VLA, Physical AI의 실현

조준호 (sudoremove )

250814 Physical AI Workshop

TESLA

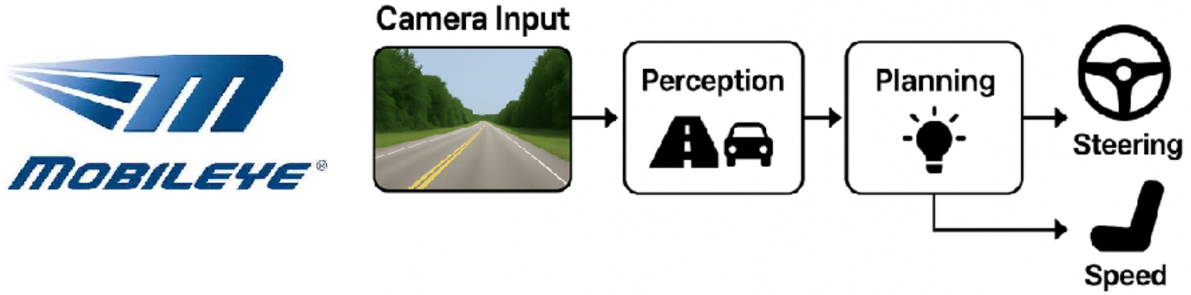
FIGURE

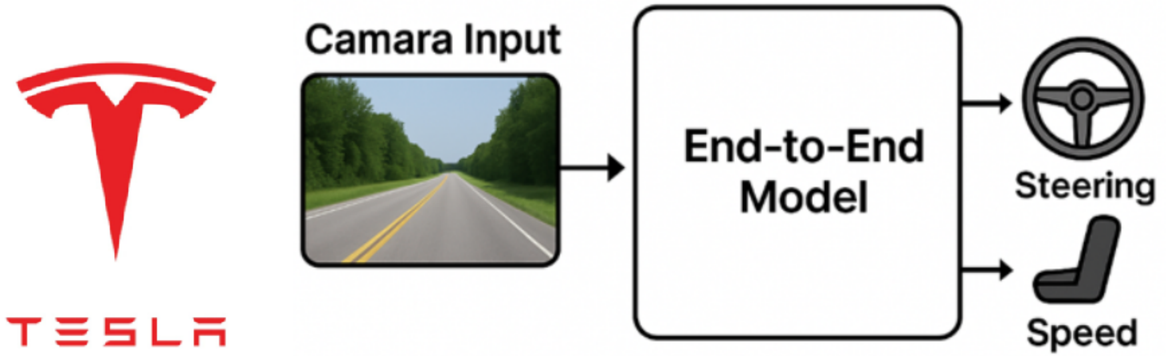
자율주행의 첫번째 레슨

Modular vs **End-to-End**

ex) Mobileye, Tesla

모듈형 접근은 규칙과 단계에 의존하지만,
End-to-End 접근은 복잡하고 비정형적인 환경에서의
해법을 방대한 주행 데이터를 통째로 학습해 찾아낸다



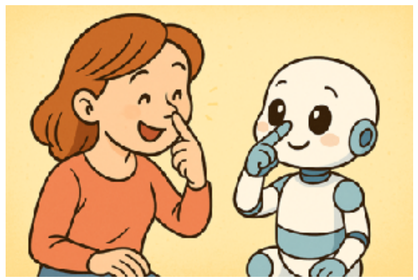


자율주행의 첫번째 레슨

End-to-End 접근은 복잡하고 비정형 환경의 해법

- 예외/돌발 상황이 너무 많아
  Rule 기반으로는 대응 불가

- 사람이 감각기관으로 운전을 한다면,
  기계도 가능하다는 가정



- 데이터 기반 모방 학습
  Imitation Learning

Elon Musk

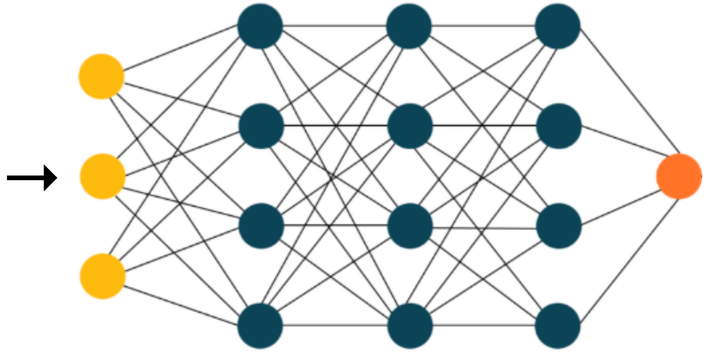
"인간은 눈과 생물학적 신경망으로 운전합니다.
따라서 **자율주행의 일반화된 해법**은
카메라와 실리콘 신경망을 사용하는 것이 합리적입니다."

자율주행의 첫번째 레슨

End-to-End 자율주행 모델의 입력과 출력

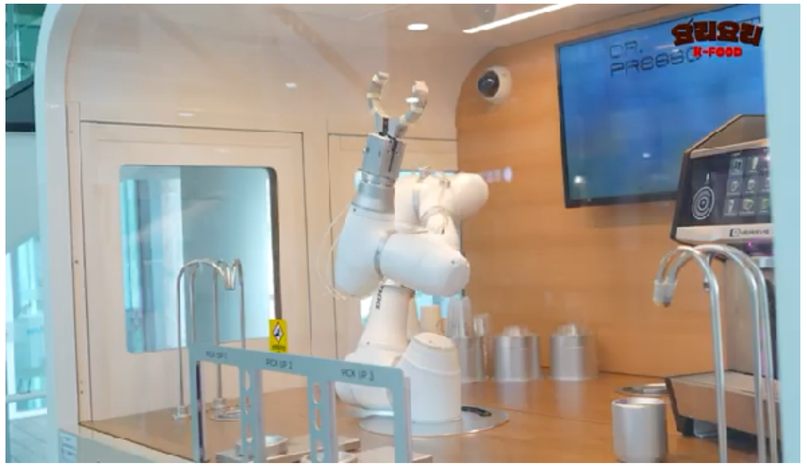
주행 영상 + 현재 위치 + 목적지

자율주행 모델



조향값 + 속도

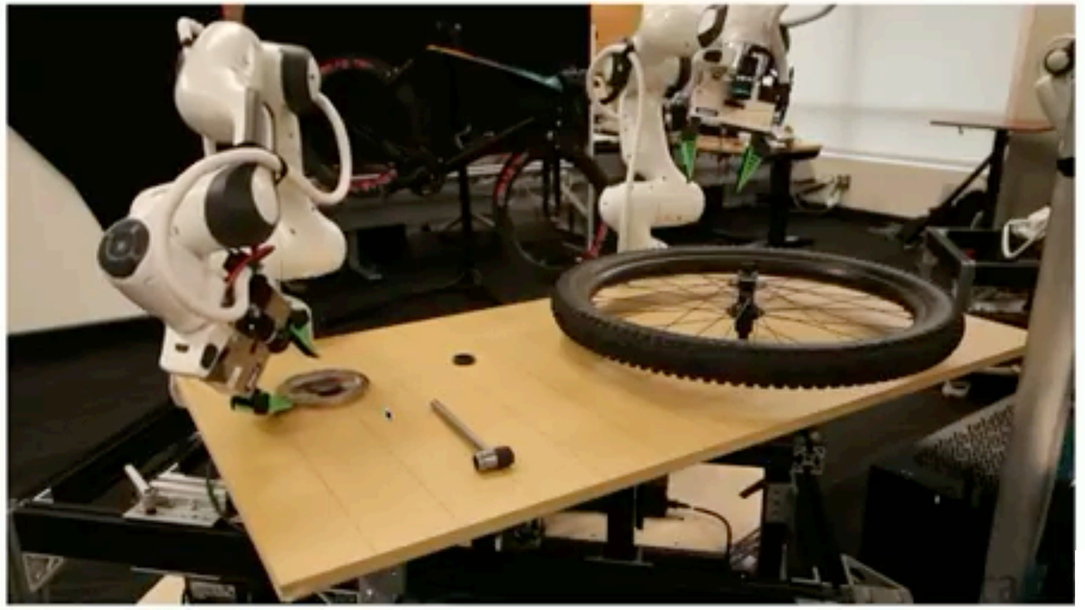
$\Delta v,\ \Delta \delta$



coffee making

정형 Task

비정형 Task

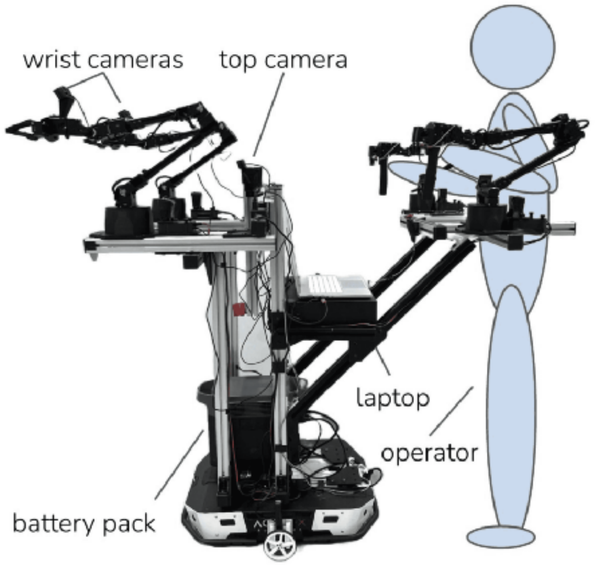


fix bike?

