Will data solve robotics?
Foundation models / vision-language action models for robotics

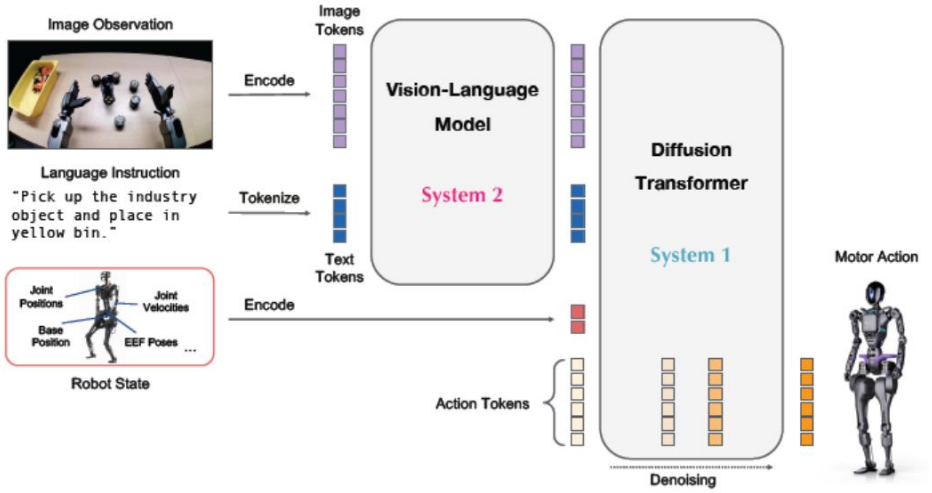


Figure 1: **Data Pyramid for Robot Foundation Model Training.** GR00T N1's heterogeneous training corpora can be represented as a pyramid: data quantity decreases, and embodiment-specificity increases, moving from the bottom to the top.

doc. Matěj Hoffmann, Ph.D.
Department of Cybernetics, Faculty of Electrical Engineering
Czech Technical University in Prague
matej.hoffmann@fel.cvut.cz
https://sites.google.com/site/matejhof
https://cyber.felk.cvut.cz/research/groups-teams/humanoids/

Outline

- Motivation and context – Robotics & AI
- History and evolution of key building blocks
- VLA Architectures
- Data collection and datasets
- Embodiment and cross-embodiment

Motivation – why would we want data to solve robotics?

The world is too varied and too messy to hand-engineer one robot behavior at a time.

Data is attractive because it lets robotics reuse the scaling logic that worked in vision and language: learn broad priors first, then adapt cheaply to new tasks, scenes, and embodiments.

1. Long-tail variation
Objects, layouts, occlusion, lighting, and human behavior vary too much to script exhaustively.

2. Semantics from language
Data links pixels to task intent: not just where the mug is, but what "put it in the sink" means.

3. Reuse across tasks
One pretrained policy or representation can be post-trained faster than building each skill from scratch.

4. Less reward engineering
Demonstrations and logs often scale more easily than hand-designed objectives for every manipulation task.

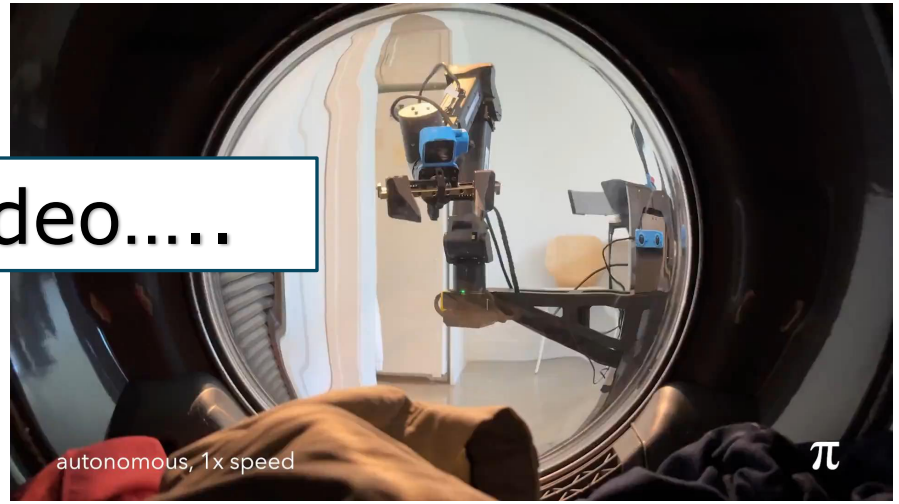
5. Online improvement
If robots collect more trajectories in the field, the system can keep improving instead of freezing at shipment.

What can robot foundation models do?

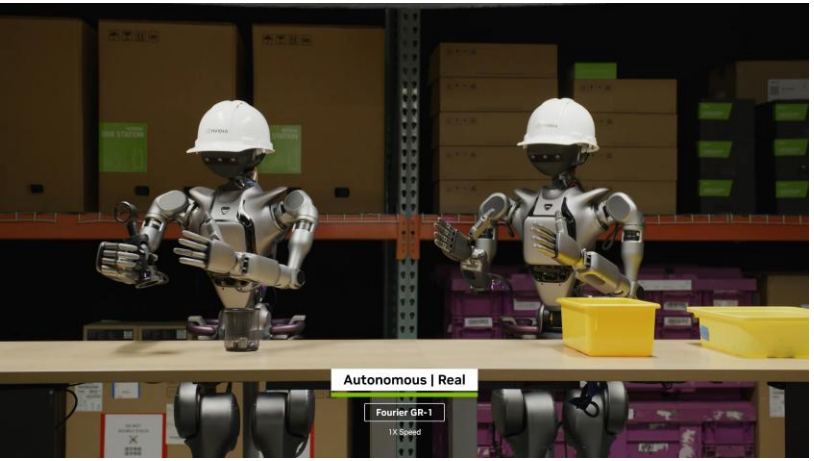
RT-1 [Google Research, 2022]

AutoRT [Google DeepMind, 2024]

Or, at least, "do" on a video…..



[Physical Intelligence (π), 2024]



GR00T N1 [NVIDIA, 2025]

Helix [Figure, 2025]

[Physical Intelligence (π*0.6), 2025]

How to read robot demo videos critically

Questions to ask immediately

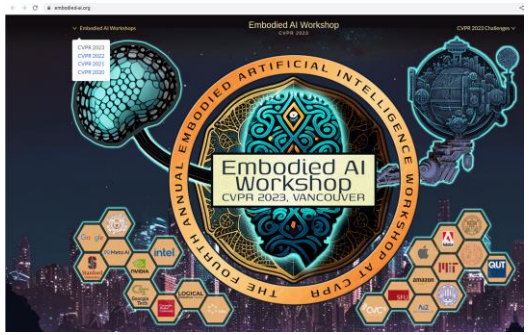
- Was the clip edited, sped up, or heavily choreographed?
- Was the task solved on-board, off-board, or partly by teleoperation?
- How many resets, retries, or human interventions were hidden?
- Can the robot recover after a slip, miss, or perception error?
- Is the task benchmarked, deployed, or only shown once?
- Are the success criteria clear and measurable?

See also TRI LBM Team (2025). A careful examination of large behavior models for multitask dexterous manipulation. https://arxiv.org/abs/2507.05331

Now, what's that thing called Embodied AI / Physical AI?

IROS KYOTO 2022

~Embodied AI for a Symbiotic Society~



Retrospectives on the Embodied AI Workshop

Matt Deitke[1,17], Dhruv Batra[3,6], Yonatan Bisk[3], Tommaso Campari[4,16], Angel X. Chang[13], Devendra Singh Chaplot[6], Changan Chen[19], Claudia Pérez-D'Arpino[9], Kiana Ehsani[1], Ali Farhadi[2,17], Li Fei-Fei[14], Anthony Francis[6], Chuang Gan[11,15], Kristen Grauman[19,8], David Hall[20], Winson Han[1], Unnat Jain[8], Aniruddha Kembhavi[1,17], Jacob Krantz[12], Stefan Lee[12], Chengshu Li[14], Sagnik Majumder[19], Oleksandr Maksymets[6], Roberto Martín-Martín[19], Roozbeh Mottaghi[8,17], Sonia Raychaudhuri[13], Mike Roberts[7], Silvio Savarese[14], Manolis Savva[15], Mohit Shridhar[17], Niko Sünderhauf[20], Andrew Szot[5], Ben Talbot[20], Joshua B. Tenenbaum[10], Jesse Thomason[16], Alexander Toshev[2], Joanne Truong[3], Luca Weihs[1], Jiajun Wu[14]
[1]Allen Institute for AI, [2]Apple, [3]Carnegie Mellon University, [4]FBK, [5]Georgia Tech, [6]Google, [7]Intel Labs, [8]Meta AI, [9]NVIDIA, [10]MIT, [11]MIT-IBM Watson AI Lab, [12]Oregon State University, [13]Simon Fraser University, [14]Stanford University, [15]UMass Amherst, [16]University of Padova, [17]University of Washington, [18]University of Southern California, [19]UT Austin, [20]QUT Centre for Robotics

Abstract

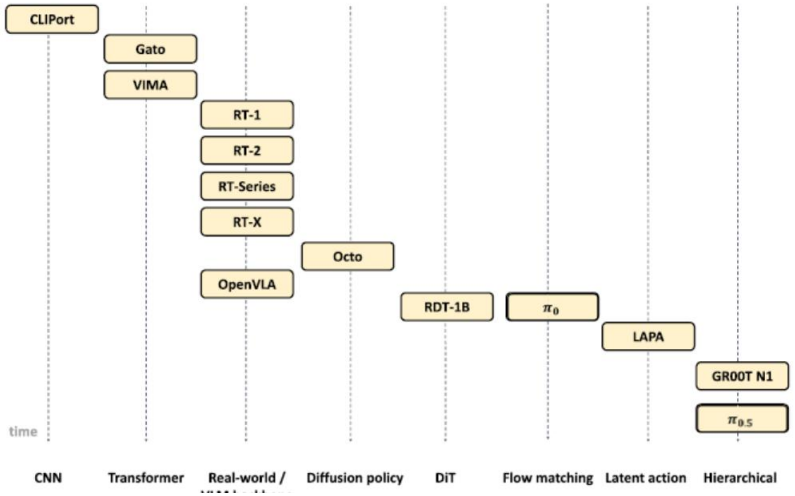
We present a retrospective on the state of Embodied AI research. Our analysis focuses on 13 challenges presented at the Embodied AI Workshop at CVPR. These challenges are grouped into three themes: (1) visual navigation, (2) rearrangement, and (3) embodied vision-and-language. We discuss the dominant datasets within each theme, evaluation metrics for the challenges, and the performance of state-of-the-art models. We highlight commonalities between top approaches to the challenges and identify potential future directions for Embodied AI research.

1. Introduction

Within the last decade, advances in deep learning, coupled with the creation of massive datasets and high-capacity models, have resulted in remarkable progress in computer vision, audio, NLP, and the broader field of AI. This progress has enabled models to obtain superhuman performance on a wide of researchers and research challenges.

Consider asking a robot to *'Clean my room'* or *'Drive me to my favorite restaurant'*. To succeed at these tasks in the real world, the robots need skills like *visual perception* (to recognize scenes and objects), *audio perception* (to receive the speech spoken by the human), *language understanding* (to translate questions and instructions into actions), *memory* (to recall how items should be arranged or to recall previously encountered situations), *physical intuition* (to understand how to interact with other objects), *multi-agent reasoning* (to predict and interact with other agents), and *navigation* (to safely move through the environment). The study of embodied agents both provides a challenging testbed for building intelligent systems and tries to understand how intelligence emerges through interaction with an environment. As such, it involves many disciplines, such as computer vision, natural language processing, acoustic learning, reinforcement learning, developmental psychology, cognitive science, neuroscience, and robotics.

https://embodied-ai.org/



Vision-Language-Action Models for Robotics: A Review Towards Real-World Applications

KENTO KAWAHARAZUKA[1], (Member, IEEE), JIHOON OH[1], JUN YAMADA[2], (Graduate Student Member, IEEE), INGMAR POSNER[2], (Member, IEEE), AND YUKE ZHU[3], (Senior Member, IEEE)
[1]Department of Mechano-Informatics, The University of Tokyo, Tokyo 113-8656, Japan

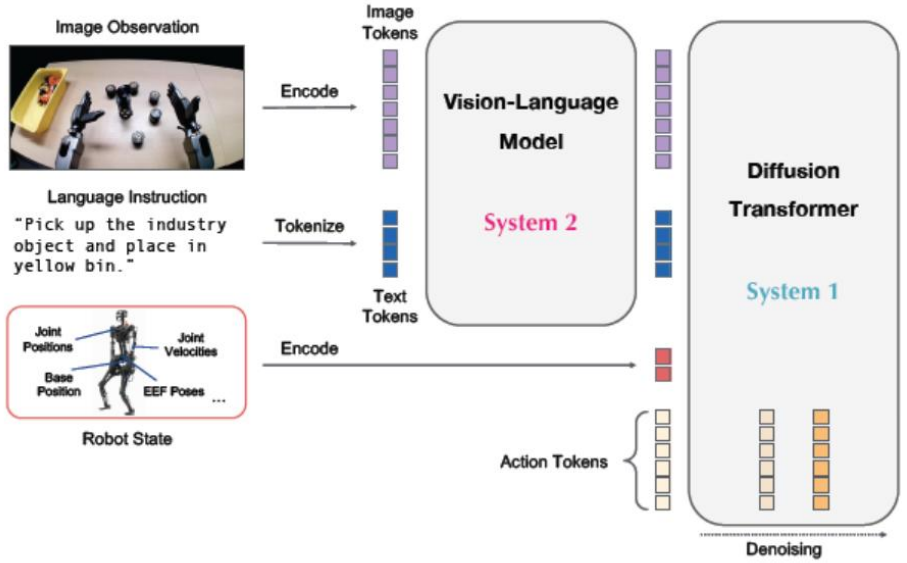


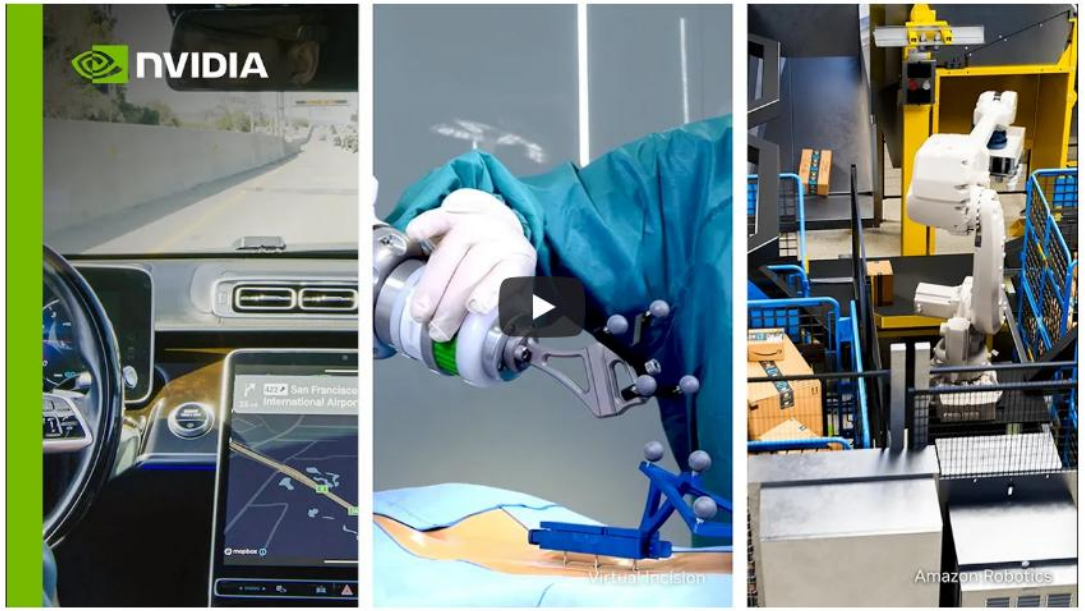
Figure 1: **Data Pyramid for Robot Foundation Model Training.** GR00T N1's heterogeneous training corpora can be represented as a pyramid: data quantity decreases, and embodiment-specificity increases, moving from the bottom to the top.

Gr00t n1: An open foundation model for generalist humanoid robots. 2025. https://arxiv.org/abs/2503.14734

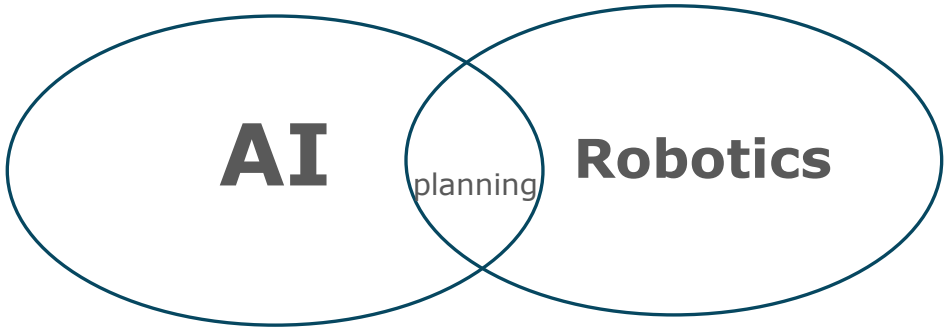
Embodied AI / Physical AI

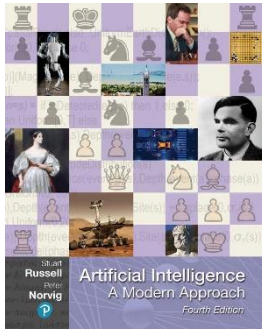
- Large-scale, so-called foundation models, in Natural Language Processing (NLP) and computer vision can enable capable AI systems by providing general-purpose pretrained models which can often outperform their narrowly targeted counterparts trained on smaller but more task-specific data.
- Applying the same strategy to control robots is appealing.
- The community originating in computer vision, machine learning, and NLP, but now connecting their models to robots uses the label “Embodied AI” (e.g., (Deitke et al. 2022; Liu et al. 2024; Vanhoucke 2024)).
- The idea is to leverage the reasoning (“common sense”) capabilities of such models as well as their capability to understand visual scenes (images) in order to produce plans that a robot can execute.

Physical AI lets autonomous systems like cameras, robots, and self-driving cars perceive, understand, reason, and perform or orchestrate complex actions in the physical world.



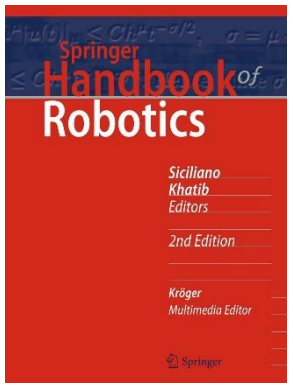
Physical AI = Robotics?





- until ~2020, AI and robotics have developed quite independently
- meeting when complex high-level plans and hence reasoning were required

Russell, S. J., & Norvig, P. (2021). *Artificial intelligence: a modern approach*.

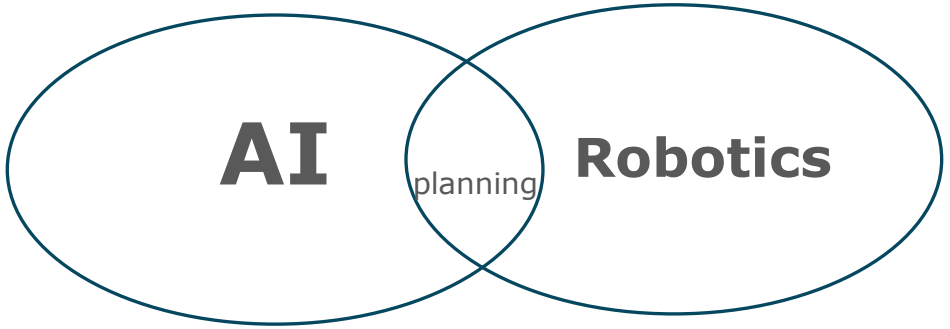


One out of 80 chapters.

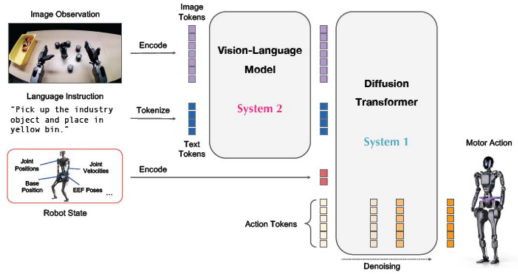
Khatib, O., & Siciliano, B. (Eds.). (2016). Springer handbook of robotics.

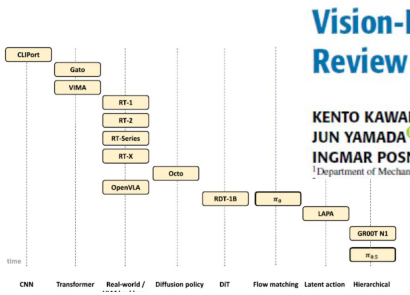
Physical AI = Robotics?

Up until recently.



Hot topic now.





Vision-Language-Action Models for Robotics: A Review Towards Real-World Applications

KENTO KAWAHARAZUKA[1], (Member, IEEE), JIHOON OH[1], JUN YAMADA[2], (Graduate Student Member, IEEE), INGMAR POSNER[2], (Member, IEEE), AND YUKE ZHU[3], (Senior Member, IEEE)
[1]Department of Mechano-Informatics, The University of Tokyo, Tokyo 113-8656, Japan

Figure 1: **Data Pyramid for Robot Foundation Model Training.** GR00T N1's heterogeneous training corpora can be represented as a pyramid: data quantity decreases, and embodiment-specificity increases, moving from the bottom to the top.

Will data solve robotics?

Gr00t n1: An open foundation model for generalist humanoid robots. 2025. https://arxiv.org/abs/2503.14734

ARTIFICIAL INTELLIGENCE

"Data will solve robotics and automation: True or false?": A debate

Nancy M. Amato[1], Seth Hutchinson[2], Animesh Garg[3], Aude Billard[4], Daniela Rus[5], Russ Tedrake[5], Frank Park[6], Ken Goldberg[7]*

Leading researchers debate the long-term influence of model-free methods that use large sets of demonstration data to train numerical generative models to control robots.

The debaters. Top row (left to right): Nancy M. Amato, Seth Hutchinson, and Ken Goldberg. Bottom row (left to right): Animesh Garg, Aude Billard, Russ Tedrake, and Frank Park.

debate at the IEEE International Conference on Robotics and Automation (ICRA), the largest research conference for the field, where many presentations celebrated algorithmic ("model-based") breakthroughs over the decades. However, since 2012, advances in deep "neural" networks that combine unprecedented amounts of example data, advances in stochastic gradient descent techniques, and advances in computing, in particular GPUs (graphics processing units), have produced remarkable results in computer vision and speech recognition. And with the emergence of ChatGPT and associated large vision language models (VLMs) in 2022, with associated advances in natural language processing and denoising diffusion methods, the paradigm of "end-to-end" (also known as "model-free") approaches to perception and control suggest an entirely new, data-driven approach to robotics and automation.

tion). K.G. formulated a catchy title and short description: "Will the future of robotics and automation be written in code or in data? Is the *Handbook of Robotics* obsolete? Are home humanoids overhyped? Join us for a high-voltage debate about physics vs pixels, theory vs terabytes."

This debate ("Data Will Solve Robotics and Automation: True or False?") was held at the 2025 IEEE International Conference on Robotics and Automation.
A video of the debate can be found at https://youtu.be/PfvctjoMPk8?si=Zwpt-QfJ3EiqferR.

TRUE

FALSE

SCIENCE ROBOTICS | VIEWPOINT

ARTIFICIAL INTELLIGENCE

"Data will solve robotics and automation: True or false?": A debate

Nancy M. Amato[1], Seth Hutchinson[2], Animesh Garg[3], Aude Billard[4], Daniela Rus[5], Russ Tedrake[5], Frank Park[6], Ken Goldberg[7]*

Leading researchers debate the long-term influence of model-free methods that use large sets of demonstration data to train numerical generative models to control robots.

The debaters. Top row (left to right): Nancy M. Amato, Seth Hutchinson, and Ken Goldberg. Bottom row (left to right): Animesh Garg, Aude Billard, Russ Tedrake, and Frank Park.

1) General-purpose intelligence, larly in physical robots, is ambiguo ended, and underspecified. Data offer models. Robotics has long drawn strength from first principles. Physics gives us elegant models for simple tasks: g ject, stabilizing a robot ar collision-free path. When structured, the material goals well defined, analyti precision and insight. The Why are data so fundame cess? Large-scale pretrainin multitask data is the best (o know to program "common s gets this from reading all of internet, and our robots are physical common sense from seeing enough demonstrations of manipulation. Most people agree that large models are useful for manipulation tasks that require common-sense language understanding (e.g., "pick up the extinct animal"), but many people underestimate how essential common sense is for low-level control.

As I noted above, I also believe that mathematical models and first principles still matter. They help structure the learning process, provide constraints, and enforce physical con- cannot capture defines the real general-purpose equations—it is e robotics and is that combine richness of data. brittle to robust, m lab demos to

There is no doubt that multimodal foundation models represent a transformational moment in human history, but expecting a parallel revolution in robotics is premature at best and wishful thinking at worst. Robotics is a different beast entirely: Real-world data are scarce, simulations remain unreliable, and robots in their many embodiments must contend with an endless variety of environments and tasks. Real-world performance

This debate ("Data Will Solve Robotics and Automation: ational Conference on Robotics and Automation.
A video of the debate can be found at https://youtu.be/

Outline

- Motivation and context – Robotics & AI
- History and evolution of key building blocks
- VLA Architectures
- Data collection and datasets
- Embodiment and cross-embodiment

History – take 1 (Vanhoucke, up until 2023)

The State of Robot Learning

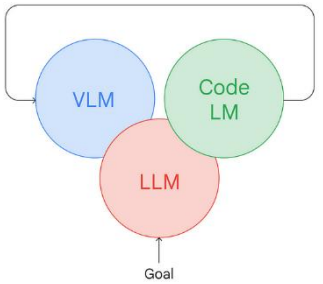
A partially observed, semi-stochastic, egocentric view.

