ETH zürich

SoftRobotics Laboratory

VLA models

Davide Liconti
Prof. Robert Katzschmann
RWR, 24-11-2025

Imitation Learning (IL)

why does it dominate today

Today **IL** is by far the dominant approach for manipulation

- **No** complex or arbitrary **reward shaping**
- **No sim2real gap** (for real world teleop data)
- Simpler, easier to debug and interpret
- Can theoretically **scale** with data and compute

LLMs (e.g., GPT) are trained in a similar form as IL.
They are trained to predict the next token given a ”context” of recent tokens.

Why is learning actions different than learning next token?

What are LLMs?

What are LLMs?

LLM works with **TOKENS** → **DISCRETE** numbers encoding words (or parts of words)

Language Learning Models (LLMs) have revolutionized the field of natural language processing, enabling machines to understand and generate human-like text. At the core of LLMs lies the concept of tokens, which serve as the fundamental building blocks for processing and representing text data. In this blog post, we'll demystify tokens in LLMs, unraveling their significance and exploring how they contribute to the power and flexibility of these remarkable models.

↓

[145, 232, 12, 111, 1563, 66673, 42, 1358, 9534, 5123,…]

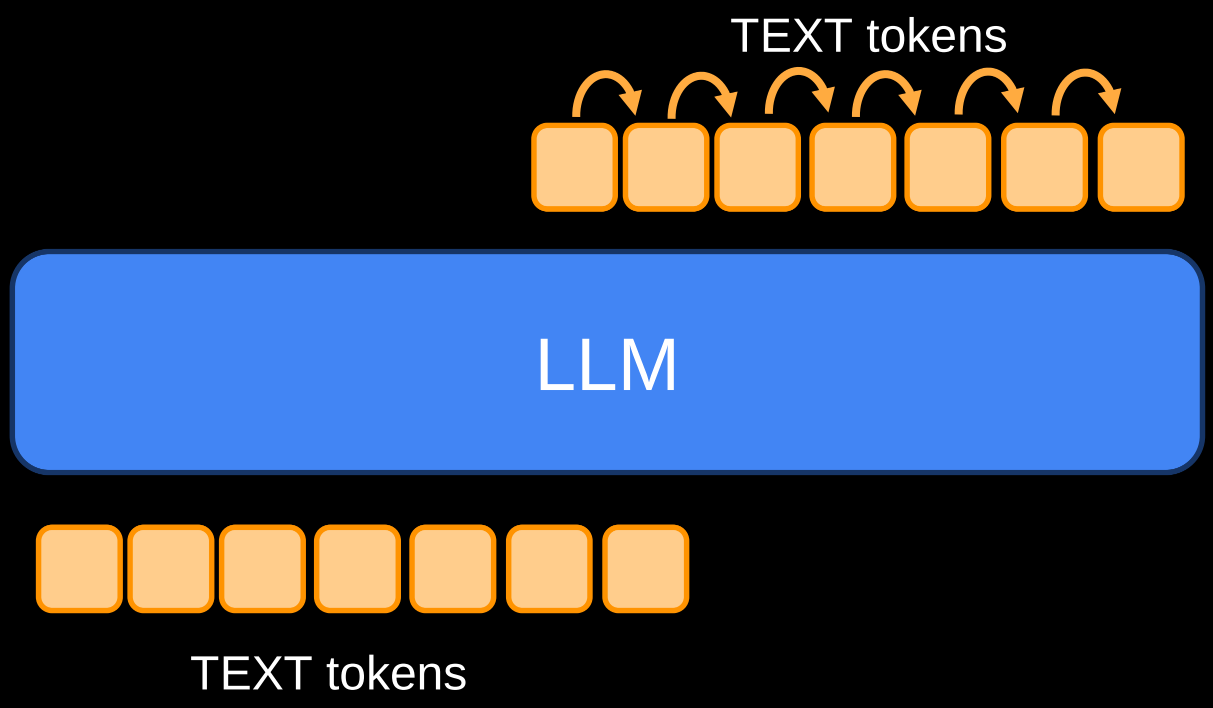
What are LLMs?

LLM works with tokens $\rightarrow$ discrete numbers encoding words (or parts of words)

Training loss: Cross-Entropy on shifted token sequence

↓

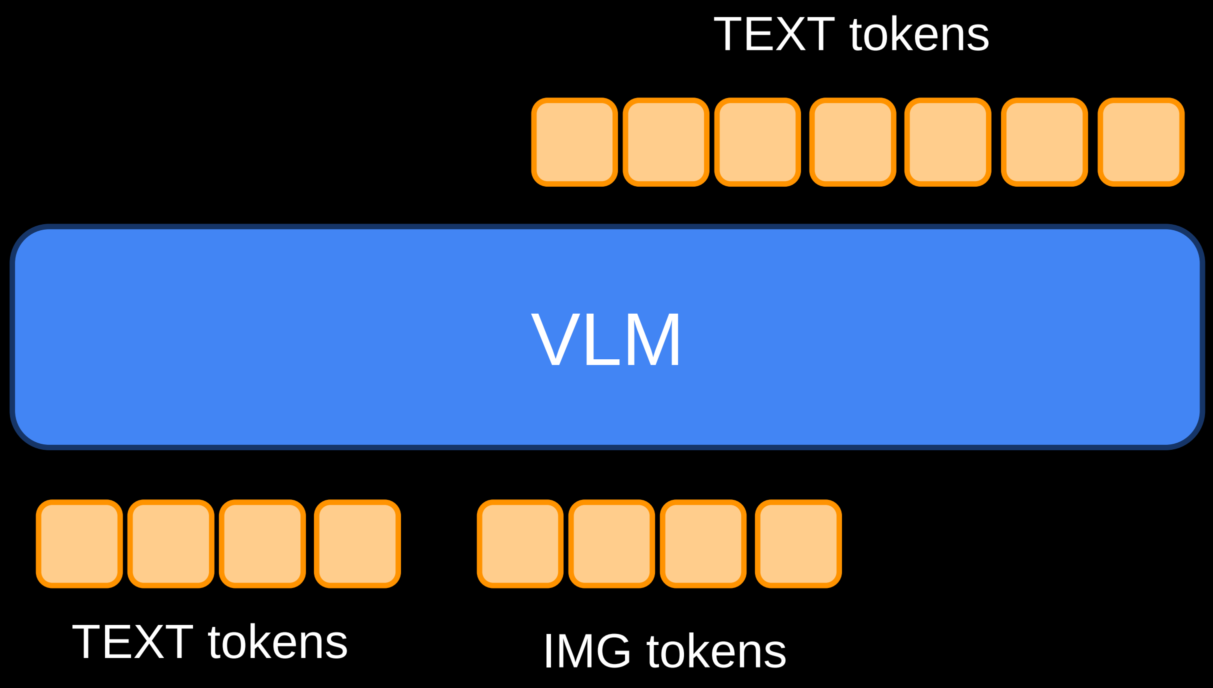
Autoregressive inference: predict one token after another



What are VLMs?

Extend LLM with Images

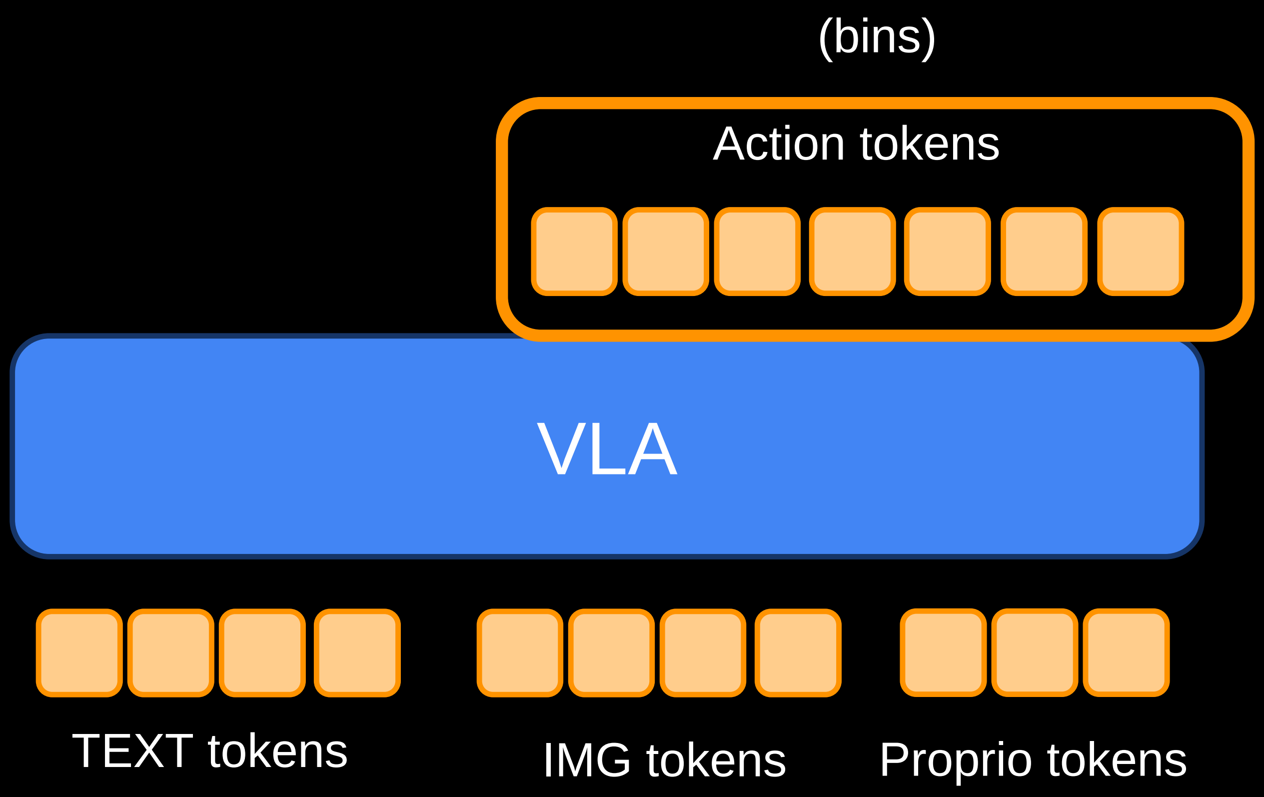
How to get Image Tokens? → Extract features (ViT backbone) and discretize



What are VLAs?