fold laundry

자율주행의 첫번째 레슨

End-to-End 로봇 (robot policy)의 입력과 출력



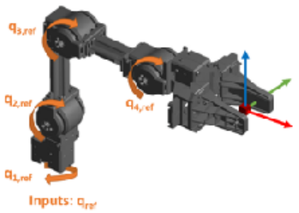
관측 영상 + 지시어 + 로봇 관절값


전방 카메라

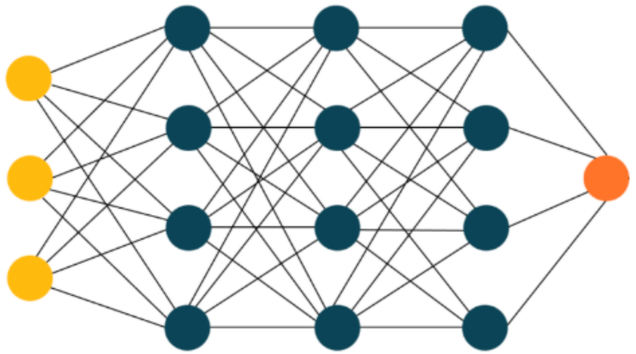

왼쪽 그리퍼 카메라


오른쪽 그리퍼 카메라

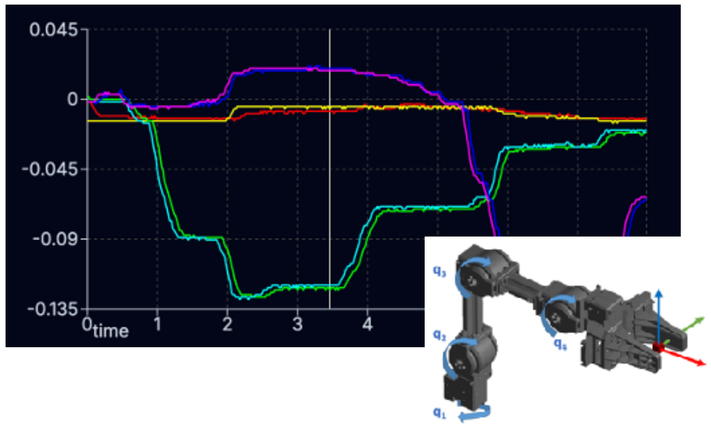
"티셔츠를 접어라"



Robot policy



Robot action



$\Delta x,\ \Delta\theta,\ \Delta grip$

첫번째 레슨

자율주행 e2e → 로봇 e2e

공통점: **복잡하고 비정형 환경**의 해법

용어 교통 정리

- **VLA (Vision Language Action model):**
  Vision-Language 입력으로 Action을 출력하는 모델

  인터넷 규모의 데이터로 학습한 *VLM (Vision-Language Model)* 의 **시각적·언어적 지식**을 활용하면서,
  로봇의 저수준 동작*(low-level actions)*을 예측할 수 있도록 확장한 모델

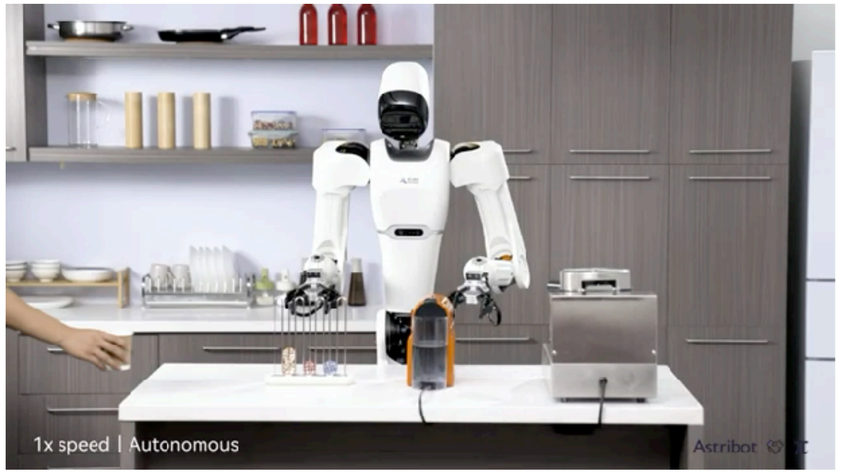
- **RFM (Robot Foundation Model)**

  어떤 로봇이든, 어떤 작업*(task)*이든 수행할 수 있도록
  다양한 환경·센서·action 데이터를 대규모로 사전학습한 범용 로봇 지능 모델.
  VLA의 궁극적인 지향점

- LAM / LBM: VLA로 취급

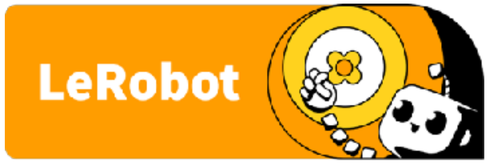
  Large Action Model
  Large Behavior Model

지금 VLA에 대한 기대감의 이유
🔥🔥🔥

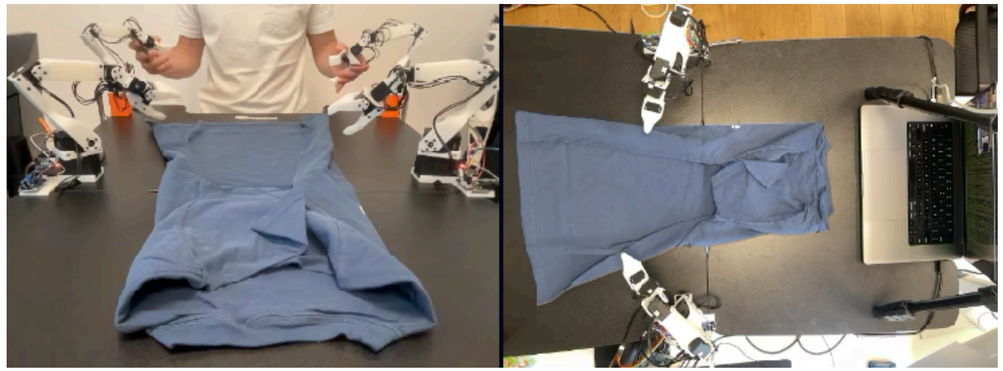


오픈소스된 pi0 모델을 Astribot 로봇에서 캡슐커피 내리는 Task를 파인튜닝

- 휴머노이드, 로봇 비정형 Task데모의 급진적 향상
- 오픈 weight VLA모델 (ACT, OpenVLA, π_0 pizero) 공개로 진입장벽 하락
- 저가형 오픈소스 로봇
  + 텔레옵/학습/구동 플랫폼
  🤗 HuggingFace LeRobot

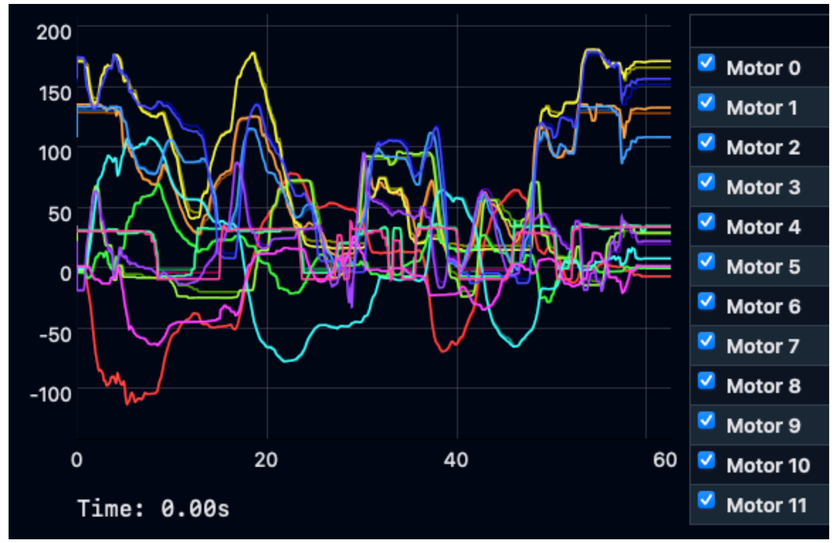


Vision-Language-Action model



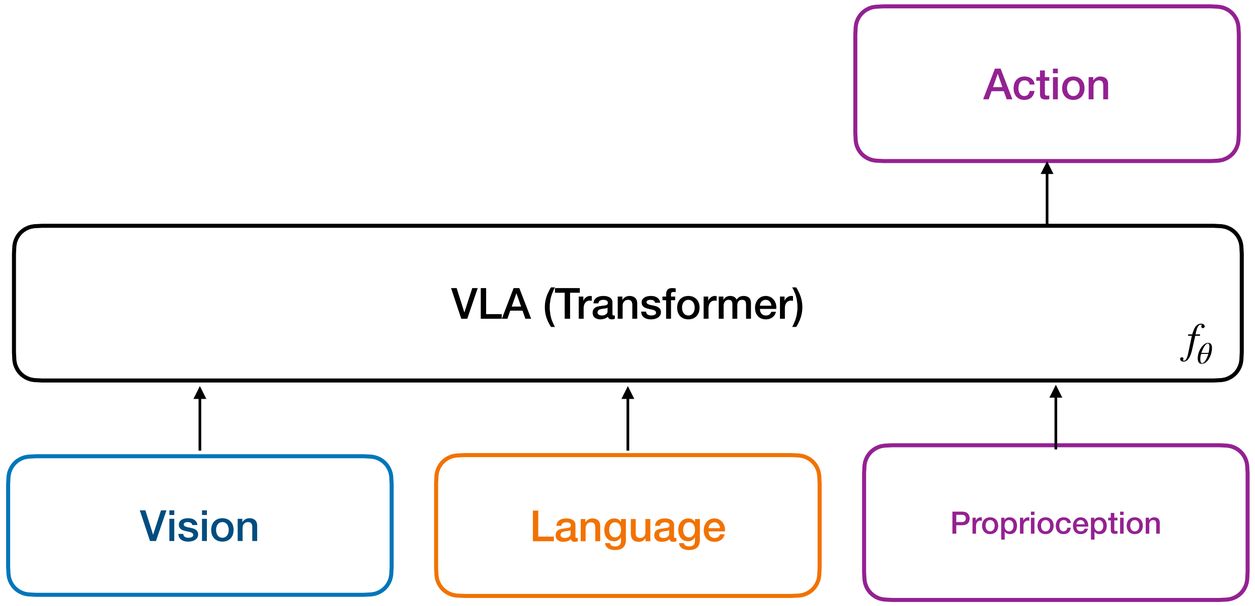
"티셔츠를 접어줘"