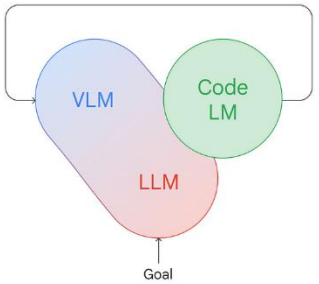
Vincent Vanhoucke · Follow

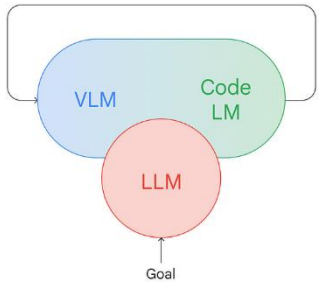
11 min read · Mar 7, 2024

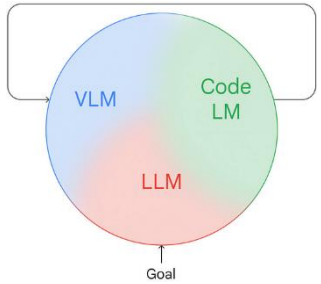
356

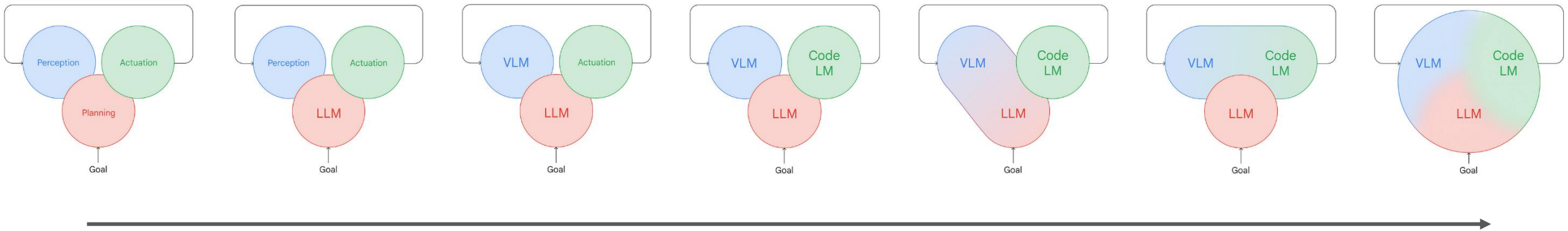
https://vanhoucke.medium.com/the-state-of-robot-learning-639dafffbcf8

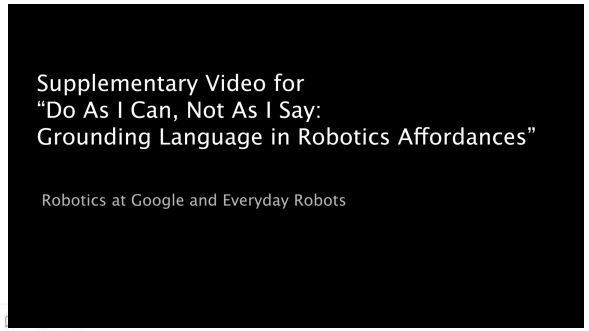






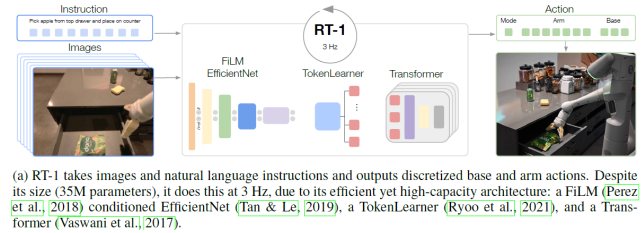






“Say-Can” (Ahn et al. 2022)

ChatGPT for Robotics (Vemprala et al. 2024)



(a) RT-1 takes images and natural language instructions and outputs discretized base and arm actions. Despite its size (35M parameters), it does this at 3 Hz, due to its efficient yet high-capacity architecture: a FiLM (Perez et al., 2018) conditioned EfficientNet (Tan & Le, 2019), a TokenLearner (Ryoo et al., 2021), and a Transformer (Vaswani et al., 2017).

RT-1, DeepMind 2023

“Inner Monologue” (Huang et al. 2022)

PALM-E (Driess et al. 2023)

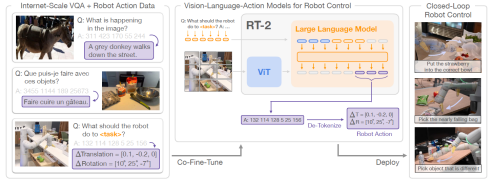
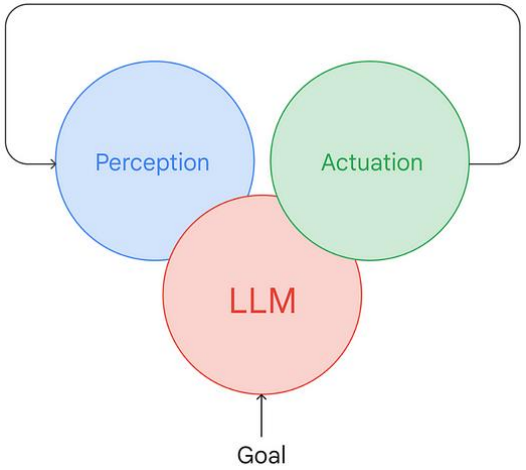


Figure 1 | RT-2 overview: we represent robot actions as another language, which can be cast into text tokens and trained together with Internet-scale vision-language datasets. During inference, the text tokens are de-tokenized into robot actions, enabling closed loop control. This allows us to leverage the backbone and pretraining of vision-language models in learning robotic policies, transferring some of their generalization, semantic understanding, and reasoning to robotic control. We demonstrate examples of RT-2 execution on the project website: robotics-transformer2.github.io.

RT-2, DeepMind 2023

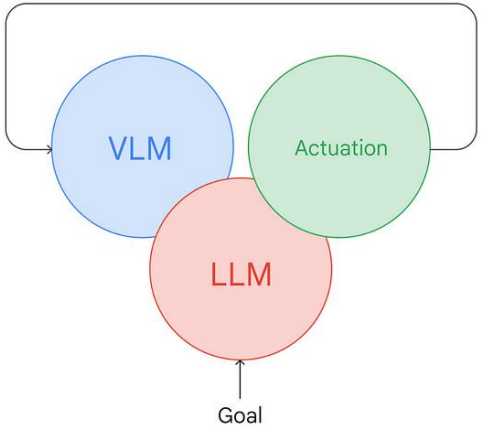
LLMs for planning

- The deployment of Large Language Models (LLMs) into robotics has naturally started from planning, which was lifted into "semantic space" from geometric space.
- An example is the "Say-Can" model (Ahn et al. 2022). The reasoning and planning capabilities traditionally handled by automatic inference systems in the GOFAI era now leverage the "common sense" power of LLMs.



https://say-can.github.io/

VLMs for perception



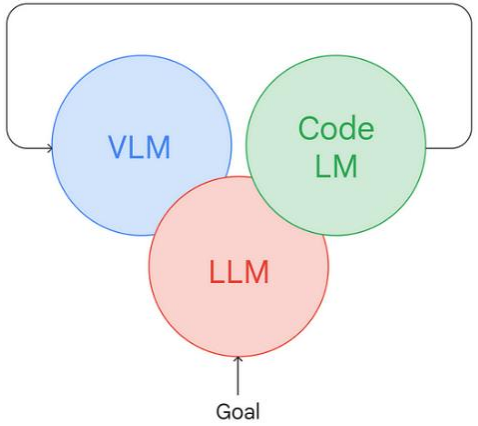
- The world state, now represented in language, had to be extracted from the world (perception or state estimation module). The next logical step thus was to deploy Visual Language Models (VLMs), which can convert images into language representations, as the Perception module (e.g., Socrates, (Zeng et al. 2022)).
- To keep the representation in sync with the real world (addressing the frame problem), periodic state reestimation and replanning is necessary, as exemplified in the "Inner Monologue" (Huang et al. 2022).

Inner Monologue:
Embodied Reasoning through Planning with Language Models

Robotics at Google

https://innermonologue.github.io/

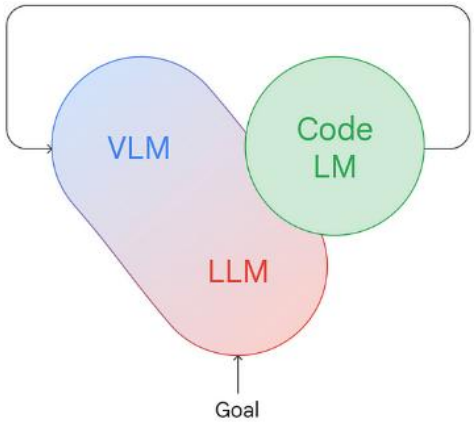
Code-LMs for action



- The final bastion of traditional robotics that was given LLM treatment (in the words of (Vanhoucke 2024)) was the action.
- This was achieved by having the language model generate code to be executed by the robot ("Code LM"), as in "Code as Policies" (Liang et al. 2023), Text2Motion (Lin et al. 2023), ProgPrompt (Singh et al. 2023), or ChatGPT for Robotics (Vemprala et al. 2024).

https://www.microsoft.com/en-us/research/articles/chatgpt-for-robotics/

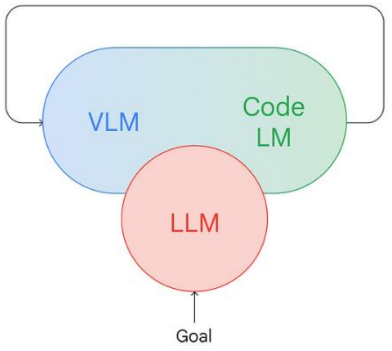
Blending it in…

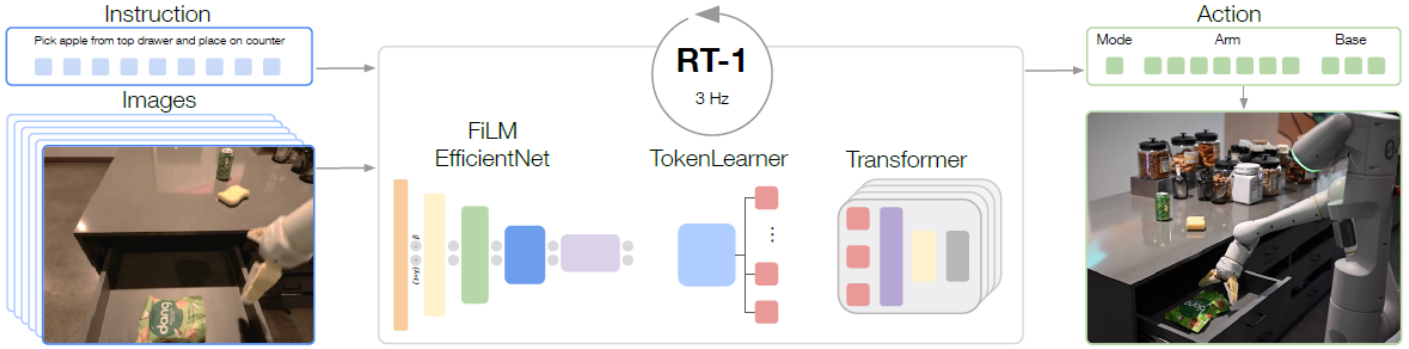


- The interfaces between the modules—LLM as a planner, VLM as state estimator, and Code LM as action generator—have become a bottleneck of these architectures. Therefore, the next step was to "blend these modules in". PALM-E (Driess et al. 2023) blended the LLM and VLM.

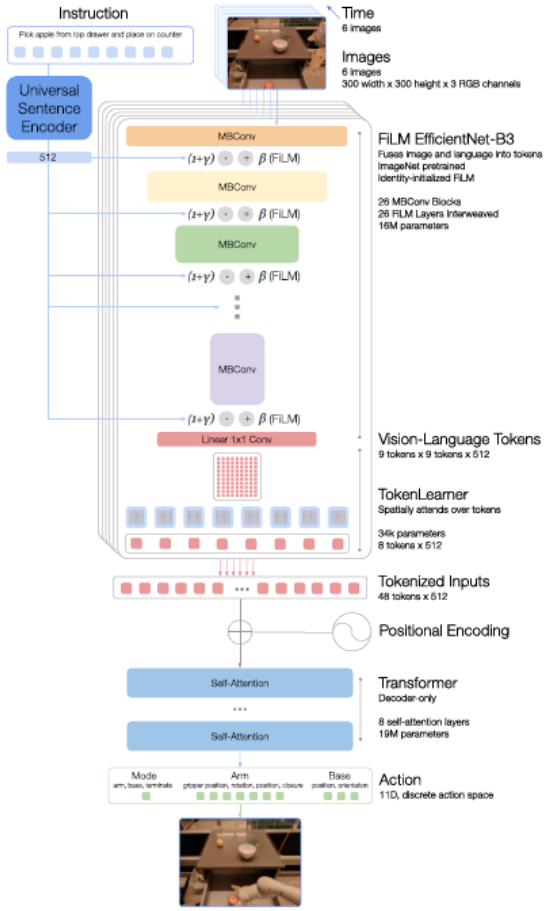
Blending it in...

- The interfaces between the modules—LLM as a planner, VLM as state estimator, and Code LM as action generator—have become a bottleneck of these architectures. Therefore, the next step was to “blend these modules in”.
- RT-1 (Brohan, Brown, Carbajal, Chebotar, Dabis, et al. 2023) blended the Perception and Actuation (VLM and Code LM).
  - Specialized for robots and capable of real-time execution on real-world tasks.
  - Image sequences go through EfficientNet, language features from USE provide FiLM conditioning, TokenLearner compresses tokens, and the model outputs actions.



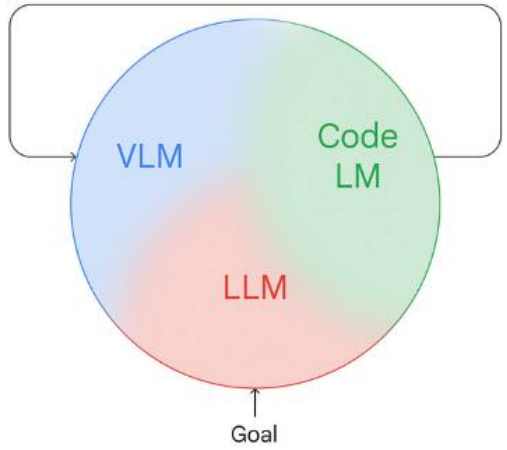


(a) RT-1 takes images and natural language instructions and outputs discretized base and arm actions. Despite its size (35M parameters), it does this at 3 Hz, due to its efficient yet high-capacity architecture: a FiLM (Perez et al., 2018) conditioned EfficientNet (Tan & Le, 2019), a TokenLearner (Ryoo et al., 2021), and a Transformer (Vaswani et al., 2017).



Blending it all in

- The final logical next step was to reason jointly about the entire problem, blending all the module boundaries while still leveraging "internet data". Examples of this approach are RT-2 (Brohan, Brown, Carbajal, Chebotar, Chen, et al. 2023) or VC-1 (Yokoyama et al. 2023).



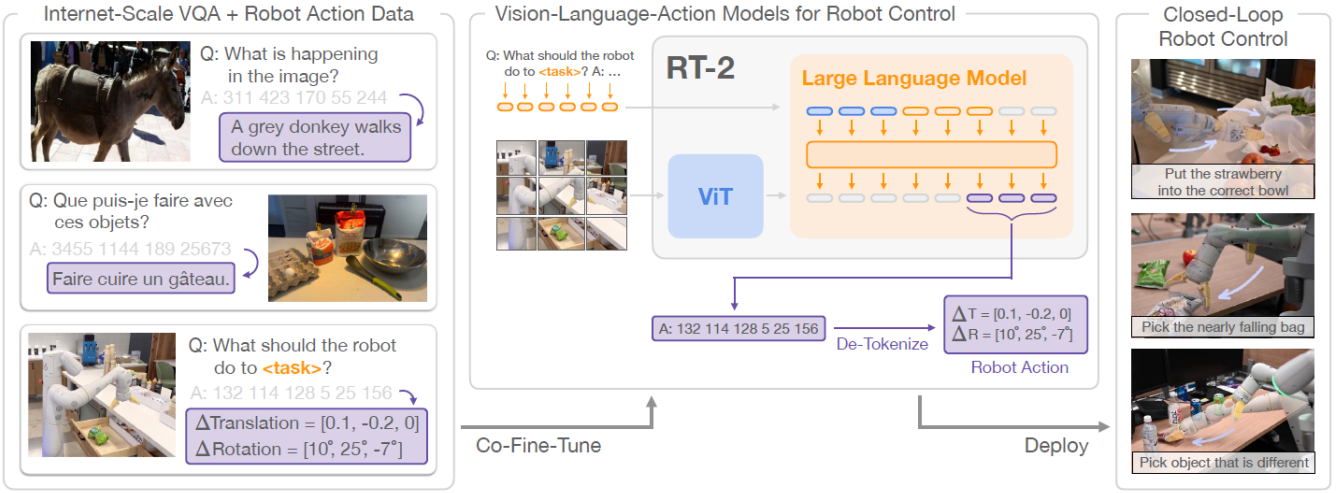


Figure 1 | RT-2 overview: we represent robot actions as another language, which can be cast into text tokens and trained together with Internet-scale vision-language datasets. During inference, the text tokens are de-tokenized into robot actions, enabling closed loop control. This allows us to leverage the backbone and pretraining of vision-language models in learning robotic policies, transferring some of their generalization, semantic understanding, and reasoning to robotic control. We demonstrate examples of RT-2 execution on the project website: robotics-transformer2.github.io.

RT-2

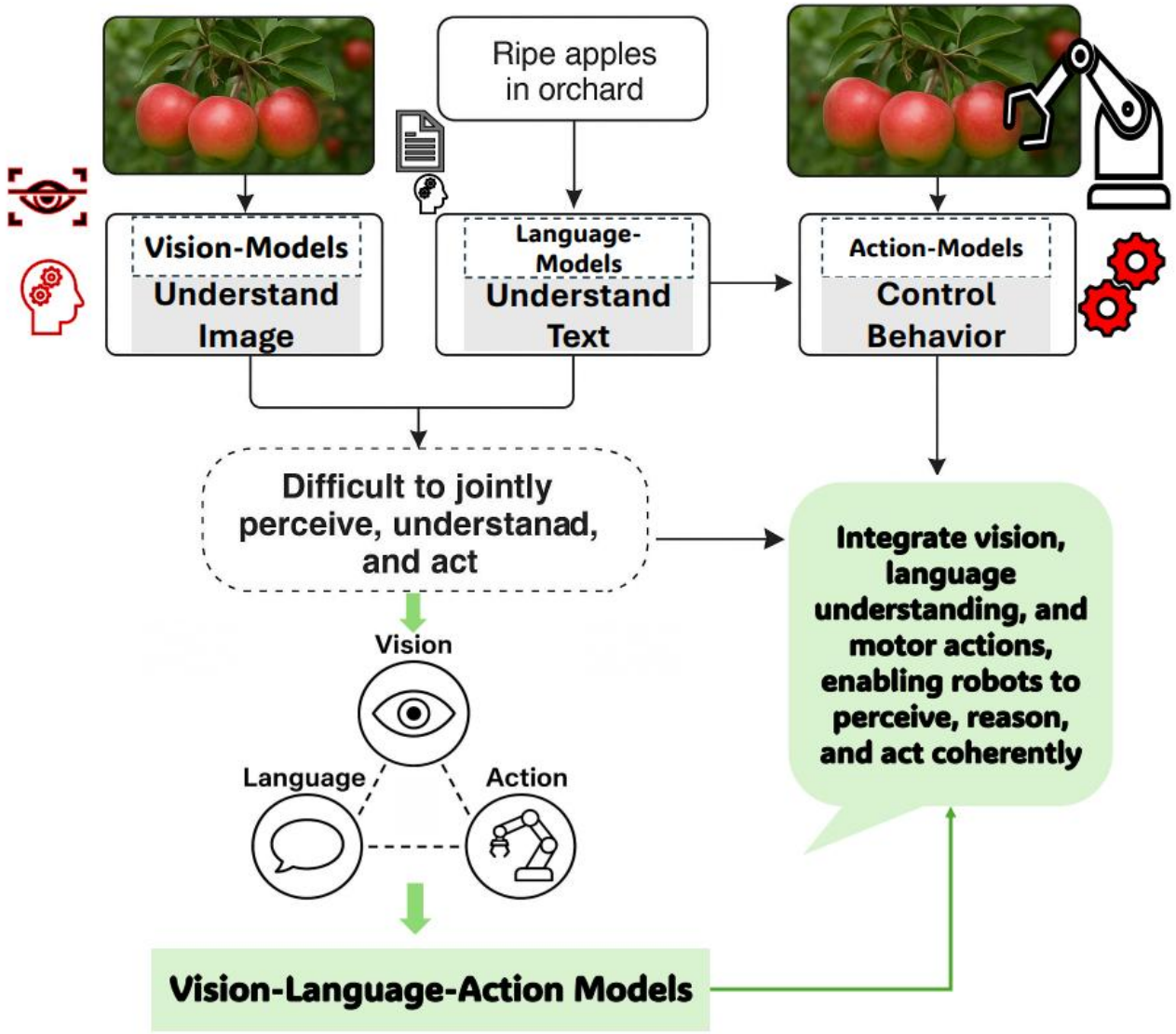
- Uses large web-pretrained VLM backbones - PaLM-E / PaLI-X - to improve generalization to unseen environments.
- Fine-tunes the VLM using both RT-1-style robot data and internet-scale vision-language tasks.
- This helped establish the now-common idea of a VLM-backed VLA.

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., ... & Zitkovich, B. (2022). Rt-1: Robotics transformer for real-world control at scale. arXiv preprint arXiv:2212.06817. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., ... & Zitkovich, B. (2023). Rt-2: Vision-language-action models transfer web knowledge to robotic control. arXiv preprint arXiv:2307.15818.

Until recently, we extracted information separately using large models for each modality.

This often leads to shallow image / language understanding and weak links to action.

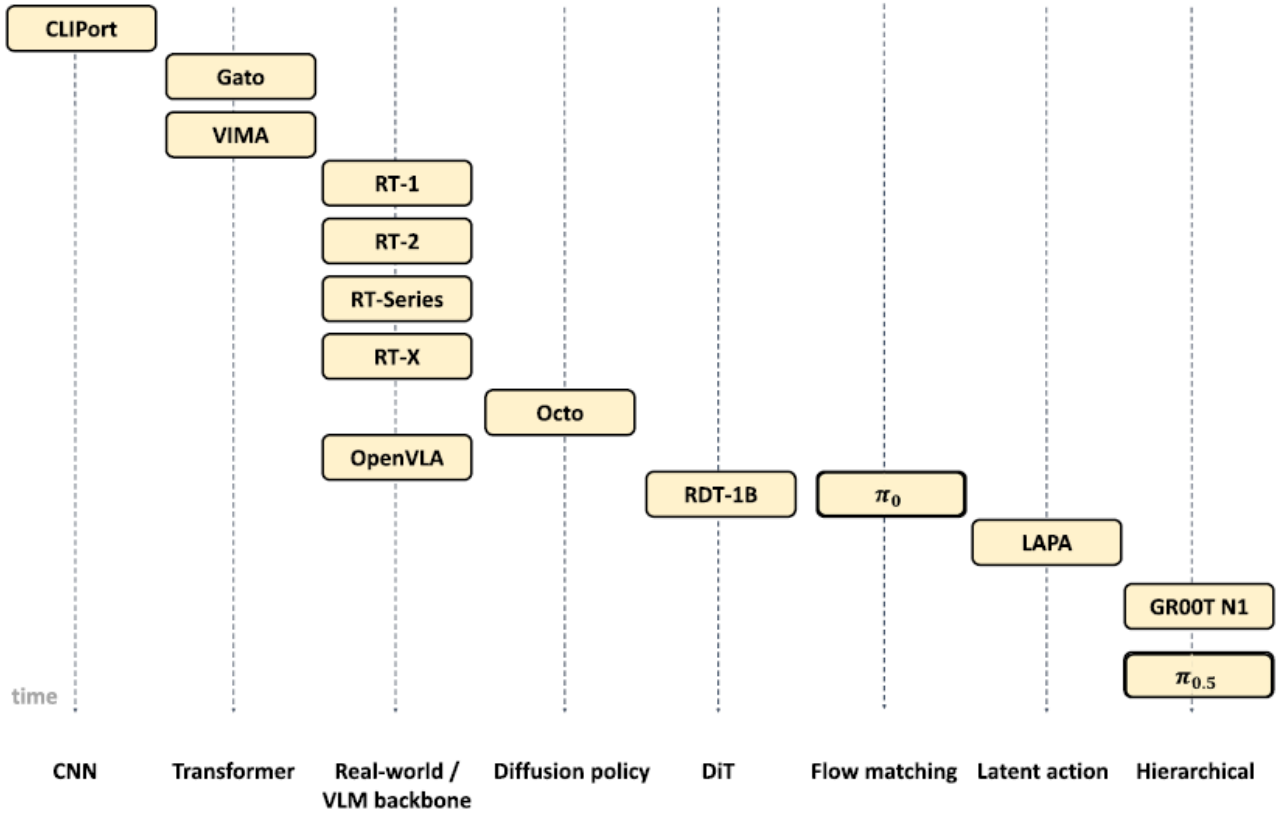
So we should learn image, language, and action jointly.



VLA Survey [R. Sapkota, arXiv, 2025]

Slide: courtesy Kento Kawaharazuka

History – take 2
Kento Kawaharazuka



Vision-Language-Action Models for Robotics: Review Towards Real-World Applications

KENTO KAWAHARAZUKA[1], (Member, IEEE), JIHOON OH[1], JUN YAMADA[2], (Graduate Student Member, IEEE), INGMAR POSNER[2], (Member, IEEE), AND YUKE ZHU[3], (Senior Member, IEEE)
[1]Department of Mechano-Informatics, The University of Tokyo, Tokyo 113-8656, Japan

Kawaharazuka, K., Oh, J., Yamada, J., Posner, I., & Zhu, Y. (2025). Vision-language-action models for robotics: A review towards real-world applications. *IEEE Access*. https://ieeexplore.ieee.org/abstract/document/11164279

CLIPort

[M. Shridhar, CoRL2021]

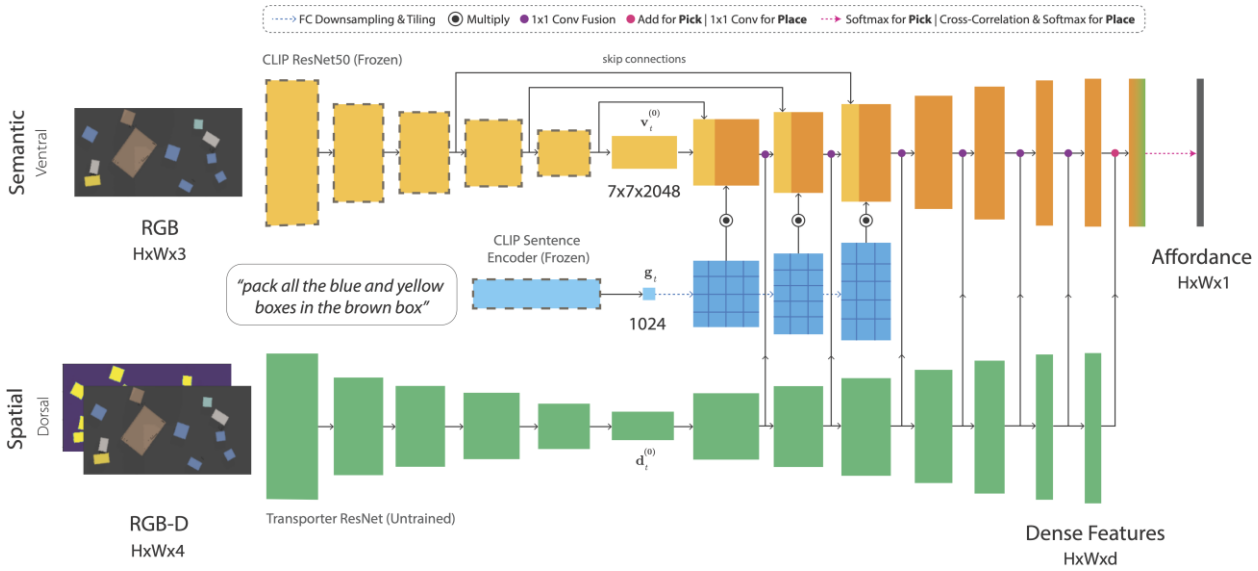
One of the earliest end-to-end VLAs.

Extracts language and visual information with CLIP.