Extend VLMs with Actions

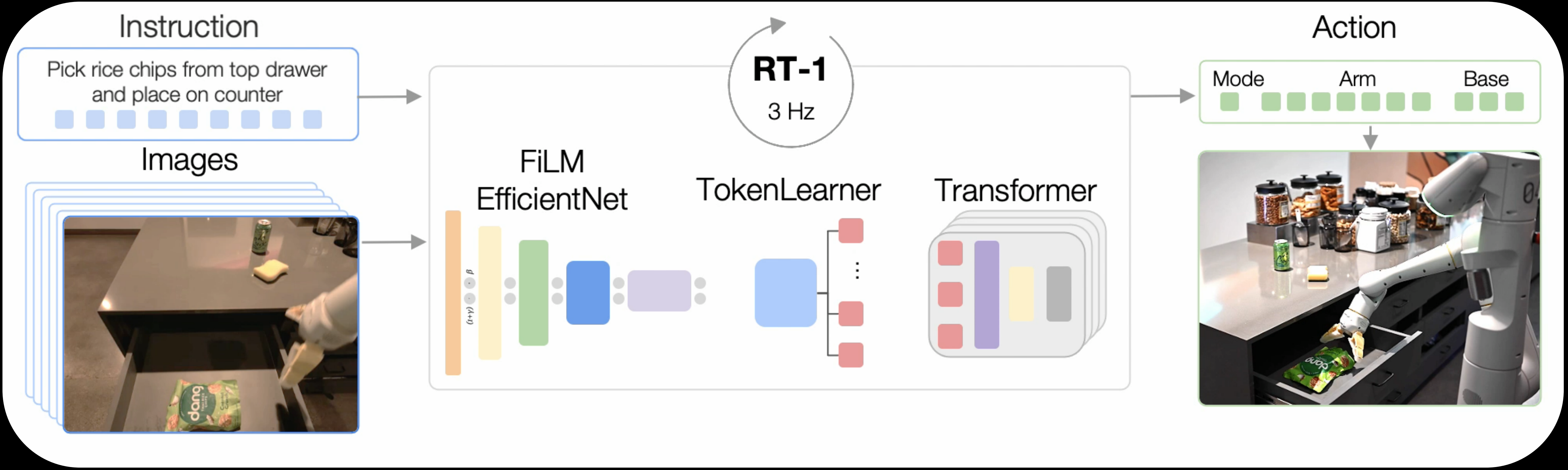
How to tokenize the Actions?
Can we afford slow autoregressive inference?



RT-1: Robotics Transformer for Real-World Control at Scale

Google Research first effort into Foundation Models (task-agnostic models) for robotics (2022)

- From single-task models to **multi-task** models
- Actions are discretized into 256 **bins** for each dimension
- RGB + Language inputs

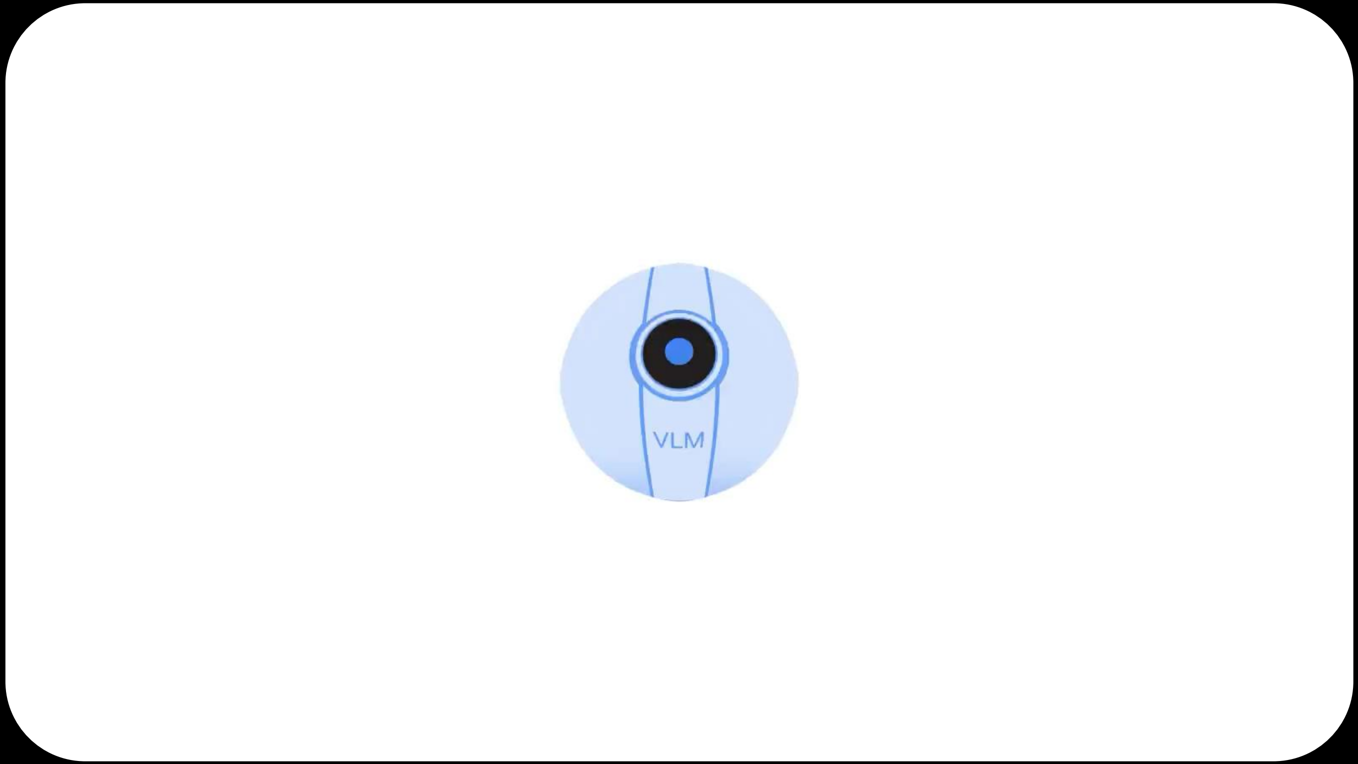


Brohan et Al, RT-1: Robotics Transformer for Real-World Control at Scale, 2022


20x speed
RT-1 Controlling the robot
Instruction: Bring me all the graspable objects from the counter.

RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control

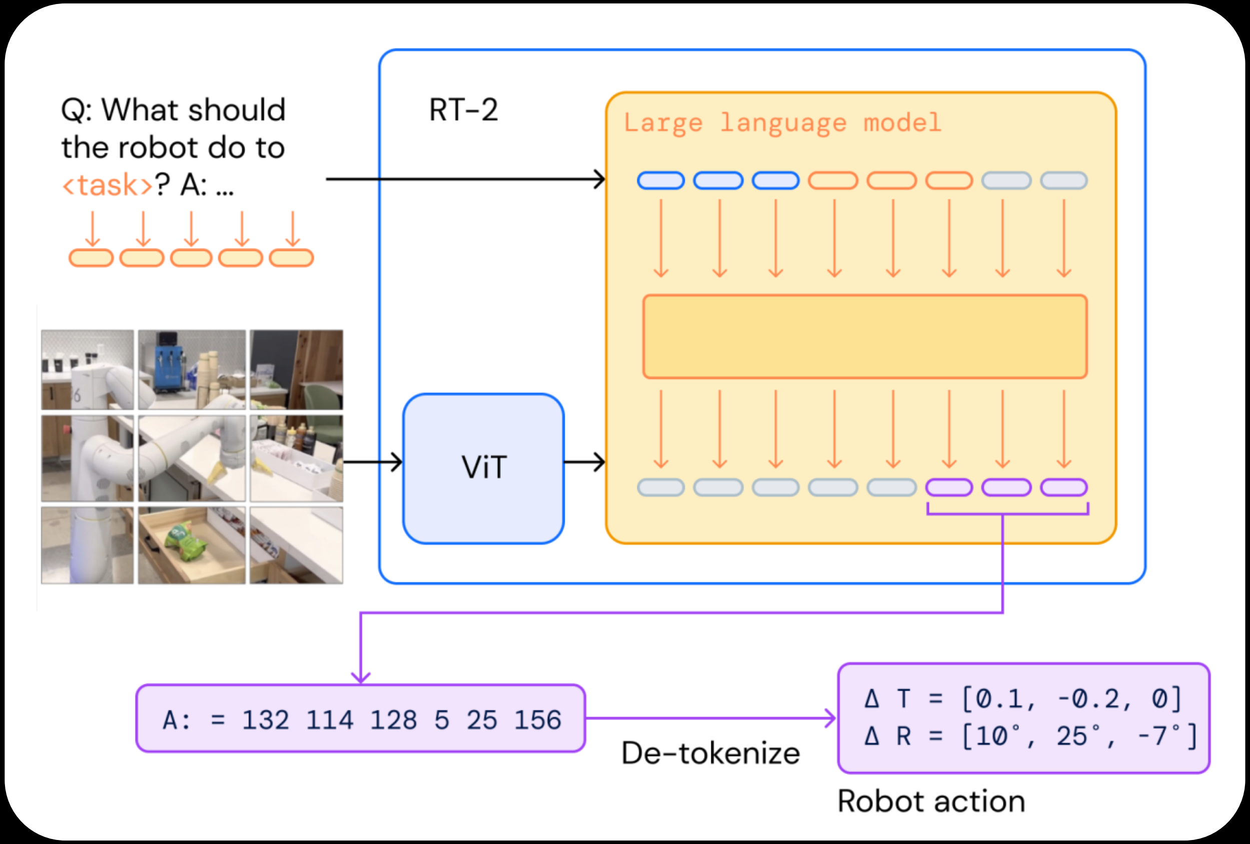
- Same concept of RT-1 but with a **VLM pretrained backbone** -> VLA model