f_θ



Vision-Language-Action model

(Vision+Language → Action) 의 방대한 데이터셋에서 모방 학습 전략



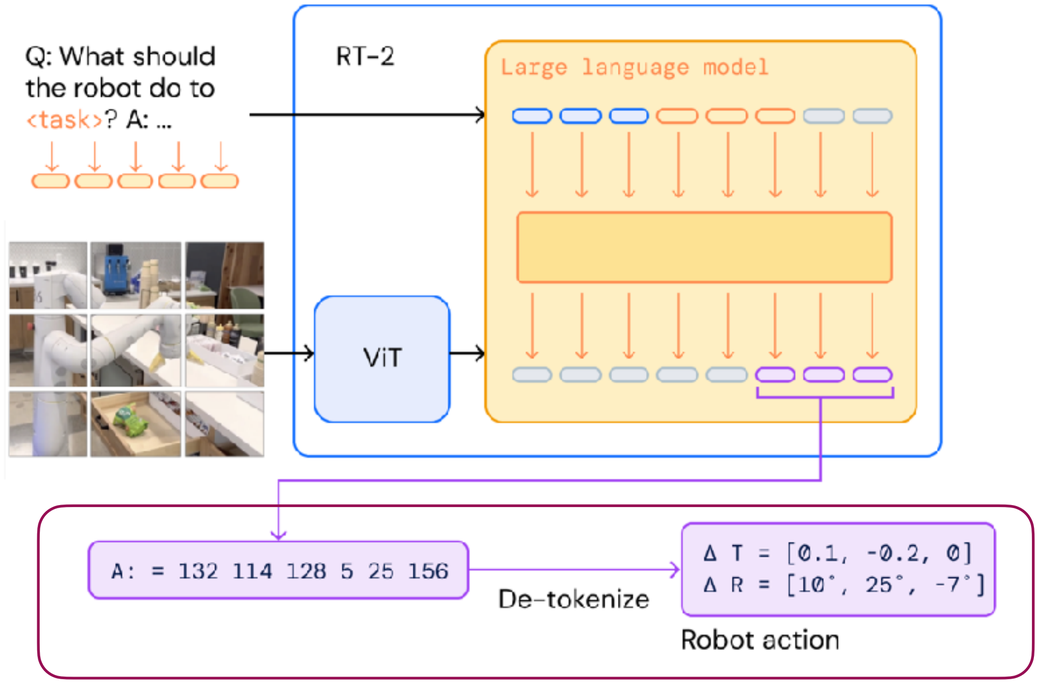
VLA 진화 과정

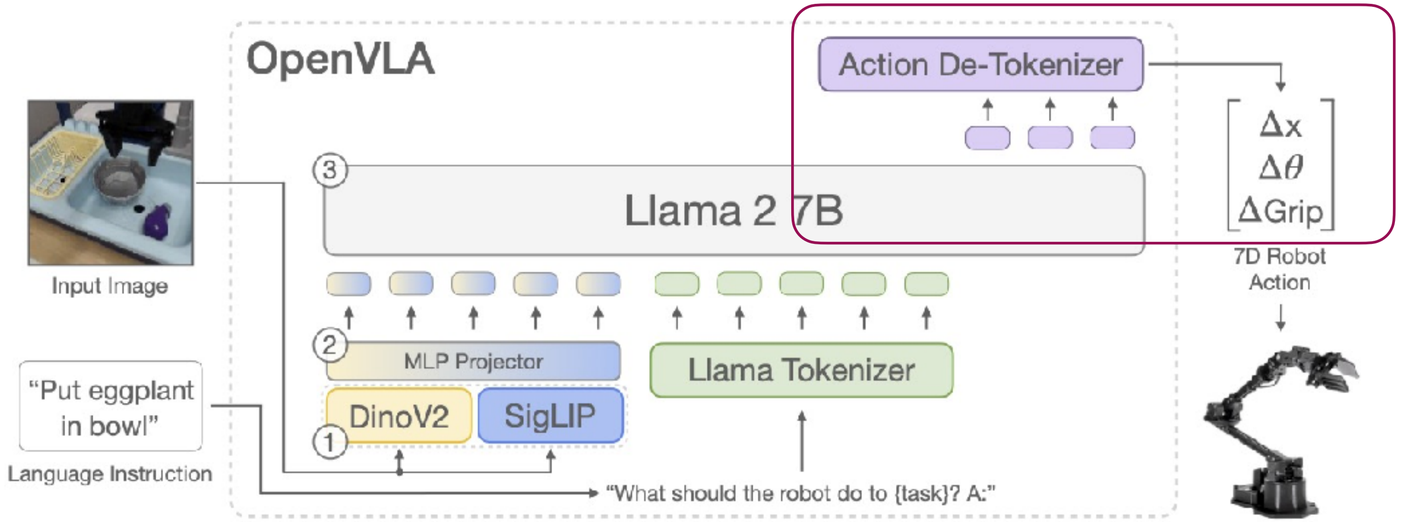
Action 생성의 발전 : binning

- Action binning (RT1,2 OpenVLA) : 0~1 값의 action을 1/256 binning, discrete token 으로 출력

(+) LLM으로 액션 시퀀스 예측 가능

(−) 속도와 성능이 아쉽





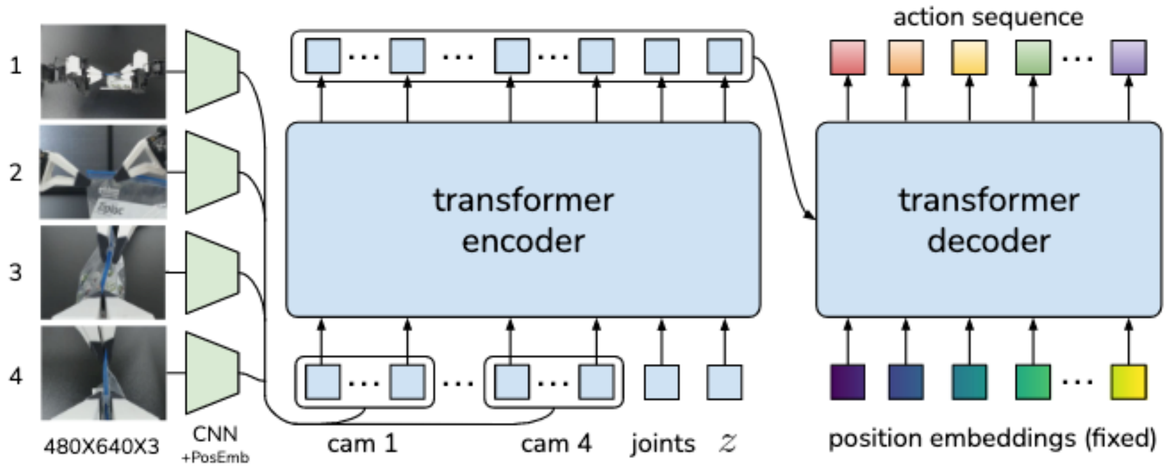
OpenVLA: An Open-Source Vision-Language-Action Model, 2024
https://openvla.github.io/

VLA 진화 과정

Action 생성의 발전 : binning→**chunking**→**diffusion/flow**

Action Chunking Transformer

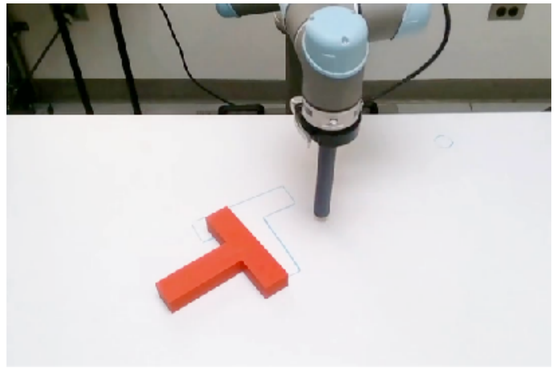
action sequence를 chunk로 예측

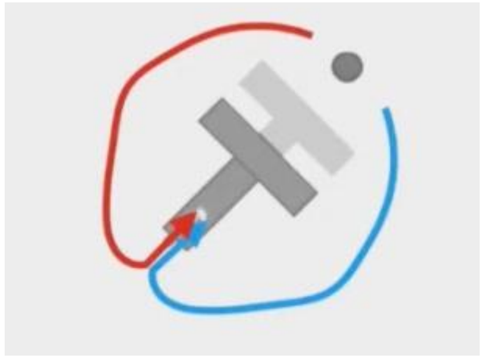


Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware, RSS2023
https://tonyzhaozh.github.io/aloha/

Diffusion policy, Flow matching

*빨간, 파란 경로 둘다 가능한 경로





Diffusion Policy: Visuomotor Policy Learning via Action Diffusion, RSS2023
https://diffusion-policy.cs.columbia.edu/

VLA 진화 과정

더 많은 robot embodiment / task 데이터

- Open-x embodiment

21개 기관, 22개 Robot Embodiments, 60개 Datasets
1M+ 로봇 궤적, 160k+ Tasks

Open X-Embodiment: Robotic Learning Datasets and RT-X Models
https://robotics-transformer-x.github.io/

