UCL ROBOTICS SOCIETY

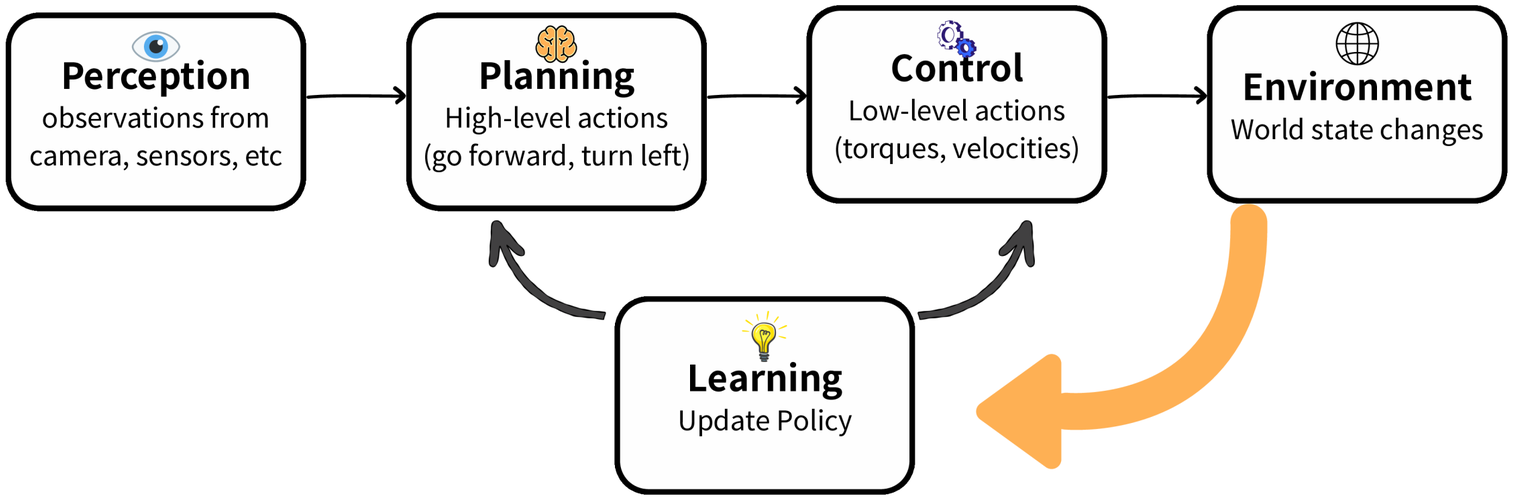
VLA WORKSHOP

Today’s Roadmap

- Intro & Motivation
- Vision, Language & Action
- How VLA Works
- Use Cases, Challenges & Future

Important Components in Robotic System

- **Perception:** Processes sensory data (vision, LiDAR, audio) to create a meaningful representation of the environment.
- **Planning:** Generates a strategy or sequence of actions to achieve a goal
- **Control:** Translates high-level plans into motor commands for physical actuation with stability and precision.
- **Learning:** Adapts behavior from data using supervised, reinforcement, or self-supervised methods.



Understanding observation, state, action, policy

Observations (o_t)

What the Robot "sees"

- **Definition:** The data that the robot collects from sensors (often noisy and partial).
- **Examples:** Camera images, LiDAR point clouds, joint encoders
- **Key:** Observation ≠full state. It provides an incomplete view of reality.

Actions (a_t)

What the Robot "does"

- **Definition:** The control output sent to the robot actuators to influence the environment.
- **Examples:** turn left, turn right.
- **Key:** Each action updates the state: $s_{t+1} = f(s_t, a_t)$. It is how the robot makes a difference.

State (s_t)

The System state

- **Definition:** It is the description of the environment and the robot itself.
- **Examples:** Positions, velocities, etc.
- **Analogy:** Driving in fog: what you see (observation) is partial, but the road layout (state) still exists fully.