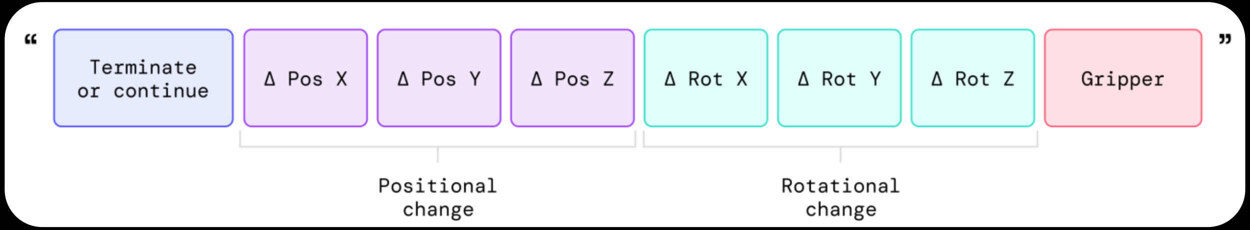


Brohan et Al, RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control, 2023

RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control

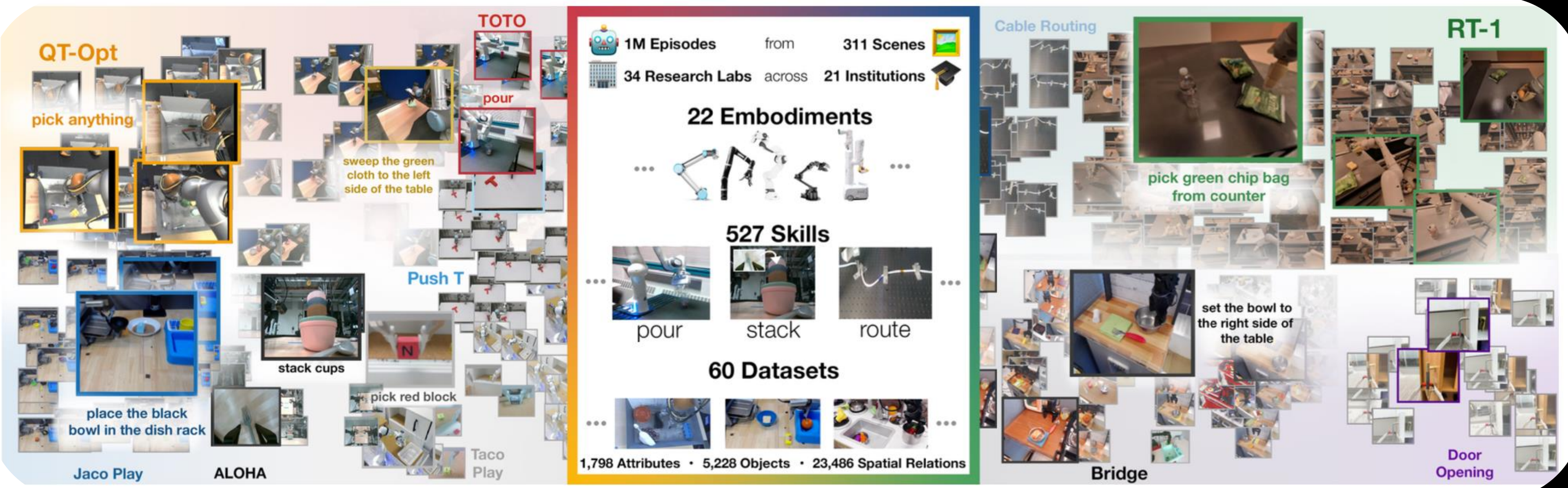
- **Co-fine-tuning** on both web-scale vision-language data *and* robot trajectory data
- Significantly improved **generalization** and emergent semantic reasoning for robotics: e.g., handling novel objects or instructions not seen in the robot training data
- First work that showed that large vision-language backbones (e.g., up to **55 B** parameters) can be adapted for robotic control

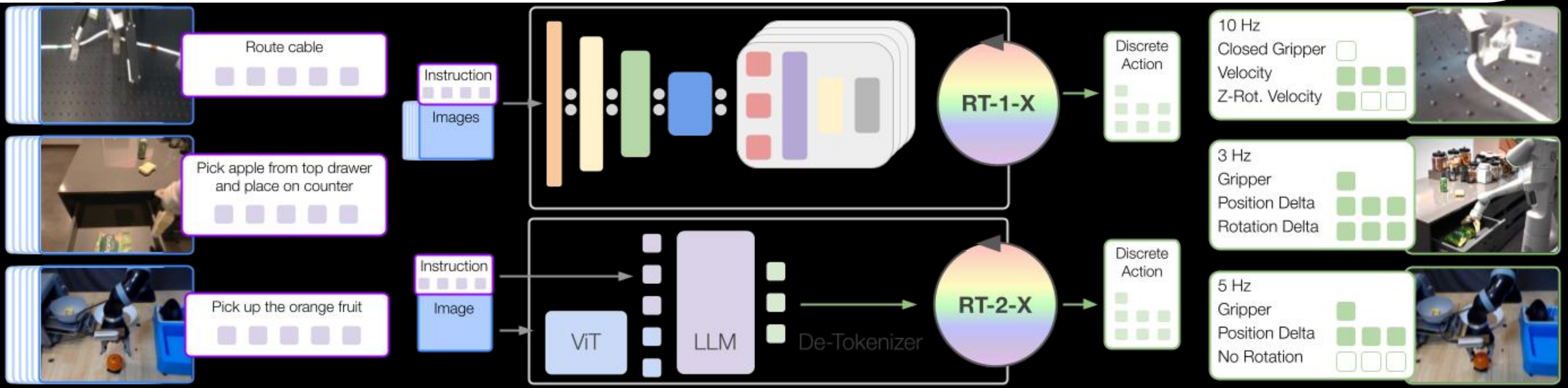




Open X-Embodiment: Robotic Learning Datasets and RT-X Models

- **Open X-Embodiment Dataset:** 1M+ trajectories from 22 embodiments
- RT-X Generalist Models: Transformers (RT-1-X / RT-2-X) trained jointly on multi-embodiment data
- Shift to Foundation Robotics: Demonstrates that **data diversity > data quantity** for generalization across tasks and embodiments.



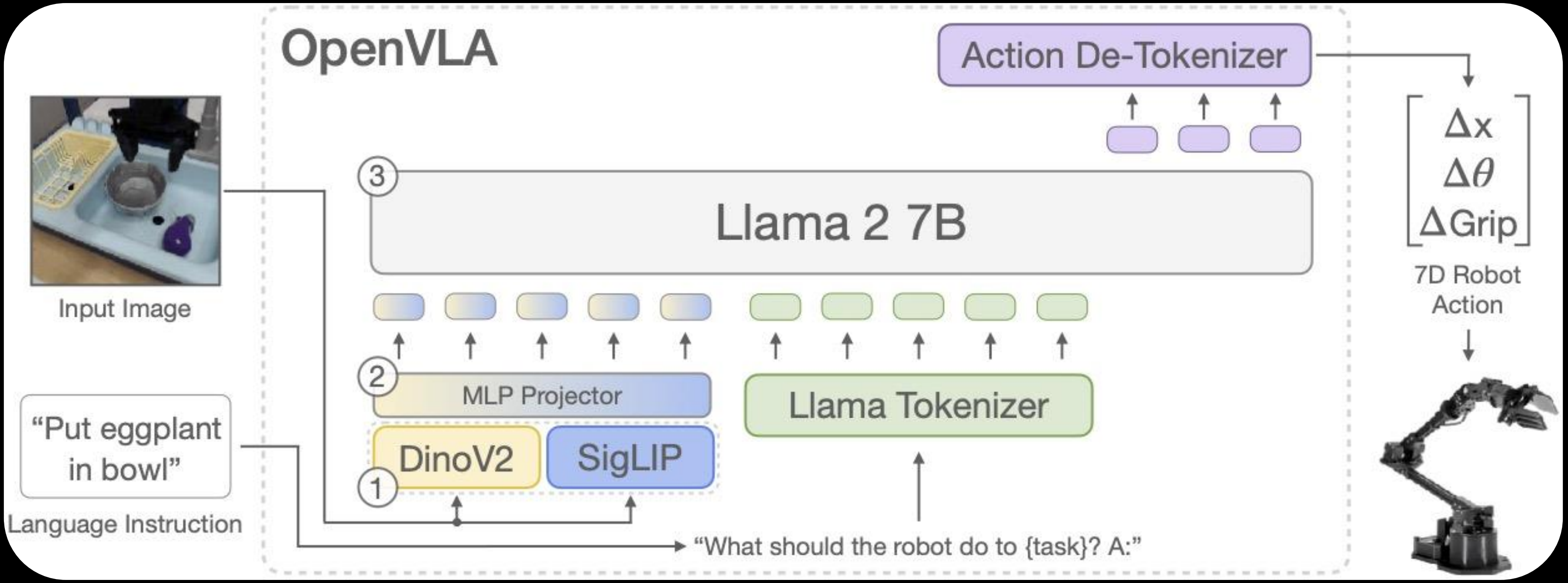


Open X-Embodiment: Robotic Learning Datasets and RT-X Models, 2024

ETH zürich   SoftRobotics Laboratory

OpenVLA: An Open-Source Vision-Language-Action Model

- Trains a **7B-parameter** vision-language-action (VLA) model on ~970 k robot manipulation episodes from the Open X-Embodiment Dataset
- Uses a **fused vision encoder** (combining features from DINOv2 + SigLIP) feeding into a large language model backbone (LLaMA 2 7B) to directly output robot action tokens
- Demonstrates **efficient fine-tuning** (LoRA + quantization) to adapt to new robot setups with less data
- Open-Data (open-x) and Open-Weights (model and code available)



Kim et Al. OpenVLA: An Open-Source Vision-Language-Action Model, 2024

OpenVLA

- Trains a **7B-parameter** vision-language-action (VLA) model on ~970 k robot manipulation episodes from the Open X-Embodiment Dataset
- Uses a **fused vision encoder** (combining features from DINOv2 + SigLIP) feeding into a large language model backbone (LLaMA 2 7B) to directly output robot action tokens
- Demonstrates **efficient fine-tuning** (LoRA + quantization) to adapt to new robot setups with less data
- Open-Data (open-x) and Open-Weights (model and code available)

