DMQA Seminar 20250214

Introduction to Robot Learning

Training generalist robot policy using vision-language model

일반대학원 산업경영공학과
김재훈

고려대학교
KOREA UNIVERSITY

Introduction

- 발표자 소개

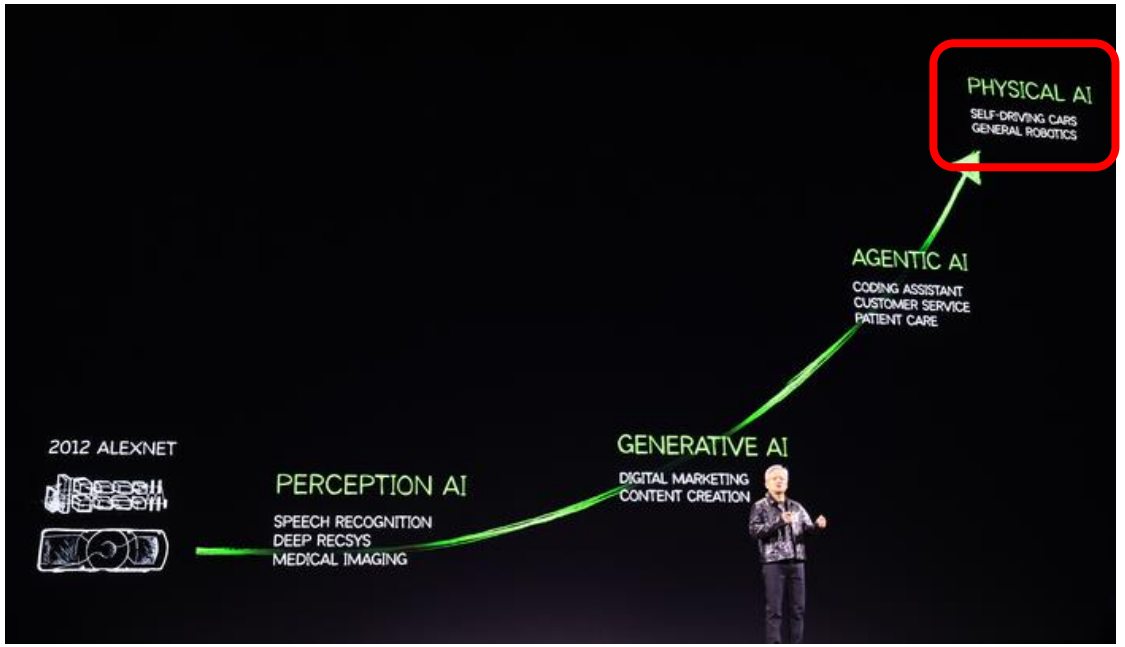
- 이름: 김재훈
- 학력
  - ✓ 2020.03 – 현재 | 석박사통합과정 | 고려대학교 산업경영공학과 (지도교수: 김성범)
- 관심분야
  - ✓ Reinforcement learning
  - ✓ Natural Language Processing
  - ✓ Self-supervised Learning
- e-mail : jhoon0418@korea.ac.kr

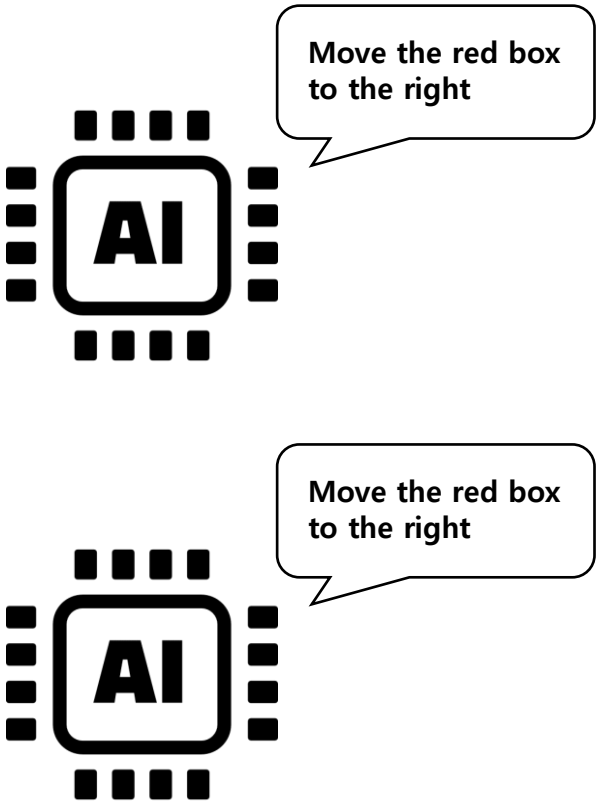
Robot Learning

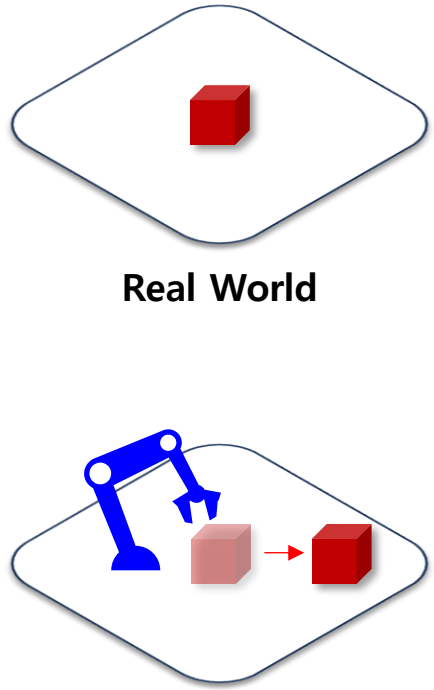
Introduction

❖ Physical Intelligence

- 하드웨어가 현실 세계를 인식 및 이해하고 복잡한 행동을 수행하여 현실 세계에 영향을 줄 수 있도록 하는 인공지능







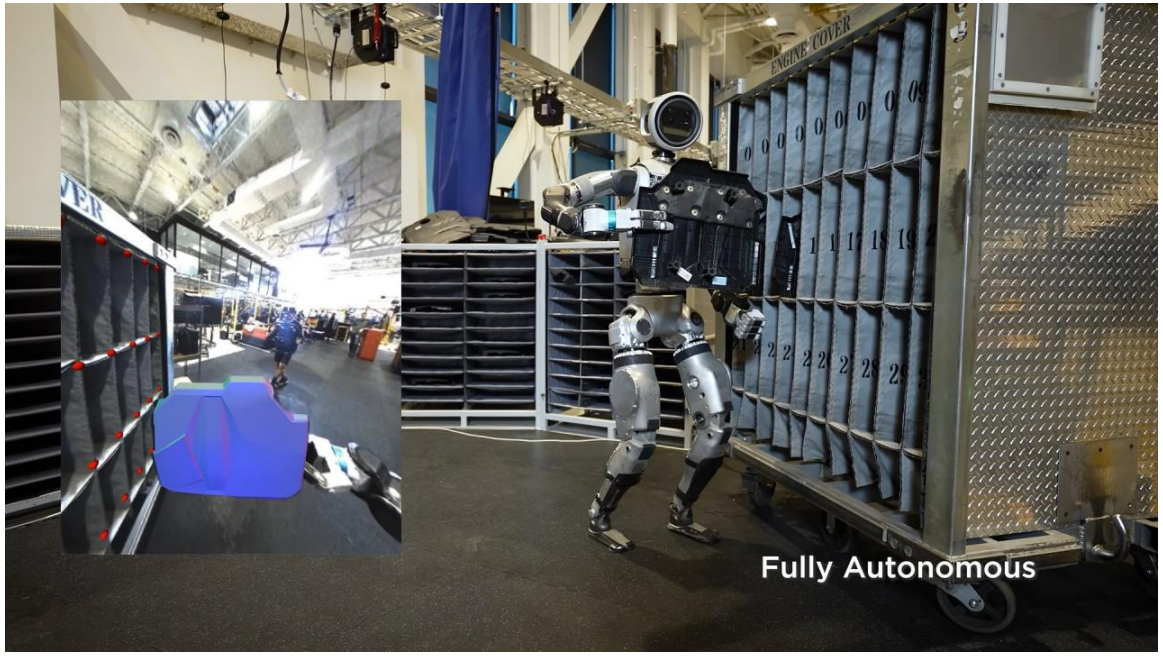
https://www.nvidia.com/en-us/glossary/generative-physical-ai/
https://weekly.khan.co.kr/khnm.html?mode=view&art_id=202501200600001&dept=114

