Vision Language to Actions: A practical guide

Danilo Pau

Technical Director, System Research

IEEE, AAIA, ST Fellow

Full professor ASN09/E3 Italy

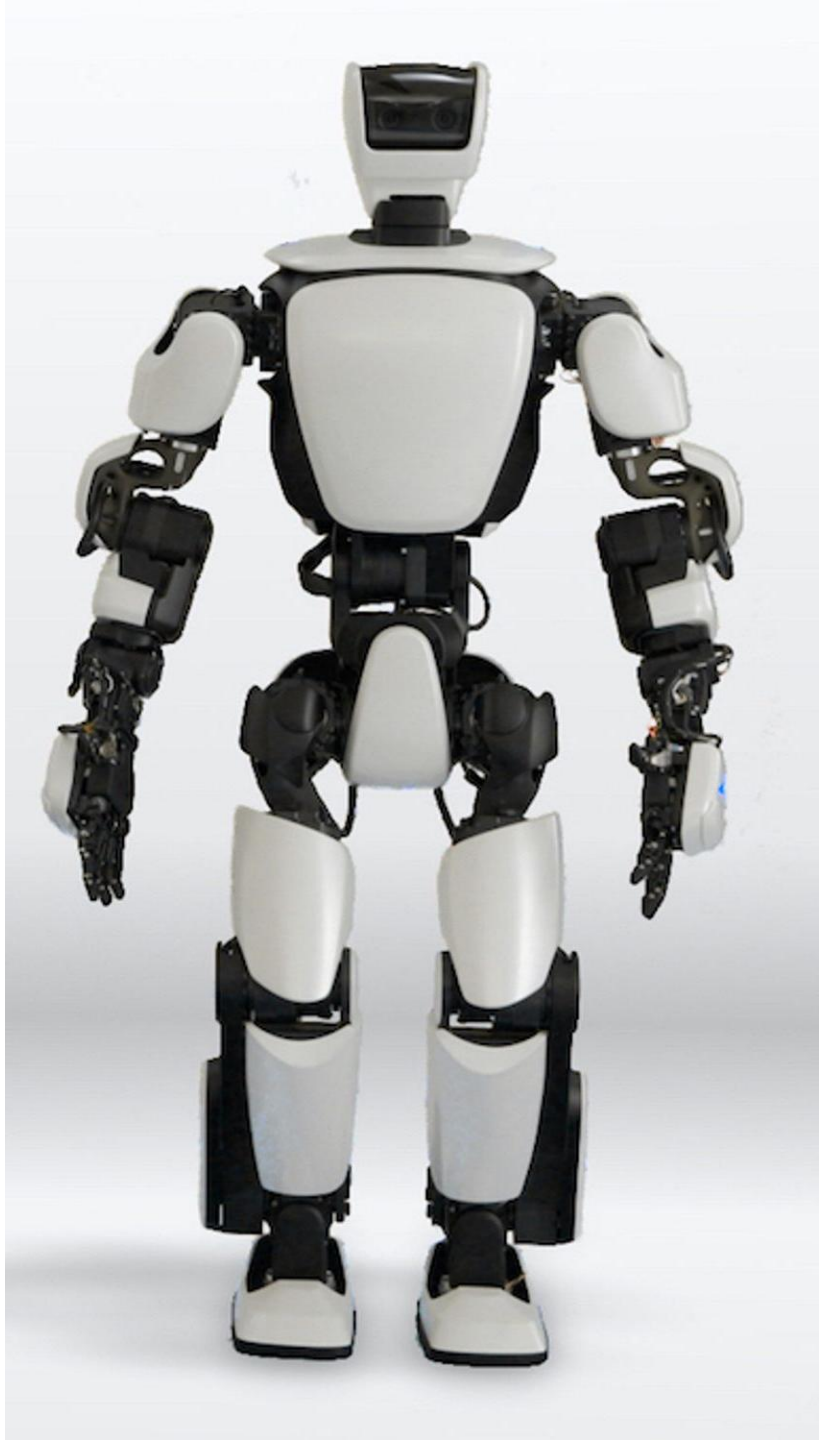
IEEE Distinguished Industry Lecturer

Scientific Committee 2026-2028 IEEE SPS Italy

Global Governance and AI Safety NAAI Committee

Editorial Board Member MDPI Electronics

APSIPA Life Member



Is TOYOTA T-RH3 humanoid robot ?

The teleoperation paradigm

- The approach
  - Bypass the “**autonomy challenge**” by keeping the human in the loop
  - The robot is **the physical avatar** deployed for safe, precise interaction in complex, unstructured environments (medical, emergency etc)

- **Master Maneuvering System (MMS)**
  - **Immersive seat**: equipped with wearable controls, data gloves, and HMD
  - **Master alarms**: grants full 1-to-1 range of motion over the robot axes and fingers
  - **Master foot**: allows operator to walk in place to drive lateral and forward mobility
  - **Self-interference prevention:** automatically ensures robot and user do not disrupt each other’s movements

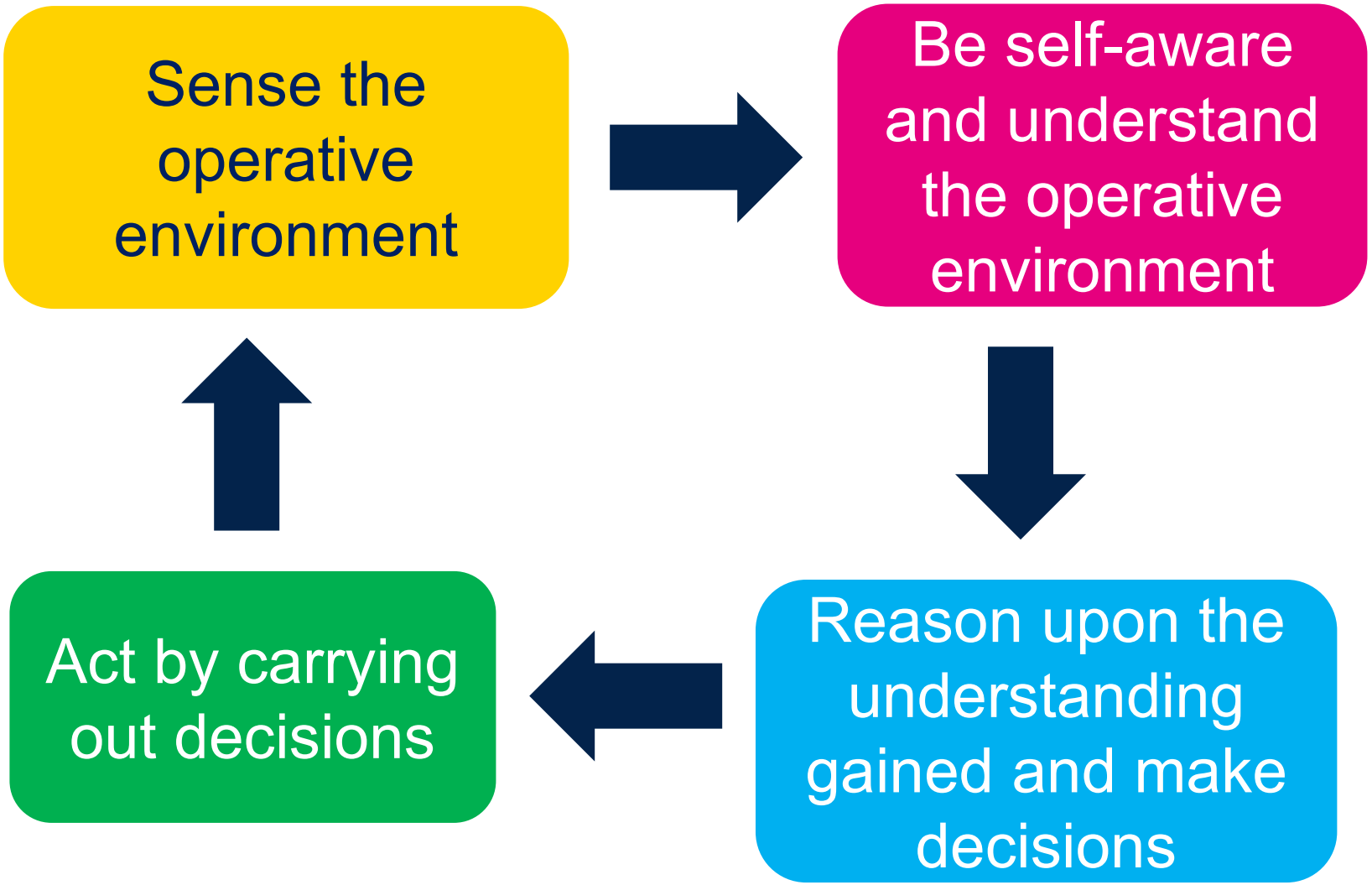
The example

	Toyota T-HR3
AI approach	Human in the loop
Primary training approach	Human guidance
Key hardware differentiator for human-like delicacy	• Cr-N Alloy strain gauges • Torque Servo Modules • Force feedbacks
Target use	Safe assistant (Medical, emergency, home)

How to address the autonomy challenge?

The example

	Figure 03	Toyota T-HR3
AI approach	**VLA end-to-end neural net**	Human in the loop
Primary training approach	Nvidia Isaac Lab	Human guidance
Key hardware differentiator for human-like delicacy	• 3 grams tactile, • 7th Gen Hands 16 dof, • mmWave low latency data comms	• Cr-N Alloy strain gauges • Torque Servo Modules • Force feedbacks
Target use	True general purpose (Home & Commercial)	Safe assistant (Medical, emergency, home)

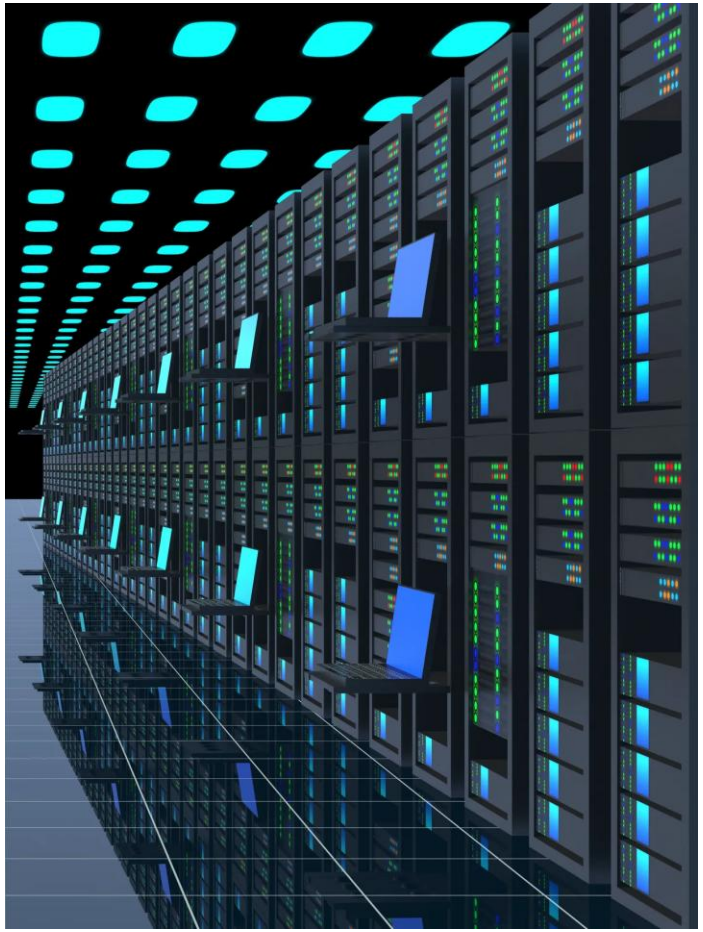

Sense the operative environment
Be self-aware and understand the operative environment
Reason upon the understanding gained and make decisions
Act by carrying out decisions