"Pick up the coke can and place on top of Taylor Swift"

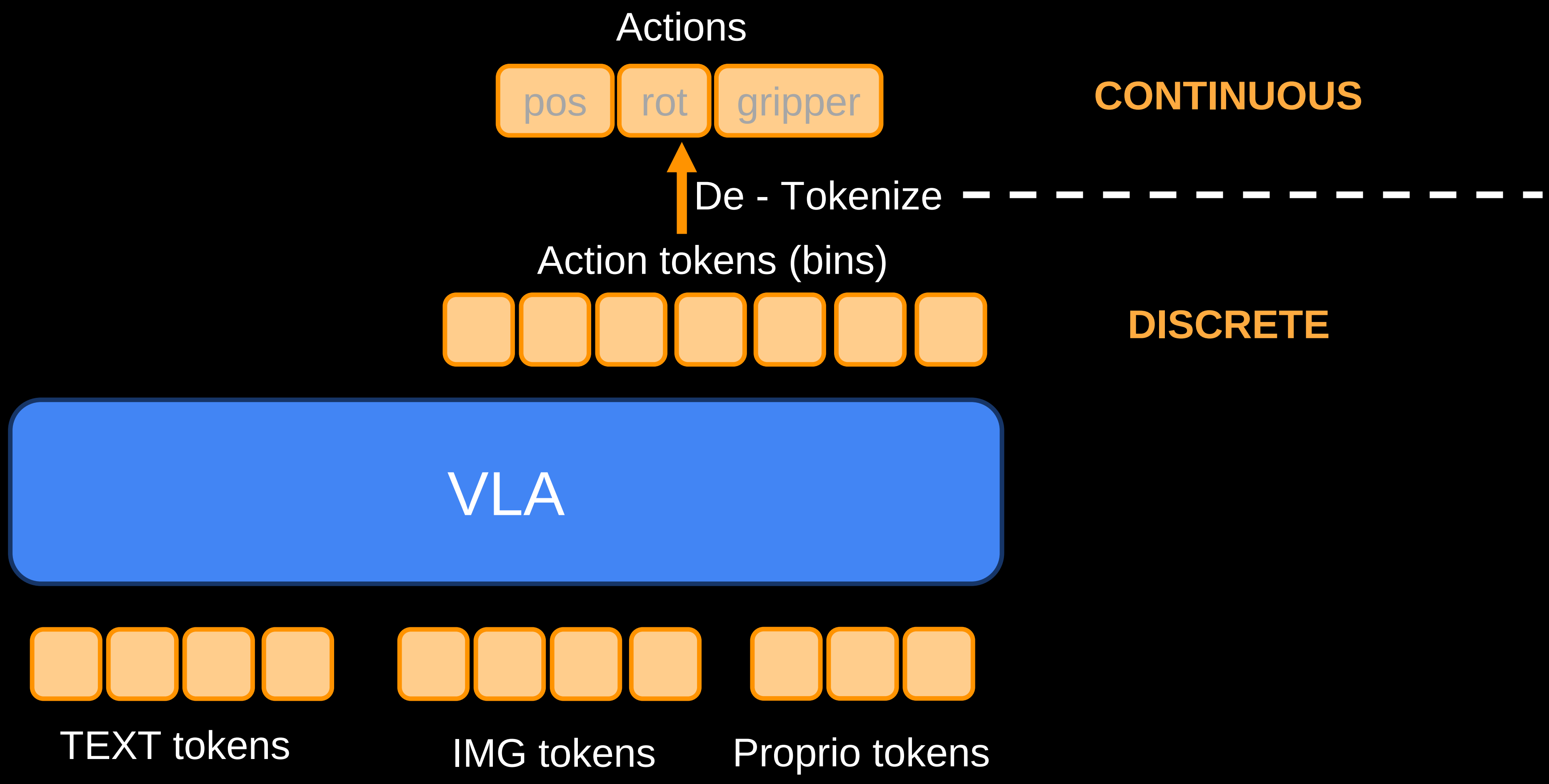
RT-2-X

OpenVLA

Kim et Al. OpenVLA: An Open-Source Vision-Language-Action Model, 2024

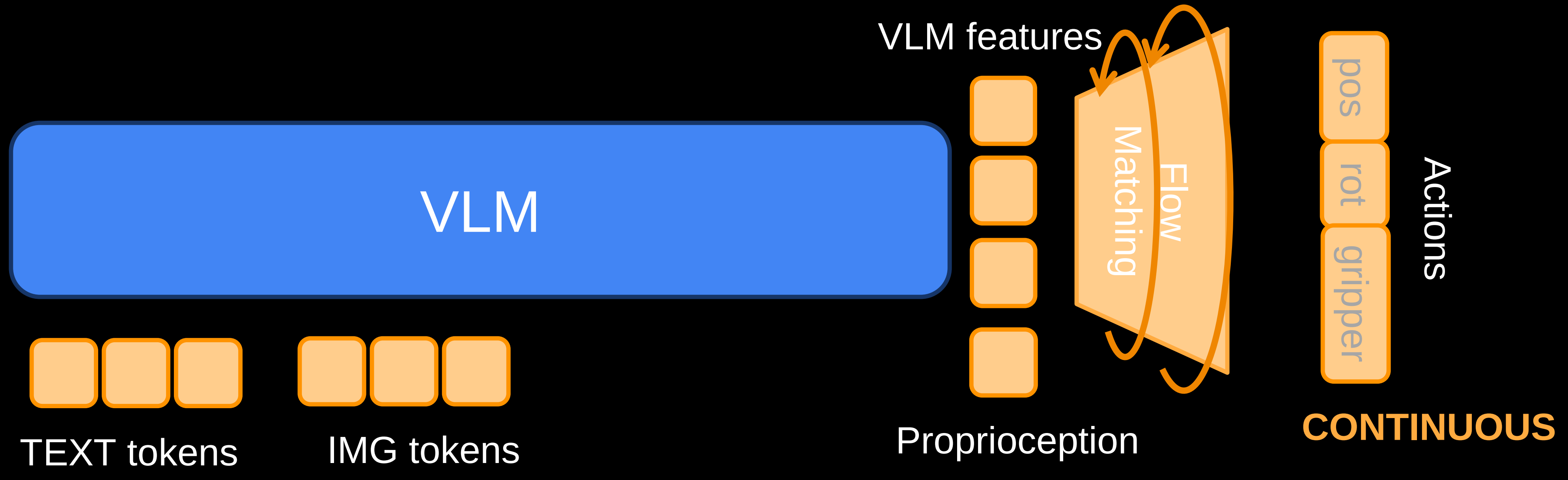
Action Representation: Discrete Bins

- Binning fails for highly dexterous tasks, as we are **losing action resolution**



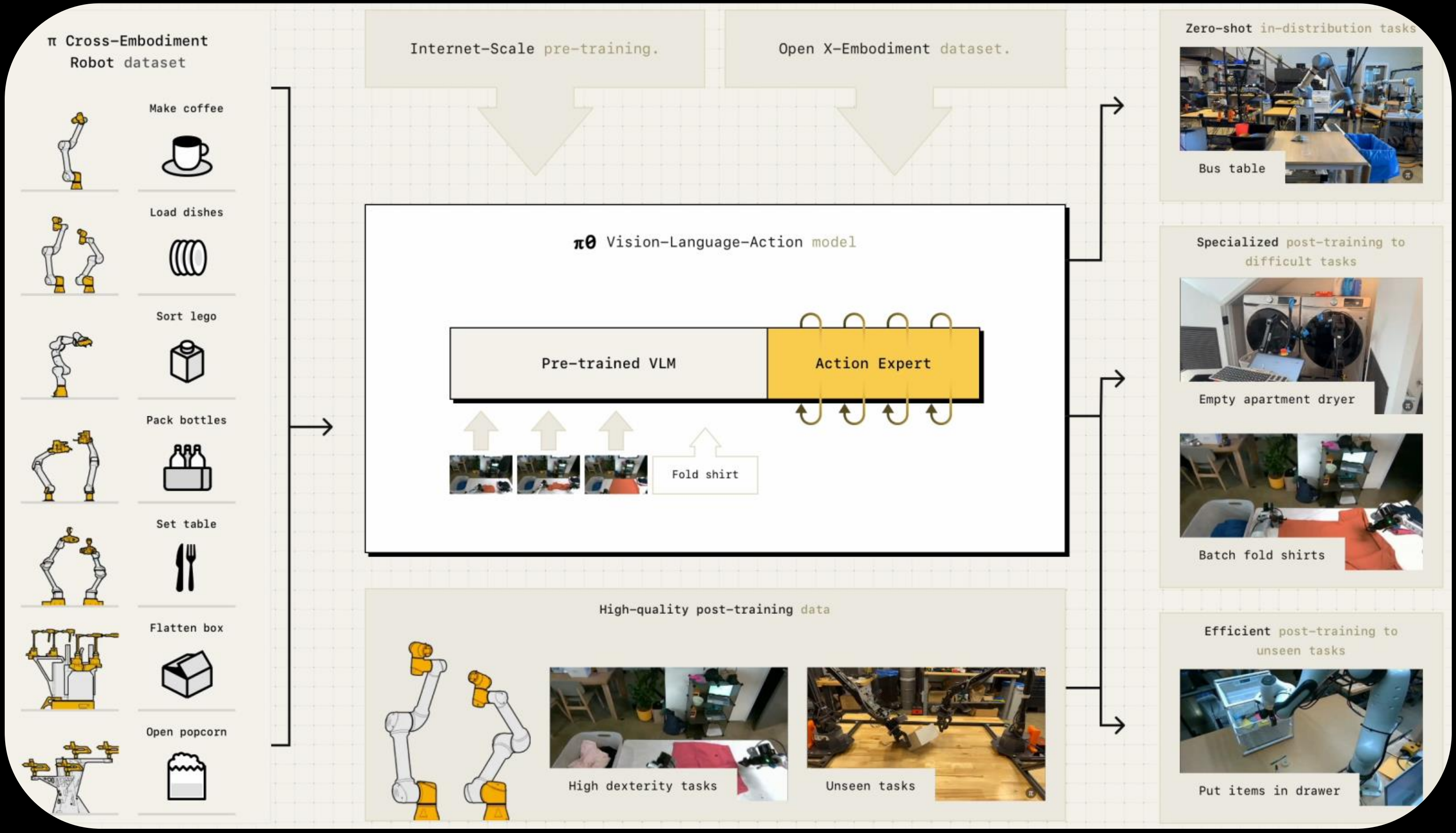
VLA → VLM + Action Head

- Binning fails for highly dexterous tasks, as we are **losing action resolution**
- Can use the VLM as very general and powerful backbone and use **diffusion or flow matching** based action head to predict continuous actions

