



VLA 001

Advanced Topics in Embodied Learning and Vision

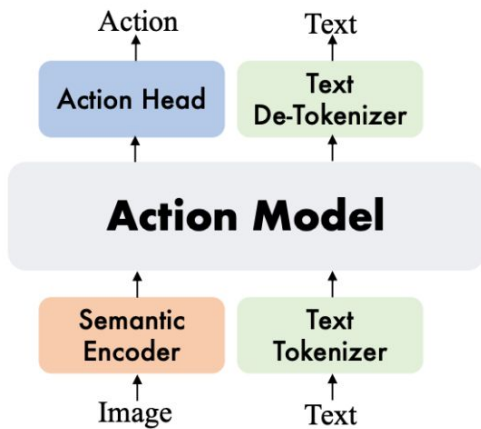
Ying Wang

2026.03.24

Moravec's paradox

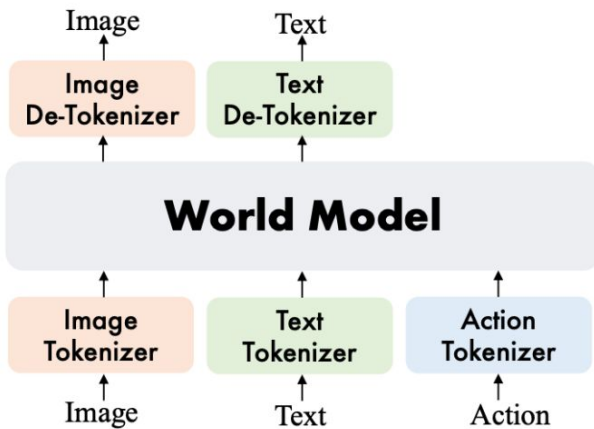
Moravec's paradox ([Hans Moravec](#), 1988): “It is comparatively easy to make computers exhibit adult level performance on intelligence tests or playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility”.

- Adult level performance on intelligence tests or playing checkers → LLMs!
- Skills of a one-year-old when it comes to perception and mobility → robotics! → Can we scale robotic learning in a similar way as LLMs?



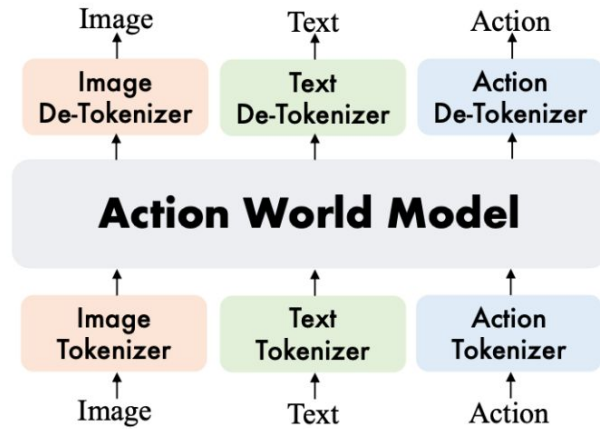
(a) Action Model
(e.g., OpenVLA)

- Image Understanding ✓
- Image Generation ✗
- Action Understanding ✗
- Action Generation ✓



(b) World Model
(e.g., iVideoGPT)

- Image Understanding ✓
- Image Generation ✓
- Action Understanding ✓
- Action Generation ✗



