清華大學 智能产业研究院
Institute for AI Industry Research, Tsinghua University

Vision-Language-Action (VLA) Models

A Review of Recent Progress

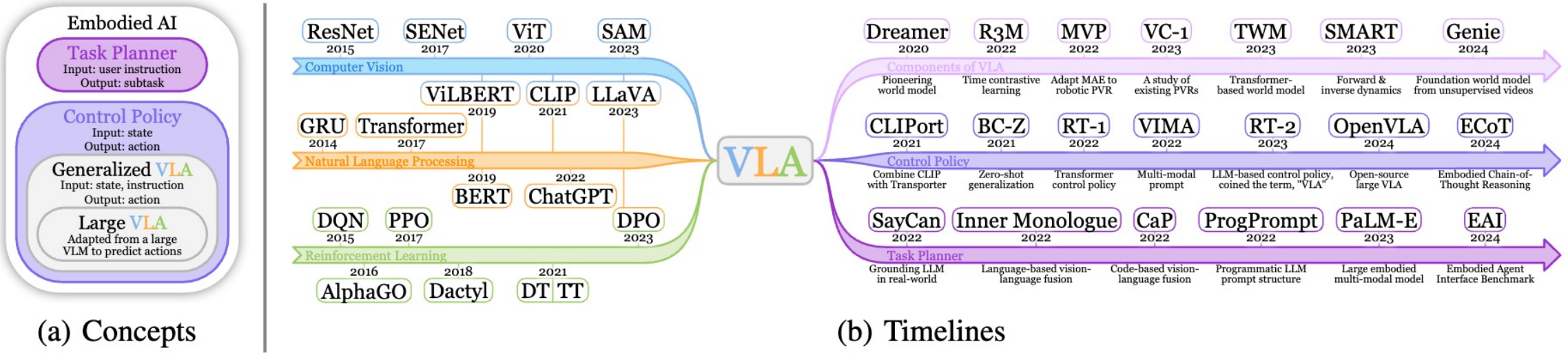
Xiangyu Li
Institute for AI Industry Research (AIR), Tsinghua University
lixiangy22@mails.tsinghua.edu.cn

Outline

- **Background & Overview**
- **Descrete VLA**
- **Continuous VLA**
- **Dual-System VLA**

What is a Vision-Language-Action (VLA) Model?

- A **multi-modal foundation model** for embodied AI.
  - Input modalities: **vision** (observation) & **language** (instruction)
  - Output modalities: **action** (low-level robot control policy)
- A VLA model utilizes a **VLM** for VL-conditioned action generation.

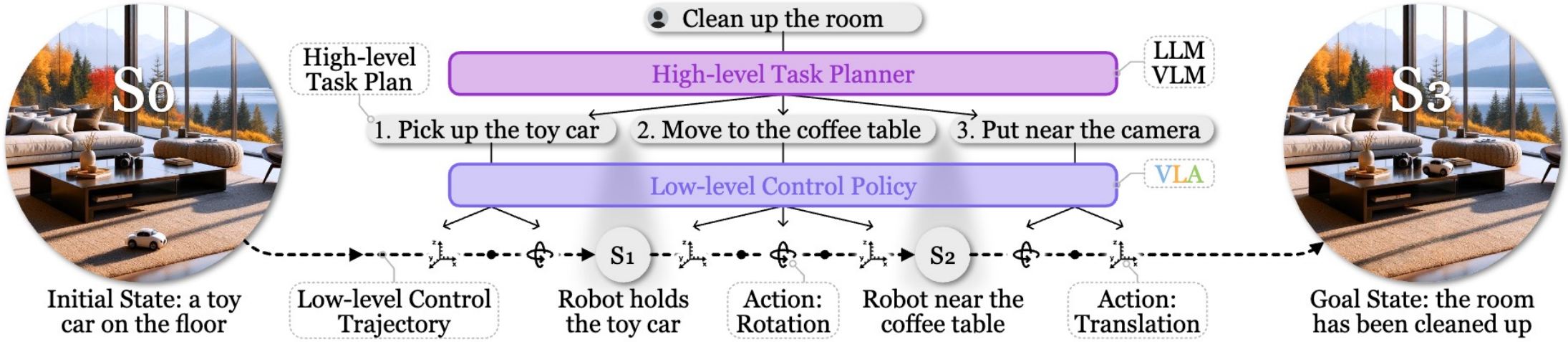


The concepts (a) and timelines (b) related to the development of VLA.