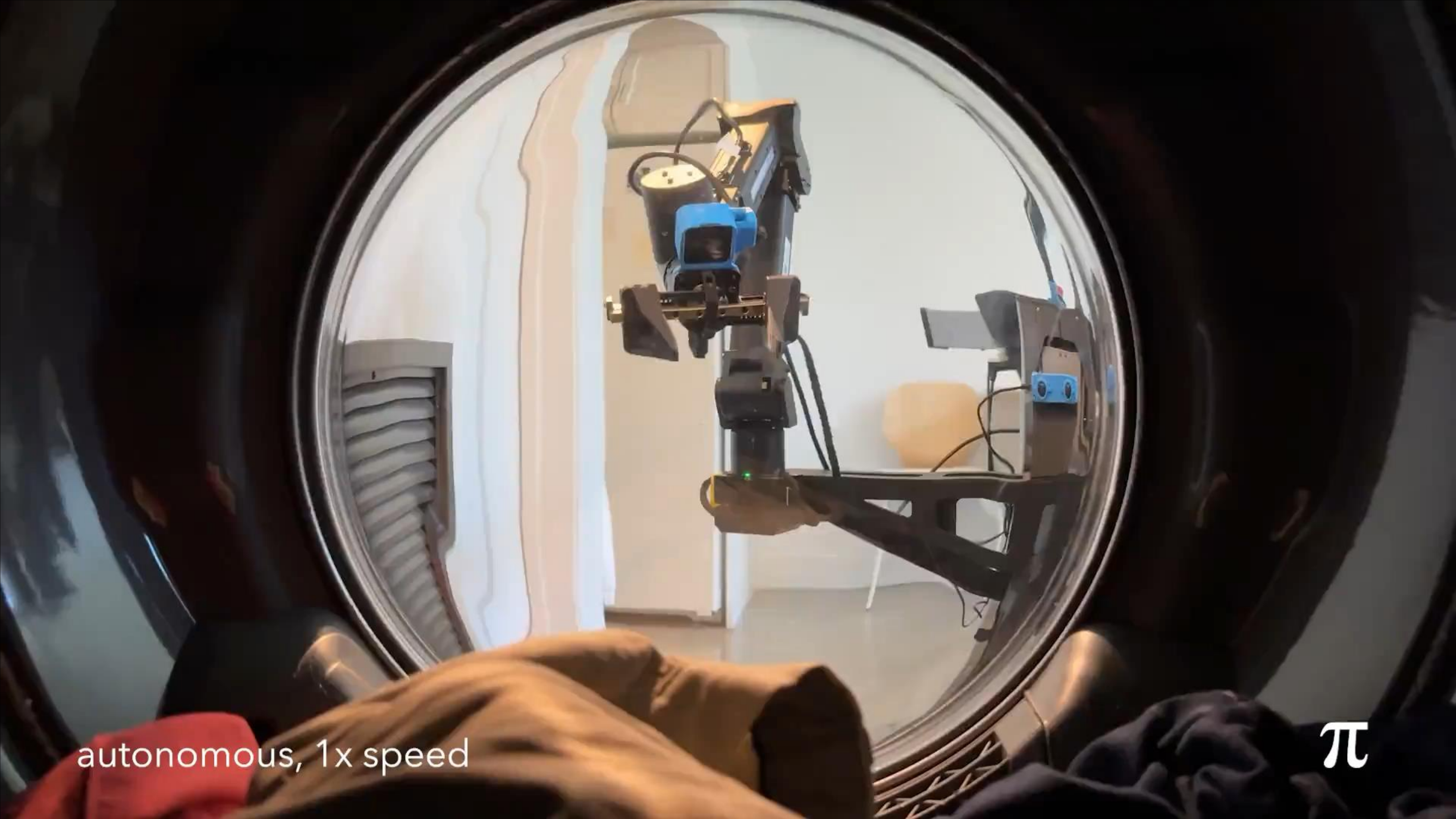


π0

- Trained across **multiple robots and tasks** **(8 robots in-house + OpenX data)**
- VLM pre-trained on web-scale image+text data (Paligemma), and then augmenting it with **continuous action output** capability (via flow-matching) so it can output motor commands at up to ~50 Hz