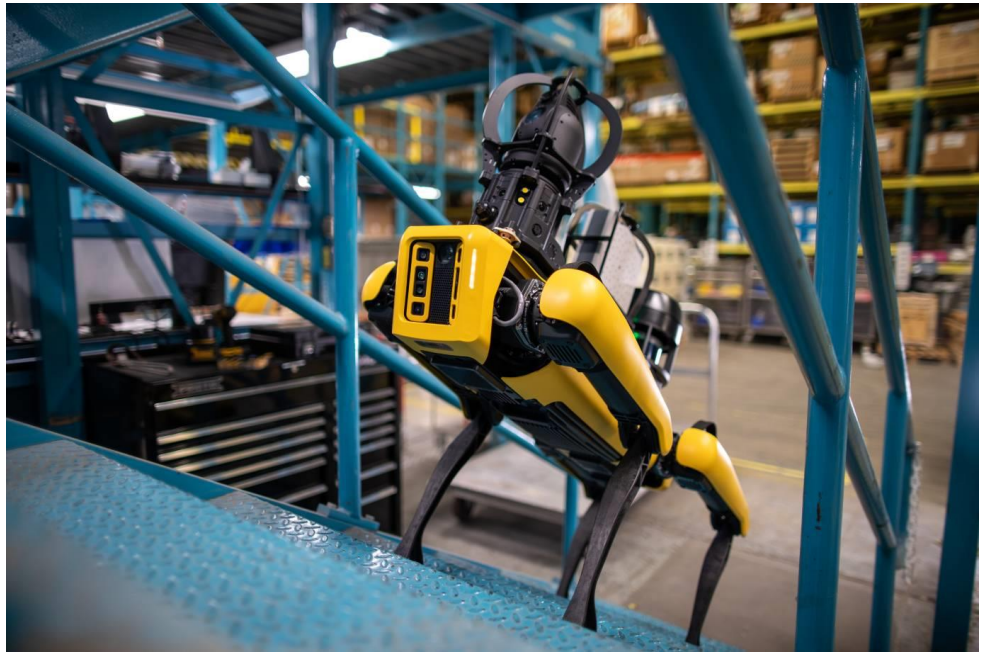
Robot Learning

Introduction

❖ Robot Intelligence

- Robotics는 physical intelligence의 대표적인 분야
- 최근에 robot의 하드웨어도 큰 발전이 있었으며 로봇개나 휴머노이드가 많은 관심을 받고 있음
- 산업에서도 사용되기 시작했으며 더 고차원의 작업을 수행할 수 있도록 인공지능을 적용하는 robot intelligence가 활발히 연구되고 있음





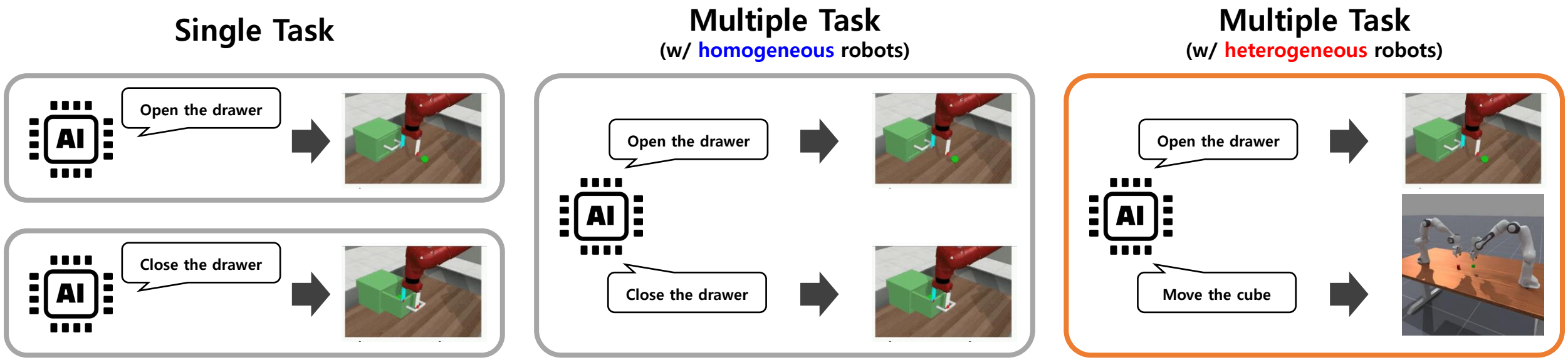
https://www.koreaherald.com/article/3852438
https://bostondynamics.com/products/spot/

Robot Learning

Introduction

❖ General Robot Intelligence (= Robot Foundation Model)

- Foundation model은 대규모 데이터로 사전학습되어 다양한 세부 과제에 범용적으로 활용할 수 있는 머신러닝 모델 (ex. GPT, Segment Anything)
- 다양한 로봇 하드웨어와 로봇 과제에서 범용적으로 활용될 수 있도록 사전학습된 모델
- 학습하지 않은 작업도 기존 지식을 바탕으로 추론하여 성공적으로 수행할 수 있는 모델



https://gymnasium.farama.org/index.html#

Robot Learning

Introduction

❖ Reinforcement Learning vs. Imitation Learning

- 강화학습
  - 환경과의 상호작용을 통해 최적의 정책을 학습
  - 주어진 상태에서 누적 보상을 최대화하는 방향으로 정책을 업데이트
- 모방학습
  - 전문가의 행동을 따라하는 정책을 학습
  - 주어진 상태에서 정답 행동을 잘 맞추도록 지도학습으로 정책 업데이트

➡ Robot foundation model 연구는 모방학습을 이용해서 연구되고 있음

Robot Learning

Introduction

❖ Research Timeline

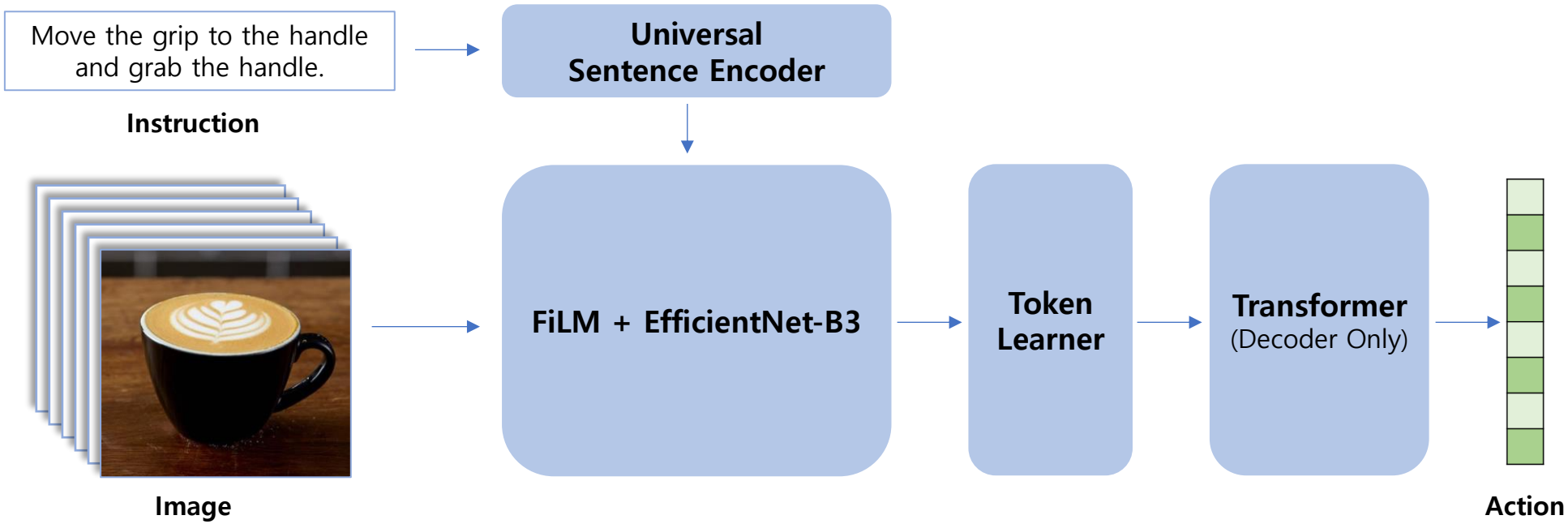
- Robot learning에서 중요하게 다루는 측면 두 가지
  1. 다양한 환경에서 다양한 작업을 강건하게 수행할 수 있는 능력
  2. 현재 상황에서 수행할 다음 행동을 계산하는데 걸리는 시간
- 이에 연구는 모델의 일반화 성능 개선과 연산 시간 단축을 중점으로 수행
- Low level control을 수행하는 제어 모델 RT-1, RT-2, $\pi 0$ 모두 모방 학습을 기반으로 정책을 학습

Robot Foundation Models

Robotics Transformer 1 (RT-1)

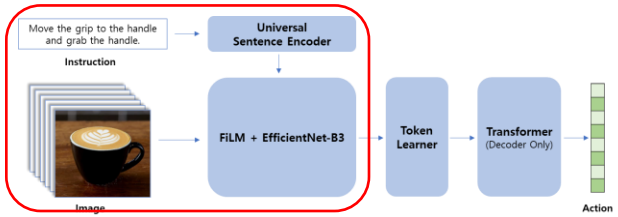
❖ RT-1: Robotics Transformer for Real-World Control at Scale (Google Research, 2022, 914회 인용)