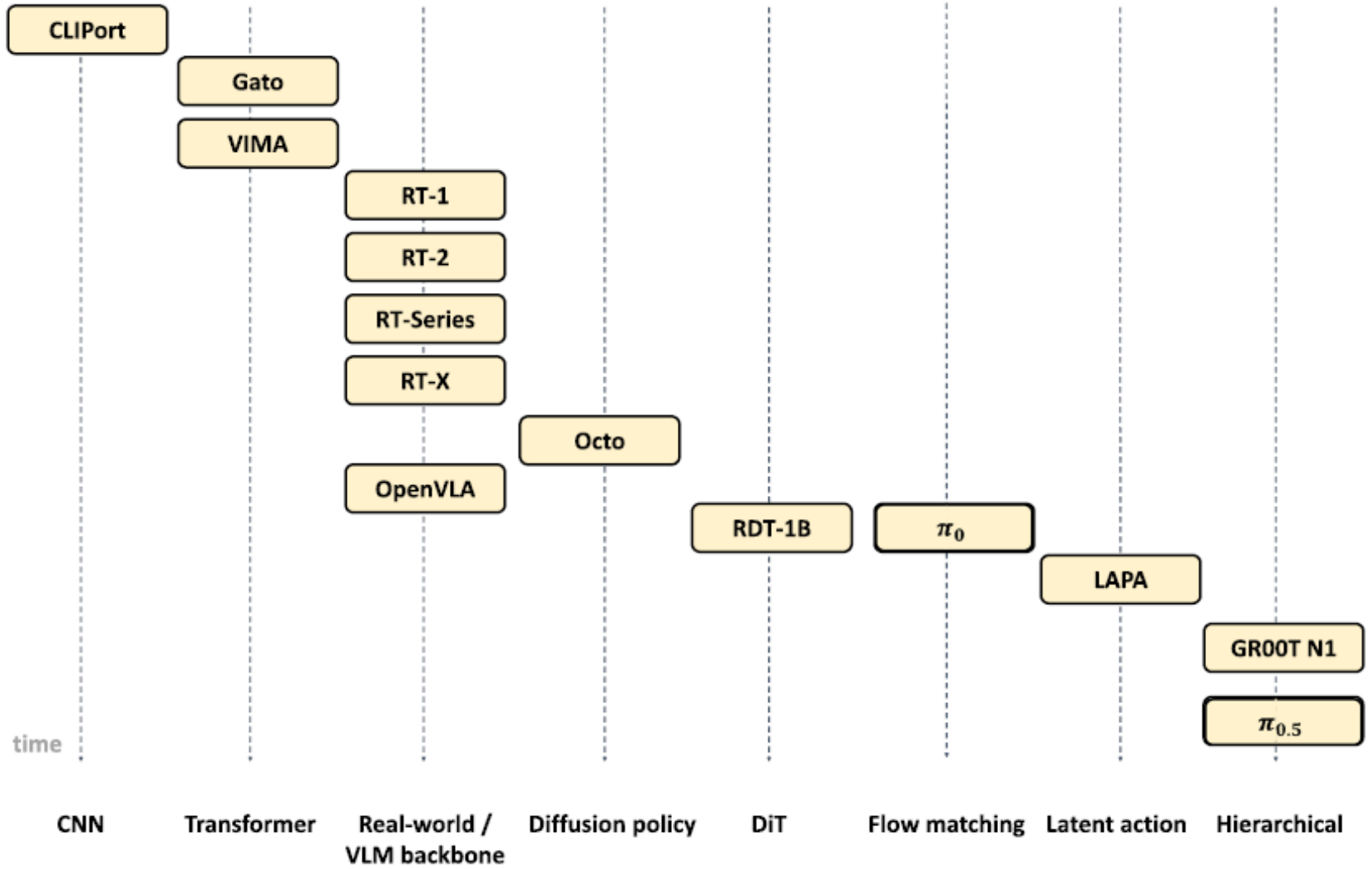
Combined with a Transporter Network specialized for pick-and-place.

Generates what object to place and where from RGB-D images and language.

Its CNN/MLP design limits modality handling and scalability.



Slide: courtesy Kento Kawaharazuka



Kawaharazuka, K., Oh, J., Yamada, J., Posner, I., & Zhu, Y. (2025). Vision-language-action models for robotics: A review towards real-world applications. *IEEE Access*. https://ieeexplore.ieee.org/abstract/document/11164279

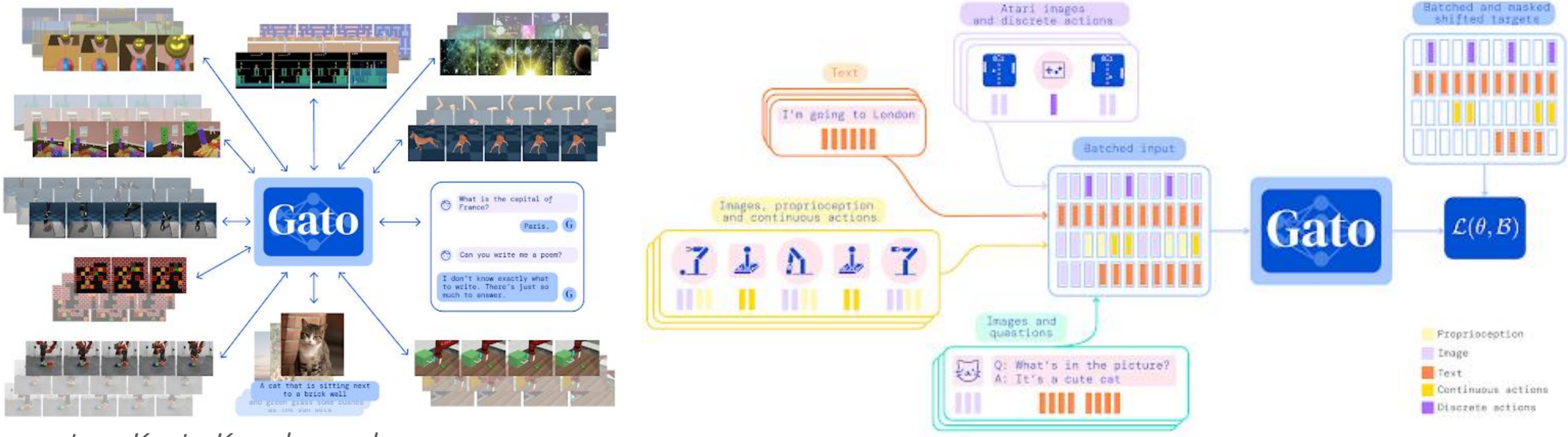
Gato

[DeepMind, TMLR2022]

A single Transformer model can perform many tasks, including text chat, VQA, image captioning, gameplay, and robot control.

Language is tokenized with SentencePiece, images with ViT, and action tokens are generated autoregressively by a decoder-only Transformer.

For robotics, it covered only simple block-stacking tasks.



Slide: courtesy Kento Kawaharazuka

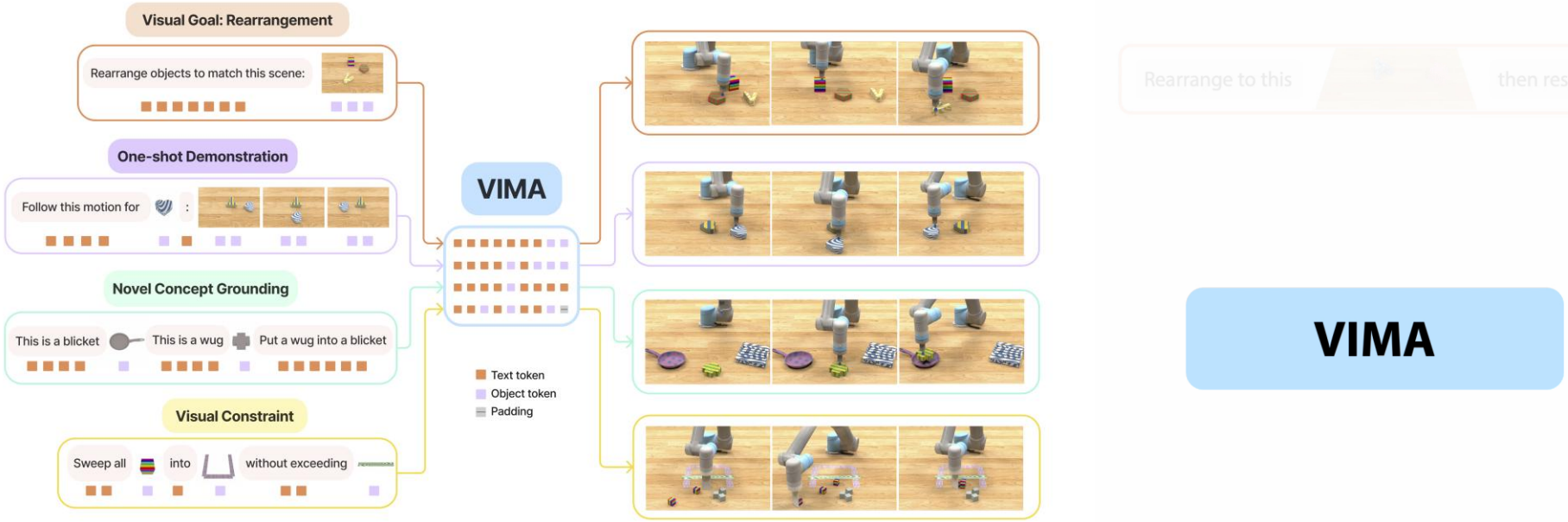
VIMA

[Y. Jiang+, ICML2023]

An encoder-decoder Transformer that supports diverse task instructions, including goal images and text.

Objects are detected with Mask R-CNN; object crops are tokenized with ViT; language instructions with a T5 encoder; bounding boxes are also tokenized.

Supports diverse robot tasks, but experiments were only in simulation.



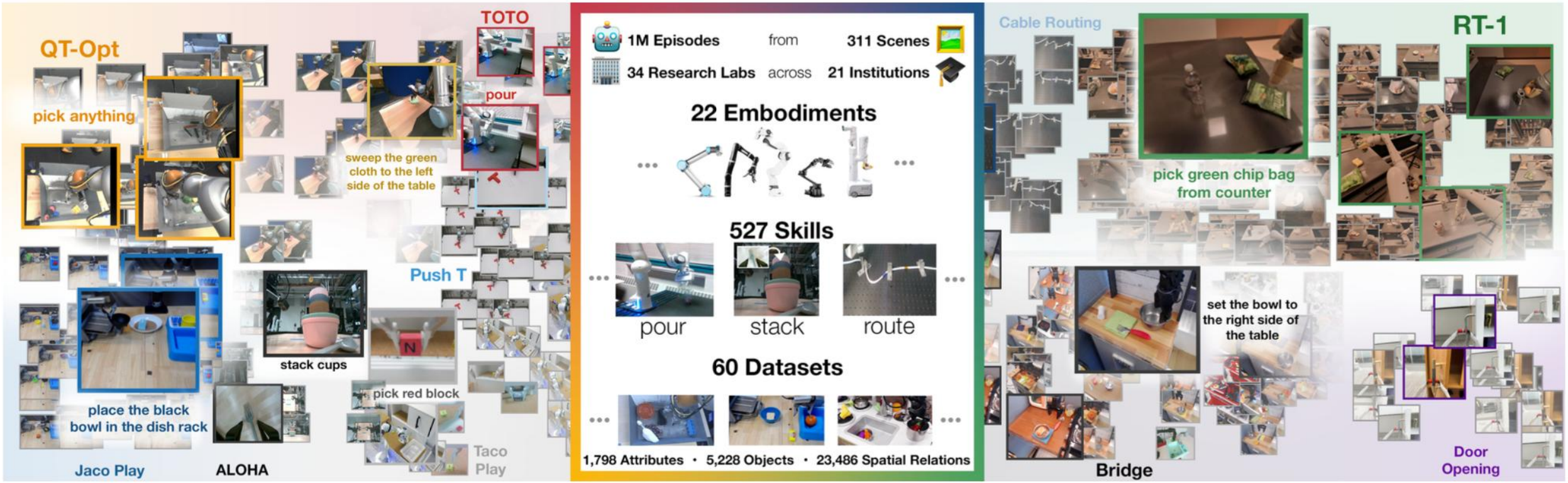
Slide: courtesy Kento Kawaharazuka

RT-X

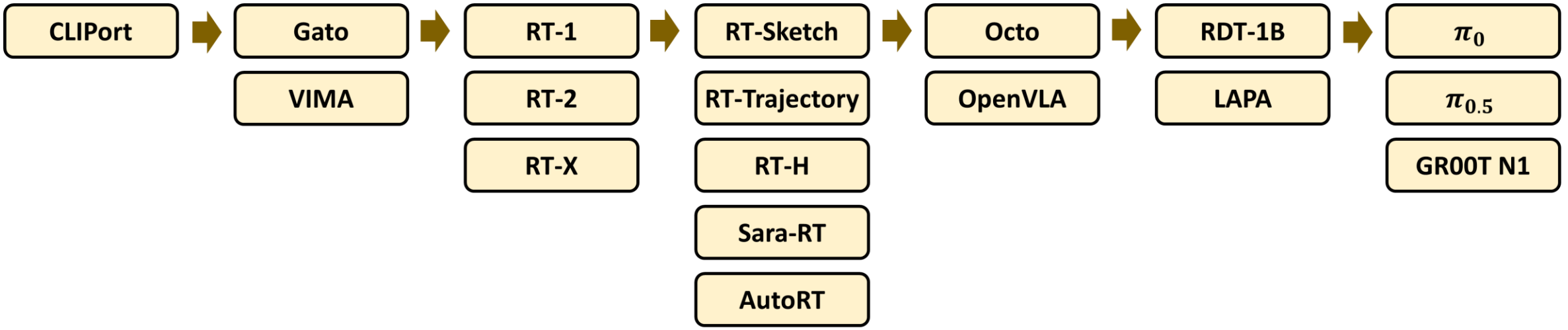
[Open X-Embodiment, ICRA2024]

Improves RT-1 / RT-2 by learning jointly from robot data with diverse embodiments rather than a single embodiment.

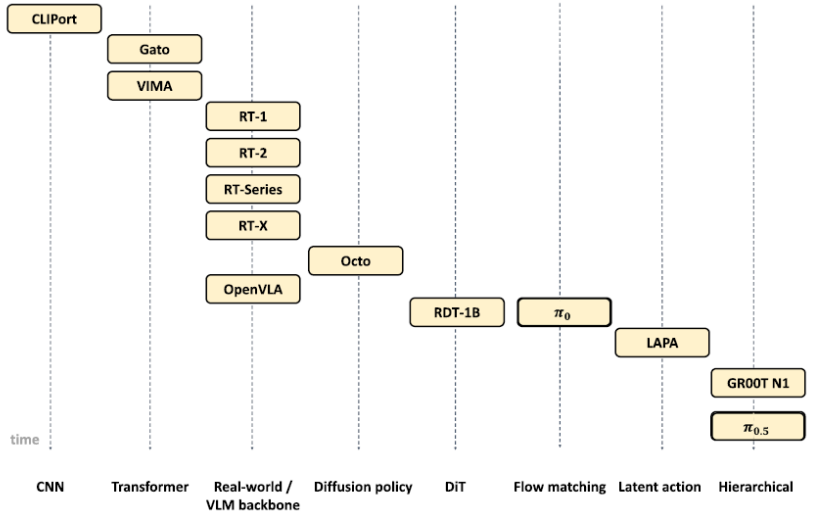
Built from 60 datasets across 22 robots, with 21 institutions and 173 authors.



Slide: courtesy Kento Kawaharazuka



Schematics: courtesy Kento Kawaharazuka



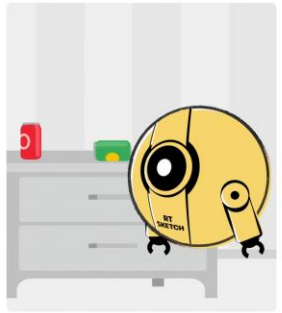
Kawaharazuka, K., Oh, J., Yamada, J., Posner, I., & Zhu, Y. (2025). Vision-language-action models for robotics: A review towards real-world applications. *IEEE Access*. https://ieeexplore.ieee.org/abstract/document/11164279

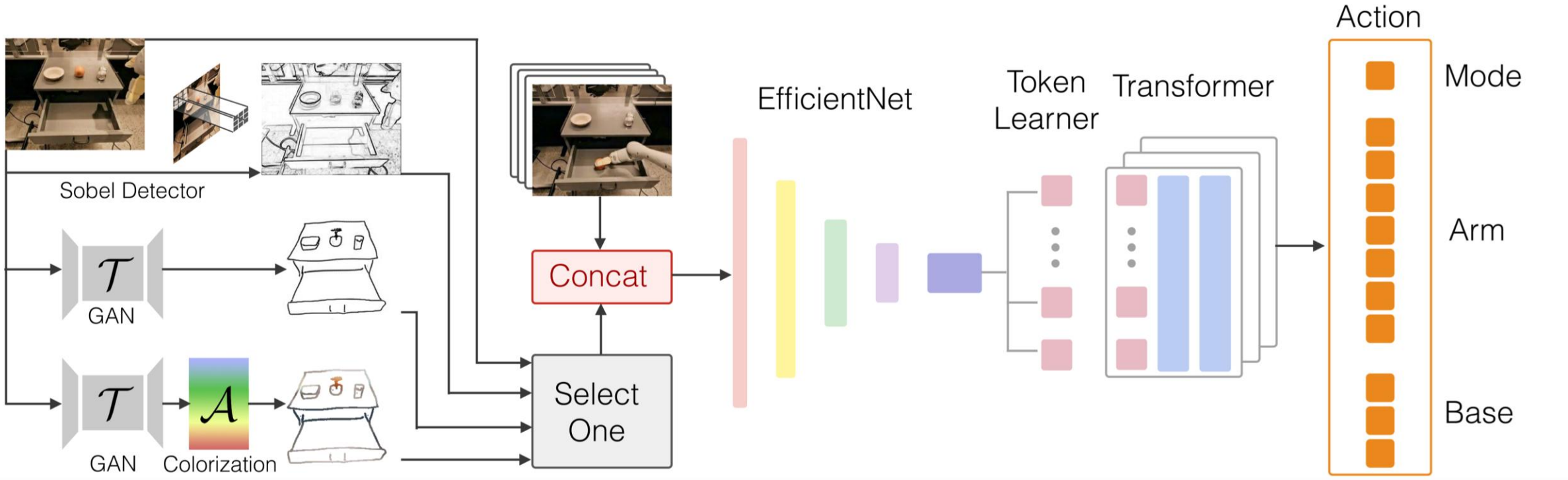
RT-Sketch

[P. Sundaresan+, CoRL2024]

There are various Robotics Transformer variants.

Instead of a goal image - and more constrained than language alone - this model generates actions from a goal sketch image.





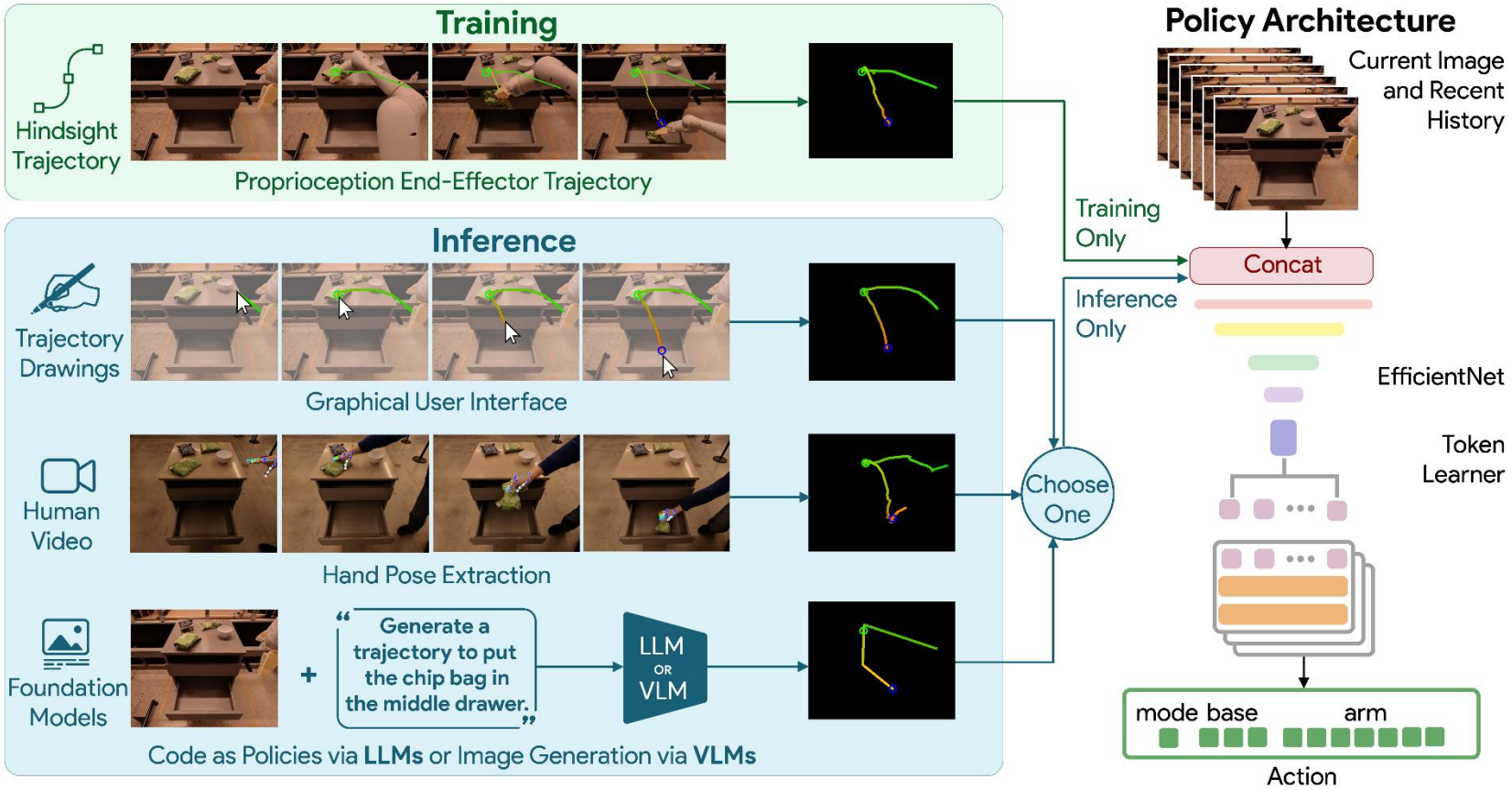
Slide: courtesy Kento Kawaharazuka

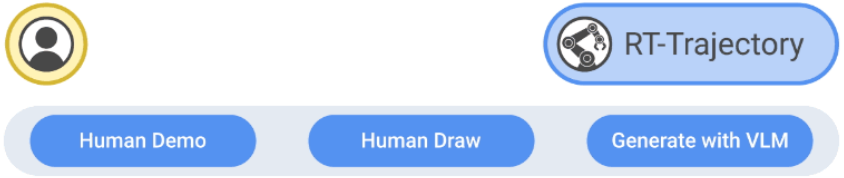
RT-Trajectory

[J. Gu+, ICLR2024]

Generates actions from an end-effector trajectory input.

Other RT-family extensions include AutoRT for automatic data collection and SARA-RT for higher speed.





Slide: courtesy Kento Kawaharazuka

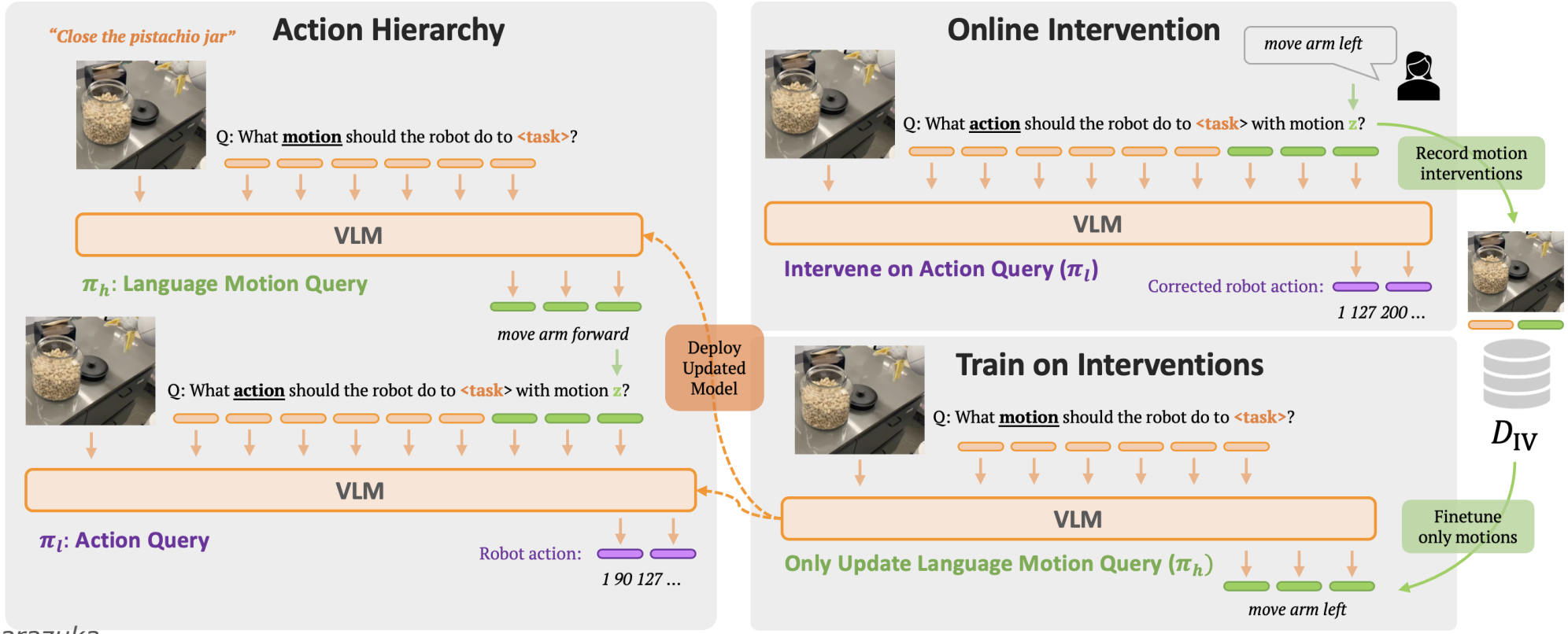
RT-H

[S. Belkhale, RSS2024]

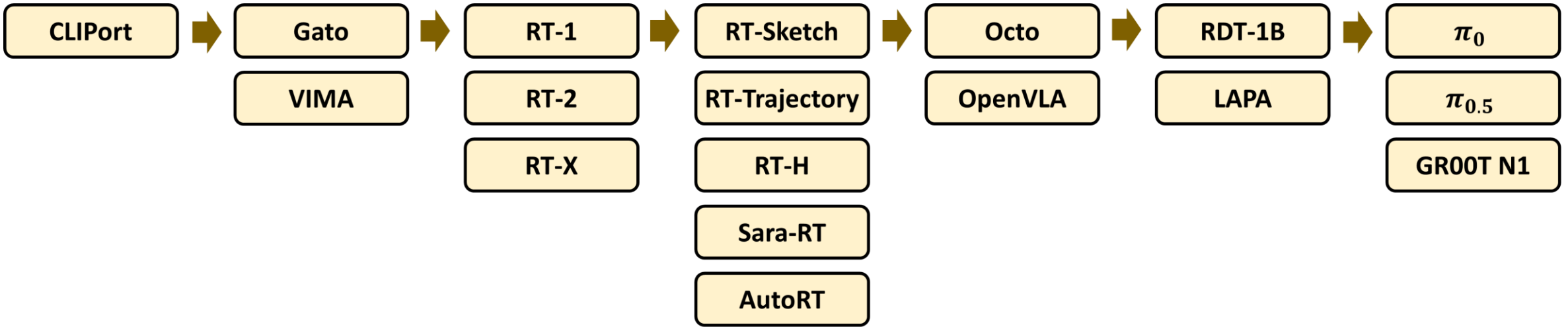
Hierarchical structure: a high-level policy predicts an intermediate representation called language motion, and a low-level policy predicts actions from language motion.