Policy (π)

The Brain of Robot

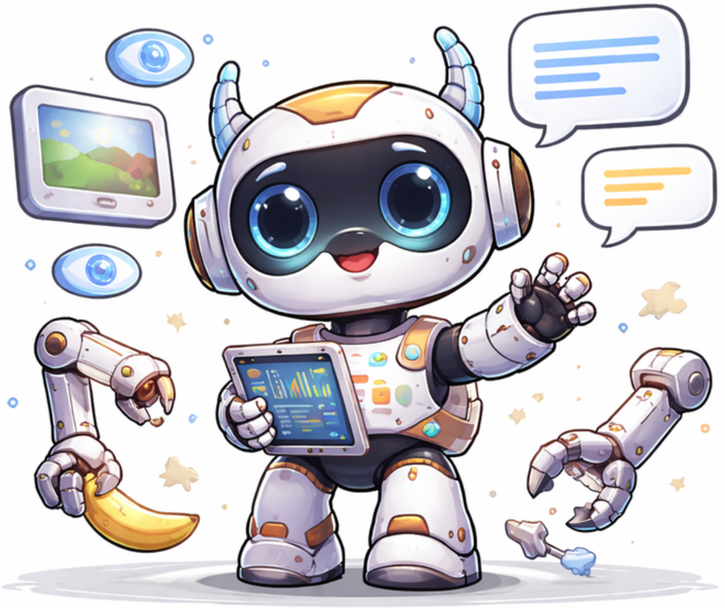
- **Definition:** Policy maps observations or states to actions: $\pi(a_t \mid o_t)$ for Stochastic Policy and a = π(s) for Deterministic Policy.
- **Examples:** Rule-based Controllers, a Neural Network, a RL agent, or a VLA model.
- **Learning the Policy:** Robots can learn the best actions through data.

Introduction to Vision-Language-Action (VLA) in Robotics

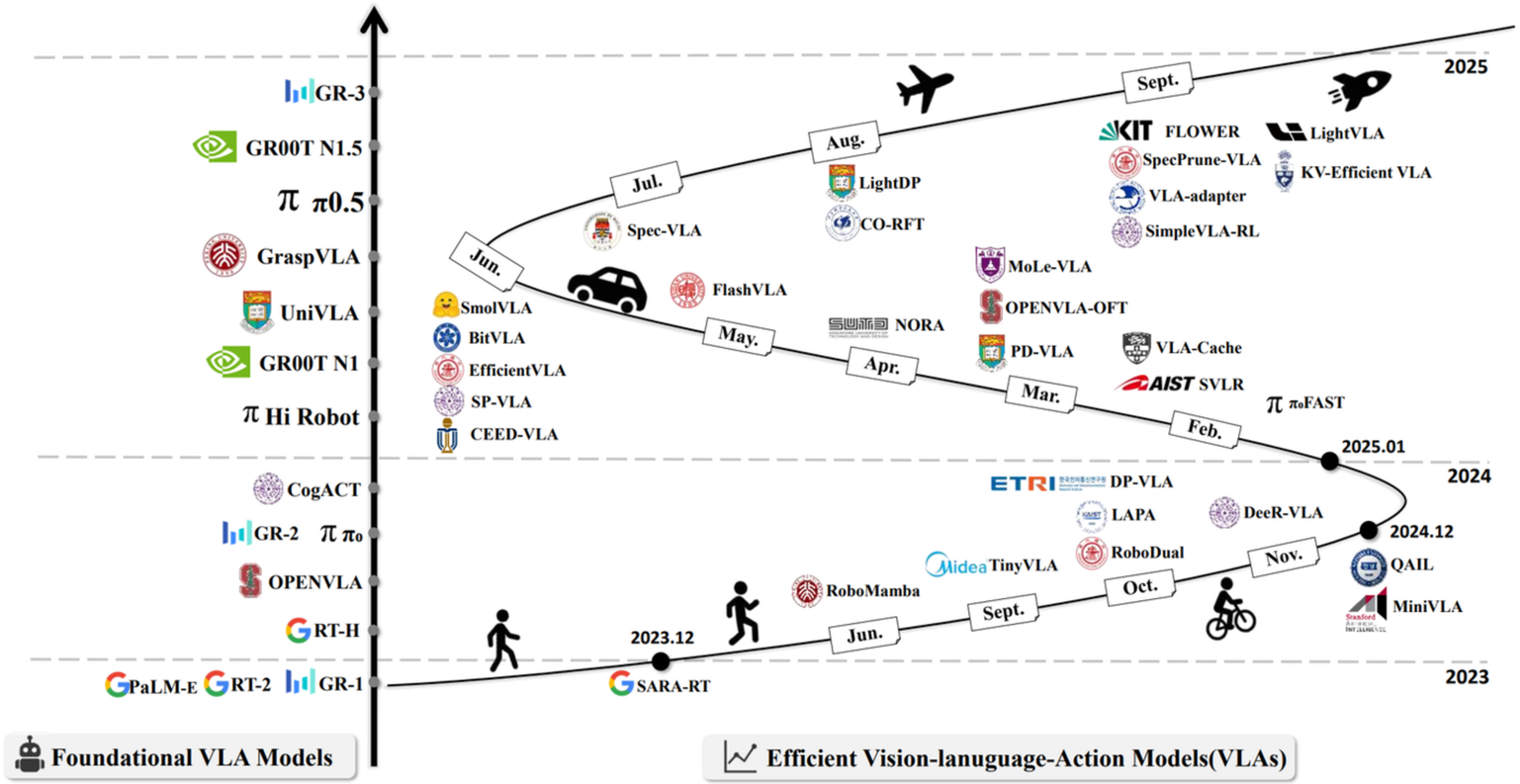
A New Frontier in Intelligent Machine Interaction