- 로봇 학습에서는 일반화 성능과 추론 속도가 중요함 → RT-1은 대규모 데이터셋과 모델 크기를 확장하여 해당 목표를 달성하고자 함
- 보통 모델 파라미터의 크기가 증가할 수록 일반화 성능이 좋아지지만 동시에 추론 속도가 하락하는 경향이 있음
- RT-1은 이 문제를 완화할 수 있는 로봇 학습에 적합한 아키텍처를 제시함



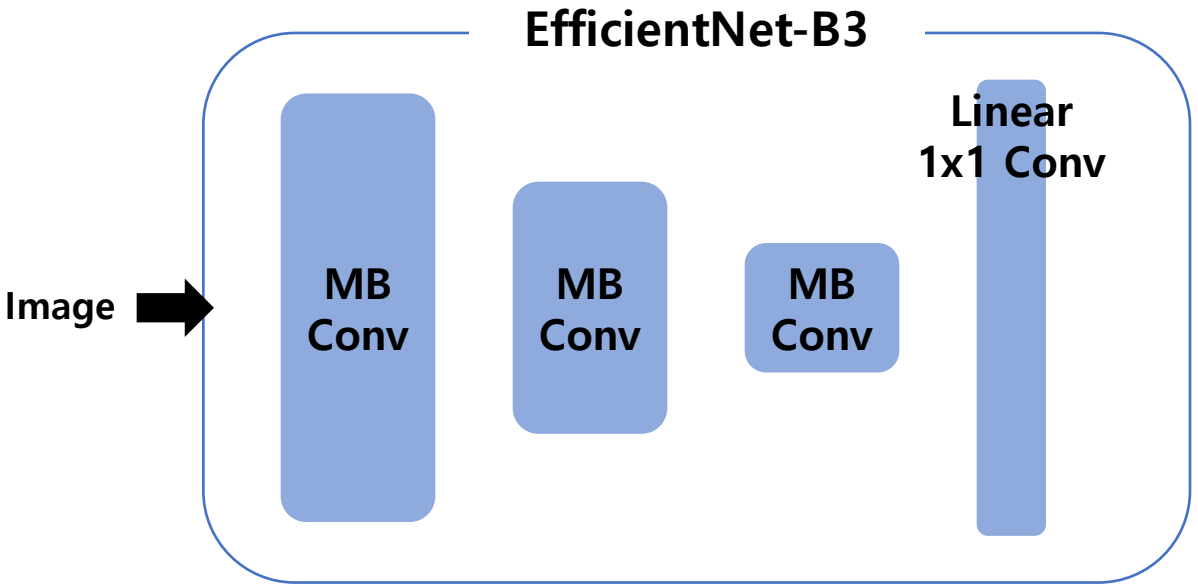
Robot Foundation Models

Robotics Transformer 1 (RT-1)

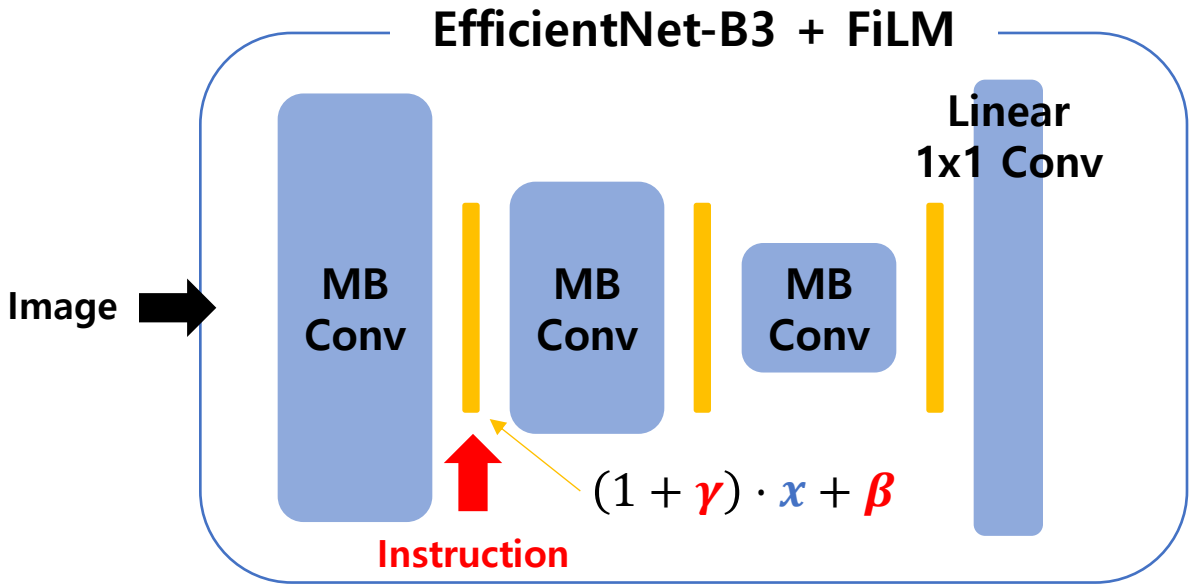


❖ Detailed Architecture of RT-1

- Feature-wise linear modulation (FiLM)은 서로 다른 두 정보를 하나의 정보로 합치는데 효과적인 방법론
- EfficientNet-B3는 사전학습된 이미지 인코더이며 파라미터 수가 작지만 뛰어난 성능을 보여주고 추론 속도가 빠르다는 특징이 있음
- FiLM을 EfficientNet-B3의 각 블록 사이마다 삽입하여 multimodal learning을 수행
- 이 때 사전학습으로 얻은 성능을 훼손하지 않기 위하여 처음엔 FiLM의 가중치를 항등함수와 동일하도록 초기화한 뒤 점차 업데이트를 진행



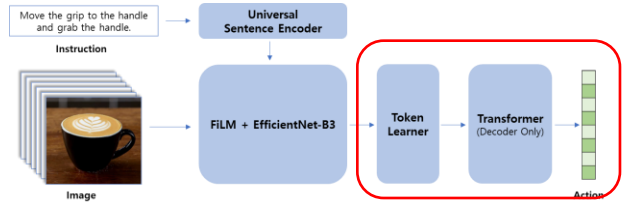
< 기 존 >



< 변 경 >

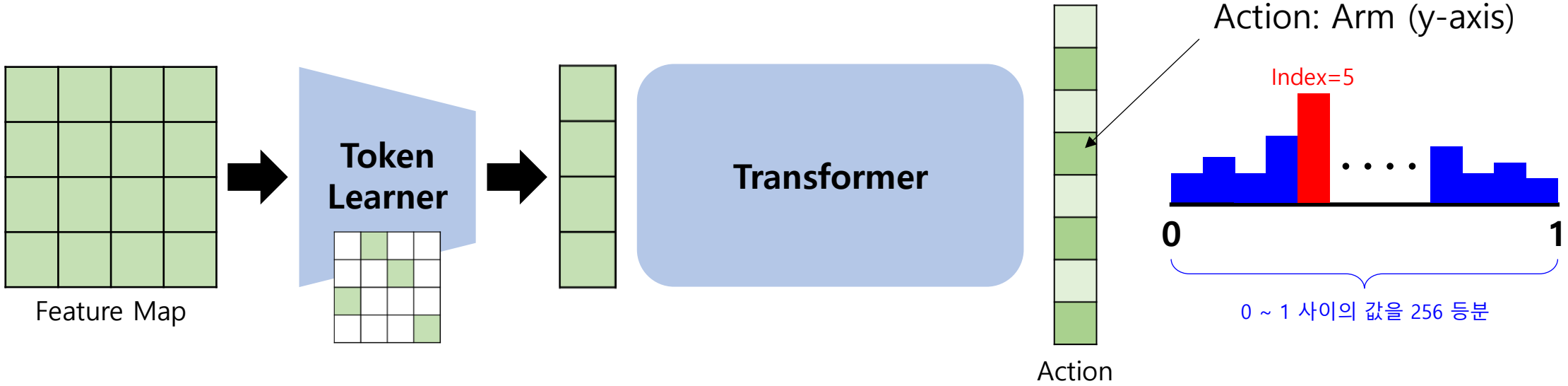
Robot Foundation Models

Robotics Transformer 1 (RT-1)



❖ Detailed Architecture of RT-1

- 인코더에서 특징을 추출한 뒤 transformer에 입력하기 전, token learner를 사용하여 중요한 일부 토큰을 선별함
  → 연산량을 줄여 추론 속도 확보
- 이후 transformer를 사용하여 action token을 예측
- 원래 연속적인 값으로 행동이 구성되지만 이를 256개의 구간으로 나누어 이산적인 값으로 표현함
  → 해당 테크닉이 더 복잡한 분포를 학습했다고 서술 (연속적인 값으로 학습할 경우 단일 모드의 정규 분포를 학습)



Ryoo, Michael S., et al. "Tokenlearner: What can 8 learned tokens do for images and videos?." *arXiv preprint arXiv:2106.11297* (2021).

