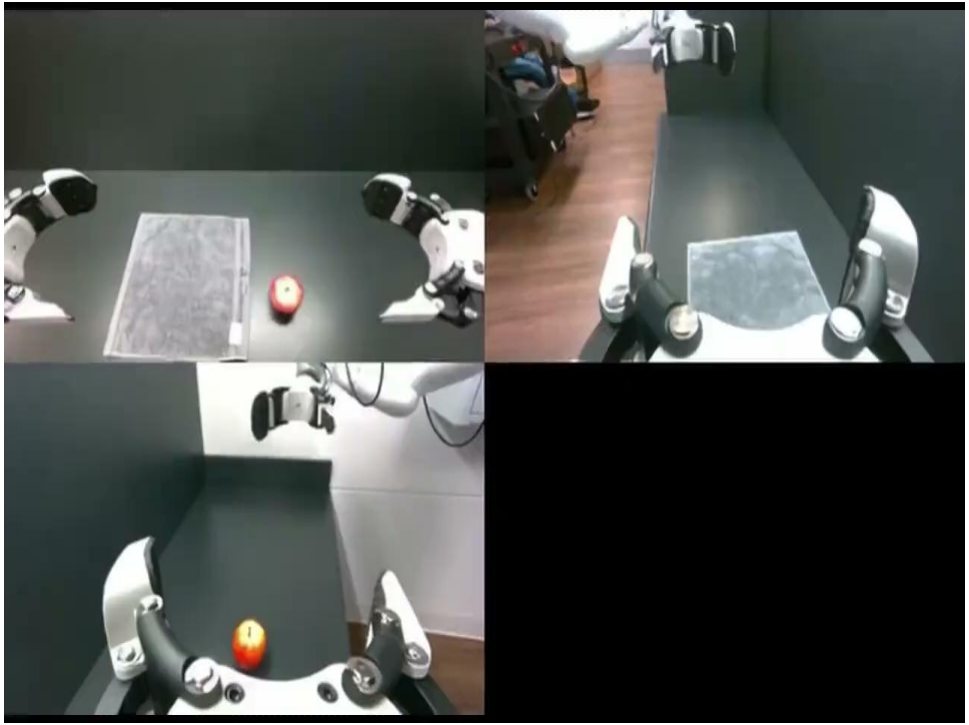


DreamZero WAM (Model Architecture)

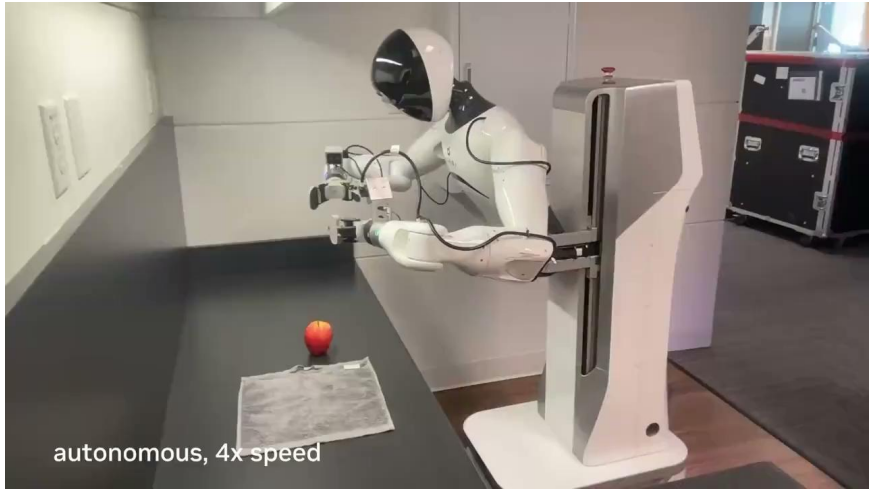


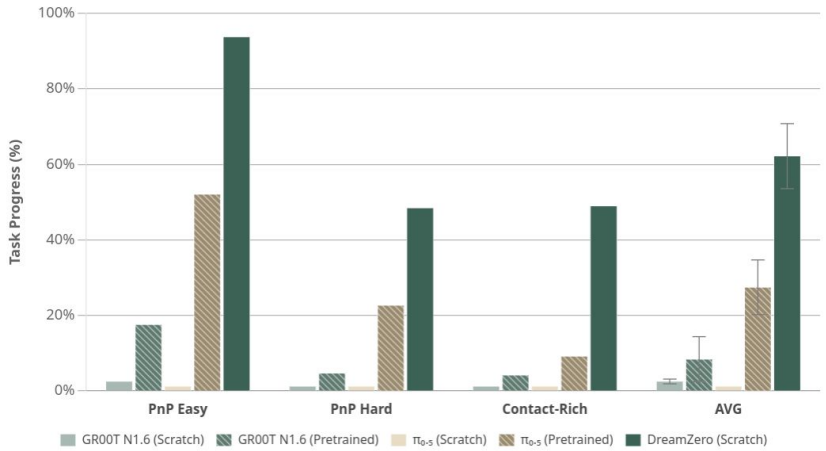
DreamZero WAM (Results)