- **What is VLA?:** VLA models integrate visual perception, language understanding, and physical action into a unified framework. It enables robots to perceive, reason, and act in complex environments.

- **Why it Matters Now?:** VLA is the bridge from language intelligence to embodied intelligence.



Development of VLA Systems



reference paper: Yu, Z., Wang, B., Zeng, P., Zhang, H., Zhang, J., Gao, L., Song, J., Sebe, N. and Shen, H.T., 2025. A survey on efficient vision-language-action models. arXiv preprint arXiv:2510.24795.

The Three Pillars of VLA: Vision, Language, Action

Vision: Seeing the World

Robots use computer vision to recognize objects, understand scenes, and interpret spatial relationships, forming the foundation for perception.

Language: Understand Instructions

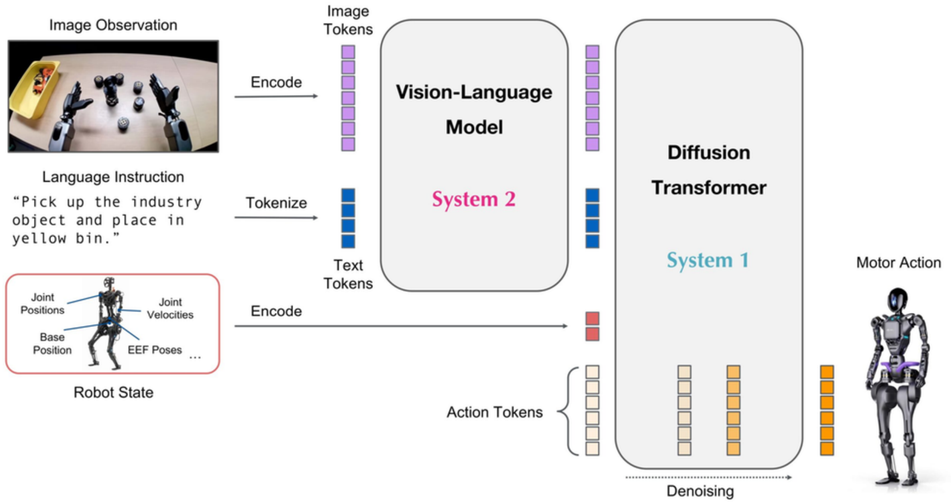
Language models allow robots to process natural language commands, ask questions, and link abstract tasks to grounded actions.

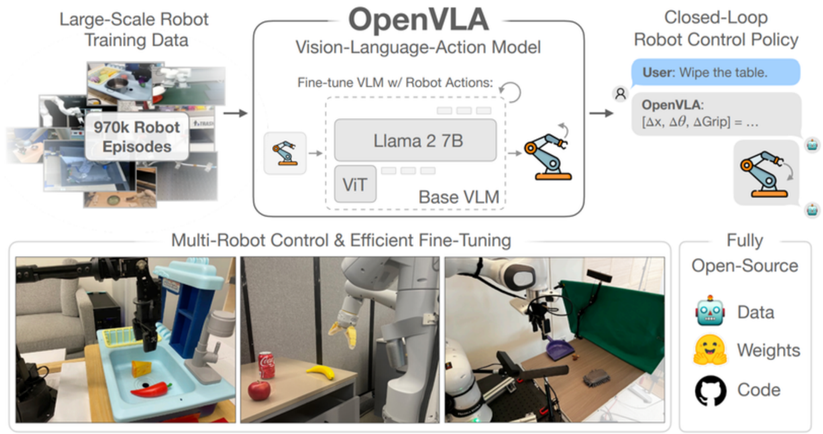
Action: Executing Tasks

Actuation systems transform interpreted data into motion, whether it's navigating, grasping, or manipulating objects.