Robot Foundation Models

Robotics Transformer 1 (RT-1)

❖ Experiment Settings (dataset & robot)

- 17개월 동안 13대의 로봇(EverydayRobot)으로 현실 세계에서 수집한 13만 여개의 동작 데이터(총 744개의 과제)를 사용
- EverydayRobot에 학습한 정책을 탑재하여 성능 검증을 수행

Skill	Count	Description	Example Instruction
Pick `Object`	130	Lift the object off the surface	pick iced tea can
Move `Object` Near `Object`	337	Move the first object near the second	move pepsi can near rxbar blueberry
Place `Object` Upright	8	Place an elongated object upright	place water bottle upright
Knock `Object` Over	8	Knock an elongated object over	knock redbull can over
Open `Drawer`	3	Open any of the cabinet drawers	open the top drawer
Close `Drawer`	3	Close any of the cabinet drawers	close the middle drawer
Place `Object` into `Receptacle`	84	Place an object into a receptacle	place brown chip bag into white bowl
Pick `Object` from `Receptacle` and Place on the Counter	162	Pick an object up from a location and then place it on the counter	pick green jalapeno chip bag from paper bowl and place on counter
Section 6.3 and 6.4 tasks	9	Skills trained for realistic, long instructions	open the large glass jar of pistachios pull napkin out of dispenser grab scooper
Total	744		

모델에 입력되는 형식

실제 입력 예시

Dataset Description

Everyday Robot

Brohan, Anthony, et al. "Rt-1: Robotics transformer for real-world control at scale." *arXiv preprint arXiv:2212.06817* (2022).

Robot Foundation Models

Robotics Transformer 1 (RT-1)

❖ Experiment Results

- RT-1의 학습 성능과 일반화 성능을 평가
  - ✓ Unseen Tasks는 학습에 사용된 적 없는 스킬과 물체를 가지고 지시하는 경우
  - ✓ Distractors는 지시사항에 없는 물체를 작업대에 진열하는 경우 (기본적으로 2~5개는 존재)
  - ✓ Backgrounds는 데이터에 없던 배경에서 작업을 수행하는 경우
- 비교 방법론에 비해서 더 좋은 일반화 성능을 가짐

목표 외 객체를 배치

본적 없는 배경으로 변경

처음보는 과제 수행

Model	Seen Tasks	Unseen Tasks	Distractors	Backgrounds
Gato (Reed et al., 2022)	65	52	43	35
BC-Z (Jang et al., 2021)	72	19	47	41
BC-Z XL	56	43	23	35
RT-1 (ours)	**97**	**76**	**83**	**59**

일반화 성능 평가
(성공률)

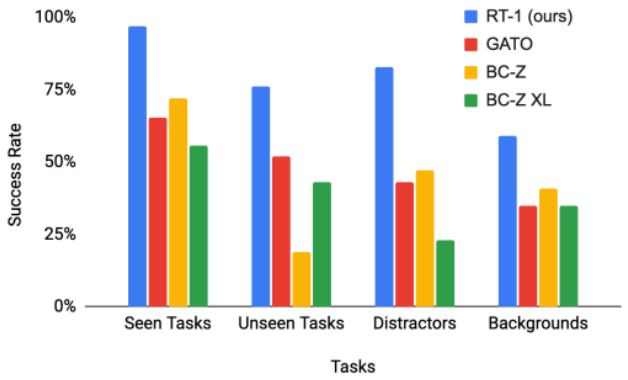
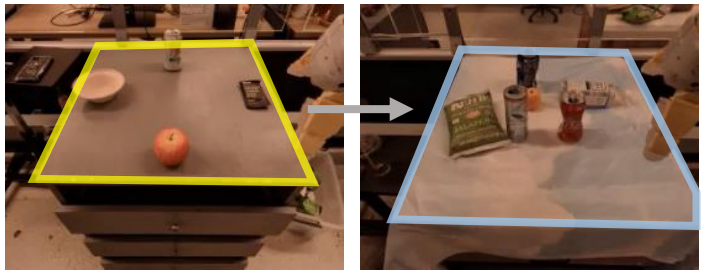


Table 2: Overall performance of RT-1 and baselines across seen tasks, generalization to unseen tasks, and robustness to distractors and backgrounds.

Placing Distracting Objects



Changing the Texture of the Table

Brohan, Anthony, et al. "Rt-1: Robotics transformer for real-world control at scale." *arXiv preprint arXiv:2212.06817* (2022).

Robot Foundation Models

Robotics Transformer 1 (RT-1)

❖ Experiment Results

- Ablation study를 통해 연속적인 행동 값을 이산적으로 표현하여 학습하는 것이 보다 효과적임을 보임

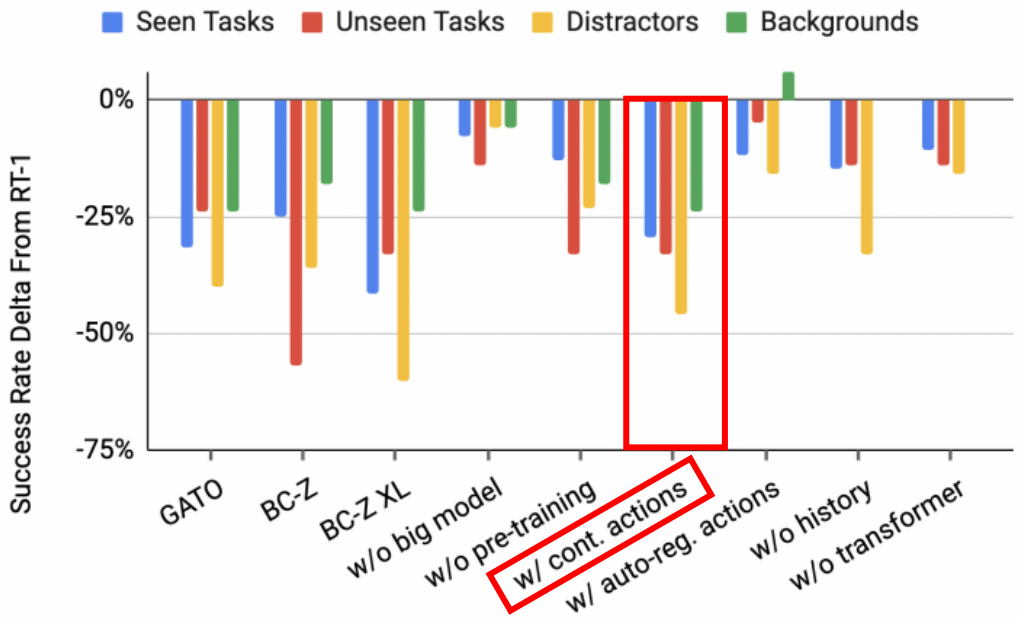


Table 13: Various model ablations of RT-1 across seen tasks, generalization to unseen tasks, and robustness to distractors and backgrounds.

Brohan, Anthony, et al. "Rt-1: Robotics transformer for real-world control at scale." *arXiv preprint arXiv:2212.06817* (2022).

Robot Foundation Models

Robotics Transformer 1 (RT-1)

❖ Experiment Results

- 형태가 다른 이종의 로봇 데이터를 사용하여도 성능 향상이 나타남 (단, 본래의 데이터와 함께 학습 해야함)
- KUKA 로봇으로 쓰레기를 집는(pick object) 과제를 수행한 데이터를 학습한 결과 RT-1에서도 해당 과제에 대한 성능이 향상됨

Models	Training Data	Classroom eval	Bin-picking eval
RT-1	Kuka bin-picking data + EDR data	90(-2)	**39(+17)**
RT-1	EDR only data	92	22
RT-1	Kuka bin-picking only data	0	0

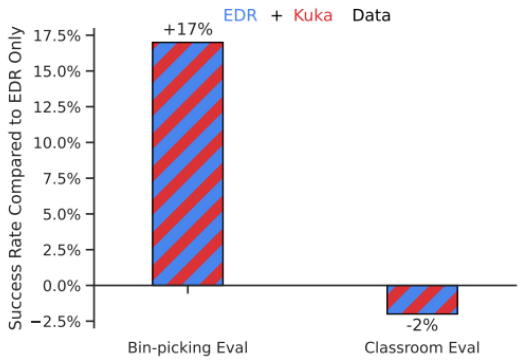


Table 5: Experimental results for mixing data from two different robots. Incorporating Kuka bin-picking data from QT-Opt (Kalashnikov et al., 2018) in RT-1 minimally impacts the standard classroom evaluation performance and results in almost a 2x improvement in generalization to the Bin-picking evaluation (that is similar to the setup in the Kuka data) on the Everyday Robots manipulator. This demonstrates an effective transfer across two different robot morphologies.