A single model learns both policies by changing the prompt.

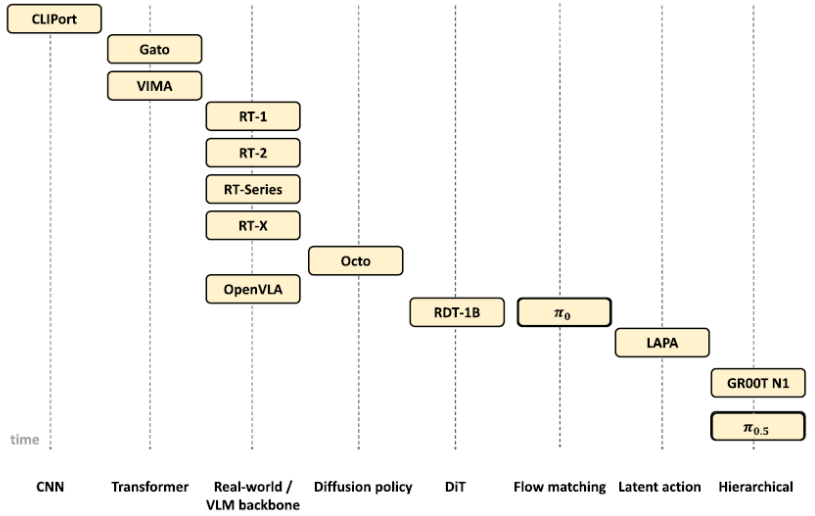
Hierarchical models became popular afterward!



Slide: courtesy Kento Kawaharazuka



Schematics: courtesy Kento Kawaharazuka



Kawaharazuka, K., Oh, J., Yamada, J., Posner, I., & Zhu, Y. (2025). Vision-language-action models for robotics: A review towards real-world applications. *IEEE Access*. https://ieeexplore.ieee.org/abstract/document/11164279

Octo

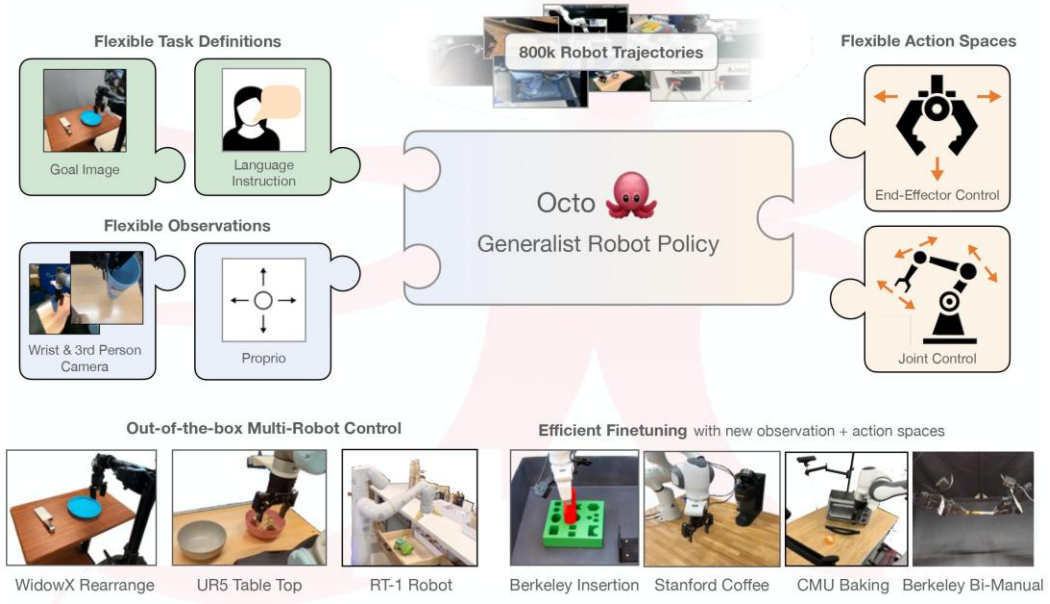
[D. Ghosh+, RSS2024]

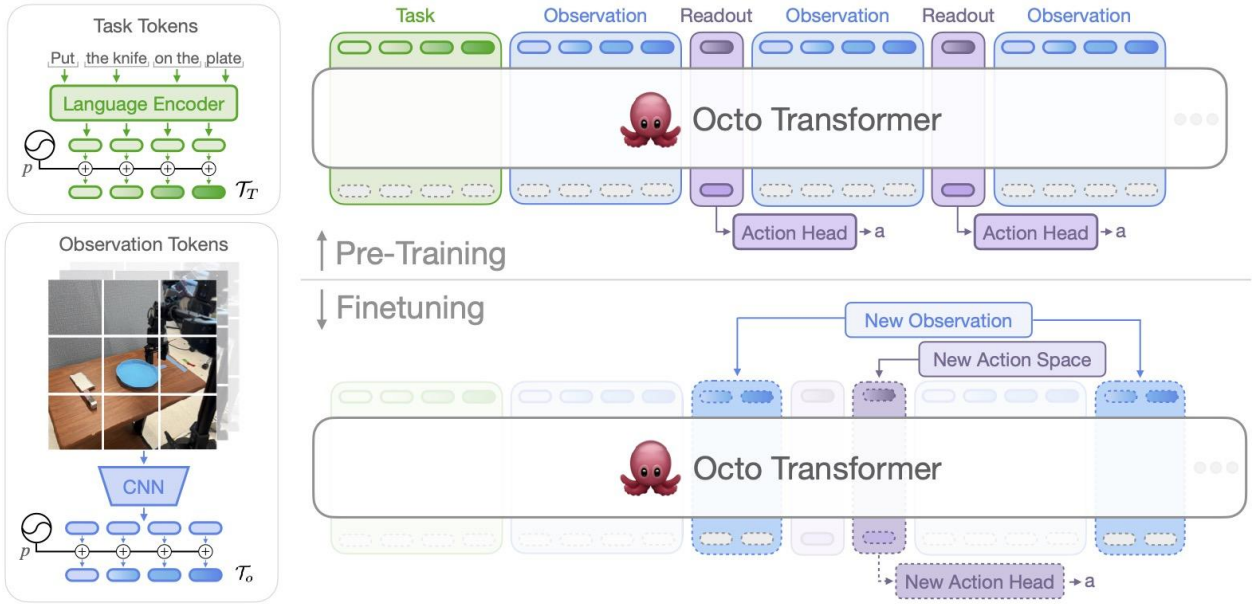
First VLA to incorporate **Diffusion Policy**.

All tokens are concatenated and fed into a Transformer, with a Diffusion Action Head conditioned by a readout token.

This enables continuous-valued actions rather than discrete tokens.

It drew major attention because all source code was open.





Slide: courtesy Kento Kawaharazuka

OpenVLA

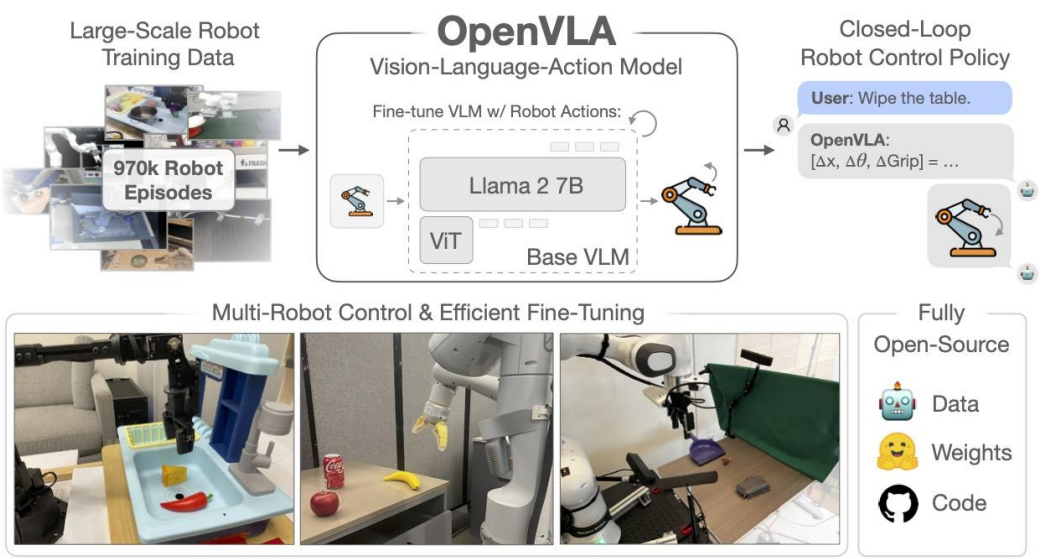
[M. J. Kim+, CoRL2024]

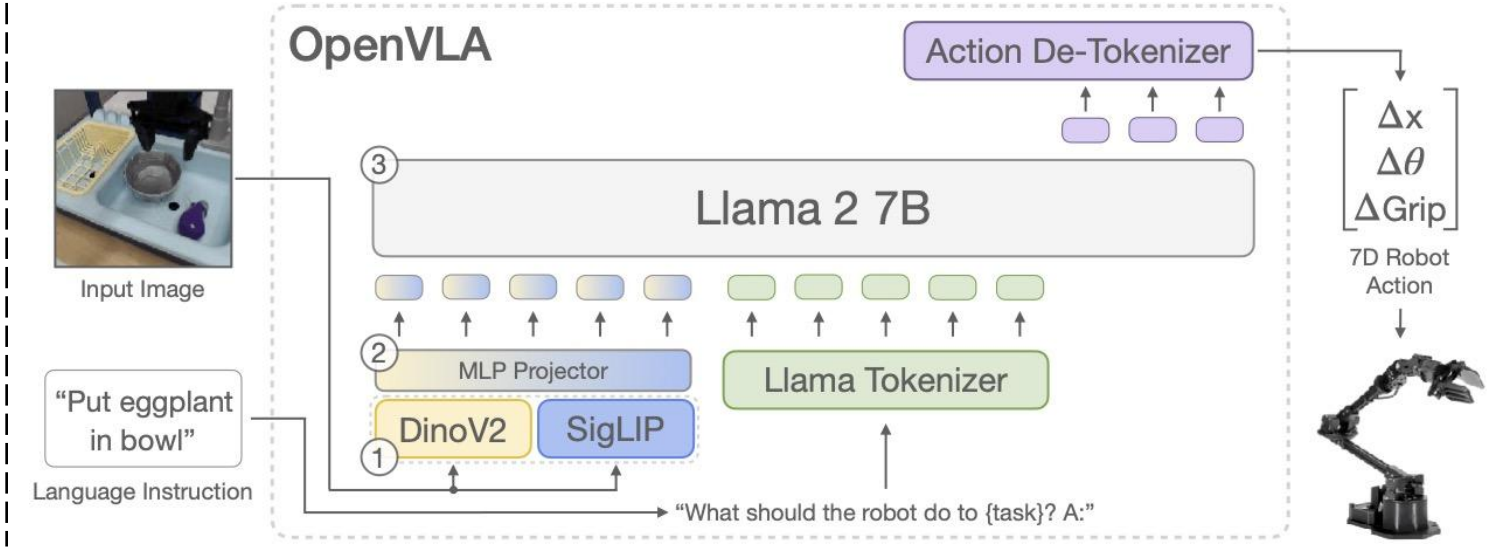
Released as open source, like Octo.

Image inputs are encoded with DINOv2 and SigLIP; the backbone is Prismatic VLM based on LLaMA 2.

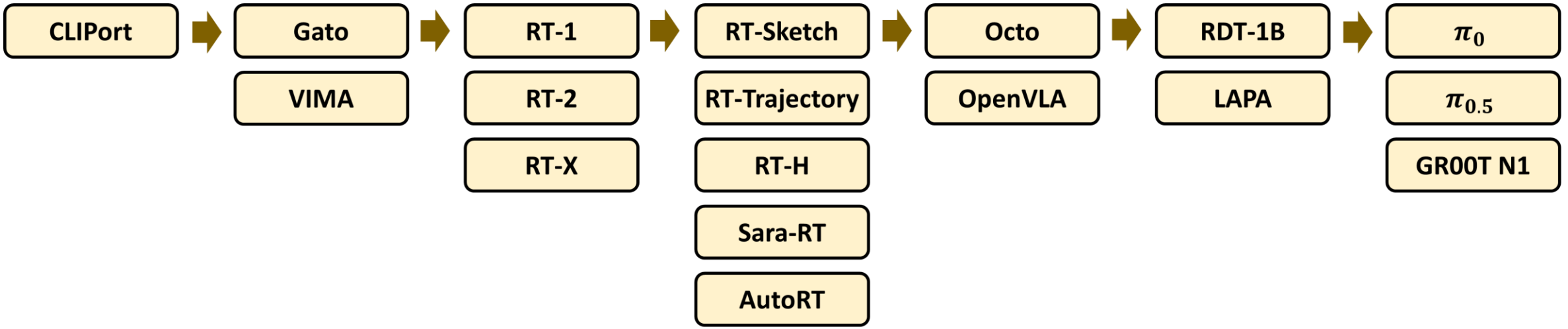
Full fine-tuning on the RT-X dataset yields higher performance than RT-2 and Octo.

OpenVLA / Prismatic VLM is frequently used as a base model.

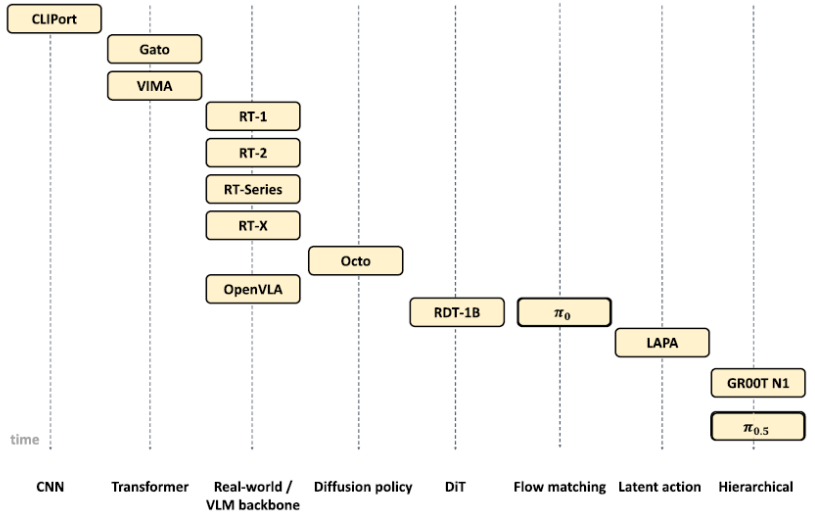




Slide: courtesy Kento Kawaharazuka



Schematics: courtesy Kento Kawaharazuka



Kawaharazuka, K., Oh, J., Yamada, J., Posner, I., & Zhu, Y. (2025). Vision-language-action models for robotics: A review towards real-world applications. *IEEE Access*. https://ieeexplore.ieee.org/abstract/document/11164279

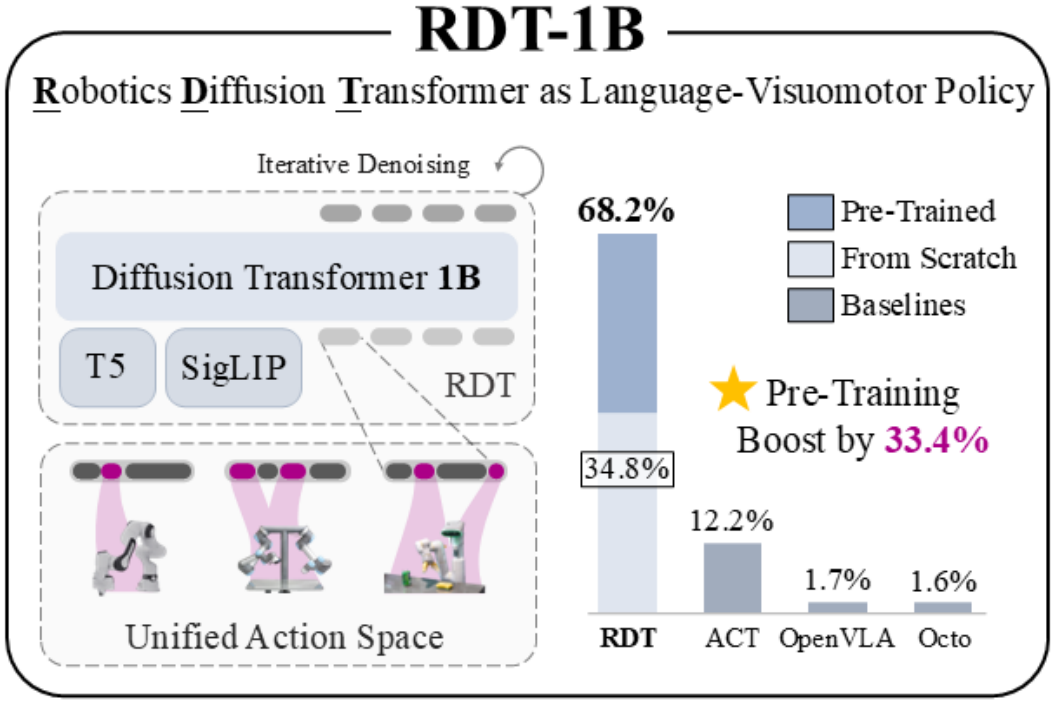
RDT-1B

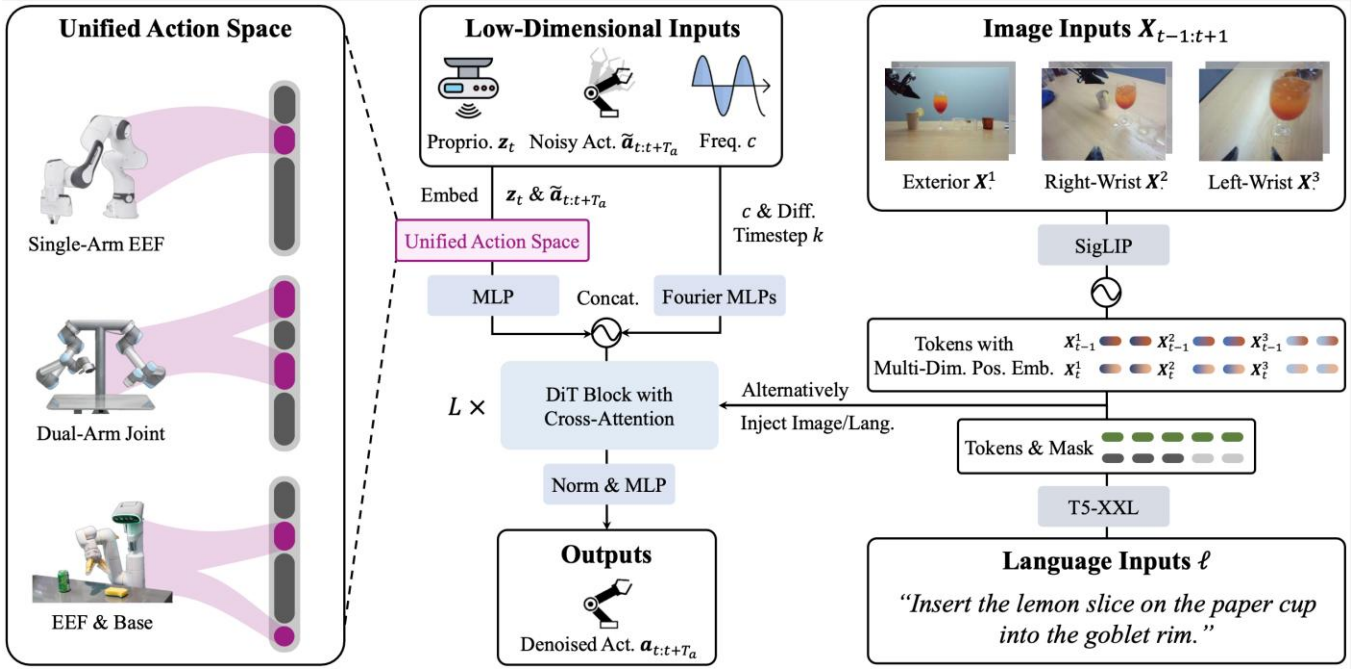
[S. Liu+, ICLR2025]

A large-scale **Diffusion Transformer** for robots.

Instead of using Diffusion Policy as the action head, it directly uses a Transformer to represent the diffusion process conditioned on image and text.

This ties vision and language more tightly to action.



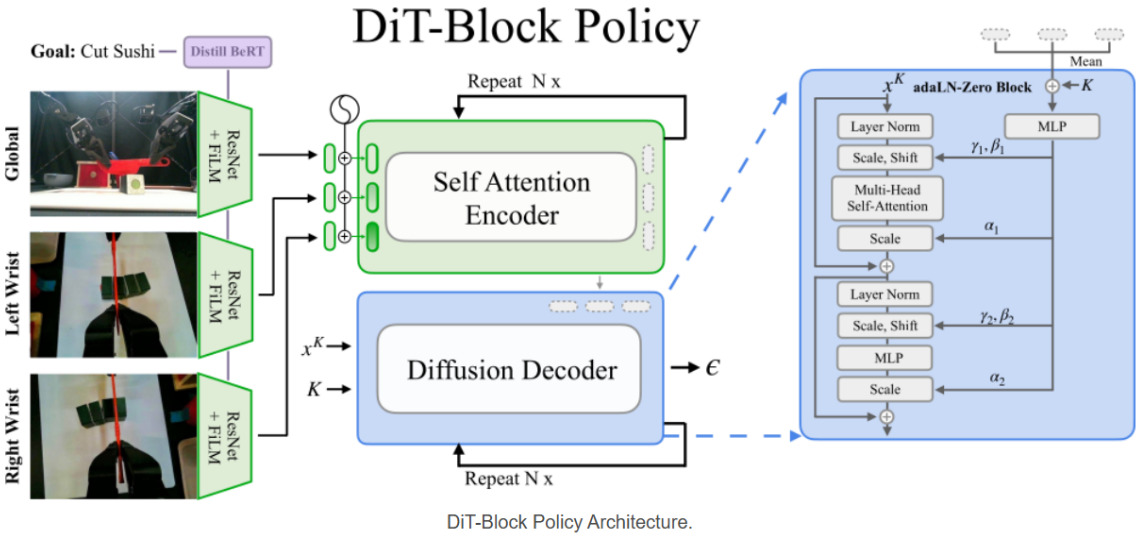


Slide: courtesy Kento Kawaharazuka

Building blocks: Diffusion transformer

- Start from random (noisy) action trajectory
- Iteratively refine ("denoise") into a valid action
- Transformer attends to:
  - Vision (scene)
  - Language (task)
- Each step improves the action

Key idea:
Refine noise into structured behavior



DiT-Block Policy Architecture.

https://dit-policy.github.io/

Scalable Diffusion Models with Transformers | DiT Explanation and Implementation https://youtu.be/aSLDXdc2hkk?si=KWVFSmbTDYfJvyvV
Peebles, W., & Xie, S. (2023). *Scalable diffusion models with transformers*. https://arxiv.org/abs/2212.09748
Dasari, S., Mees, O., Zhao, S., & Levine, S. (2024). *Diffusion transformer policy*. OpenReview. https://openreview.net/forum?id=PvvXDazPMs
Mees, O., et al. (2024). *The ingredients for robotic diffusion transformers*. https://www.oiermees.com/publication/dit_policy/
Lightly AI. (2025). *Diffusion transformers explained*. https://www.lightly.ai/blog/diffusion-transformers-dit

LAPA

[S. Ye+, ICLR2025]

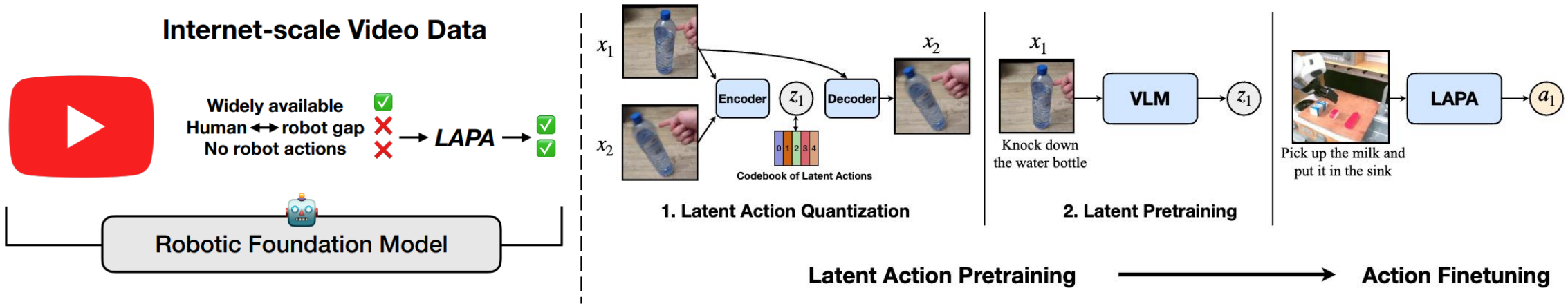
Extract **latent actions** from unlabeled human demonstration videos and use them for **VLA pretraining**

Compute the feature difference between x_t and $x\_\{t+H\}$, tokenize it into z_t with a **VQ-VAE**, and learn to reconstruct $x\_\{t+H\}$ from x_t and z_t.

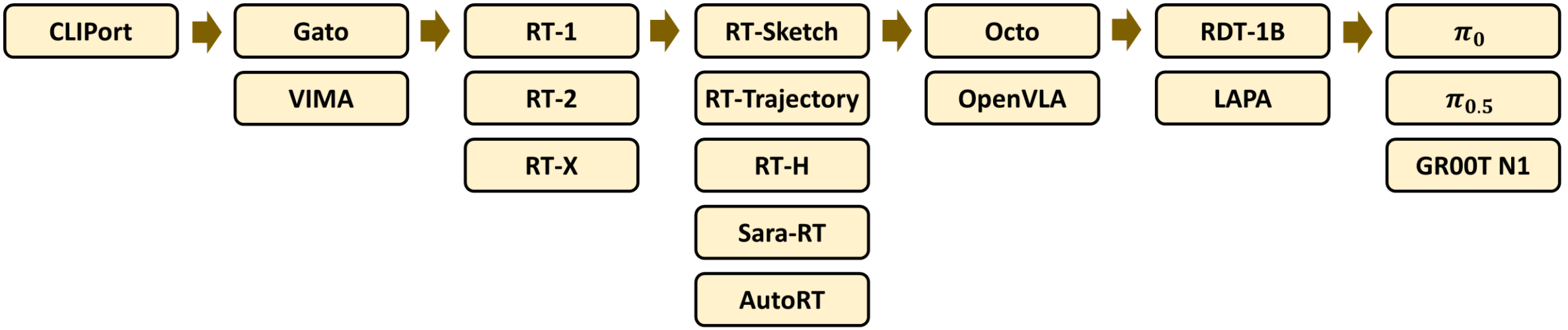
Train the model so that z_t can be predicted from the VLA **readout token** through an MLP.

