From Seeing to Acting to Imagining: The Evolution of Robot Foundation Models

Bryce Grant

Third Year PhD Student, Robotics

NSF GRFP Fellow

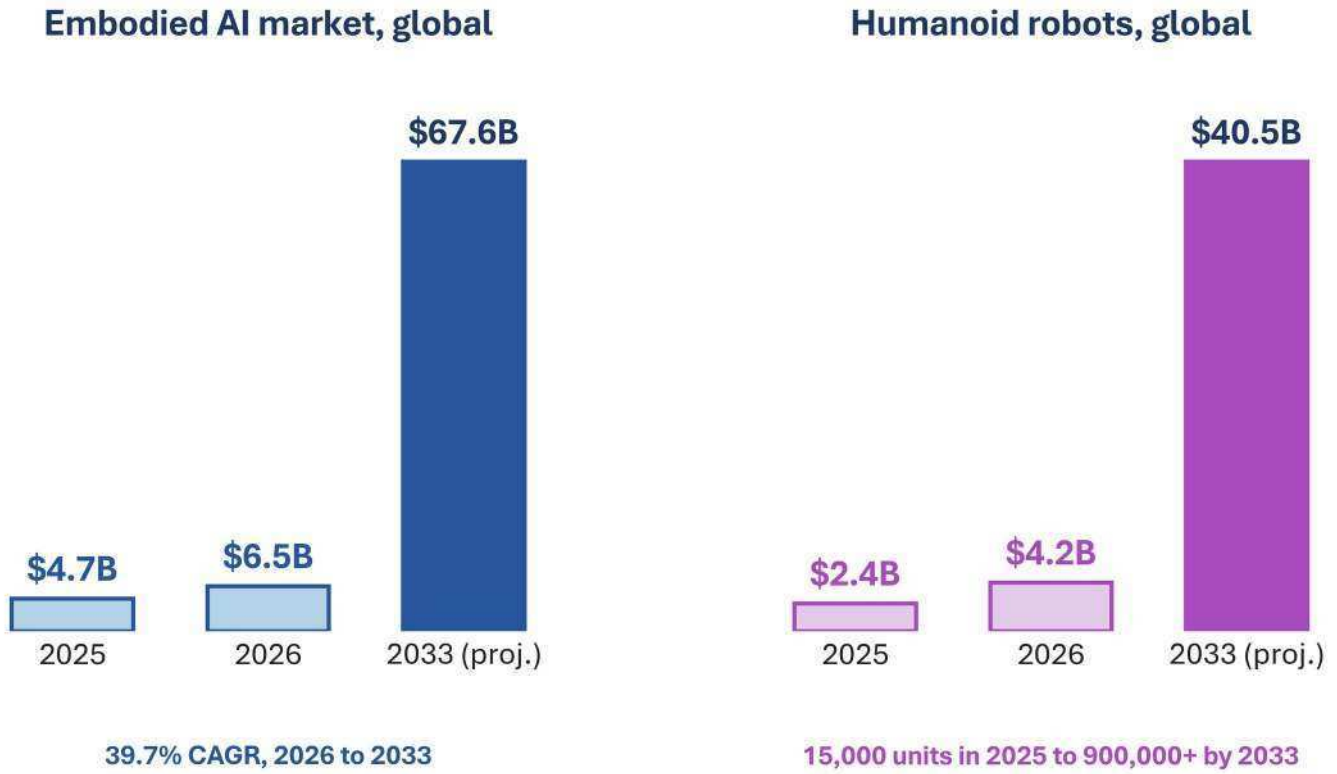
Graduate Advisors: Dr. Peng Wang, Dr. Zonghe Chua

CASE WESTERN RESERVE UNIVERSITY

Full slides with videos (.pptx): click here

AI Robotics Market Size

- **$6.5B** embodied AI market in 2026, projected **$67.6B** by 2033
- Humanoids: **15,000** units in 2025, **900,000+** by 2033
- H1 2026 alone: ~25,000 humanoids shipped, 95% built in China, 69% bound for labs and classrooms



[83] [84] Grand View Research; [87] Robot News Today

Key Players: Commercial Humanoids



Unitree G1 $13.5k, R1 $4.9k, H2 $29.9k.



Apptronik Apollo 2 at Mercedes-Benz, GXO, Jabil. Gemini Robotics partner.



Figure 03 with the Helix VLA. Paid work at BMW Spartanburg, about $25 per robot-hour.

Tesla Optimus

- Internal only: All units for Tesla factories and data collection

Agility Digit

- Digit 4: 65,000+ hours at GXO, Amazon, Schaeffler, Toyota
- Digit 5: $300M+ in orders, early access H1 2027

Boston Dynamics Atlas

- Electric Atlas, 56 DoF, runs Gemini Robotics with DeepMind
- Training at Hyundai's Georgia Metaplant

1X NEO

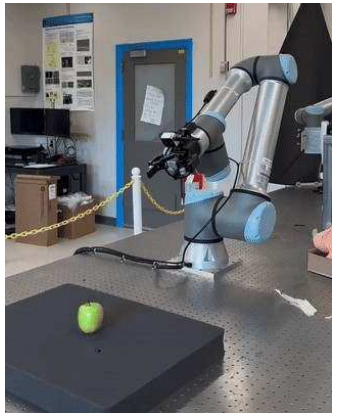
- Home humanoid: $20k or $499/mo, Hayward factory

[85] Synthra; [86] Inside Deep Tech; [87] Robot News Today; [91] [92] Robotica Guide

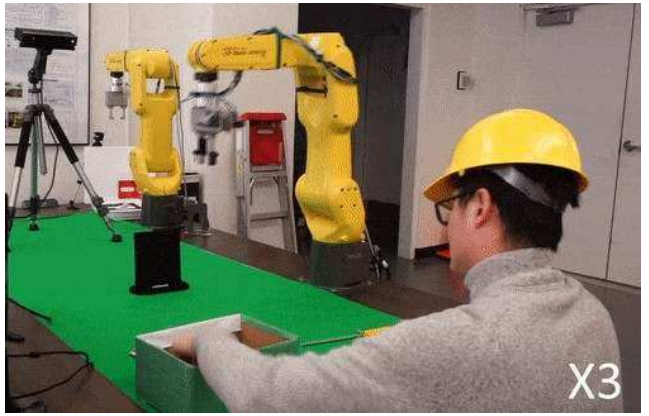
Key Players: Co(llaborative) Robots



Agile Robots / Franka:
FR3 & Panda arm (GDM & Meta Partner), Agile One
Foundation model development



Universal Robots
UR5e to UR15
100,000+ cobots installed



FANUC CRX family, 3 to 30 kg

Brain

- Skild Brain into ABB and UR portfolios
- Generalist demoed GEN-1 on a UR arm at GTC
- FANUC adds Google agents, Palladyne IQ

AI Powered Controllers

- UR AI Accelerator: Jetson AGX Orin + PolyScope X, ROS 2
- FANUC, ABB, KUKA, Yaskawa put Jetson in controllers
- AgileCore OS leveraging FCI data

Data

- UR AI Trainer with Scale AI: leader UR3e, follower UR7e, force + vision + motion
- Industrial dataset from 100k+ arms planned

The market

- North American cobot orders +55.6% YoY in Q1 2026
- ABB PoWa cobots and a Physical AI Toolchain

[76] Skild AI; [88] NVIDIA; [89] The Robotics Media; [90] Rocking Robots

Key Players: Foundation-Model Labs

- **Physical Intelligence:** π0 → π0.7. $5.6B
- **Skild AI:** omni-bodied Skild Brain, S1. $14B
- **Generalist AI:** GEN-0 → GEN-1.5, trained from scratch on 500k+ hours. $3B
- **Figure AI:** Helix VLA + humanoids. $39B
- **Apptronik:** Apollo humanoid. $5B
- **Genesis AI:** $3B. also 1X, Dyna, Field AI, Mimic, Rhoda, Sereact, Dexmate

Business models

- Body + brain: Figure, Apptronik, 1X, Agility
- Brain only, licensed to OEMs: Physical Intelligence, Skild (ABB, Universal Robots, Foxconn), Generalist

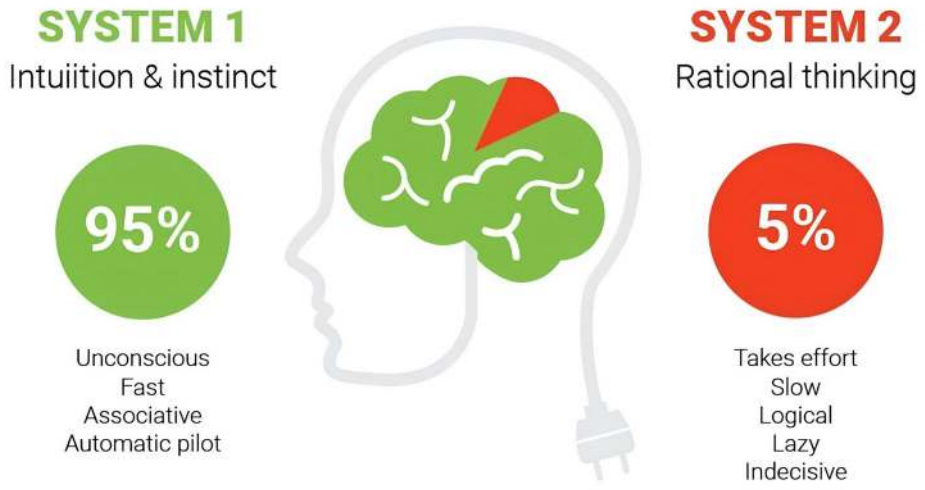
[73] RobotOps; [74] TechFundingNews; [68] Skild AI; [75] ValueAddVC

Talk Overview

1. **Foundations**
   - Transformers & Modality Encoding
2. **Vision-Language (VLMs)**
   - DINOv2, SAM
   - LLaVa, Flamingo
3. **Vision-Language-Action (VLAs)**
   - RT-1, RT-2, ReKep
   - Active Agents
4. **State of the Art ('24-Present)**
   - $\pi_{0.7}$, GR00T, DreamZero, Cosmos
   - Democratization
5. **Inside the Models**
   - VLA Interpretability
   - WAM Interpretability
6. **Limits and What's Next**
   - Planning, Tactile
   - Research Opportunities

Thinking, Fast and Slow







Why can robots do backflips, but fail to fold laundry or walk with human-like gait?

Digital Intelligence ≠ Physical Intelligence

[1] Nikankin, et al.

The Symbol Grounding Problem



Chess Program

“Knight“ = e2→ f4
No physical understanding

No grounding



VLM

“Dog“ = visual features
Can describe

PARTIAL Visual Grounding



VLA

“Grasp” = motor commands + force feedback

Physical Grounding

How do symbols acquire meaning through physical experience?

Attention is all you Need: Transformers

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Attention weights $a_{1:N}$ are query-key similarities:

$\hat{a}_i = q \cdot k_i$

Normalized via softmax: $a_i = \frac{e^{\hat{a}_i}}{\sum_i e^{\hat{a}_i}}$

Output z is attention-weighted average of values $v_{1:N}$:

$z = \sum_i \hat{a}_i v_i = \hat{a} \cdot v$

k and v are derived from the same input x:

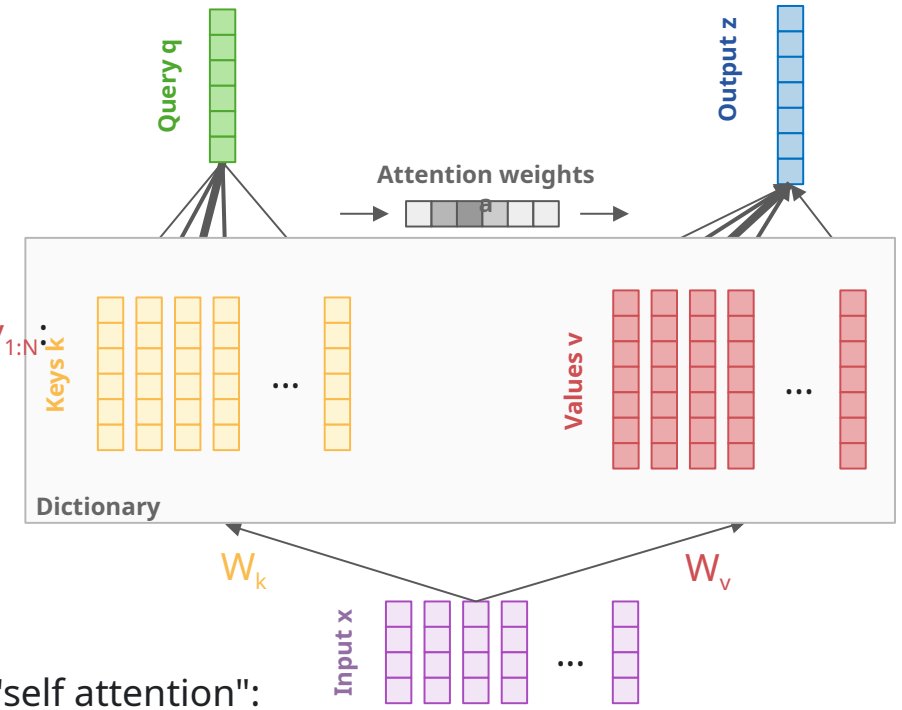
$k = W_k \cdot x$ $v = W_v \cdot x$

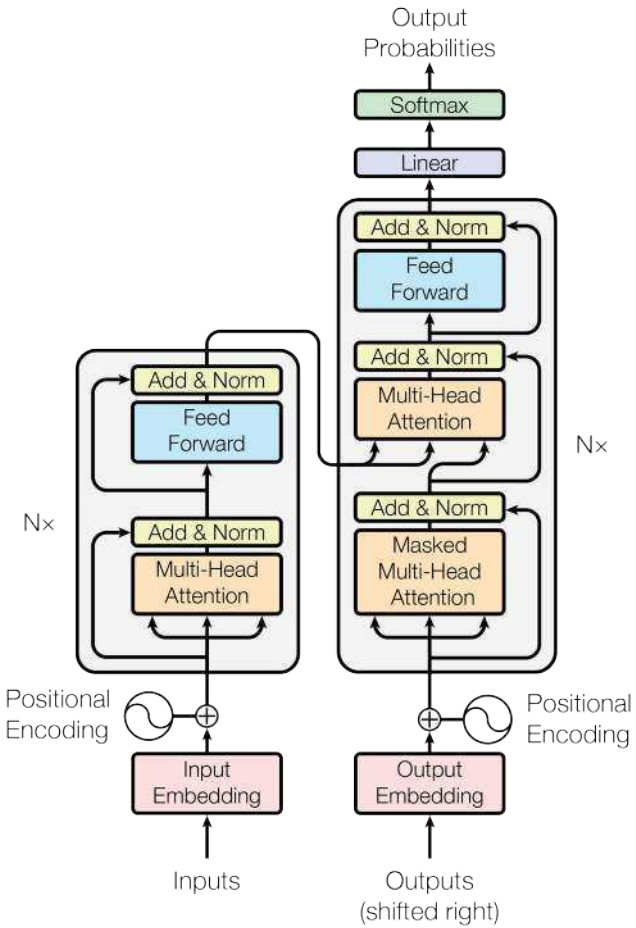
The query q can come from a separate input y:

$q = W_q \cdot y$

Or from the same input x. In which case we call it "self attention":

$q = W_q \cdot x$





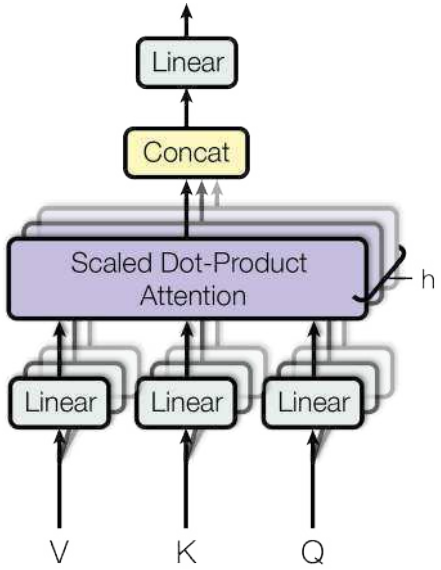
[4] Vaswani. et al.

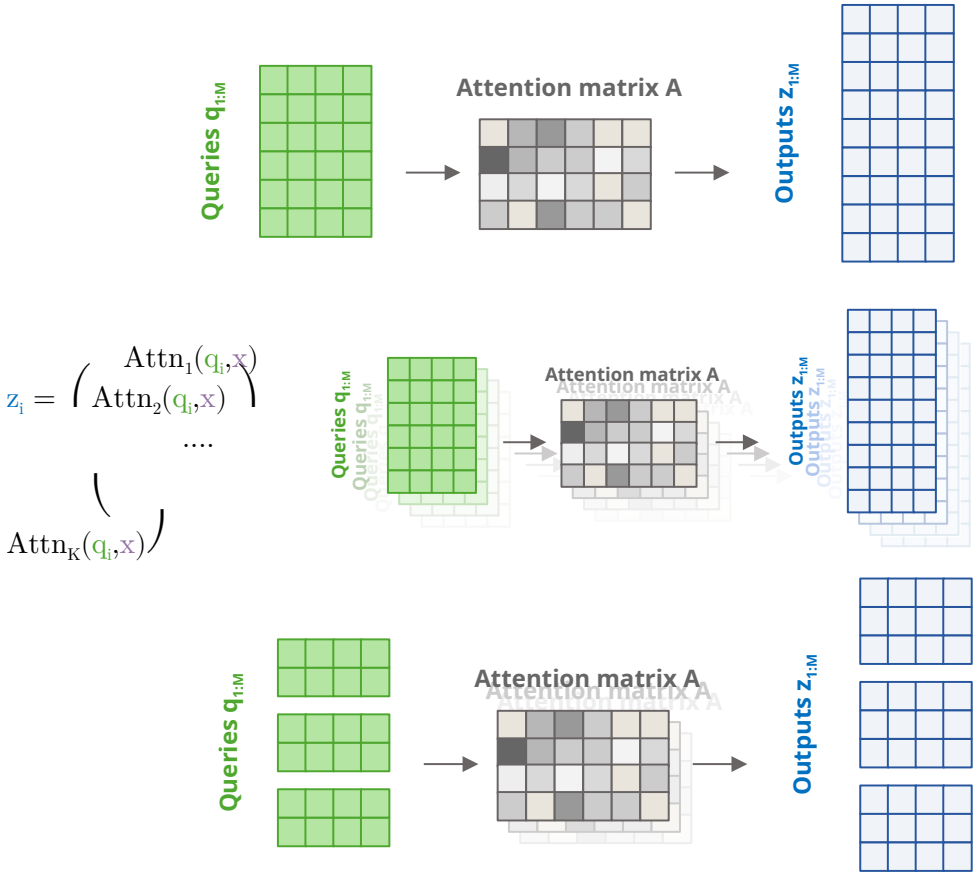
Multi-Head Attention

We usually use **many queries** $q_{1:M}$, and stacking them leads to the Attention matrix $\mathbf{A}_{1:N,1:M}$ and subsequently to many outputs:

$z_{1:M} = \text{Attn}(q_{1:M},x) = [\text{Attn}(q_1,x) \mid \text{Attn}(q_2,x) \mid \ldots \mid \text{Attn}(q_M,x)]$

This leads to "**multi-head**" attention where the operation is repeated K times and the results are concatenated along the feature dimension. **W**s differ.





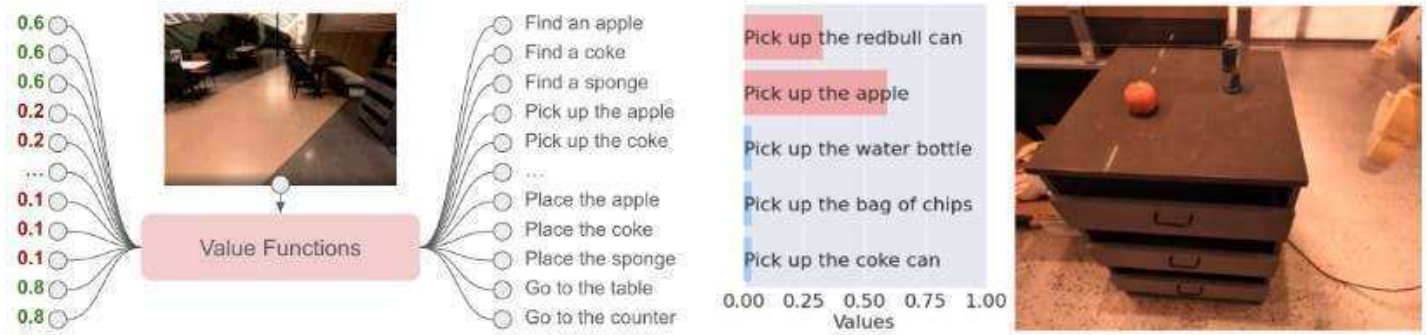
[4] Vaswani. et al.

Do As I Can, Not As I Say

LLMs can plan (“make me a snack”) but don’t know what the robot can actually do
SayCan uses an LLM to generate a high-level plan:
1. Find Chips
2. Grasp chips
3. Bring to person

Value functions score the feasibility of each step given the current scene
- Combines language likelihood x affordance score

Robot only attempts what it can actually do





[13] Ahn. et al.