Kuka Robot

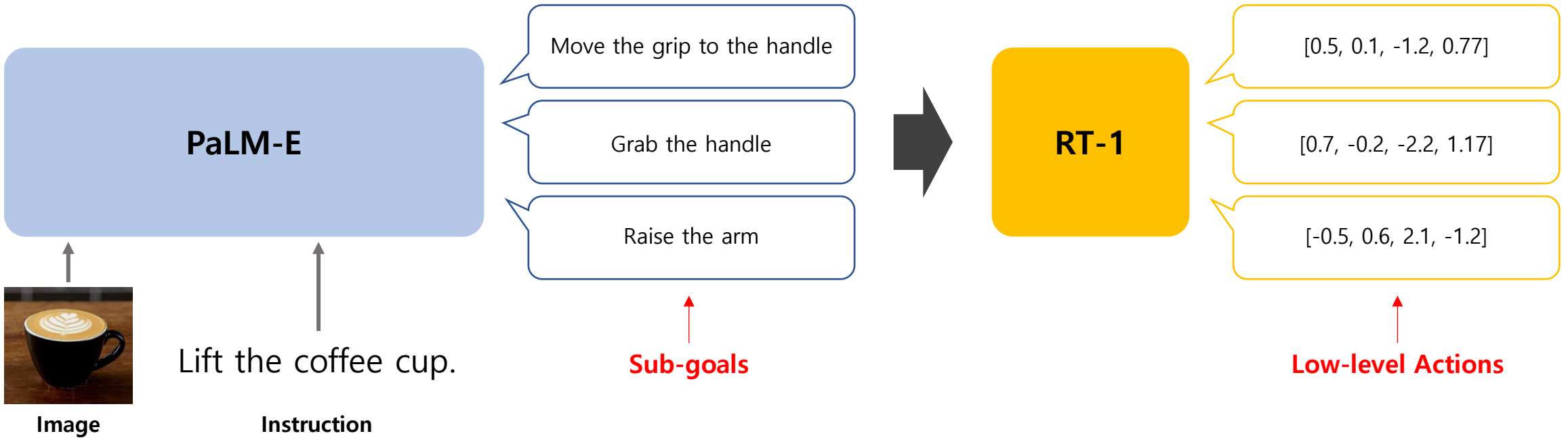
Brohan, Anthony, et al. "Rt-1: Robotics transformer for real-world control at scale." *arXiv preprint arXiv:2212.06817* (2022).
https://www.kuka.com/ko-kr/%EC%A0%9C%ED%92%88/%EB%A1%9C%EB%B4%87-%EC%8B%9C%EC%8A%A4%ED%85%9C/%EC%82%B0%EC%97%85%EC%9A%A9-%EB%A1%9C%EB%B4%87/lbr-iiwa

Robot Foundation Models

Pathways Language Model – Embodied (PaLM-E)

❖ PaLM-E: An Embodied Multimodal Language Model (ICML, 2023, 1630회 인용)

- 사전학습된 거대언어모델 PaLM에 이미지-텍스트, 제어 데이터를 기반으로 multi-modal 학습
- 현실 세계를 효과적으로 표현할 수 있도록 vision-language-model을 활용하여 로봇 제어 수행
- 다만 PaLM-E는 지시사항을 다수의 세부 절차로 나누어 주는 역할을 하며 실질적인 움직임은 RT-1이 수행

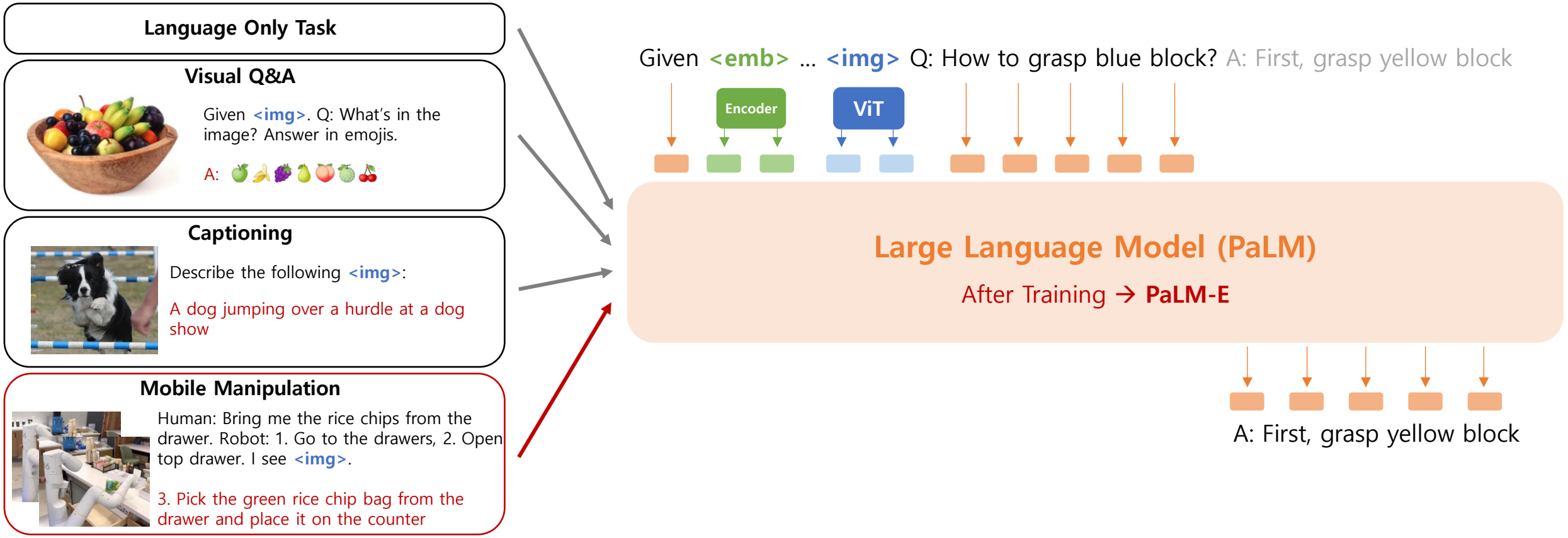


Robot Foundation Models

Pathways Language Model – Embodied (PaLM-E)

❖ Co-training Strategy

- 사전학습된 PaLM (8B, 62B, 540B)에 이미지, 텍스트, 로봇 조종이 유기적으로 연관된 데이터를 학습하여 multi-modal large language model 생성



Driess, Danny, et al. "PaLM-E: an embodied multimodal language model." Proceedings of the 40th International Conference on Machine Learning. 2023.

Robot Foundation Models

Pathways Language Model – Embodied (PaLM-E)

❖ Experiment Settings (dataset & robot)

- 제시된 데이터 조합으로 학습된 PaLM-E (8B, 62B, 540B)를 high level controller로 사용, low level control은 RT-1을 사용
- 클라우드 시스템에서 지시사항을 추론 한 뒤 EverydayRobot에 전송하는 방식으로 성능 검증을 수행

Dataset in full mixture	Sampling frequency	%
Webli (Chen et al., 2022)	100	52.4
VQ^{2}A (Changpinyo et al., 2022)	25	13.1
VQG (Changpinyo et al., 2022)	10	5.2
CC3M (Sharma et al., 2018)	25	13.1
Object Aware (Piergiovanni et al., 2022)	10	5.2
OKVQA (Marino et al., 2019)	1	0.5
VQAv2 (Goyal et al., 2017)	1	0.5
COCO (Chen et al., 2015)	1	0.5
Wikipedia text	1	0.5
(robot) Mobile Manipulator, real 주방에서의 과제를 수행	6	3.1
(robot) Language Table (Lynch et al., 2022), sim and real 물체를 밀어 이동하는 법	8	4.2
(robot) TAMP, sim 물체를 쌓는 법	3	1.6