During post-training, replace only the MLP and learn the robot action outputs.

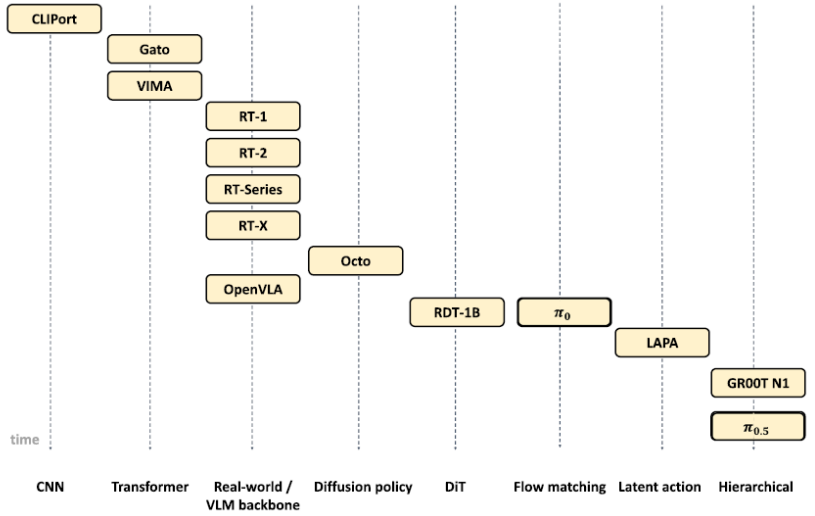
Makes it possible to use large-scale human demonstration videos as training data.



Slide: courtesy Kento Kawaharazuka



Schematics: courtesy Kento Kawaharazuka



Kawaharazuka, K., Oh, J., Yamada, J., Posner, I., & Zhu, Y. (2025). Vision-language-action models for robotics: A review towards real-world applications. *IEEE Access*. https://ieeexplore.ieee.org/abstract/document/11164279

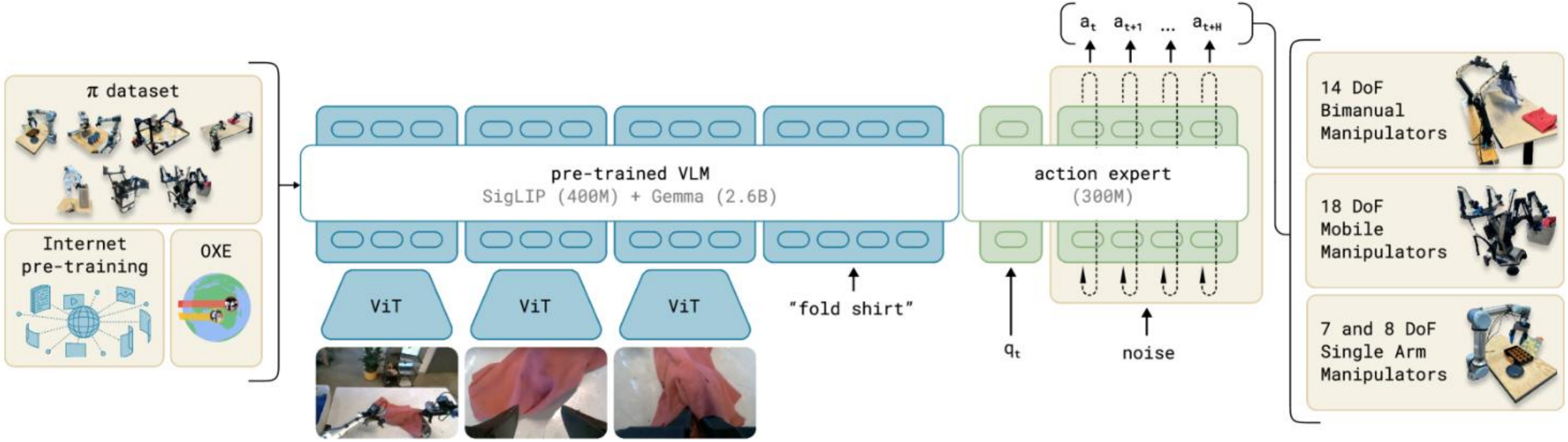
π_0

[Physical Intelligence, 2024]

Uses **Flow Matching** instead of a diffusion process, enabling 50 Hz action generation.

The base model is PaliGemma, using Gemma and SigLIP.

Conditioned on proprioception and the Transformer readout token, it outputs a vector field and recovers the desired action.



Slide: courtesy Kento Kawaharazuka

π_0

nce (π), 2024]

https://www.physicalintelligence.company/blog/pi0

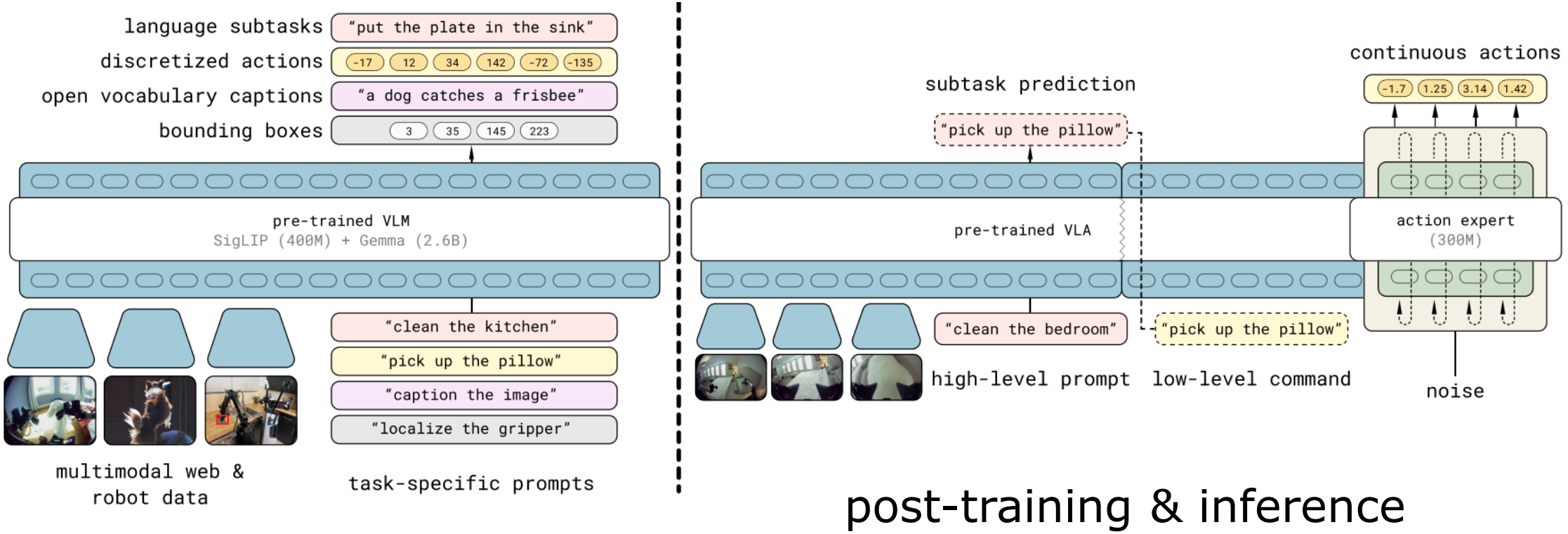
$\pi_{0.5}$

[Physical Intelligence, 2025]

During pretraining, it learns subtask prompts and discrete action tokens.

During post-training, it takes a subtask prompt as input and learns with Flow Matching.

It integrates two levels: discrete actions are easier to align with language, but the final output should be smooth and continuous.



Slide: courtesy Kento Kawaharazuka

$\pi^*_{0.6}$

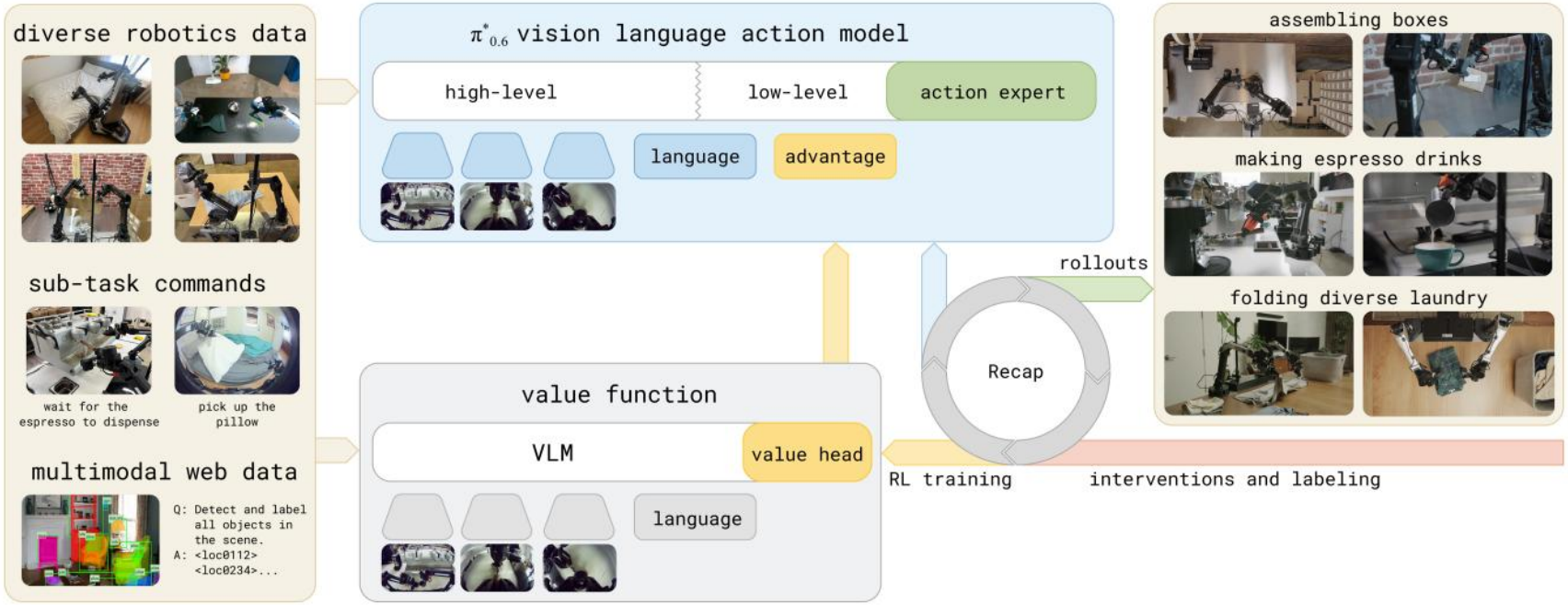
[Physical Intelligence, 2025]

How can robots keep learning behaviors from real experience?

Imitation learning -> deployment / assistance / value learning -> reinforcement learning

Continue tuning the VLA with batch offline RL.

Succeeded in making espresso continuously for 13 hours.



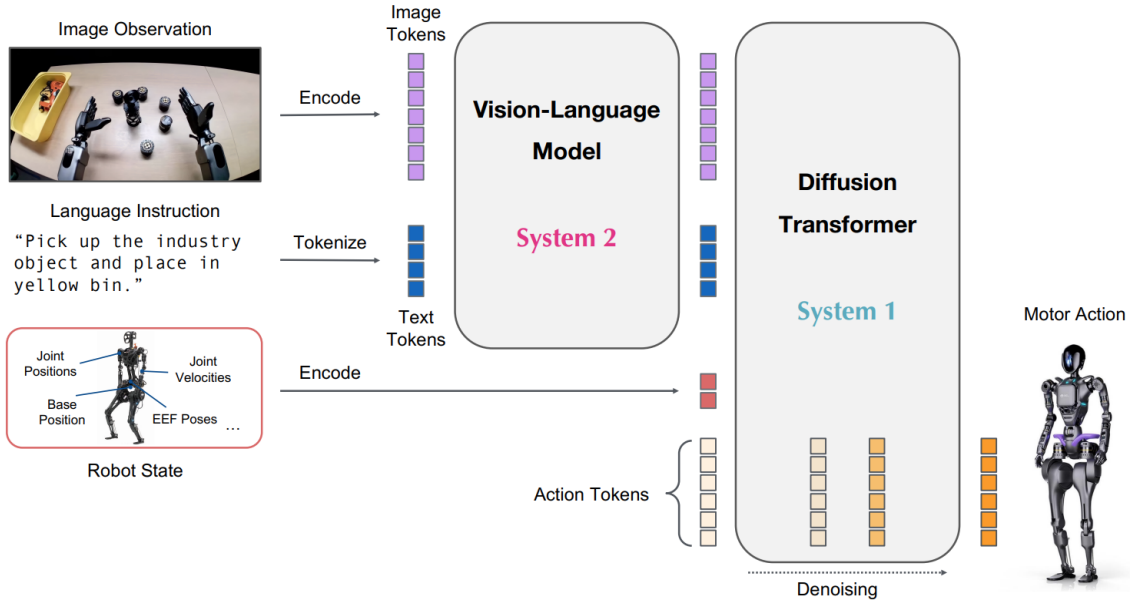
Slide: courtesy Kento Kawaharazuka

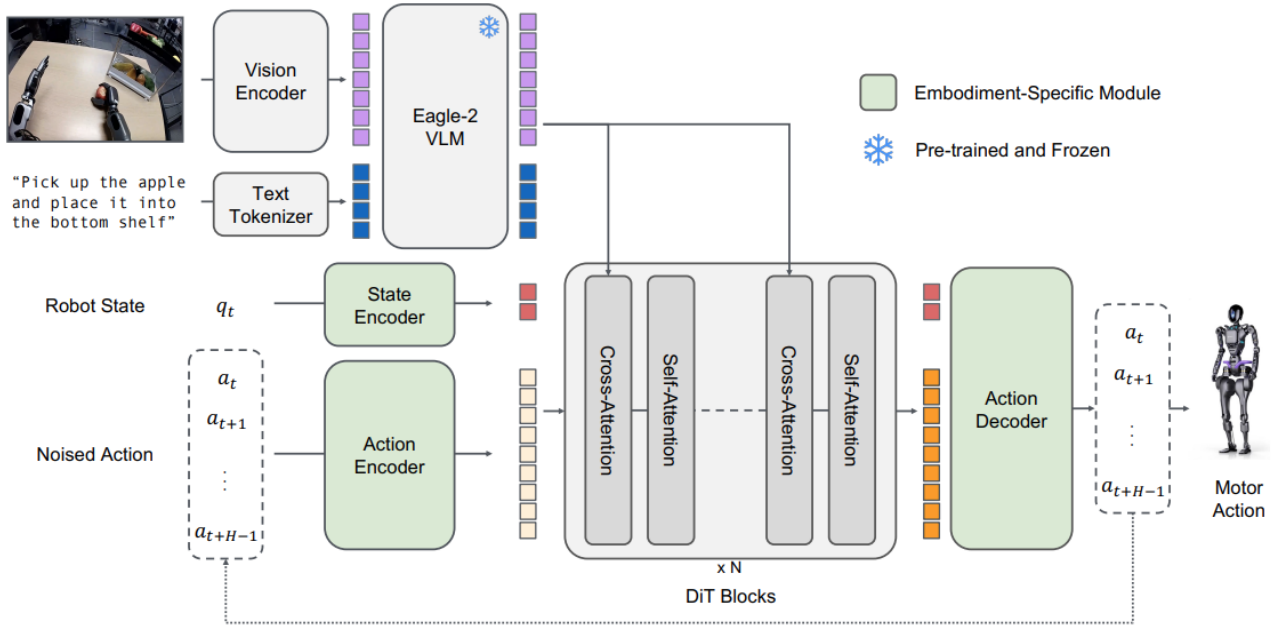
GR00T N1

[NVIDIA, 2025]

Combines essentially everything we have seen so far: hierarchy, Flow Matching, Diffusion Transformer, and data usage inspired by LAPA.

Applies cross-attention from VLM tokens to a Diffusion Transformer and outputs continuous actions with Flow Matching.





Slide: courtesy Kento Kawaharazuka

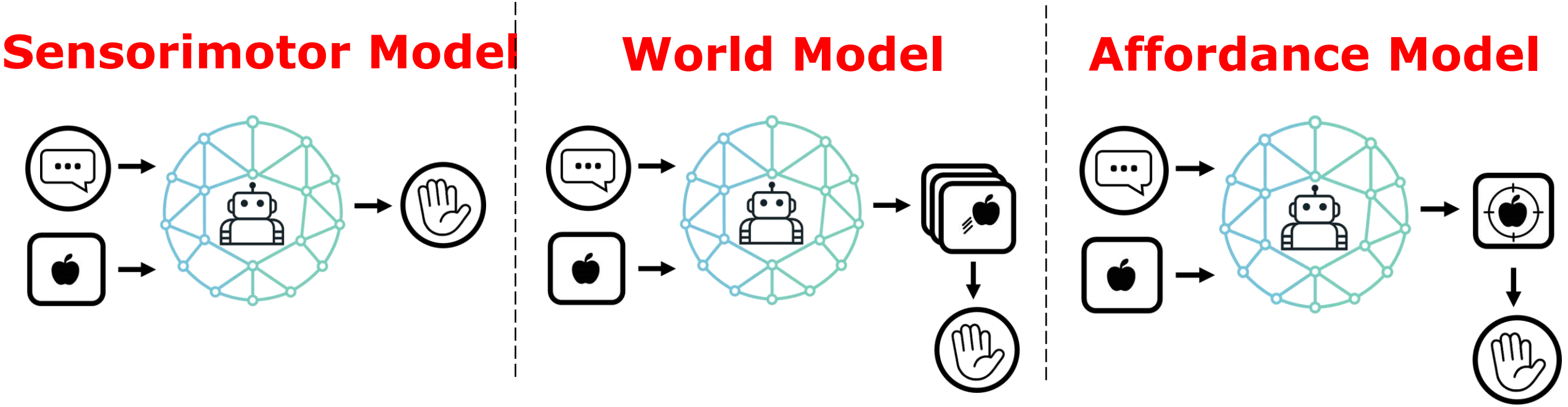
Outline

- Motivation and context – Robotics & AI
- History and evolution of key building blocks
- VLA Architectures
- Data collection and datasets
- Embodiment and cross-embodiment

Categories of VLA architectures

There are three main forms.

Most VLAs are sensorimotor, but other forms are also used.

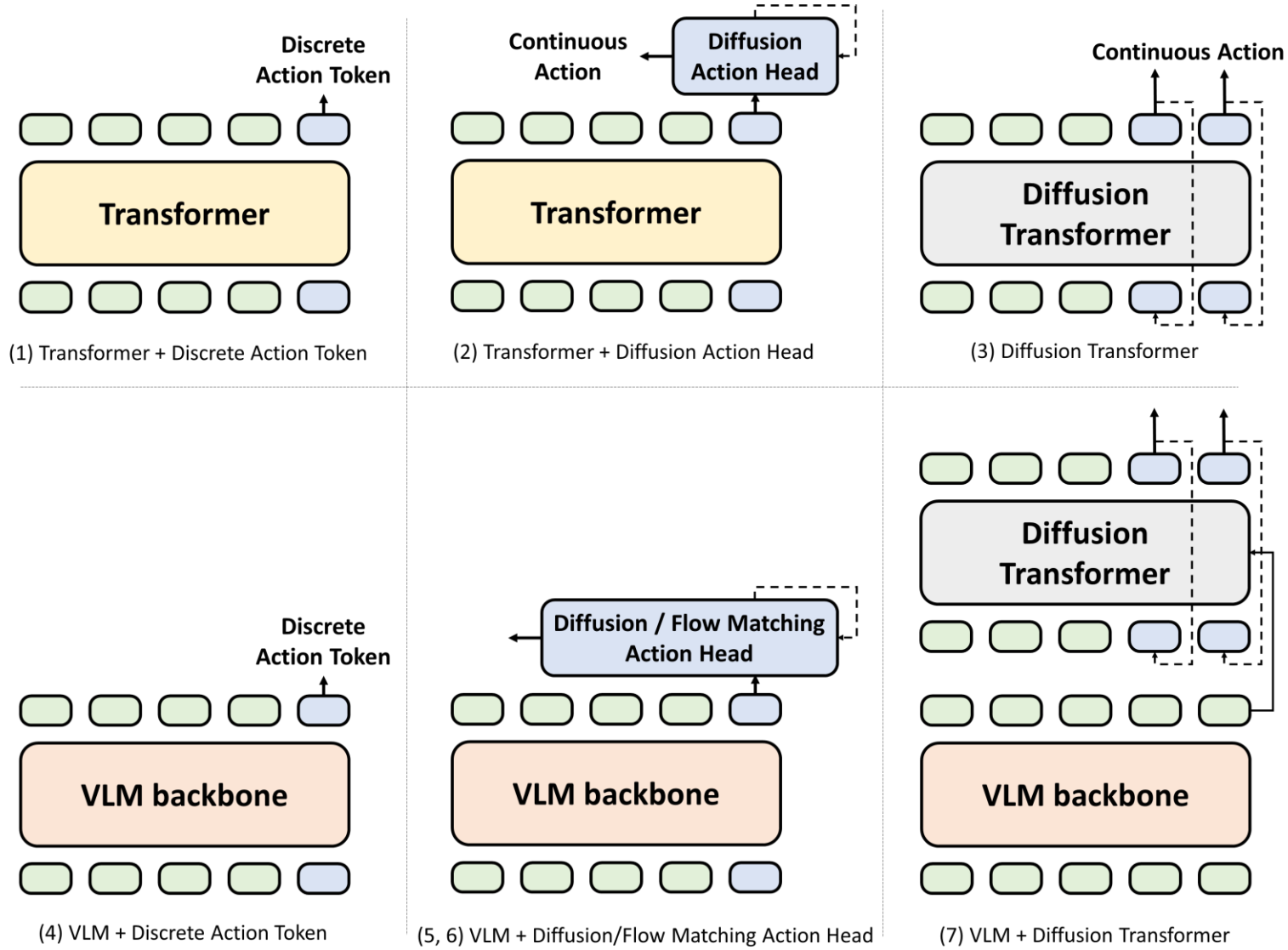


Slide: courtesy Kento Kawaharazuka

Categories of sensorimotor-model VLAs

A typical example

1. RT-1, Gato
2. Octo, NoMAD
3. RDT-1B, LBMs
4. RT-2, GR-1
5. π_0, GO-1
6. GR00T N1



Slide: courtesy Kento Kawaharazuka

Outline

- Motivation and context – Robotics & AI
- History and evolution of key building blocks
- VLA Architectures
- Data collection and datasets
- Embodiment and cross-embodiment

ARTIFICIAL INTELLIGENCE

Good old-fashioned engineering can close the 100,000-year "data gap" in robotics

Large vision-language models (VLMs) based on internet-scale data can now pass the Turing test for intelligence. In this sense, data have "solved" language, and many claim that data have solved speech recognition and computer vision.

Will data also solve robotics? Rich Sutton points out in "The Bitter Lesson" (*1*) that data and black-box "end-to-end" models have surpassed all the best-laid analytic work in artificial intelligence (AI). I accept that this trend will eventually produce general-purpose robots. But the question is… when?

Using commonly accepted metrics for converting word and image tokens into time, the amount of internet-scale data (texts and images) used to train contemporary VLMs is on the order of 100,000 years—it would take a human that long to read or view these data (*2*). However, the data needed to train robots are a combination of video inputs with robot motion commands: Those data do not exist on the internet.

Goldberg, K. (2025). Good old-fashioned engineering can close the 100,000-year "data gap" in robotics. *Science Robotics*, *10*(105), eaea7390.

Data collection methods

Figure 1: **Data Pyramid for Robot Foundation Model Training.** GR00T N1's heterogeneous training corpora can be represented as a pyramid: data quantity decreases, and embodiment-specificity increases, moving from the bottom to the top.

Bjorck, J., Castañeda, F., Cherniadev, N., Da, X., Ding, R., Fan, L., ... & Zhu, Y. (2025). Gr00t n1: An open foundation model for generalist humanoid robots. https://arxiv.org/abs/2503.14734

Data collection methods

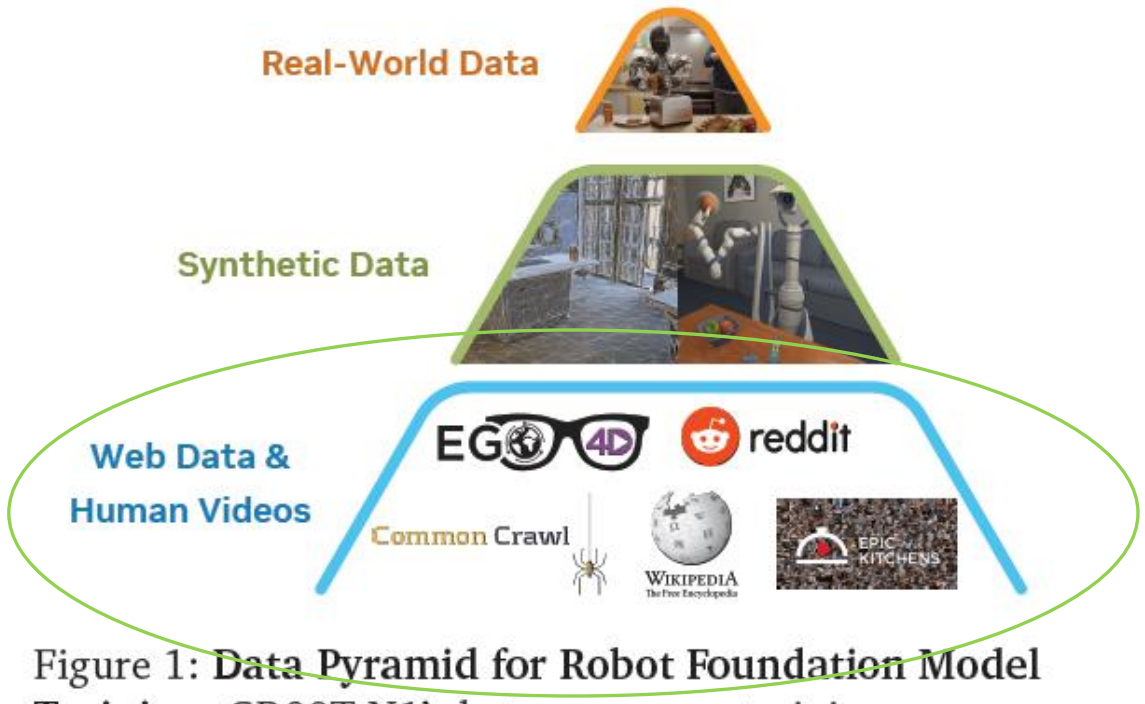
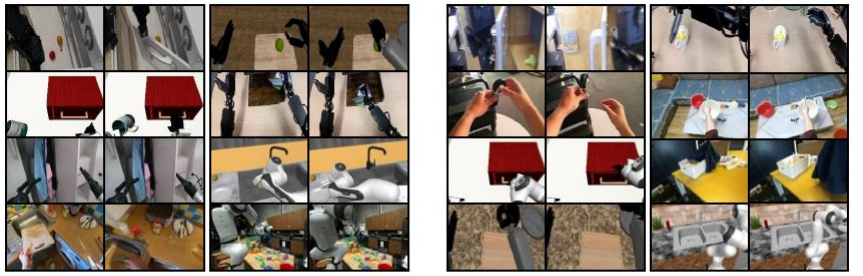


Figure 1: **Data Pyramid for Robot Foundation Model Training.** GR00T N1's heterogeneous training corpora can be represented as a pyramid: data quantity decreases, and embodiment-specificity increases, moving from the bottom to the top.