Sense the operative environment

Be self-aware and understand the operative environment

Reason upon the understanding gained and make decisions

Act by carrying out decisions



NO CLOUD IN LATENCY CRITICAL LOOPS

Figure.AI Helix

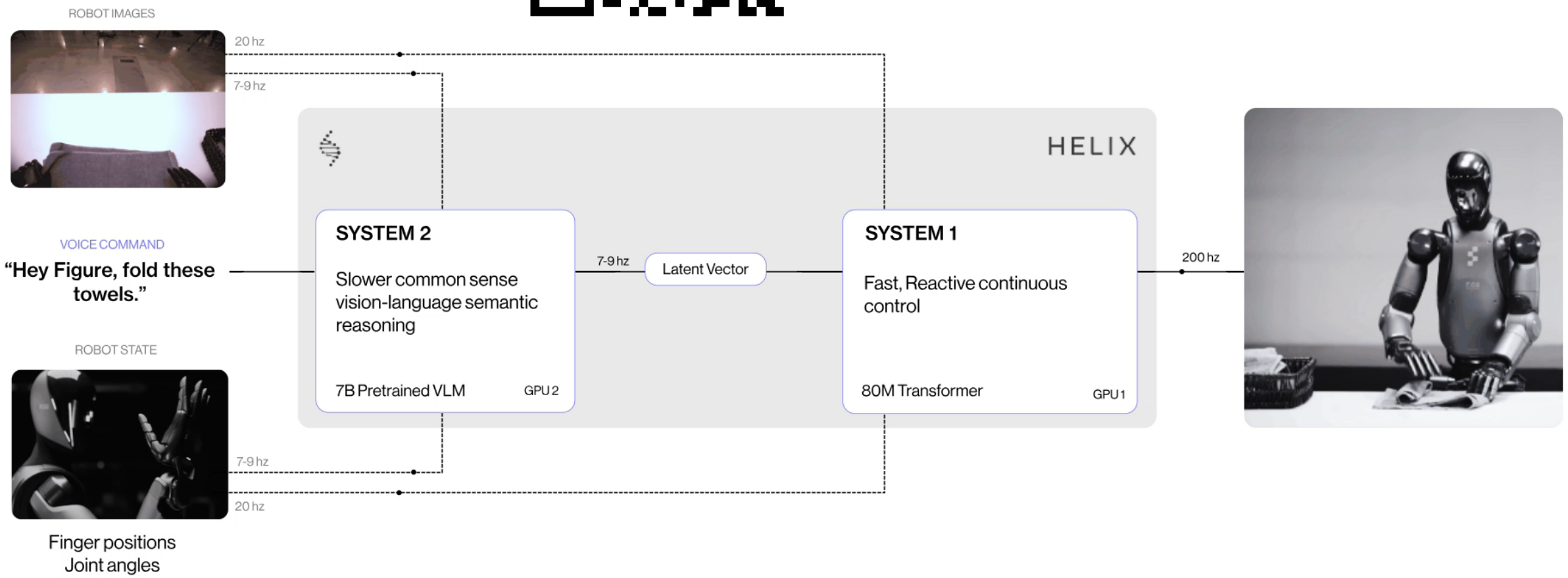
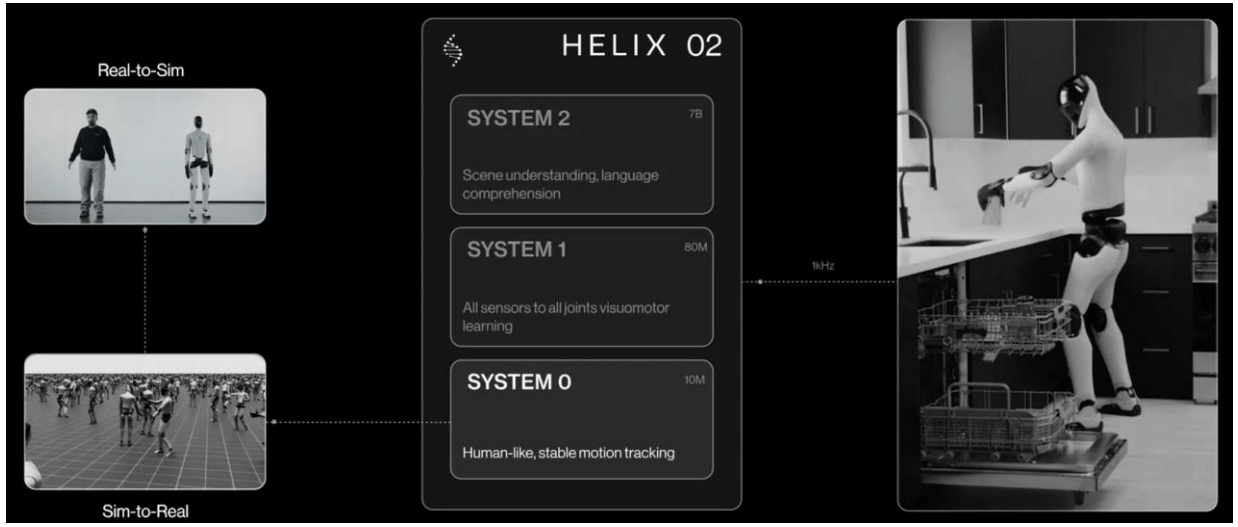


Figure.AI Helix-2

S0: Human-like, whole-body Control

- Drives human body control on the ground of S1 inputs.
- Uses a **10M-parameter** neural network mapping full-body joint state and base motion to joint-level actuator commands at **1 kHz**.
- Learns loco-manipulation by human motion imitation from 1,000+ hours of retargeted joint-level data.

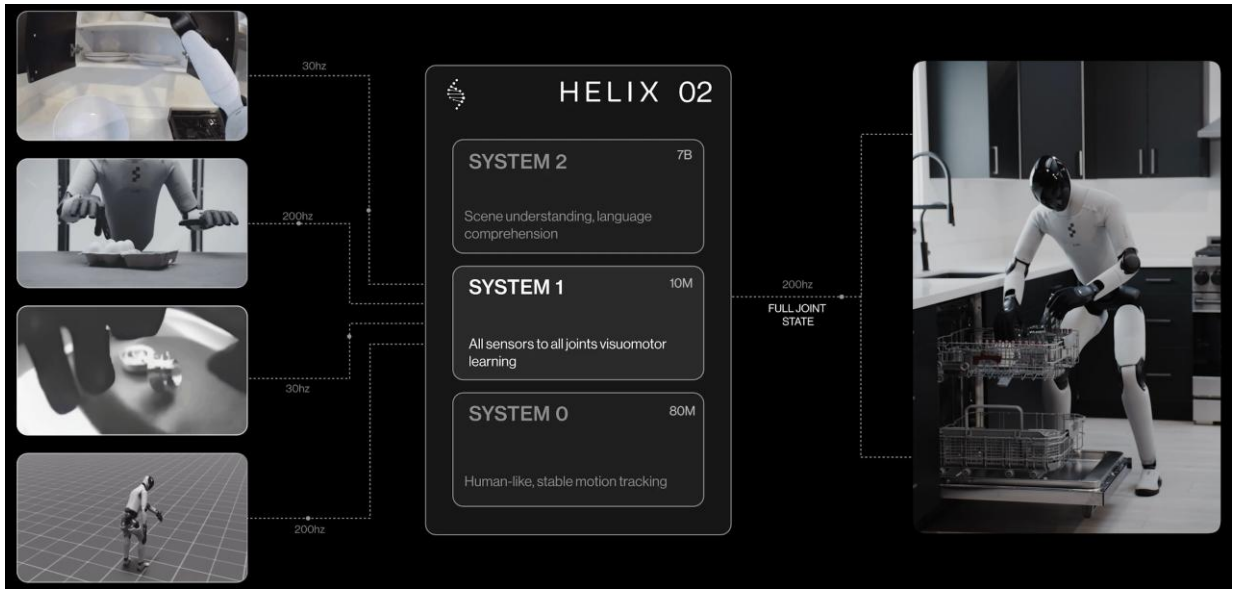


Introducing Helix 02: Full-Body Autonomy

Figure.AI Helix-2

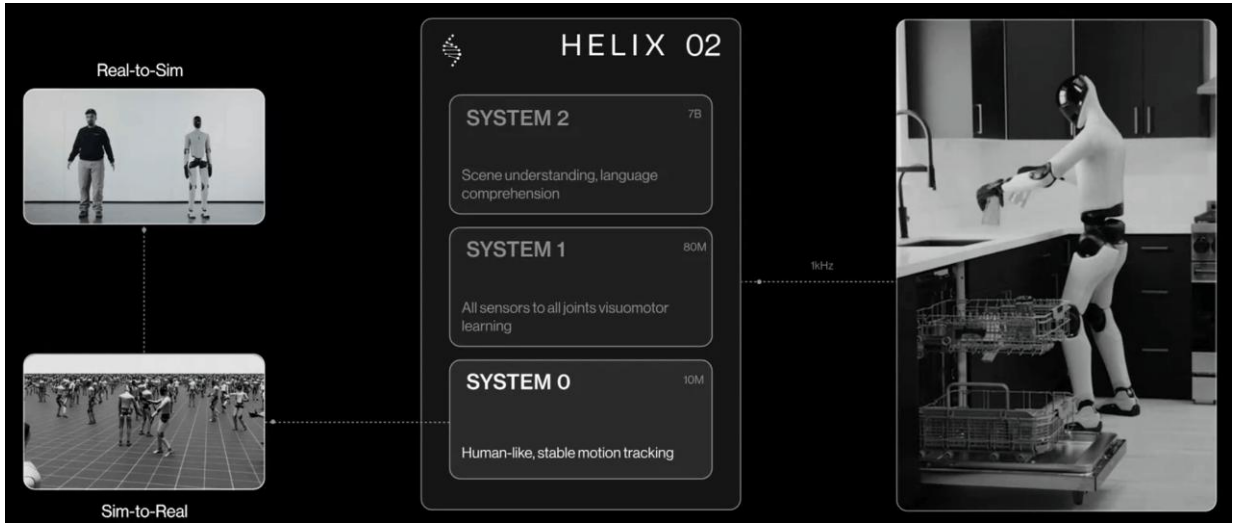
S1: Robot state analysis from Sensors

- On the ground of S2 inputs on main goal, it analyses the inputs from sensors for instance head camera, hand palm camera/TOF, hand tactile/pressure sensors, to provide inputs for S0 to drive complex manipulation.
- Architecture: **80M Transformer** working at **200 Hz**.



S0: Human-like, whole-body Control

- Drives human body control on the ground of S1 inputs.
- Uses a 10M-parameter neural network mapping full-body joint state and base motion to joint-level actuator commands at 1 kHz.
- Learns loco-manipulation by human motion imitation from 1,000+ hours of retargeted joint-level data.

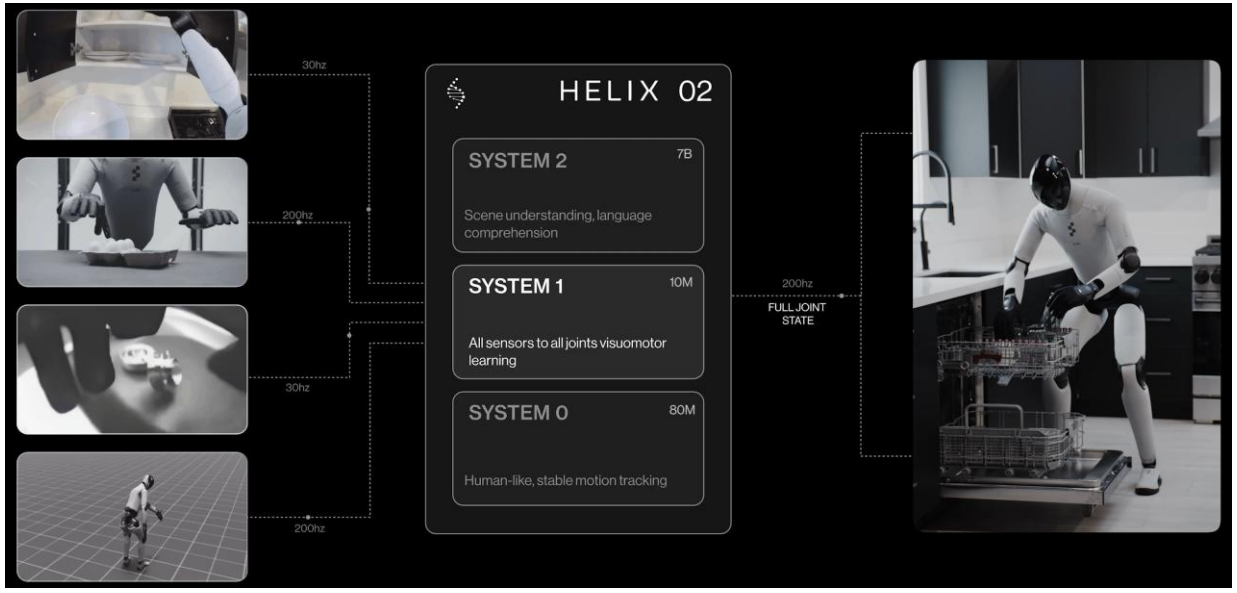


Introducing Helix 02: Full-Body Autonomy

Figure.AI Helix-2

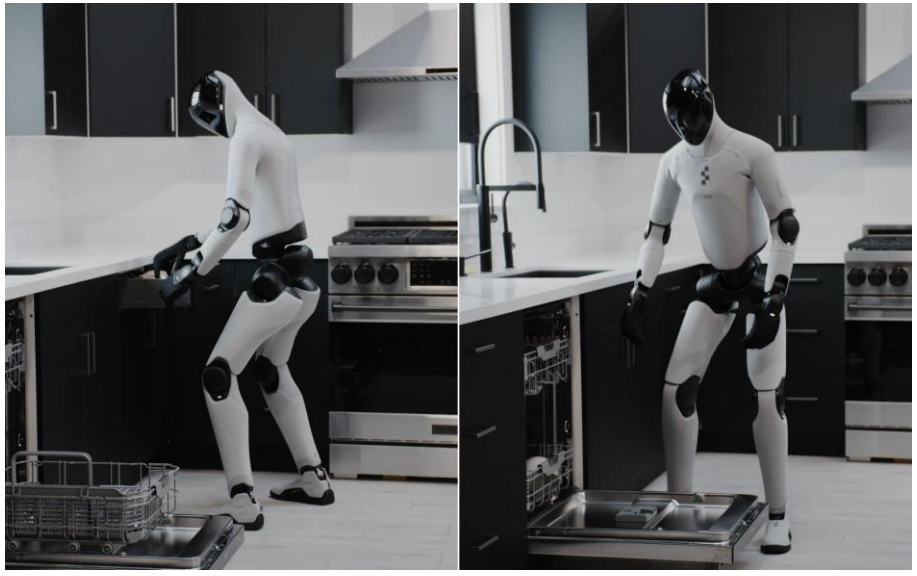
S1: Robot state analysis from Sensors

- On the ground of S2 inputs on main goal, it analyses the inputs from sensors for instance head camera, hand palm camera/TOF, hand tactile/pressure sensors, to provide inputs for S0 to drive complex manipulation.
- Architecture: 80M Transformer working at 200 Hz.