Generated



Real-world task execution



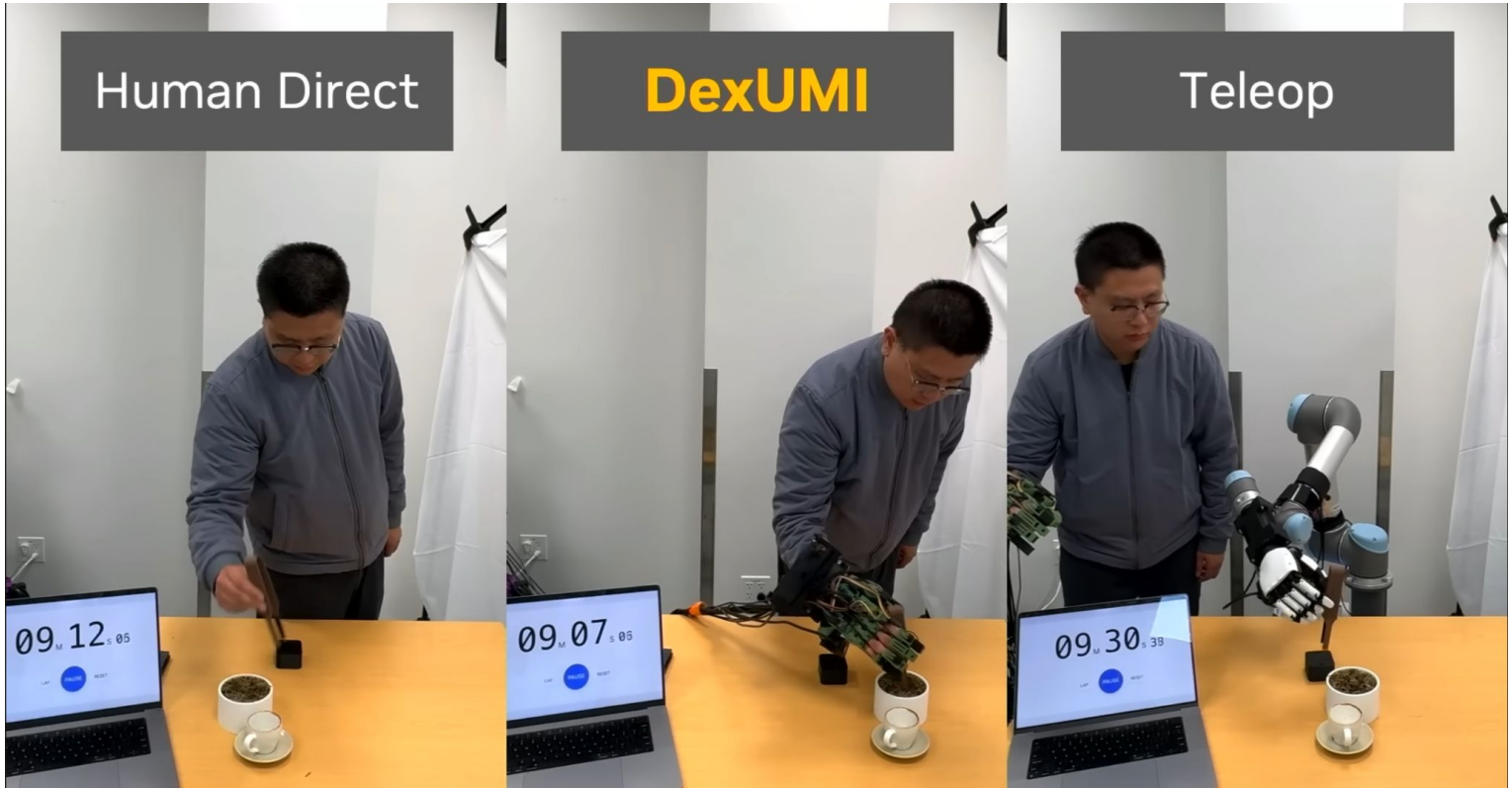


Robotics' End Game