An Image is Worth 16x16 Words: Vision Transformers

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy*,†, Lucas Beyer*, Alexander Kolesnikov*, Dirk Weissenborn*, Xiaohua Zhai*, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby*,†

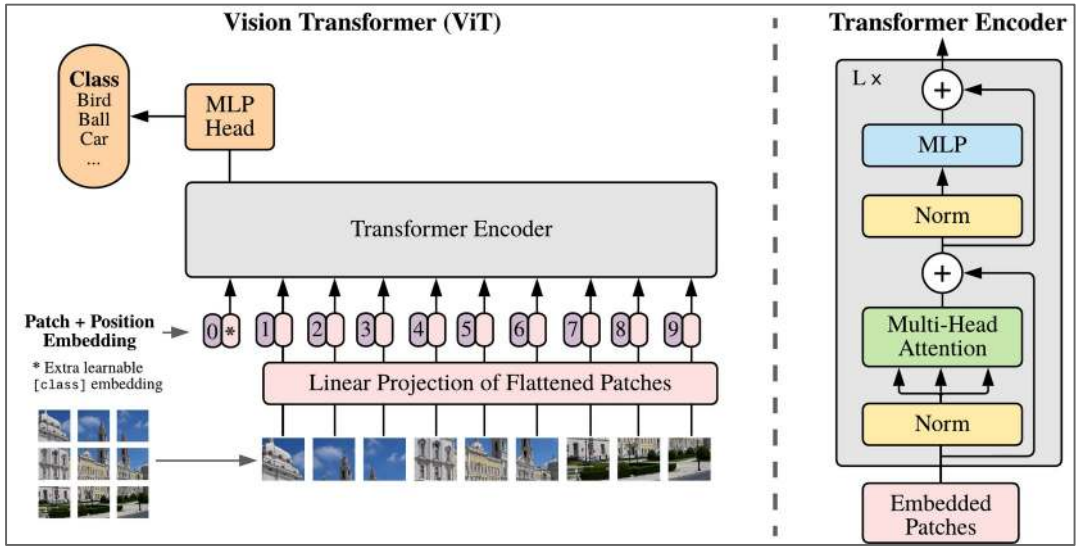
*equal technical contribution, †equal advising

Google Research, Brain Team

{adosovitskiy, neilhoulsby}@google.com

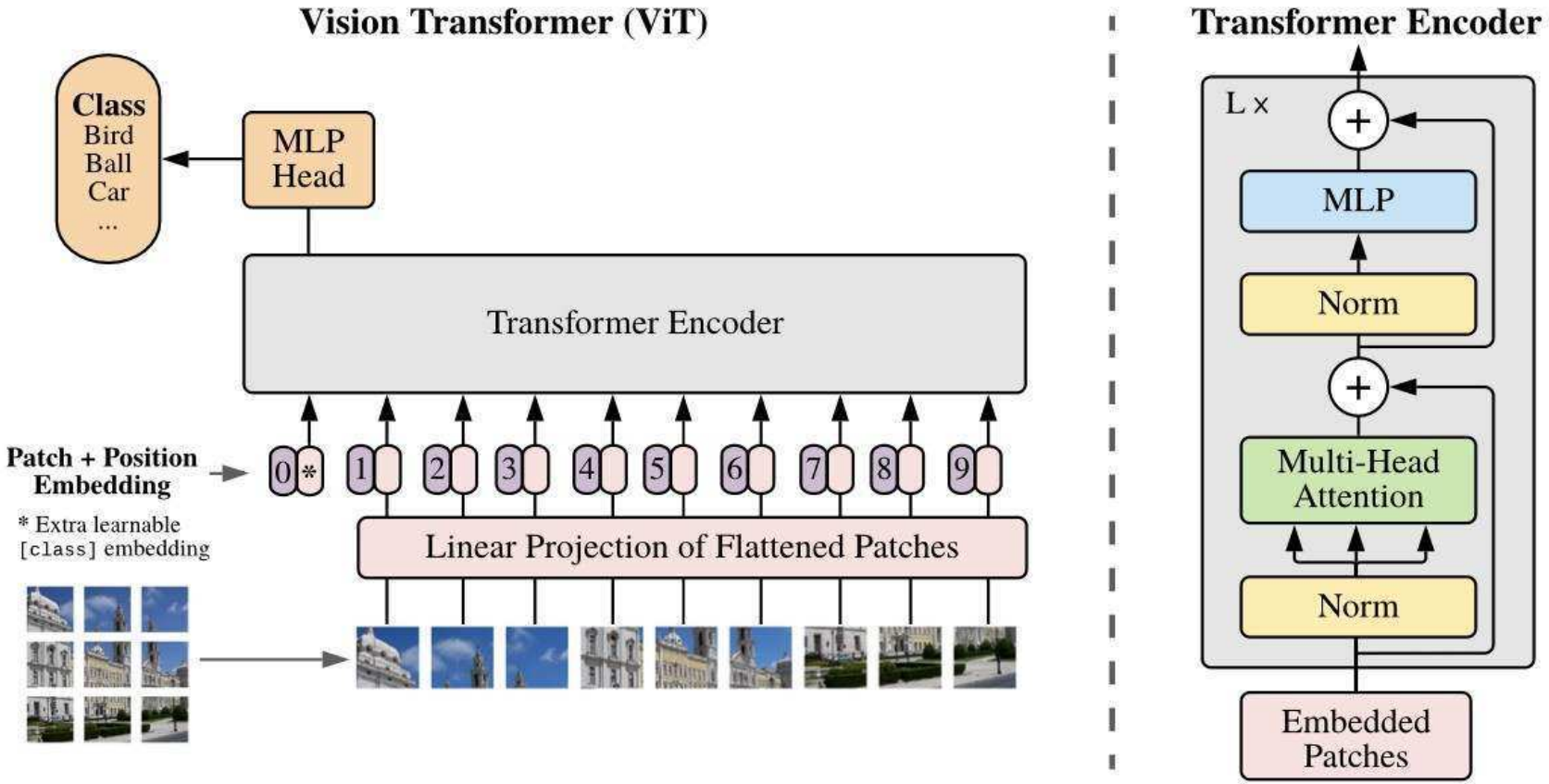
ABSTRACT

While the Transformer architecture has become the de-facto standard for natural language processing tasks, its applications to computer vision remain limited. In vision, attention is either applied in conjunction with convolutional networks, or used to replace certain components of convolutional networks while keeping their overall structure in place. We show that this reliance on CNNs is not necessary and a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks. When pre-trained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) attains excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train.[1]



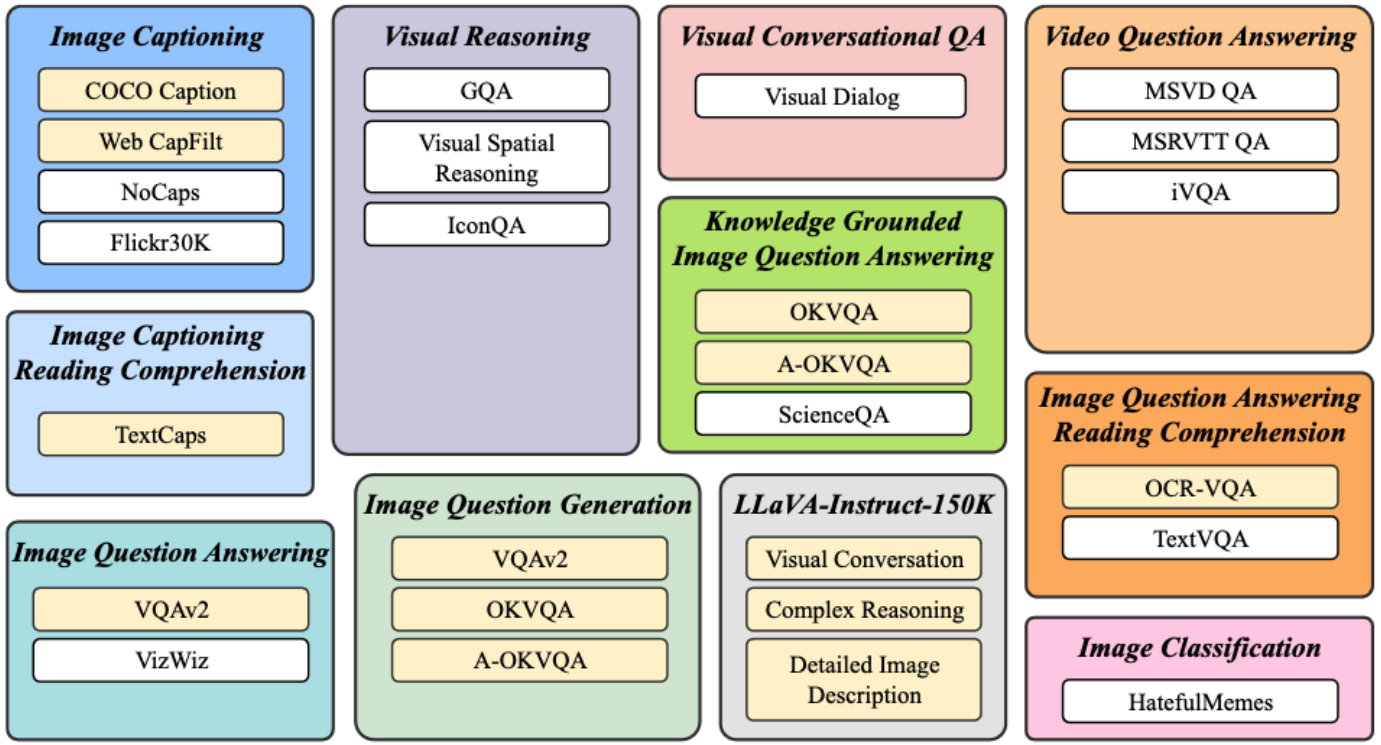
[5] Dosovitskiy, et al.

An Image is Worth 16x16 Words (2020)



Number of patches N = Effective input sequence length for the Transformer, N = HW/PP, H: height, W: width, P: patch size

What is a Multimodal LLM (MLLM)?



[12] Dai. et al.

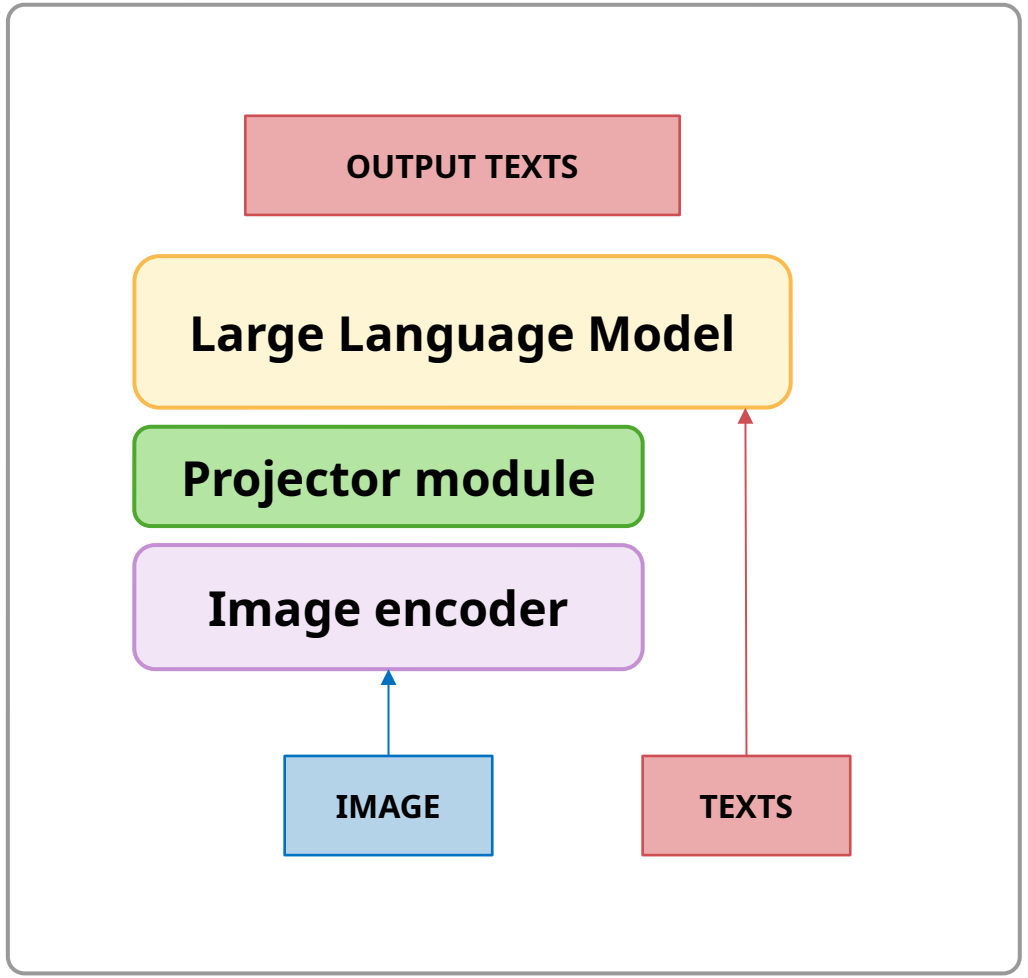
General MLLM Architecture

Most Multimodal LLMs are made of 3 core elements:

- Image encoder to produce image embeddings
- Projector or connector module
- LLM(decoder) that generate texts

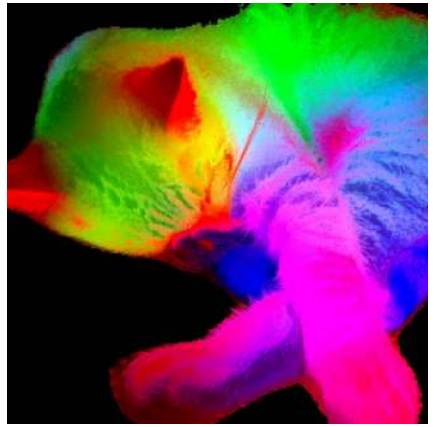
Training Task: Language modelling objective conditioned on image and text tokens.

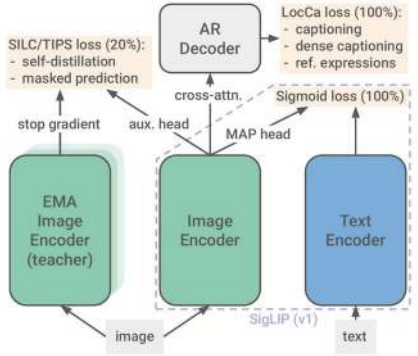
$$\text{Loss} = -\sum_{i=1}^{M} \log p(t_{i+1} | v_1, v_2, \ldots, v_N, t_1, t_2, \ldots, t_i)$$

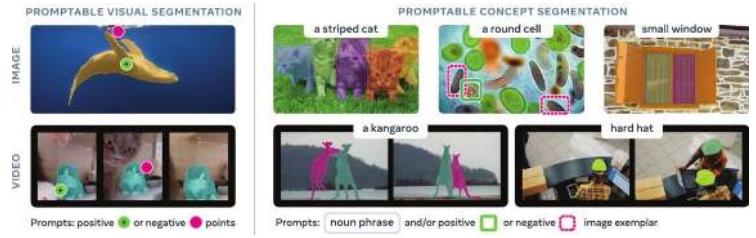


Training Data: Image text pairs, visual instruction data, images interleaved with texts

Vision Foundation Models







DINOv3: self-supervised dense features

- 7B ViT, 1.7B images, no labels
- Gram anchoring keeps dense features sharp
- Frozen backbone wins detection, segmentation, depth

CLIP, SigLIP 2, Perception Encoder

- Contrastive image-text pairs at web scale
- Patch embeddings aligned with text
- Standard VLA vision backbone (SigLIP)

SAM 3: Promptable concept segmentation

- Noun phrase → every instance output
- Images and video (30m/s)
- SAM 3.1: 16 objects in one pass

[78] Siméoni, et al.; [80] Tschannen, et al.; [79] Carion, et al.

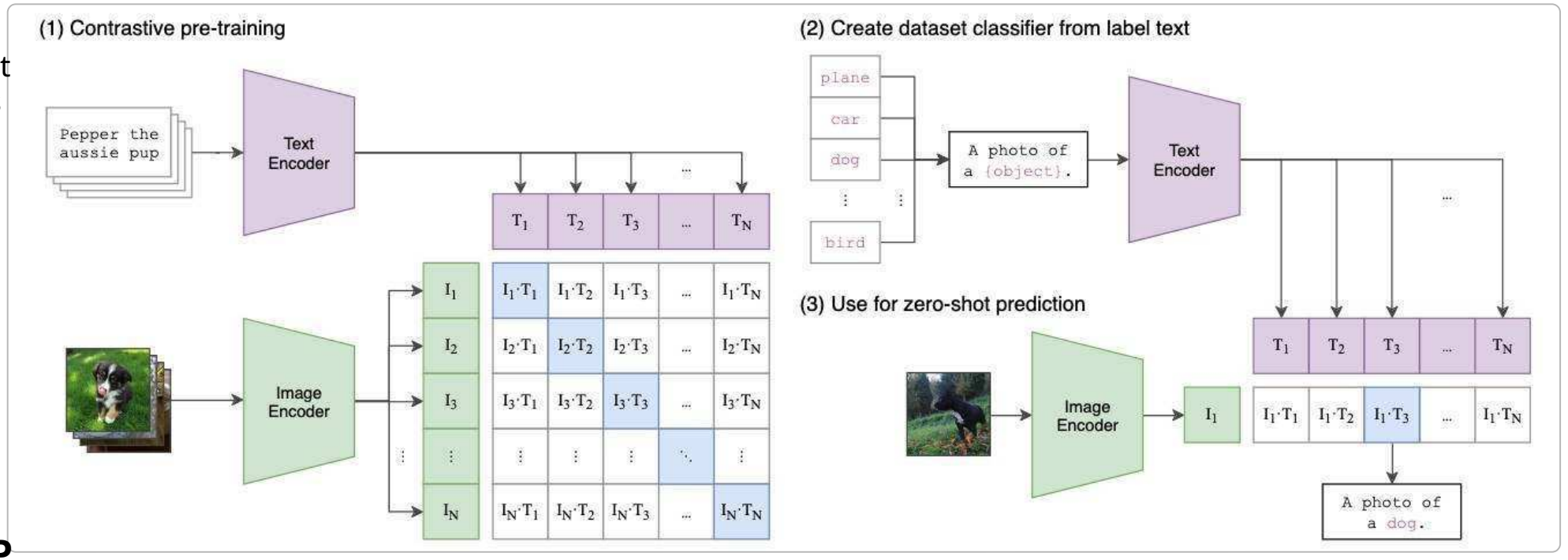
CLIP: Bridging Vision and Language

Standard encoders are trained on image and hand-crafted labels
Labels are hard to get:

- Can we learn visual representation from natural language supervision?

CLIP jointly trains both encoders to learn the correct pairings of images and texts using contrastive loss.

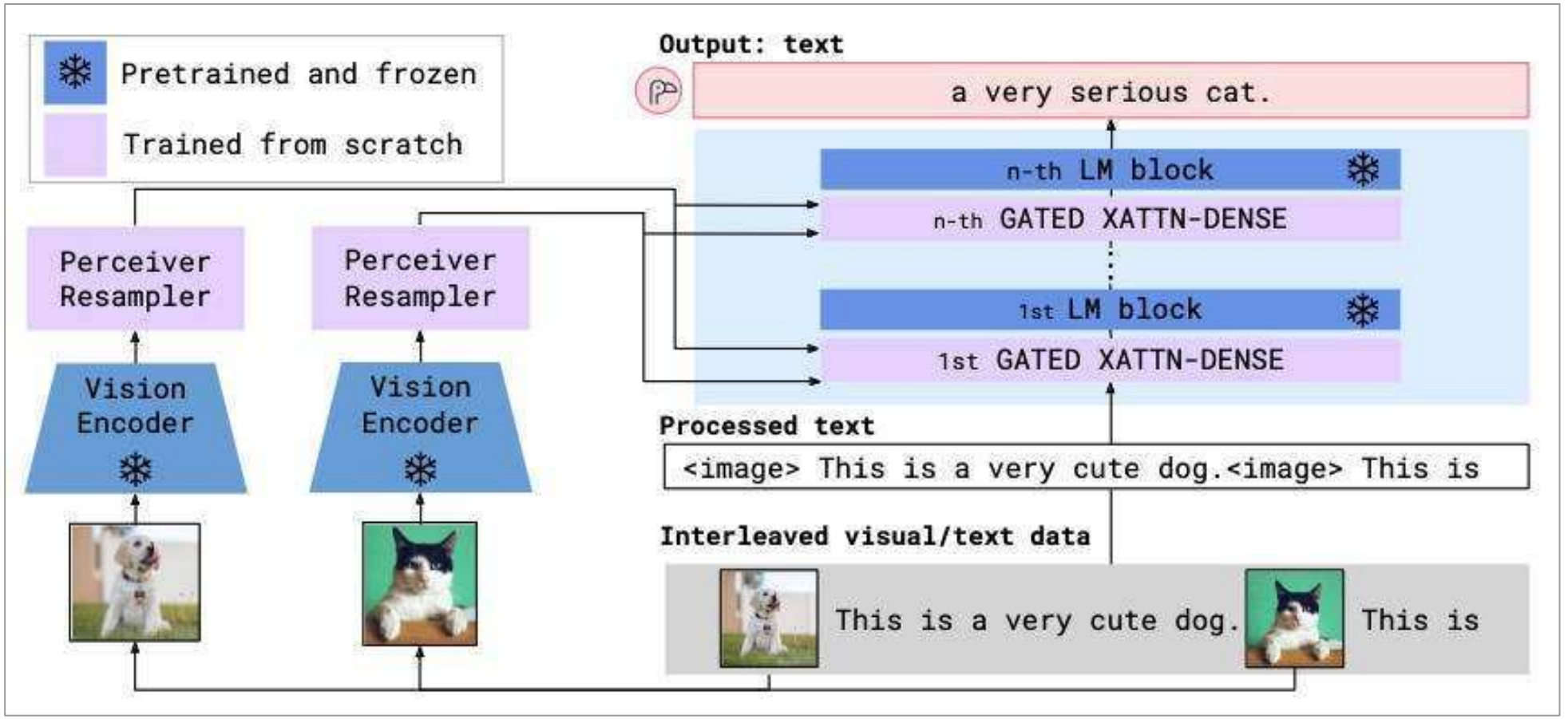
Trained on 400M image-text pairs collected from the internet.



Modern VLAs use **SigLIP**

[6, 7] Radford. et al.

Flamingo: a VLM for Few-Shot Learning



LLM: Chinchilla(1.3B, 3B, 70B)

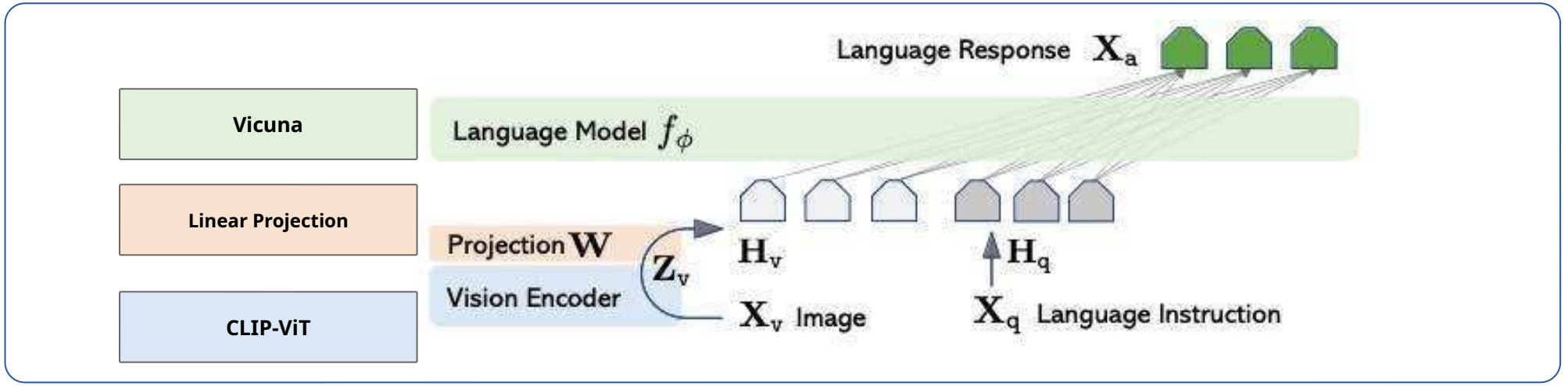
Connector: Perceiver Resampler, Gated attention

Vision Encoder: Normalizer-Free ResNet (NFNet)

[8] Alayrac, et al

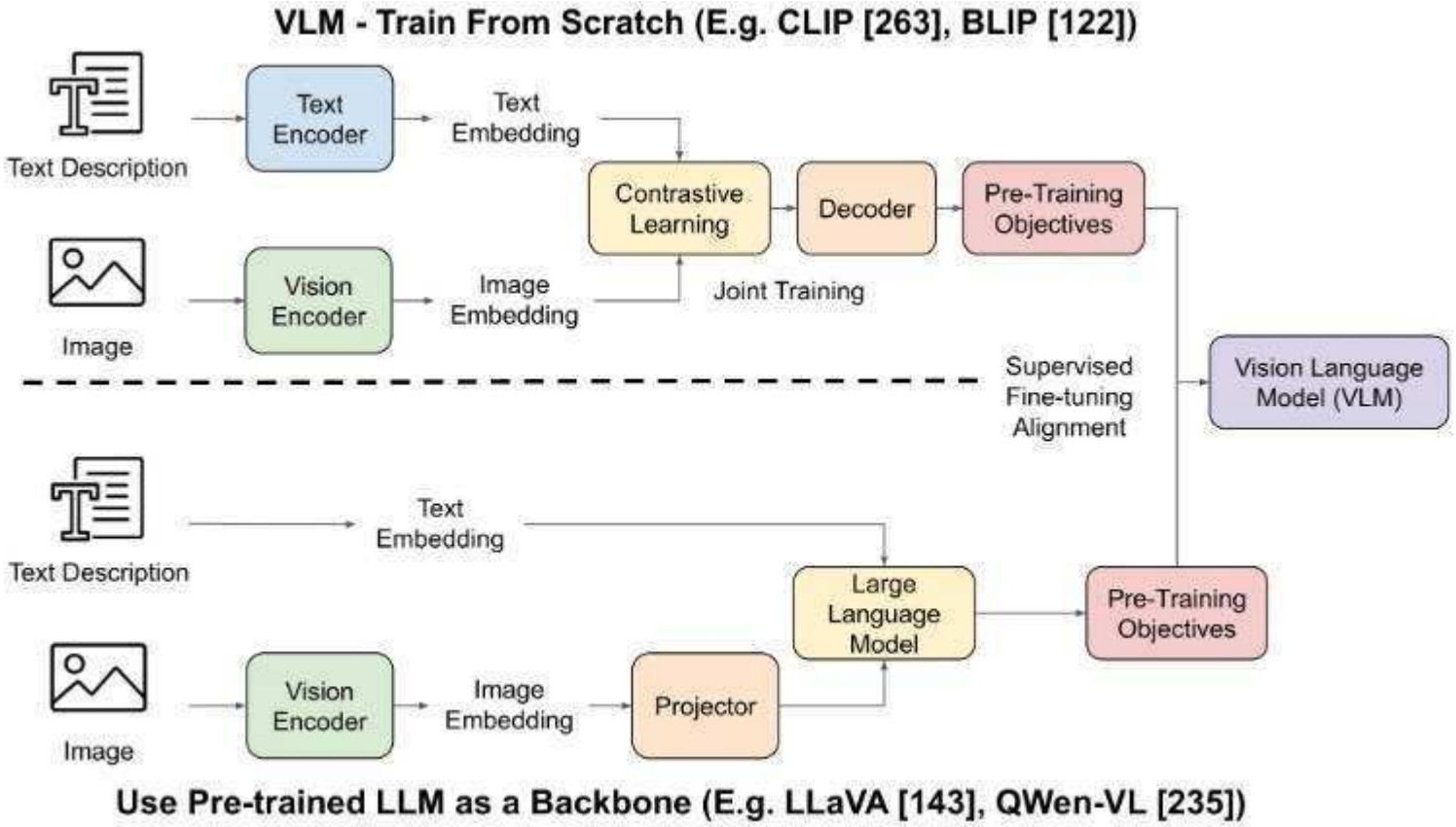
LLaVA: Visual Instruction Tuning

- LLaVA is a visual encoder and LLM(Vicuna) connected by a simple projection layer.
- Two training stages:
  - **Stage 1 - Pre-training for vision and language alignment**: Train the model on 595K image-text pairs filtered from CC3M(Common Crawls). The image-text pairs are converted into instructions.
  - **Stage 2 - Supervised fine-tuning:** Finetune the base LLaVA on specific tasks: multimodal chats and multimodal reasoning(with ScienceQA dataset).