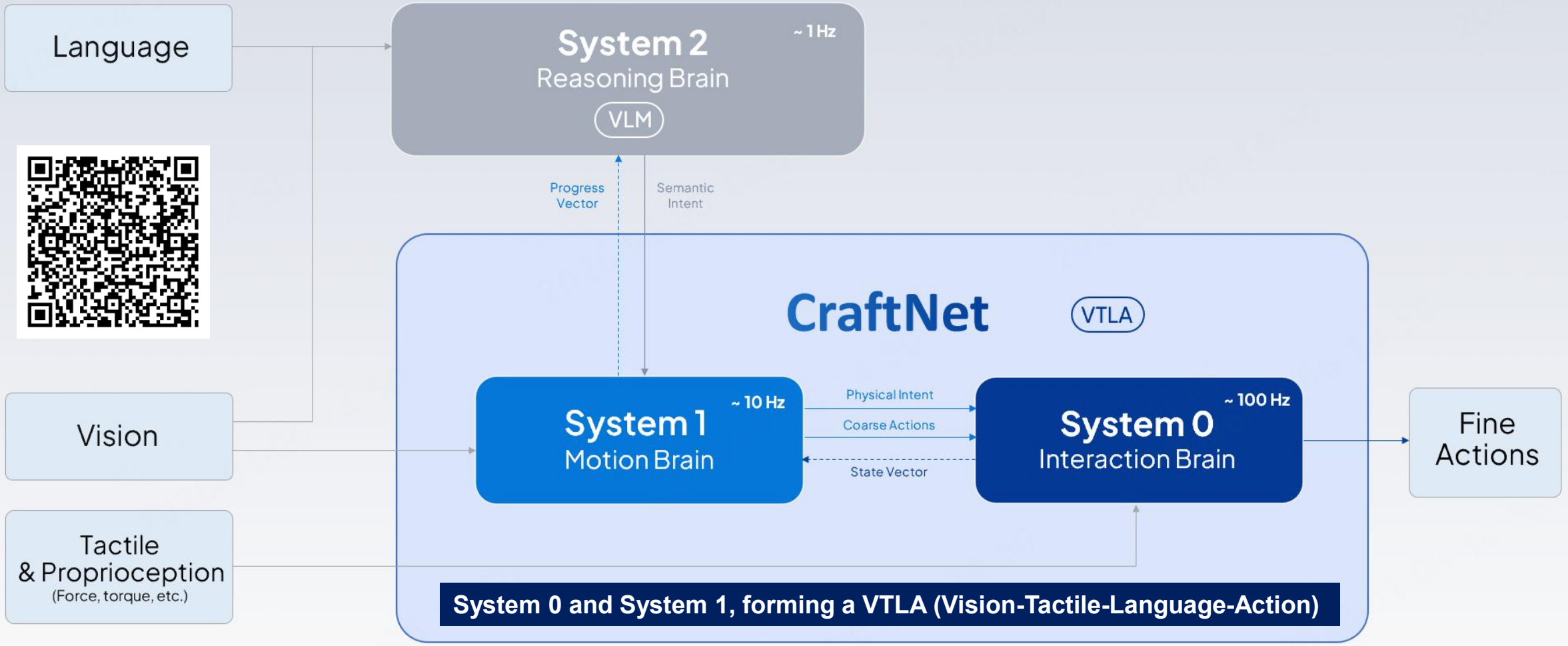
S2: Scene Understanding and Language

- The semantic reasoning layer: processing scenes, understanding language, and producing latent goals for S1. Example: "Go back to the top rack and pick up the cups“
- Architecture: **7B VLM** working at **10Hz**

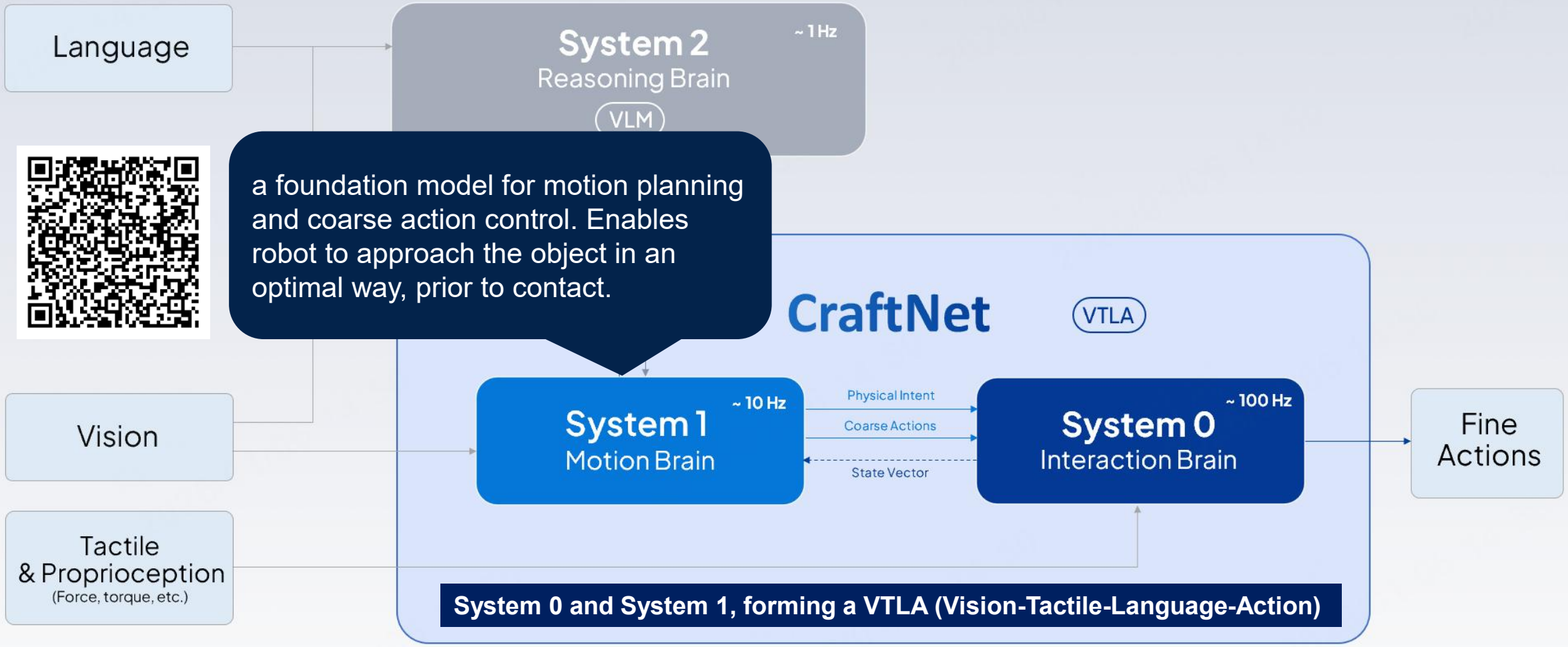


Introducing Helix 02: Full-Body Autonomy

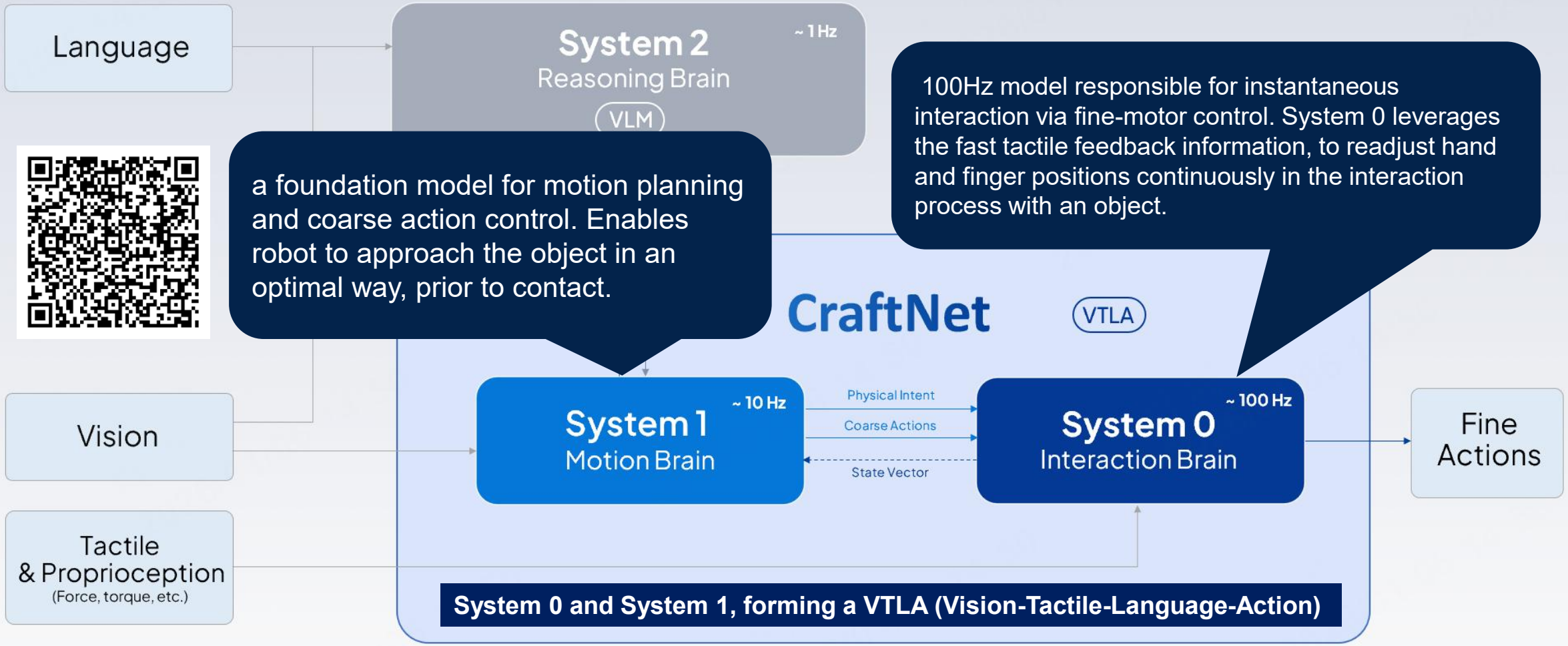
CraftNet is an end-to-end, hierarchical VTLA model for fine manipulation, enabling native anthropomorphic last-millimeter interaction.