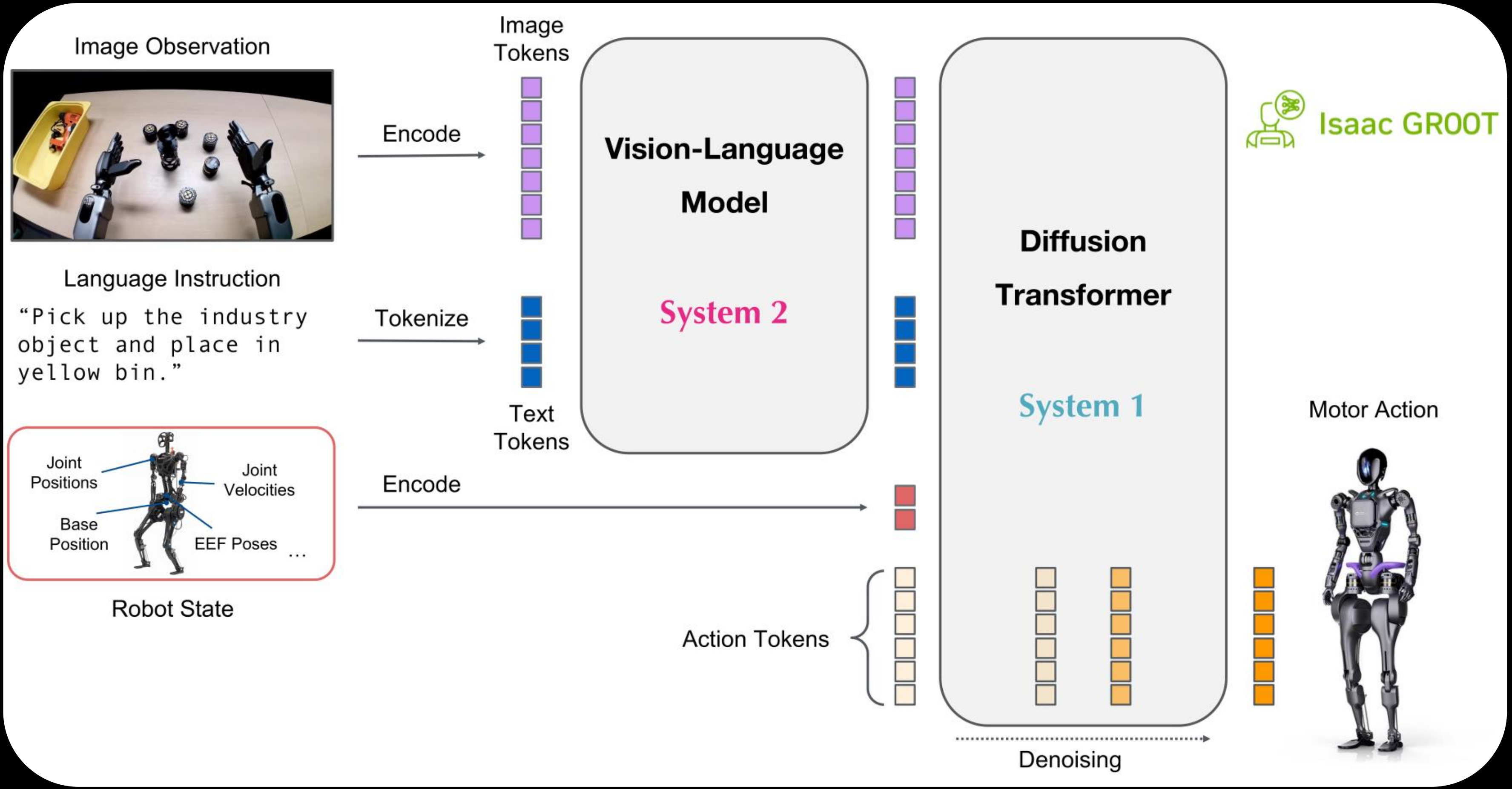


physicalintelligence.company


autonomous, 1x speed

GR00T N1 - NVIDIA

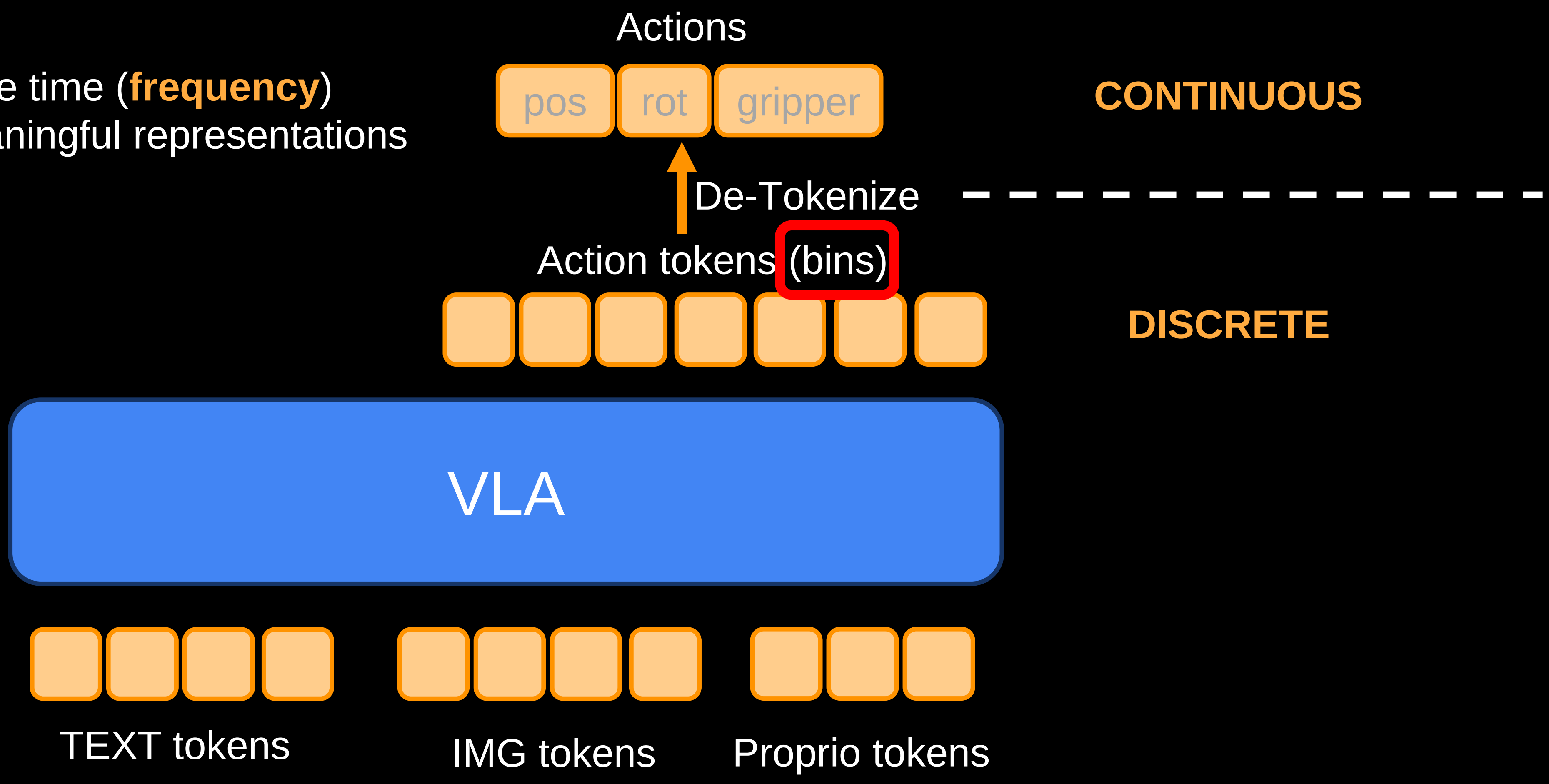
- GR00T N1: a general-purpose VLA model for humanoids (vision + language → motor actions)
- Dual-system architecture: **reasoning (System 2 - VLM) + action generation (System 1- Diffusion)**
- Heterogeneous training data pyramid: web videos → synthetic trajectories → real-robot data



physicalintelligence.company

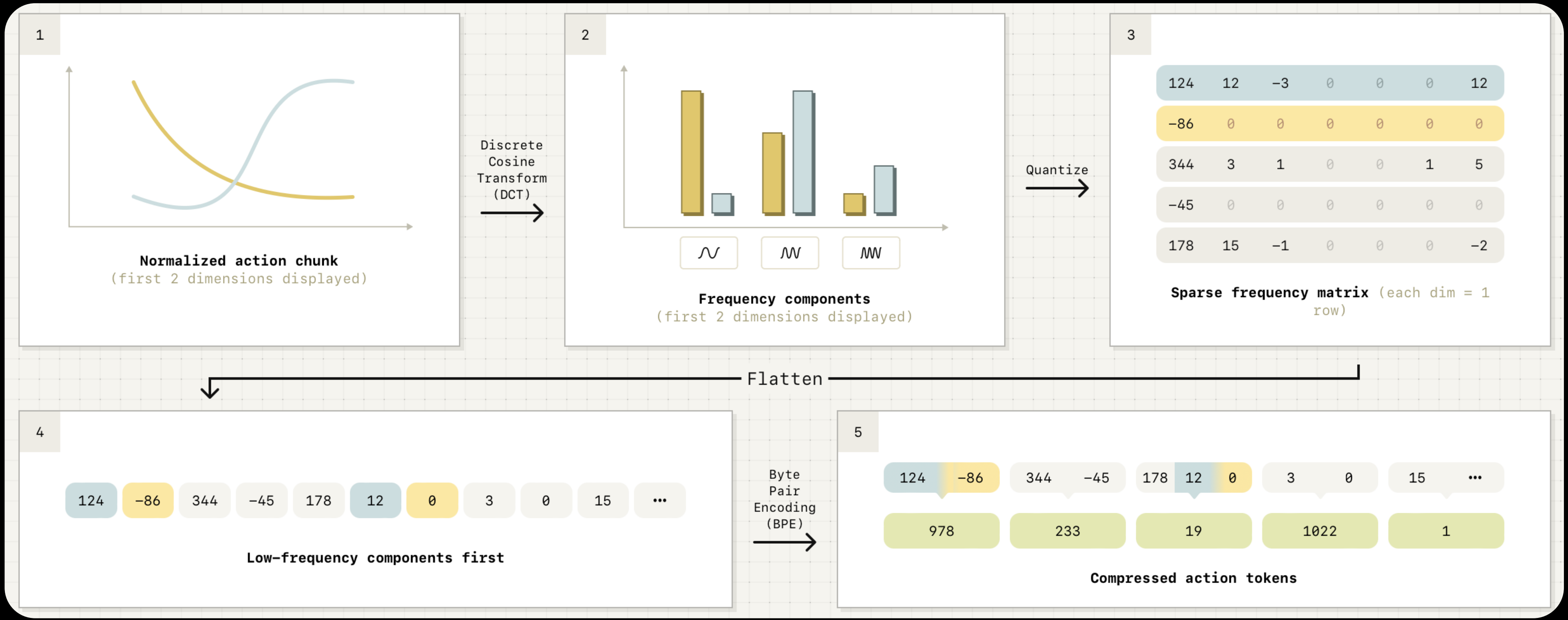
Action Representation

- Is there a better way to encode robot actions into discrete tokens?
- We can reason about the time (**frequency**) domain to get more meaningful representations



π0 - FAST

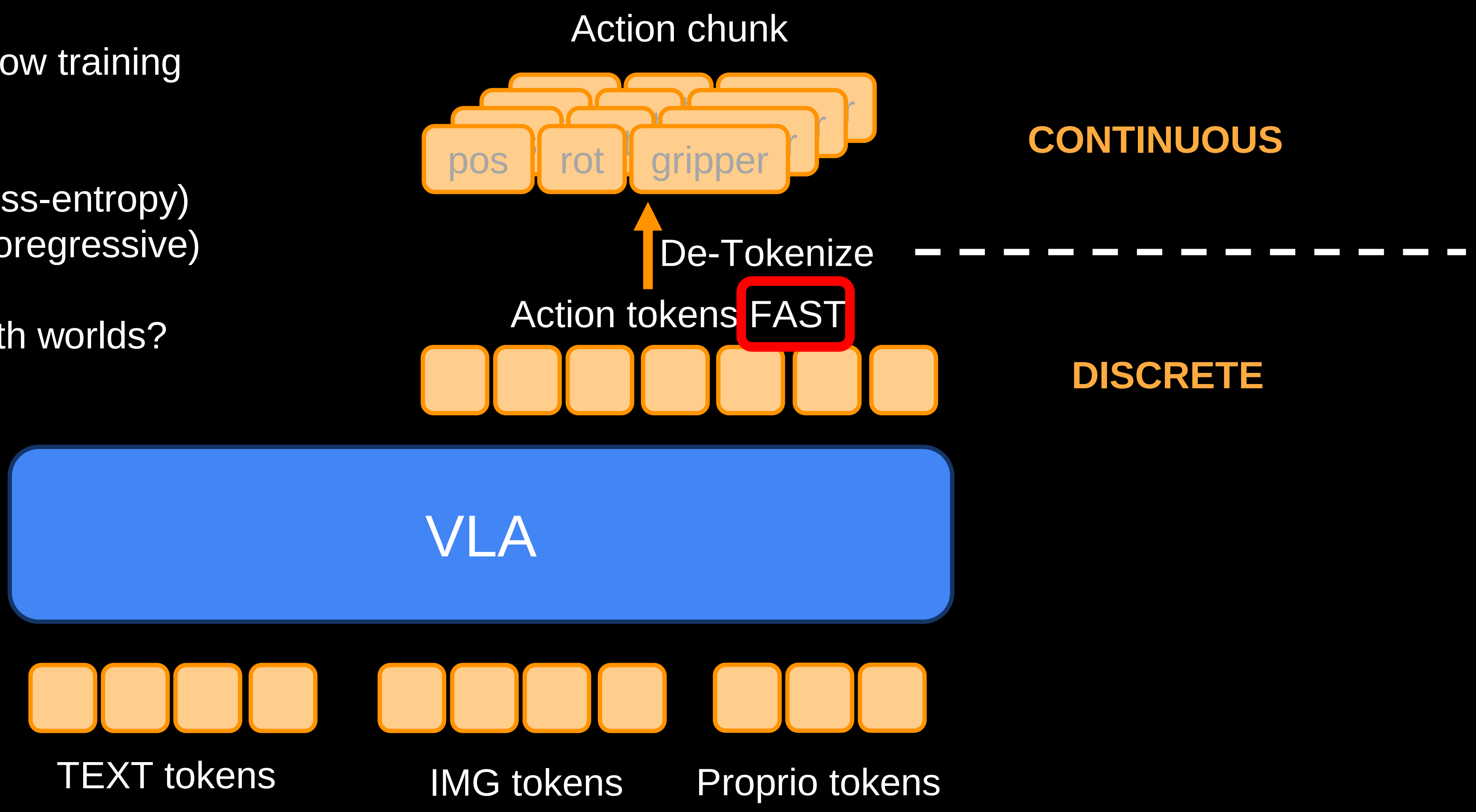
- FAST: transforms continuous robot action chunks into dense discrete tokens via **DCT + BPE**
- Significantly accelerates training (≈5× faster) of generalist VLA policies (cross-entropy loss)
- **Universal tokenizer** trained on 1 M real robot trajectories → supports transfer across embodiments and control frequencies.
- **Autoregressive inference of VLA** policies built with FAST match diffusion-based models on complex tasks while being simpler to train and deploy.