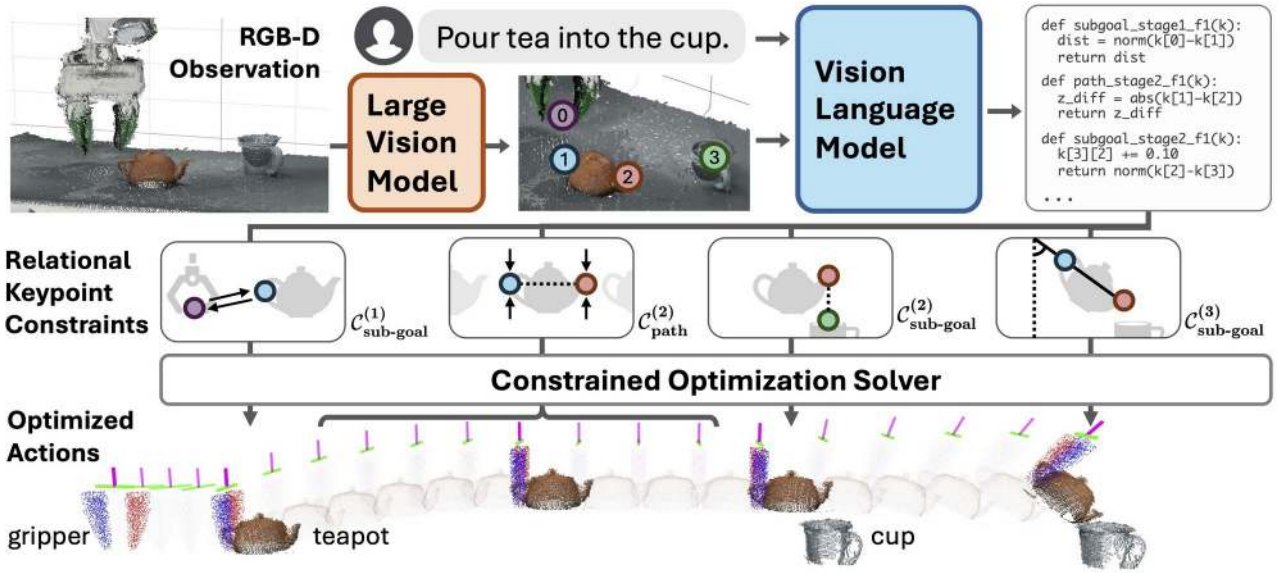
[9] Liu, et al

Two Recipes for a VLM, and Which One VLAs Use



[10] Li. et al.

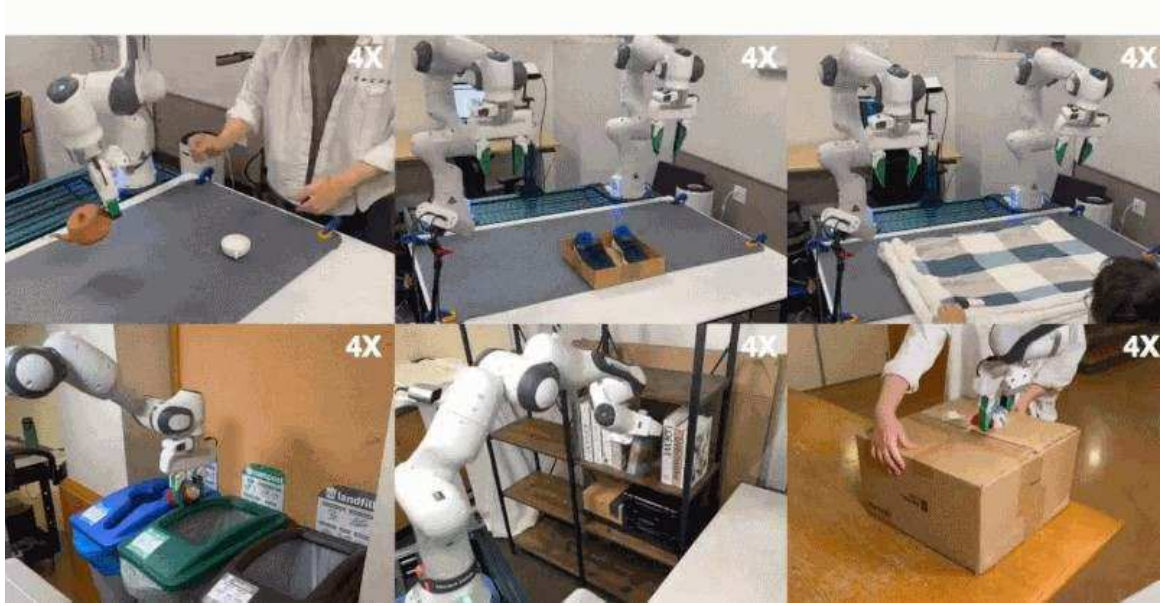
ReKep: VLMs with robots?



Relational Keypoint Constraints represents manipulation as a sequence of constraints defined on keypoints

- Each constraint is a function of a set of 3D keypoints that return a cost
- Sub-goals obtained by stringing these functions together

Achieve tasks by minimizing cost



[14] Huang, et al.

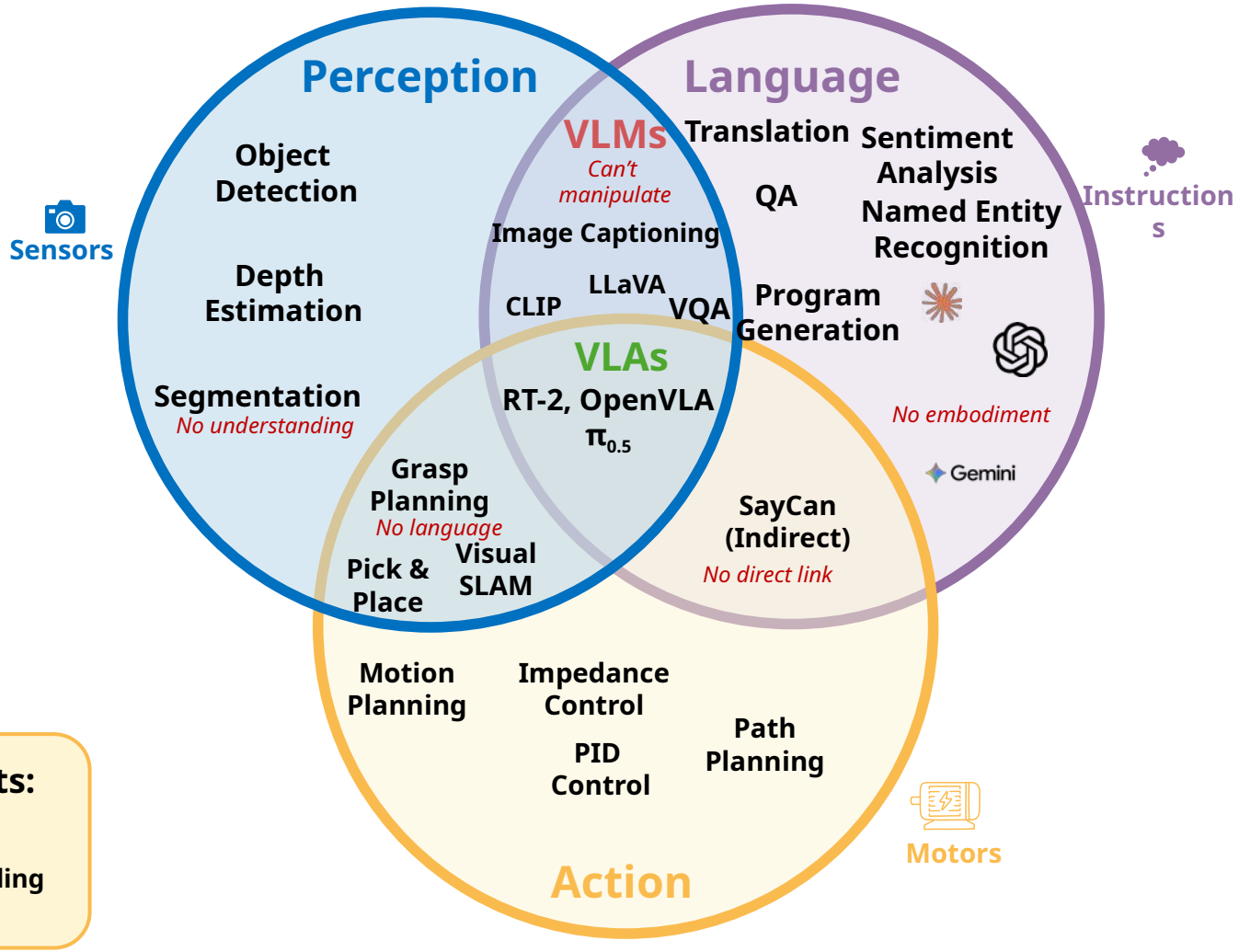
Perception vs Action Ga

Problem 1:
VLMs can describe but can't do

Problem 2:
Robots can do but can't understand

Solution: VLAs understand AND do!

VLM Capabilities:
✓ Zero-shot recognition
✓ Scene understanding
✓ Visual reasoning
✗ Physical interaction

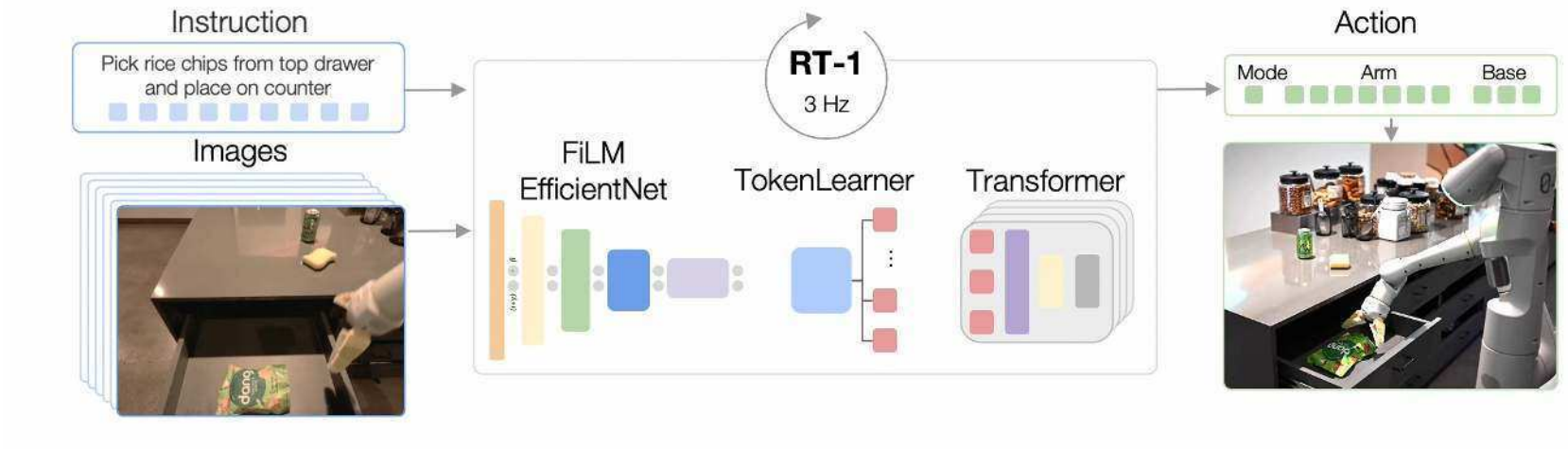


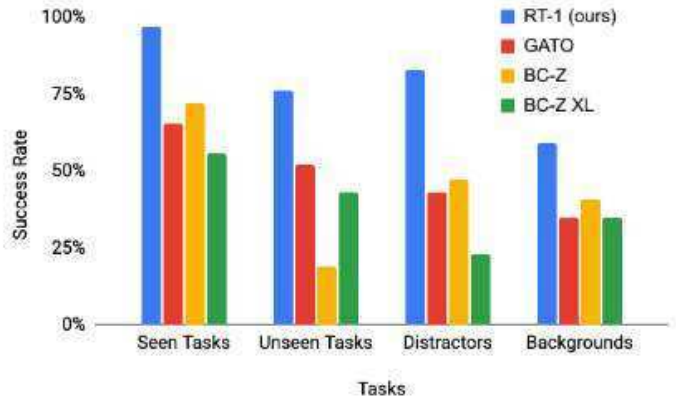
Traditional Robots:
✓ Precise control
✓ Force feedback
✗ Semantic understanding
✗ Generalization

VLA Breakthrough:
Semantic manipulation
Language instructions
Physical grounding
Zero-shot transfer

The Breakthrough of Vision Language Action Models

RT-1

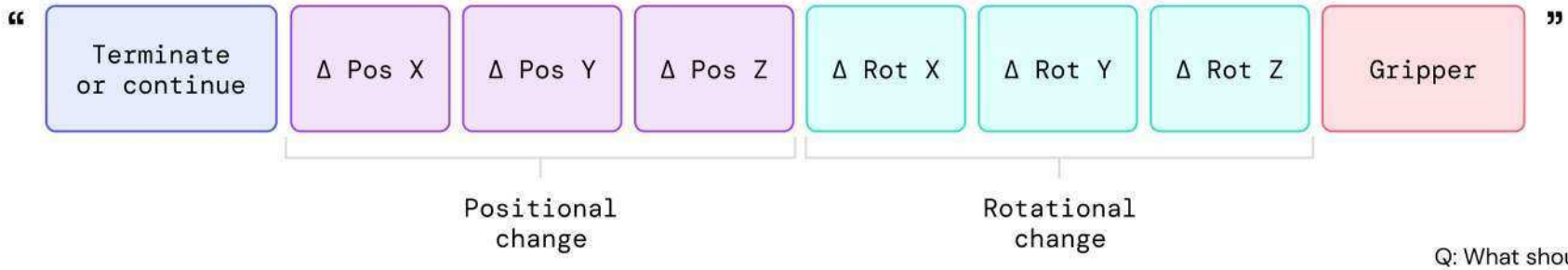




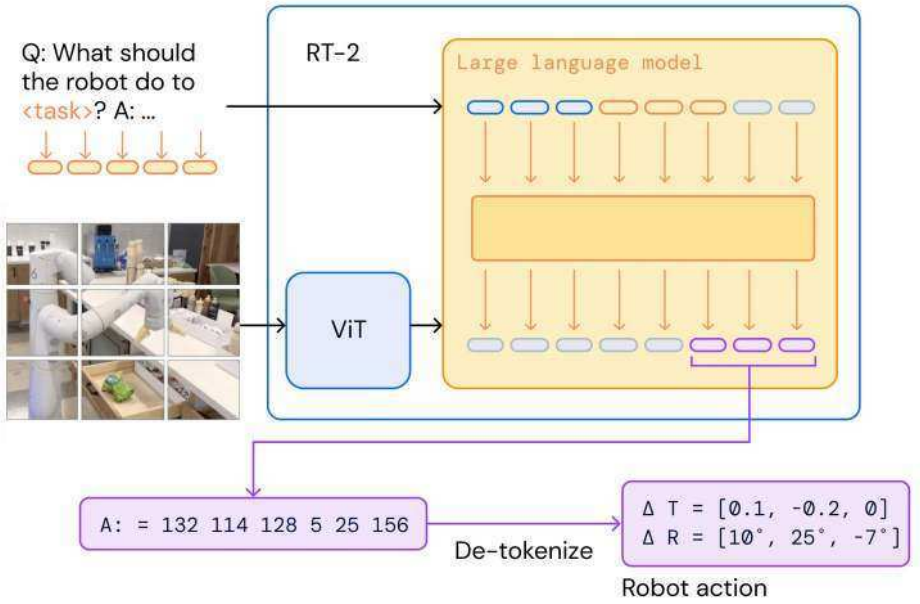


[15] Brohan, et al.

RT-2: Patient Zero VLA

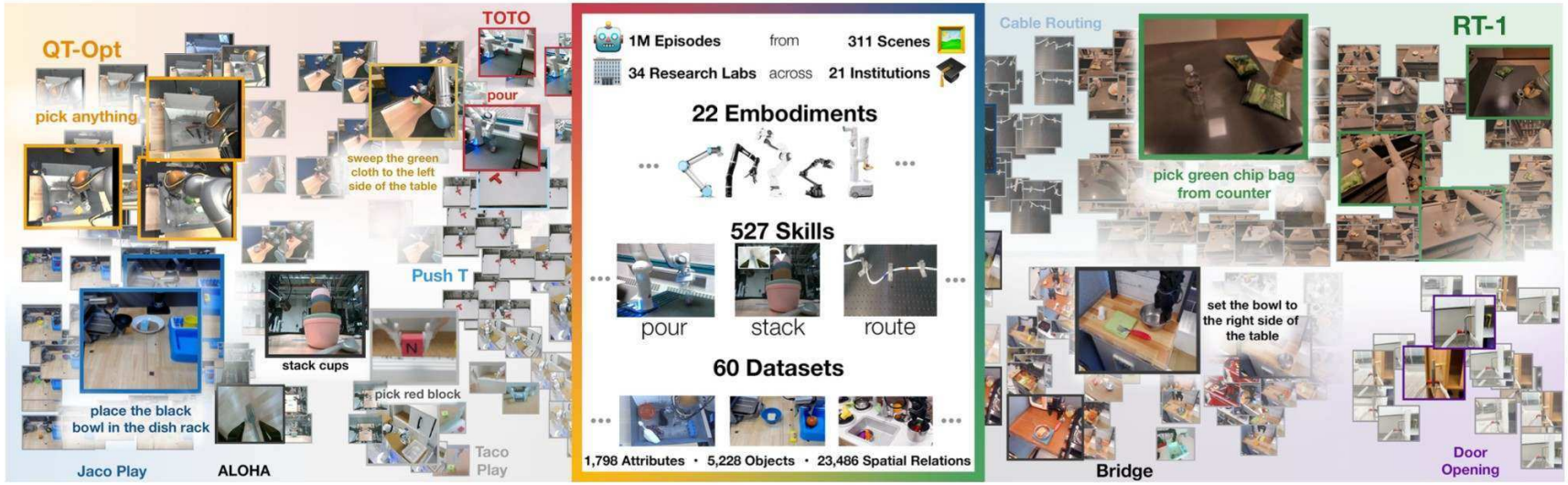


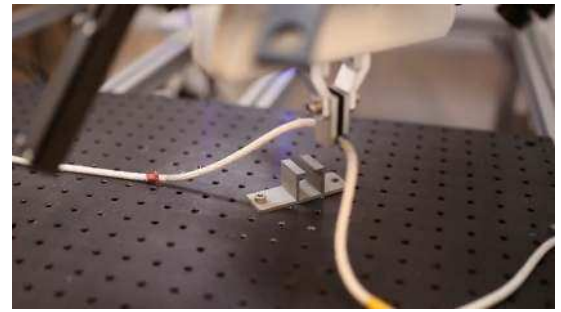




[16] Brohan, et al.

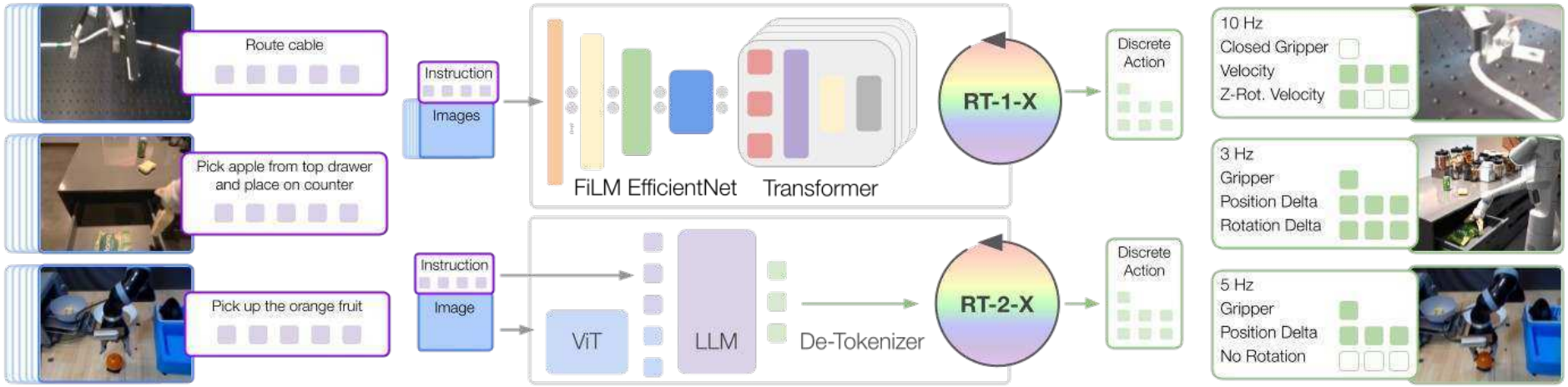
RT-X: Scaling Robotic Transformers





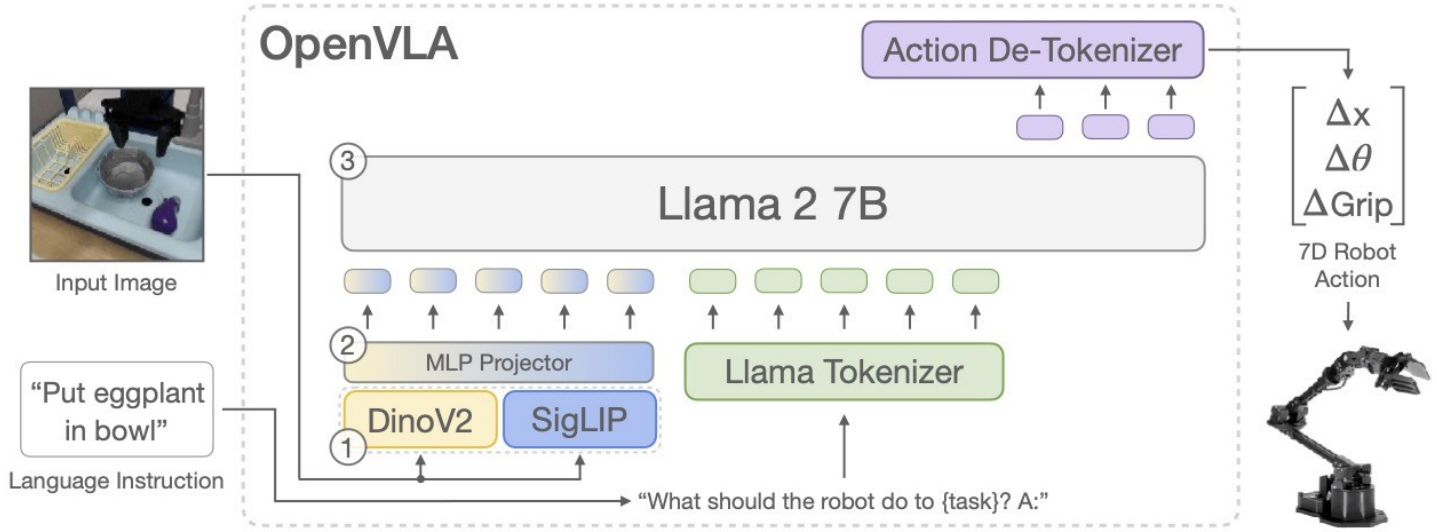






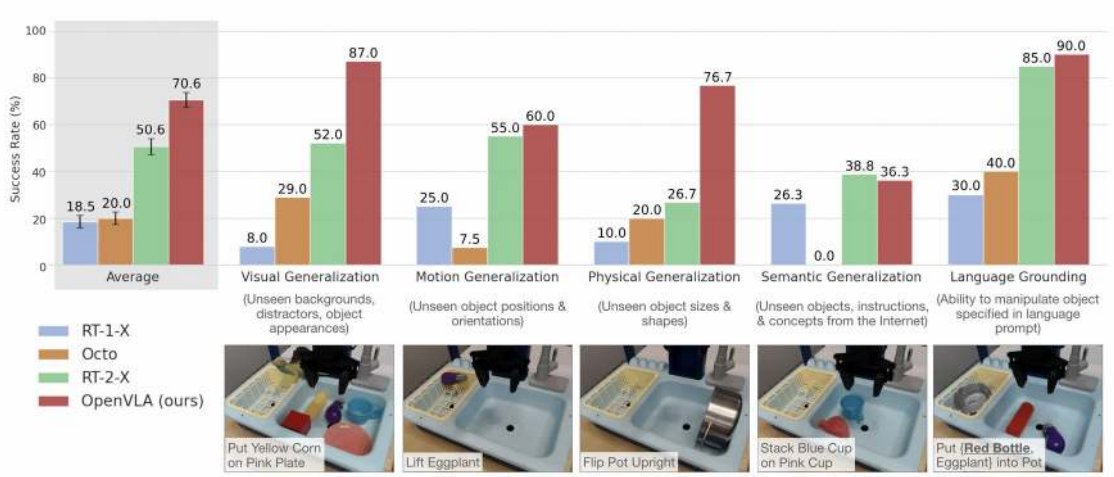
[17] Open X-Embodiment Collaboration

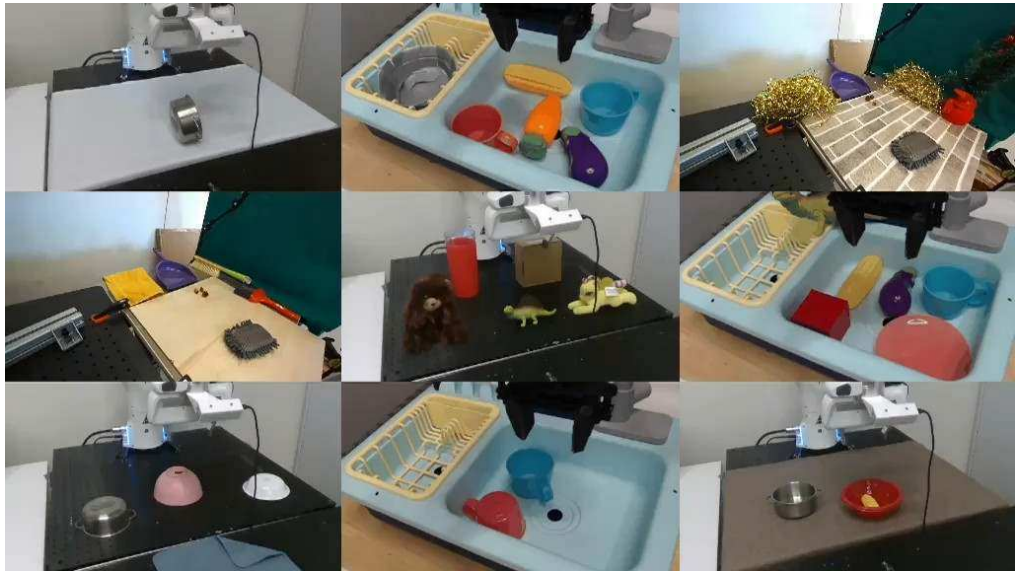
OpenVLA and the Open-Source Revolution



$$\Delta W = BA, B \in R^{m \times r}, A \in R^{r \times n}$$

$$h = W_0 x + \Delta W x = W_0 x + BAx$$

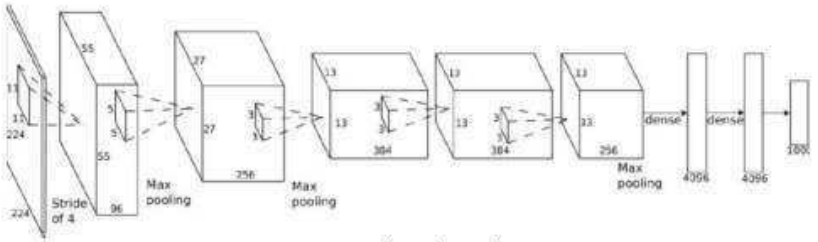




[18] Kim, et al.