Latent action 1: Move the right arm to the left

Latent action 2: Move the right arm to the right

Figure 4: **Latent Actions.** We retrieve similar latent embeddings across various embodiments. The left images illustrate the latent action that corresponds to moving the right arm (or hand) to the left, while the right images illustrate the latent action that corresponds to moving the right arm (or hand) to the right. Note that this general latent action is not only consistent in different robot embodiments, but also in human embodiment.

- **Ego4D** is a large-scale egocentric video dataset that includes diverse recordings of everyday activities (Grauman et al., 2022);
- **Ego-Exo4D** adds complementary exocentric (third-person) views alongside first-person recordings (Grauman et al., 2024);
- **Assembly-101** focuses on complex assembly tasks by providing detailed videos of step-by-step object assembly (Sener et al., 2022);
- **EPIC-KITCHENS** includes first-person footage of culinary activities (Damen et al., 2018);
- **HOI4D** captures human-object interactions with frame-wise annotations for segmentation, hand and object poses, and actions (Liu et al., 2022);
- **HoloAssist** captures collaborative and assistive tasks within augmented reality environments (Wang et al., 2023);
- **RH20T-Human** includes recordings of fine-grained manipulation tasks with an emphasis on natural hand-object interactions across diverse real-world scenarios (Fang et al., 2023).

Bjorck, J., Castañeda, F., Cherniadev, N., Da, X., Ding, R., Fan, L., ... & Zhu, Y. (2025). Gr00t n1: An open foundation model for generalist humanoid robots. https://arxiv.org/abs/2503.14734

Data collection methods

Figure 1: **Data Pyramid for Robot Foundation Mode Training.** GR00T N1's heterogeneous training corpora can be represented as a pyramid: data quantity de creases, and embodiment-specificity increases, moving from the bottom to the top.

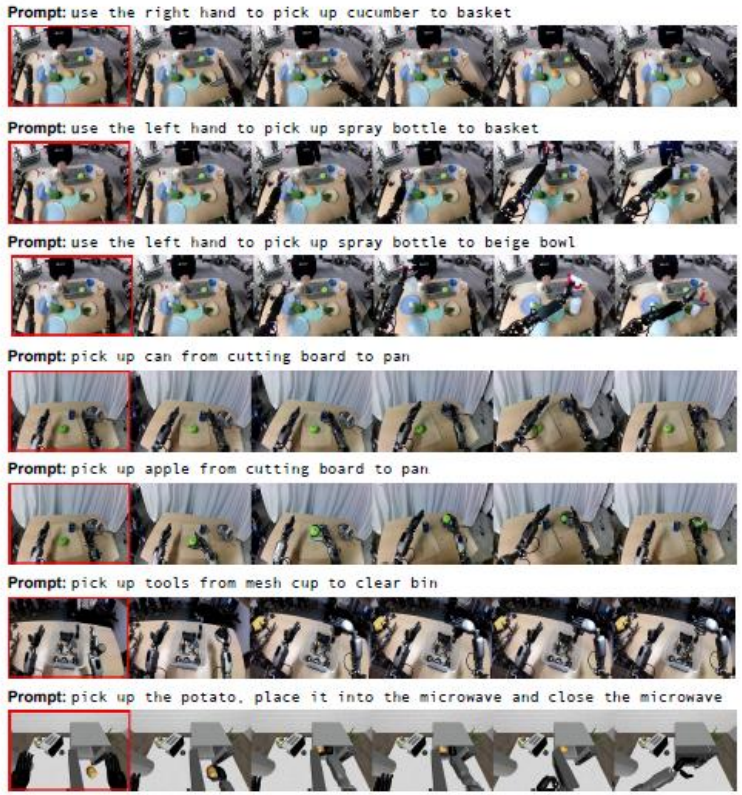


Figure 5: **Synthetically Generated Videos.** We leverage off-the-shelf video generation models to create neural trajectories to increase the quantity and diversity of our training datasets. These generated data can be used for both pre- and post-training of our GR00T N1. (1) The first three rows are generated from the same initial frames but with different prompts (change left or right, the location to place the object), (2) the following two are from the same initial frames but replace the object to pick up, (3) the next row showcases the video model generating a robot trajectory which is very challenging to generate in simulation (spilling contents inside a mesh cup into a bin), and (4) the last row is generated from an initial frame from simulation data. We use the red rectangles to indicate the initial frames.

Figure 7: **Simulation Tasks.** Our simulation experiments use tasks from two open-source benchmarks (RoboCasa (Nasiriany et al., 2024) in the top row and DexMimicGen (Jiang et al., 2024) in the middle row) and a newly developed suite of tabletop manipulation tasks that closely resemble our real-world tasks (bottom row). We provide Omniverse renderings of the tasks above.

Bjorck, J., Castañeda, F., Cherniadev, N., Da, X., Ding, R., Fan, L., ... & Zhu, Y. (2025). Gr00t n1: An open foundation model for generalist humanoid robots. https://arxiv.org/abs/2503.14734

Data collection methods

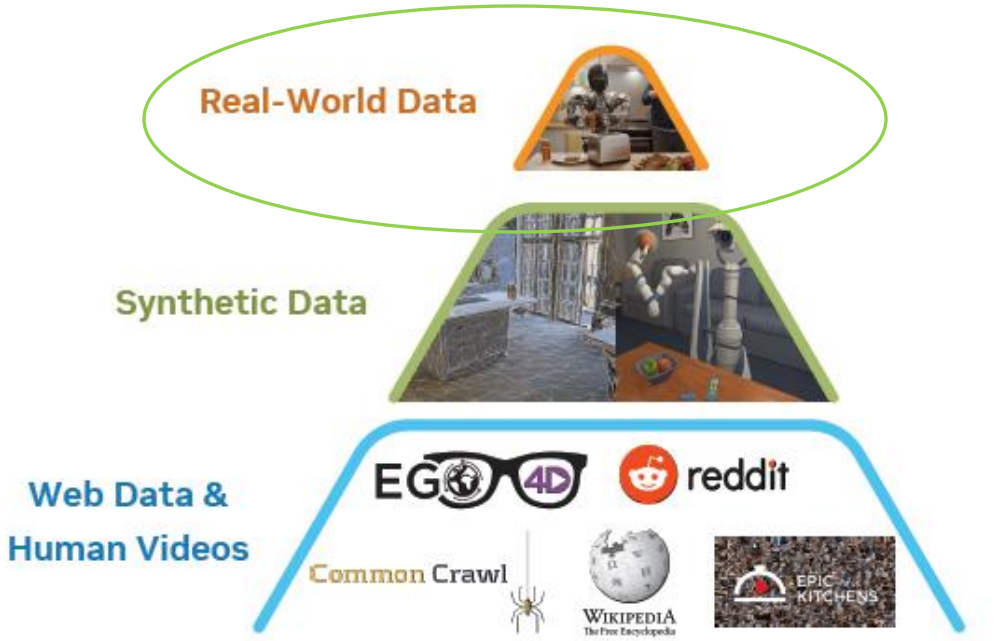


Figure 1: **Data Pyramid for Robot Foundation Model Training.** GR00T N1's heterogeneous training corpora can be represented as a pyramid: data quantity decreases, and embodiment-specificity increases, moving from the bottom to the top.

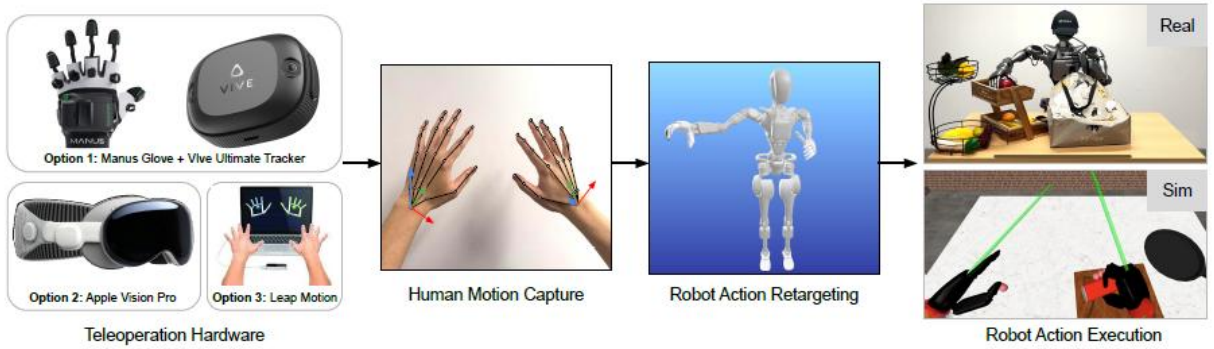


Figure 6: **Data Collection via Teleoperation.** Our teleoperation infrastructure supports multiple devices to capture human hand motion, including 6-DoF wrist poses and hand skeletons. Robot actions are produced through retargeting and executed on robots in real and simulation environments.

Bjorck, J., Castañeda, F., Cherniadev, N., Da, X., Ding, R., Fan, L., ... & Zhu, Y. (2025). Gr00t n1: An open foundation model for generalist humanoid robots. https://arxiv.org/abs/2503.14734

Coaching with corrections and practicing with reinforcement

RECAP enables two ways to get good training signals from "bad" experiential data: *coaching* to provide corrections, where an expert shows the robot how it can fix a mistake or do better, and *reinforcement learning*, where the robot judges for itself which of its behaviors were better or worse based on the overall outcome of an episode, and iteratively learns to perform the good behaviors while avoiding the bad ones.

For coaching to be useful, an expert teleoperator needs to provide corrections showing how to recover from the mistakes that the robot actually makes in the real world. In practice, this means running our best current policy and "taking over" with manual teleoperation when the robot makes a mistake. This intervention can be used as supervision, but unlike the demonstrations used to train the original policy, the intervention provides supervision for the situations that the policy actually puts the robot into, addressing the compounding mistakes issue.

An expert teleoperator takes over and provides a real-time correction for the robot's mistake.

https://www.pi.website/blog/pistar06

Outline

- Motivation and context – Robotics & AI
- History and evolution of key building blocks
- VLA Architectures
- Data collection and datasets
- Embodiment and cross-embodiment

Good Old-Fashioned AI, Embodied Intelligence, and Embodied AI

Classical: “intelligence as computation”

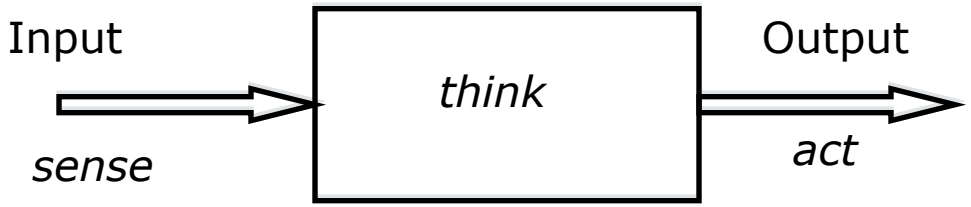
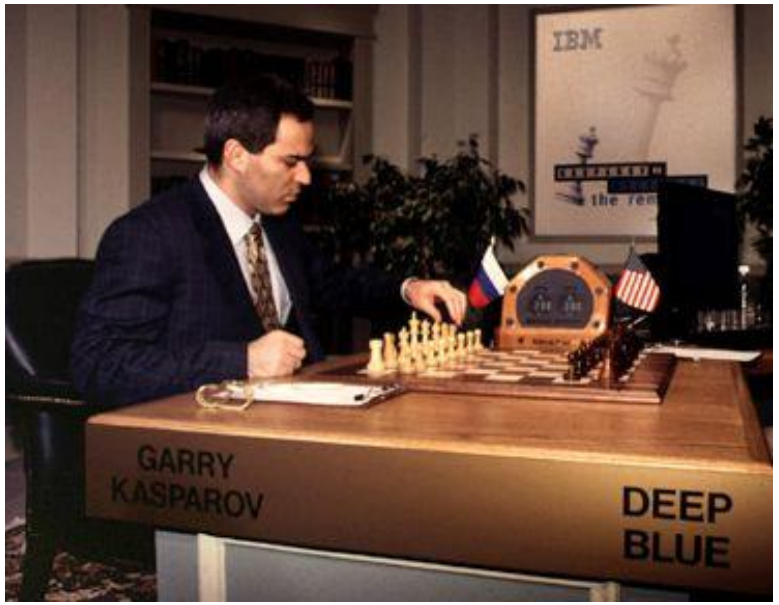


Illustration: Shun Iwasawa, from Pfeifer, R: How the body shapes the way we think, 2007

Where it works nicely… search

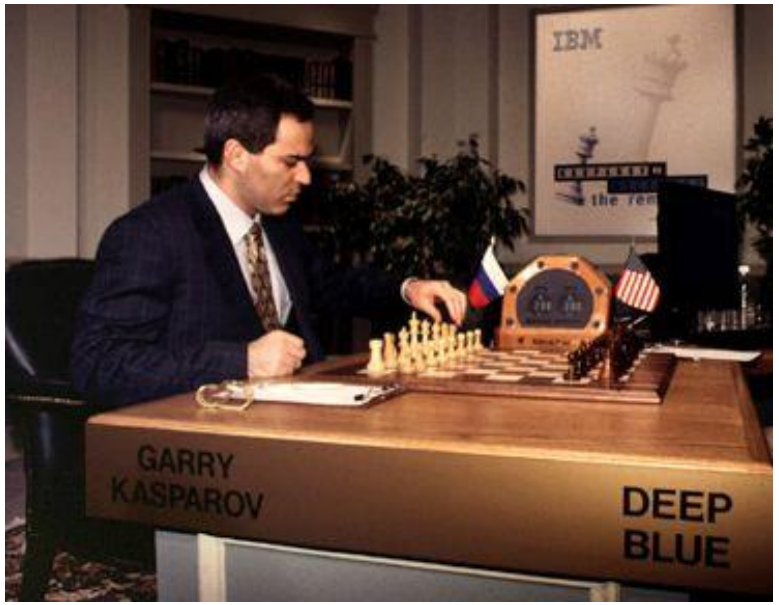


IBM Deep Blue chess computer, 1997 Google Deep Mind AlphaGo, 2016

- formally precisely defined discrete state space
- program has access to complete information (fully observable)
- deterministic state evolution
- not real-time (or soft real time)

- Premiere methods – e.g.: **search**, planning

Where it works nicely… search



IBM Deep Blue chess computer, 1997 Google Deep Mind AlphaGo, 2016

- formally precisely defined discrete state space
- program has access to complete information (fully observable)
- deterministic state evolution
- not real-time (or soft real time)

- Premiere methods – e.g.: **search**, planning

Classical AI – theoretical positions

Intelligence ~ abstract symbol processing

Functionalism
- Algorithm / software matters
- Hardware (on which it runs)
  does not matter

Physical Symbol Systems

Hypothesis (Newell and Simon 1976)

Digital computer
- Key tool
- Metaphor for the mind!

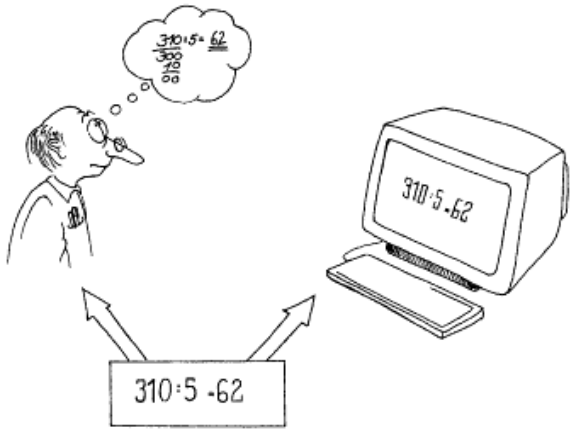


Fig. 2.4 from Pfeifer & Scheier 1999

Nicknamed GOFAI – Good Old-Fashioned Artificial Intelligence (Haugeland 1989)

Connecting the representation to the real world

Example : Blocks World

•STRIPS : A planning system – Has rules with precondition deletion list and addition list

on(B, table)	on(C, table)
on(A, table)	on(B, C)
on(C, A)	on(A, B)
hand empty	hand empty
clear(C)	clear(A)
clear(B)	

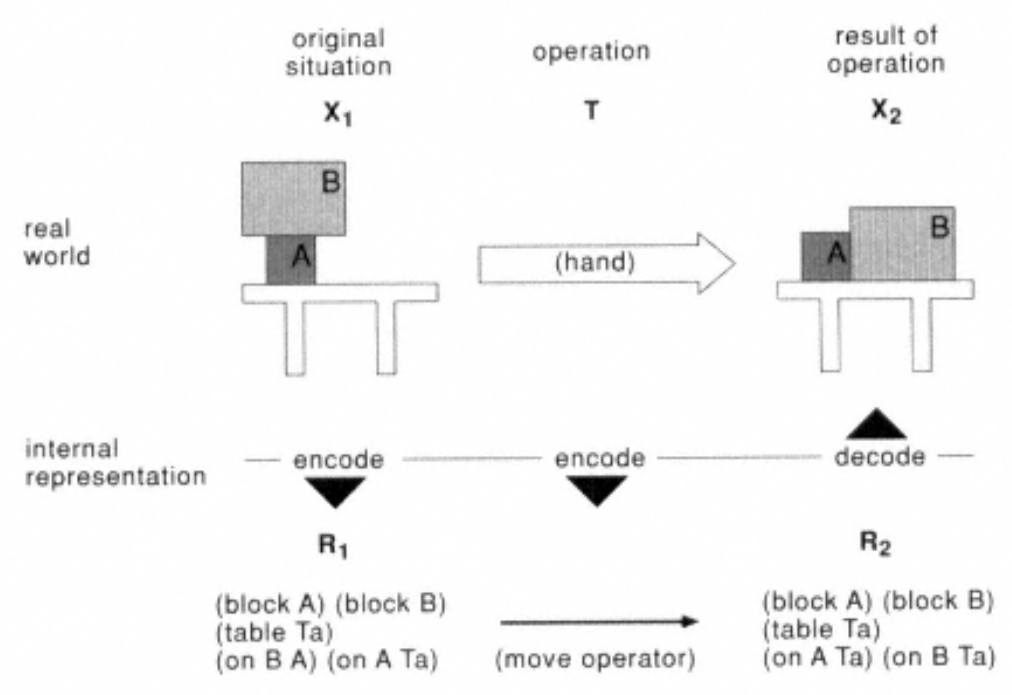
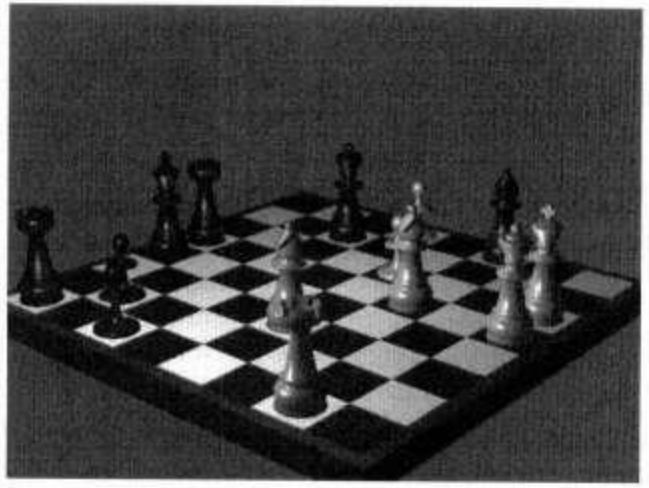
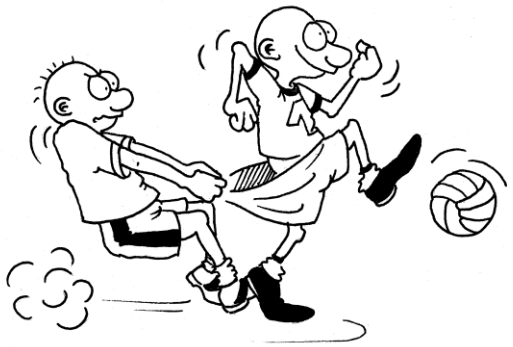


Fig. 2.5 from Pfeifer & Scheier 1999

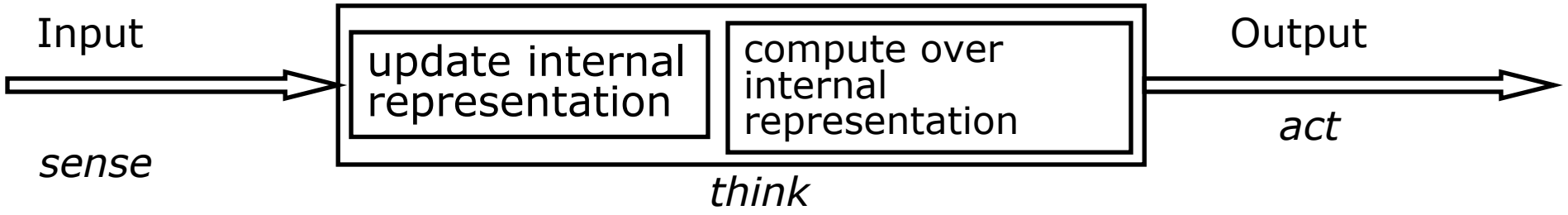
From formal world to the real world

GOFAI for robotics?



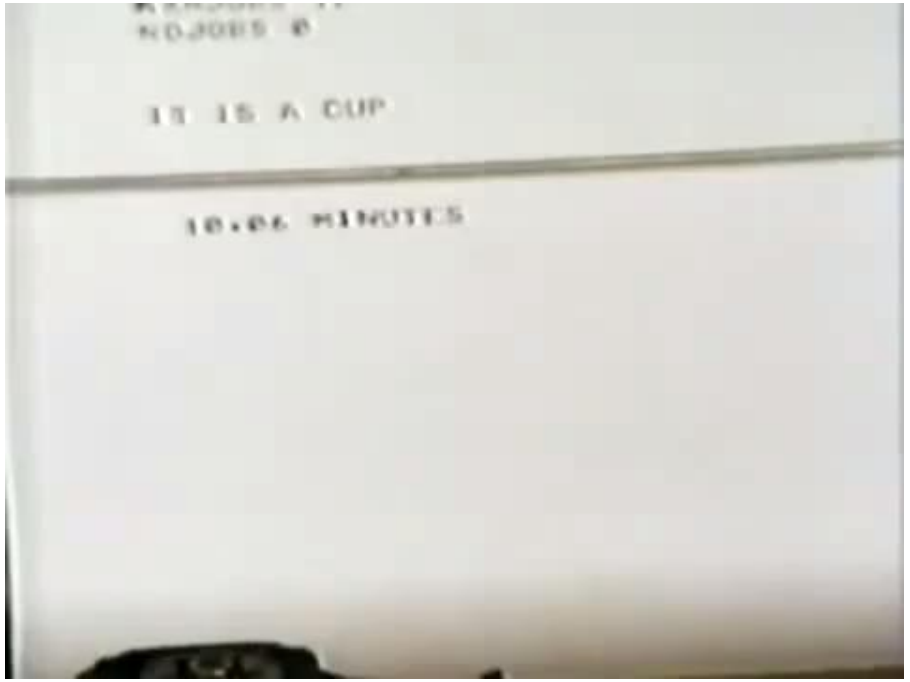


From formal world to the real world



Ancient times:

Stanford Cart, 1975



Each flash ~ 15 min. thinking.

GOFAI for robotics - problems

- The “interfaces” with the real world—previously uninteresting and underestimated—became fundamental and practical challenges.
- Frame problem
  - keeping the internal representation of the world consistent with the real world outside
- Symbol grounding problem (Harnad, 1990)
  - relationship of the symbolic representation and the outside world

STRIPS: A New Approach to the Application of Theorem Proving to Problem Solving[1]

Richard E. Fikes
Nils J. Nilsson
Stanford Research Institute, Menlo Park, California

Recommended by B. Raphael

Presented at the 2nd IJCAI, Imperial College, London, England, September 1–3, 1971.

The initial hopes of applying for example theorem solving (e.g., (Fikes and Nilsson 1971)) to real-world robot tasks have never really materialized.

TABLE 1. Formulation for STRIPS Tasks.

Initial World Model

$(\forall x \forall y \forall z)$[CONNECTS$(x,y,z) \Rightarrow$CONNECTS$(x,z,y)$]

CONNECTS(DOOR1,ROOM1,ROOM5)
CONNECTS(DOOR2,ROOM2,ROOM5)
CONNECTS(DOOR3,ROOM3,ROOM5)
CONNECTS(DOOR4,ROOM4,ROOM5)
LOCINROOM(f,ROOM4)
AT(BOX1,a)
AT(BOX2,b)
AT(BOX3,c)
AT(LIGHTSWITCH1,d)
ATROBOT(e)
TYPE(BOX1,BOX)
TYPE(BOX2,BOX)
TYPE(BOX3,BOX)
TYPE(D4,DOOR)
TYPE(D3,DOOR)
TYPE(D2,DOOR)
TYPE(D1,DOOR)
INROOM(BOX1,ROOM1)
INROOM(BOX2,ROOM1)
INROOM(BOX3,ROOM1)
INROOM(ROBOT,ROOM1)
INROOM(LIGHTSWITCH1,ROOM1)
PUSHABLE(BOX1)
PUSHABLE(BOX2)
PUSHABLE(BOX3)
ONFLOOR
STATUS(LIGHTSWITCH1,OFF)
TYPE(LIGHTSWITCH1,LIGHTSWITCH)

Operators

goto1(*m*): Robot goes to coordinate location *m*.
Preconditions:
(ONFLOOR) $\wedge$ $(\exists x)$[INROOM(ROBOT,*x*) $\wedge$ LOCINROOM(*m*,*x*)]
Delete list: ATROBOT($),NEXTTO(ROBOT,$)
Add list: ATROBOT(*m*)

goto2(*m*): Robot goes next to item *m*.
Preconditions:
(ONFLOOR) $\wedge$ {$(\exists x)$[INROOM(ROBOT,*x*) $\wedge$ INROOM(*m*,*x*)] $\vee$ $(\exists x)(\exists y)$ [INROOM(ROBOT,*x*) $\wedge$ CONNECTS(*m*,*x*,*y*)]}
Delete list: ATROBOT($),NEXTTO(ROBOT,$)
Add list: NEXTTO(ROBOT,*m*)

pushto(*m*,*n*): robot pushes object *m* next to item *n*
Precondition:
PUSHABLE(*m*) $\wedge$ ONFLOOR $\wedge$ NEXTTO(ROBOT,*m*) $\wedge$ {$(\exists x)$[INROOM(*m*,*x*) $\wedge$ INROOM(*n*,*x*)] $\vee$ $(\exists x,\exists y)$[INROOM(*m*,*x*) $\wedge$ CONNECTS(*n*,*x*,*y*)]}
Delete list: AT ROBOT ($) NEXTTO (ROBOT $) NEXTTO ($,*m*)
AT (*m*$) NEXTTO (*m*$)
Add list: NEXTTO(*m*,*n*)
NEXTTO(*n*,*m*)
NEXTTO(ROBOT,*m*)

Embodiment

behavior is not in the brain

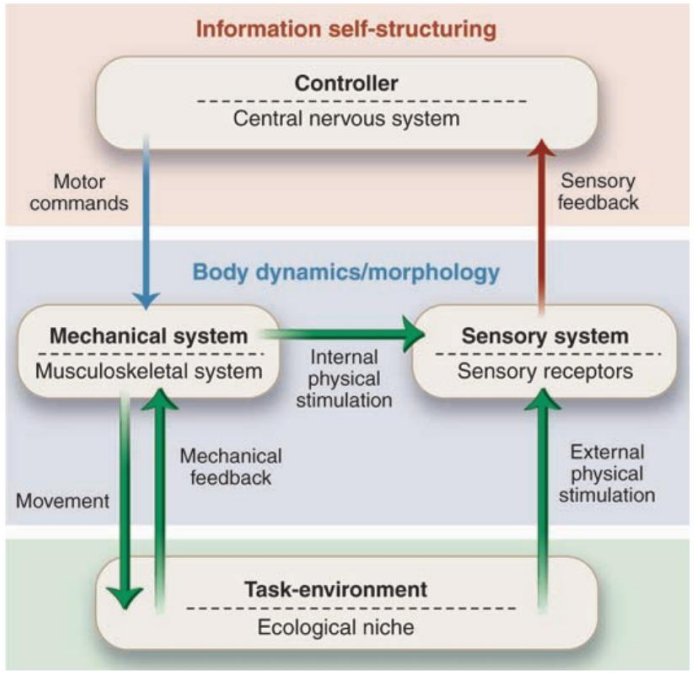
(or cell, molecule…)

cf. "all behavior is a result of brain function"

Eric R. Kandel, Ch. 1: The Brain and Behavior, in Kandel, E.R., Schwartz, J.H. and Jessell, T.M. eds., 2000. Principles of neural science (Vol. 4, pp. 1227-1246)

it is in the interaction

Illustrations: Shun Iwasawa, from R. Pfeifer & J. Bongard: How the body shapes the way we think, 2007



Pfeifer, R., Lungarella, M., & Iida, F. (2007). Self-organization, embodiment, and biologically inspired robotics. Science.