Action Representation : FAST

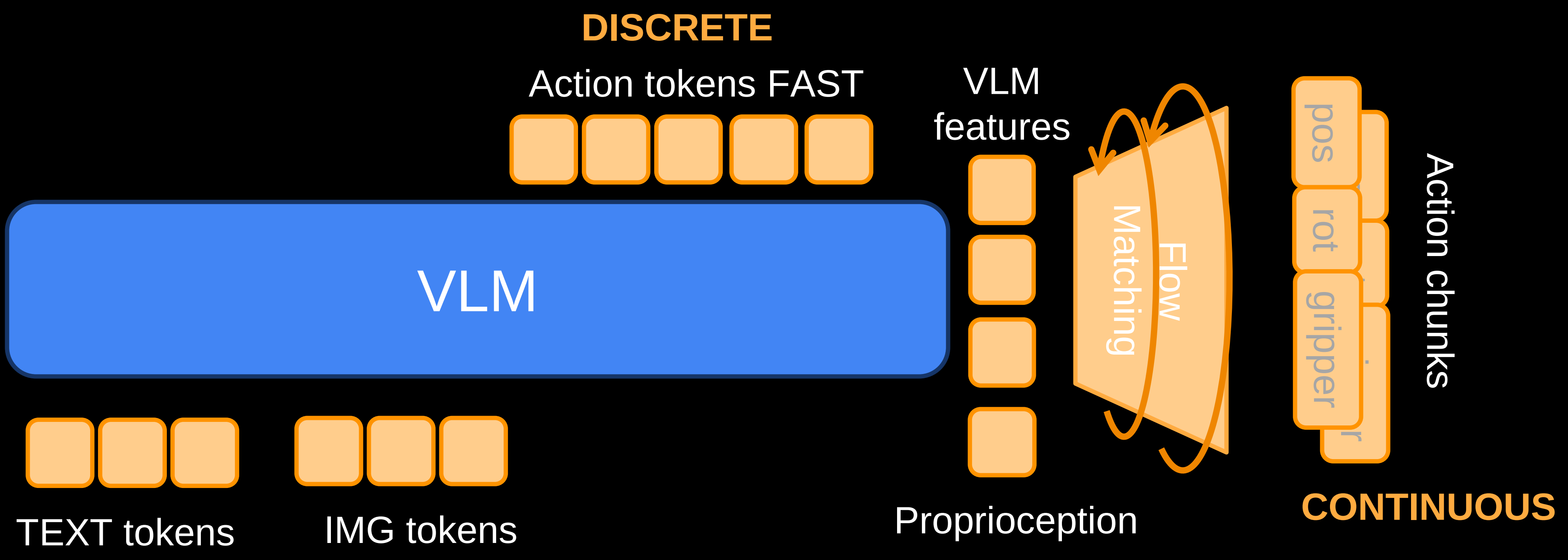
- Flow matching head: slow training but fast inference
- FAST: fast training (cross-entropy) but slow inference (autoregressive)

Can we get the best of both worlds?



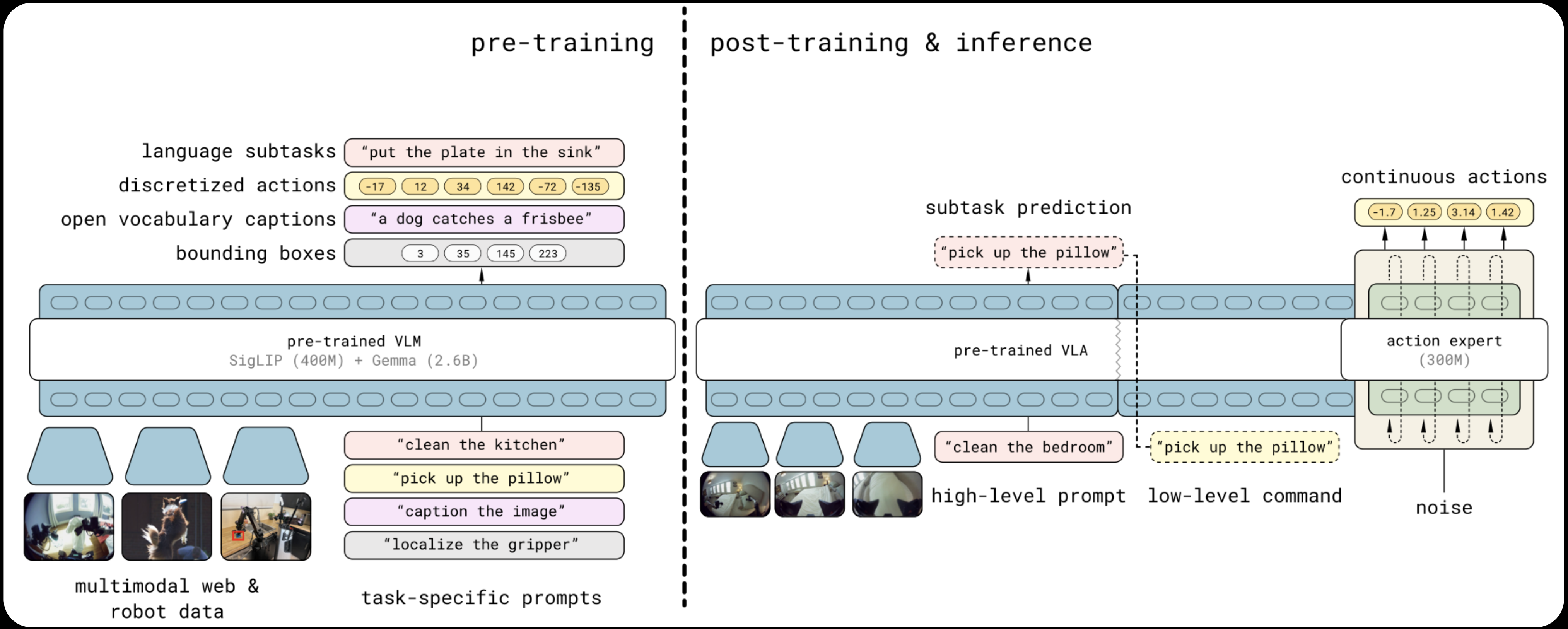
Action Representation : FAST + Action Head

- Use **both discrete and continuous** action representations



π0.5

- **Discrete FAST** tokens for training + **continuous action** expert for fast inference
- Trained on web-vision-language + multi-robot + mobile manipulation datasets
- Sub-task decomposition via high-level/low-level prompts. Step toward open-world generalist robotics



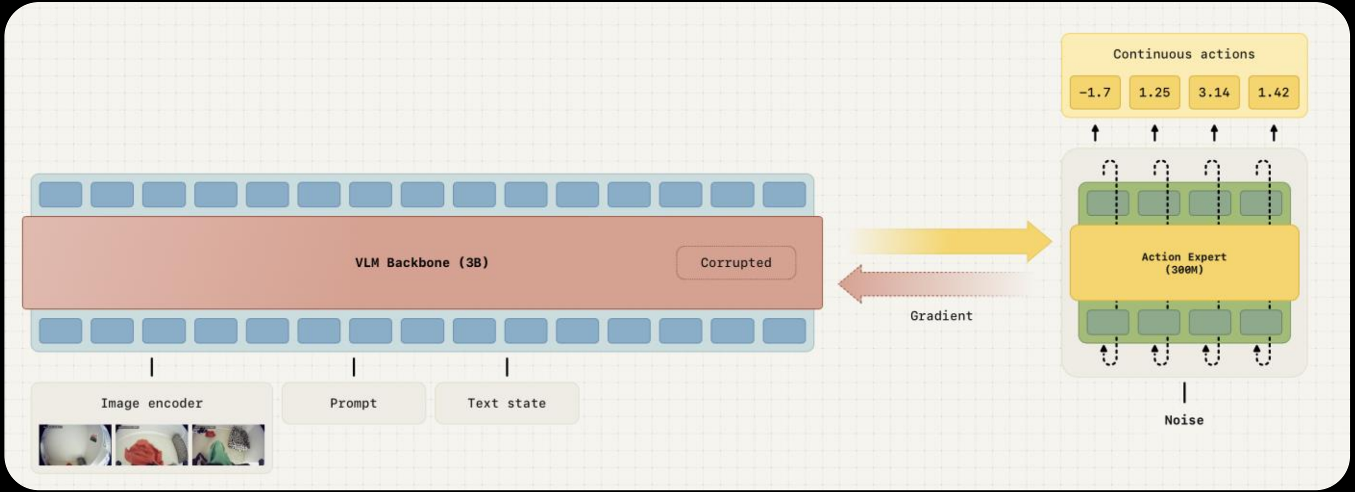
physicalintelligence.company


Douglass
30

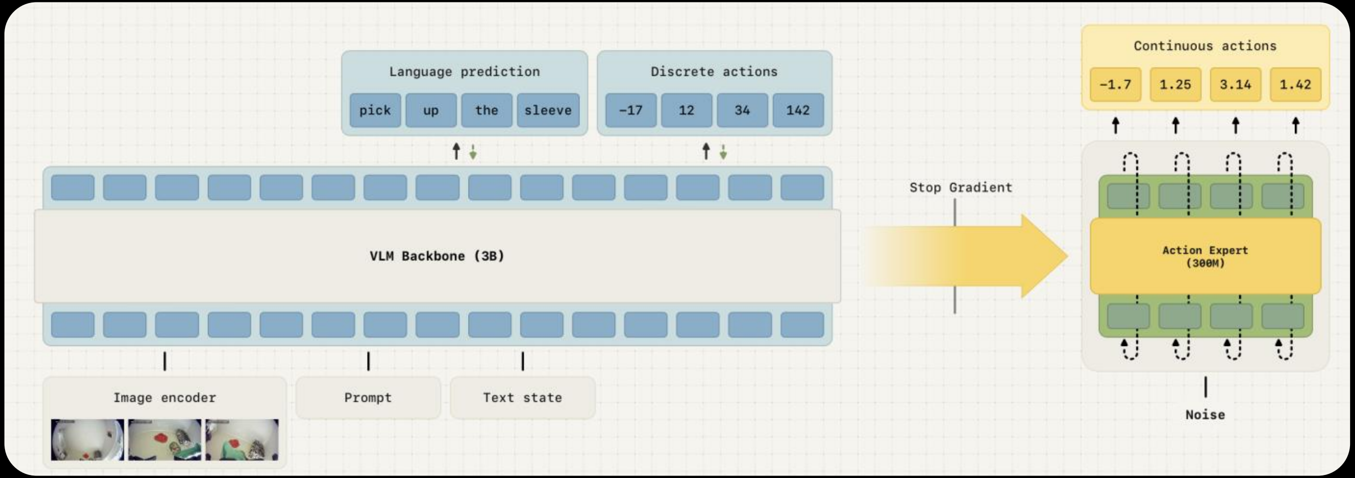
π0.5 – Knowledge insulation

Knowledge Insulation: decouple VLM backbone from action expert gradients
→ retain semantic knowledge + faster training
Significant gains: up to ~7.5× **faster training**, **strong generalisation**, and **real-time** continuous control