[1] Ma, Yueen, et al. "A Survey on Vision-Language-Action Models for Embodied AI." *arXiv preprint arXiv:2405.14093* (2024).

VLA + Task Planner for Long-Horizon Tasks

- VLA is initially optimized for **low-level robot control policy**.
- To complete complex, long-horizon tasks, an effective approach is to add a **LLM/VLM-based task planner** to decompose them to simple subtasks.
  - Earlier works usually adopt a separate model as the task planner.
  - Recent works utilize a shared VLM for both task planning and control policy (i.e., *dual-system*).

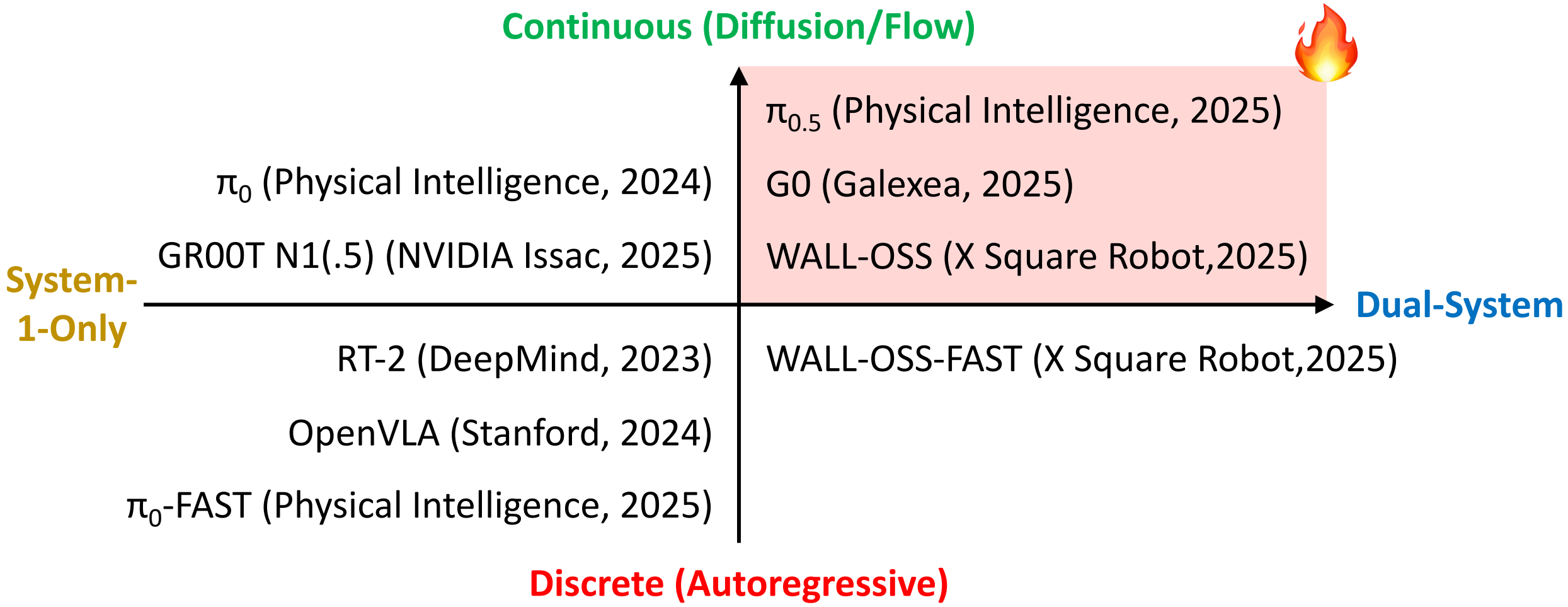


An example of a hierarchical robot policy (high-level planning + low-level control).