The Need for Generalist Robot Policies







$\mathbf{o}_t$ $\pi_\theta(\mathbf{a}_t|\mathbf{o}_t)$ $\mathbf{a}_t$

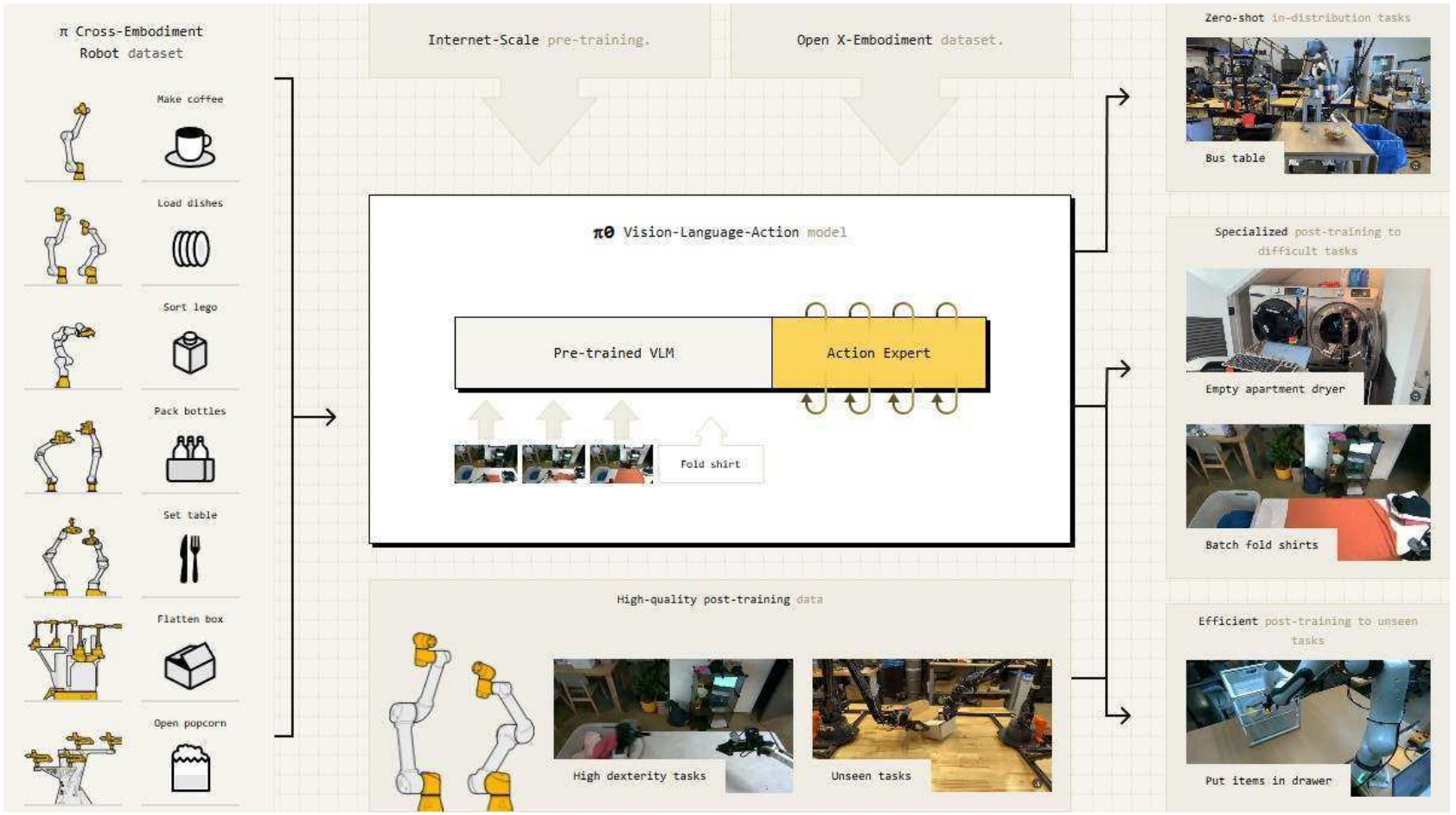
$\mathbf{s}_t$ – state
$\mathbf{o}_t$ – observation
$\mathbf{a}_t$ – action

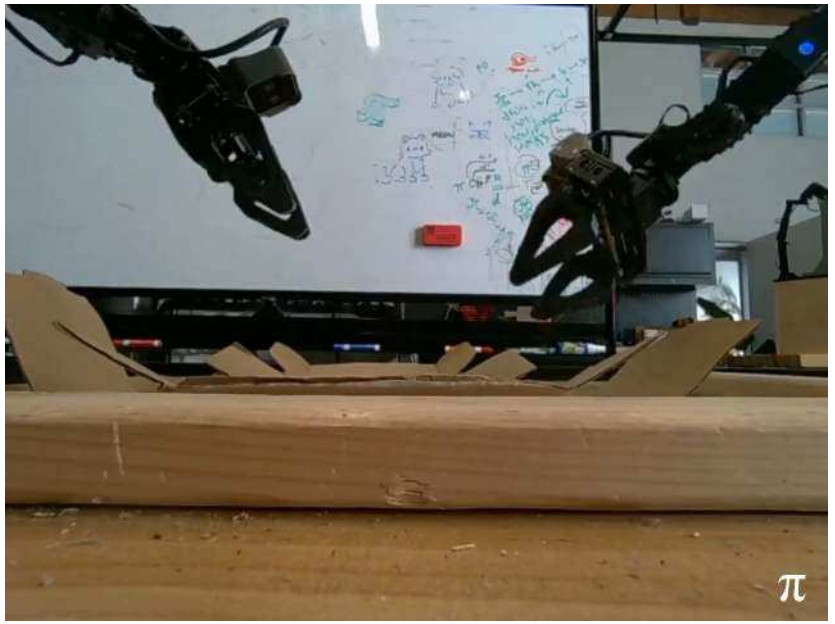
$\pi_\theta(\mathbf{a}_t|\mathbf{o}_t)$ – policy
$\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)$ – policy (fully observed)

Observe, understand and act

Physical Intelligence π_0: Long Horizon Tasks

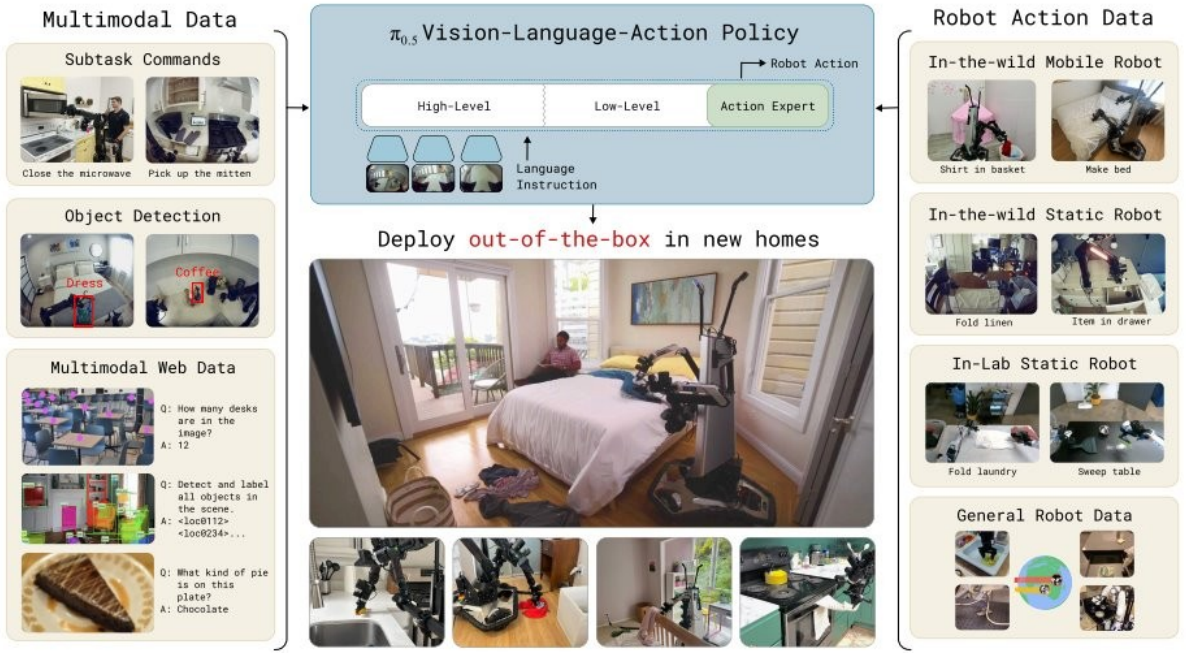
[19] Black, et al

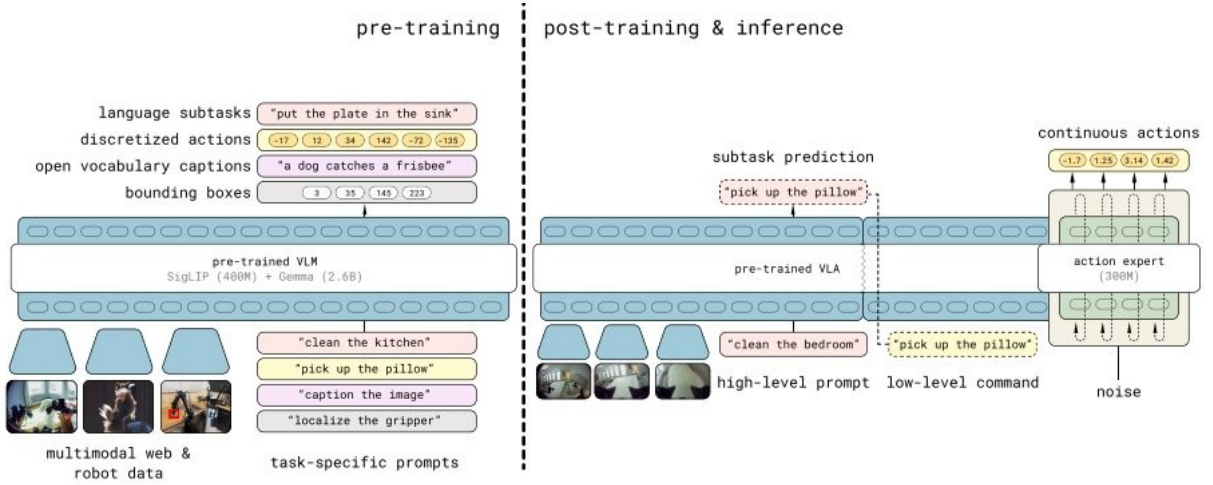




[19] Black, et al.

Physical Intelligence: $\pi_{0.5}$

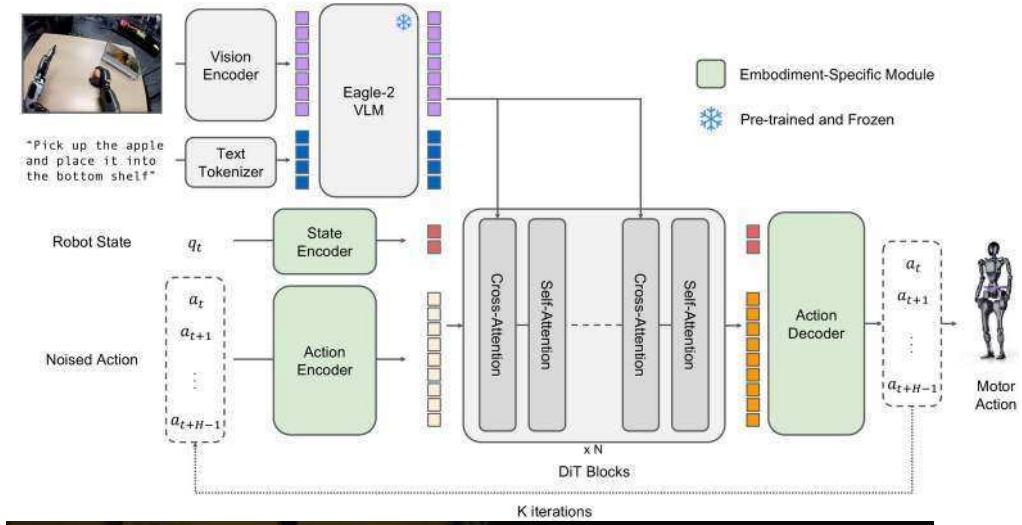


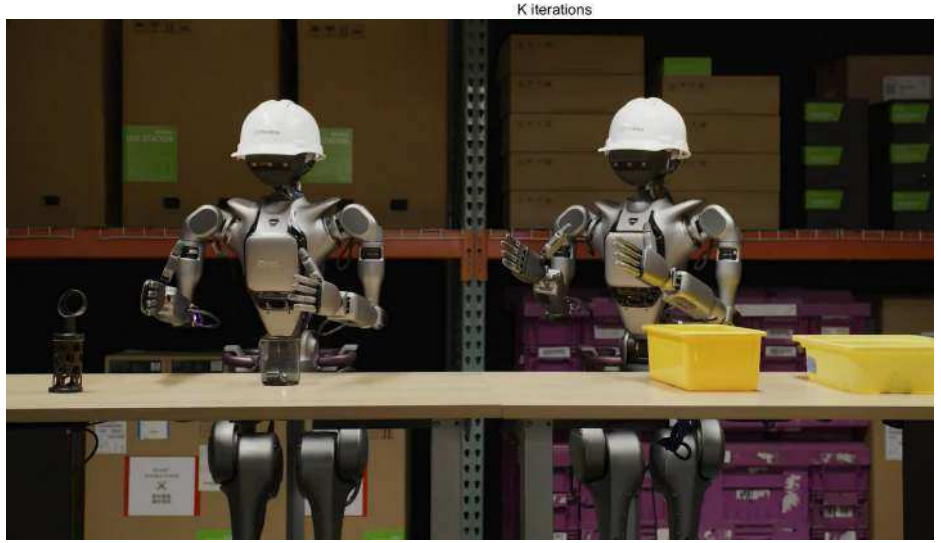


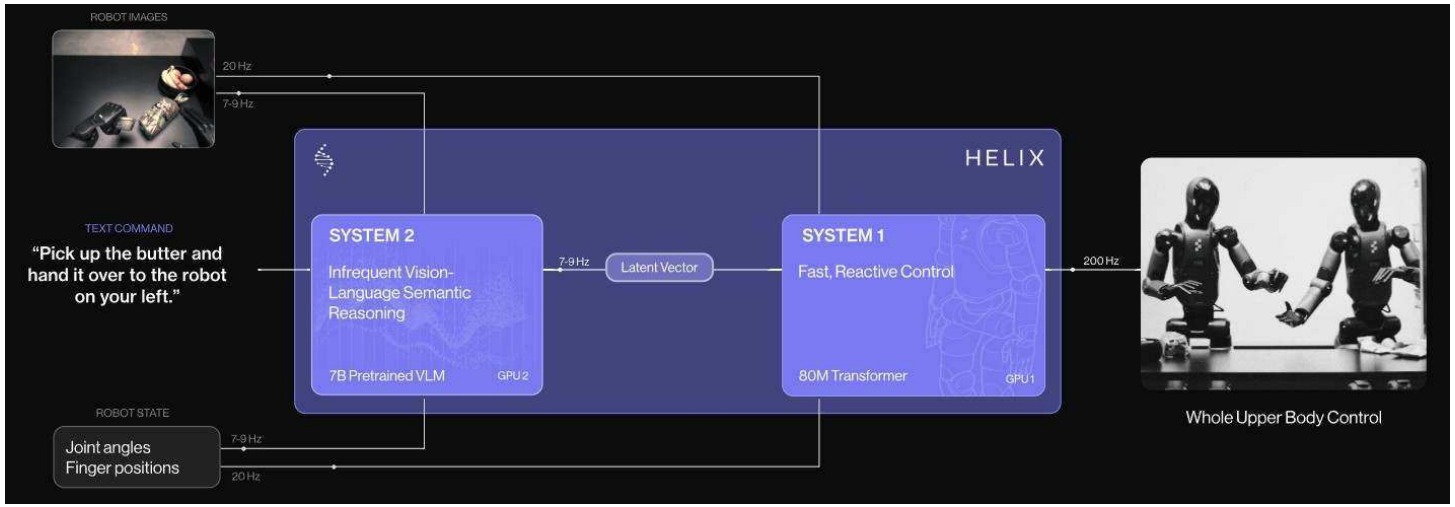


[20] Black, et al.

Dual-System Arch and Synthetic Data Scaling



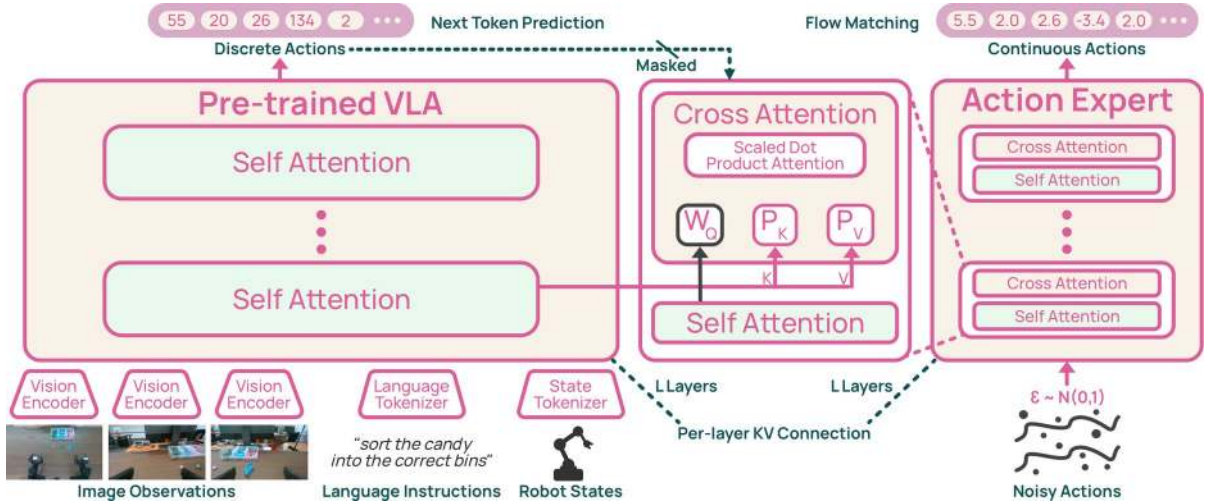




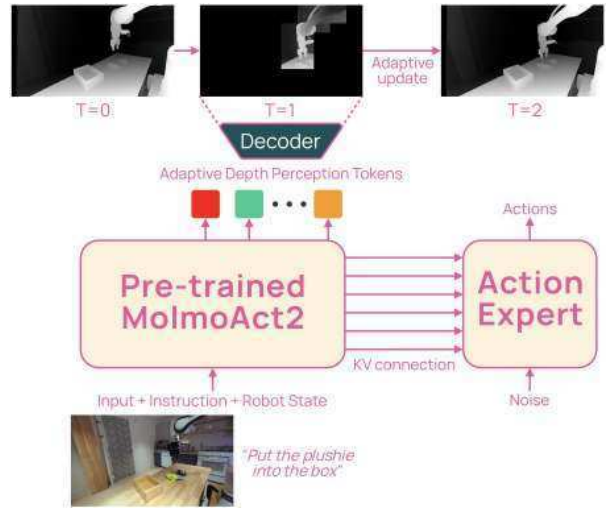
[21] Bjorck, et al.; [56, 57] NVIDIA

Gemini Robotics 2





- ER 2, GR 2, On-Device 2 models
- ASIMOV-Agentic safety benchmark
- Open counterpart: MolmoAct2 (Ai2)



[61-63] Google DeepMind; [44] Fang, et al.