Research questions

Classical AI

- Thinking, reasoning, abstract problem solving

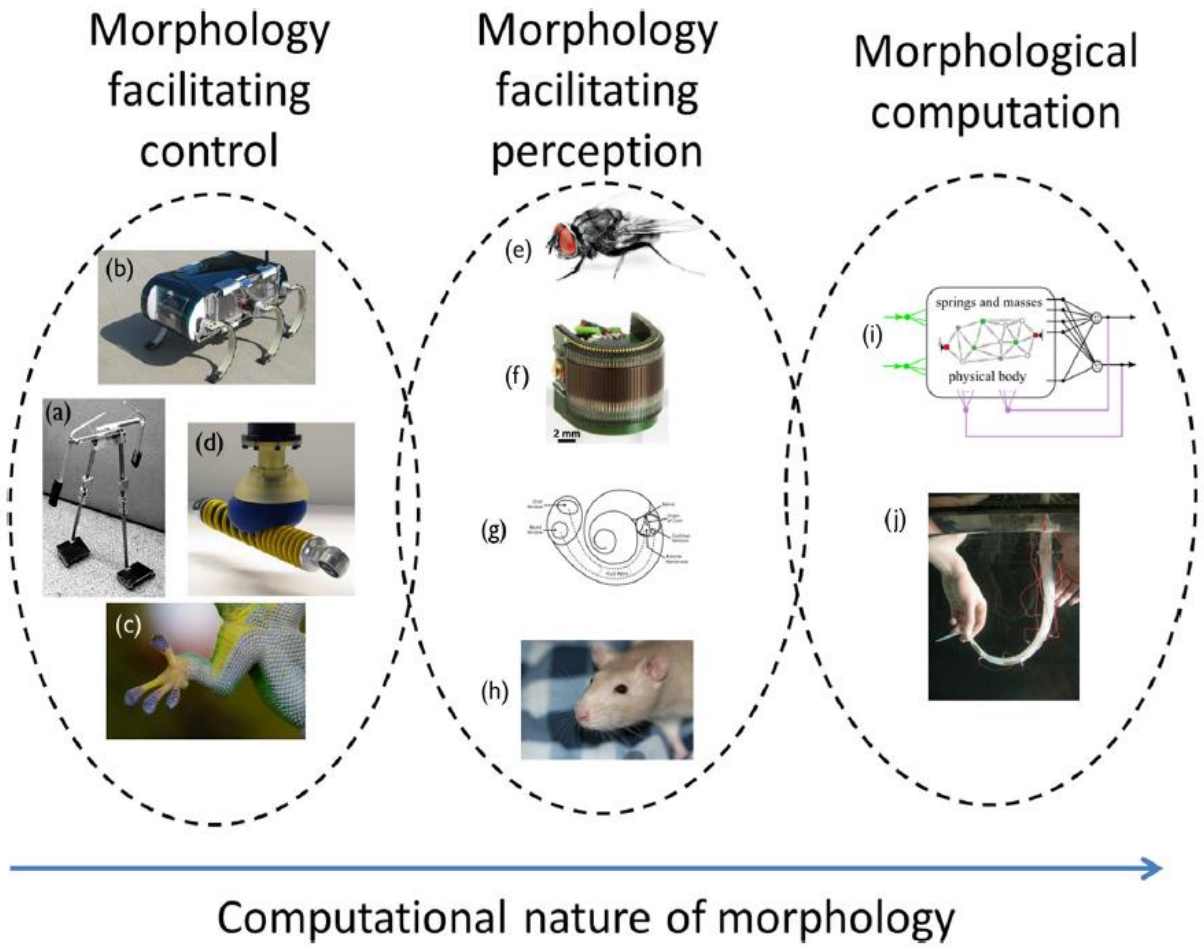
Embodied Intelligence

- Movement, physical interaction with the real world

"Why do plants not have brains? The answer is actually quite simple: they don't have to move."

Lewis Wolpert, UCL

Morphological “computation”



Müller, V.C. & Hoffmann, M. (2017), 'What is morphological computation? On how the body contributes to cognition and control', *Artificial Life* **23** (1), 1-24.

Physical implications of embodiment

~ morphology facilitating control

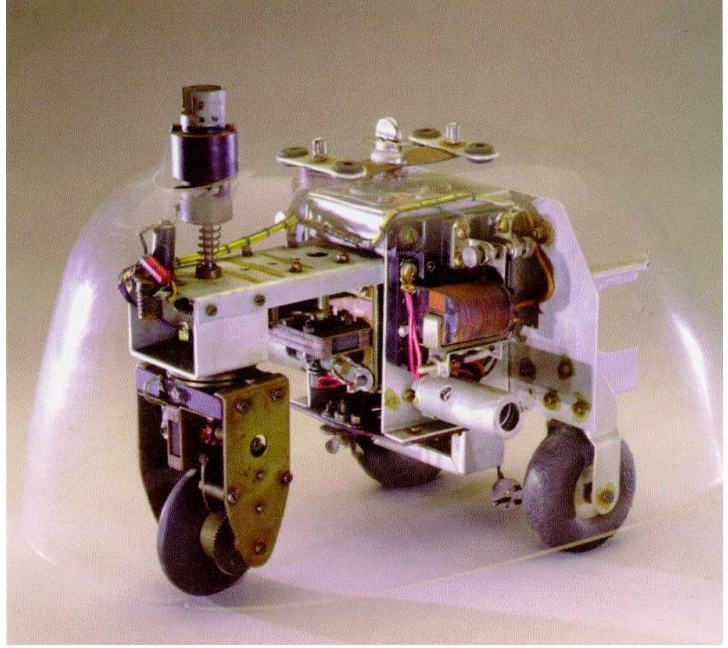
Is brain/computation needed for walking?

Passive dynamic walkers (McGeer 1990)

- "pure physics walking"
- No computer
- No motors
- No sensors

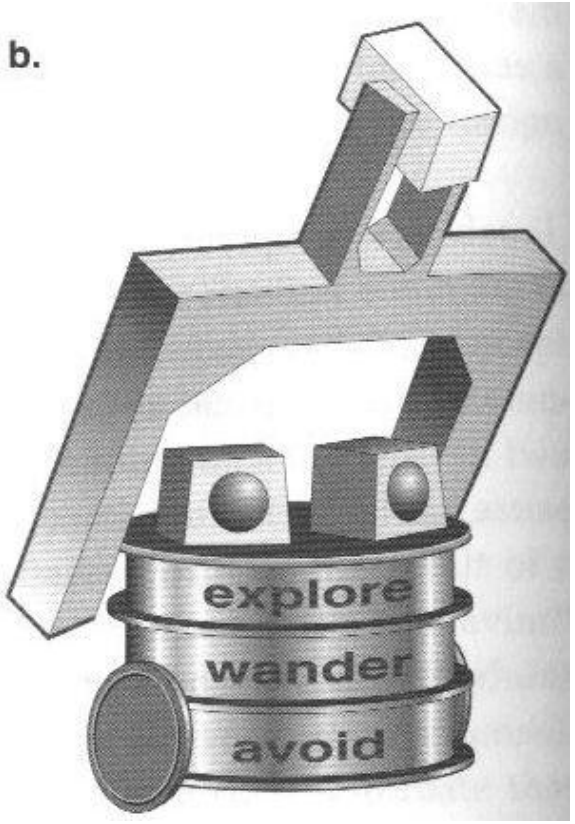
Morphology:
- shape of feet
- counterswing of arms
- friction on bottom of feet

Cornell PDW with arms, Collins et al. 2001



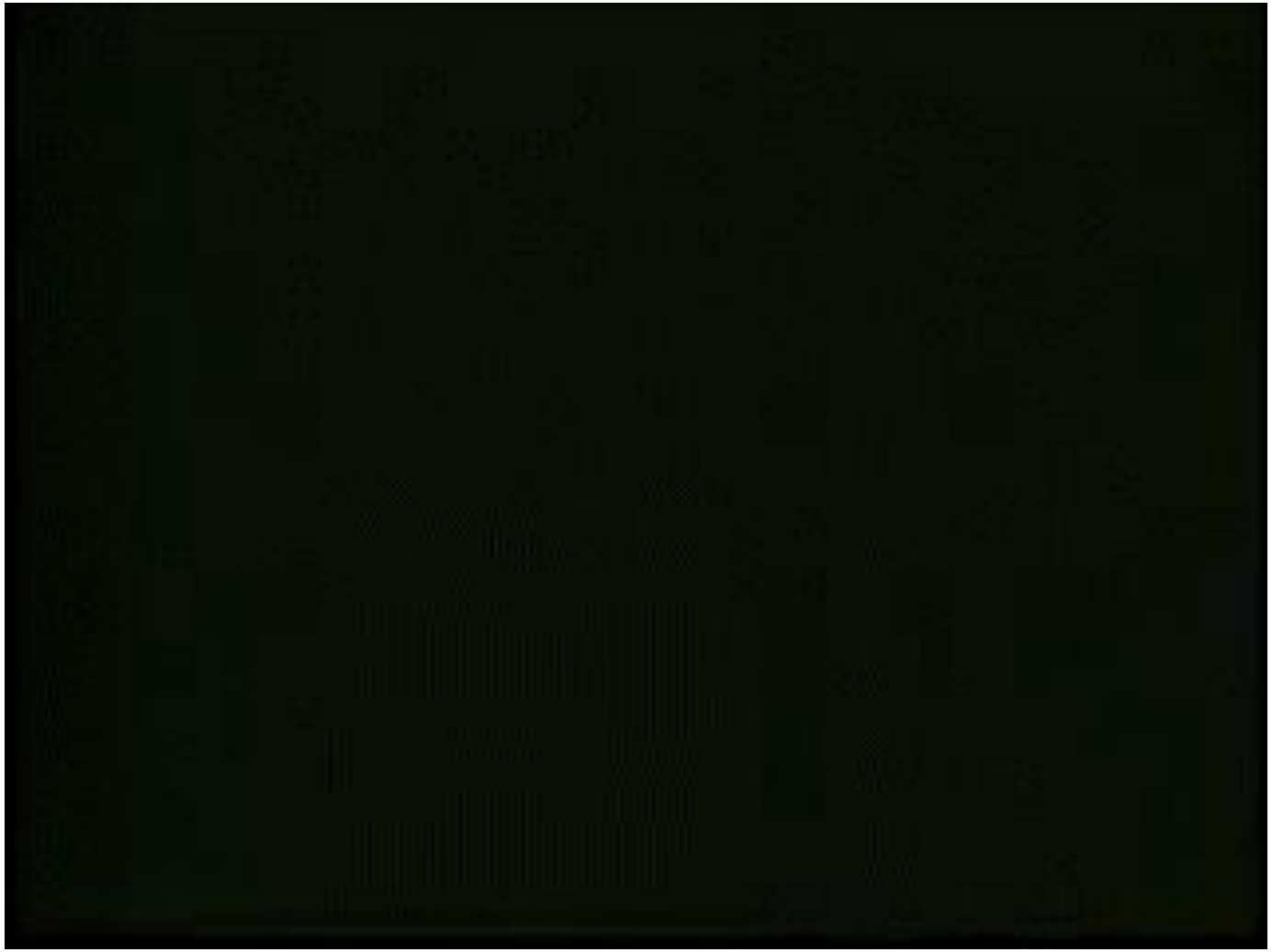
Grey Walter
Tortoises, 1940s

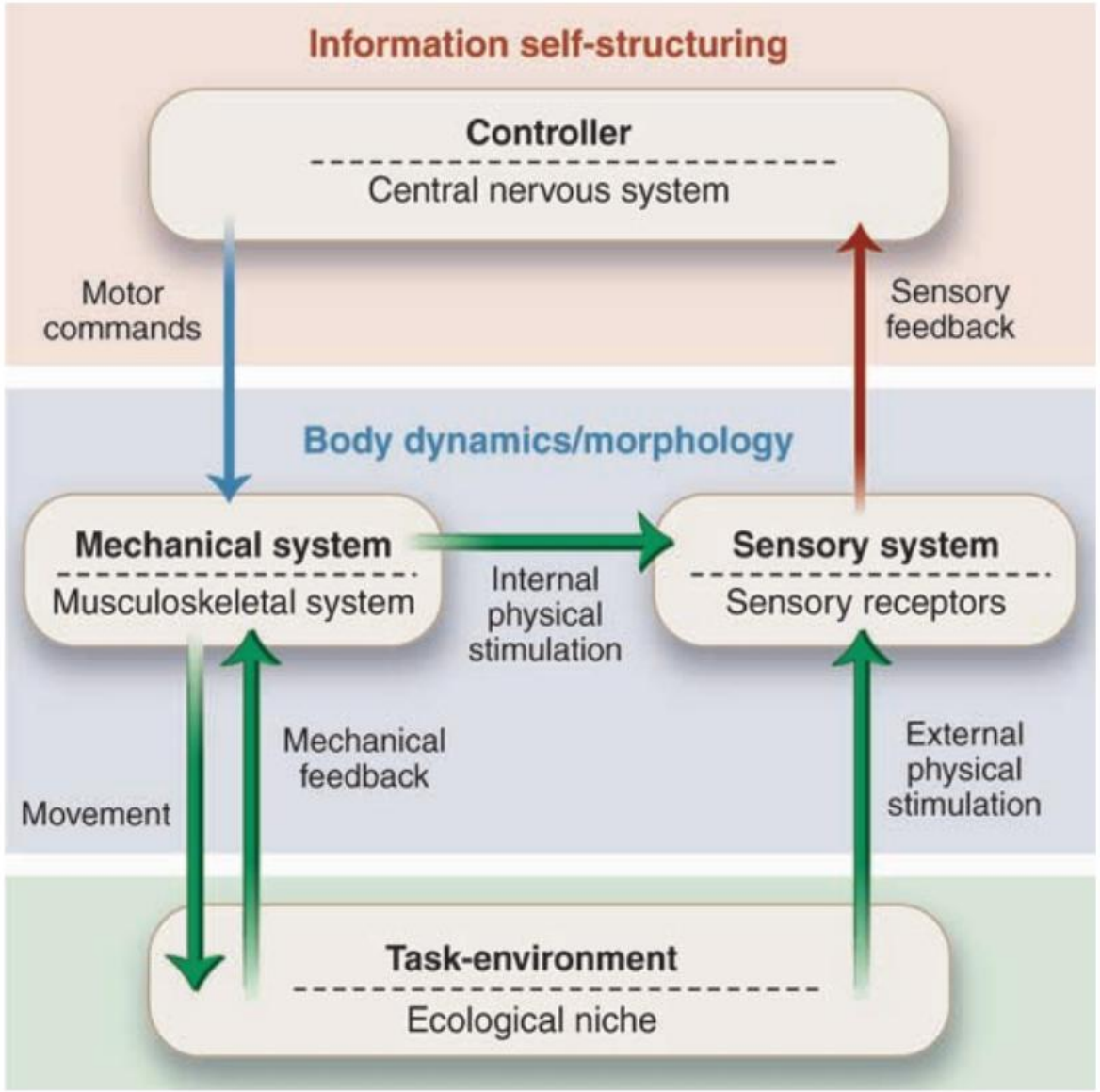
V. Breitenberg, 1980s



R. Brooks, 1980s
subsumption
architecture

Grey Walter’s tortoises





Pfeifer, R., Lungarella, M., & Iida, F. (2007). Self-organization, embodiment, and biologically inspired robotics. Science.

Behavior-based robotics manifestos

Intelligence without representation*

Rodney A. Brooks

MIT Artificial Intelligence Laboratory, 545 Technology Square, Rm. 836, Cambridge, MA 02139, USA

Received September 1987

Brooks, R.A., Intelligence without representation, Artificial Intelligence 47 (1991), 139–159.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

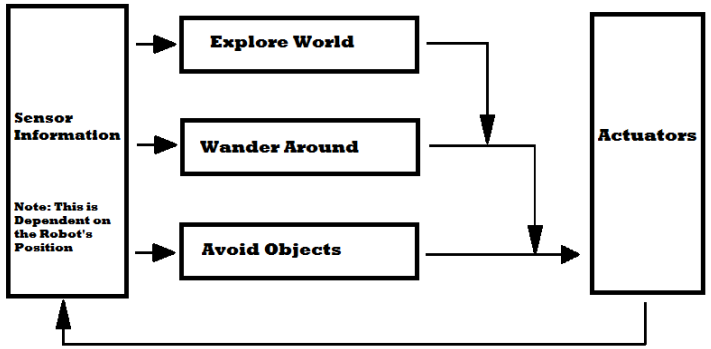
A.I. Memo No. 1293 April, 1991

Intelligence Without Reason

Rodney A. Brooks

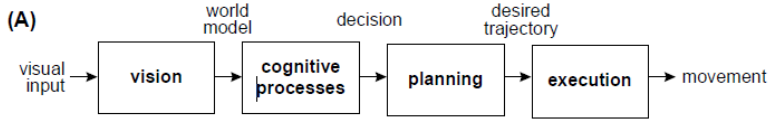
Prepared for *Computers and Thought*, IJCAI-91

Parallel loosely coupled processes

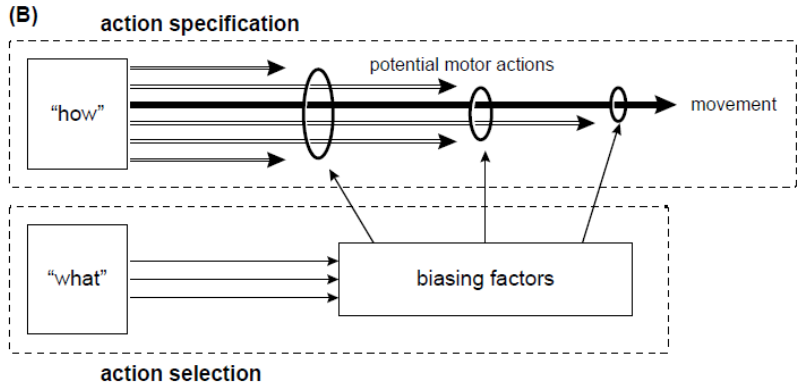


Subsumption architecture

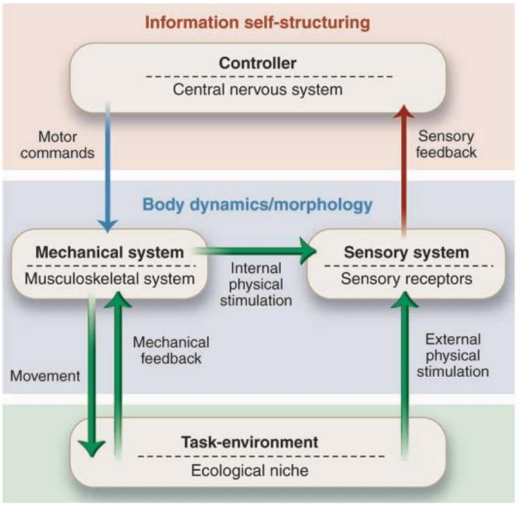
Brooks, Rodney (1999). Cambrian Intelligence: The Early History of the New AI. Cambridge, Massachusetts: The MIT Press.



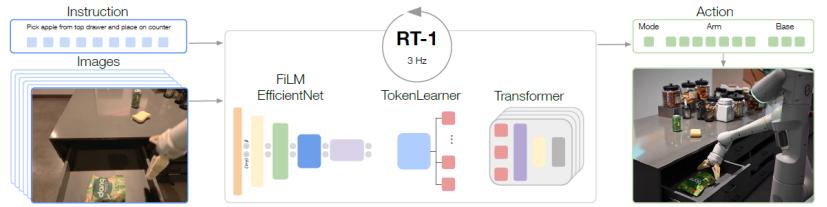
Cisek, P., & Kalaska, J. F. (2003). Reaching movements: implications for computational models. *Handbook of Brain Theory and Neural Networks*, 945-948.



How embodied is this embodied AI?



- Rt-2: Vision-language-action models **transfer web knowledge to robotic control**
- Reminiscent of Good Old-Fashioned AI (GOFAI, ~ 1950s-1970s)
- Will it inherit its problems?



(a) RT-1 takes images and natural language instructions and outputs discretized base and arm actions. Despite its size (35M parameters), it does this at 3 Hz, due to its efficient yet high-capacity architecture: a FiLM (Perez et al., 2018) conditioned EfficientNet (Tan & Le, 2019), a TokenLearner (Ryoo et al., 2021), and a Transformer (Vaswani et al., 2017).

Coming up as a chapter in Macrine, S.L., Fugate, J.M.B., Abdulali, A. and Hughes, J. (eds.) (2026): Embodied Intelligence, MIT Press.

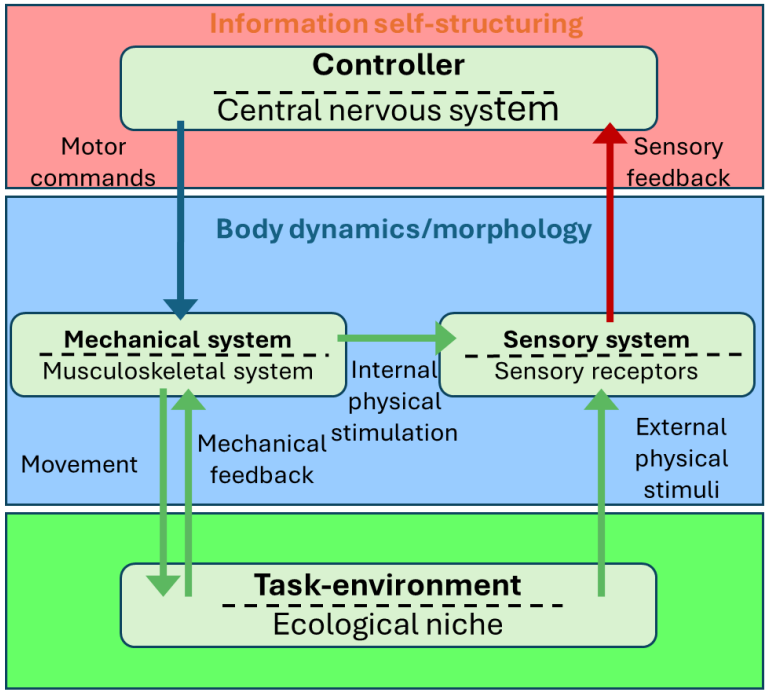
Embodied AI in Machine Learning – is it Really Embodied?

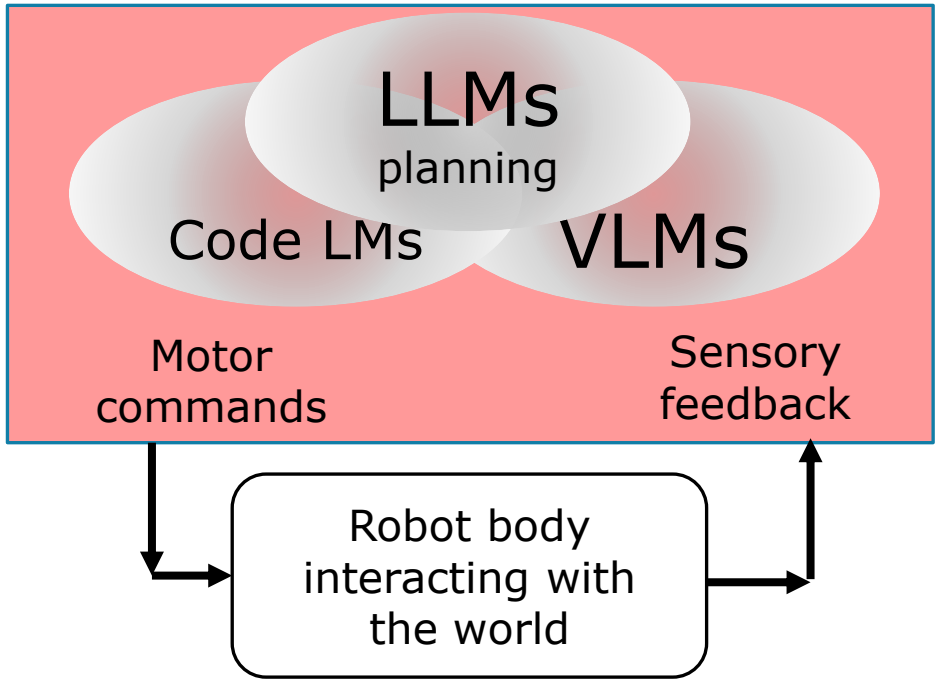
Matej Hoffmann and Shubhan Parag Patni

https://arxiv.org/abs/2505.10705

“Embodied AI” is weakly embodied.

I call it Weakly Embodied AI (WEAI).



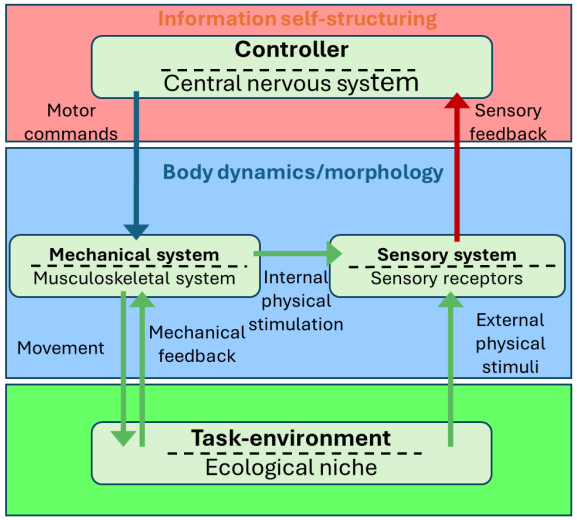


Embodiment checklist (1)

Embodied intelligence

0. Behavior is not in the brain/controller but in the **closed-loop interaction of the controller, the body, and the environment**.

1. **Body morphology facilitates control**.