[1] Ma, Yueen, et al. "A Survey on Vision-Language-Action Models for Embodied AI." *arXiv preprint arXiv:2405.14093* (2024).

Recent Progress of VLA: an Overview

Trend: dual-system (planning + control) with continuous action generation.



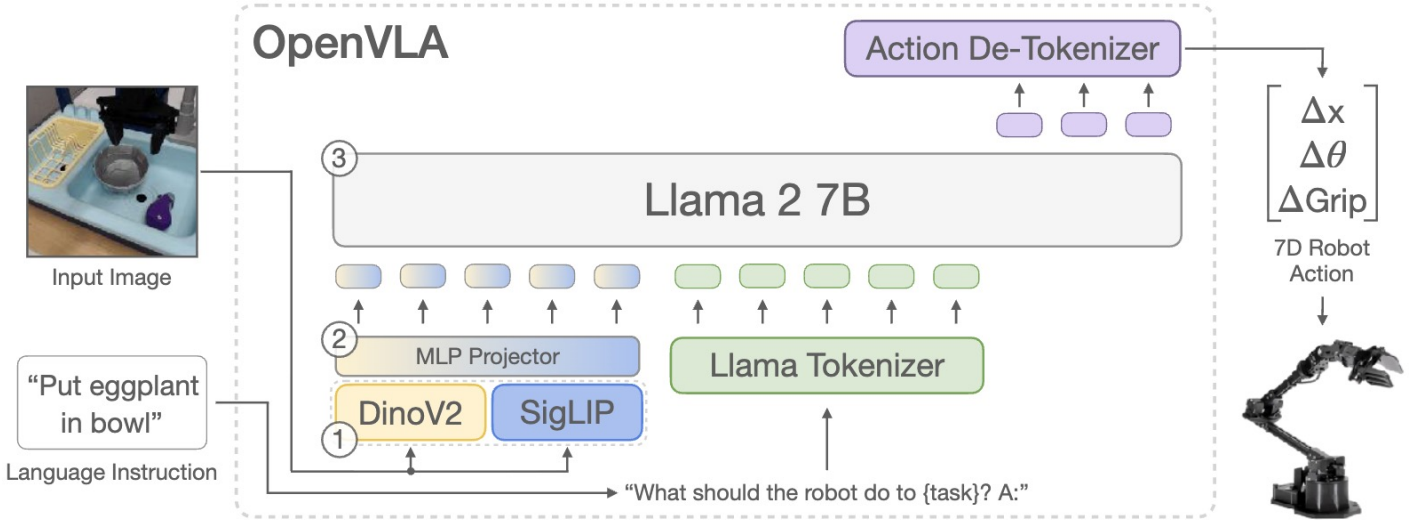
Outline

- Background & Overview
- **Descrete VLA**
- Continuous VLA
- Dual-System VLA

Discrete (Autoregressive) Action Generation

- Train the VLM to generate **discrete tokens** that directly map to actions.
  - **Pros:** friendly for autoregressive VLM to learn and generate.
  - **Cons:** high latency and low FPS, not suitable for high frequency control.
- Representative methods:
  - **RT-2** (ViT + PALI-X/PALM-E): a pioneer work that proposes and popularizes the term “VLA”.
  - **OpenVLA** (DinoV2 & SigLIP + Llama2): an influential open-source VLA model (3.8k stars).
  - **FAST**: an action tokenizer that compresses action sequences with DCT.