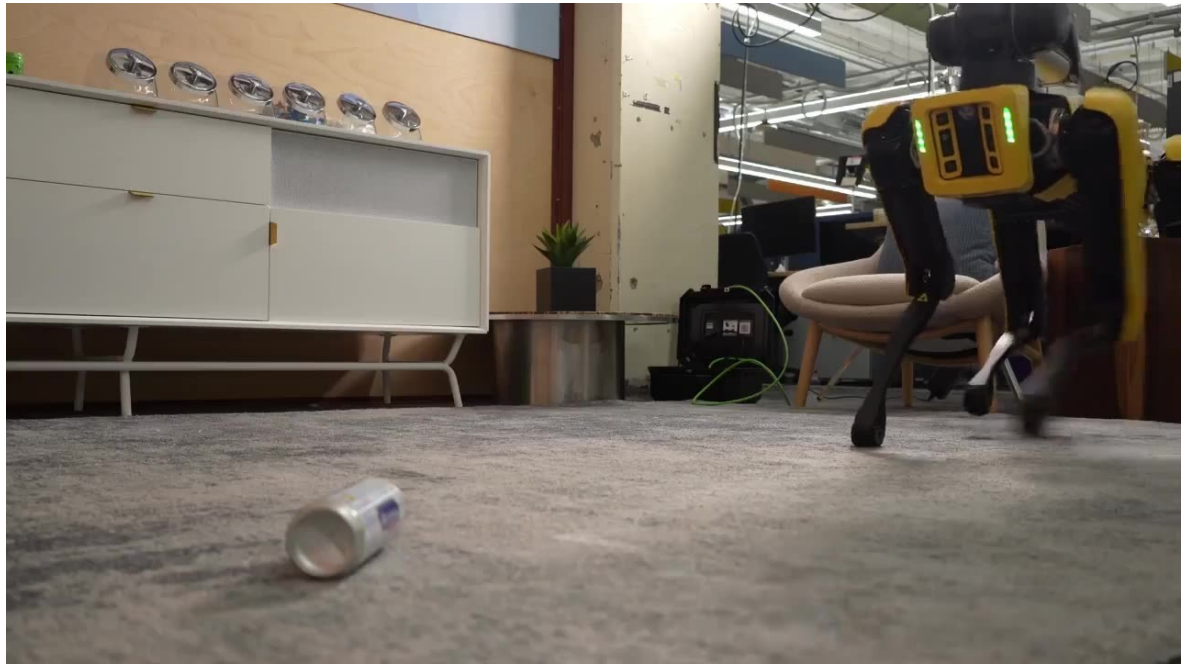
Skild AI S1: “In-Context Learning for Robotics”



- **One model for all embodiments!**
- Low-rate high-level policy drives a high-rate low-level controller
- **In-Context Learning (ICL):** Handles jammed wheels, lost limbs, additional payload, or a new body
- **S1**: video demonstration as input→ actions

[68, 69, 76] Skild AI; [70] NVIDIA

Gemini 4 Argon

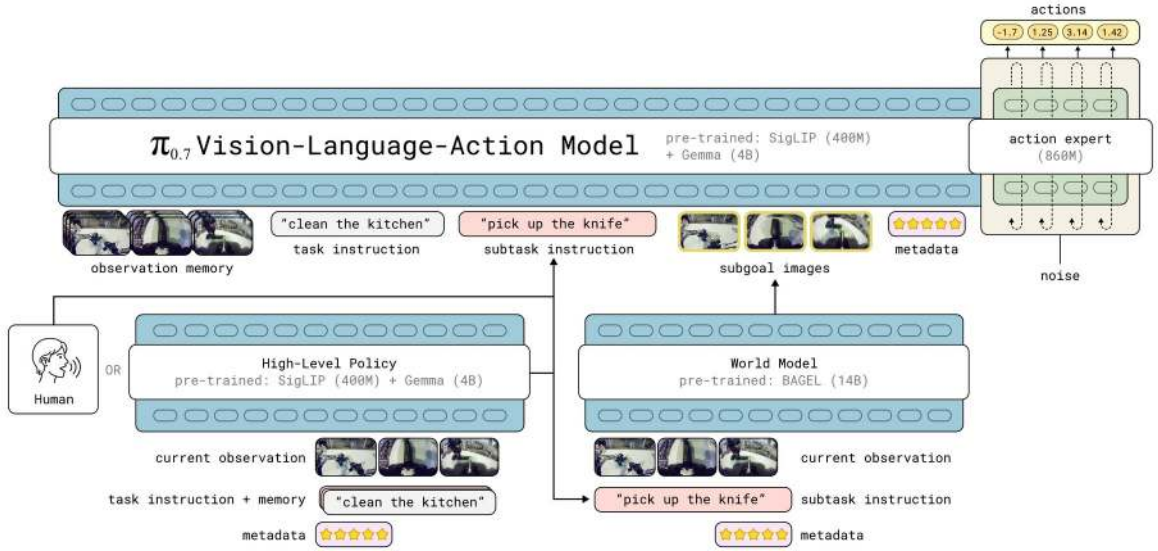


- **Real2Sim:** physically calibrated twins from robot cameras, egocentric human video, or a pen sketch
  - Spot builds digital twin in MuJoCo of pedal bin, practices on 3 legs and deploys
  - Closed-loop steering of video world models
- **Physical understanding:** hindsight labeling, dense value and progress estimation
- **Self-improvement:** one photo → sim → 4,500 episodes → VLA fine-tune: real success 56% → 75

[71, 72] Google DeepMind

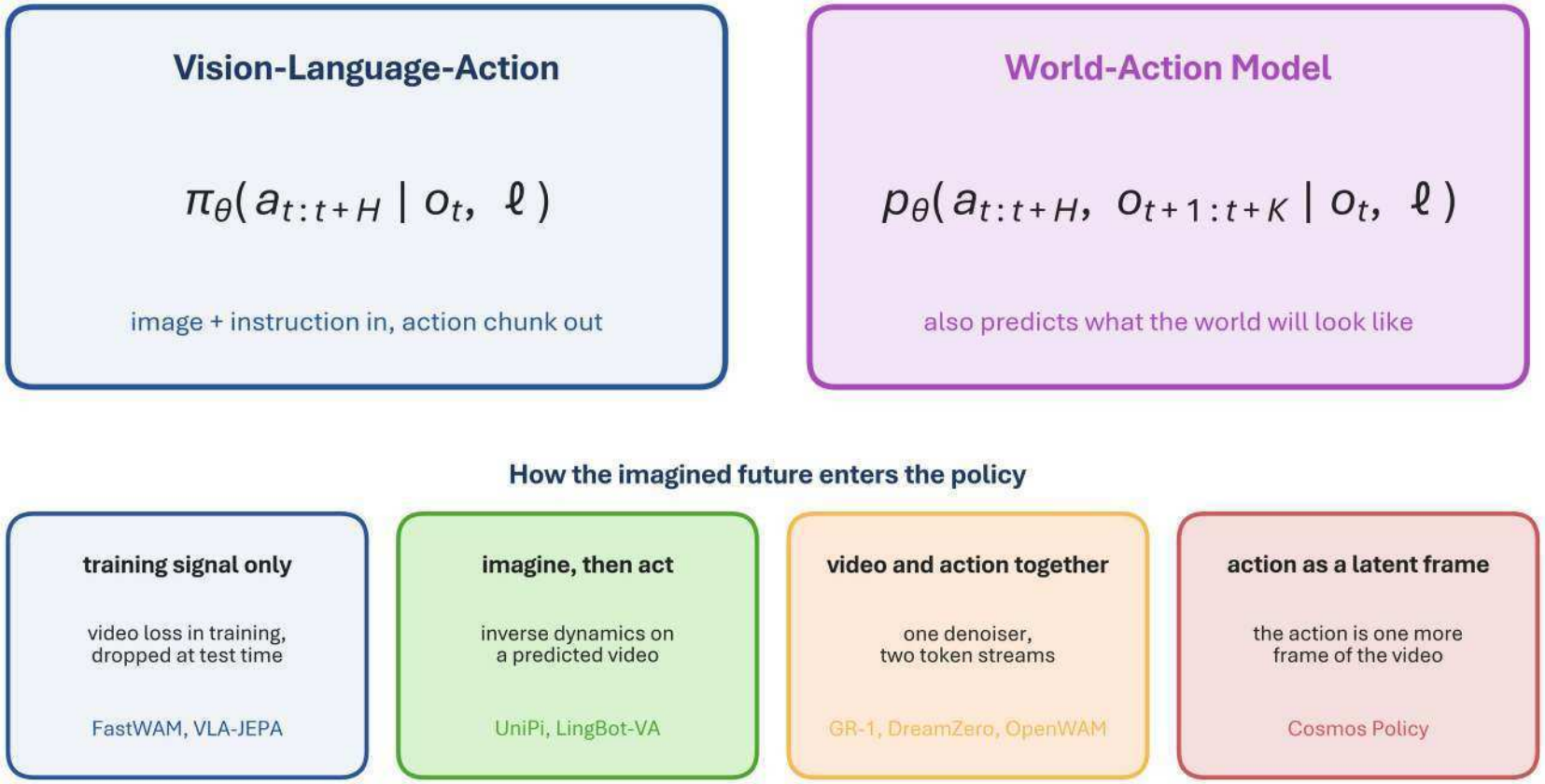
Physical Intelligence: $\pi_{0.7}$





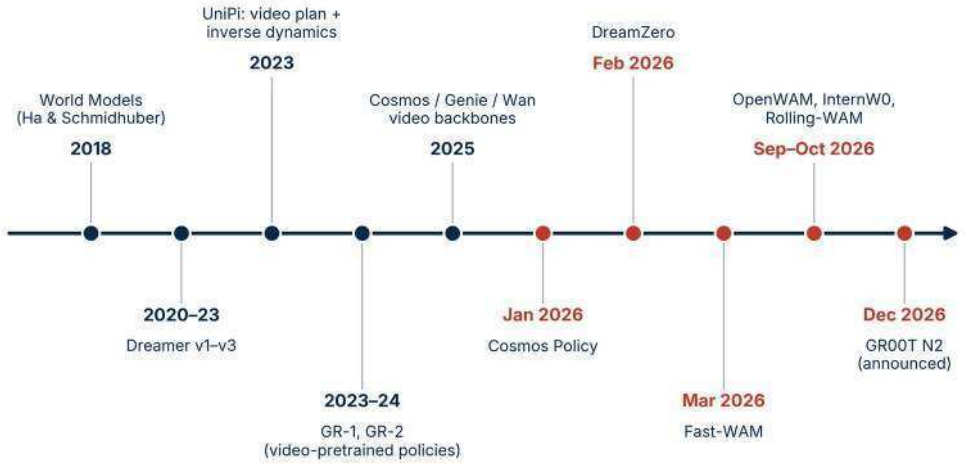
[31] Physical Intelligence

Pretrained to Imagine, Fine-Tuned to Act



[60] Reuss, NVIDIA

Video Models as World Simulators



- Video backbone robustness: Wan 2.x, Cosmos, Genie 3
- Action heads moved from per-step MLPs to chunked flow-matching heads
- Cross-embodiment robot datasets proliferation: OXE, DROID, RoboTwin 2.0, InternW0
- Video is the cheapest physics prior

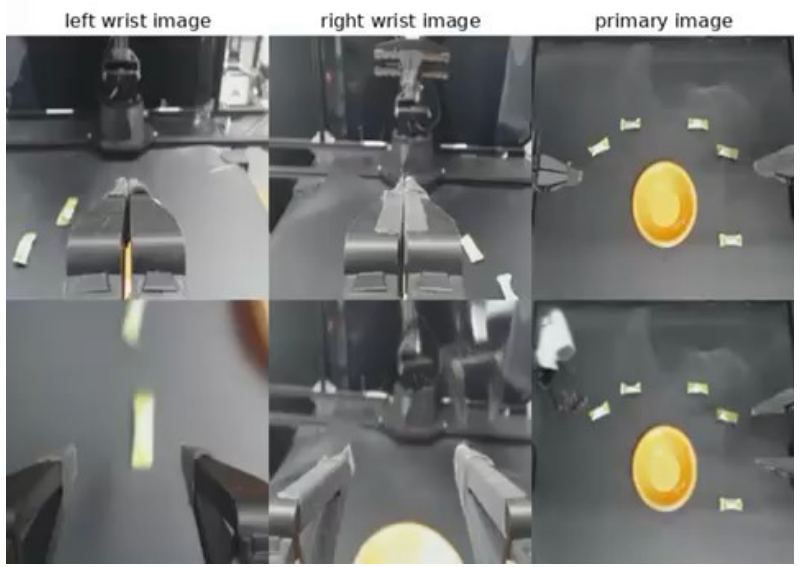
[49] Ha and Schmidhuber; [45] Du, et al.; [46] Wu, et al.; [60] Reuss, NVIDIA

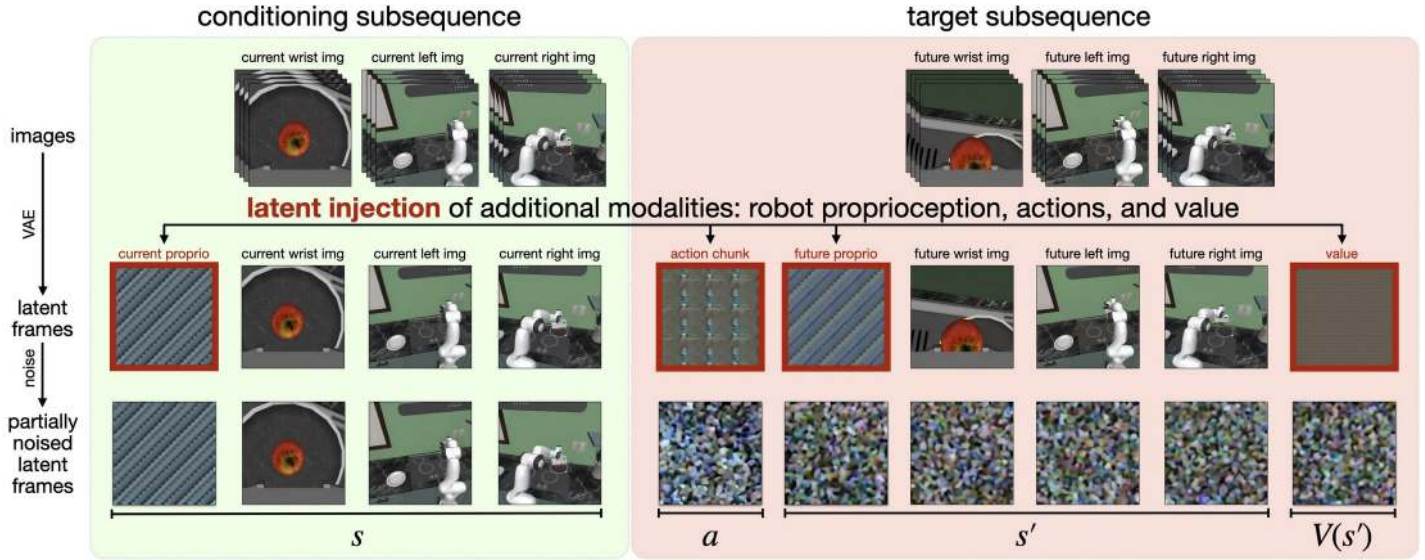
A Taxonomy of World-Action Models

Family	What is predicted	Used at test time?	Examples
Video-backbone policy	future pixels (video diffusion)	**dropped (train-only) or generated**	FastWAM, LingBot-VA, GR-2, InternW0-Δ
Joint latent-frame diffusion	action, future images and value as frames of one video	**generated jointly**	Cosmos Policy, DreamZero, PatchWAM
Latent-predictive (JEPA)	future embeddings, not pixels	**planning or auxiliary**	V-JEPA 2-AC, VLA-JEPA, V-JEPA Policy, LeWorldModel
Plan, then act	a video plan, then inverse dynamics	**yes**	UniPi, OpenWAM (video-then-action mode)
World model as simulator	rollouts for evaluation or RL	**offline**	Ctrl-World, Hydra-0, WorldEval

[27-30, 33, 34, 39, 40, 42, 43, 52, 53]

Cosmos Policy



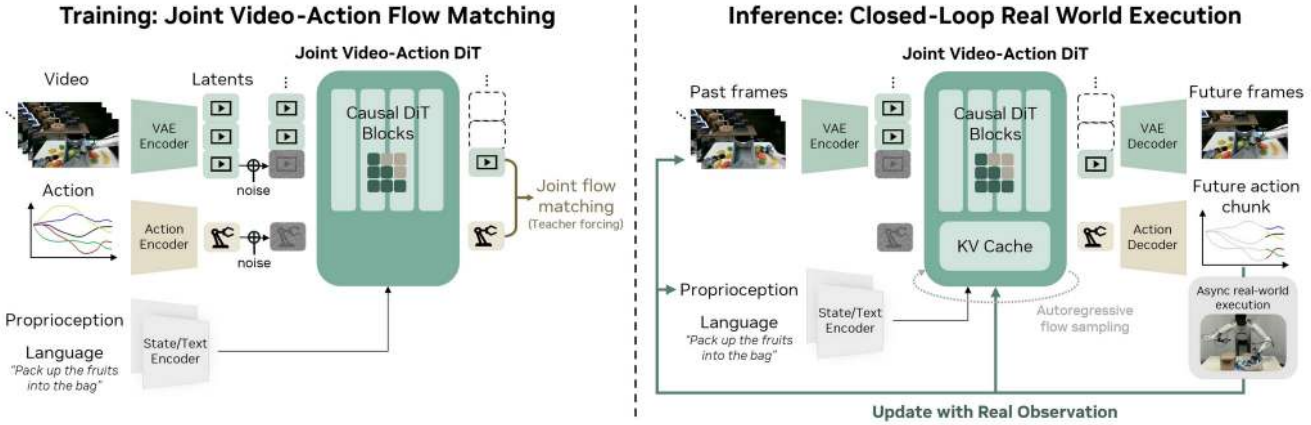




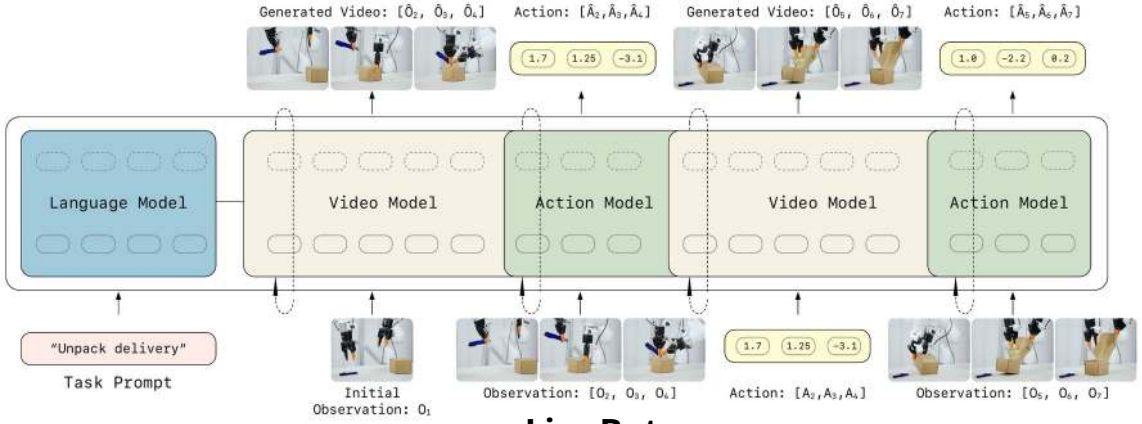
- Post-trains Cosmos-Predict2 2B
- Actions, proprioception and value become latent frames of the video
- DiT denoises all nine frames
- 97.6% on LIBERO-10

[27] Kim, et al.

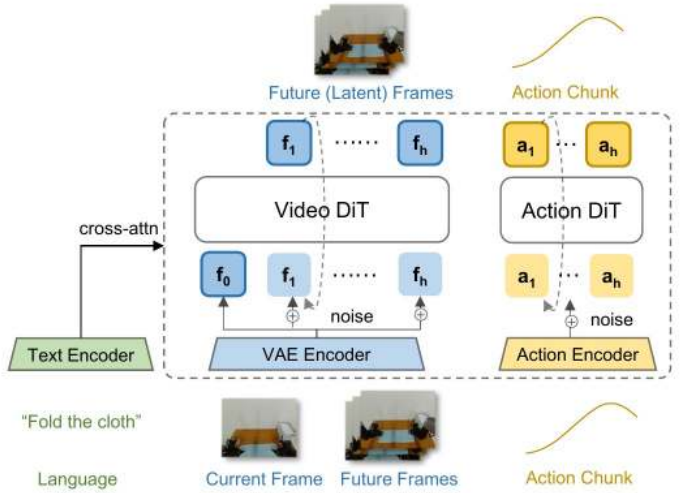
Same Backbone Family, Test Time Disparities



DreamZero



LingBot-VA



FastWAM

[30] Ye, et al.; [28] Yuan, et al.; [29] Li, et al.

Generalist AI: GEN-0 → GEN-1.5

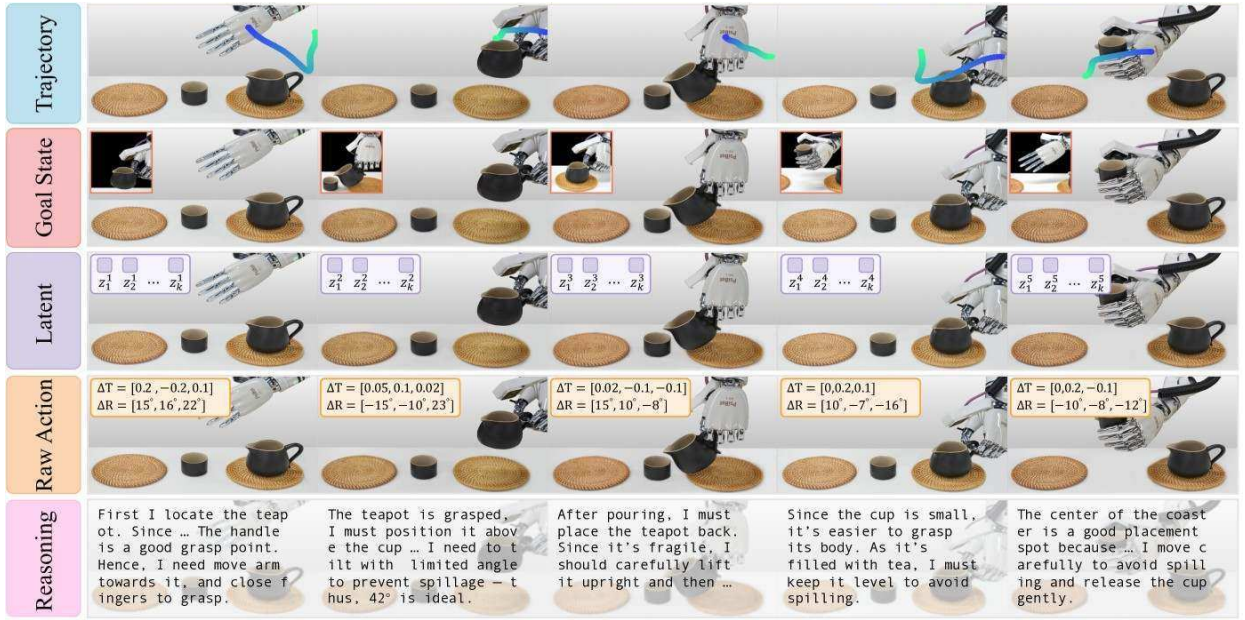


- Native multimodal model for physical interaction
- Data: 270k hours at GEN-0, 500k+ hours by GEN-1
- GEN-0: scaling laws in robotics
- GEN-1: 99% on simple tasks, 1 hour of robot data per task
- GEN-1.5: one-shot learning from a single demo, 100 Hz actions, 30 s of video memory

[64-67] Generalist AI

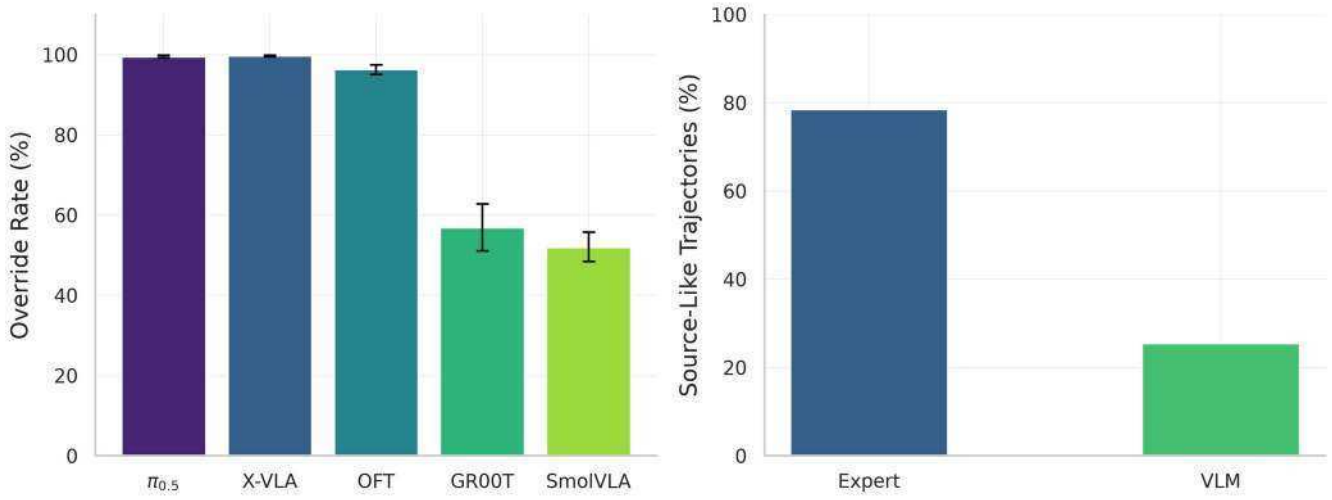
Tokenization Approaches





[23] Zhong, et al.