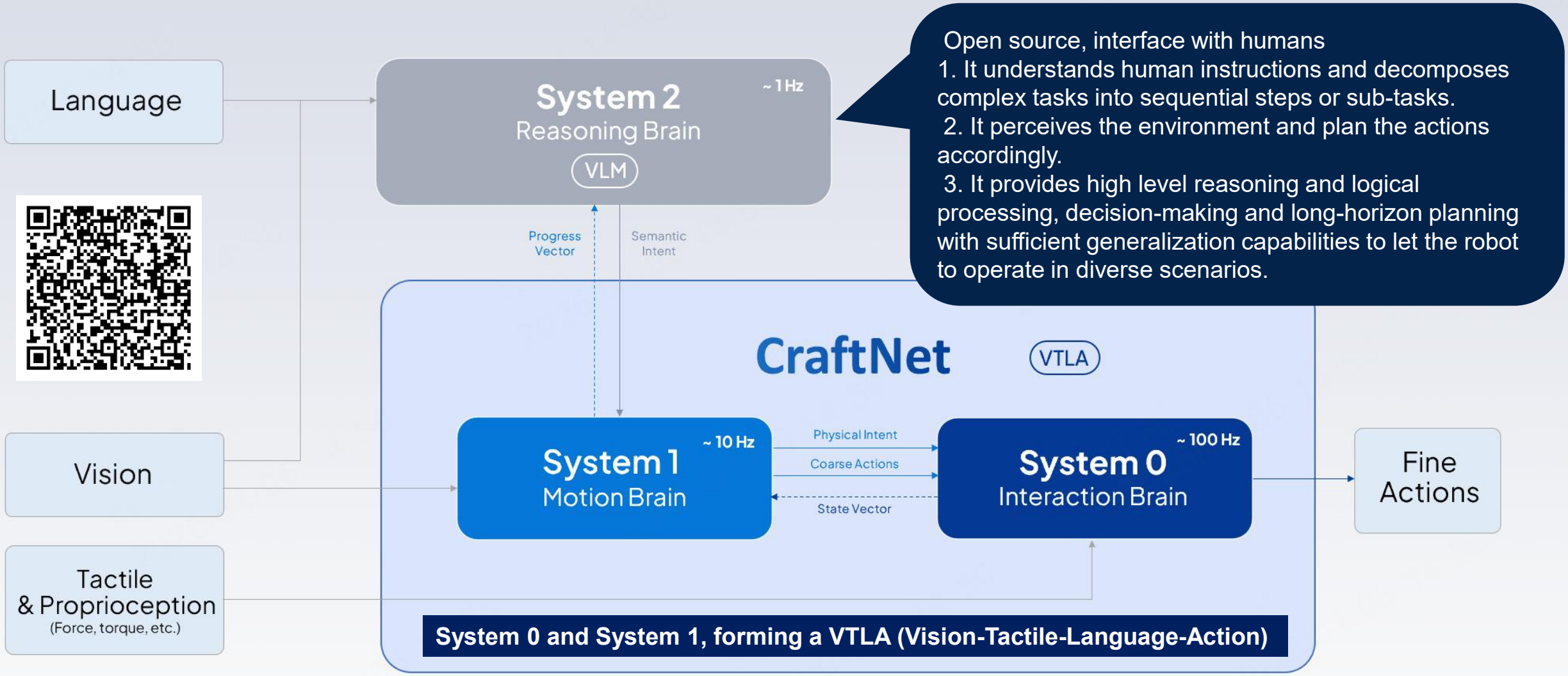
CraftNet is an end-to-end, hierarchical VTLA model for fine manipulation, enabling native anthropomorphic last-millimeter interaction.



CraftNet is an end-to-end, hierarchical VTLA model for fine manipulation, enabling native anthropomorphic last-millimeter interaction.

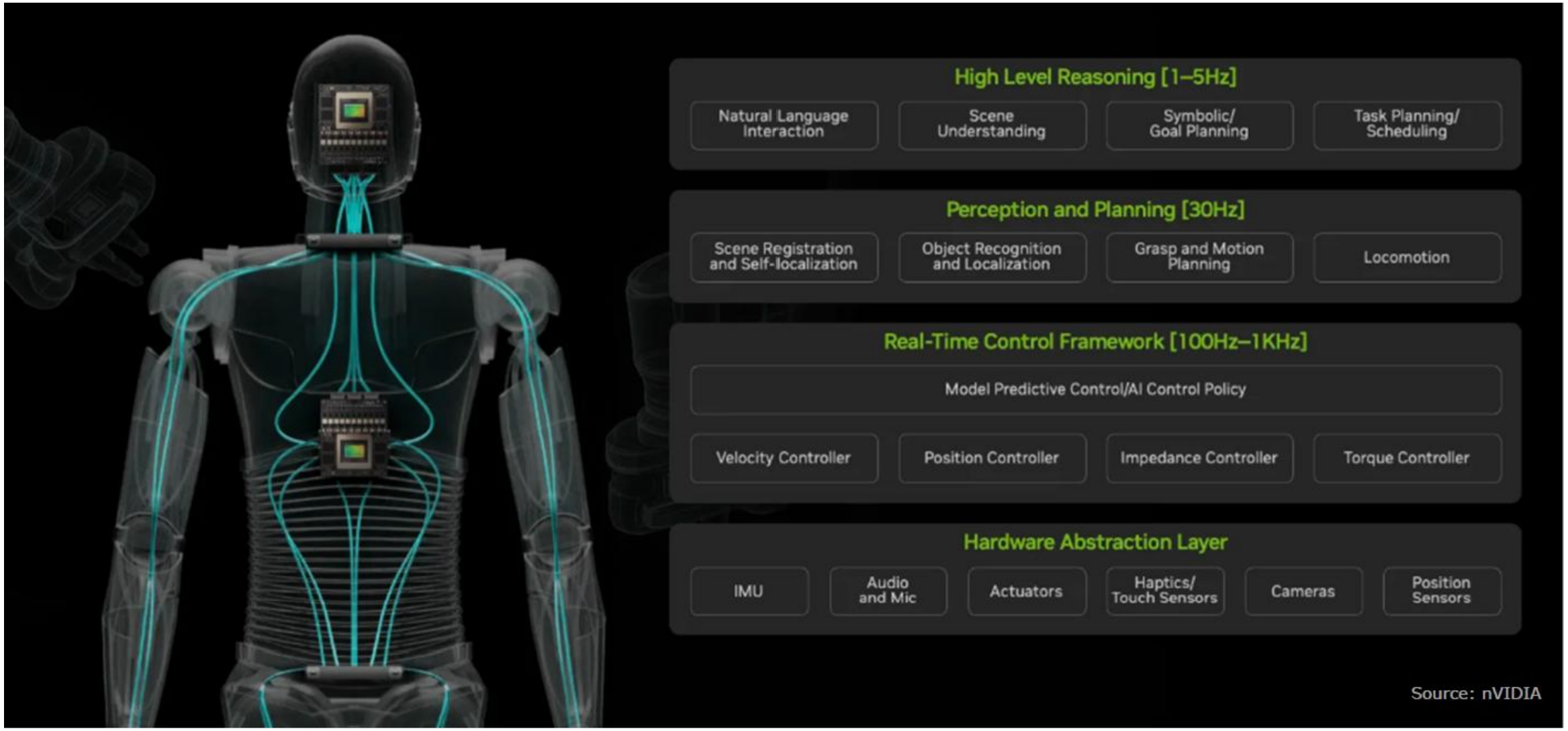


CraftNet is an end-to-end, hierarchical VTLA model for fine manipulation, enabling native anthropomorphic last-millimeter interaction.



Open source, interface with humans
1. It understands human instructions and decomposes complex tasks into sequential steps or sub-tasks.
2. It perceives the environment and plan the actions accordingly.
3. It provides high level reasoning and logical processing, decision-making and long-horizon planning with sufficient generalization capabilities to let the robot to operate in diverse scenarios.

The nVIDIA summary



Requirements summary

	System 0	System 1	System 2
HELIX		200 Hz 80 M Transformer	9 Hz 7b VLM
HELIX 2	**1000 Hz** **10 M**	**200 Hz** **80 M Transformer**	**9 Hz** **7b VLM**
CraftNet	100Hz	10Hz	1Hz VLM
nVidia	100-1000 Hz	30 Hz	5 Hx

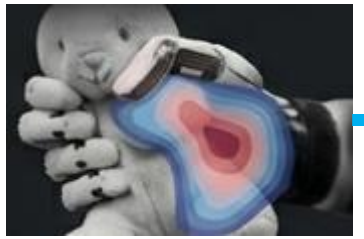
Requirements summary

	System 0	System 1	System 2
HELIX		200 Hz 80 M Transformer	9 Hz 7b VLM
HELIX 2	**1000 Hz** **10 M**	**200 Hz** **80 M Transformer**	**9 Hz** **7b VLM**
CraftNet	100Hz	10Hz	1Hz VLM
nVidia	100-1000 Hz	30 Hz	5 Hx