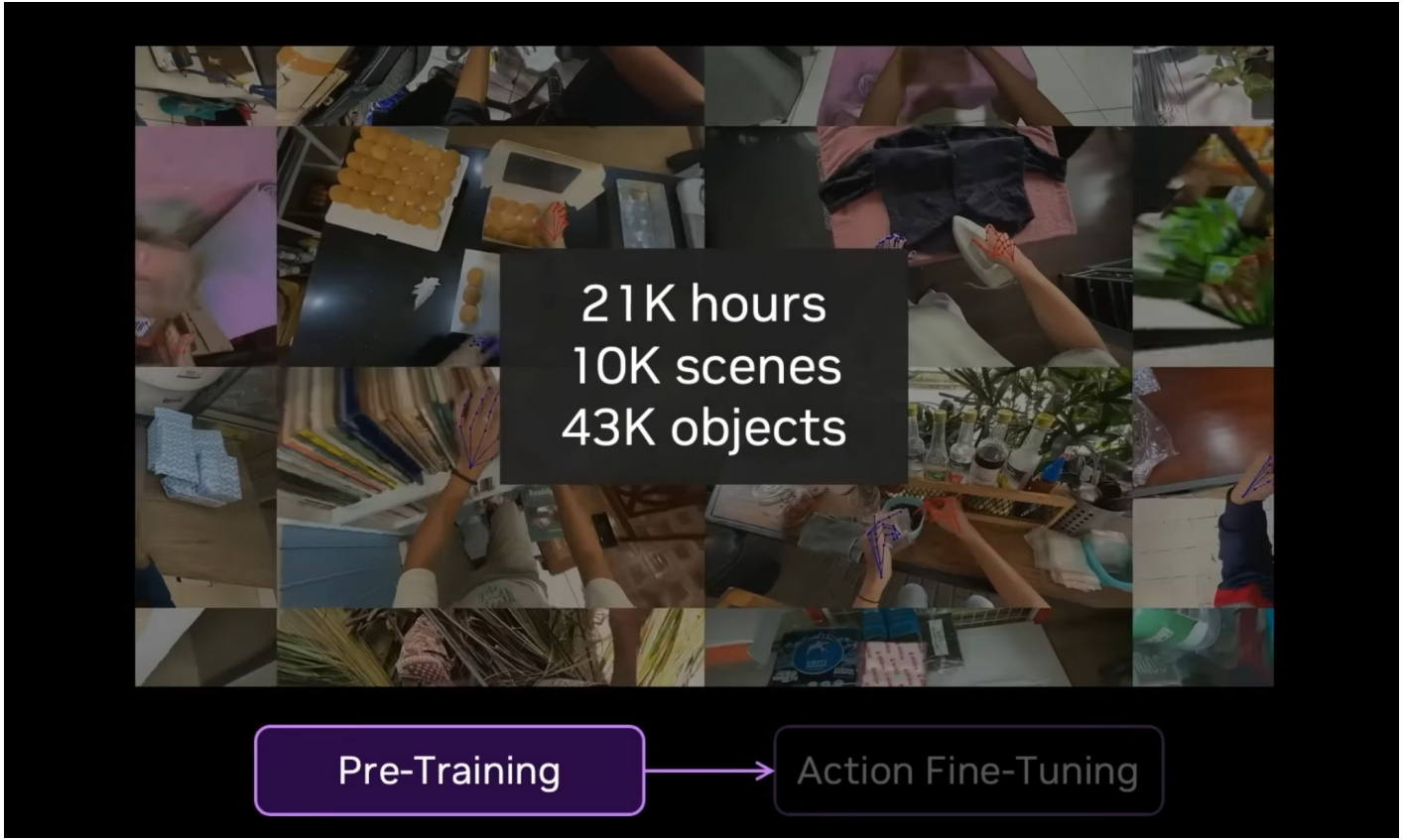
1st prediction



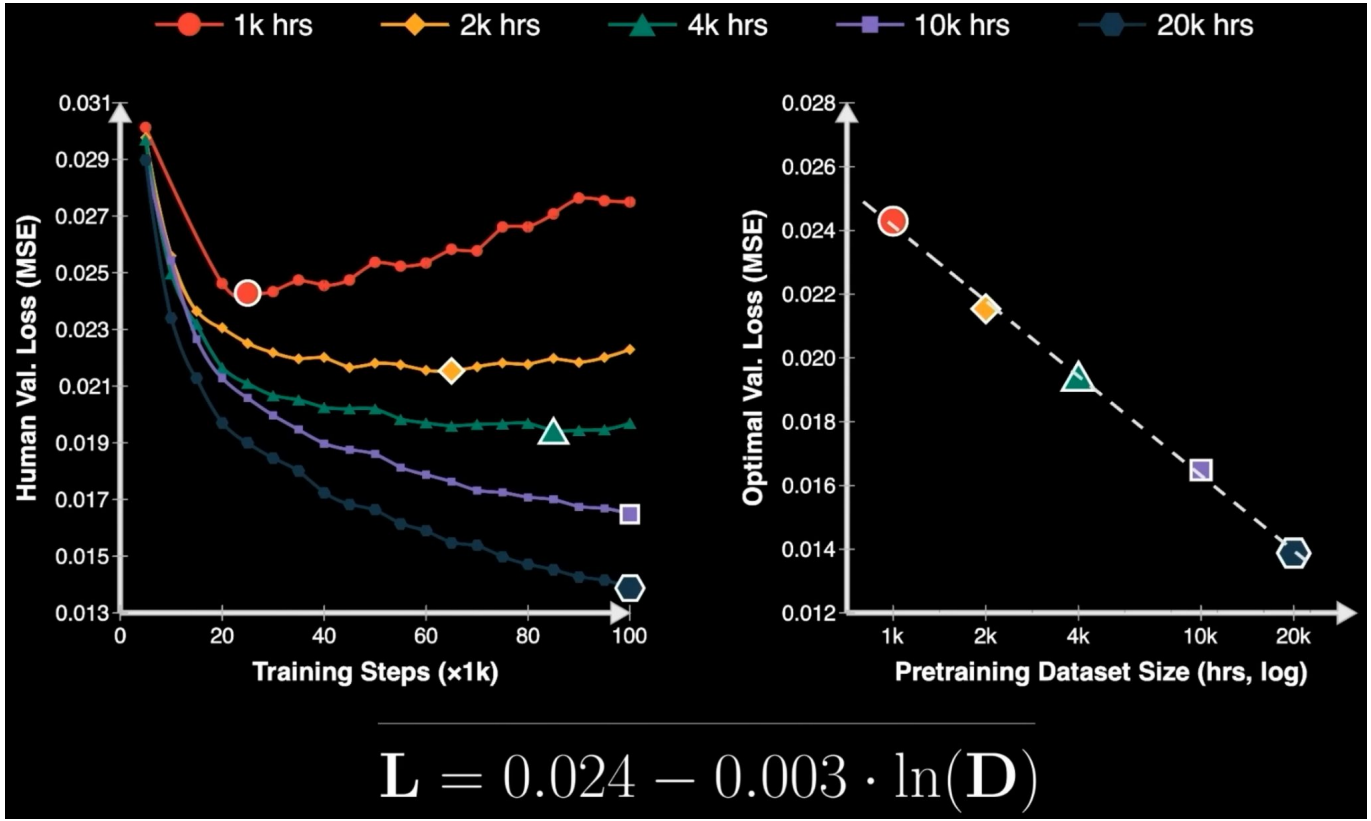
What data to use at aligning phase?