Example: OpenVLA



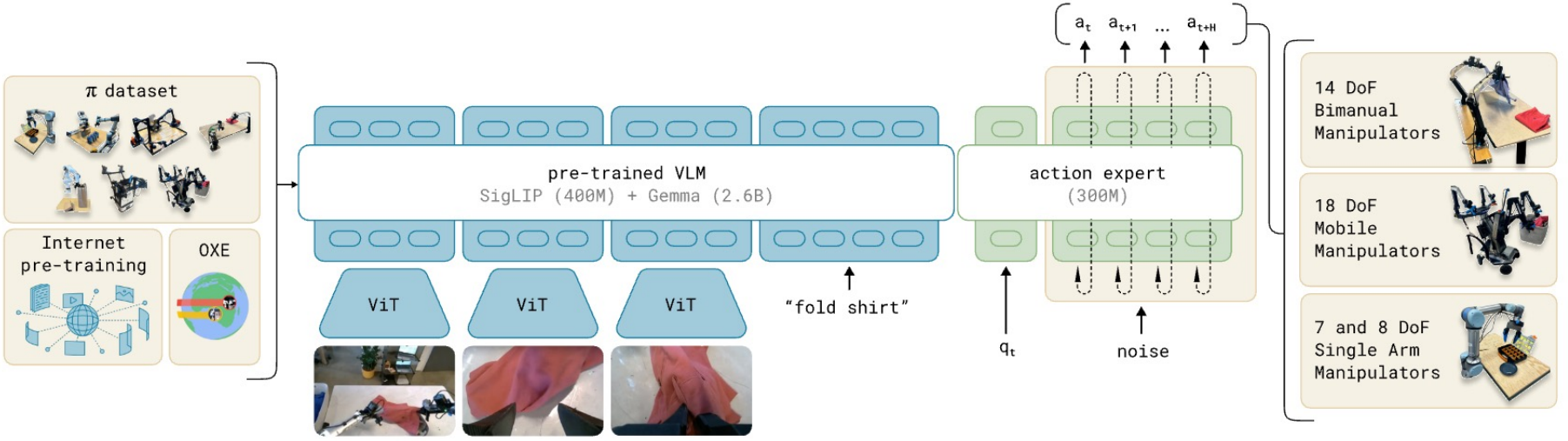
[1] Kim, Moo Jin, et al. "OpenVLA: An Open-Source Vision-Language-Action Model." *arXiv preprint arXiv:2406.09246* (2024).

Outline

- Background & Overview
- Descrete VLA
- **Continuous VLA**
- Dual-System VLA

π_0 | Physical Intelligence (π) | Oct. 2024 (Opened in Feb. 2025)