Most compelling requirements

The architecture of action

INPUTS



Tactile

Tactile Encoder

Image Observation

Visual Encoder

Put bowl in plate

Language instruction

Text Encoder

Robot state

State Encoder

Projections

Edge Language Model Backbone

Action Decoder / Diffusion Transformer

Temporal Convolutions

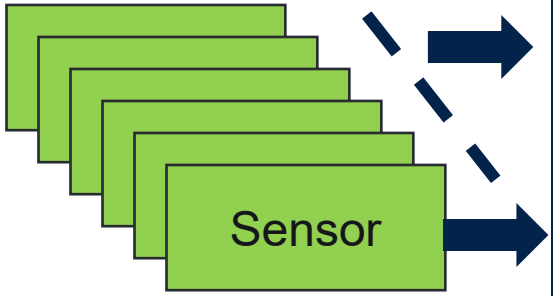
Action data

Action data

through time

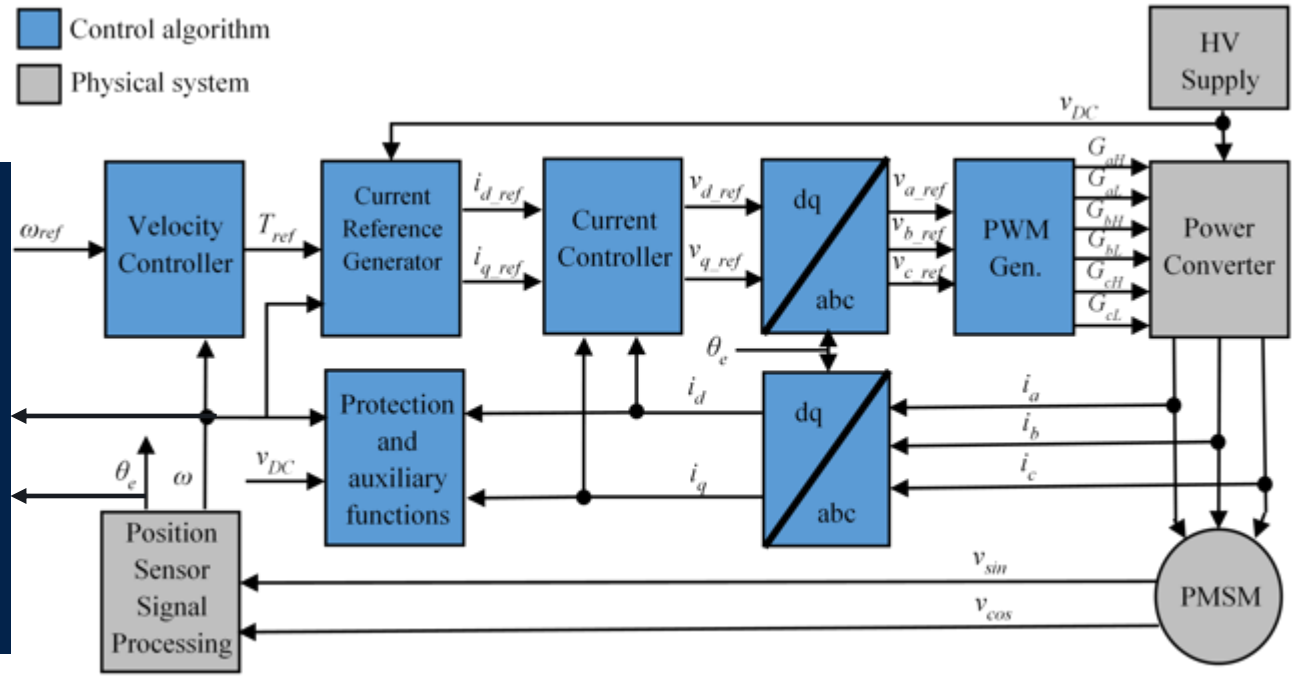
Motor Control Reshaped by AI

Generalized Sensing to Action Model

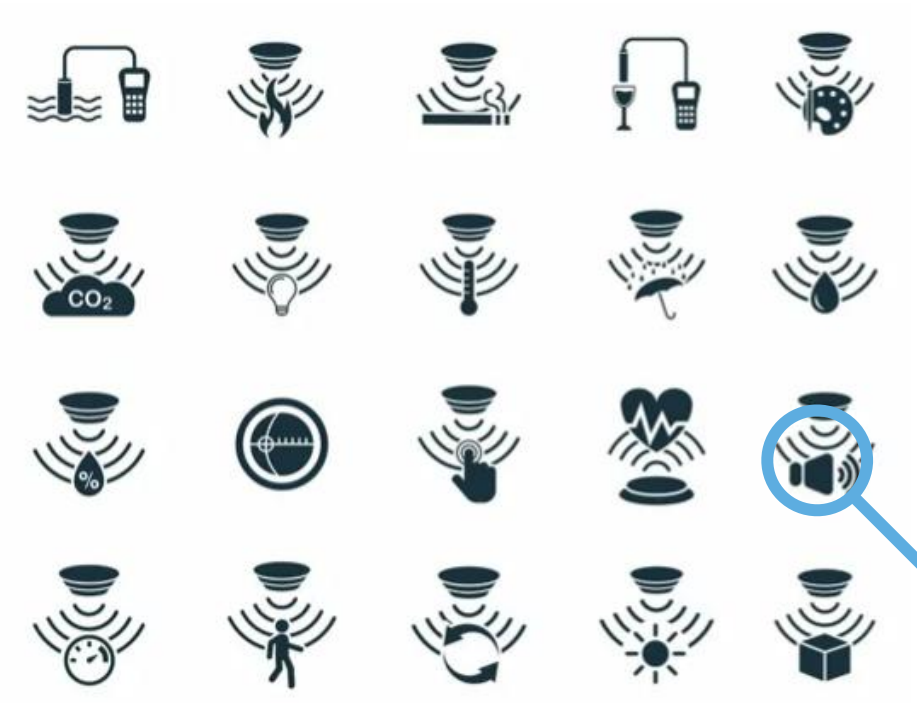


Helix 2 like

Field Oriented Control

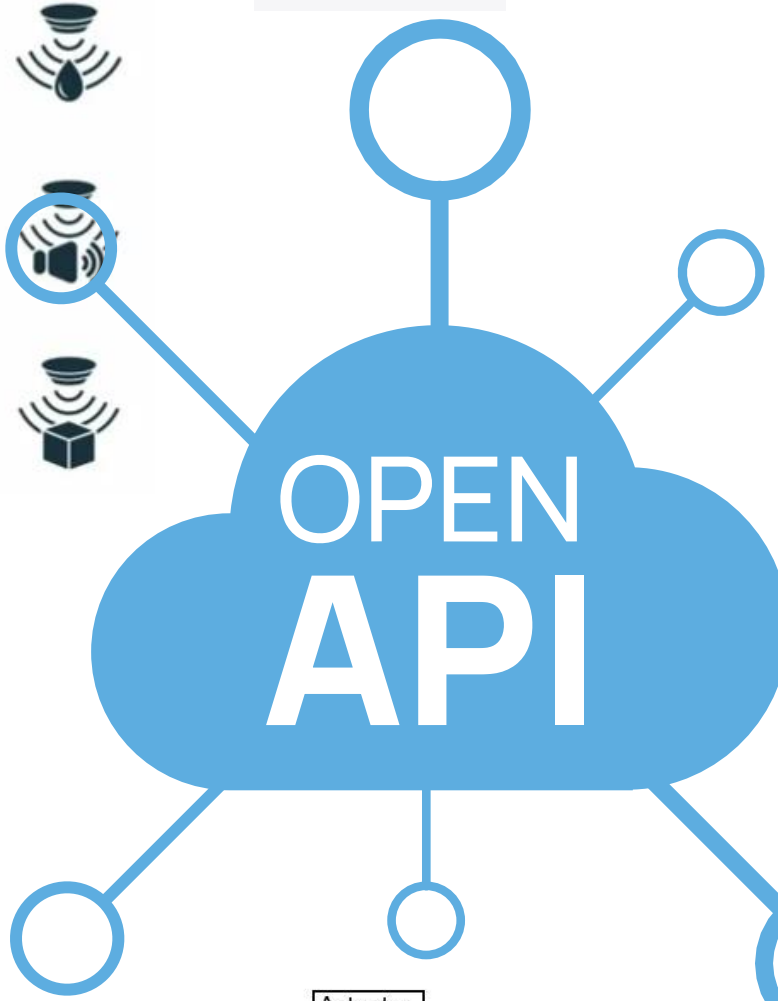


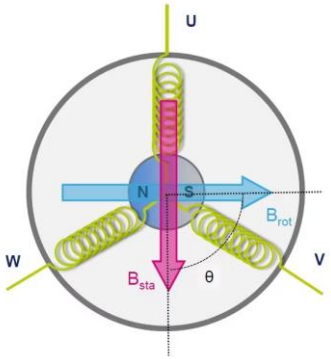
Call for action



Sensors

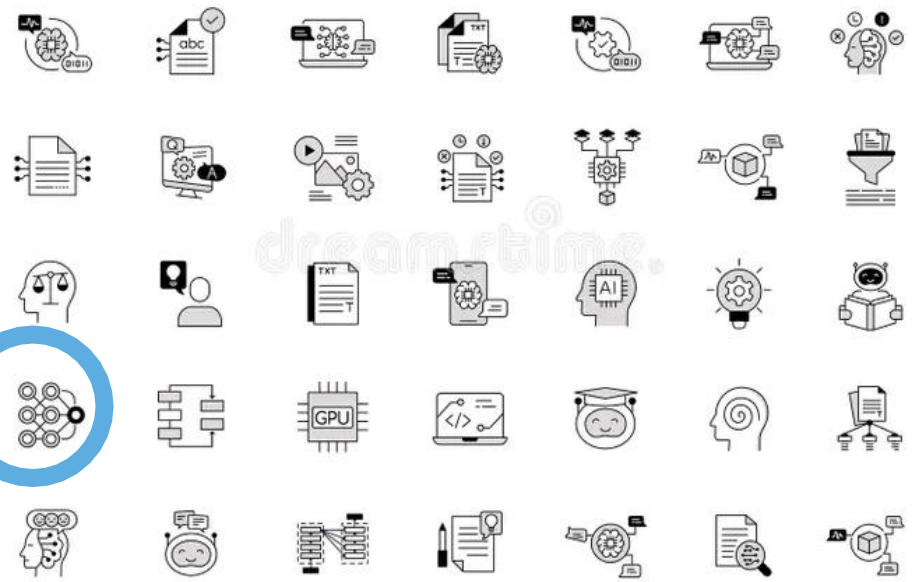
Foundation Models

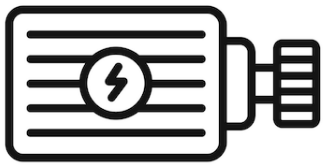




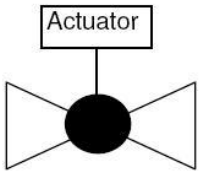
Motor control

Gen Edge AI workloads





Actuators



gpustack/gguf-parser-go

VLA @ HF

"vla": [**# Core terms** "vla gguf", "vision-language-action gguf", "vision language action gguf", "vision-to-action gguf", "vision to action gguf", "vision action gguf", "vl action gguf", "vision-language action gguf", "vision language policy gguf", "vision-language policy gguf", **# Known model families / projects** "openvla gguf", "smolvla gguf", "spatialvla gguf", "seed-vla gguf", "openpi gguf", "pi0 gguf", "pi-0 gguf", "gr00t gguf", "groot gguf", "gr00t-n1 gguf", "rfm-1 gguf", "gato gguf", "nitrogen gguf", **# Embodied / robotics phrasing** "robotics policy gguf", "robotics control gguf", "robot action gguf", "robot control gguf", "embodied ai gguf", "embodied agent gguf", "embodied navigation gguf", "embodied manipulation gguf", "robot manipulation gguf", "manipulation policy gguf", "control policy gguf", "robot policy gguf", "embodied policy gguf", "action model gguf", "action policy gguf", "embodied foundation model gguf", "multimodal robotics gguf", "multimodal for robotics gguf", "vision-to-action model gguf", "affordance prediction gguf", "driving planning gguf", **# Other common names in this space** "rt-1 gguf", "rt1 gguf", "rt-2 gguf", "rt2 gguf", "octo gguf", "vima gguf“], ….. **etc.**

gpustack/gguf-parser-go

About gguf-parser-go

- Accordingly, to the authors of the tool gguf-parser-go, the evaluation results usually deviate from the actual hardware usage by about **100MiB**.

- The tool prediction capabilities have been validated on UMA and NUMA hardware chips such as
  - UMA → Apple Mac Studio (M2)
  - NUMA → Intel i5-14600k and NVIDIA GeForce RTX 4080