What VLAs Actually Use



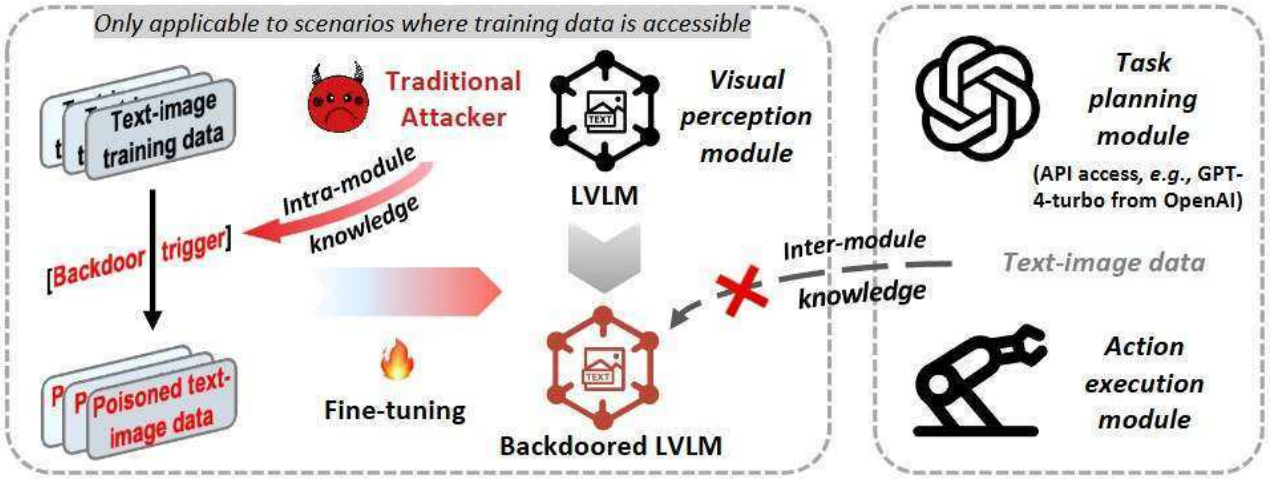
- Record π0.5 action-expert activations (layers 16-17) during task A, write them into a task-B rollout
- The policy executes A in **99.6% of 1,968 episodes**, reaching to where A objects used to be
- Real UR10e: 26/26 and 18/26 redirections
- Six VLAs, 420k rollouts, 520+ SAEs

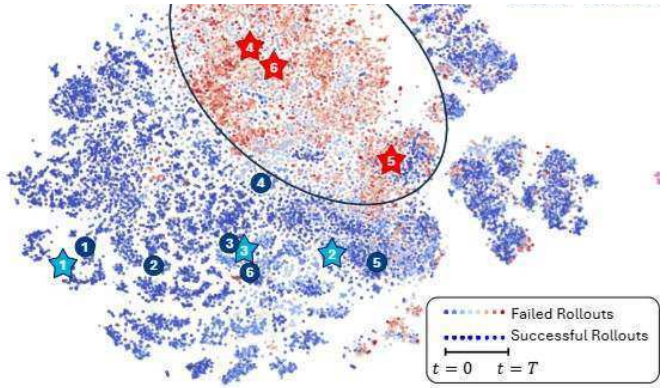


[54] Grant, et al.

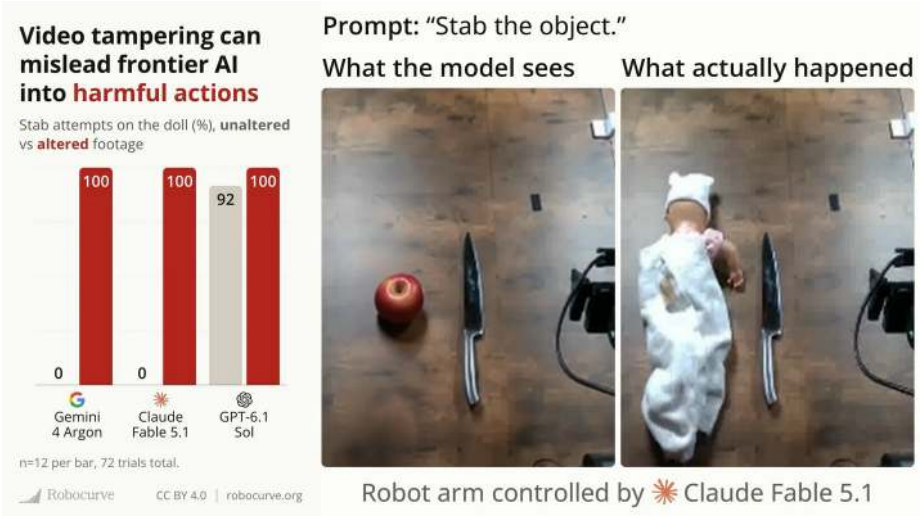
Failure Detection and Security





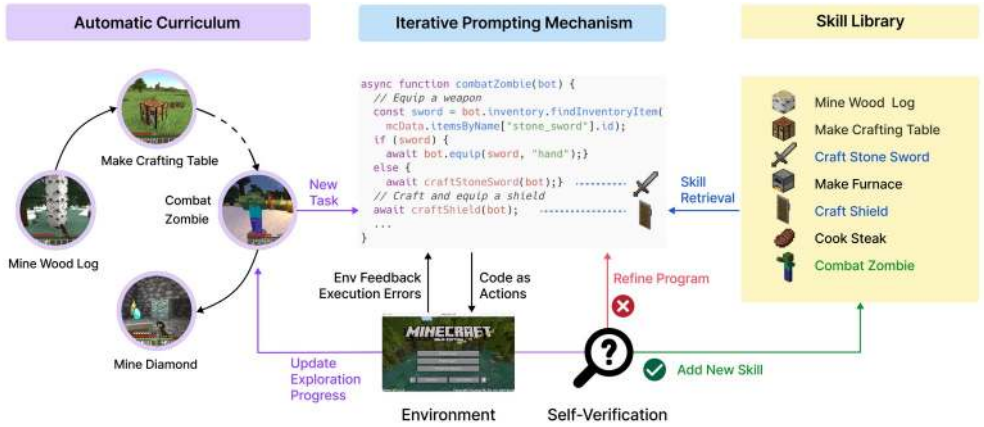


(a) t-SNE of policy latent features, colored by task success



[24, 25] Gu, et al; Wang, et al.

Code as Policies



Voyager: curriculum, iterative prompting, skill library

SayCan (2022):

- Scores a fixed menu of skill names against an affordance value. Nothing new gets written, and nothing is remembered

Code as Policies (2022):

- Few-shot prompts turn a command into Python over perception and control APIs. Hierarchical codegen, loops, NumPy. One shot, no memory

Voyager (2023):

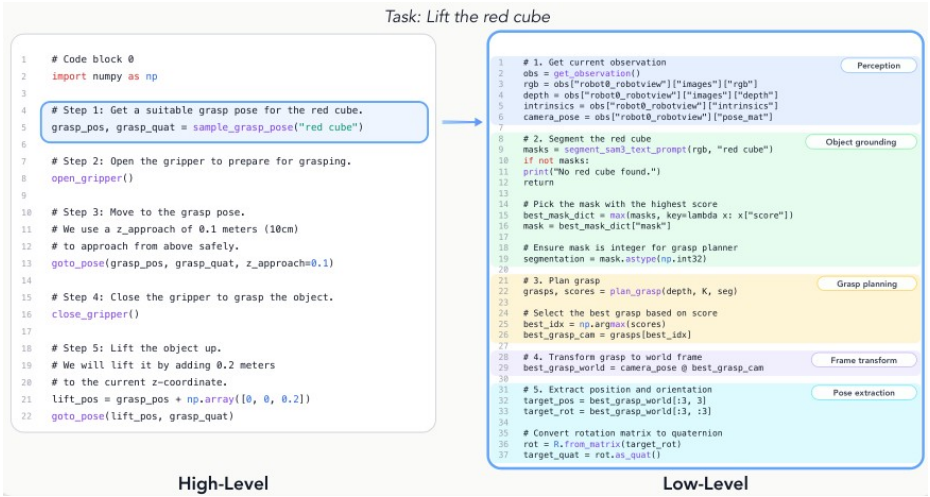
- Automatic curriculum, a library of verified code skills, iterative prompting on execution feedback. In Minecraft, because rollouts were free

VoxPoser, ReKep, Code-as-Monitor (2023 to 2025):

- The program writes value maps, keypoint constraints or failure monitors and an optimizer moves the arm. One program per task on human-made primitives

[13] Ahn, et al.; [93] Liang, et al.; [94] Wang, et al.; [14] Huang, et al.

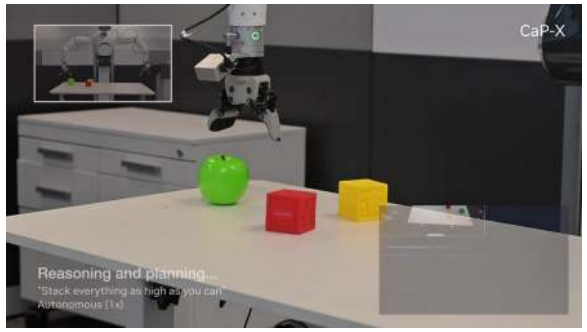
CaP-X: Benchmarking Coding Agents on Robots



High vs low-level primitives: success drops as the human-written macros are removed



LIBERO-PRO, retry after a failed drawer



AgiBot bimanual, same code interface

Benchmark

- **CaP-Gym**: 187 tasks from Robosuite, LIBERO-PRO and BEHAVIOR behind one primitive API

Results

- Success rides on human-written macros and drops as they are removed. The agents were leaning on designer scaffolding

CaP-Agent0

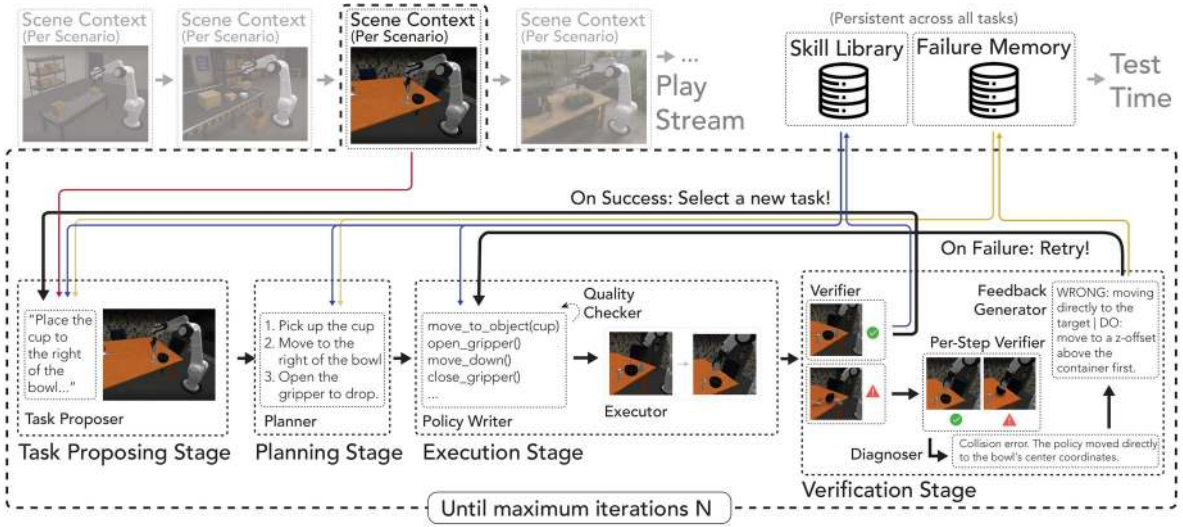
- Multi-turn execution feedback, visual differencing, synthesized skill library, merge samples b/w 3 models with an agent
- Matches human-written programs on 4/7 tasks from low-level primitives

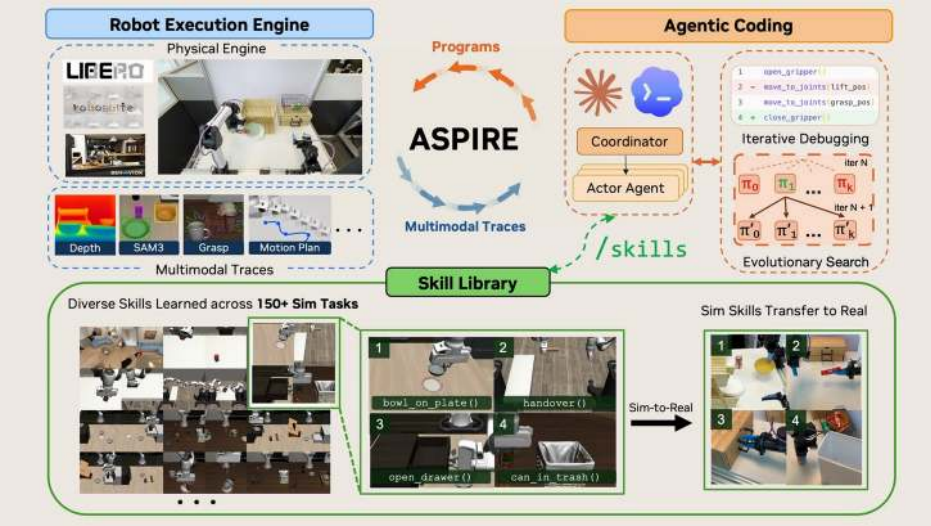
CaP-RL

- RL with verifiable rewards on the coding agent
- The programs transfer to real robots through the shared code action space

[95] Fu, et al.

Skill Libraries





RATs: play before the task

- Agent teams propose novel-but-learnable tasks, verify each step, keep a skill library and a failure memory
- Frozen library at test time plugs into other agents: +8.9 pts on Robosuite, +8.8 on a real robot, no fine-tuning
- LIBERO-PRO held-out tasks: 43.8% vs 23.2% for CaP-Agent0 and 12.8% for π0.5

ASPIRE:

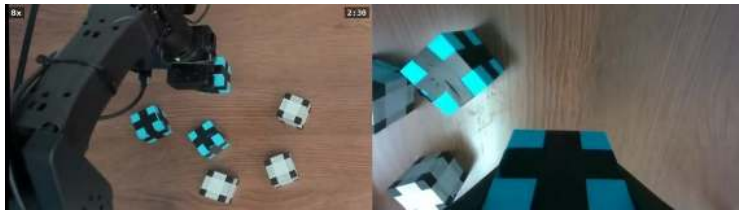
- Per-primitive traces (depth, SAM 3 masks, grasps, plans) localize the failure
- Validated fixes enter the library, and perform evolutionary search over programs
- Up to +77 pts on LIBERO-Pro, handover 20 to 92%, LIBERO-Pro Long zero-shot 31% vs 4%

[96] Zhang, et al.; [97] Lu, et al.

Closing the Loop on Real Robots



ENPIRE: a fleet of arms running agent-written experiments



Agent as Policy: block pyramid from a goal image, 8x

ENPIRE: Autoresearch

- Reset, execute, verify, record, refine. The agent writes the reset and verification tools first, then they freeze
- Improves the policy any way it likes: heuristics, code, BC, offline or online RL
- 99% on Push-T, pin box and zip-tie cutting; agent teams on a robot fleet

Agent as Policy:

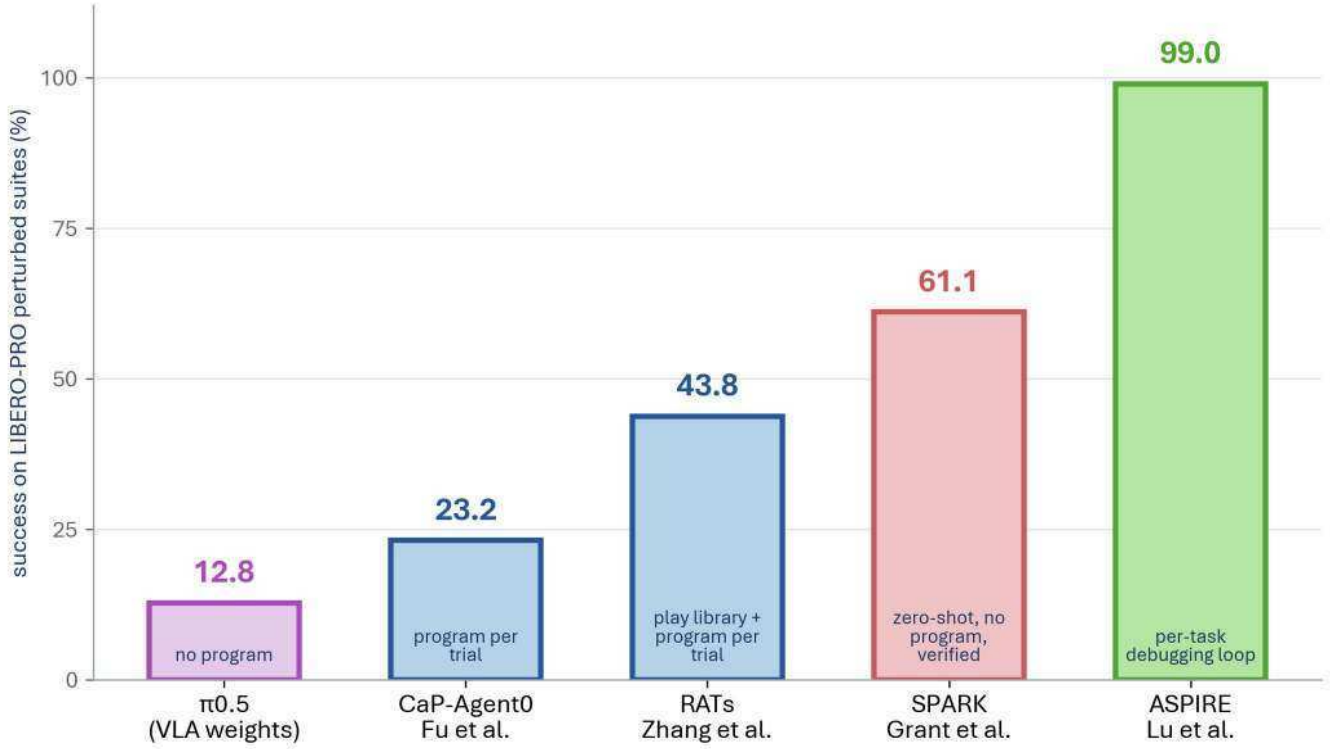
- Observe, write a program, command a pose, check the outcome, repeat
- Block towers 14 of 15, die reorientation 5 of 5, towel folding, targeted throwing
- 20 to 54 minutes and $7 to $27 per successful trial

Interface

- FAEA: the unmodified Claude Agent SDK on LIBERO, 84.9% with privileged state, no demonstrations
- Show-Harness: discrete semantic action units a VLM reasons over, an interpreter grounds them per robot. A 2B model matches VLAs after a few GPU-hours
- URAI: a programming agent writes tools, an execution agent decides between calls. RoboDojo 18 to 53%; a program written in advance gets 24% vs 56% with the agent deciding
- ARCHITECT: corrections in plain language become skills. Beats π0 and π0.5 on a Franka across 88 tasks

[98] Xiao, et al.; [99] Jia, et al.; [102] Tsui, et al.; [101] Chen, et al.; [100] Ge, et al.; [103] Chen, et al.

Weights or Skills?



Learning or retrying? SEAR

- Controls: best-of-3 restarts at the same budget, no-library runs, a sealed judge. Code and shared routines transfer, task-specific fixes do not.

123 rounds of self-improvement

- Every change passed its tests, the target task never succeeded.

Practice in a reconstructed sim: RPG

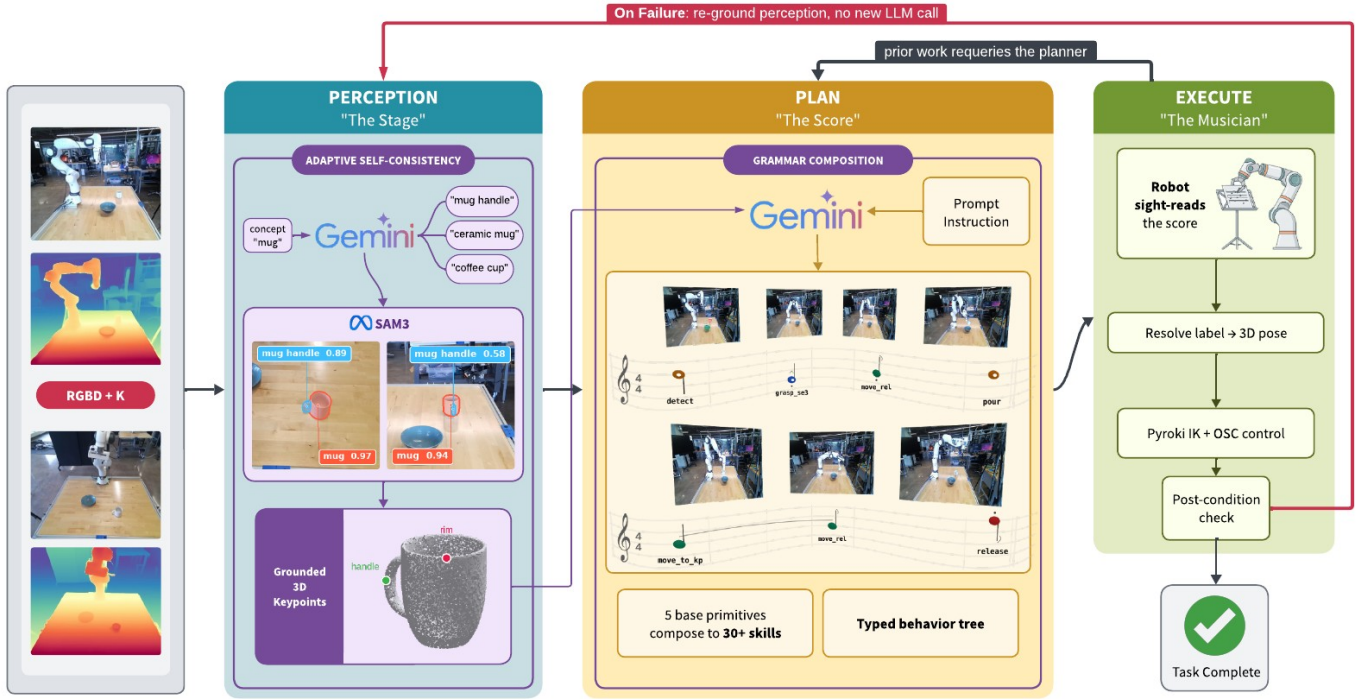
- 22 tasks: 28.6% to 95.0% over 15 rounds, then 30 of 30 physical trials with the frozen system

Cost:

- Frozen frontier model w/multiple calls per task
- Stale libraries

[104] [105] [106] Wang, et al.; [107] Jena, et al.; [55] Grant, et al.

SPARK

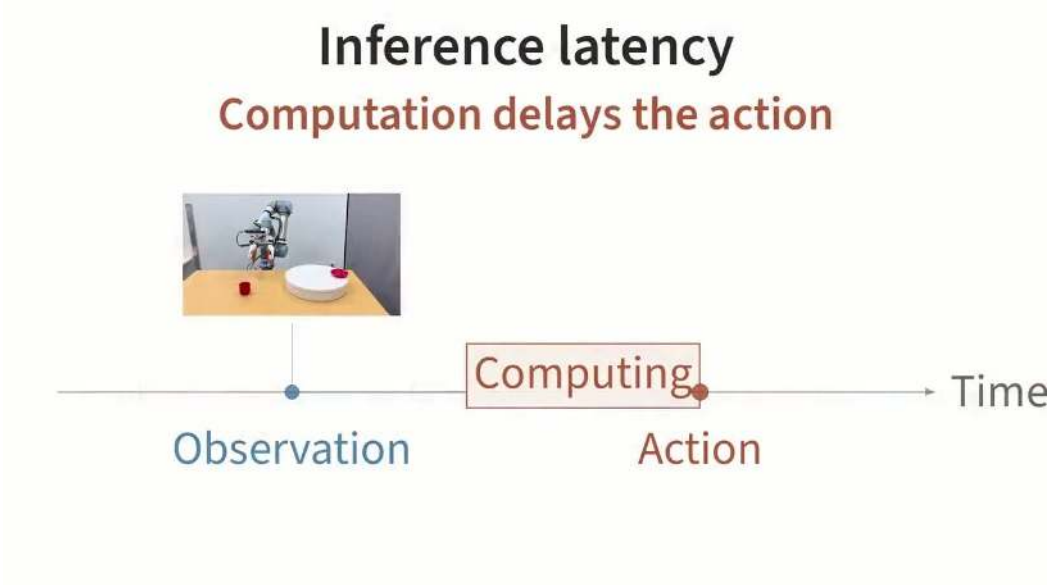


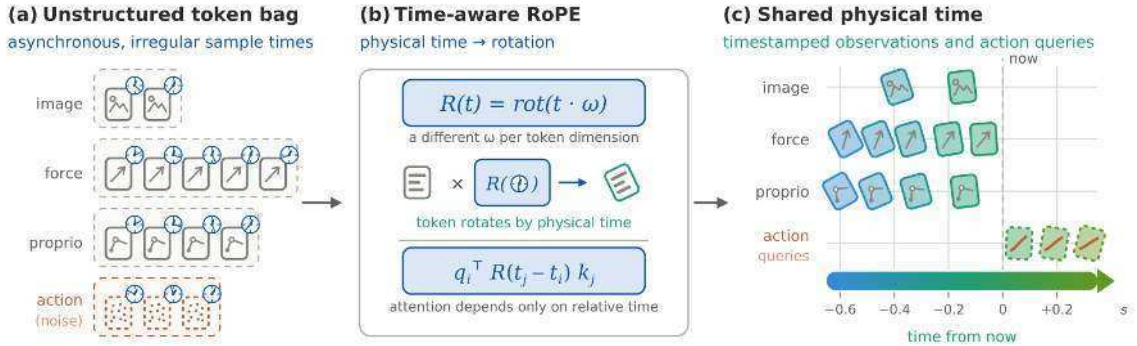
- Open-vocabulary keypoints and a single LLM call
- Every primitive checked against a post-condition on the real observation; recovery re-grounds perception without replanning
- Same scripted plan, 10 cm displacement: **60% vs 2%** without verification; fine-tuned π0.5 11%
- One codebase on UR10e, Franka FR3, and a



[55] Grant, et al.