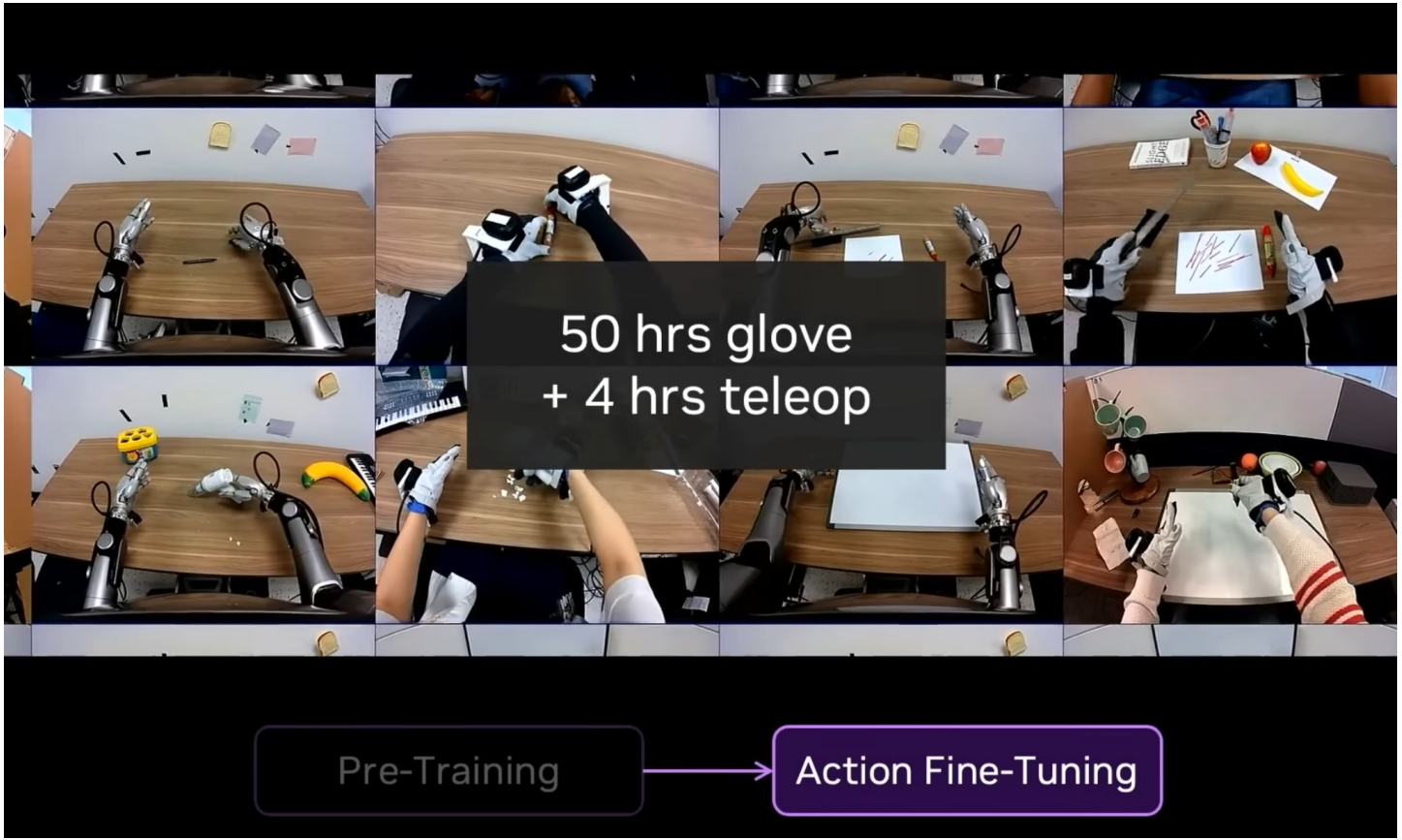
EgoScale Pre-Training dataset



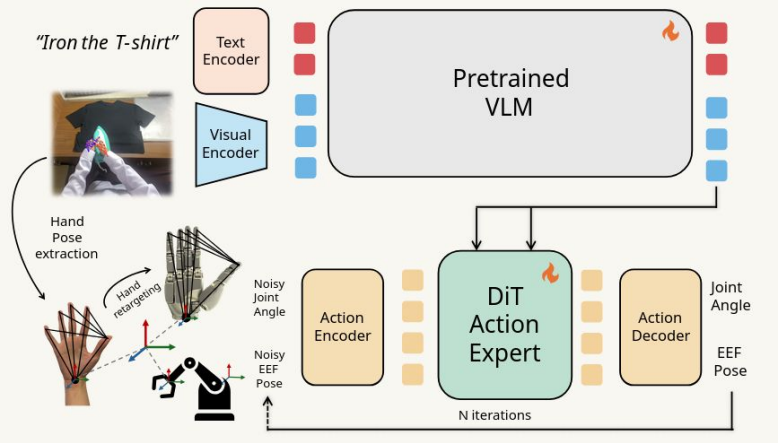
EgoScale scaling law



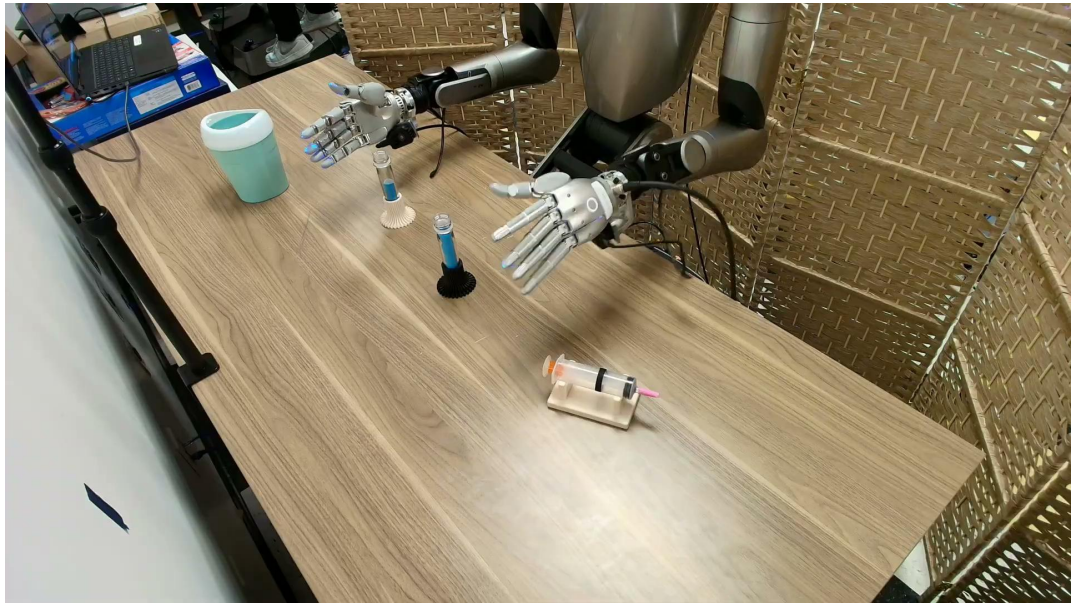