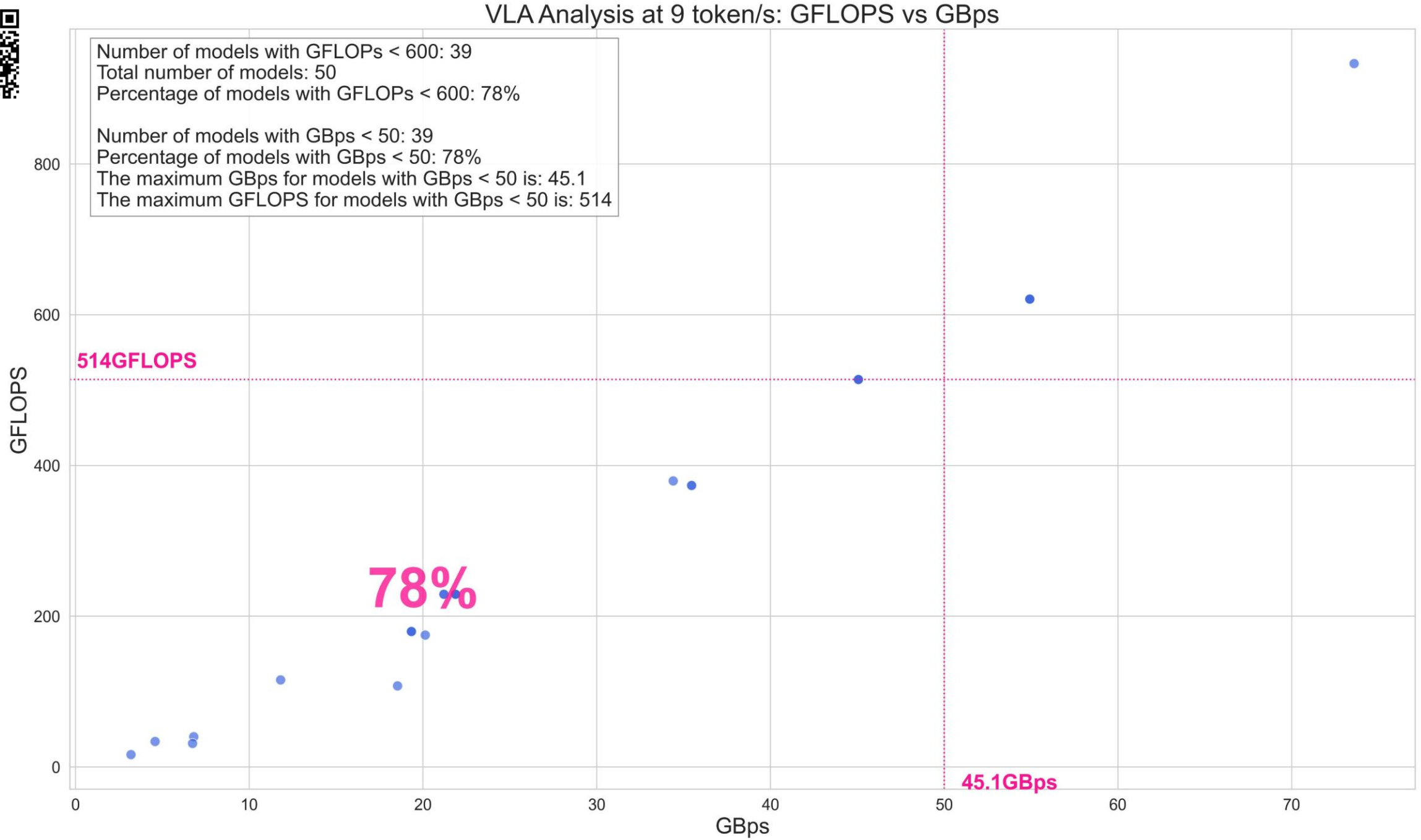
gpustack/gguf-parser-go

Summary statistics for VLAs

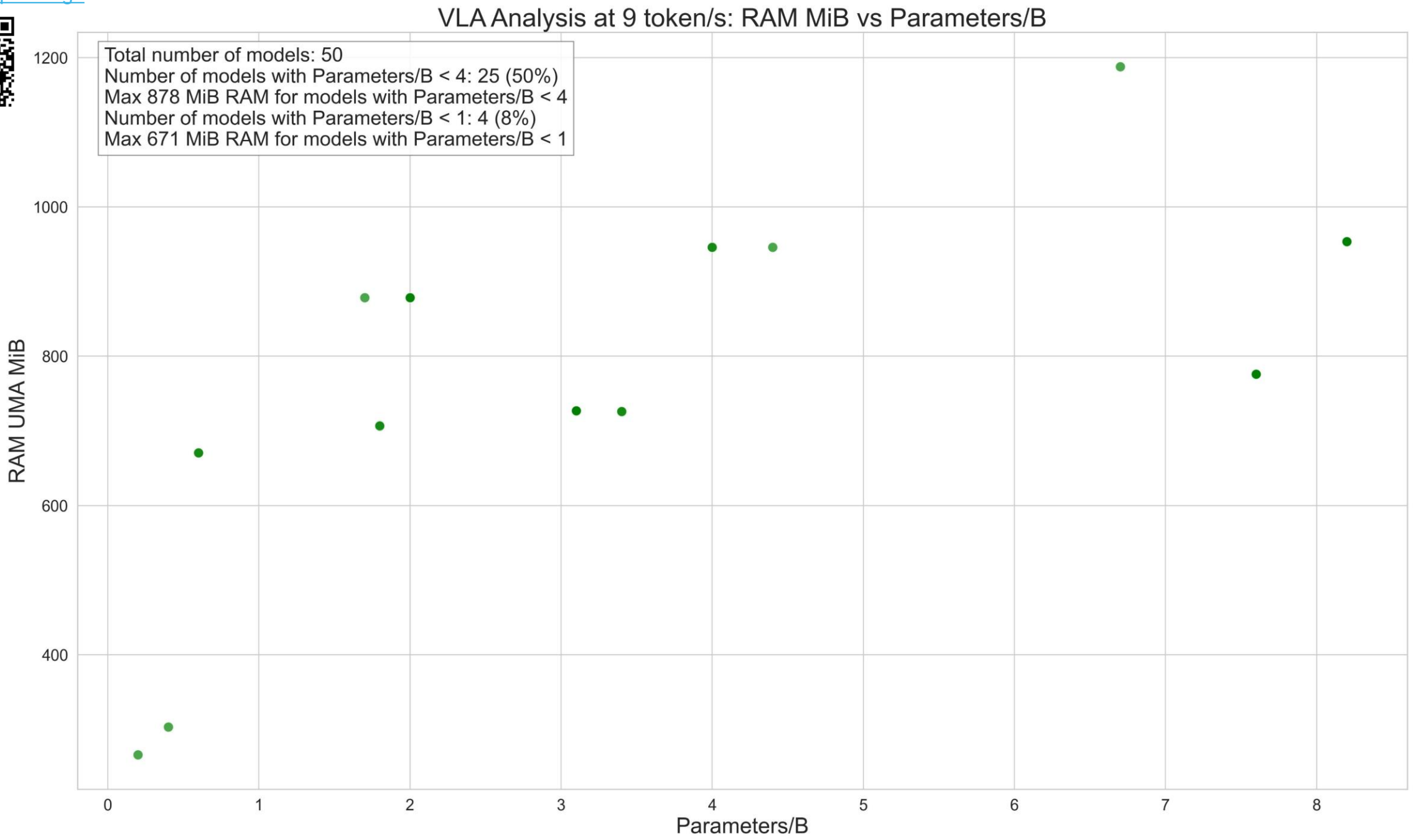
Regime	Models	Param./B ($\mu \pm \sigma$ [Min, Max])	GFLOPS ($\mu \pm \sigma$ [Min, Max])	GBps ($\mu \pm \sigma$ [Min, Max])	UMA/MiB ($\mu \pm \sigma$ [Min, Max])
Overall	50	4.90 ± 2.77 [0.20, 8.20]	368.55 ± 212.54 [16.56, 933.40]	33.60 ± 17.14 [3.18, 73.58]	796.80 ± 152.76 [266.15, 1187.84]
Parameters/B < 4	25	2.41 ± 1.03 [0.20, 3.40]	180.06 ± 75.18 [16.56, 229.40]	18.25 ± 6.17 [3.18, 21.86]	709.65 ± 141.89 [266.15, 878.46]
Parameters/B < 1	4	0.45 ± 0.19 [0.20, 0.60]	30.53 ± 10.04 [16.56, 40.29]	5.32 ± 1.76 [3.18, 6.80]	477.69 ± 223.23 [266.15, 670.56]

Assumption: 9 tokens per second (Helix 2)

gpustack/gguf-parser-go



gpustack/gguf-parser-go

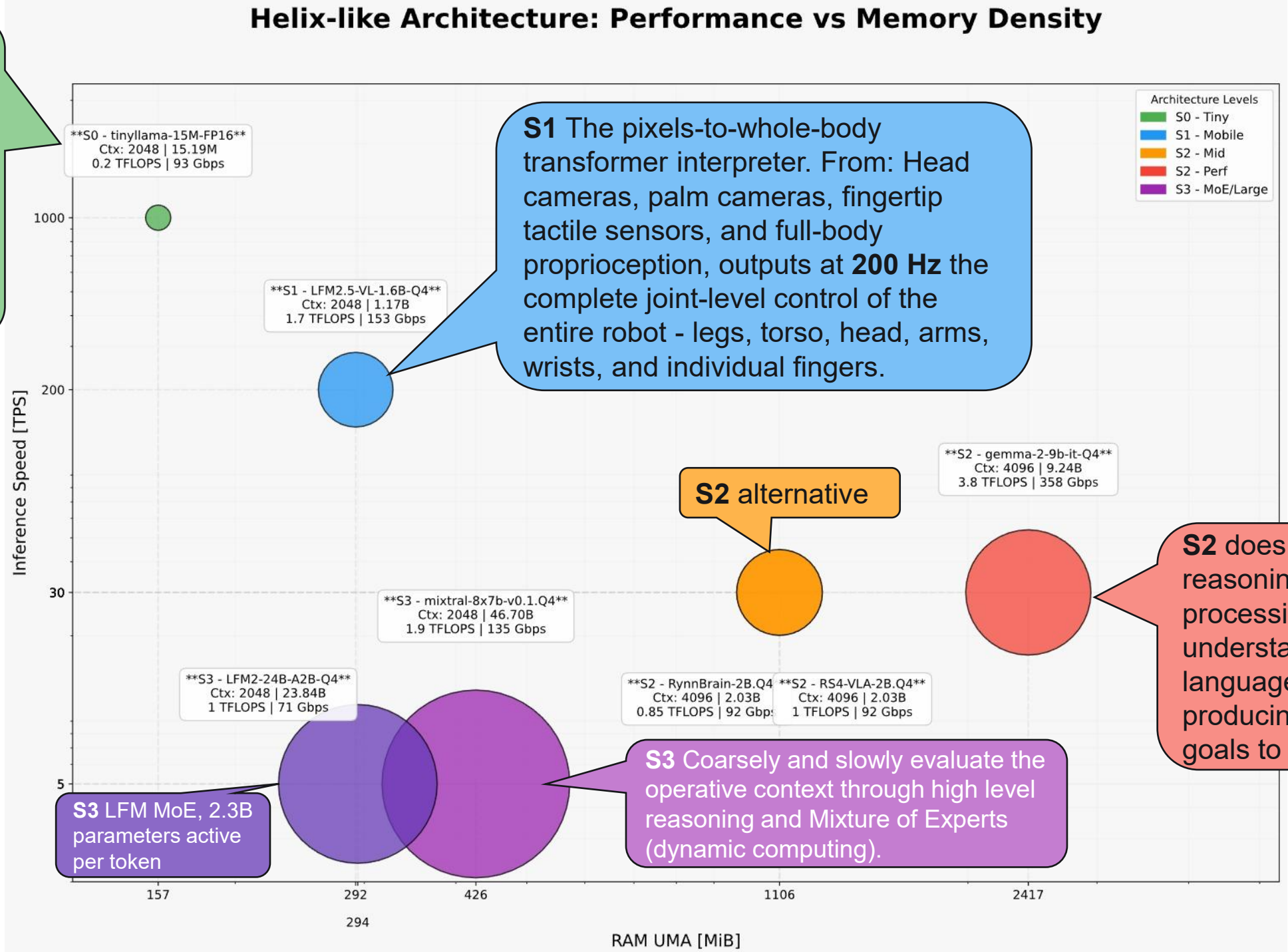


Backbone-unique VLA - Params/B < 4

Model	Backbone	Param./B	GBps	GFLOPS	UMA/MiB
SmolVLM2-256M-Video-Instruct-GGUF-BPU	llama	0.20	3.183	16.563	266.15
InternVL2_5-1B-GGUF-BPU	qwen2	0.60	4.566	33.866	670.56
ShowUI-2B-Q4_K_M-GGUF	qwen2vl	1.80	11.797	115.687	707.57
RynnBrain-2B-GGUF	qwen3vl	2.00	19.329	179.957	878.46

Assumption: 9 tokens per second (Helix 2)

S0 a 10M-parameter executor that read full-body joint state and base motion and outputs joint-level actuator commands at **1 kHz**.



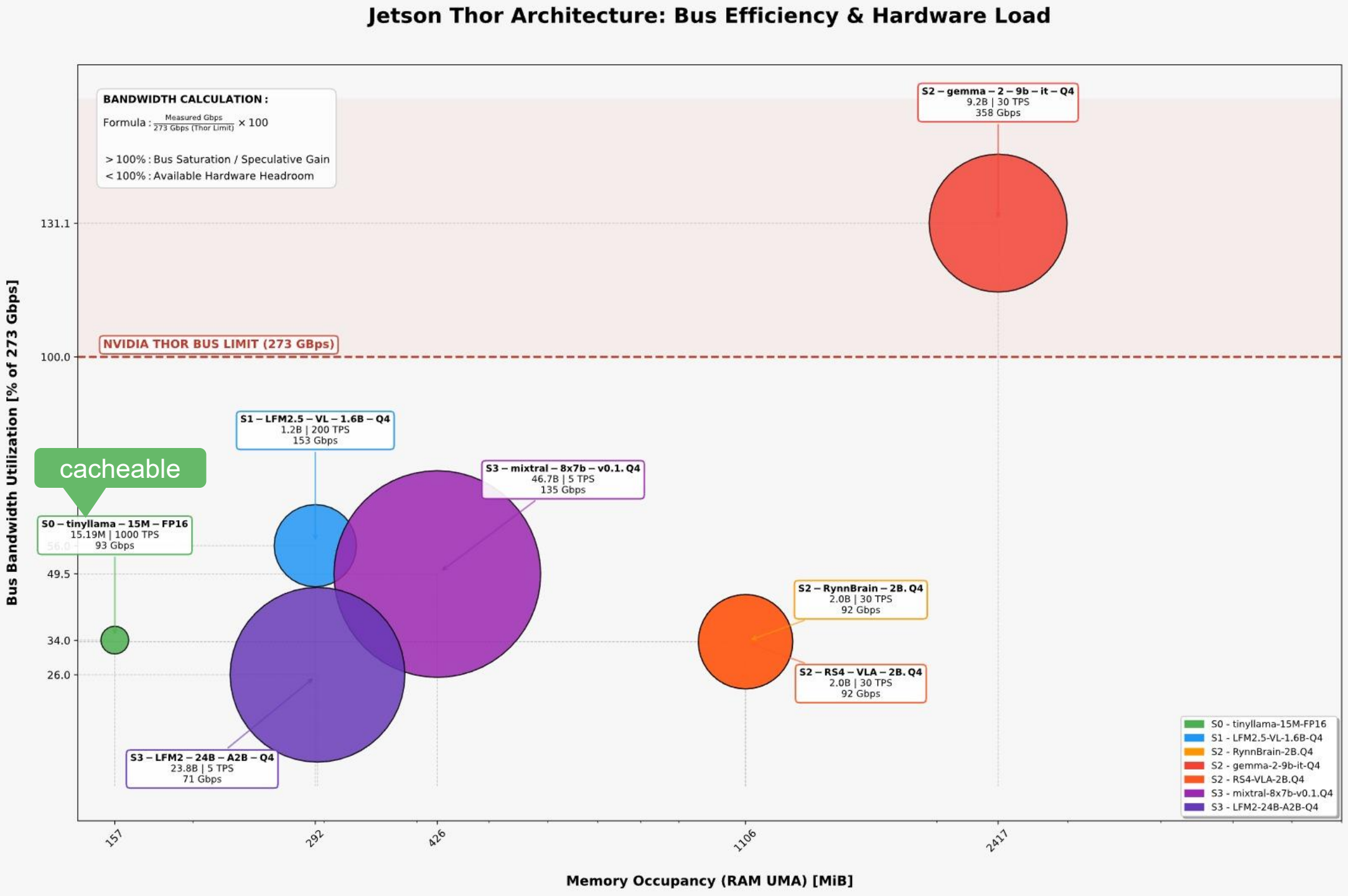
S1 The pixels-to-whole-body transformer interpreter. From: Head cameras, palm cameras, fingertip tactile sensors, and full-body proprioception, outputs at **200 Hz** the complete joint-level control of the entire robot - legs, torso, head, arms, wrists, and individual fingers.