Architecture of a VLA Model

- **Vision Encoder:** Processes camera images into visual tokens (z_v), usually uses pretrained Vision Transformers (ViT).
- **Language Encoder:** Processes natural language instructions into text tokens (z_l), usually uses LLM, such as OpenVLA uses Llama2
- **Vision Language Model:** Combines z_v and z_l into a shared representation.
- **Action Model:** Outputs actions such as joint movement or gripper control. Action Model can be autoregressive transformer, diffusion policy, etc.



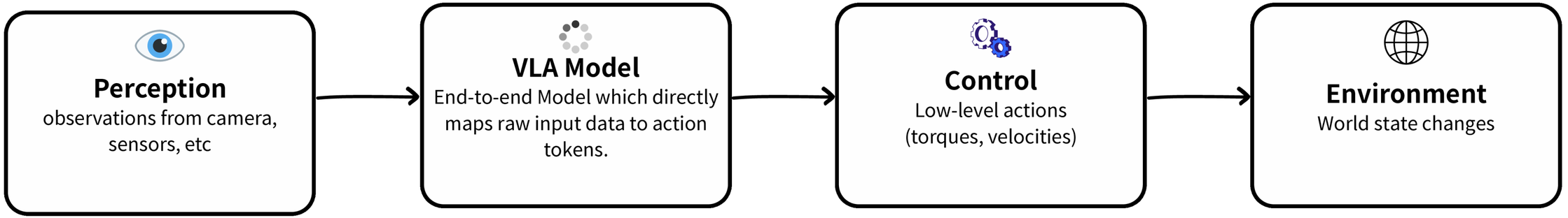


Paper Reference:

[1] NVIDIA, Johan Bjorck, Fernando Castañeda, N., Xingye Da, Runyu Ding, Linxi "Jim" Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, & Yuke Zhu (2025). GR00T N1: An Open Foundation Model for Generalist Humanoid Robots. In ArXiv Preprint.

[2] Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P. and Vuong, Q., 2024. Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246.