Dataset Description

Everyday Robot

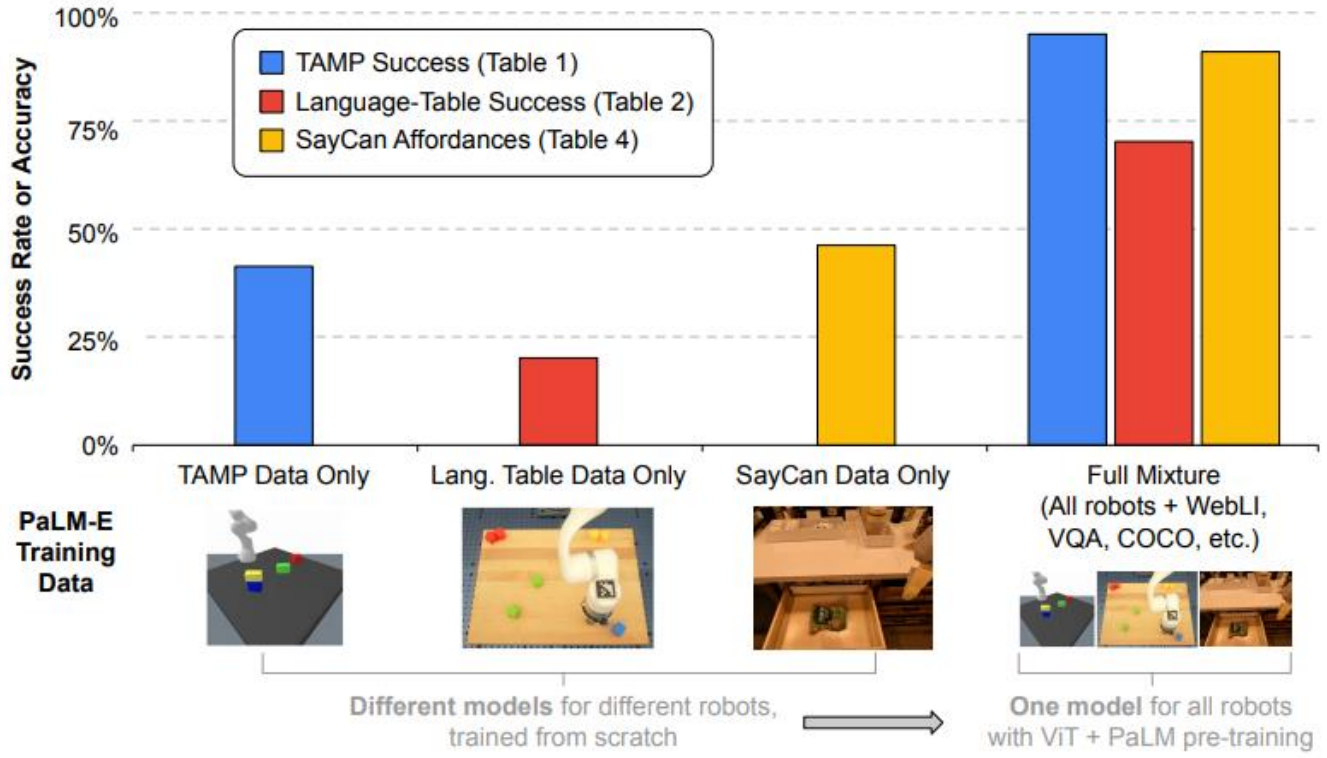
Driess, Danny, et al. "PaLM-E: an embodied multimodal language model." Proceedings of the 40th International Conference on Machine Learning. 2023.

Robot Foundation Models

Pathways Language Model – Embodied (PaLM-E)

❖ Experiment Results

- 로봇 조종 과제에 있어서 이미지와 텍스트 데이터를 함께 학습하는 것이 성능 향상에 더 도움이 됨



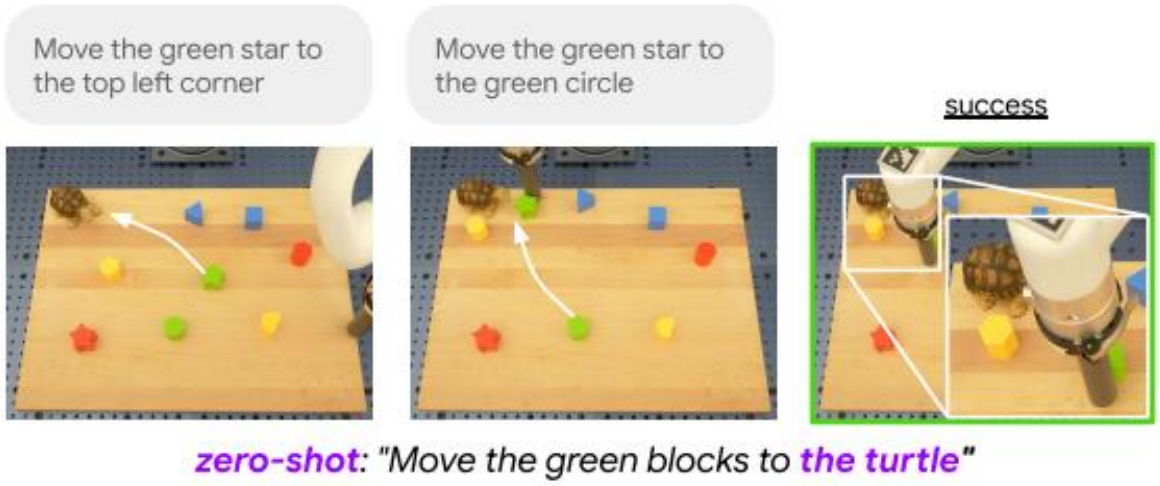
Driess, Danny, et al. "PaLM-E: an embodied multimodal language model." Proceedings of the 40th International Conference on Machine Learning. 2023.

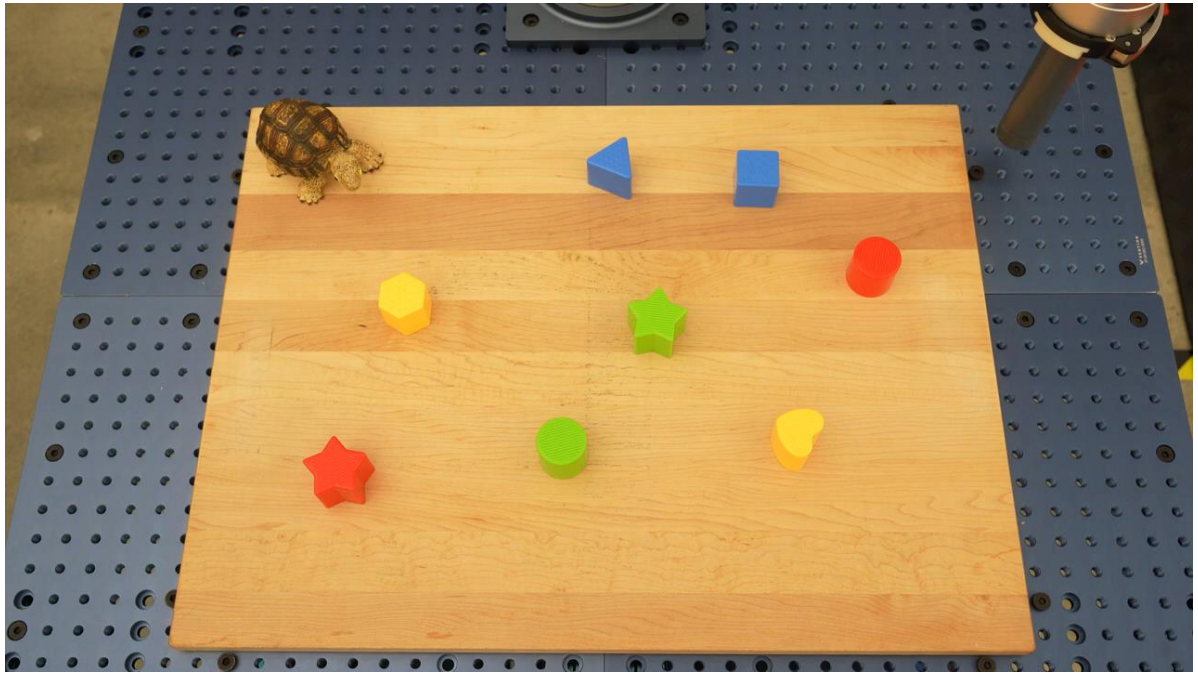
Robot Foundation Models

Pathways Language Model – Embodied (PaLM-E)

❖ Experiment Results

- Language Table 환경에서 PaLM-E의 일반화 성능을 검증: Robot 데이터에 없던 물체가 있을 때 + 환경에 갑작스런 인위적인 변화를 줄 때
- Robot manipulation 학습 데이터에는 '거북이'가 전혀 없었으나 잘 인지해서 주어진 지시사항을 성공적으로 수행