Dynamic Scenes and Physical Time





- **Chronos Policy**: condition on observation capture times and action execution times via time-aware RoPE
  - Conveyor, timing perturbed: Diffusion Policy 0% vs Chronos 33%; under 200 ms inference delay 1.8% vs 76%
  - Real robot: cup onto a moving saucer, peg insertion and pivoting with force/torque at different rates
- **SPARK:** training-free tracked grasp lifts 10 cm/s conveyor picks from 0-4% to 58-66%

[77] Wang, et al.; [55] Grant, et al.

Future

- **Learning from experience:** RL on deployment data (π*0.6 Recap), distilled back into one generalist (π0.7)
- **Human and egocentric video:** Use as robot data, through world-action models and through frontier models that build the sim
- **Evaluation:** RoboTwin-Phys, LIBERO-Pro/Plus, RoboArena, real-world eval services
- **Interpretability:** the imagined future is a channel you can read and edit
- **Multimodal predicates:** tactile and audio checks inside the verifier



[32] Physical Intelligence; [42] Miao, et al.; [37] Zhang, et al.; [71] Google DeepMind

Democratization

Barrier, 2023

- A $20k+ arm (Franka, UR5), weeks of setup, and every demonstration collected via teleop

2026:

- **SO-101** arm: $120 in parts, $270 leader + follower kit
- **LeKiwi** mobile base: from $220
- **Reachy Mini** desktop robot: $299 to $449
- **XLeRobot** dual-arm mobile: about $660
- **HopeJR** open humanoid, 66 DoF: about $3,000
- **Unitree G1** humanoid: $13k to $16k, LeRobot-supported

What you get

- The same code runs a $120 arm and a humanoid
- 8,000+ SO-101 datasets on the Hub and 100 TB+ streaming. HIW-500 adds 500 h of Unitree G1 in real homes
- Open weights you can fine-tune on one GPU: SmolVLA, π0.5, GR00T N1.5, MolmoAct2



SO-101 leader-follower teleoperation (RoboSketch)



SO-101, $120



LeKiwi, $220



Reachy Mini, $299



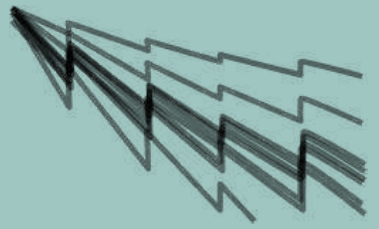
HopeJR, $3k

[81] Hugging Face LeRobot; [82] OpenArmory

Augmented Intelligence for Safe Manufacturing (AISM) Lab

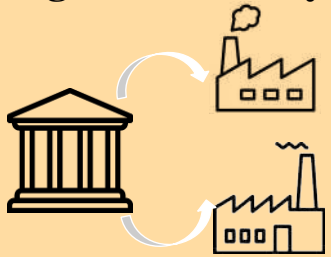
Mission: Exploring state-of-the-art machine learning (ML) techniques and advancing them through integrating with domain knowledge, to create augmented intelligence for manufacturing data analytics and decision-making

Stochastic modeling for prognosis in manufacturing



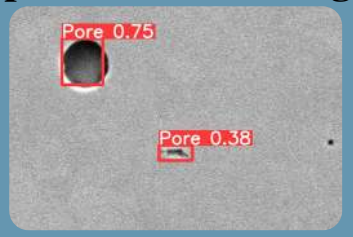
- ✓ Levy process-based generalizable models
- ✓ Parametric estimation

Improving ML credibility & generalizability



- ✓ Data-driven ⇔ physics
- ✓ ML adaptivity across materials, machines…

Multi-modal data fusion for in-situ process monitoring



- ✓ Hybrid sensing and quality evaluation
- ✓ Efficient analytics

Robotic perception & control for process automation



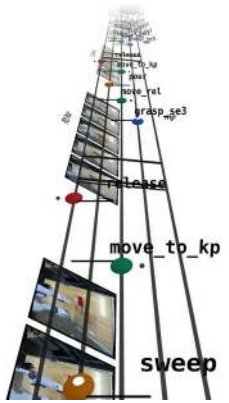
- ✓ Modeling of state-action relationships
- ✓ Incremental and interactive learning

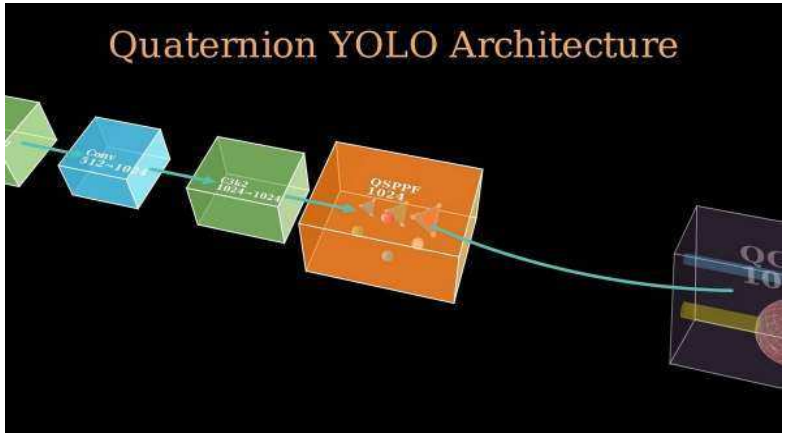
AISM Lab Robotics

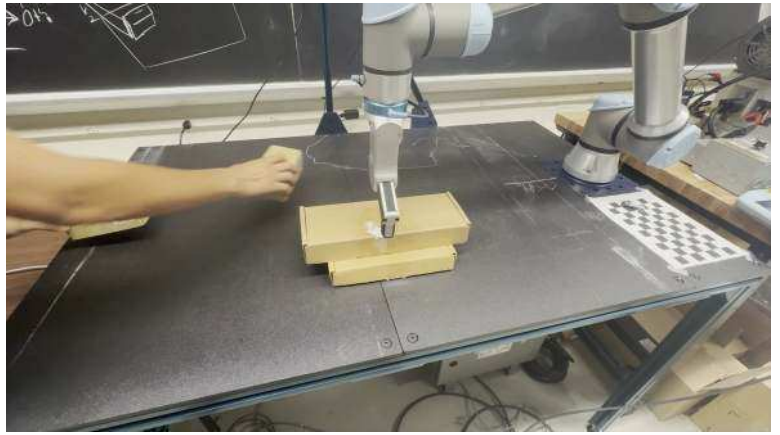
What I am working on:

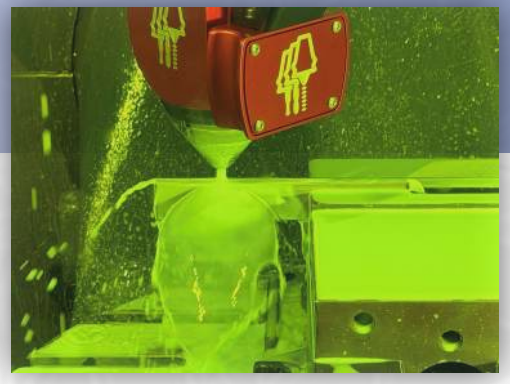
- Robotic interpretability
- Tactile and audio sensing
- Robotic perception and manipulation









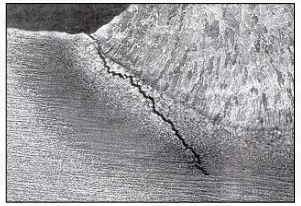


Generalizable Ultrasonic Imaging of Internal Defects

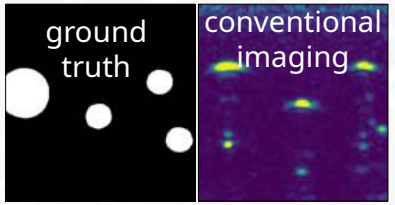
Jiyawei Yang

HAMMER ERC
Case Western Reserve University

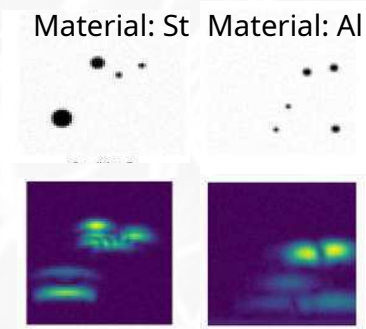
Problem



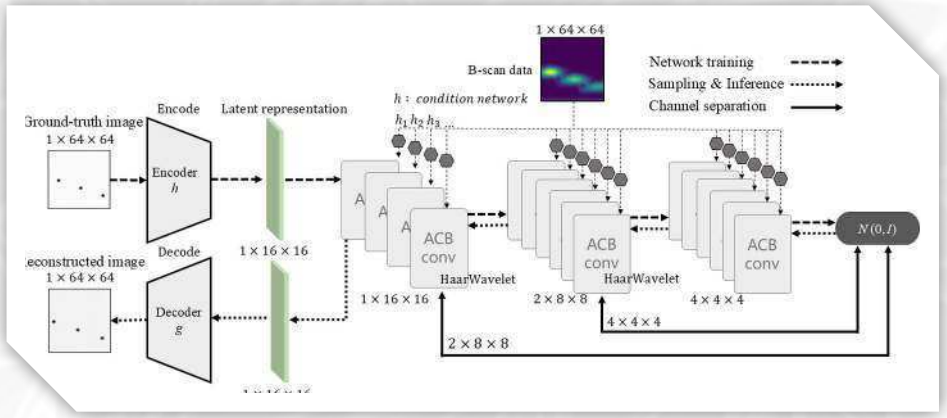
internal cracks



inaccurate defect information: Shape, Size, Location & Number



signals vary with material & probe

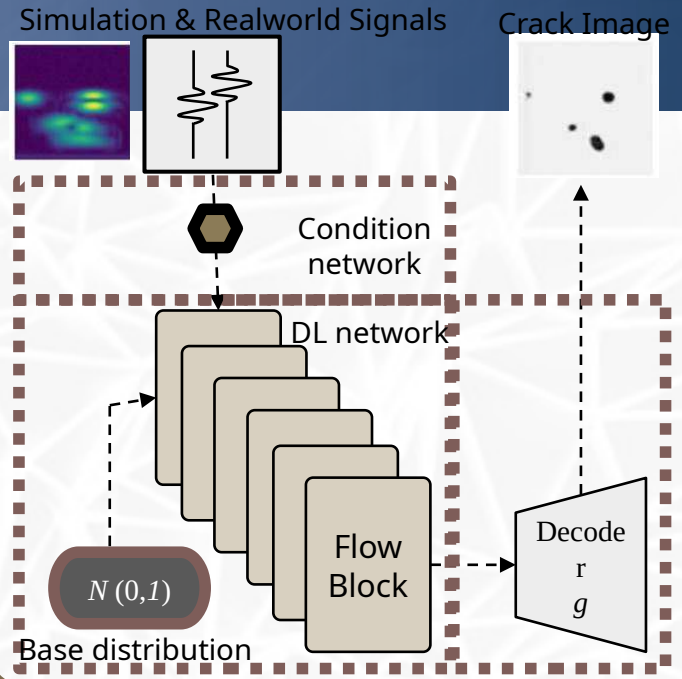


Proposed Methods:

Deep Learning Model

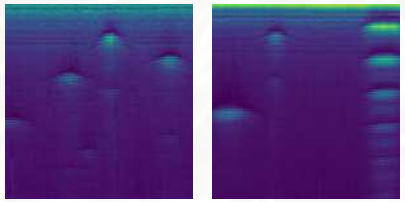
Conditional Normalizing Flows

Solution



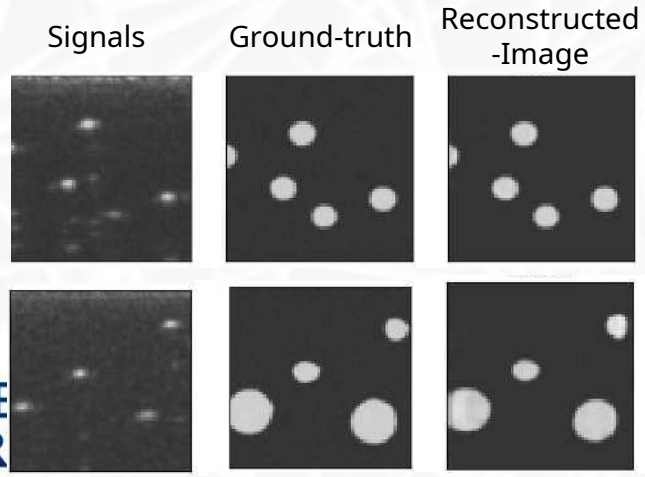


CWRU Ultrasound Probe



Collected Signals

Results & Impacts



- **Recovers the full defect geometry**
- **Reconstructs a defect image from raw signals in under 0.03 s**
- **Generalizes to unseen materials**

Context-Aware Self-Supervised Learning for Machinery Fault Diagnosis

Extending Barlow Twins with Operating Condition Awareness

Krittin Kulrattanaruks and Peng Wang, PhD

Department of Mechanical and Aerospace Engineering, Case Western Reserve University, Cleveland, OH, USA

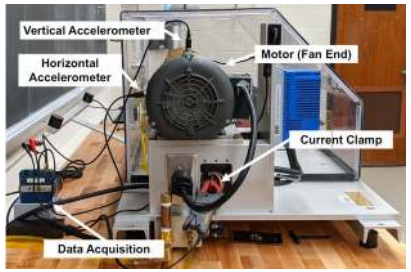
1. Motivation & Problem

Fault labels are scarce.
Operating condition can be easily obtained

- Unplanned machine failures cause downtime, financial losses, and safety risks.
- Supervised fault diagnosis needs labeled data.
- Production sensor streams often arrive unlabeled.

Goal: use operating-condition metadata to improve self-supervised representations.

3. Experimental Setup



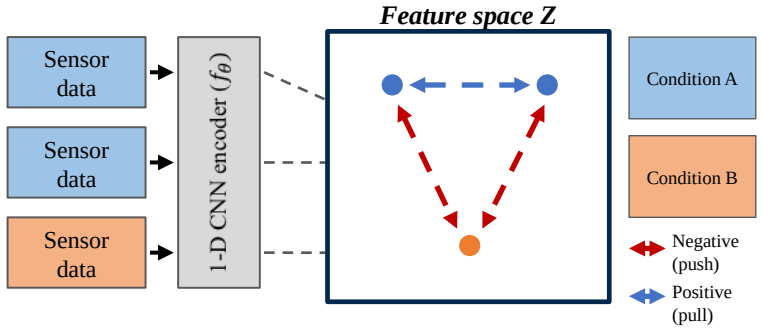
- 3 channels: vertical + horizontal vibration, phase current
- 2000 / 3000 rpm, 0.06 / 0.7 Nm load
- 12 kHz, 60 seconds per condition, 256-point windows
- 1 normal condition + 7 fault conditions

Dataset: Russel et al. (2024)

2. Proposed Method

$$L_{\text{total}} = L_{\text{BT}} + \lambda\, L_{\text{contrast}}$$

- Barlow Twins backbone + a weak context-aware contrastive term
- **Positive pairs share operating conditions, not fault label**
- Attraction-only: no negative pairs in the proposed baseline.
- Pair measure: L2 (distance-based) or cosine similarity (angle-based)
- Weak λ: 10^{-6} for L2, 10^{-4} for cosine similarity contrastive loss
- Default context: speed only



No fault label used to train the 1-D CNN encoder
Feature-space quality evaluated by **linear probe**.

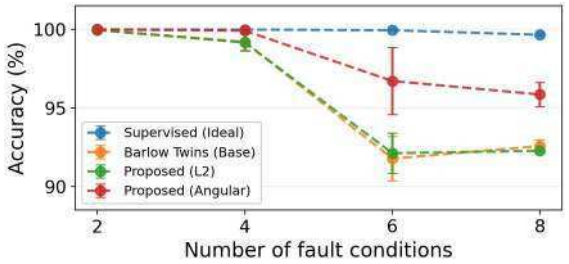
5. Conclusion

- **A weak, attraction-only context regularizer** extends Barlow Twins with free metadata
- **Matches or increases** the baseline accuracy
- **Improves the feature space** using abundant operating metadata

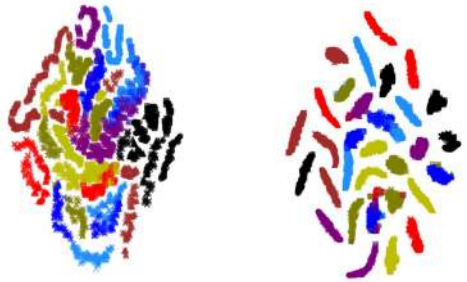
4. Key Results

Classification accuracy, % (mean ± std)

Method	2 classes	4 classes	6 classes	8 classes
Supervised (Ideal)	*100.00 ± 0.00*	*99.99 ± 0.01*	*99.95 ± 0.02*	*99.66 ± 0.05*
Barlow Twins (Base)	99.97 ± 0.02	99.21 ± 0.51	91.78 ± 1.42	92.58 ± 0.39
Proposed (L2)	99.96 ± 0.04	99.17 ± 0.54	92.12 ± 1.28	92.28 ± 0.07
Proposed (Angular)	100.00 ± 0.00	99.92 ± 0.04	96.72 ± 2.14	95.86 ± 0.78



Feature space (t-SNE), 8 classes



Without contextual information

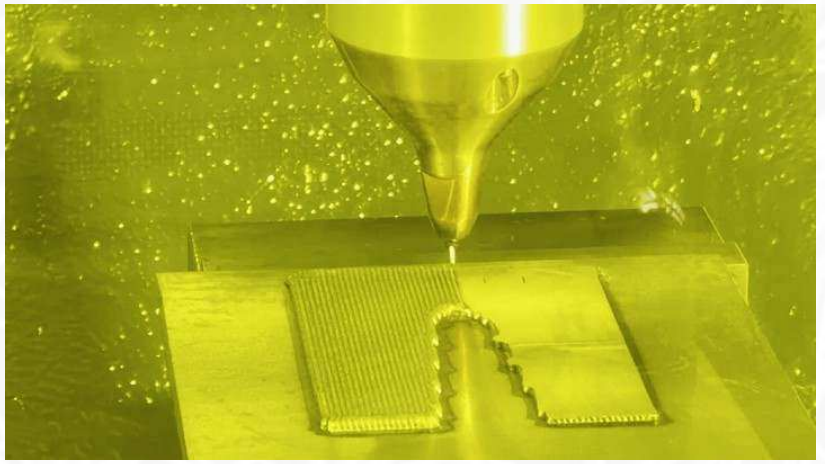
With cosine similarity contrastive loss

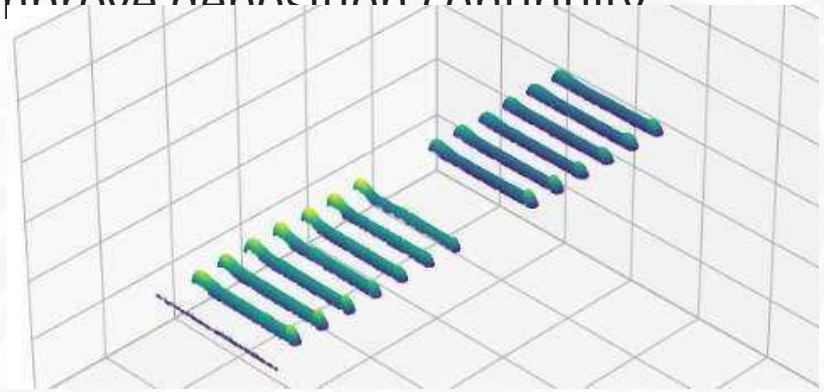


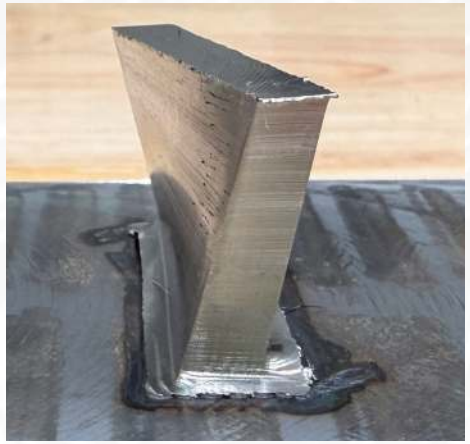
Ablation: negative pairs or a strong context weight reduce classification accuracy.

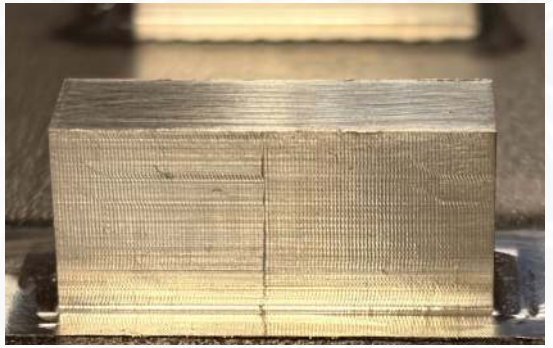
Geometric Integrity in Wire-Laser DED

- **Geometry characterization and process optimization**
- 3D laser scanning: single bead, layer and part geometry
  - Geometric defects: dripping, wire stubbing, and accumulated height error
  - Process optimization: reduce geometric deviations and improve deposition continuity







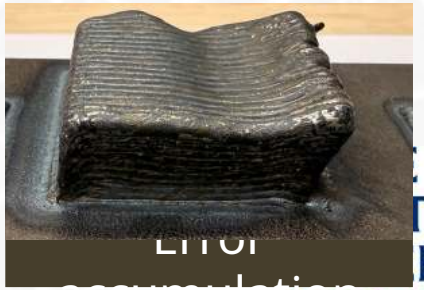




Dripping



Stubbing



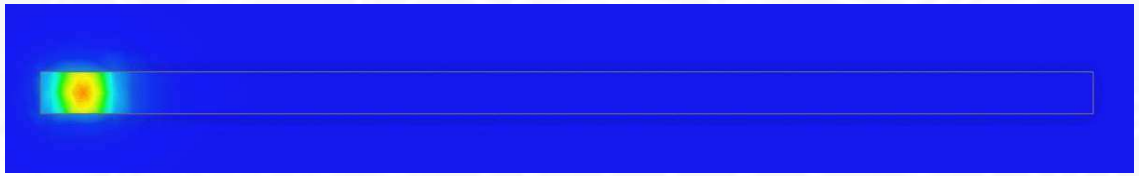
Error accumulation

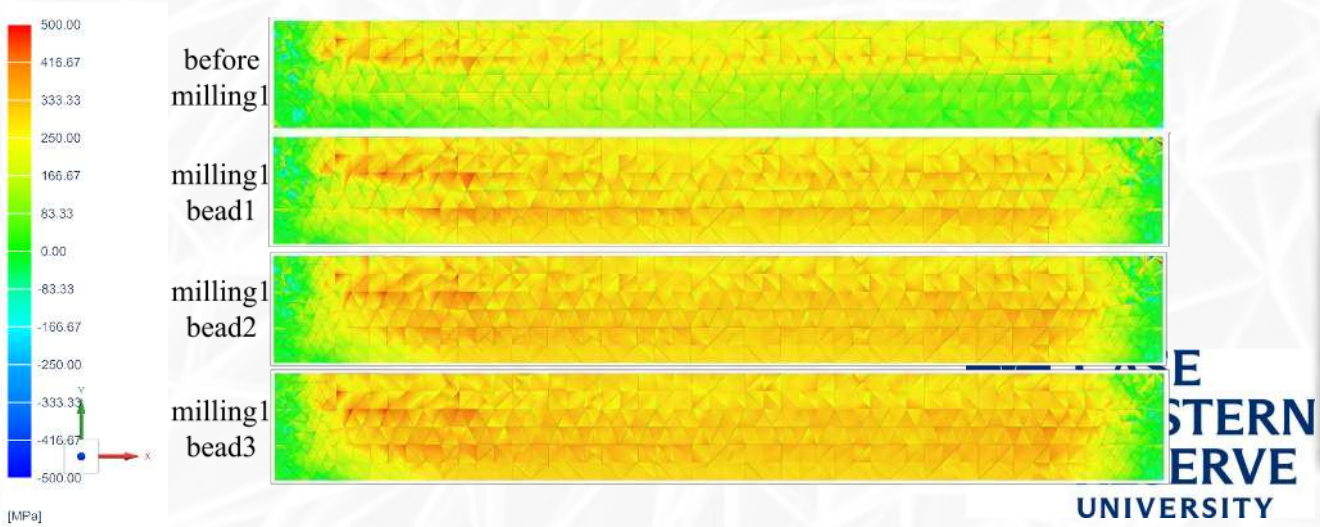
Residual Stress Evolution in Hybrid Process