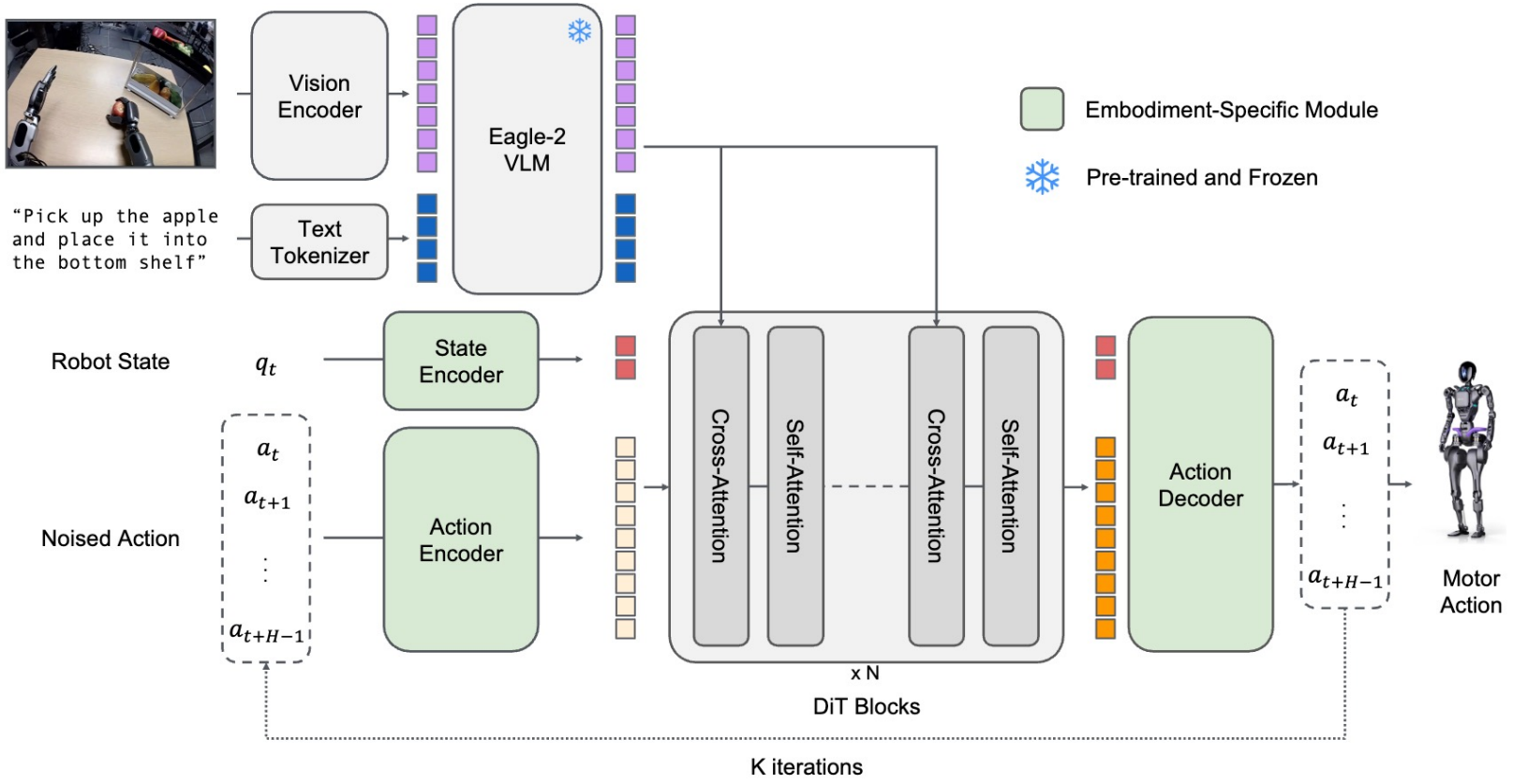
- The first to integrate a **flow-matching action expert** to a **pre-trained VLM**.
  - **Pretrained VLM:** Internet-scale semantic understanding.
  - **Flow-matching action expert:** high-frequency (up to 50Hz) control policy.
- Training recipe matters!
  - **VLM pre-training:** Internet-scale dataset.
  - **VLA pre-training:** Open X Embodiment dataset + π Cross-Embodiment Robot dataset.
  - Post-training (optional): high-quality post-training data (difficult/unseen tasks).



[1] Black, Kevin, et al. "$\pi_0 $: A Vision-Language-Action Flow Model for General Robot Control." *arXiv preprint arXiv:2410.24164* (2024).

GR00T N1(.5) | NVIDIA Isaac | Mar. 2025 (N1.5 in Jun. 2025)

- The first open foundation model for generalist humanoid robots.
  - Similar to π_0 in model architecture and training method.
  - A difference: <u>only passes hidden states from a certain VLM layer (π is each layer).</u>