EgoScale Fine-Tuning dataset



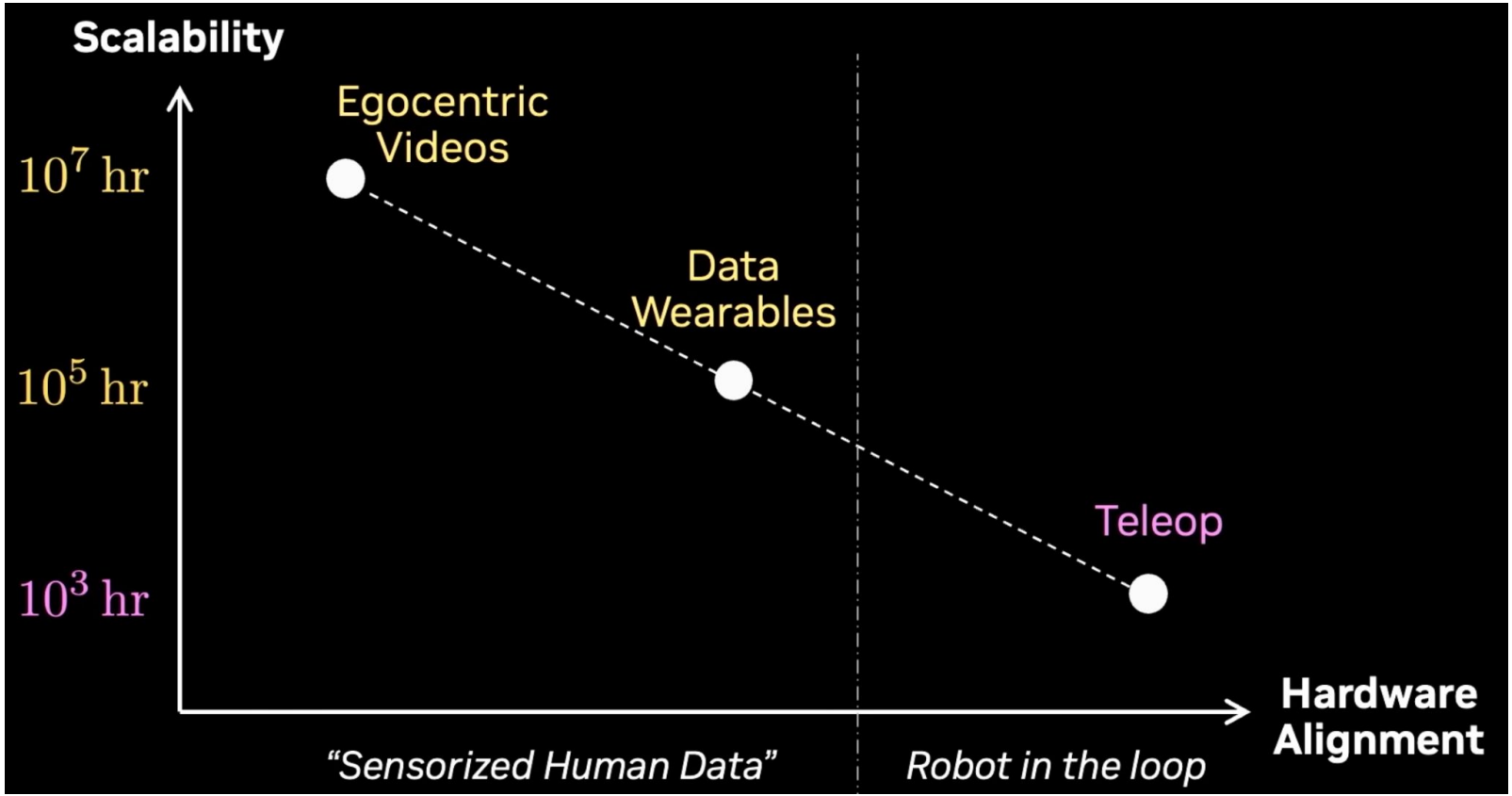
EgoScale. Trained model performance



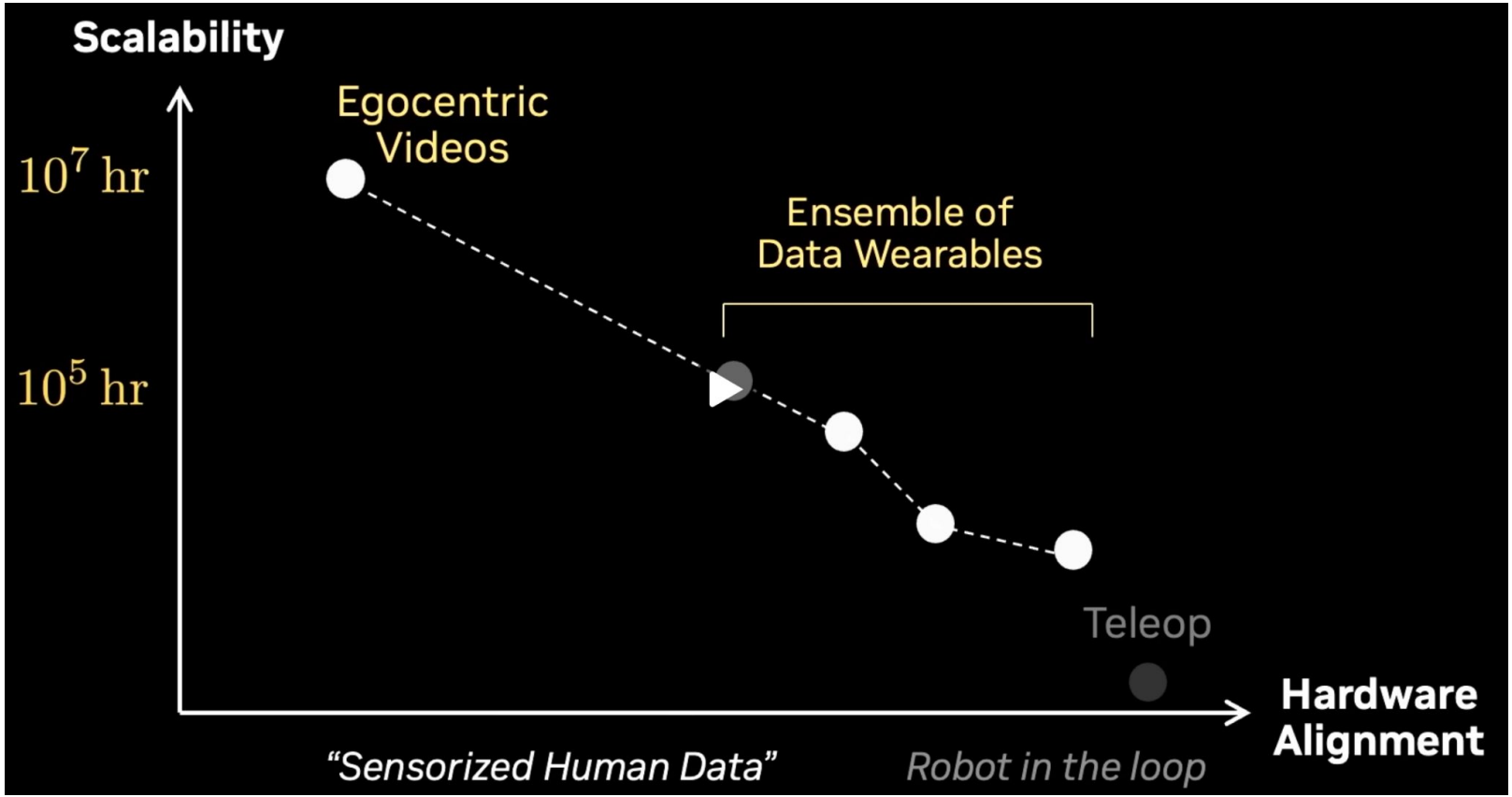
flow-based VLA architecture
similar to GR00T N1