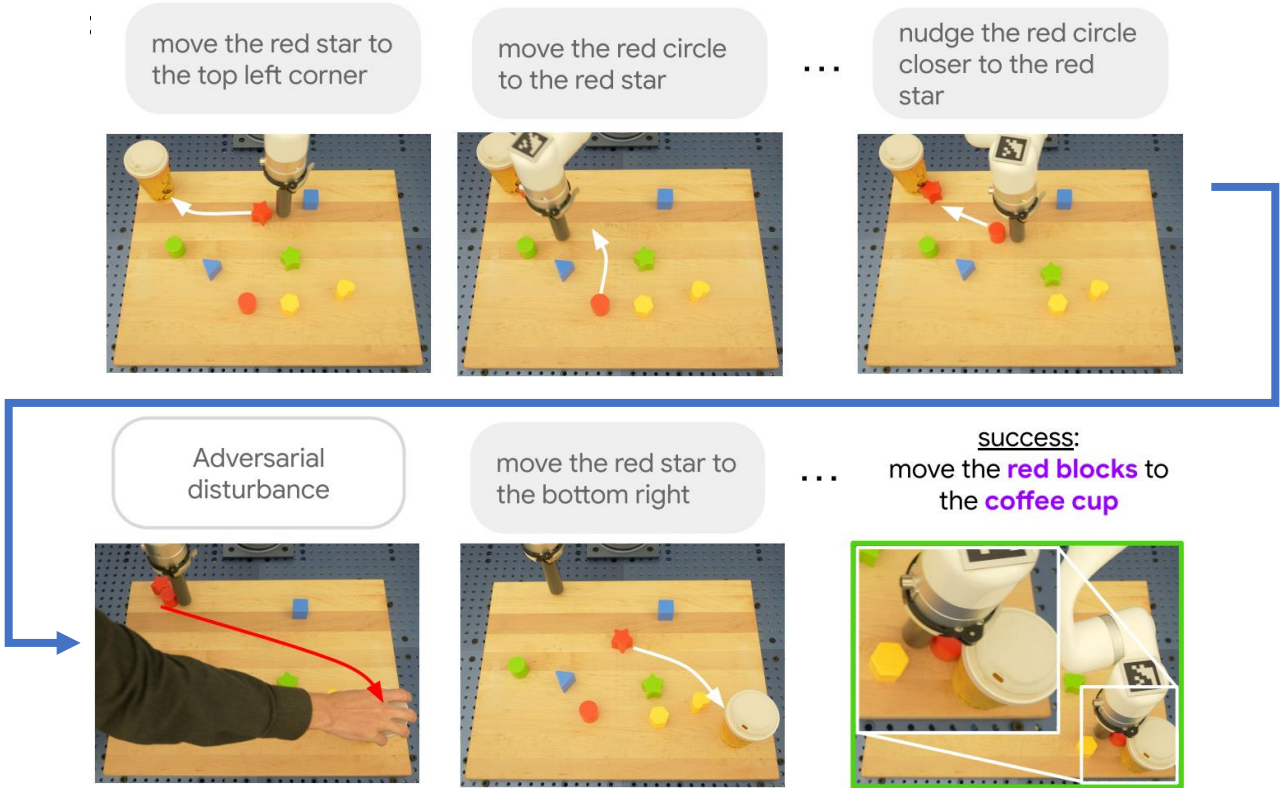
Driess, Danny, et al. "PaLM-E: an embodied multimodal language model." Proceedings of the 40th International Conference on Machine Learning. 2023.

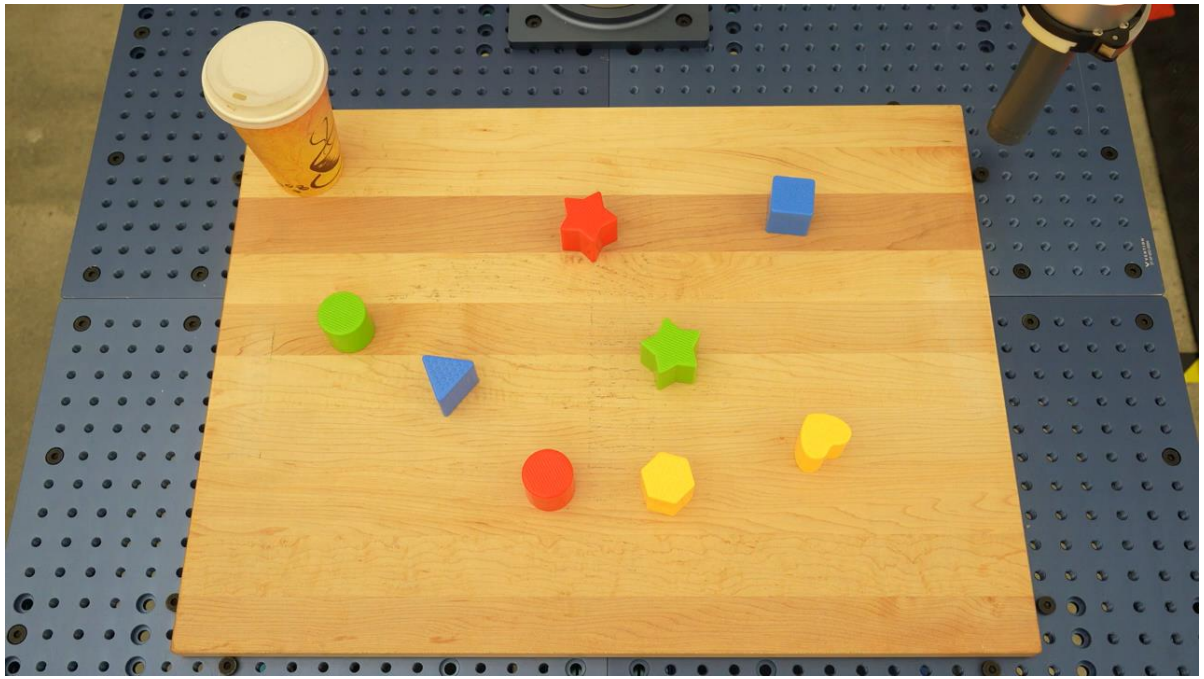
Robot Foundation Models

Pathways Language Model – Embodied (PaLM-E)

❖ Experiment Results

- Language Table 환경에서 PaLM-E의 일반화 성능을 검증: Robot 데이터에 없던 물체 조합이 있을 때 + 환경에 갑작스런 인위적인 변화를 줄 때
- 학습 시 red block과 coffee cup이 하나의 과제 안에 있는 경우가 없었으나 성공적으로 수행





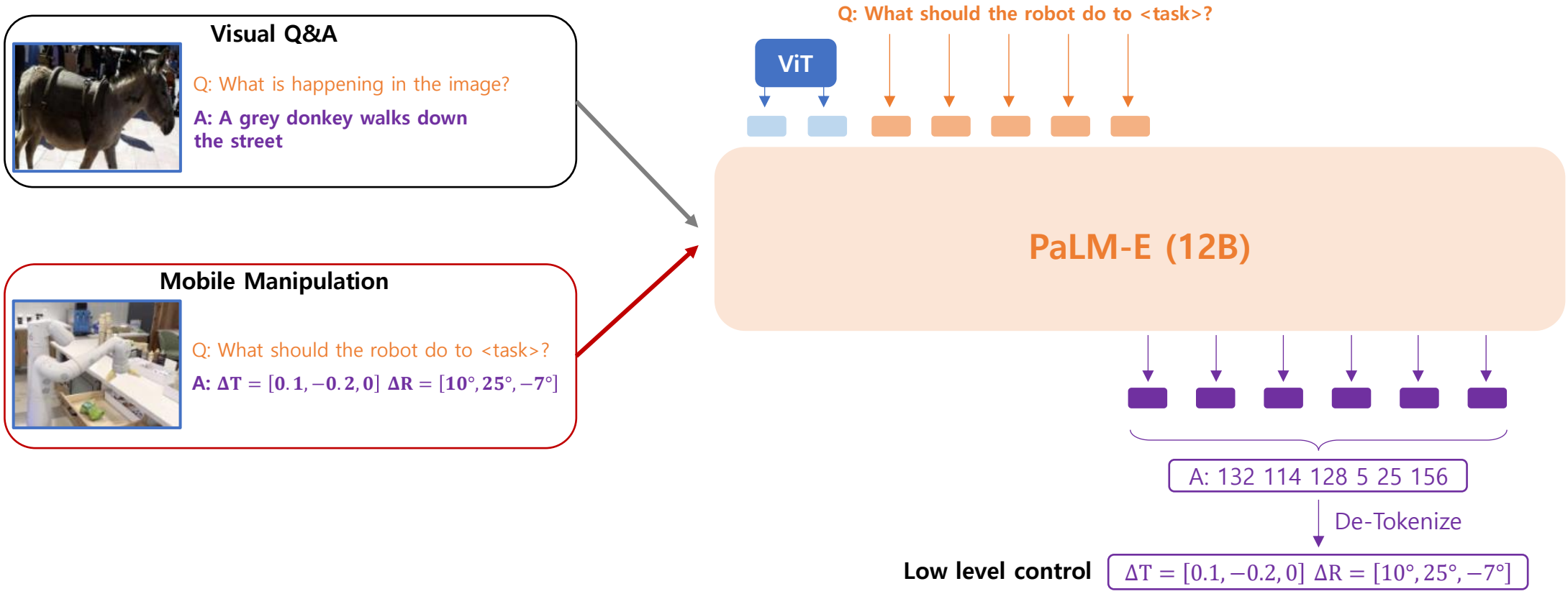
Driess, Danny, et al. "PaLM-E: an embodied multimodal language model." Proceedings of the 40th International Conference on Machine Learning. 2023.