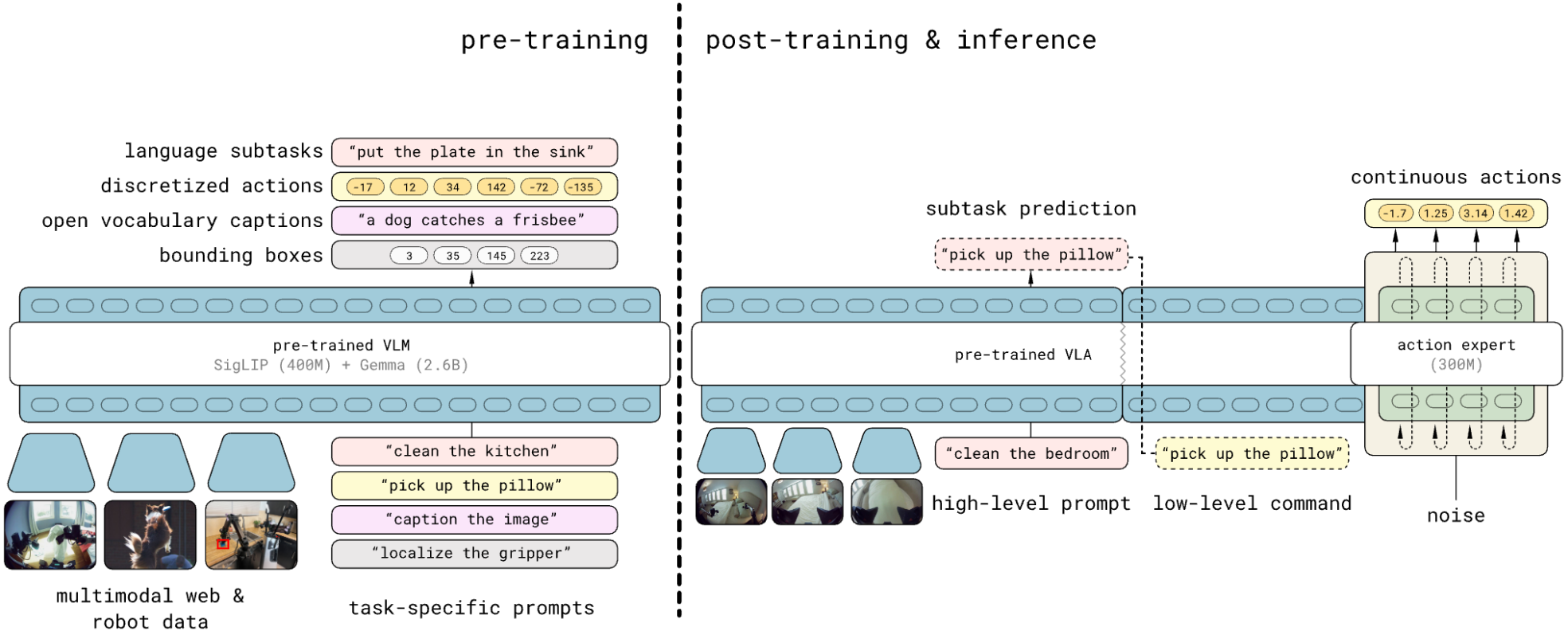
[1] Bjorck, Johan, et al. "GR00T N1: An Open Foundation Model For Generalist Humanoid Robots." *arXiv preprint arXiv:2503.14734* (2025).

Outline

- Background & Overview
- Descrete VLA
- Continuous VLA
- **Dual-System VLA**

$\pi_{0.5}$ | Physical Intelligence (π) | Apr. 2025 (Opened in Sep. 2025)