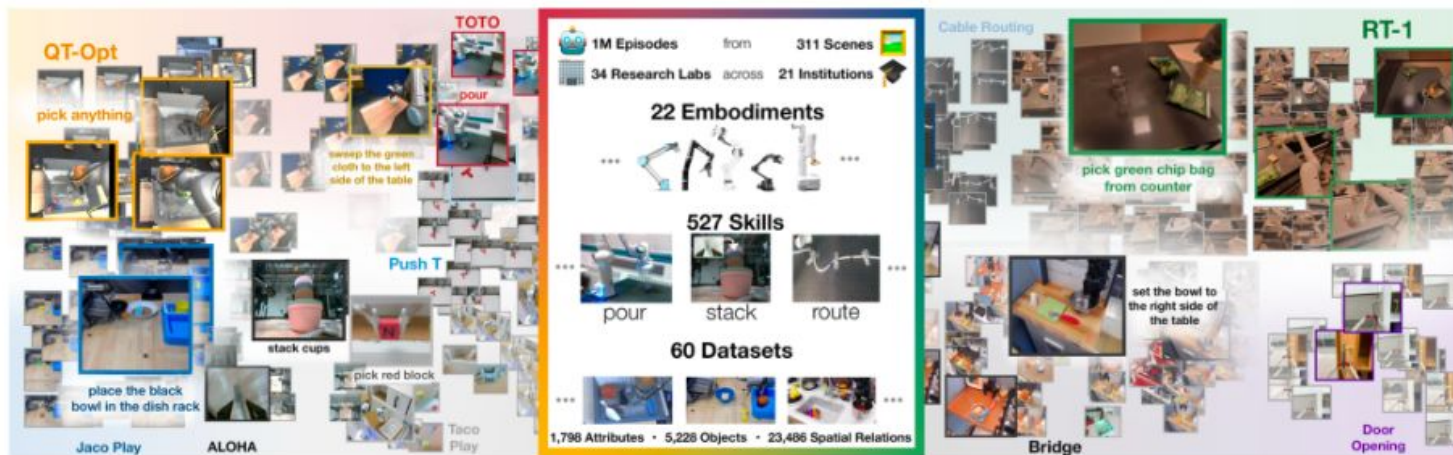
(c) Action World Model
(e.g., WorldVLA)

- Image Understanding ✓
- Image Generation ✓
- Action Understanding ✓
- Action Generation ✓

today!

Fig taken from <https://arxiv.org/pdf/2506.21539>

Data



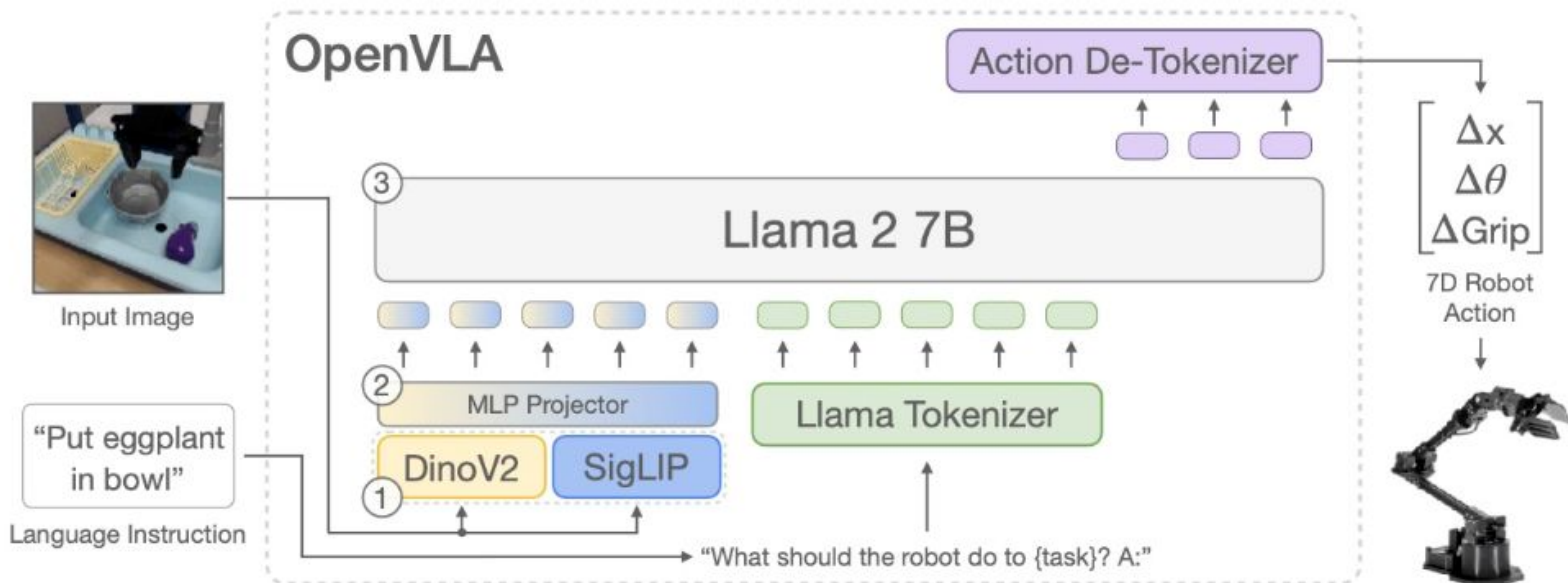
Open X-Embodiment: the largest open-source real robot dataset to date.