Robotics' End Game



Robotics' End Game



Robotics' End Game

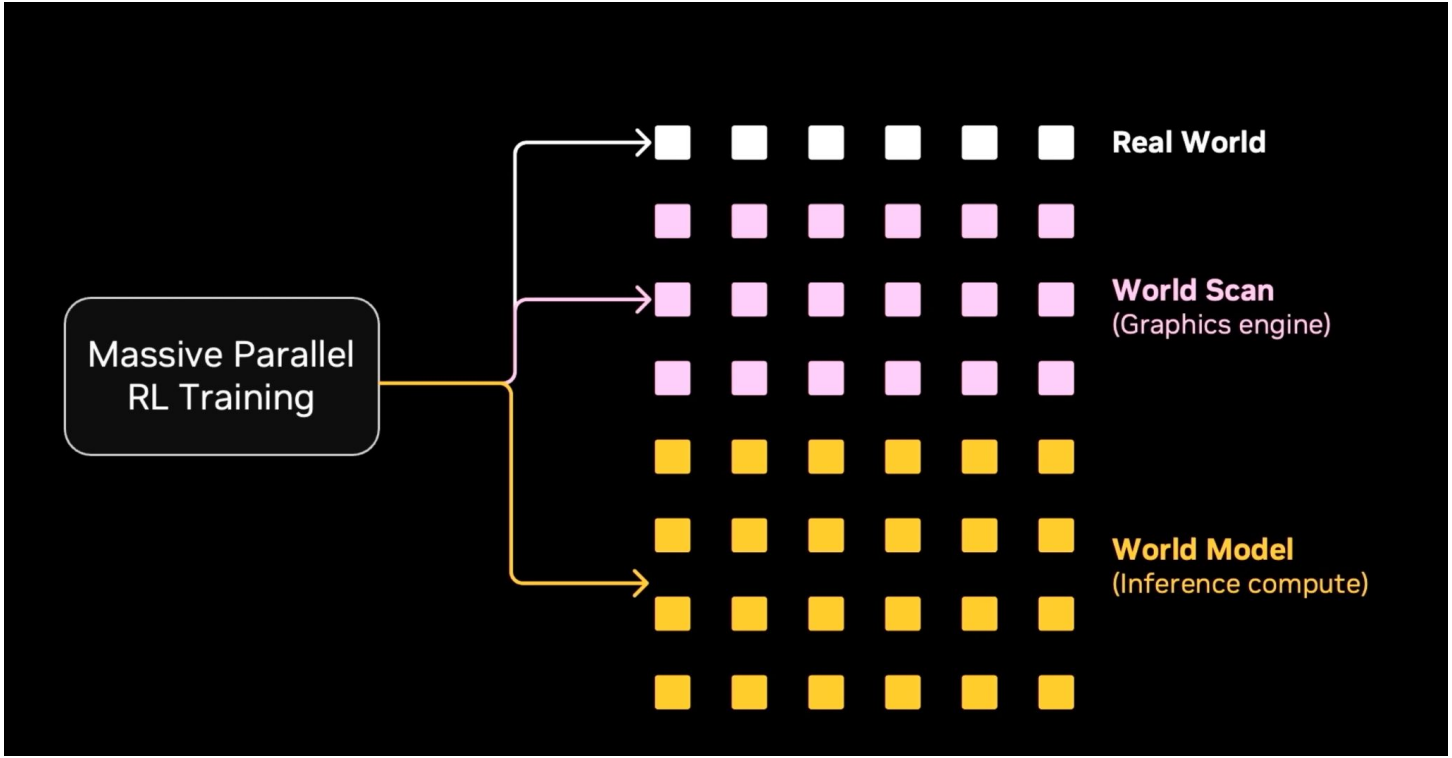
1st prediction



2nd prediction



New Post Training Paradigm
(surpassing human-level performance)

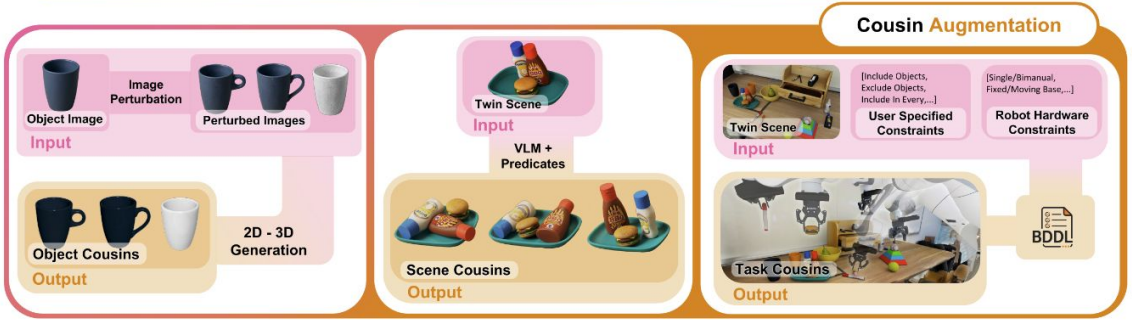


SimFoundry: Modular and Automated Scene Generation for Policy Learning and Evaluation