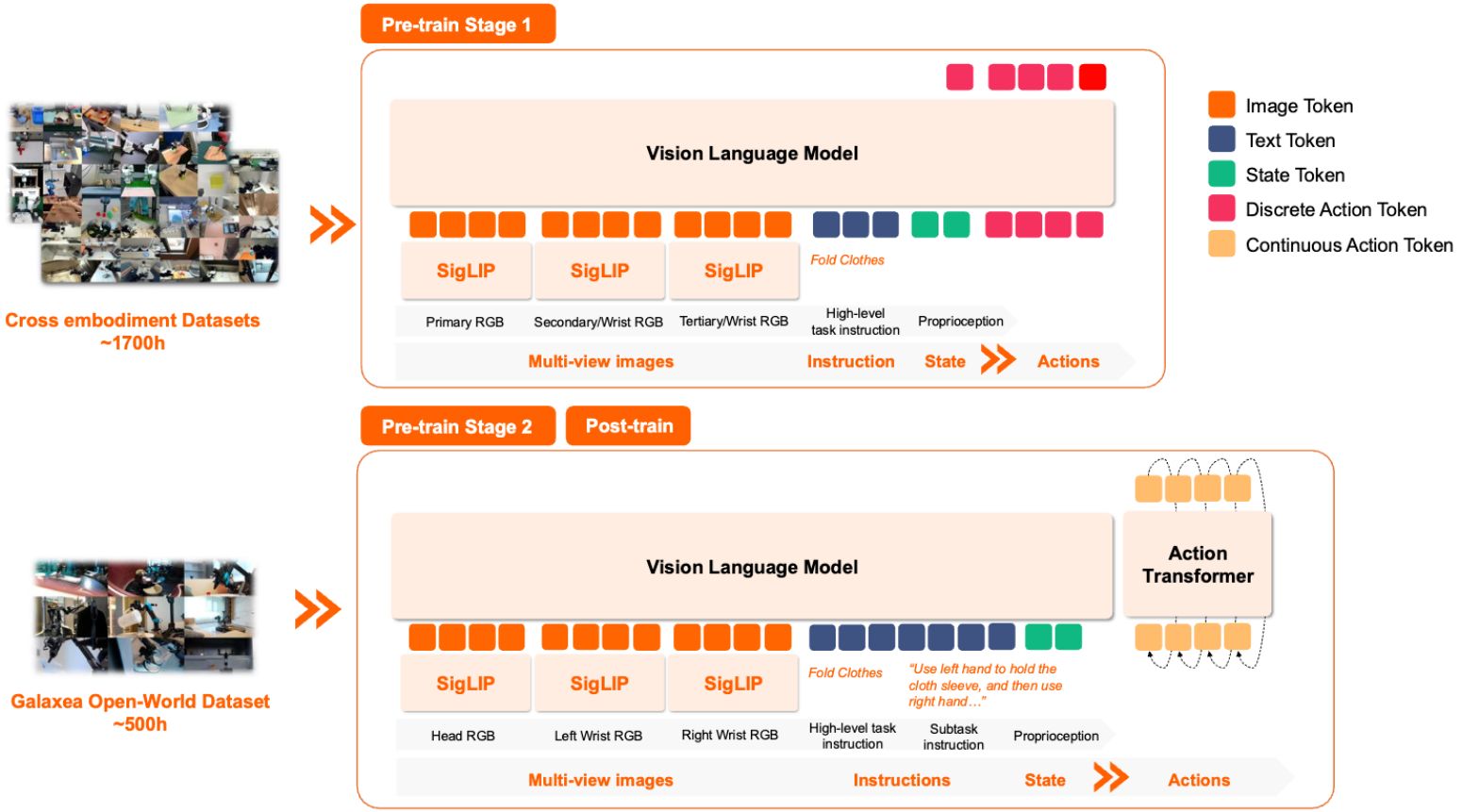
- What's new, compared to π_0? — for **open-world generalization**.
  - **New training data**: object detection, instructions & subtask commands, etc.
  - **New inference flow**: subtask prediction with the same model (earlier works use 2 separate models).
    - AKA the *dual-system* design: **system 2 for high-level planning** and **system 1 for low-level control policy.**



[1] Intelligence, Physical, et al. "$\pi_ {0.5} $: a Vision-Language-Action Model with Open-World Generalization." *arXiv preprint arXiv:2504.16054* (2025).

G0 | Galaxea 星海图 | Aug. 2025 (Opened in Sep. 2025)