S2 alternative

S2 does semantic reasoning by processing scenes, understanding language, and producing latent goals to feed S1.

S3 Coarsely and slowly evaluate the operative context through high level reasoning and Mixture of Experts (dynamic computing).

S3 LFM MoE, 2.3B parameters active per token


Jetson Thor Architecture: Bus Efficiency & Hardware Load
BANDWIDTH CALCULATION :
Formula : Measured Gbps / 273 Gbps (Thor Limit) × 100
> 100% : Bus Saturation / Speculative Gain
< 100% : Available Hardware Headroom
S2 – gemma – 2 – 9b – it – Q4
9.2B | 30 TPS
358 Gbps
NVIDIA THOR BUS LIMIT (273 GBps)
S1 – LFM2.5 – VL – 1.6B – Q4
1.2B | 200 TPS
153 Gbps
cacheable
S3 – mixtral – 8x7b – v0.1. Q4
46.7B | 5 TPS
135 Gbps
S0 – tinyllama – 15M – FP16
15.19M | 1000 TPS
93 Gbps
S2 – RynnBrain – 2B. Q4
2.0B | 30 TPS
92 Gbps
S2 – RS4 – VLA – 2B. Q4
2.0B | 30 TPS
92 Gbps
S3 – LFM2 – 24B – A2B – Q4
23.8B | 5 TPS
71 Gbps
Bus Bandwidth Utilization [% of 273 Gbps]
131.1
100.0
56.0
49.5
34.0
26.0
157
292
426
1106
2417
Memory Occupancy (RAM UMA) [MiB]
S0 - tinyllama-15M-FP16
S1 - LFM2.5-VL-1.6B-Q4
S2 - RynnBrain-2B.Q4
S2 - gemma-2-9b-it-Q4
S2 - RS4-VLA-2B.Q4
S3 - mixtral-8x7b-v0.1.Q4
S3 - LFM2-24B-A2B-Q4

Key trends in 2026

Integration of Multimodal Sensing (Tactile & Force)

. **Beyond Vision:** New architectures are incorporating **tactile and force sensing** to handle "contact-rich" tasks where vision alone fails (e.g., detecting slips or managing insertion force).

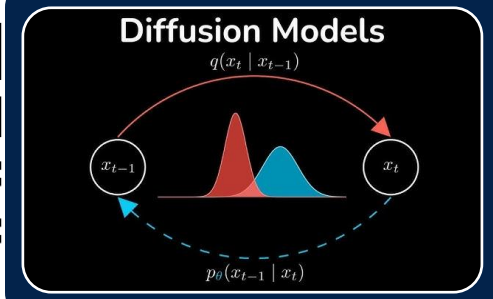
. **Tactile-VLA Models:** Models like **Rho-Alpha** and **VLA-Touch** are demonstrating significant success in high-precision tasks such as charger insertion and bimanual manipulation.

Key trends in 2026

Integration of Multimodal Sensing (Tactile & Force)

. **Beyond Vision:** New architectures are incorporating **tactile and force sensing** to handle "contact-rich" tasks where vision alone fails (e.g., detecting slips or managing insertion force).

. **Tactile-VLA Models:** Models like **Rho-Alpha** and **VLA-Touch** are demonstrating significant success in high-precision tasks such as charger insertion and bimanual manipulation.



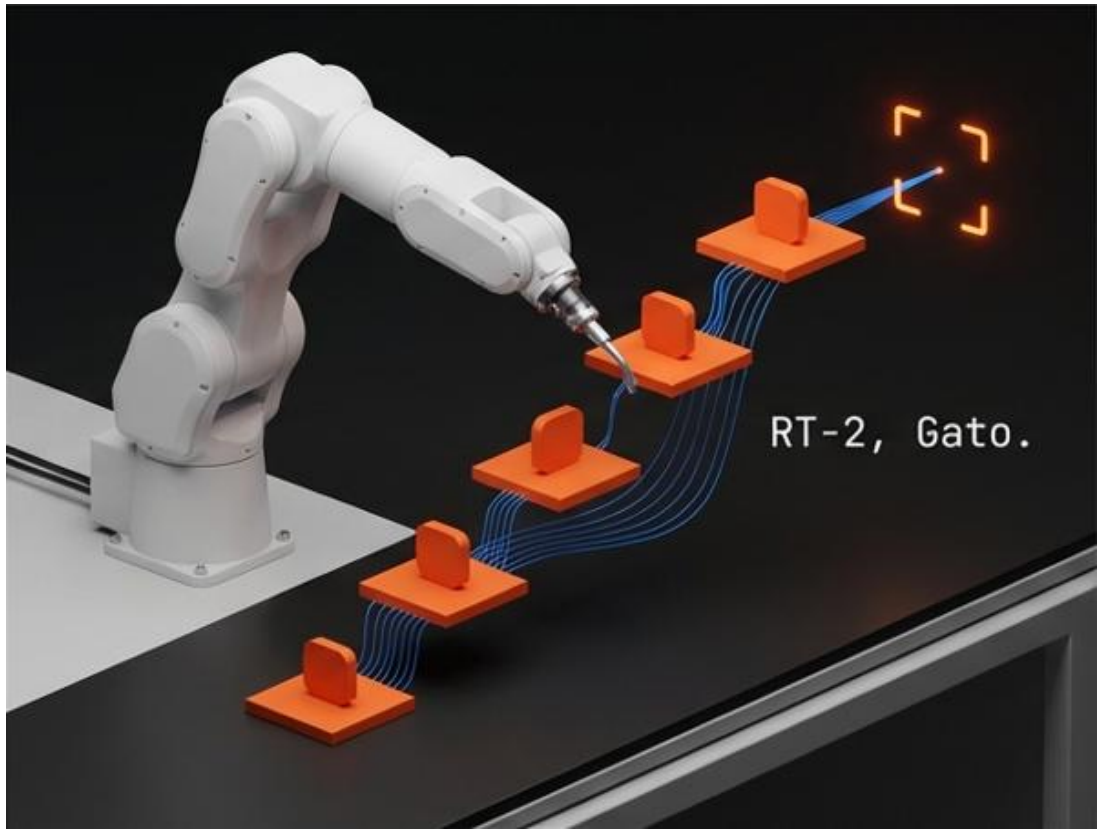
Shift from Autoregressive to Diffusion & Hybrid Models

. **Efficiency & Control:** Researchers are moving away from purely autoregressive policies toward **diffusion-based and hybrid models**. These approaches generate action sequences more efficiently and provide better alignment between abstract reasoning and low-level motor control.

. **Discrete Tokenization:** Some systems, like **Cortex VLA**, are tokenizing both actions and perception into a language-like vocabulary, allowing them to leverage the reasoning power of large language models for robot control. **)**

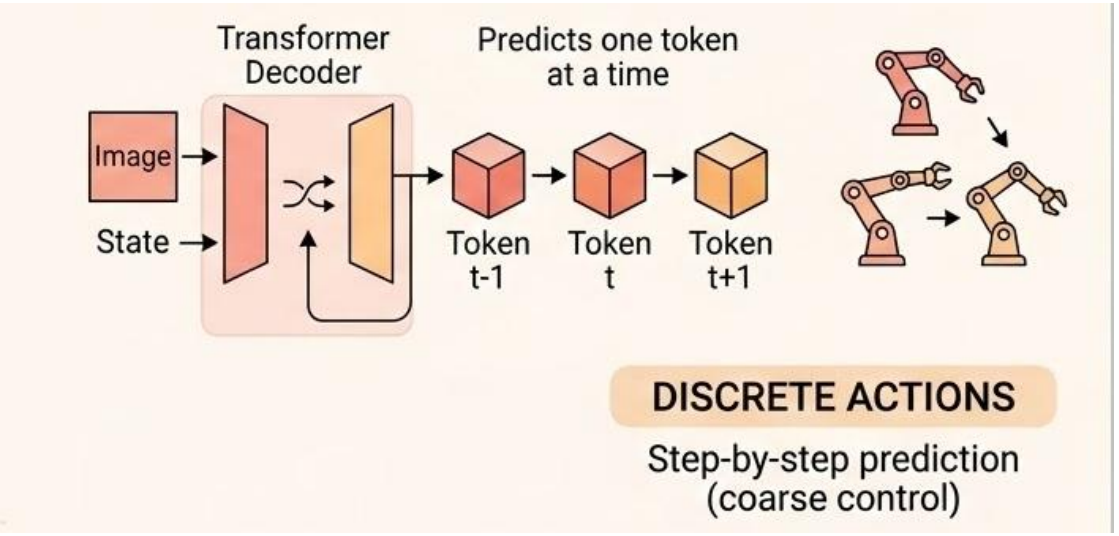
Architectural Divergence: Tokens vs. Diffusion

Autoregressive (Discrete Tokens)



Decoder-Based Autoregressive Generation

- Treats actions as sequences of discrete tokens and predicts them autoregressively, one token at a time
- Learn and generate P_{data}
- step-by-step actions too coarse for fast, precise, dexterous control
- **Models : RT-2, OpenVLA**

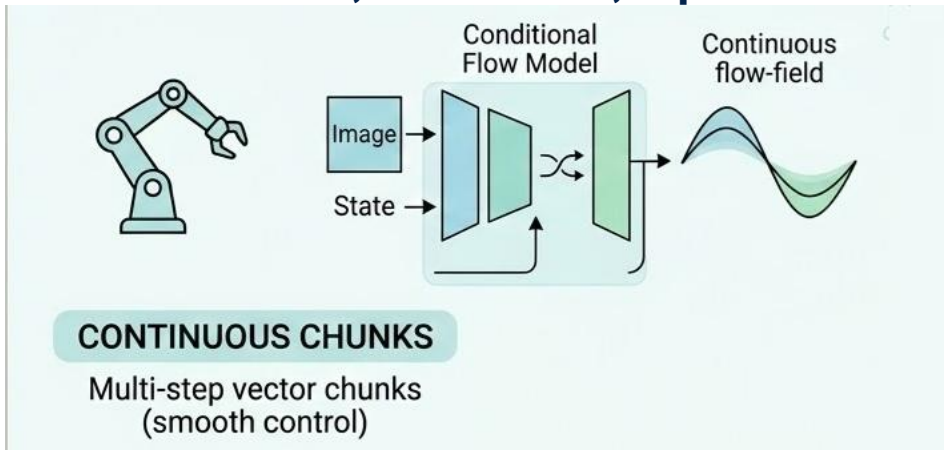


Architectural Paradigm Shift in VLA Models

Diffusion (Continuous Denoising)

Conditional Flow Matching based action chunk Generation

- Treats actions as continuous chunks (multi-step control outputs) and generates these continuous action vectors directly.
- **Gradually convert one distribution to another being compute-efficient**
- produce smooth, high-frequency actions enabling precise, dexterous control
- **Models : smolVLA, VLA-JEPA, OpenHelix**

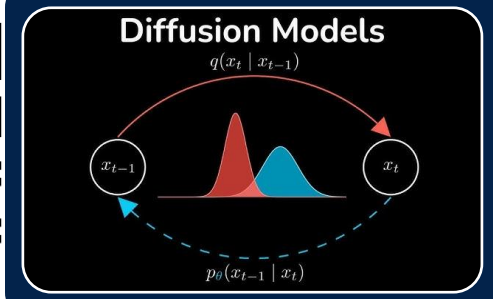


Key trends in 2026

Integration of Multimodal Sensing (Tactile & Force)

. **Beyond Vision:** New architectures are incorporating **tactile and force sensing** to handle "contact-rich" tasks where vision alone fails (e.g., detecting slips or managing insertion force).

. **Tactile-VLA Models:** Models like **Rho-Alpha** and **VLA-Touch** are demonstrating significant success in high-precision tasks such as charger insertion and bimanual manipulation.



Shift from Autoregressive to Diffusion & Hybrid Models

. **Efficiency & Control:** Researchers are moving away from purely autoregressive policies toward **diffusion-based and hybrid models**. These approaches generate action sequences more efficiently and provide better alignment between abstract reasoning and low-level motor control.