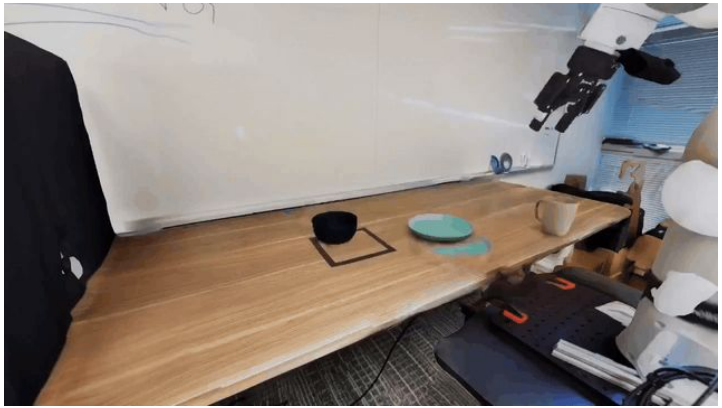




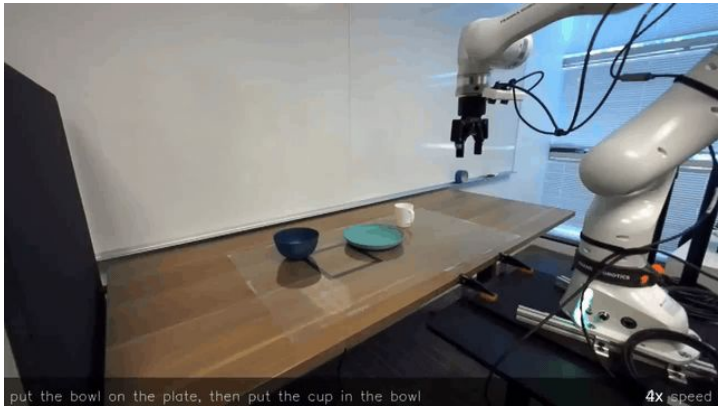


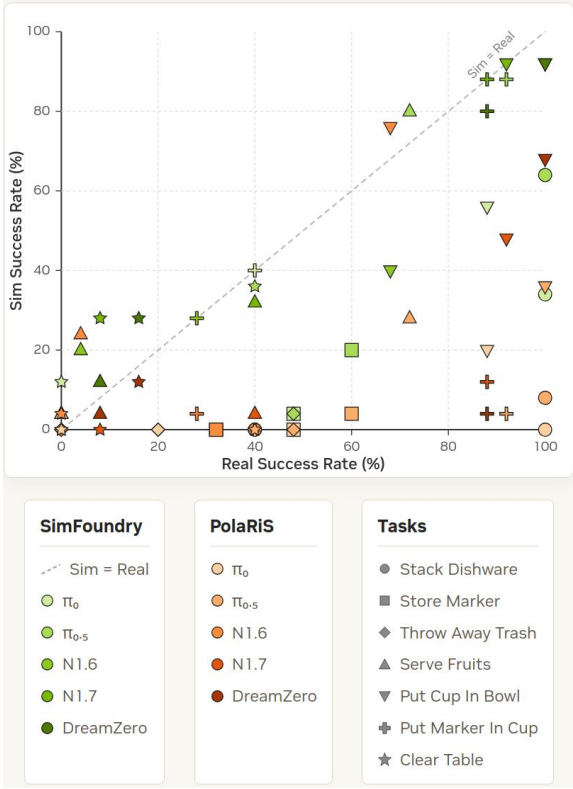
SimFoundry: Modular and Automated Scene Generation for Policy Learning and Evaluation