$\pi 0$



$\pi 0.5 - KI$



physicalintelligence.company

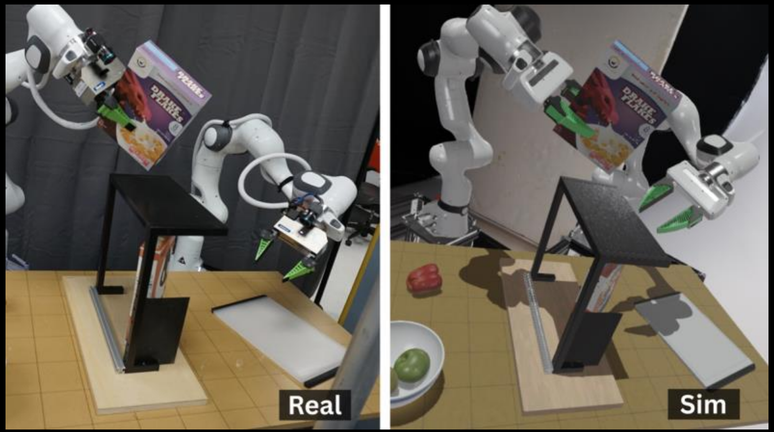
ETH zürich SoftRobotics Laboratory

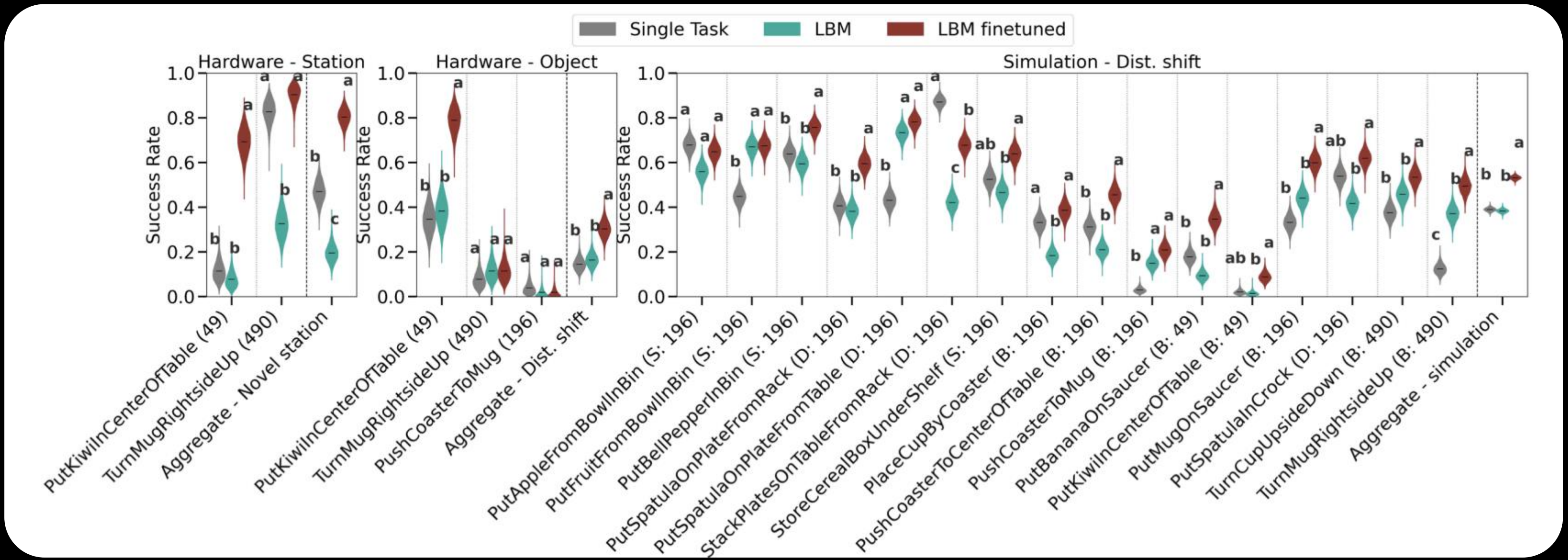
Gemini Robotics

Multi-embodiment VLA model (Gemini Robotics 1.5) with **Motion Transfer** → unified across robots
Unified **agentic framework** enables generalist robot behaviours: perception → reasoning → motion

Large Behavior Models (LBM) - TRI

- Multi-task visuomotor policies (diffusion policy) trained on ~1,700 h of data across ~500 tasks
- **Rigorous evaluation**: simulation + real-world trials (~1,800), **blind A/B testing** for statistical confidence
- Key results: fewer fine-tuning samples needed + higher performance + better robustness under shift
- **Scale matters**: larger and more diverse pre-training datasets → better manipulation generalisation





Very rigorous and well written paper, recommended!

Why making ChatGPT for robotics is not straightforward

Data

Cannot simply scrape internet
→ Human videos
→ Synthetic Data
→ Simulation

Control Frequency

Need real time control
→ System 1/ System 2
→ Chunk Quantization (FAST)

Cross-embodiment

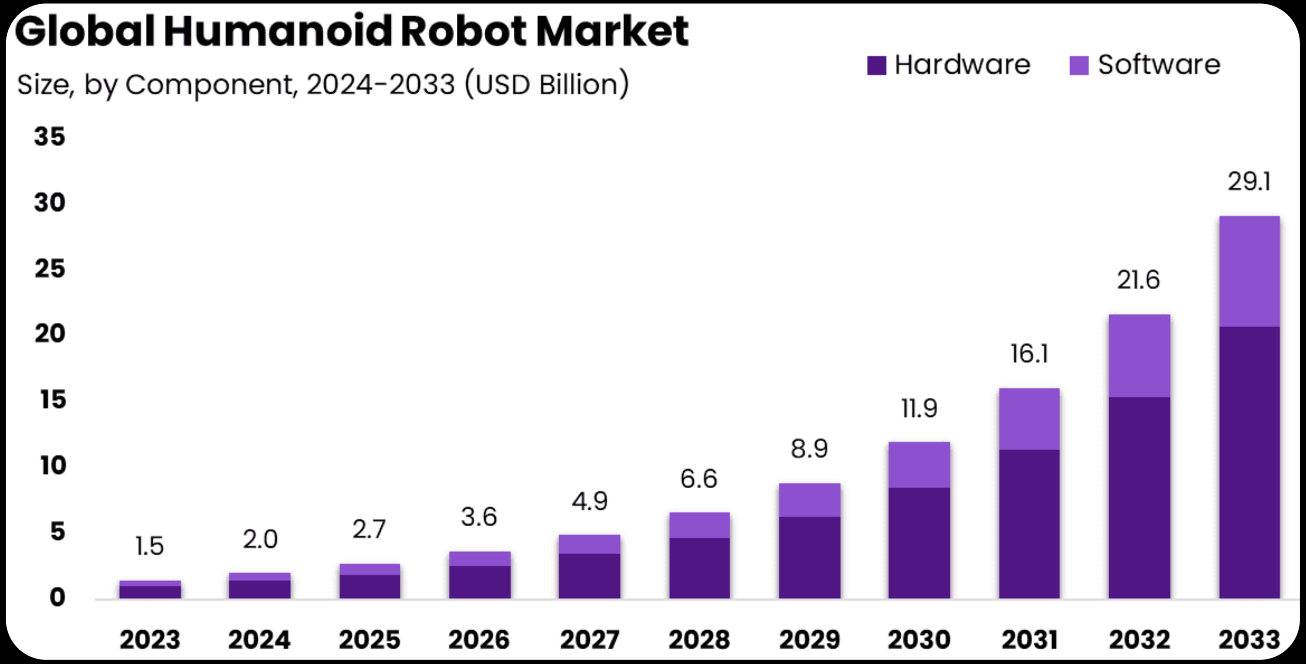
Observation and Action Space depends on the Robot Embodiment
→ Scaling Data
→ Latent Actions

ETH zürich SoftRobotics Laboratory

State of VLA research

Strong Industry interest (Google, NVIDIA, Tesla, Figure, 1X, Generalist, Physical Intelligence, …)

Hot topics now: new tokenization techniques, video generation (i.e., world models), evaluation, cross-embodiment, learning from human videos, RL finetuning



Useful Resources

- Robotics Transformers series:
  - https://robotics-transformer1.github.io
  - https://robotics-transformer2.github.io
  - https://robotics-transformer-x.github.io
- Pi0: https://www.physicalintelligence.company/download/pi0.pdf
- FAST tokenizer: https://arxiv.org/pdf/2501.09747
- Pi0.5 : https://www.physicalintelligence.company/download/pi05.pdf
- Gr00t Model: https://arxiv.org/pdf/2503.14734.
- Large Behavior Models (TRI) https://arxiv.org/pdf/2507.05331
  https://www.youtube.com/watch?v=TN1M6vg4CsQ&t=3936s
- Gemini Robotics 1.5 https://arxiv.org/pdf/2510.03342
- Knowledge Insulation: https://www.physicalintelligence.company/download/pi05_KI.pdf