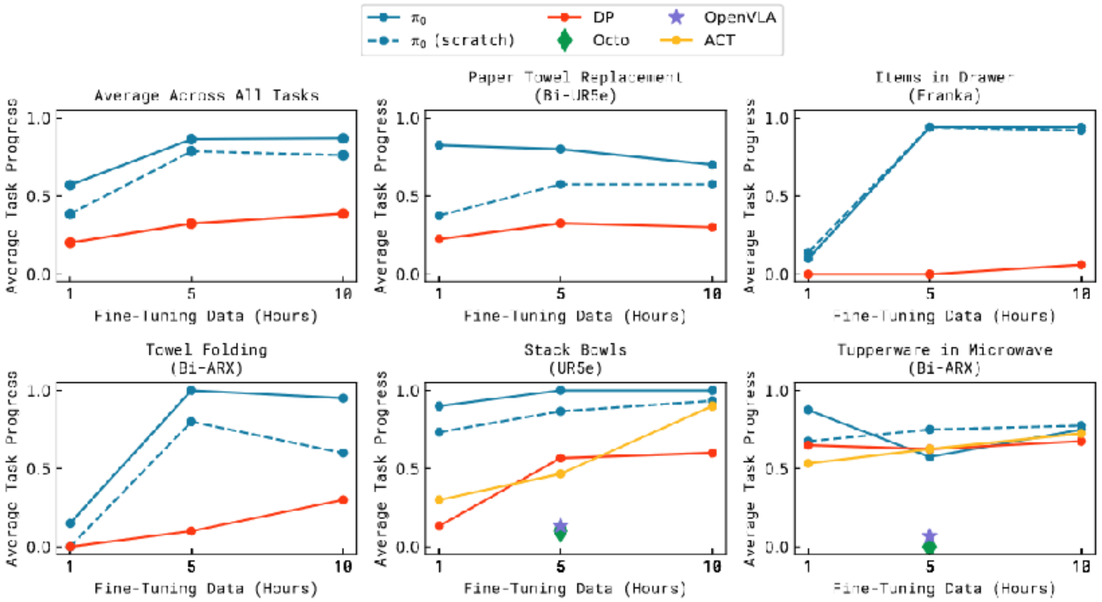
VLA 진화 과정

Multi-task learning

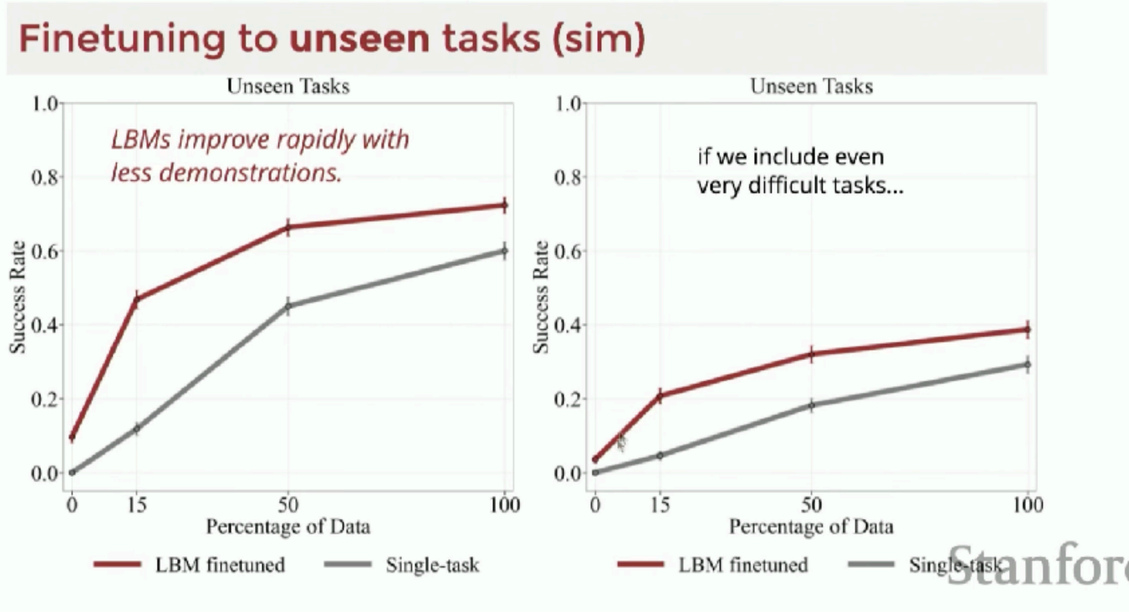
Pre-trained VLA > scratch VLA 보다 우수

Train "single-task" Diffusion Policy as one long task with 229 demonstrations

Stanfor



π0: A Vision-Language-Action Flow Model for General Robot Control
https://www.physicalintelligence.company/blog/pi0



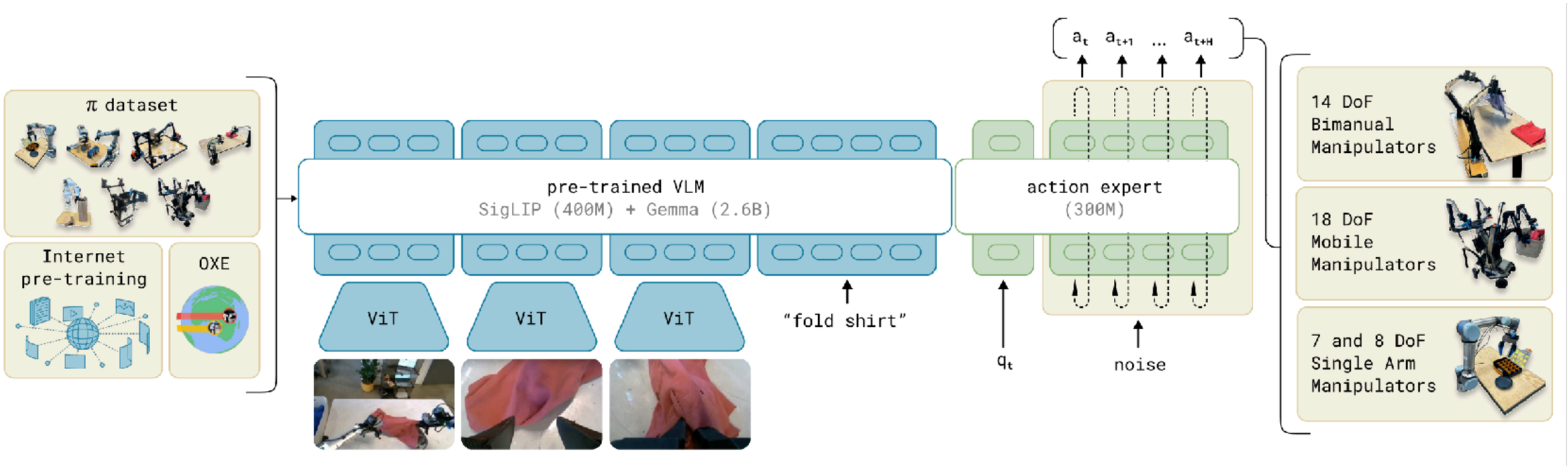
A Careful Examination of Large Behavior Models for Multitask Dexterous Manipulation
https://toyotaresearchinstitute.github.io/lbm1/

VLA 수렴 진화

Pre-trained VLM + DiT + Multi-task

Physical Intelligence, π_0

Physical Intelligence (π)

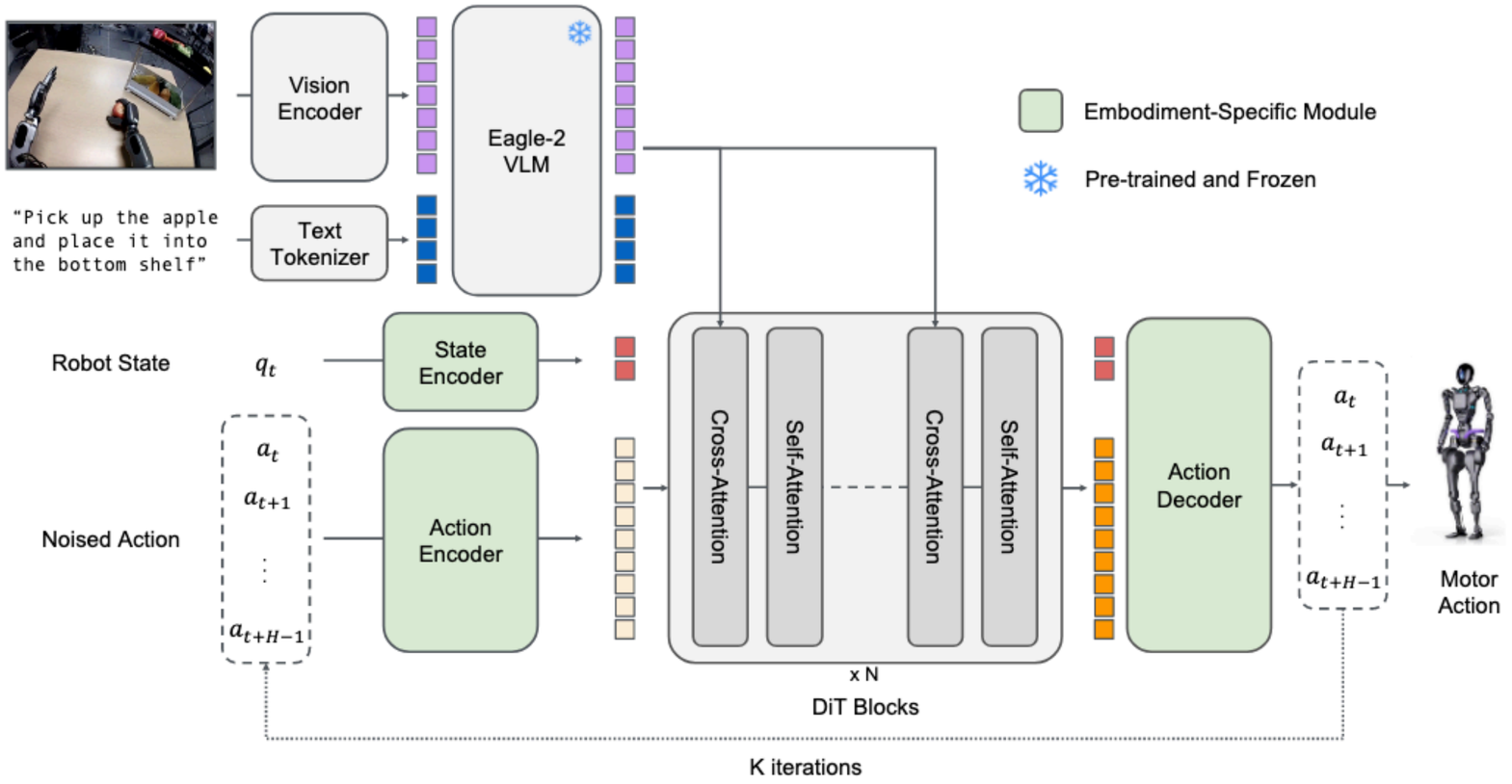


π0: A Vision-Language-Action Flow Model for General Robot Control
https://www.physicalintelligence.company/blog/pi0

VLA 수렴 진화

Pre-trained VLM + DiT + Multi-task

NVIDIA, `n1.5 gr00t`

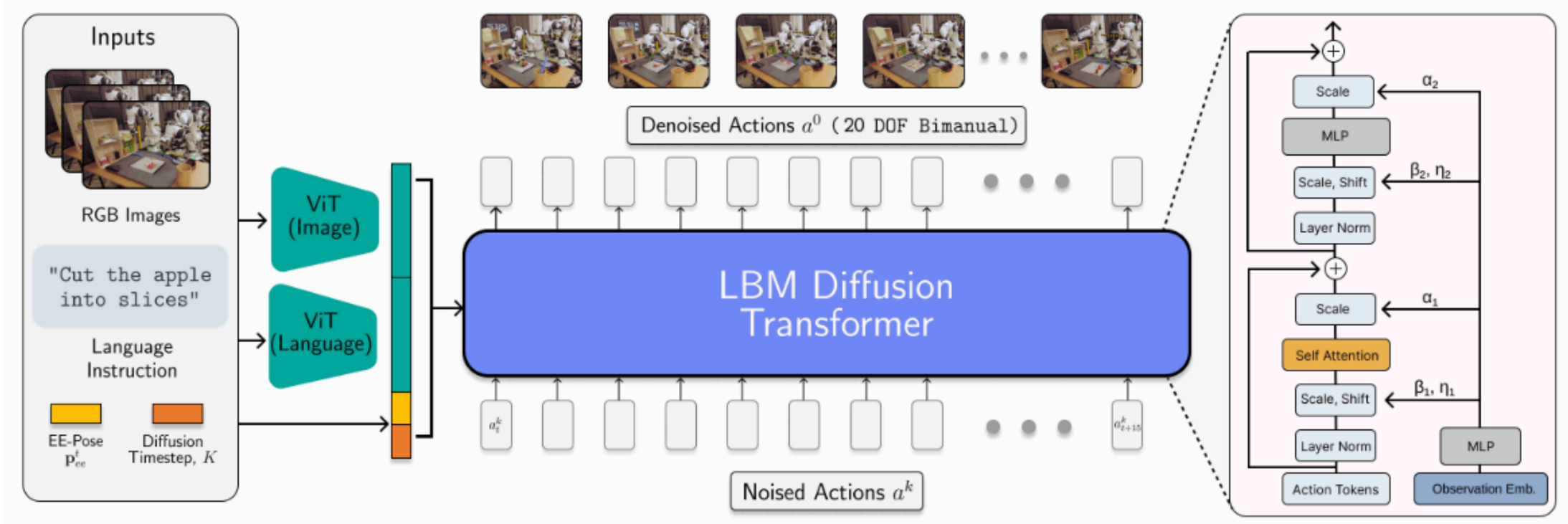


GR00T N1: An Open Foundation Model for Generalist Humanoid Robots
https://arxiv.org/abs/2503.14734