Sim



Real

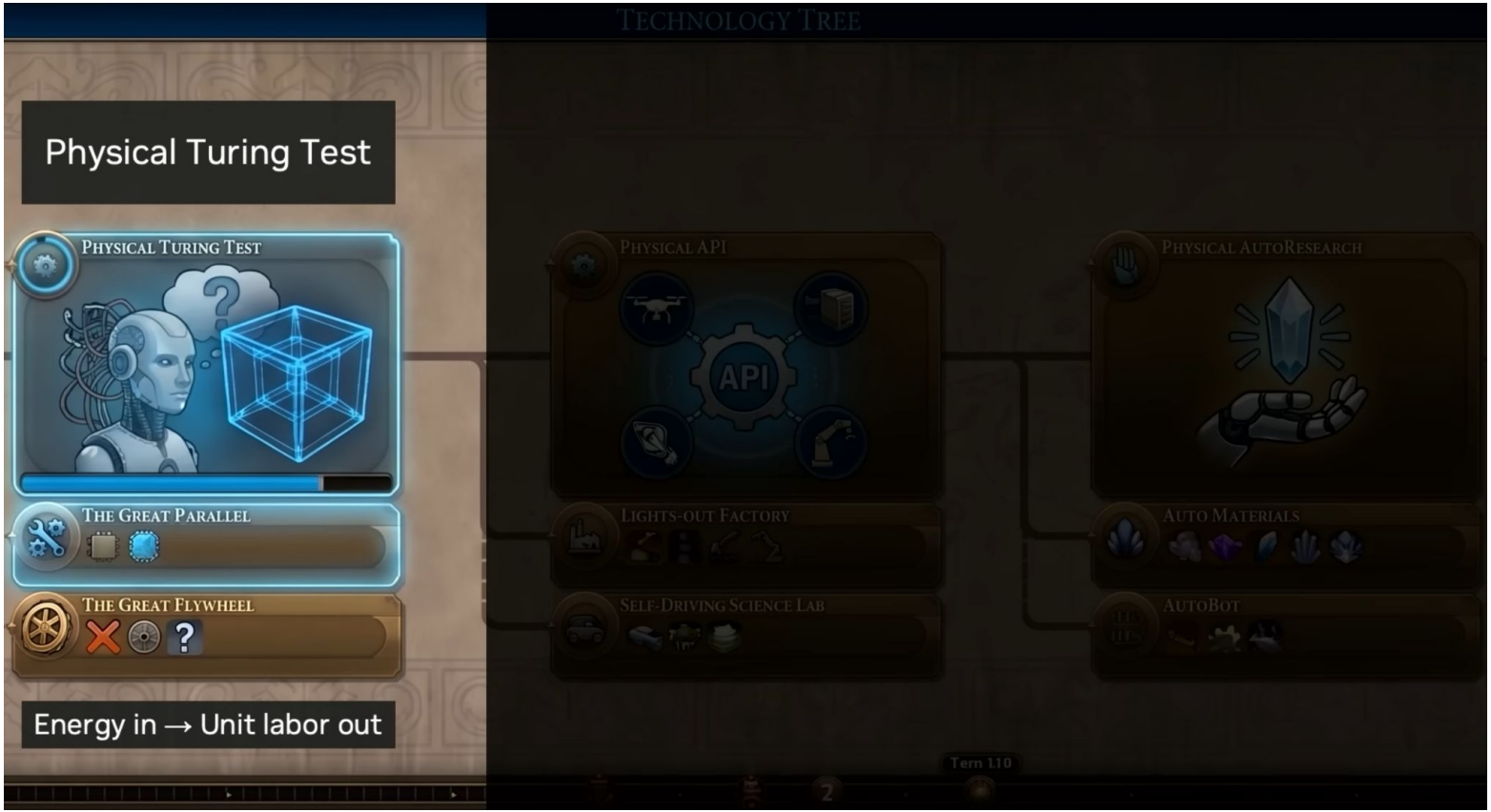




Robotics' End Game



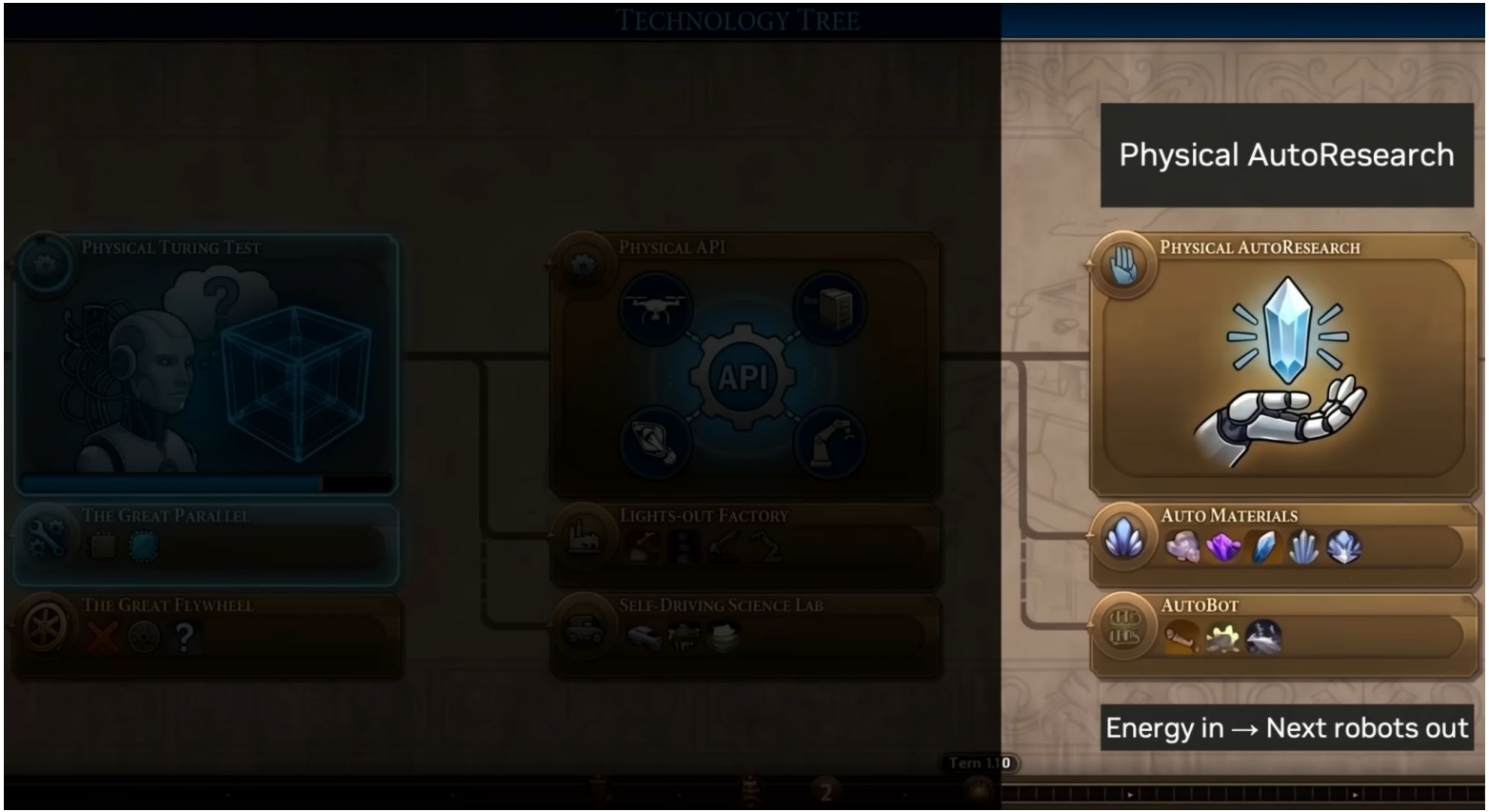
Robotics' End Game



Robotics' End Game



Robotics' End Game



Robotics' End Game