1M+ real robot trajectories spanning 22 robot embodiments

<https://robotics-transformer-x.github.io/>

Method

predicts tokenized output actions → get decoded into continuous output actions
✗ high-frequency, precise, or fluent motions

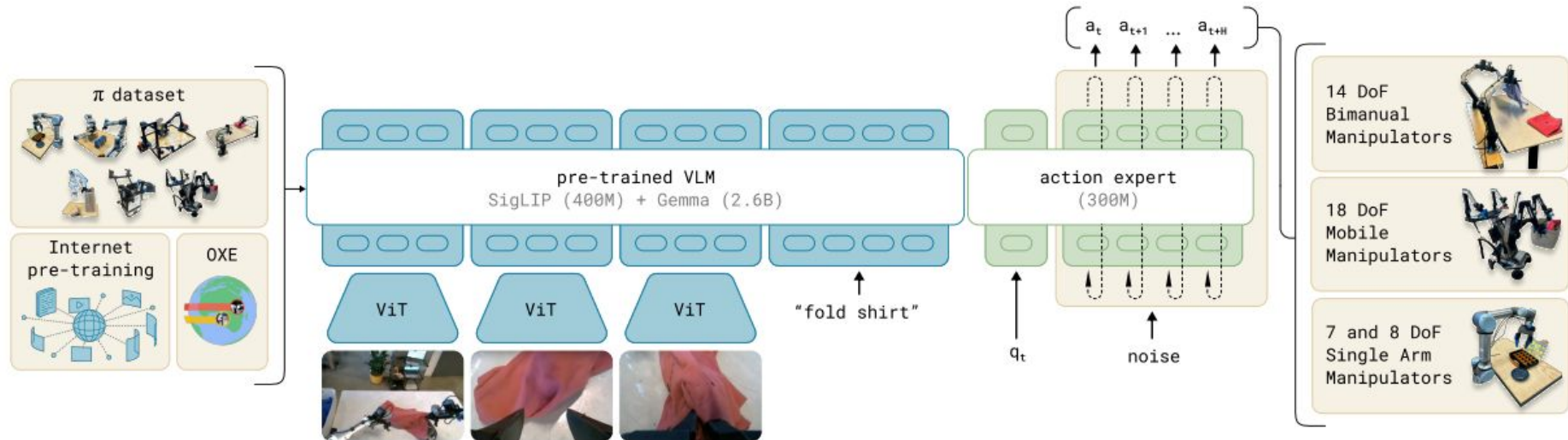


OpenVLA: <https://openvla.github.io/>

RT-2: <https://robotics-transformer2.github.io/>

Method

Generative modeling → continuous actions
✗ slower training; might degrade pretraining knowledge in VLM



pi0: <https://www.pi.website/download/pi0.pdf> → pi0.5
GR00T N1: <https://robotics-transformer2.github.io/>

Challenges

Data & Scaling:

- Heterogeneity of Embodiment, e.g. different action spaces
- Sample Efficiency: Humans learn to use a new tool in 1–2 trials; current VLAs still require thousands of demonstrations
- Sim-to-Real Gap

Generation & Robustness

- Performance degrades significantly under varying lighting, lens flares, or cluttered backgrounds that weren't represented in the clean training demonstrations.

Deployment Constraints

- Inference latency

...

Building a VLA from scratch

[\[colab notebook\]](#) adapted from Prof.Glen Berseth's tutorial in World Model Workshop