VLA 수렴 진화

Pre-trained VLM + DiT + Multi-task

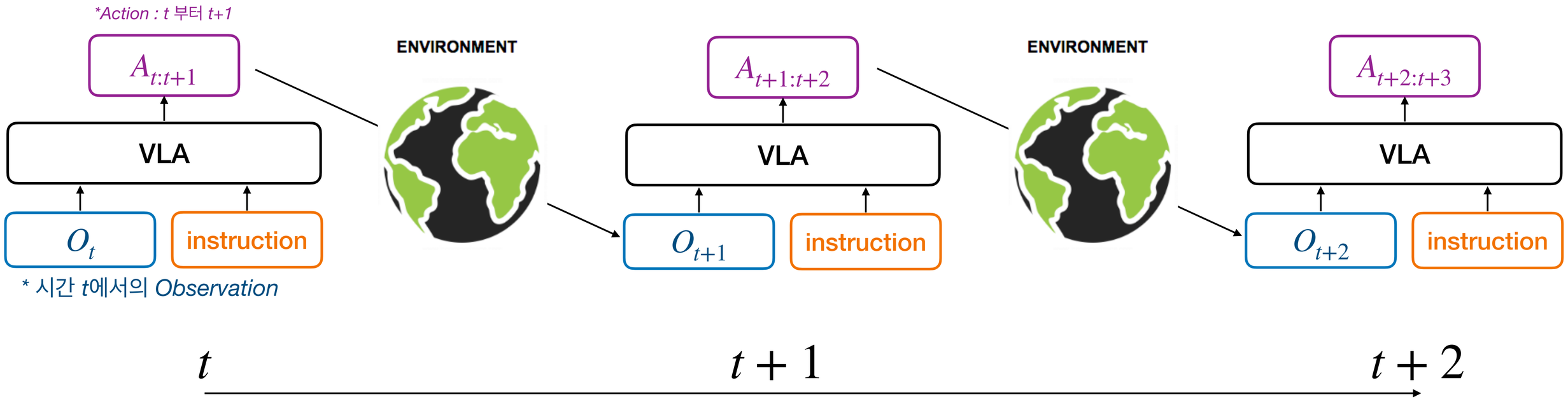
Toyota Research Institute, Large Behavior Model



A Careful Examination of Large Behavior Models for Multitask Dexterous Manipulation
https://toyotaresearchinstitute.github.io/lbm1/

VLA의 Deployment

Act & Observe



VLA의 장점, 포텐셜

+ data-driven imitation learning

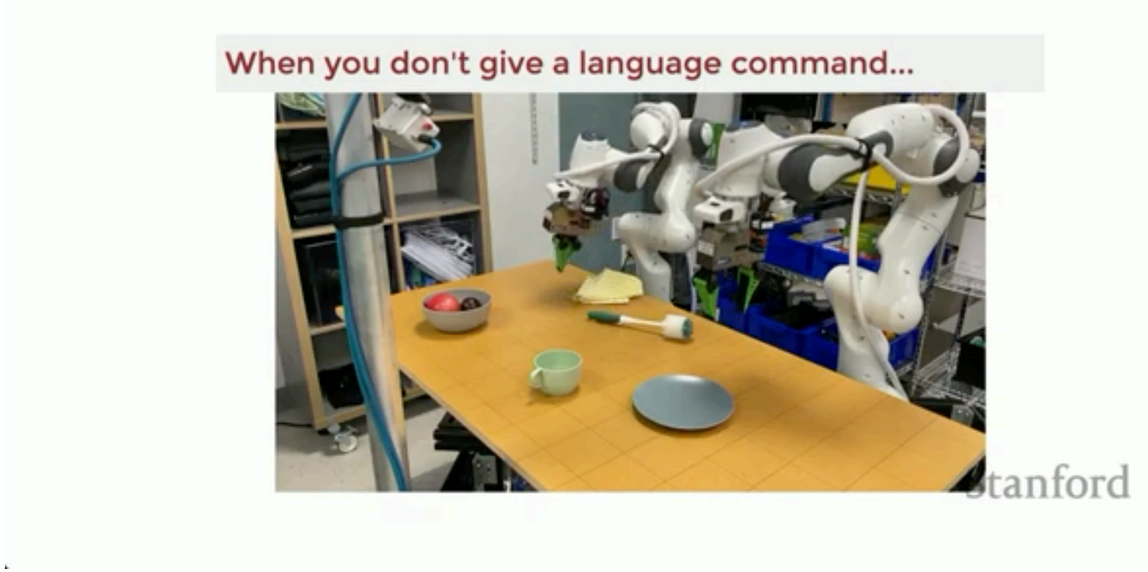
https://youtu.be/vBx6_E97Jxo?si=OmOZTHoNDaaLip-p

- 로봇의 **비정형 task가 가능성**
- Generalist policy, **Robot Foundation Model** 가능성,
- Finetuned Specialist policy는 scratch training 보다 우수

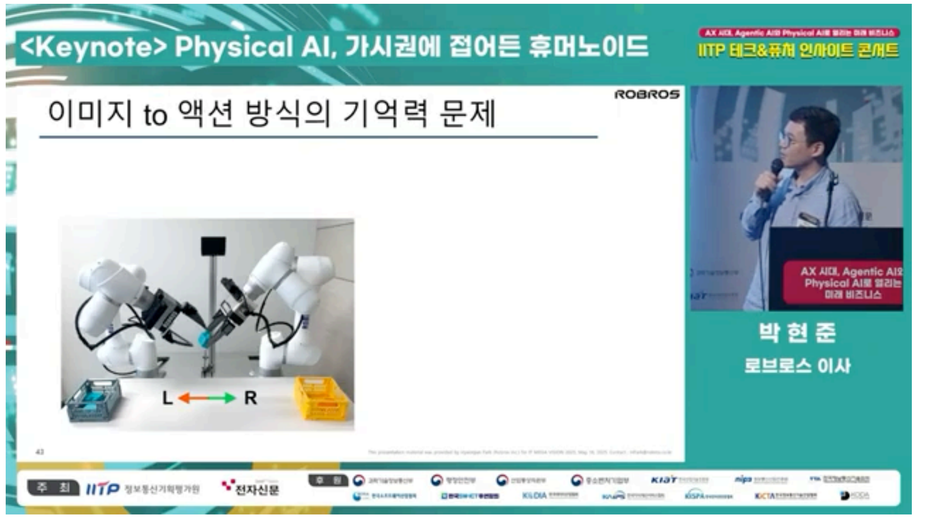
VLA의 현재 한계

생각보다 ... 많다

- Data: teleoperation cost, quantity
- Poor instruction following
- Lack of memory / history
- Evaluation cost
- Reliability, hard to achieve 99.99%



https://youtu.be/TN1M6vg4CsQ?si=QjpnhjG9Ua5Si1fg&t=2900



https://youtu.be/PZpf17oIVXk?si=u7PC_YIIbsW3cSOy&t=1370

두번째 레슨: *Inductive Bias for VLA*

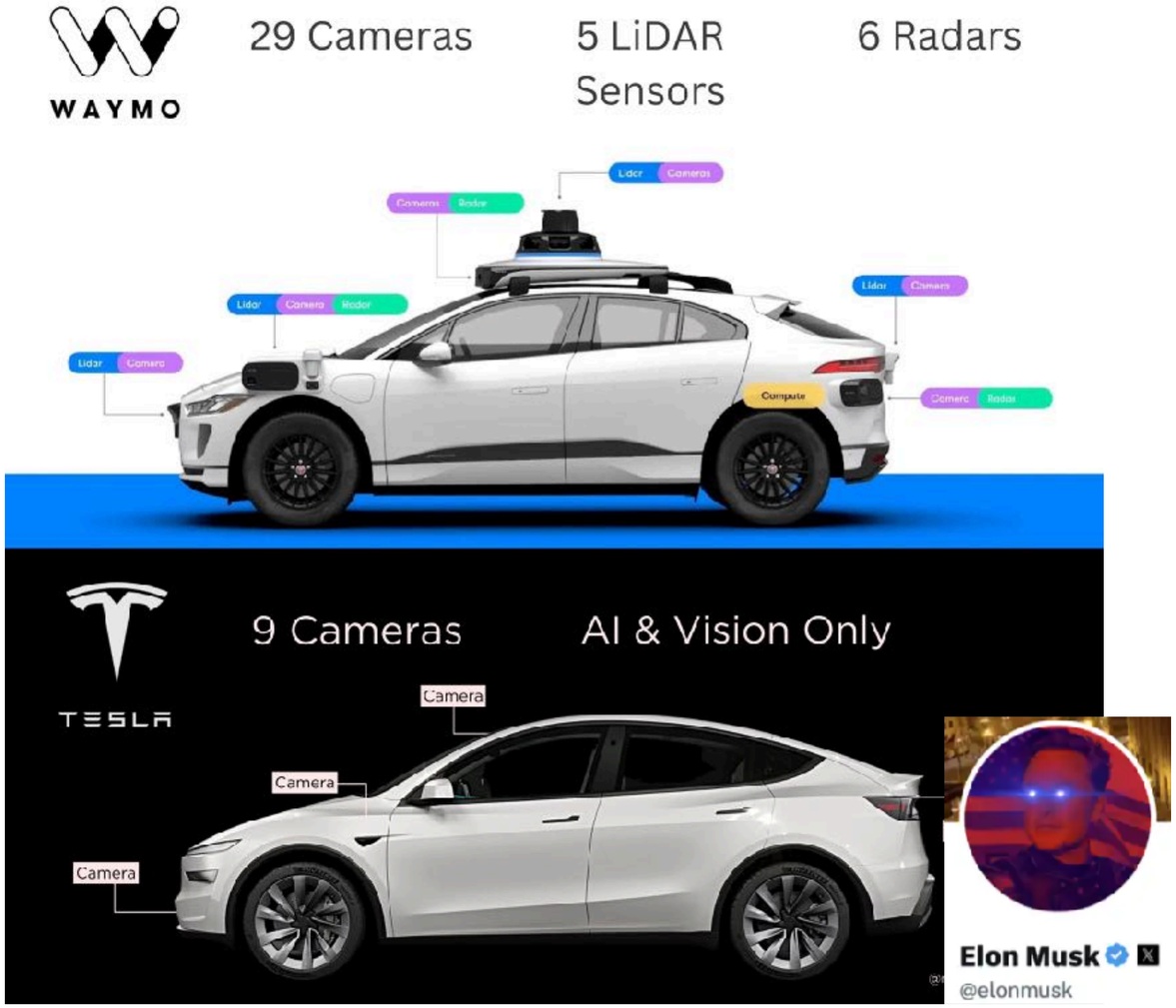
센서/하드웨어 선택

- 자율주행의 오래된 논쟁
  LiDAR vs Vision only

  정확도, 안정성, 비용, 환경 적응성

- VLA 시사점:

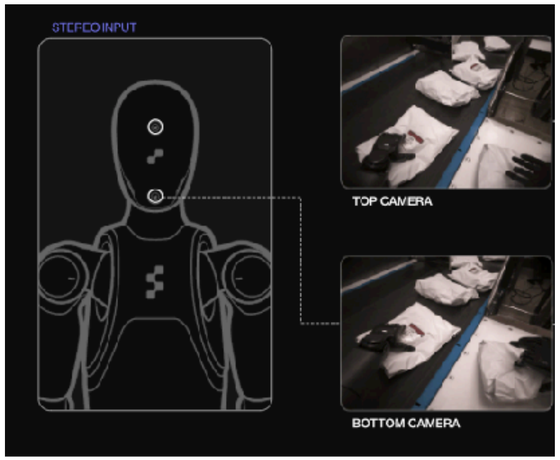
  어떤 Task에 따라
  proprioception, tactile, depth 등
  센서 조합, 하드웨어의 결정이 필요

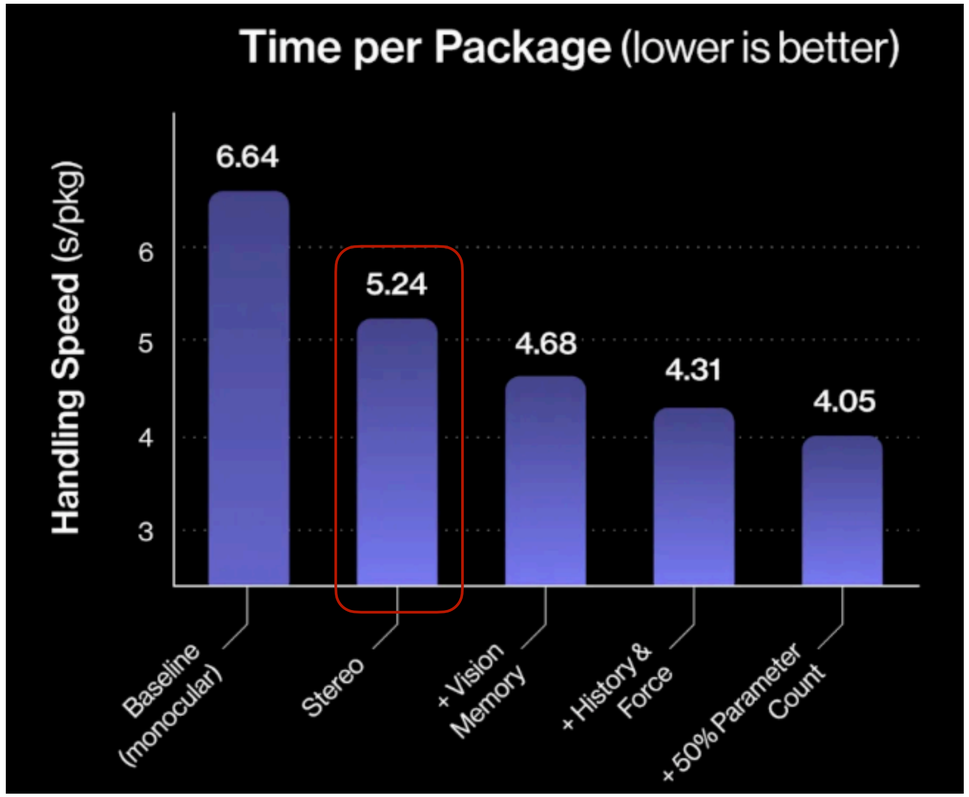


두번째 레슨: *Inductive Bias for VLA*

***Inductive bias: 모델이나 시스템이 세상을 해석하는 기본 가정과 제약, 모델 설계의 가정**

- 센서 선택, 로봇 embodiment 설계
  - **접근 (Reaching)** : Stereo camera? Depth? LiDAR?
  - **파지 (Grasp)** : Tactile force feedback? 초음파? 소리?
  - Humanoid VS Non-humanoid

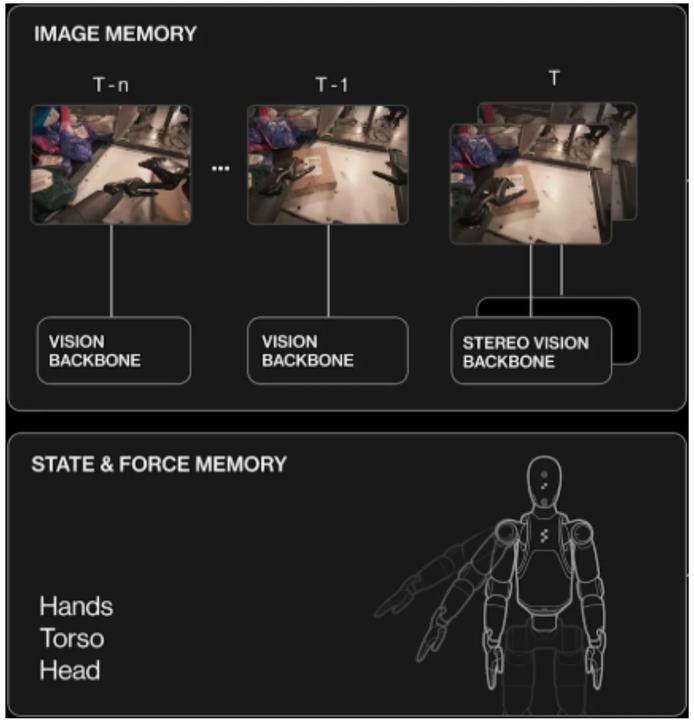


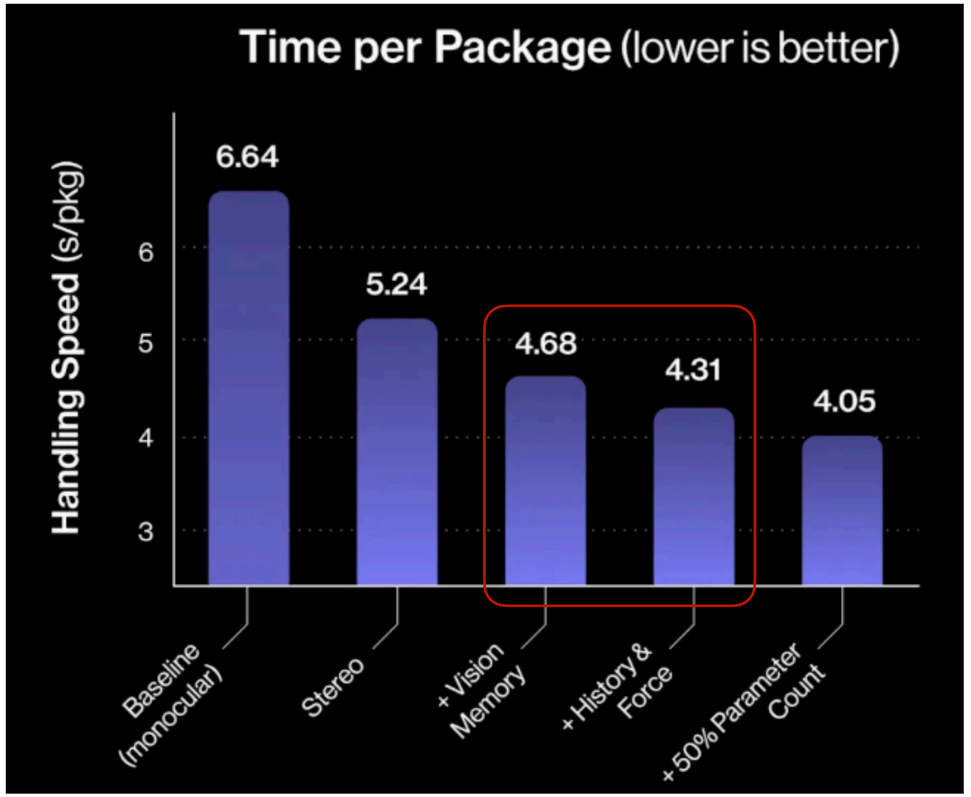


FigureAI Helix

두번째 레슨: *Inductive Bias for VLA*

- 모델 아키텍쳐
  - 예를 들어, 현재 프레임뿐만 아니라 최근 시각·동작 history

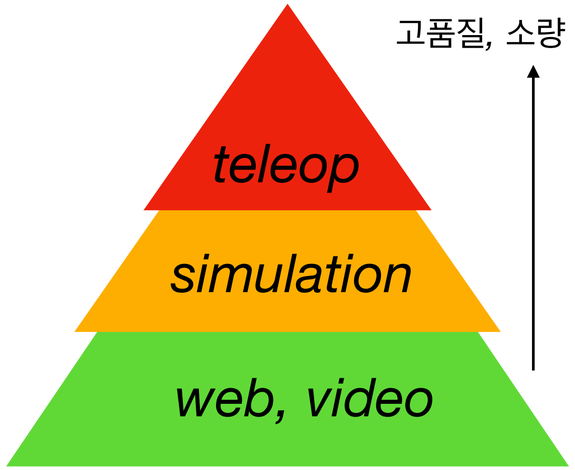




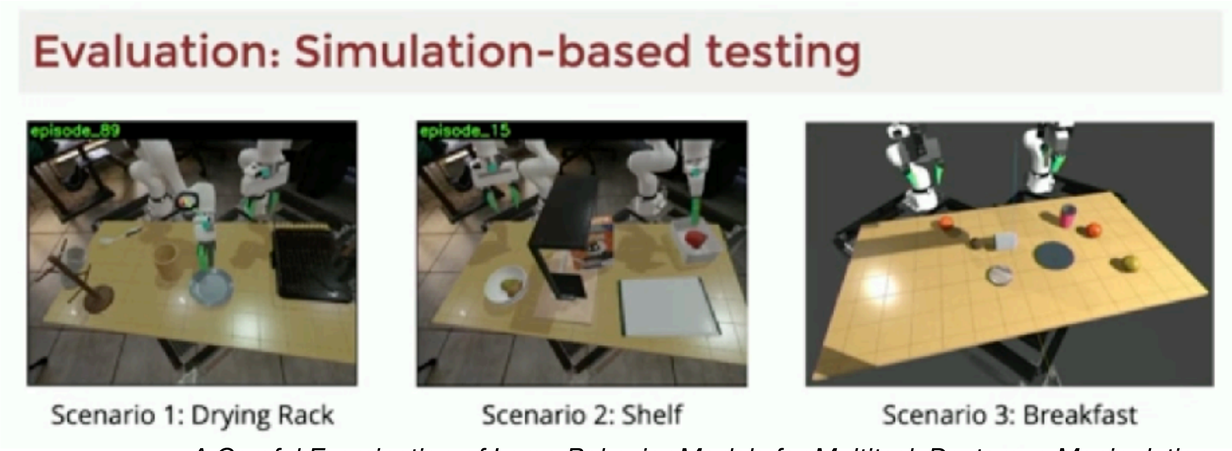
FigureAI Helix

세번째 레슨: *Scalability*

대규모 데이터 확보와 확장성 있는 모델 평가 방법



- 자율주행 수십~수백만 km 주행 데이터 필요. **Tesla는 fleet 도 운용.*
- VLA 시사점: 고품질 다량 Tele-operation, 시뮬레이터 데이터 혼합 전략 필요