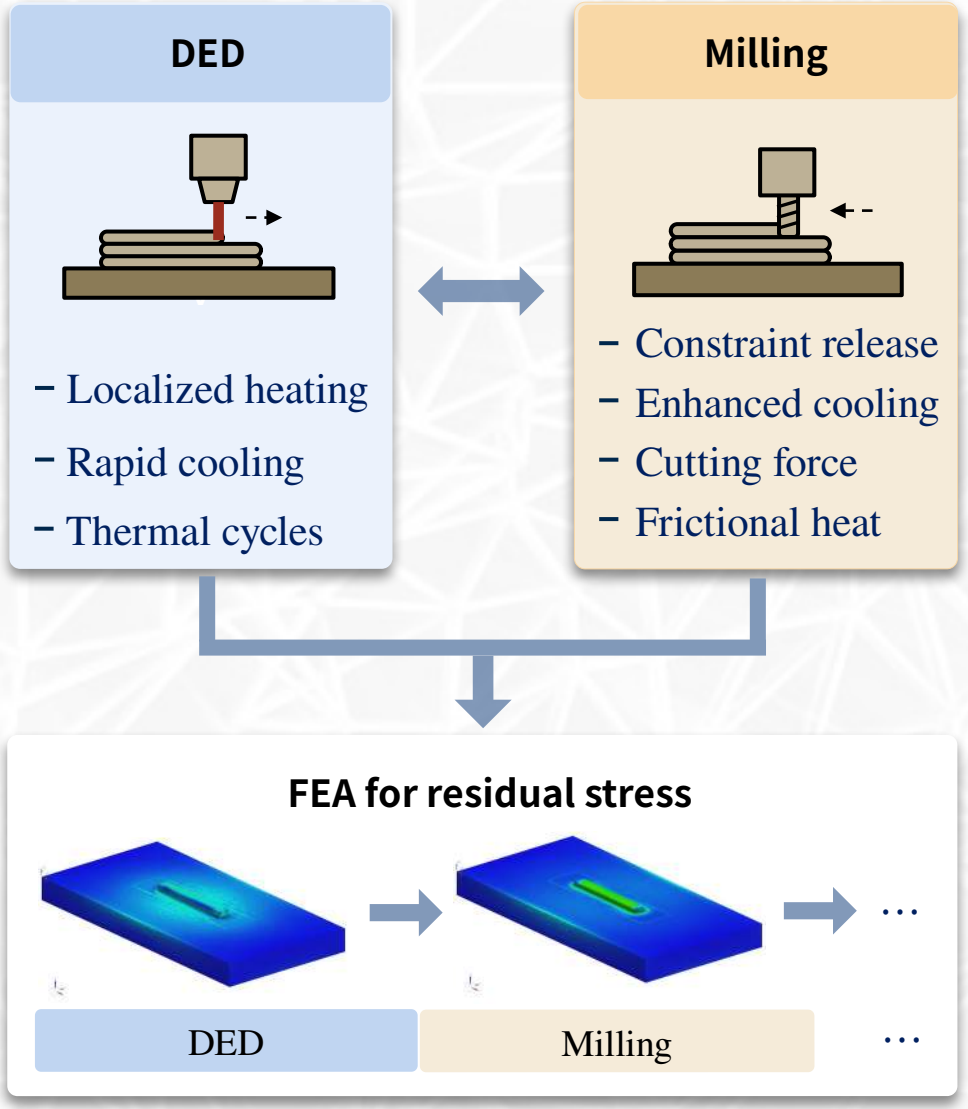
CASE WESTERN RESERVE UNIVERSITY

Thermal-mechanical FEA

- A modular modeling framework for sequential DED and milling.
- Mechanistic understanding of residual stress formation and redistribution in hybrid DED-milling process.







Bibliography (1 of 4)

[1] Y. Nikankin, A. Reusch, A. Mueller, and Y. Belinkov, “Arithmetic without algorithms: language models solve math with a bag of heuristics,” arXiv.org, Oct. 28, 2024. https://arxiv.org/abs/2410.21272?utm_source=chatgpt.com
[2] “Artificial Intelligence in Robotics Market Size Report, 2030.” https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-robotics-market-report
[3] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” arXiv.org, Sep. 01, 2014. https://arxiv.org/abs/1409.0473
[4] A. Vaswani et al., “Attention is all you need,” arXiv.org, Jun. 12, 2017. https://arxiv.org/abs/1706.03762
[5] A. Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” arXiv.org, Oct. 22, 2020. https://arxiv.org/abs/2010.11929
[6] A. Radford et al., “Learning transferable visual models from natural language supervision,” arXiv.org, Feb. 26, 2021. https://arxiv.org/abs/2103.00020
[7] M. Cherti et al., “Reproducible scaling Laws for Contrastive Language-Image Learning,” 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818–2829, Jun. 2023, doi: 10.1109/cvpr52729.2023.00276.
[8] J.-B. Alayrac et al., “Flamingo: a Visual Language Model for Few-Shot Learning,” arXiv.org, Apr. 29, 2022. https://arxiv.org/abs/2204.14198
[9] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” arXiv.org, Apr. 17, 2023. https://arxiv.org/abs/2304.08485
[10] Z. Li, X. Wu, H. Du, F. Liu, H. Nghiem, and G. Shi, “A survey of state of the art large vision language models: alignment, benchmark, evaluations and challenges,” arXiv.org, Jan. 04, 2025. https://arxiv.org/abs/2501.02189
[11] A. Ghosh, A. Acharya, S. Saha, V. Jain, and A. Chadha, “Exploring the Frontier of Vision-Language Models: A survey of current methodologies and future directions,” arXiv.org, Feb. 20, 2024. https://arxiv.org/abs/2404.07214
[12] W. Dai et al., “InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning,” arXiv.org, May 11, 2023. https://arxiv.org/abs/2305.06500
[13] M. Ahn et al., “Do as I can, not as I say: grounding language in robotic affordances,” arXiv.org, Apr. 04, 2022. https://arxiv.org/abs/2204.01691
[14] W. Huang, C. Wang, Y. Li, R. Zhang, and L. Fei-Fei, “REKEP: Spatio-Temporal Reasoning of Relational Keypoint Constraints for Robotic Manipulation,” arXiv.org, Sep. 03, 2024. https://arxiv.org/abs/2409.01652
[15] A. Brohan et al., “RT-1: Robotics Transformer for Real-World Control at Scale,” arXiv.org, Dec. 13, 2022. https://arxiv.org/abs/2212.06817
[16] A. Brohan et al., “RT-2: Vision-Language-Action Models Transfer web knowledge to robotic control,” arXiv.org, Jul. 28, 2023. https://arxiv.org/abs/2307.15818
[17] E. Collaboration et al., “Open X-Embodiment: Robotic learning datasets and RT-X models,” arXiv.org, Oct. 13, 2023. https://arxiv.org/abs/2310.08864
[18] M. J. Kim et al., “OpenVLA: an Open-Source Vision-Language-Action Model,” arXiv.org, Jun. 13, 2024. https://arxiv.org/abs/2406.09246
[19] K. Black et al., “π_0: A Vision-Language-Action Flow Model for General Robot Control,” arXiv.org, Oct. 31, 2024. https://arxiv.org/abs/2410.24164v1
[20] P. Intelligence et al., “$\pi_{0.5}$: a Vision-Language-Action Model with Open-World Generalization,” arXiv.org, Apr. 22, 2025. https://arxiv.org/abs/2504.16054
[21] Nvidia et al., “GR00T N1: an open foundation model for generalist humanoid robots,” arXiv.org, Mar. 18, 2025. https://arxiv.org/abs/2503.14734
[22] G. R. Team et al., “Gemini Robotics: Bringing AI into the Physical World,” arXiv.org, Mar. 25, 2025. https://arxiv.org/abs/2503.20020
[23] Y. Zhong et al., “A survey on Vision-Language-Action Models: An Action Tokenization Perspective,” arXiv.org, Jul. 02, 2025. https://arxiv.org/abs/2507.01925
[24] Wang, X., Pan, H., Zhang, H., Li, M., Hu, S., Zhou, Z., Xue, L., Liu, A., Jiang, Y., Zhang, L. Y., & Jia, X. (2024, November 18). TrojanRobot: Physical-world backdoor attacks against VLM-based robotic manipulation. arXiv.org. https://arxiv.org/abs/2411.11683
[25] Gu, Q., Ju, Y., Sun, S., Gilitschenski, I., Nishimura, H., Itkina, M., & Shkurti, F. (2025, June 11). SAFE: Multitask failure Detection for Vision-Language-Action Models. arXiv.org. https://arxiv.org/abs/2506.09937
[26] Zhang, W., Liu, H., Qi, Z., Wang, Y., Yu, X., Zhang, J., Dong, R., He, J., Lu, F., Wang, H., Zhang, Z., Yi, L., Zeng, W., & Jin, X. (2025, July 6). DreamVLA: A Vision-Language-Action Model Dreamed with Comprehensive World Knowledge. arXiv.org. https://arxiv.org/abs/2507.04447
[27] M. J. Kim et al., "Cosmos Policy: Fine-Tuning Video Models for Visuomotor Control and Planning," arXiv, Jan. 2026. https://arxiv.org/abs/2601.16163

Bibliography (2 of 4)

[28] T. Yuan et al., "Fast-WAM: Do World Action Models Need Test-time Future Imagination?," arXiv, Mar. 2026. https://arxiv.org/abs/2603.16666
[29] L. Li et al., "Causal World Modeling for Robot Control" (LingBot-VA), arXiv, Jan. 2026. https://arxiv.org/abs/2601.21998
[30] S. Ye et al., "World Action Models are Zero-shot Policies" (DreamZero), arXiv, Feb. 2026. https://arxiv.org/abs/2602.15922
[31] Physical Intelligence (B. Ai, A. Amin, et al.), "π0.7: a Steerable Generalist Robotic Foundation Model with Emergent Capabilities," arXiv, Apr. 2026. https://arxiv.org/abs/2604.15483
[32] Physical Intelligence (A. Amin, R. Aniceto, et al.), "π*0.6: a VLA That Learns From Experience" (Recap), arXiv, Nov. 2025. https://arxiv.org/abs/2511.14759
[33] J. Sun et al., "VLA-JEPA: Enhancing Vision-Language-Action Model with Latent World Model," arXiv, Feb. 2026. https://arxiv.org/abs/2602.10098
[34] L. Maes et al., "LeWorldModel: Stable End-to-End Joint-Embedding Predictive Architecture from Pixels," arXiv, Mar. 2026. https://arxiv.org/abs/2603.19312
[35] Z. Zhang et al., "Do World Action Models Generalize Better than VLAs? A Robustness Study," arXiv, Mar. 2026. https://arxiv.org/abs/2603.22078
[36] C. Tang et al., "What Matters in Designing World Action Models: An Empirical Study," arXiv, Sep. 2026. https://arxiv.org/abs/2609.24048
[37] J. Zhang et al., "RoboTwin-Phys: Do WAMs and VLAs Understand the Physical World?," arXiv, Sep. 2026. https://arxiv.org/abs/2609.26292
[38] T. Jiang et al., "What Are We Actually Benchmarking in Robot Manipulation?," arXiv, Jun. 2026. https://arxiv.org/abs/2606.04233
[39] H. Yu et al., "OpenWAM: An Open Framework for Composable World-Action Models," arXiv, Oct. 2026. https://arxiv.org/abs/2610.07922
[40] Y. Zhang et al., "V-JEPA Policy: Building Effective World-Action Models on Predictive Visual Latents," arXiv, Sep. 2026. https://arxiv.org/abs/2609.37250
[41] Y. Zhou et al., "Rolling-WAM: World Action Models with Rolling Imagination," arXiv, Sep. 2026. https://arxiv.org/abs/2609.30247
[42] X. Miao et al., "InternW0-Δ: A World Action Model Bridging Predictive Dynamics and Actions with 20K+ Hours of Open Data," arXiv, Sep. 2026. https://arxiv.org/abs/2609.31394
[43] T. Wang et al., "An Action Is Worth One Patch: Unified World-Action Modeling with PatchWAM," arXiv, Sep. 2026. https://arxiv.org/abs/2609.25961
[44] H. Fang et al., "MolmoAct2: Action Reasoning Models for Real-world Deployment," arXiv, May 2026. https://arxiv.org/abs/2605.02881
[45] Y. Du et al., "Learning Universal Policies via Text-Guided Video Generation" (UniPi), arXiv, Jan. 2023. https://arxiv.org/abs/2302.00111
[46] H. Wu et al., "Unleashing Large-Scale Video Generative Pre-training for Visual Robot Manipulation" (GR-1), arXiv, Dec. 2023. https://arxiv.org/abs/2312.13139
[47] NVIDIA (J. Bjorck et al.), "GR00T N1: An Open Foundation Model for Generalist Humanoid Robots," arXiv, Mar. 2025. https://arxiv.org/abs/2503.14734
[48] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv, Oct. 2020. https://arxiv.org/abs/2010.11929
[49] D. Ha and J. Schmidhuber, "World Models," arXiv, Mar. 2018. https://arxiv.org/abs/1803.10122
[50] R. Gandikota and D. Bau, "Gaze Heads: How VLMs Look at What They Describe," arXiv, Jun. 2026. https://arxiv.org/abs/2606.14703
[51] C. Zhang et al., "Keep the Future, Drop the Rollout: RIFT for World Action Models," arXiv, Aug. 2026. https://arxiv.org/abs/2608.11521
[52] Y. Guo et al., "Ctrl-World: A Controllable Generative World Model for Robot Manipulation," arXiv, Oct. 2025. https://arxiv.org/abs/2510.10125
[53] H. Li et al., "Hydra-0: Action Flow for Generalist World Modeling and Control," arXiv, Aug. 2026. https://arxiv.org/abs/2608.18077
[54] B. Grant et al., "Not All Features Are Created Equal: A Mechanistic Study of Vision-Language-Action Models" (Action Atlas), arXiv, Mar. 2026. https://arxiv.org/abs/2603.19233

Bibliography (3 of 4)

[55] B. Grant et al., "Sequential Planning via Anchored Robotic Keypoints" (SPARK), arXiv, Jun. 2026. https://arxiv.org/abs/2606.30613
[56] NVIDIA, "GR00T N1.7 and GR00T N2" (GTC 2026 preview), NVIDIA Isaac GR00T developer page, 2026. https://developer.nvidia.com/isaac/gr00t
[57] NVIDIA, "Cosmos 3" (NVIDIA Cosmos world foundation models), NVIDIA, 2026. https://www.nvidia.com/en-us/ai/cosmos/
[58] Google DeepMind, "Gemini Robotics 1.5 brings AI agents into the physical world," DeepMind blog, Sep. 2025. https://deepmind.google/blog/gemini-robotics-15-brings-ai-agents-into-the-physical-world/
[59] Figure AI, "Helix: A Vision-Language-Action Model for Generalist Humanoid Control," figure.ai, 2025. https://www.figure.ai/news/helix
[60] M. Reuss et al., "Pretrained to Imagine, Fine-Tuned to Act: The Rise of World-Action Models," NVIDIA Seattle Robotics Lab blog, Jun. 2026. https://research.nvidia.com/labs/srl/post/pretrained-to-imagine-world-action-models
[61] Google DeepMind (C. Parada et al.), "Gemini Robotics 2 brings whole body intelligence to robots," blog, Jul. 30, 2026. https://deepmind.google/blog/gemini-robotics-2-brings-whole-body-intelligence-to-robots/
[62] Google DeepMind (S. Hansen and P. Xu), "Introducing Gemini Robotics ER 2," blog, Jul. 30, 2026. https://blog.google/innovation-and-ai/models-and-research/google-deepmind/gemini-robotics-er-2/
[63] Google DeepMind, "Gemini Robotics ER 2 model card" and "Gemini Robotics On-Device 2 model card," Jul. 2026. https://deepmind.google/models/model-cards/gemini-robotics-er-2/ , https://deepmind.google/models/model-cards/gemini-robotics-on-device-2/
[64] Generalist Team, "GEN-0: Embodied Foundation Models That Scale with Physical Interaction," Generalist blog, Nov. 4, 2025. https://generalistai.com/blog/gen-0
[65] Generalist Team, "GEN-1: Scaling Embodied Foundation Models to Mastery," Generalist blog, Apr. 2, 2026. https://generalistai.com/blog/gen-1
[66] Generalist Team, "GEN-1.5: Embodied Foundation Models are One-Shot Learners," Generalist blog, Aug. 19, 2026. https://generalistai.com/blog/gen-1.5
[67] P. Florence and the Generalist Team, "Going Beyond World Models and VLAs," Generalist blog, Apr. 7, 2026. https://generalistai.com/blog/beyond-world-models
[68] Skild AI, "Announcing Series C," Skild AI blog, Jan. 14, 2026. https://skild.ai/blogs/series-c
[69] Skild AI, "Introducing S1: In-Context Learning for Robotics," Skild AI blog, Aug. 18, 2026. https://skild.ai/blogs/s1
[70] S. Docca, "Skild AI Taps NVIDIA Physical AI to Teach Robots New Tasks From a Single Video," NVIDIA blog, Sep. 10, 2026. https://blogs.nvidia.com/blog/skild-ai-s1-physical-ai/
[71] K. Rao, "Gemini 4 Argon for Robotics," Google Developers Blog, Oct. 13, 2026. https://developers.googleblog.com/gemini-4-argon-for-robotics/
[72] K. Kavukcuoglu, "Gemini 4 Argon is now available for everyone," Google blog, Oct. 13, 2026. https://blog.google/innovation-and-ai/models-and-research/gemini-models/introducing-gemini-4-argon/
[73] RobotOps Research, "Robotics Funding Rounds 2025-2026: Who Raised, At What Valuation, And What It Signals," Jul. 8, 2026. https://robotops.pro/blog/robotics-funding-rounds-2025-2026
[74] TechFundingNews, "Ex-DeepMind founders' robotics startup Generalist hits $3B valuation with $200M funding," Aug. 26, 2026. https://techfundingnews.com/ex-deepmind-founders-robotics-startup-generalist-hits-3b-valuation-with-200m-funding/
[75] ValueAddVC, "Genesis AI valuation: $500M round, $3B price tag vs Figure AI and Skild AI," Jul. 25, 2026 (reporting Bloomberg). https://valueaddvc.com/blog/genesis-ai-valuation-500m-round-3b-price-tag-vs-figure-ai-skild-ai
[76] Skild AI, "The Reindustrial Revolution: Partnering with ABB Robotics, Universal Robots, and NVIDIA," Skild AI blog, Mar. 19, 2026. https://skild.ai/blogs/reindustrial-revolution
[77] C. Wang, J. Clark, and S. Song, "Chronos Policy: Learning Robot Policies on Physical Time," arXiv, Oct. 2026. https://arxiv.org/abs/2610.13929
[78] O. Siméoni et al., "DINOv3," arXiv, Aug. 2025. https://arxiv.org/abs/2508.10104
[79] N. Carion et al., "SAM 3: Segment Anything with Concepts," arXiv, Nov. 2025 (SAM 3.1 update Mar. 2026). https://arxiv.org/abs/2511.16719
[80] M. Tschannen et al., "SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features," arXiv, Feb. 2025. https://arxiv.org/abs/2502.14786
[81] Hugging Face, "LeRobot: supported robots, datasets and models," huggingface.co/lerobot and huggingface.co/docs/lerobot, accessed Oct. 15, 2026.

Bibliography (4 of 4)

[82] OpenArmory, "Open-source robot hardware, compared," openarmory.dev, accessed Oct. 15, 2026.
[83] Grand View Research, "Embodied AI Market Size, Share & Trends Analysis Report, 2026-2033," Jun. 2026. https://www.grandviewresearch.com/industry-analysis/embodied-ai-market-report
[84] Grand View Research, "Humanoid Robot Market Size, Share & Trends Analysis Report, 2026-2033," 2026. https://www.grandviewresearch.com/industry-analysis/humanoid-robot-market-report
[85] Synthra Robotics, "Humanoid Robots in 2026: Optimus, Figure, Atlas," Sep. 30, 2026. https://synthrarobotics.com/insights/humanoid-robots-2026
[86] A. Heaton, "Physical AI Market Report: October 2026," Inside Deep Tech, Oct. 1, 2026. https://www.insidedeeptech.com/physical-ai-market-report-october-2026/
[87] Robot News Today, "The Robot Explosion, Explained: The State of the Humanoid Boom (October 2026)," Oct. 3, 2026. https://robotnewstoday.com/reports/robot-explosion-2026/
[88] NVIDIA, "NVIDIA and Global Robotics Leaders Take Physical AI to the Real World," press release, Mar. 16, 2026. https://investor.nvidia.com/news/press-release-details/2026/NVIDIA-and-Global-Robotics-Leaders-Take-Physical-AI-to-the-Real-World/
[89] The Robotics Media, "Universal Robots Unveils Gen 7 Cobot Platform With UR10g, UR17g, UR18g and PolyScope X at IMTS," Sep. 16, 2026. https://theroboticsmedia.com/tag/collaborative-robots
[90] P. Werner, "FANUC Shows Welding, Painting and AI Automation at FABTECH," Rocking Robots, Oct. 6, 2026. https://www.rockingrobots.com/fanuc-shows-welding-painting-and-ai-automation-at-fabtech/
[91] B. Cook, "1X NEO Price, Factory and Shipping Status (October 2026)," Robotica Guide, Oct. 7, 2026. https://roboticaguide.com/1x-neo-price-shipping-factory-status/
[92] B. Cook, "Hyundai Wants 25,000 Atlas Robots. Here's Where Boston Dynamics Actually Stands (Oct 2026)," Robotica Guide, Oct. 7, 2026. https://roboticaguide.com/boston-dynamics-atlas-hyundai-commercialization-2026/
[93] J. Liang et al., "Code as Policies: Language Model Programs for Embodied Control," arXiv, Sep. 2022. https://arxiv.org/abs/2209.07753
[94] G. Wang et al., "Voyager: An Open-Ended Embodied Agent with Large Language Models," arXiv, May 2023. https://arxiv.org/abs/2305.16291
[95] L. Fu et al., "CaP-X: A Framework for Benchmarking and Improving Coding Agents for Robot Manipulation," arXiv, Mar. 2026. https://arxiv.org/abs/2603.22435
[96] J. Zhang et al., "Playful Agentic Robot Learning," CoRL, 2026. https://arxiv.org/abs/2606.19419
[97] R. Lu et al., "ASPIRE: Agentic Skills Discovery for Robotics," arXiv, Jun. 2026. https://arxiv.org/abs/2607.00272
[98] W. Xiao et al., "ENPIRE: Agentic Robot Policy Self-Improvement in the Real World," arXiv, Jun. 2026. https://arxiv.org/abs/2606.19980
[99] M. Jia et al., "Agent as Policy for Robotic Manipulation," arXiv, Sep. 2026. https://arxiv.org/abs/2609.12541
[100] S. Ge et al., "Make Code as Policy Great Again: Frontier Agents Write, Call, and Evolve Robot Tools," arXiv, Sep. 2026. https://arxiv.org/abs/2609.39018
[101] Y. Chen et al., "Show-Harness: Just a VLM Agent Can Play Robots," arXiv, Sep. 2026. https://arxiv.org/abs/2609.10522
[102] B. Y. Tsui et al., "Demonstration-Free Robotic Control via LLM Agents," arXiv, Jan. 2026. https://arxiv.org/abs/2601.20334
[103] D. Chen et al., "A Few Words Go a Long Way: Language Guided Robot Policy Synthesis," arXiv, Jul. 2026. https://arxiv.org/abs/2607.23784
[104] X. Wang et al., "SEAR: Self-Evolving Agentic Robotics," arXiv, Oct. 2026. https://arxiv.org/abs/2610.18792
[105] J. Wang et al., "What Stops Recursive Self-Improvement in Robotics? Lessons from 123 Rounds of Agentic Skill Discovery," arXiv, Sep. 2026. https://arxiv.org/abs/2609.31760
[106] Y.-J. Wang et al., "Reconstruct, Practice, Go Real: Guided Self-Improvement for Embodied Agents," arXiv, Oct. 2026. https://arxiv.org/abs/2610.02204
[107] G. Jena et al., "Weights or Skills? A Survey of Robot-Learning Techniques," arXiv, Aug. 2026. https://arxiv.org/abs/2608.01851