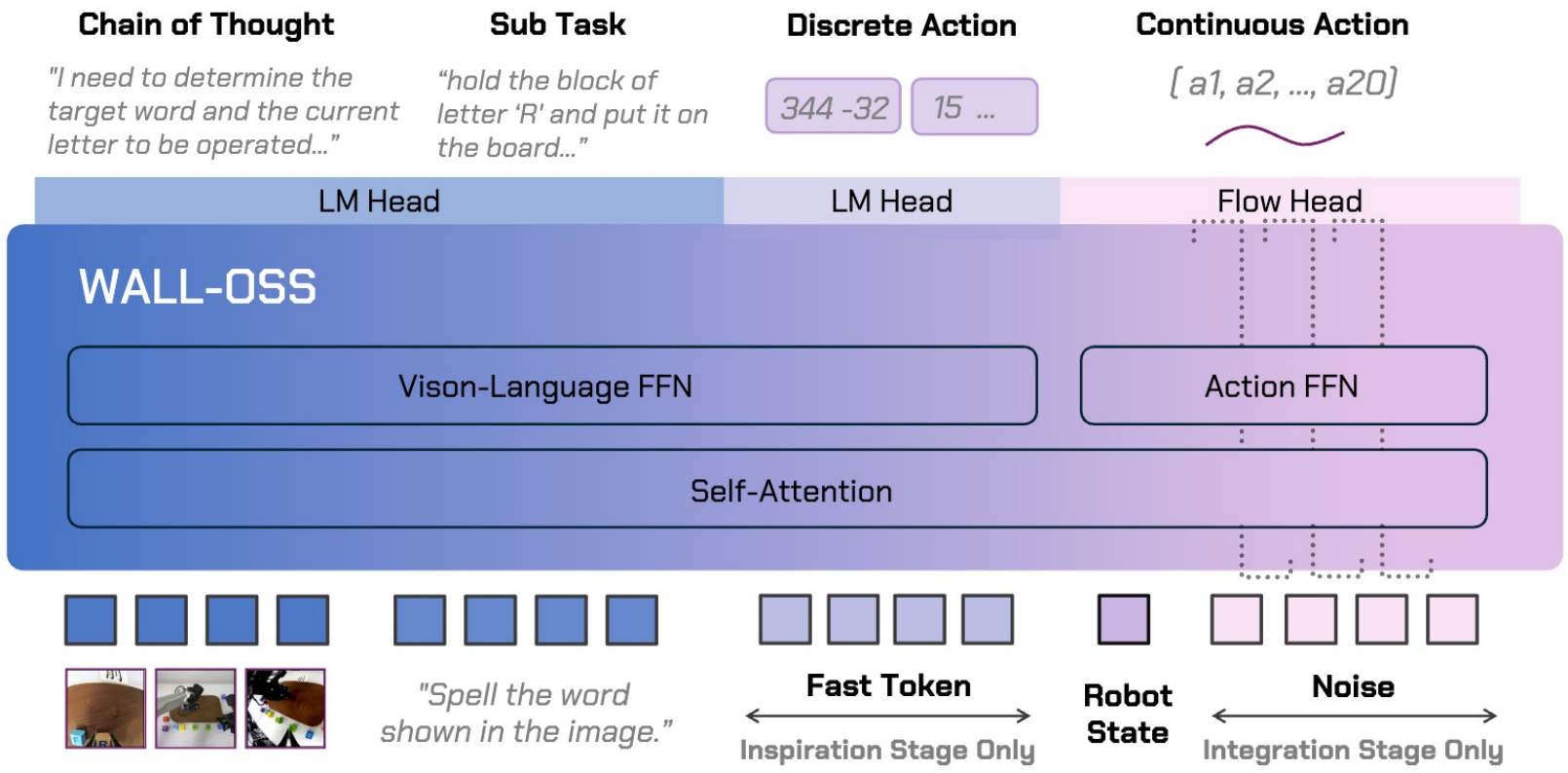
- **Similar to $\pi_{0.5}$** in model architecture and training method.
  - Single foundation model for subtask prediction, continous action generation, etc.
  - Only compared with π_0 in paper. Model weights opened on the same day as $\pi_{0.5}$.



[1] Jiang, Tao, et al. "Galaxea Open-World Dataset and G0 Dual-System VLA Model." *arXiv preprint arXiv:2509.00576* (2025).

WALL-OSS | X Square Robot 自变量 | Sep. 2025 (Opened in Sep. 2025)

- **Similar to $\pi_{0.5}$**, too.
  - Only compared with π_0, too — weights opened right before G0 and $\pi_{0.5}$.



[1] X Square Robot. "WALL-OSS: Igniting VLMs toward the Embodied Space." 2025, https://x2robot.cn-wlcb.ufileos.com/wall_oss.pdf. White paper.

Summary & Future Look

- VLAs evolve from **discrete** to **continuous**, from **single-system** to **dual-system**.
  - Only within **3 years** (from RT-2 in 2023 to G0/WALL-OSS in 2025)!!
  - The future 3 years?
- Another trend (from my perspective): **multi-tasking**.
  - **Model-side:** SOTA VLAs are built upon Internet-scale pre-trained VLMs.
  - **Task-side:** SOTA VLAs are capable of task planning in human language.
  - **A step forward:** all VLM-capable tasks, with one single model (multi-expert).
    - Currently working on — welcome to discuss & collaborate.

Thanks