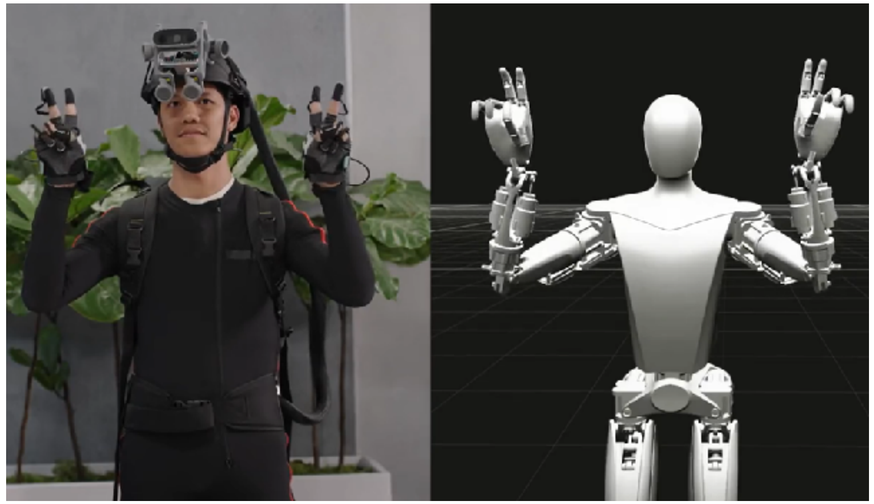
A Careful Examination of Large Behavior Models for Multitask Dexterous Manipulation
https://toyotaresearchinstitute.github.io/lbm1/

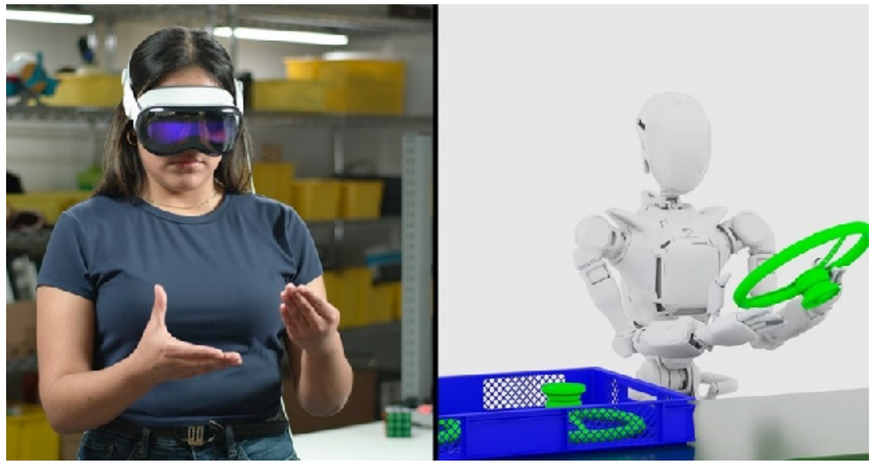
TRI LBM:
sim data를 co-train한 이유는
sim에서 eval 하기 위해서

Real World에서 Evaluation cost 매우 높음

세번째 레슨: *Scalability*

당장 좋은 퀄리티의 teleop 데이터를 모아야한다.





**teleop*도 연구 분야중 하나
teleop 데이터를 마치 *LLM*의 *web-sclae*로 모을 수 있는가?

테슬라가 시간당 48달러,
휴머노이드 로봇 ‘Optimus’ teleoperator 에게 지급

5’7″~5’11″(약 170~180cm)
하루 7시간 이상 걷기와 최대 30파운드(약 13.6kg) 하중 운반이 가능한 체력이 요구됩니다.

https://interestingengineering.com/culture/tesla-paying-to-train-optimus-robot

세번째 레슨: *Scalability*

합성·시뮬레이션 데이터

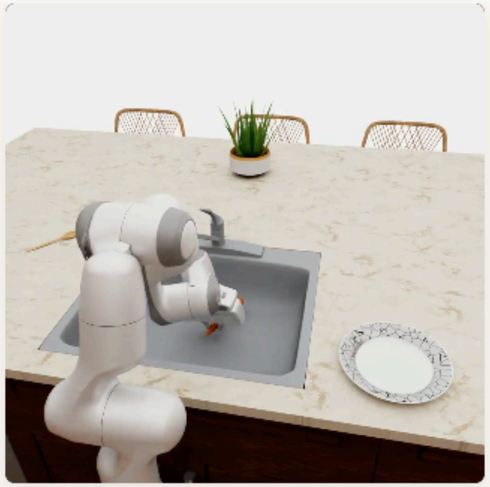
simulation

대규모 실물 시연을 모으는 데 비용이 커서,

합성 영상·시뮬레이터로 ‘무한’ 데이터를 뽑아내는 노선
(ex. GR00T N1.5)

Pick up the apple, place it into the drawer and close the drawer.

Pick up eggplant from placemat to plate.



Pick the carrot from the sink and place it on the plate located on the counter.

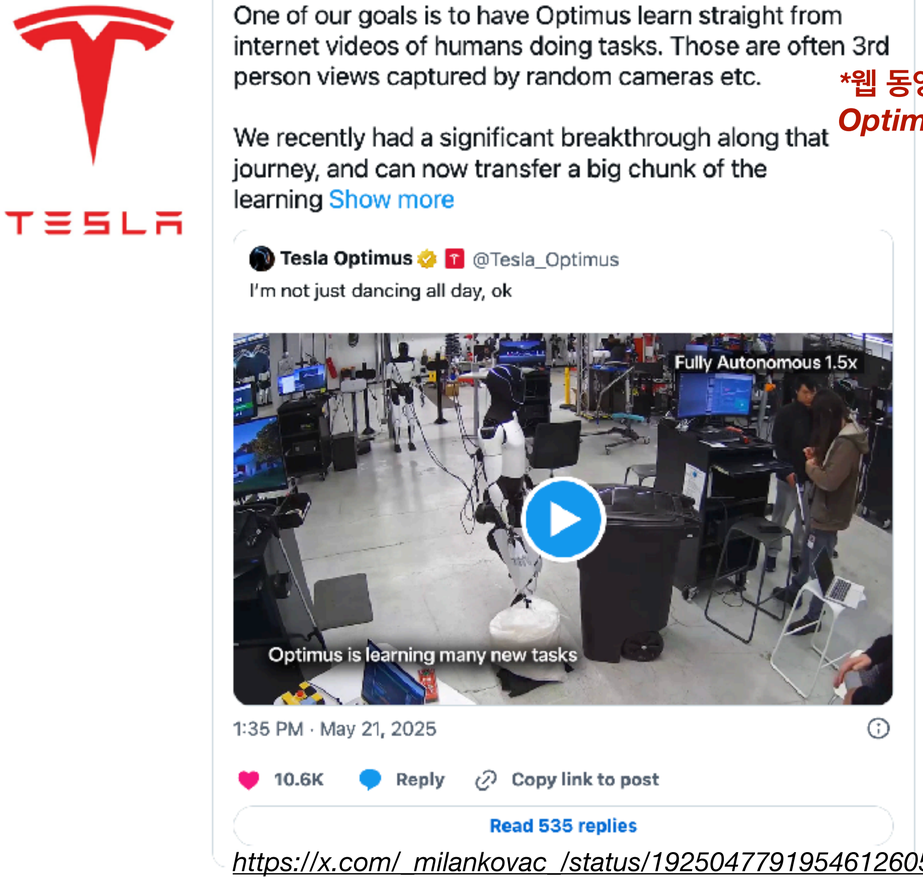
Pick the teapot from the counter and place it in the sink.

https://research.nvidia.com/labs/gear/flare/
https://research.nvidia.com/labs/gear/dreamgen/