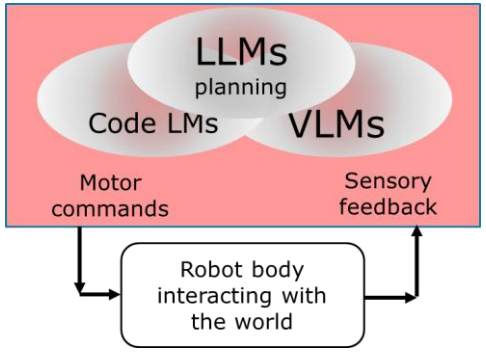
"Embodied AI"

0. **Behavior is in the controller—in the large model.**

✗

1. There is little room for the agent body to be exploited to fulfill the task. The **robot platforms** used—like mobile robots or robot arms with 2-finger grippers—**do not feature rich body dynamics** that could be exploited for the task. Moreover, the **details of** the sensory and **motor interfaces** of the robots **are abstracted away**.

✗

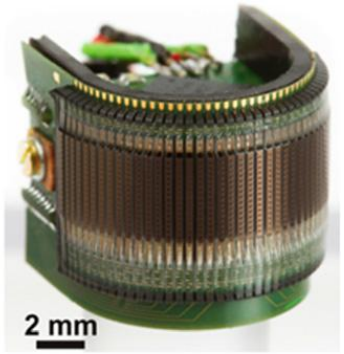


Embodiment checklist (2)

Embodied intelligence

2. **Sensor morphology facilitates perception.**

(a)



(b)

CurvACE – artificial compound eye - image courtesy of Dario Floreano

"Embodied AI"

2. WEAI leverages the power of deep neural networks to find patterns in static images (e.g., object recognition) and how what is on the images can be translated into text (VLMs). The inputs need to be more or less **standard RGB images.**

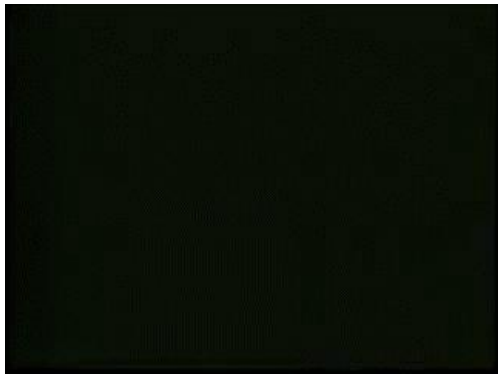
The power of Convolutional Neural Networks (CNNs) and Visual Transformers on tasks like recognition will hardly transfer to particular sensor morphologies like with non-uniform spacing of light sensitive cells.

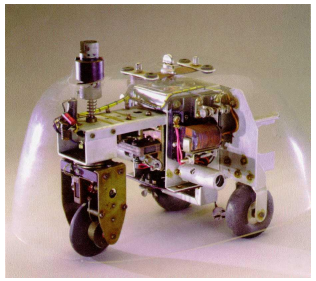
Thus, sensor morphologies designed for a particular task environment are not compatible with the WEAI approach.

Embodiment checklist (3)

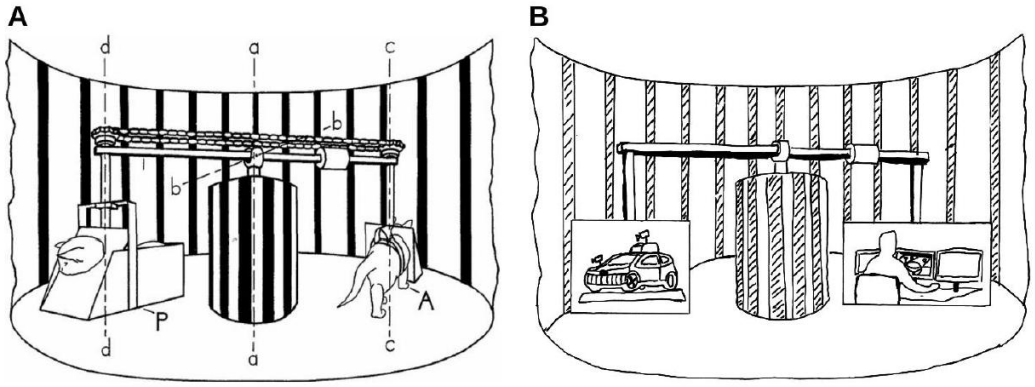
Embodied intelligence

3. **Sensorimotor coordination and active perception.**





Grey Walter Tortoises, 1940s



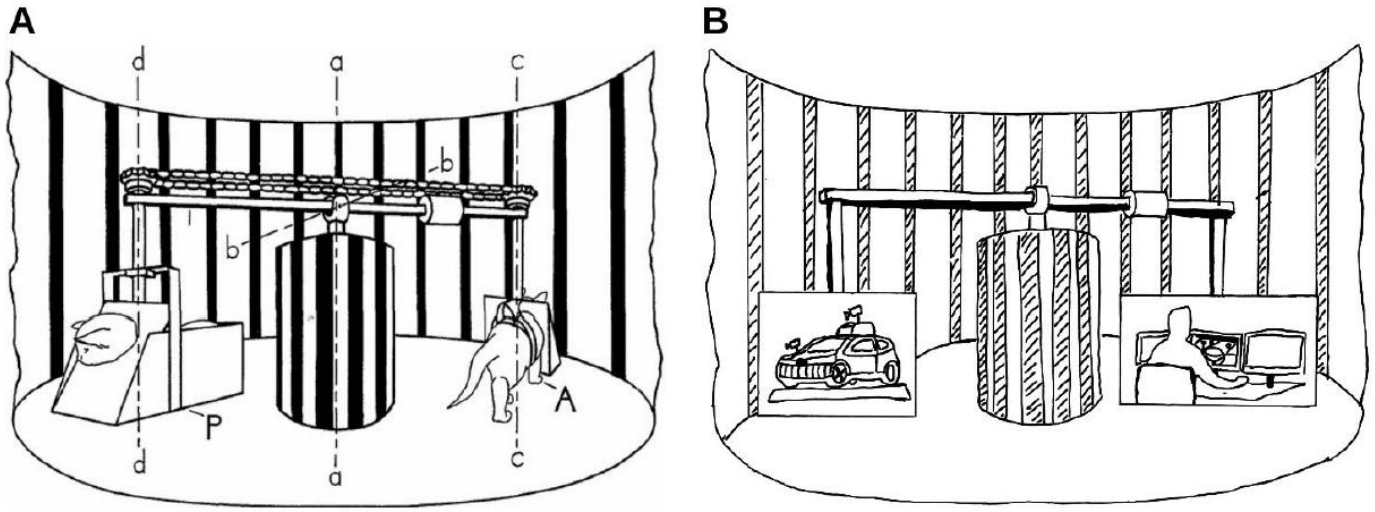
"Embodied AI"

3. Typical WEAI architectures aim to maximally exploit the "common sense" of LLMs to solve tasks without the need for additional training—also called "zero shot". In this case, there is no possibility to exploit the closed-loop embodied interaction with the environment.

✘

When additional training (finetuning) of the model takes place, it typically uses offline datasets collected when for example a human was **teleoperating the robot** to solve the task.
The **operator may exploit sensorimotor coordination** to solve the task and this would be transferred to the robot controller. However, **there has to be a match in the situatedness**, i.e. the human and the robot will have to see through similar "eyes". If the operator drives the robot using for example a third person view but the robot has an egocentric camera, the transfer will not work.

Active embodied interaction vs. offline learning



Trends in
Cognitive Sciences

CellPress
OPEN ACCESS

Opinion

Generating meaning: active inference and the scope and limits of passive AI

Giovanni Pezzulo [1,*] Thomas Parr,[2] Paul Cisek,[3] Andy Clark,[4,5,6] and Karl Friston[7,8]

Hoffmann and Patni (2026): Embodied AI in Machine Learning – is it Really Embodied? https://arxiv.org/abs/2505.10705
(coming up in Embodied Intelligence, MIT Press, 2026)

Generalization across robot bodies?

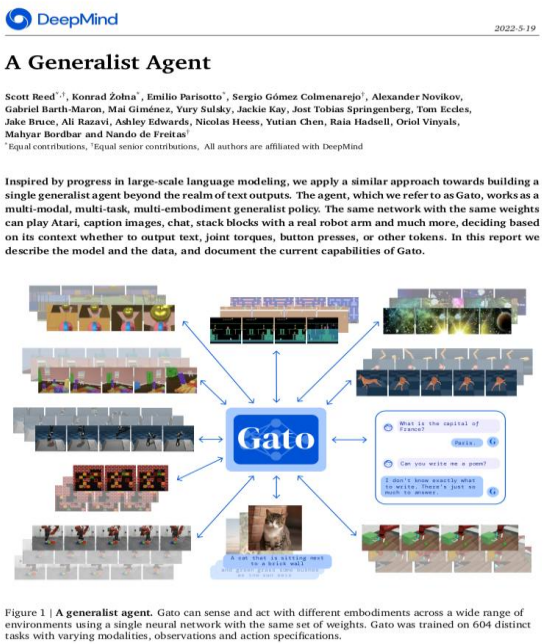


Figure 1 | **A generalist agent.** Gato can sense and act with different embodiments across a wide range of environments using a single neural network with the same set of weights. Gato was trained on 604 distinct tasks with varying modalities, observations and action specifications.

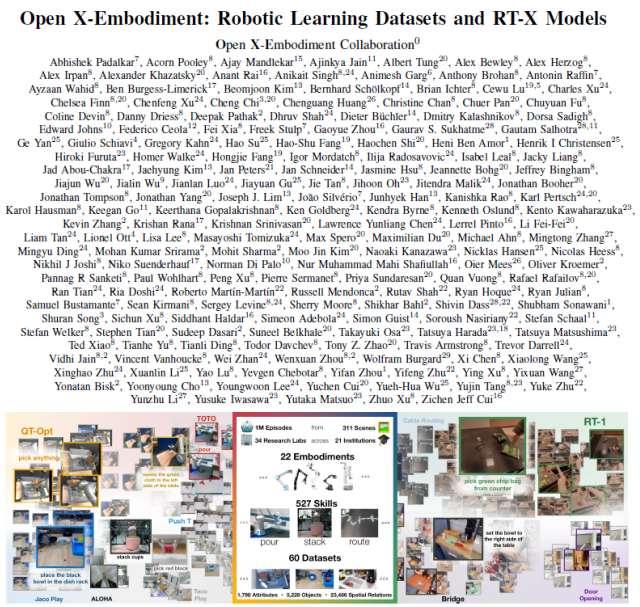


Fig. 1: We propose an open, large-scale dataset for robot learning curated from 21 institutions across the globe. The dataset represents diverse behaviors, robot embodiments and environments, and enables learning generalized robotic policies.

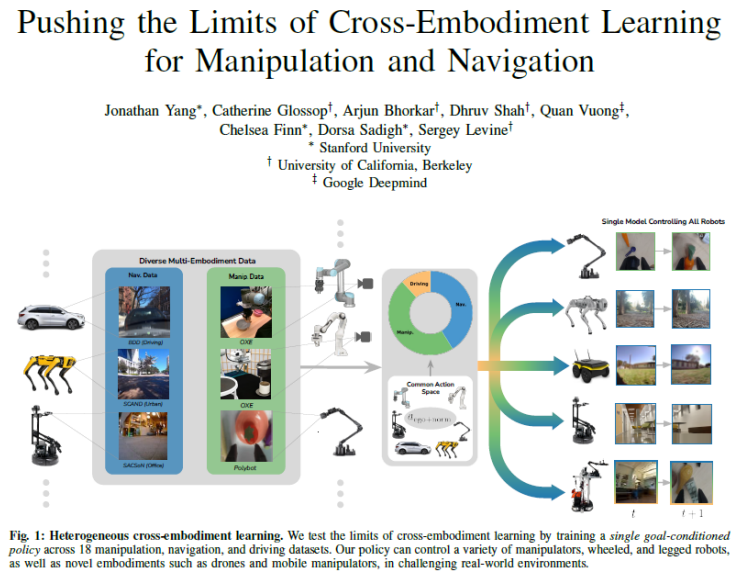


Fig. 1: Heterogeneous cross-embodiment learning. We test the limits of cross-embodiment learning by training a *single goal-conditioned policy* across 18 manipulation, navigation, and driving datasets. Our policy can control a variety of manipulators, wheeled, and legged robots, as well as novel embodiments such as drones and mobile manipulators, in challenging real-world environments.

1. the models/controllers are based on a transformer network
2. they were trained offline in a supervised manner using imitation learning
3. the action space consisted of discretized target positions and orientations for the robot (end effector) in the Cartesian space or rates of these quantities, plus gripper opening

How come the cross-embodiment controllers work?

- **the embodiment there is *shallow***
- abstraction layer that reduced every robot to
  - Cartesian coordinates of its gripper (manipulator)
  - the coordinates of the robot's center of mass (mobile robot)
- models were trained to learn a mapping from camera images to a low-dimensional action space, conditioned on a goal expressed in
  - language ("pick apple from ... and place on ..." (Padalkar et al., 2023))
  - a goal image (Yang et al., 2024).
- The fact that the robot has joints is abstracted away.
- The control has a coarse spatial (discretized low-dim. action space) and temporal resolution (3-10 Hz). Interaction forces are not considered.
- **This is not "real robotics" or "real embodiment".**
- **What are the real tasks one can do with this representation?**

Is this only a matter of getting relevant data from more robots and making a bigger model or are there principled roadblocks?

Fundamental problems of cross-embodiment

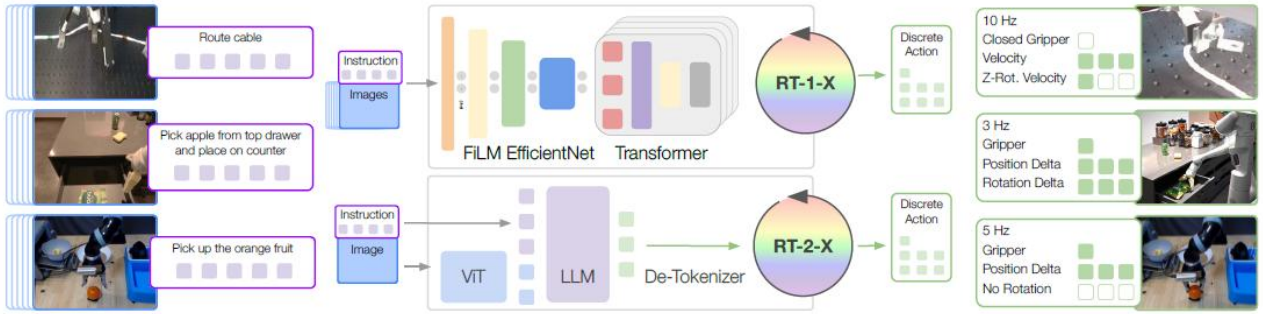
1. More profound embodiment implies larger models and less positive transfer between embodiments.
   - Cartesian space to joint space?

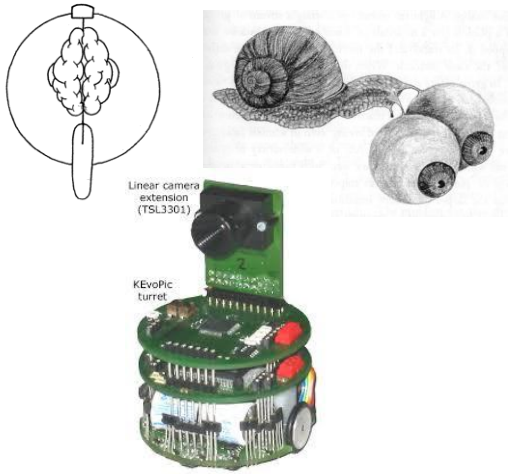
2. Principle of ecological balance: The complexity of the agent's sensory, motor, and neural systems should match.

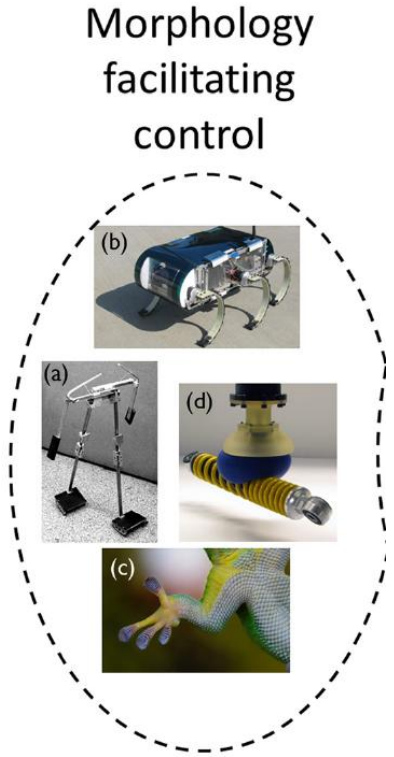
Current cross-embodiment learning:

- gigantic brain shared across embodiments, irrespective of the agent's complexity.
- mismatch between the dimensionality of the input space (raw RGB images, i.e. high-dimensional input) and the output space (7-dimensional discretized action space).

3. Cross-embodiment prevents exploitation of embodiment.

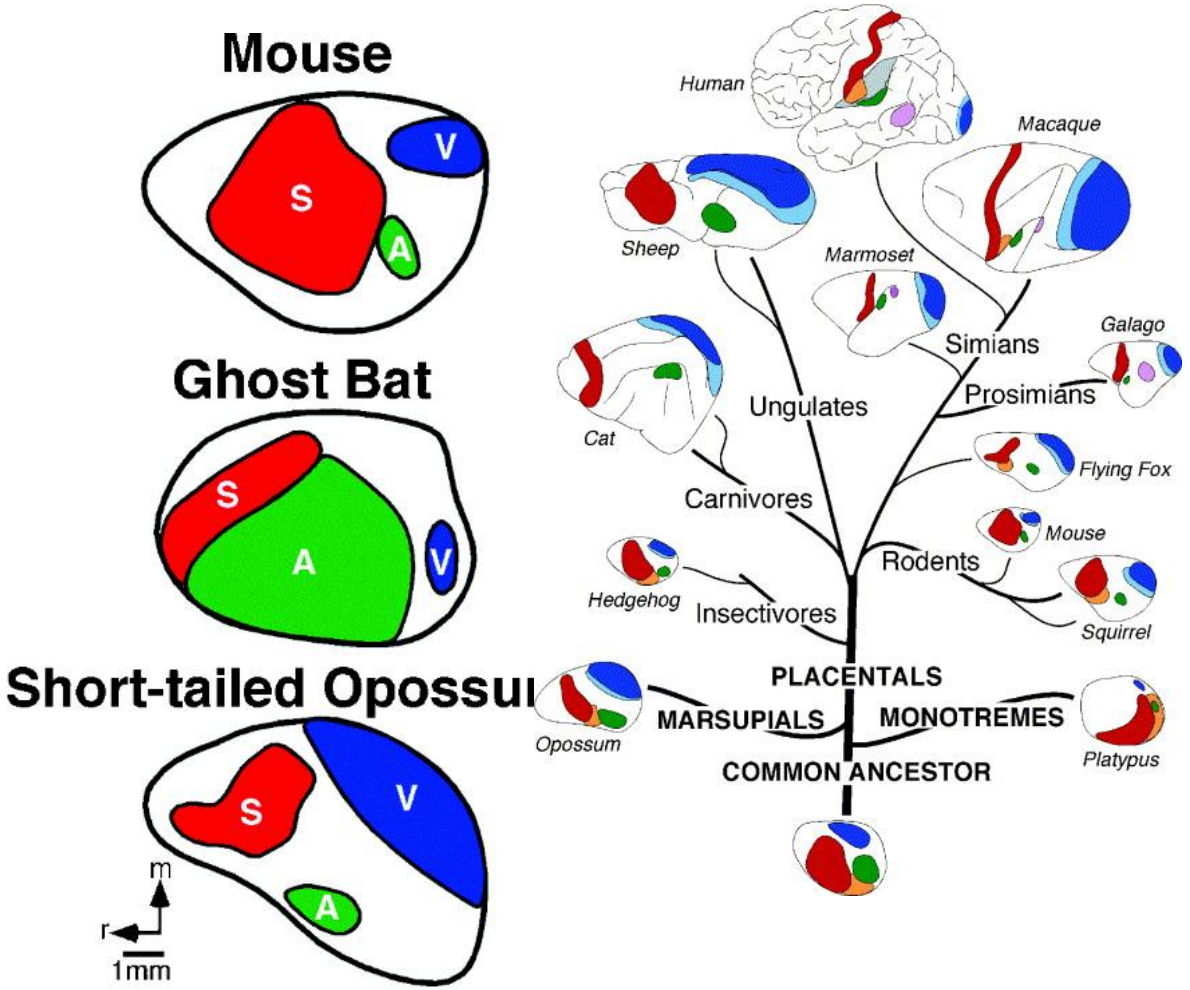






Muller & Hoffmann 2017

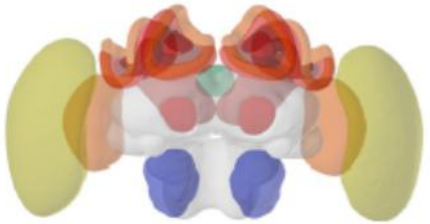
Biological intelligence – different bodies yield different brains and controllers



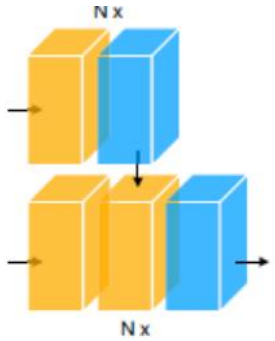
Krubitzer, L., & Kahn, D. M. (2003). Nature versus nurture revisited: an old idea with a new twist. *Progress in neurobiology*, *70*(1), 33-52.

Transformers – one brain rules them all - new kind of intelligence?

- Scaling problems – to further improve performance, exponential increase in model & compute for training and inference may be needed.
- Hugely inefficient compared to biology.
- Universal "flat" brain structure.
- Embodiment not exploited.



honey bee brain

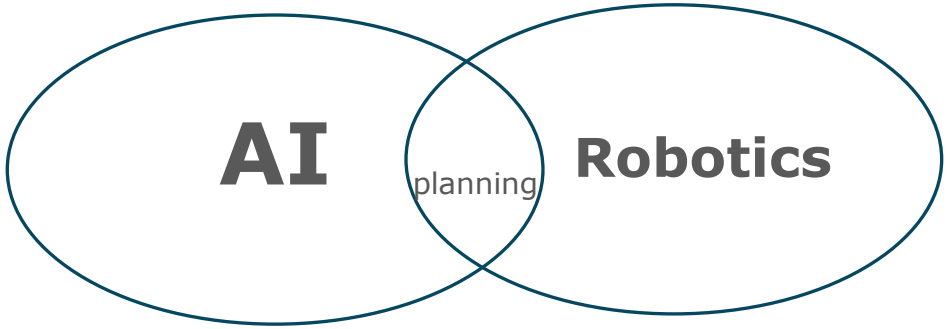


arshall & Barron (2025): Are transformers truly oundational for robotics?

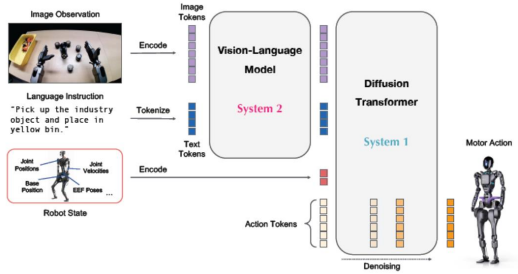
Brooks (1989, 1991)

Physical AI = Robotics?

Up until recently.



Hot topic now.



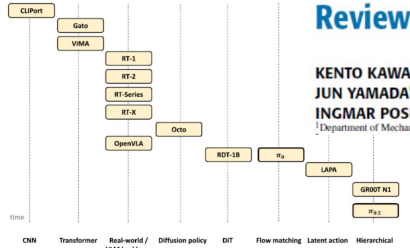


Figure 1: **Data Pyramid for Robot Foundation Model Training.** GR00T N1's heterogeneous training corpora can be represented as a pyramid: data quantity decreases, and embodiment-specificity increases, moving from the bottom to the top.

Vision-Language-Action Models for Robotics: A Review Towards Real-World Applications

KENTO KAWAHARAZUKA[1], (Member, IEEE), JIHOON OH[1],
JUN YAMADA[2], (Graduate Student Member, IEEE),
INGMAR POSNER[2], (Member, IEEE), AND YUKE ZHU[3], (Senior Member, IEEE)
[1]Department of Mechano-Informatics, The University of Tokyo, Tokyo 113-8656, Japan

Future?

Gr00t n1: An open foundation model for generalist humanoid robots. 2025. https://arxiv.org/abs/2503.14734

Summary and possible exam topics

Motivation
Why data-driven?

History
LLM→VLM→VLA arc

Architectures
VLA taxonomy

Data
Collection & scale

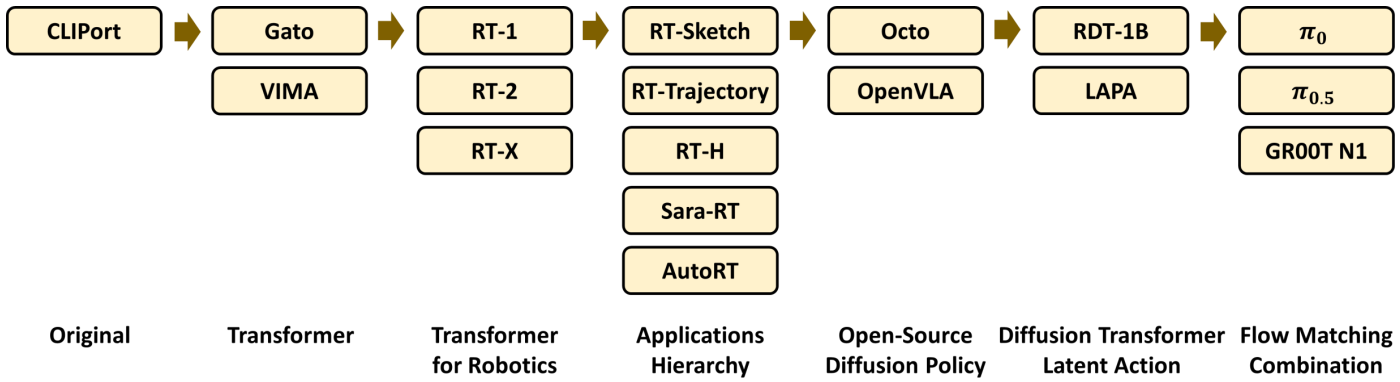


Figure 1: **Data Pyramid for Robot Foundation Model Training.** GR00T N1's heterogeneous training corpora can be represented as a pyramid: data quantity decreases, and embodiment-specificity increases, moving from the bottom to the top.

Resources & references

Online talks:

- R. Tedrake (TRI): Multitask transfer in TRI's large behavior models for dexterous manipulation," Stanford seminar, 25 April 2025; available at https://youtube.com/watch?v=TN1M6vg4CsQ
- Ken Goldberg (UC Berkeley, Ambi Robotics): What robots can (& can't) do in 2025; https://youtu.be/94v3zRfBQXQ?si=jNAq6KnAvVfm8CvK

Articles

- Amato, N. M., Hutchinson, S., Garg, A., Billard, A., Rus, D., Tedrake, R., ... & Goldberg, K. (2025). "Data will solve robotics and automation: True or false?": A debate. *Science Robotics*, *10*(105), eaea7897.
- Goldberg, K. (2025). Good old-fashioned engineering can close the 100,000-year "data gap" in robotics. *Science Robotics*, *10*(105), eaea7390.
- Kawaharazuka, K., Oh, J., Yamada, J., Posner, I., & Zhu, Y. (2025). Vision-language-action models for robotics: A review towards real-world applications. *IEEE Access*. https://ieeexplore.ieee.org/abstract/document/11164279