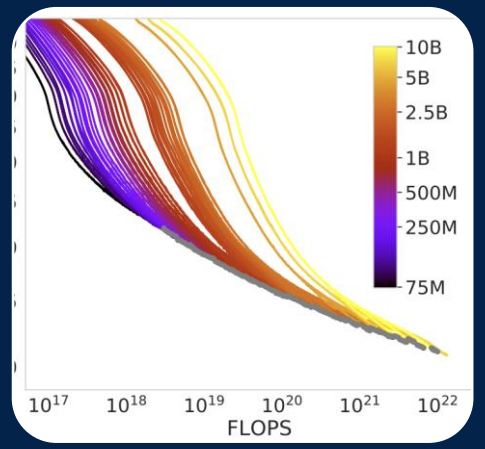
. **Discrete Tokenization:** Some systems, like **Cortex VLA**, are tokenizing both actions and perception into a language-like vocabulary, allowing them to leverage the reasoning power of large language models for robot control. **)**

Long-Horizon Reasoning and World Models

. **Goal-Oriented Missions:** Robots are evolving from executing single instructions to fulfilling broad missions (e.g., "inspect this facility") by inferring context and planning multi-step routes.

. **Visual Foresight:** Emerging models "foresee" environmental changes before acting. A major research goal for 2026 is achieving **1-hour coherent world model predictions** to allow robots to simulate "what-if" scenarios internally.

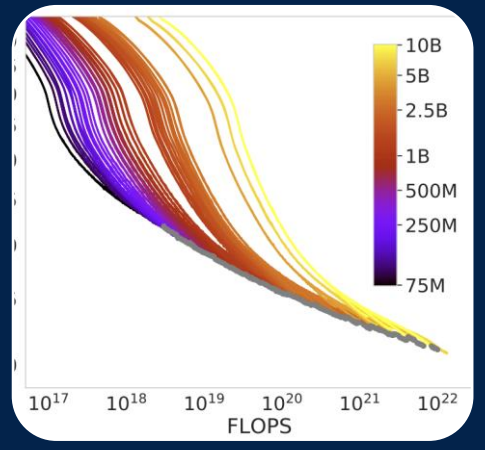
Key trends in 2026



Challenging the Scaling Laws: moving to "Small-Scale"

While the "scaling laws" for VLA models are being tested, there is a strong trend toward **lightweight efficiency**. Models like **Evo-1** (0.77B parameters) are achieving state-of-the-art results (94.8% on LIBERO) with significantly fewer parameters than larger foundations.

Key trends in 2026

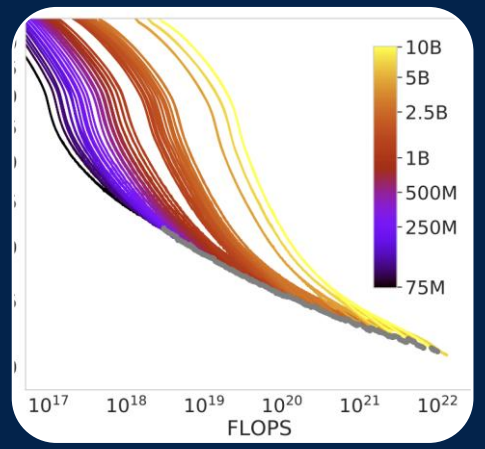


Challenging the Scaling Laws: moving to "Small-Scale"

While the "scaling laws" for VLA models are being tested, there is a strong trend toward **lightweight efficiency**. Models like **Evo-1** (0.77B parameters) are achieving state-of-the-art results (94.8% on LIBERO) with significantly fewer parameters than larger foundations.

Feature	Std Transformer	MSFT YOCO	Improvement
KV Cache Memory	Stored for every layer	Stored only once globally	~80x reduction (65B model)
Prefilling Speed	Slow	Ultra-fast	Up to 71x faster
Throughput	Low (small batch sizes)	High (large batch sizes)	~9x higher efficiency
Context Length	Extremely memory-heavy	Scalable and lightweight	Supports 1M+ tokens easily
Architecture	Std Attention, Encoder-decoder	Decoder-Decoder Design	Layer-independent cache
Energy consumption	1	4 to 20 x reduction	In memory computing friendly

Key trends in 2026



Challenging the Scaling Laws: moving to "Small-Scale"

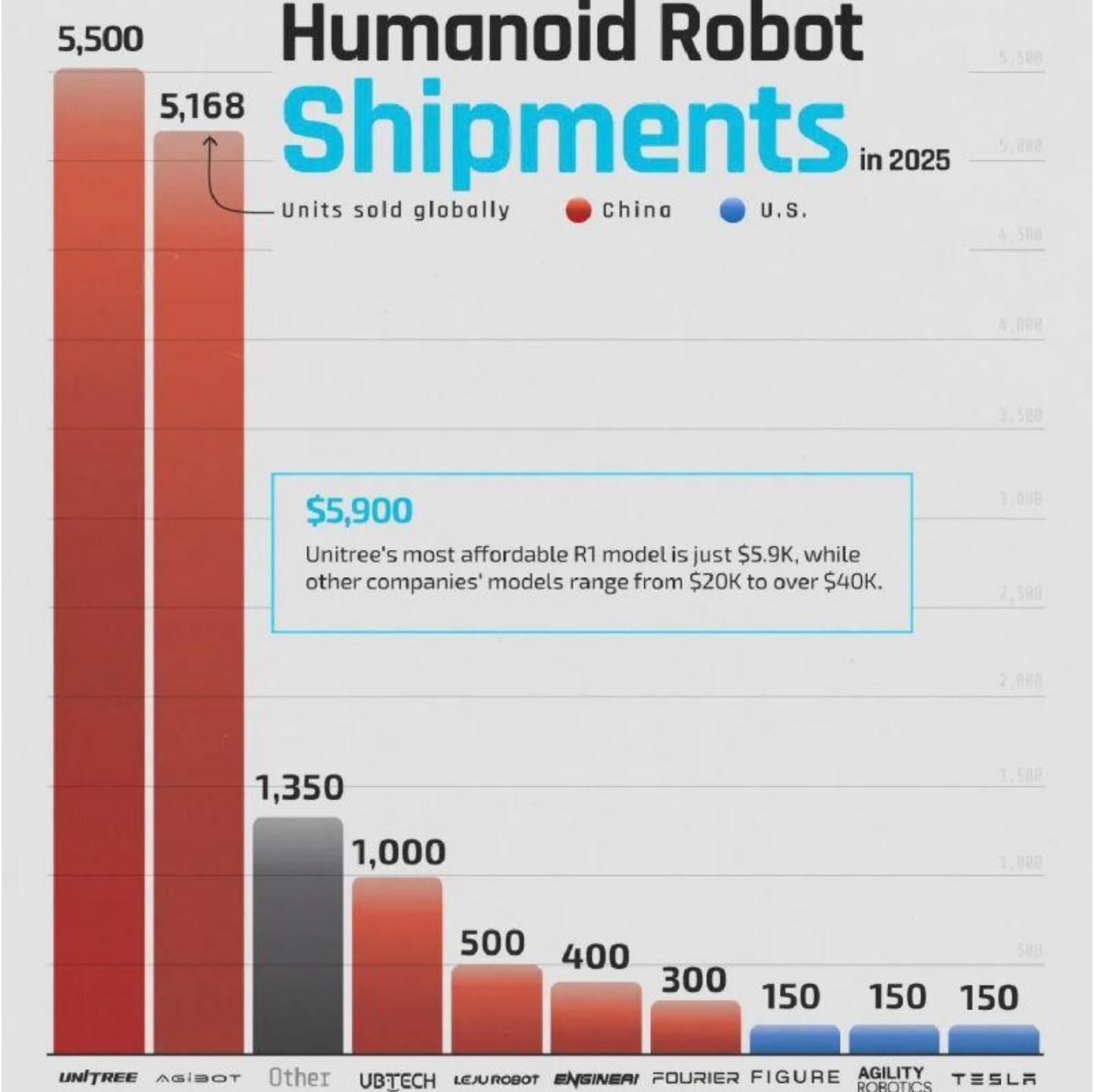
While the "scaling laws" for VLA models are being tested, there is a strong trend toward **lightweight efficiency**. Models like **Evo-1** (0.77B parameters) are achieving state-of-the-art results (94.8% on LIBERO) with significantly fewer parameters than larger foundations.

Generalization Across Robot Morphologies

. **"One Brain, Any Robot":** The industry is pushing toward generalist policies that can be deployed across different embodiments (e.g., humanoids, mobile manipulators, and robotic arms) with minimal fine-tuning.

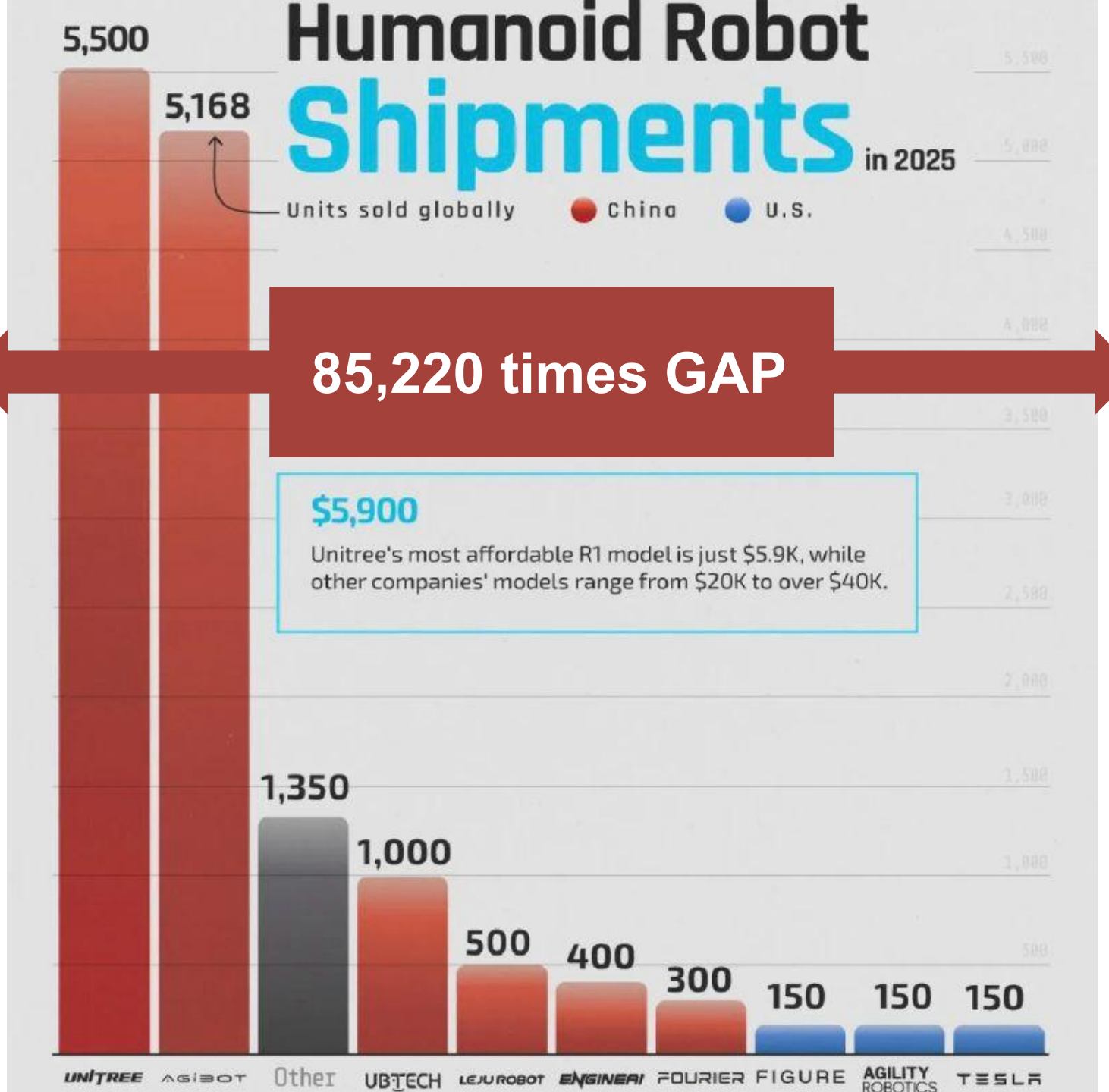
. **Commercial Rollouts:** Companies like **XPENG** are beginning to integrate VLA 2.0 into humanoid robots and autonomous vehicles for mass production scheduled for late 2026.

14,668 units sold in 2025



1.25B phones sold in 2025

14,668 units sold in 2025



85,220 times GAP

1.25B phones sold in 2025

Our technology starts with You

Find out more at www.st.com