세번째 레슨: *Scalability*

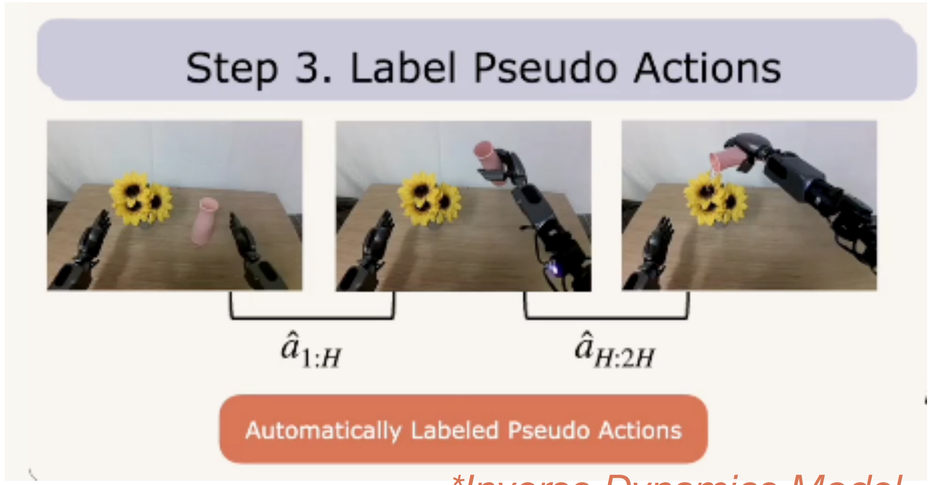
액션 라벨 없는 사람의 비디오에서 학습

web, video

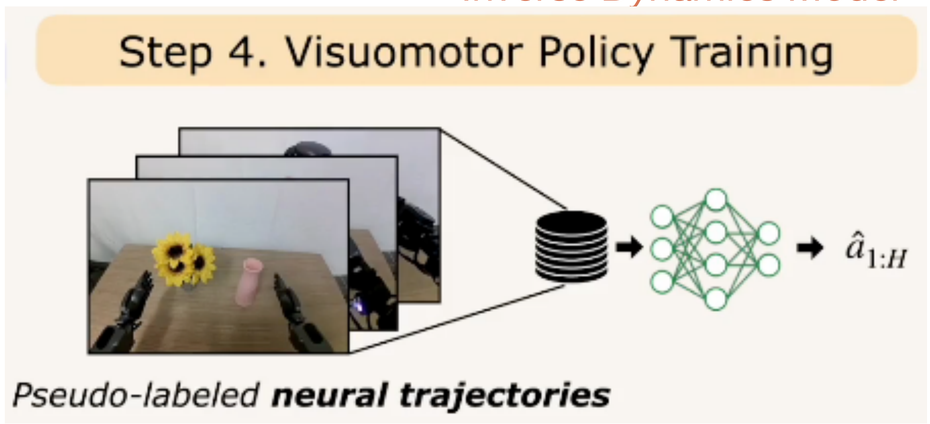


https://x.com/_milankovac_/status/1925047791954612605

*웹 동영상의 사람 작업 수행을 ***Optimus***가 학습하는게 목표



*Inverse Dynamics Model

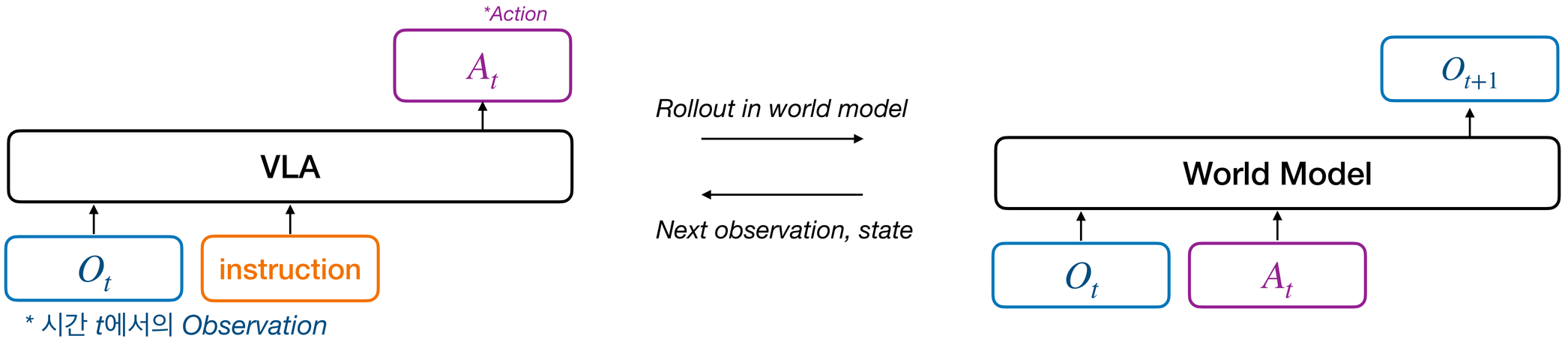


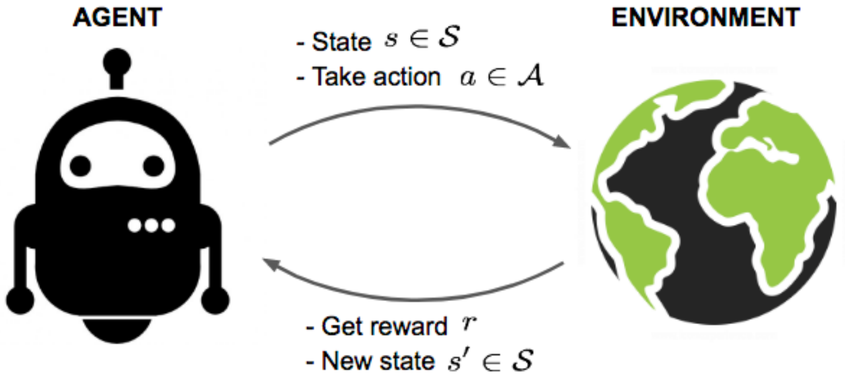
https://research.nvidia.com/labs/gear/dreamgen/

세번째 레슨: *Scalability*

World Model

* 현재 시간 *t* 에서의 *Observation, Action*으로 다음 시간 *t+1* 에서의 *Observation*을 예측 하는 모델

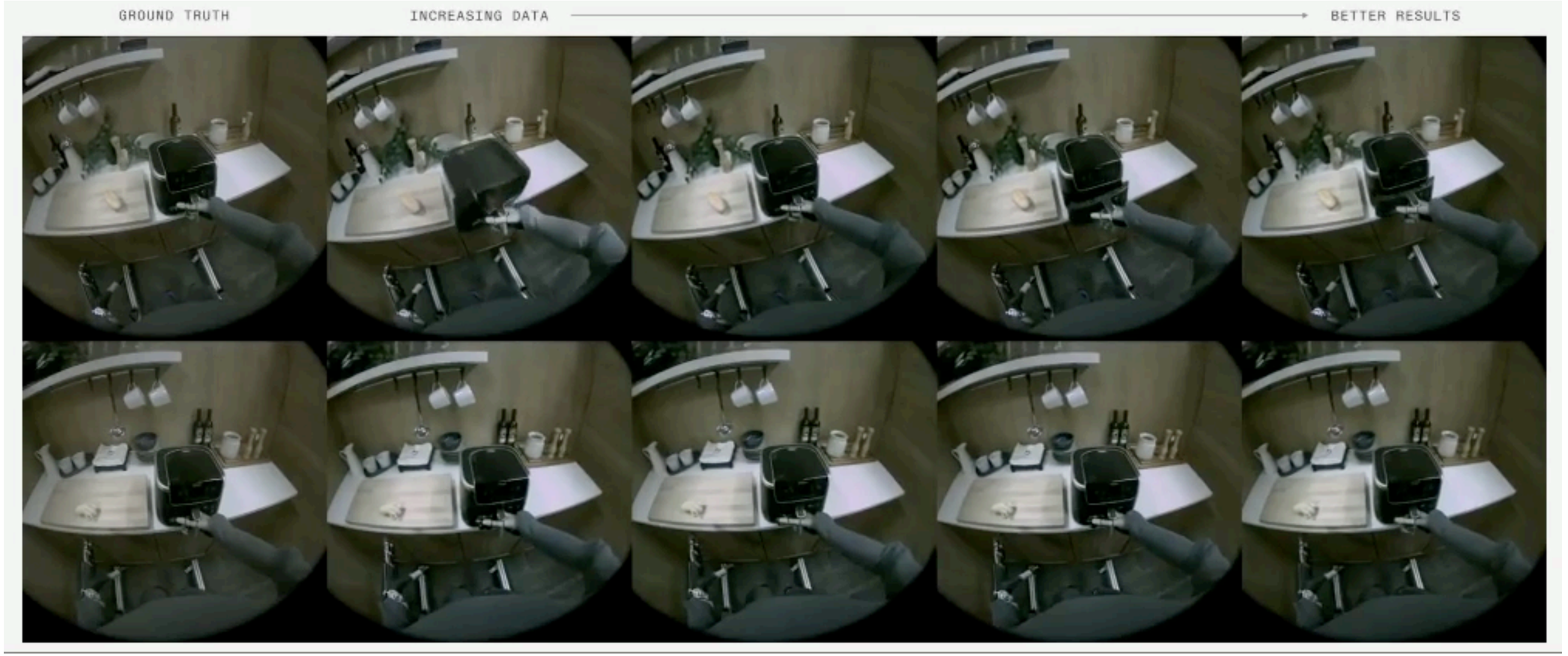




세번째 레슨: *Scalability*

World Model

* *Data* 가 많을 수록 정교한 *World Model* 학습 가능



*VLA*의 *evaluation*을 *eval cost* 높은 실제 환경 대신에 *World model* 에서 *proxy* 평가

https://www.1x.tech/discover/redwood-ai-world-model

세번째 레슨: *Scalability*

World Model, Model-based RL

- Teleop은 사실 완벽한 고품질 데이터는 아님

  *자신의 신체가 아닌 것을 조종

- 모방 학습의 한계

- 더 효율적인 행동을 Embodied Agent가 World Model 내에서 탐색할 수 있다면?

https://deepmind.google/discover/blog/genie-3-a-new-frontier-for-world-models/

Genie 3 Prompt:
"POV action camera of a tan house being painted by a first person agent with a paint roller"

* *robot agent*의 시점이라면?

세번째 레슨: *Scalability*
World Model, Model-based RL

Embodied Agent가 **World Model**에서 여러 가상 시뮬레이션(**rollout**)을 수행

이를 통해 효율적인 action을 탐색하고 **강화학습(RL)**에 활용

모방학습(IL)만으로는 어려운 일반화·탐색 문제를 극복할 수 있음

LLM 발전 과정에서의 교훈과 유사 (SFT + RLVR)

네번째 레슨: *Business Strategy*

시장 진입 순서

Specialist → Generalist

자율주행

ADAS (Advanced Driver Assistance System)

제한된 환경(고속도로, 저속 주차 등)에서 안정성 확보

Highway Pilot / 특정 환경 자율주행

‘제한된 맵’ + ‘고정된 조건’에서 신뢰성 향상

FSD (Full Self-Driving)

비정형 환경 대응

Robot Foundation Model

Specialist VLA

반복적, 구조화된 환경에서 동작 *공장, 물류센터, 단순 가사 로봇

Semi-generalist VLA

환경/작업 범위를 점차 확장 Multi-task 학습,

Generalist RFM

새로운 환경, Multi-embodiment, multi-task 수행

네번째 레슨: *Business Strategy*

전략적 유사점

- 제한된 환경에서 먼저 성공사례 만들기 (자율주행의 고속도로, 로봇의 공장/물류)
- ROI 빠른 곳부터 진입 (투자자 설득)
- 데이터를 Scalable하게 모으는 방법 고안

자율주행이 "도로" 정복하려는 도전이라면,
RFM은 "실제 세계" 정복하려는 도전

감사합니다