How VLA works



Use Cases of VLA in Robotics

Figure AI - Helix

reference link: https://www.youtube.com/watch?v=8gfuUzDn4Q8

Use Cases of VLA in Robotics

Figure AI - Helix

reference link: https://www.youtube.com/watch?v=8gfuUzDn4Q8

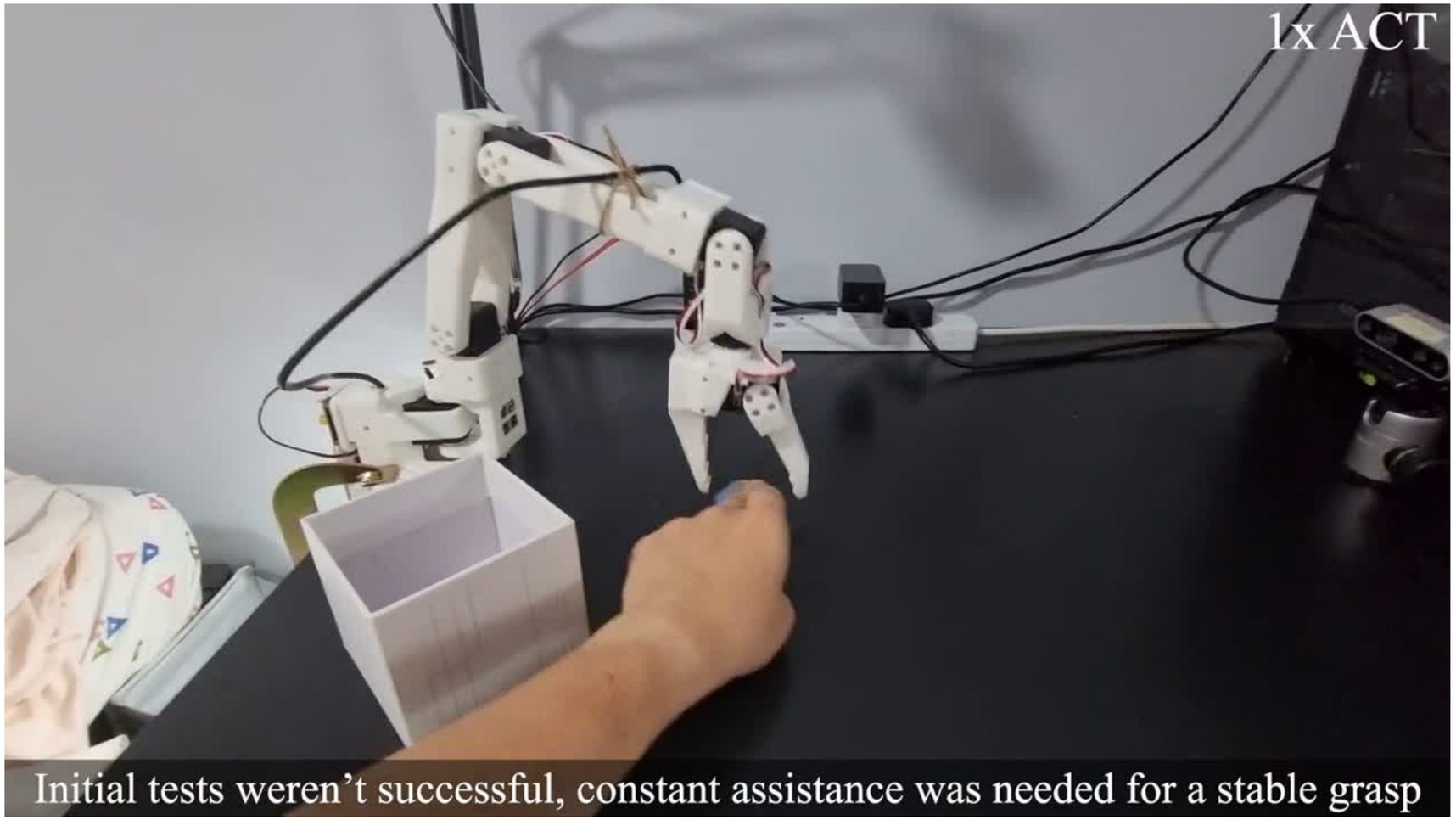
Use Cases of VLA in Robotics

LIMX Dynamics - LimX COSA

reference link: https://www.youtube.com/watch?v=0hIqs3TBb5g

What we will be doing for VLA in Robotics

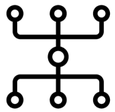
LeRobot - SO101 Arm



reference link: https://www.youtube.com/watch?v=nWNIsJBwbvU

Training Data & Real-World Challenges

Sim2Real Gap:
Robots trained in simulation often struggle in the real world due to noise, variability, and unexpected edge cases.



Multimodal Data Alignment:
Aligning vision, language, and action data is complex and prone to inconsistencies, leading to misinterpretations

Robustness & Generalization:
VLA systems can overfit to training environments and fail to generalize well to novel settings or unseen objects.

Limitations and Bottlenecks in VLA

What is Holding Back the Next Wave of Robotics

Hallucinations:
VLA systems can 'hallucinate', generating inaccurate actions or actions that can not be achieved in the environment.

Overconfidence:
Models may act with high certainty even when unsure, posing safety risks in dynamic environments.

Future Trends in VLA Robotics

Where is This All Heading?

- **Unified Agents:** Future systems aim to handle multiple tasks and environments seamlessly—blurring the line between specialized and general-purpose robots.

- **Human-Robot Collaboration:** Natural language interfaces and shared situational awareness will boost teamwork between humans and intelligent robots

Wrap-Up & Key Takeaways

What You Should Now Know About VLA in Robotics

What VLA is:

Vision-Language-Action models integrate perception, understanding, and behavior into unified, intelligent agents.

Why It Matters:

VLA is redefining how robots interact with the world, enabling flexible, human-aligned, and context-aware autonomy.

What is next:

Challenges remain in generalization, safety, and ethics, but the trajectory is clear: smarter, more collaborative robotics.