Robot Foundation Models

Robotics Transformer 2 (RT-2)

❖ **RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control (Google DeepMind, 2023, 781회 인용)**

- PaLM-E (또는 PaLI-X)가 sub-goals를 출력하는 것이 아닌 RT-1처럼 로봇 제어에 필요한 값을 직접 출력하도록 학습하는 방법론
- PaLM-E에서 진행한 co-training strategy를 수행하되 이번 manipulation 데이터는 정답에 low level action을 사용함



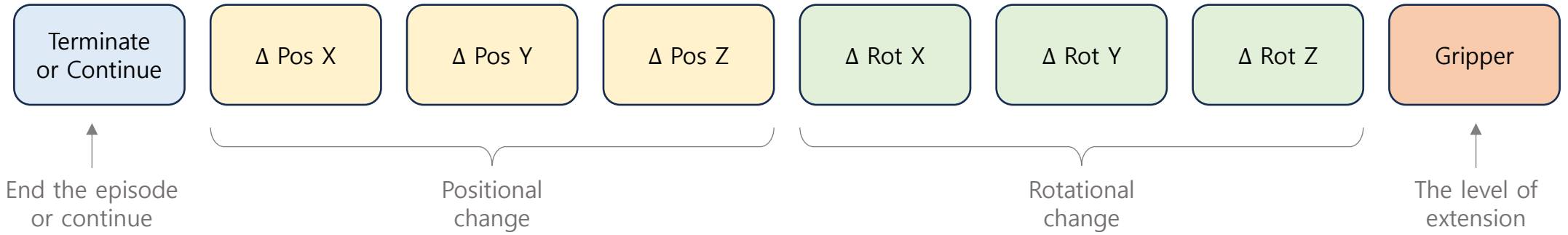
Brohan, Anthony, et al. "Rt-2: Vision-language-action models transfer web knowledge to robotic control." arXiv preprint arXiv:2307.15818 (2023).

Robot Foundation Models

Robotics Transformer 2 (RT-2)

❖ Action tokens for RT-2

- RT-1과 동일하게 연속적인 값을 이산적인 값으로 표현 (action tokens)
- 실험에서 단일 로봇 기종(Everyday Robot)을 쓰기 때문에 action tokens 길이가 고정되어 있음
- 다만, RT-1과 달리 사용하려는 vision-language model의 토큰을 사용해야 하므로 수정이 필요
  → PaLM-E를 쓰는 경우 토큰 중 잘 쓰이지 않는 토큰을 action token으로 변경하여 사용
  → PaLI-X를 쓰는 경우 자체적으로 숫자를 표현하는 토큰이 존재하여 변경 X



Robot Foundation Models

Robotics Transformer 2 (RT-2)

❖ Experiment Settings (dataset & robot)

- 17개월 동안 13대의 로봇(EverydayRobot)으로 현실 세계에서 수집한 13만 여개의 동작 데이터(총 744개의 과제)를 사용
- EverydayRobot에 학습한 정책을 탑재하여 성능 검증을 수행

Skill	Count	Description	Example Instruction
Pick Object	130	Lift the object off the surface	pick iced tea can
Move Object Near Object	337	Move the first object near the second	move pepsi can near rxbar blueberry
Place Object Upright	8	Place an elongated object upright	place water bottle upright
Knock Object Over	8	Knock an elongated object over	knock redbull can over
Open Drawer	3	Open any of the cabinet drawers	open the top drawer
Close Drawer	3	Close any of the cabinet drawers	close the middle drawer
Place Object into Receptacle	84	Place an object into a receptacle	place brown chip bag into white bowl
Pick Object from Receptacle and Place on the Counter	162	Pick an object up from a location and then place it on the counter	pick green jalapeno chip bag from paper bowl and place on counter
Section 6.3 and 6.4 tasks	9	Skills trained for realistic, long instructions	open the large glass jar of pistachios pull napkin out of dispenser grab scooper
Total	744		

모델에 입력되는 형식

실제 입력 예시

Dataset Description

Everyday Robot

Brohan, Anthony, et al. "Rt-1: Robotics transformer for real-world control at scale." *arXiv preprint arXiv:2212.06817* (2022).

Robot Foundation Models

Robotics Transformer 2 (RT-2)

❖ Experiment Results

- 학습한 상황에서는 성능이 조금 높거나 비슷한 경향을 보임
- 하지만 학습 데이터에 없던 객체, 배경, 환경이 주어진 상황에서는 RT-2가 더 좋은 일반화 성능을 보임

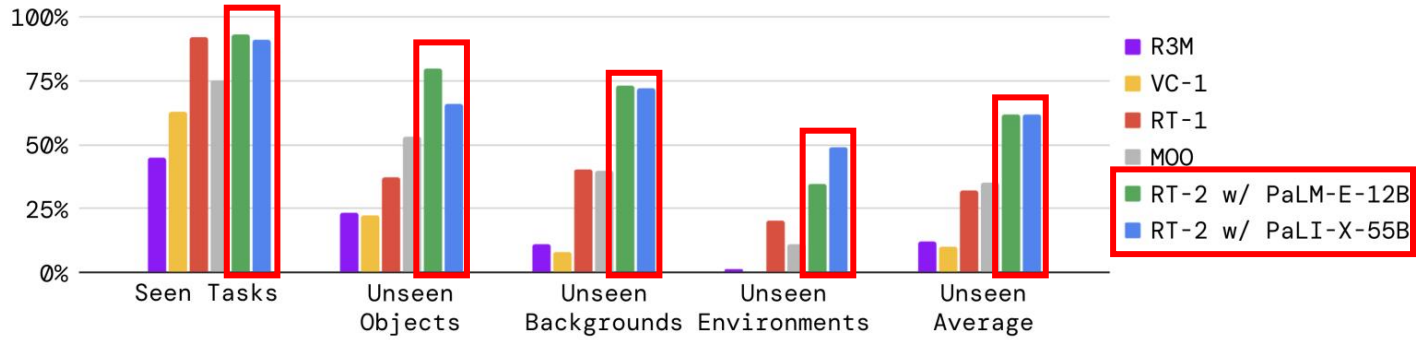
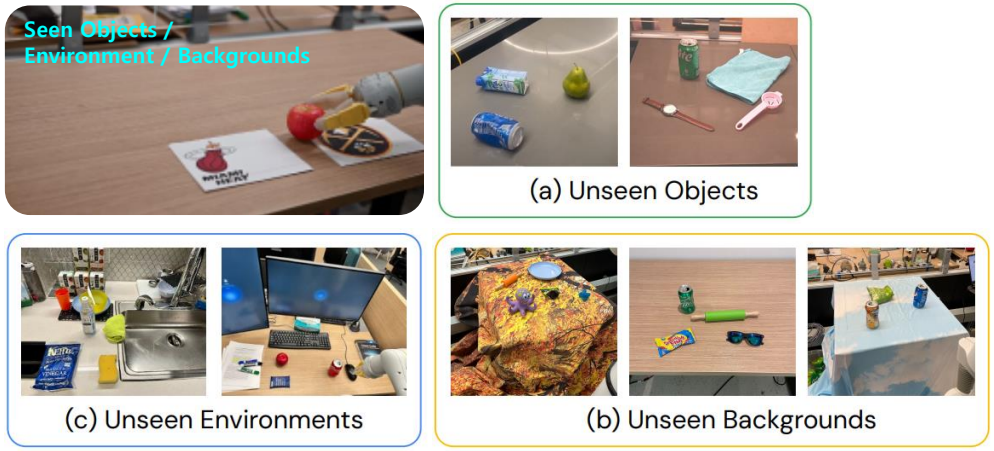


Figure 4 | Overall performance of two instantiations of RT-2 and baselines across seen training tasks as well as unseen evaluations measuring generalization to novel objects, novel backgrounds, and novel environments. Appendix Table 4 details the full results.



Brohan, Anthony, et al. "Rt-2: Vision-language-action models transfer web knowledge to robotic control." arXiv preprint arXiv:2307.15818 (2023).

Robot Foundation Models

Robotics Transformer 2 (RT-2)

❖ Experiment Results

- 학습 과정에서 데이터에는 없었던 새로운 스킬을 찾아내는지 실험을 수행
- 학습 데이터에는 미는(push) 지시사항 존재하지 않으니 해당 지시를 내렸을 때 올바르게 수행하는 것을 확인
- 따라서 웹 규모의 사전 훈련으로 얻은 지식이 제어 관련 스킬과 연결되어 새로운 스킬을 획득할 수 있음

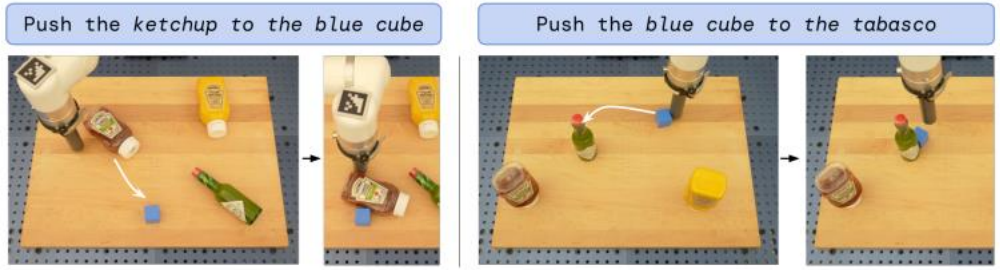


Figure 5 | Real-world out-of-distribution behaviors in the Language Table environment. Identical RT-2-PaLI-3B model checkpoint is used as in Tab. 1.

Model	Language-Table
BC-Zero (Jang et al., 2021)	72 ± 3
RT-1 (Brohan et al., 2022)	74 ± 13
LAVA (Lynch et al., 2022)	77 ± 4
RT-2-PaLI-3B (ours)	**90 ± 10**

Table 1 | Performance on the simulated Language-Table tasks (Lynch and Sermanet, 2020).

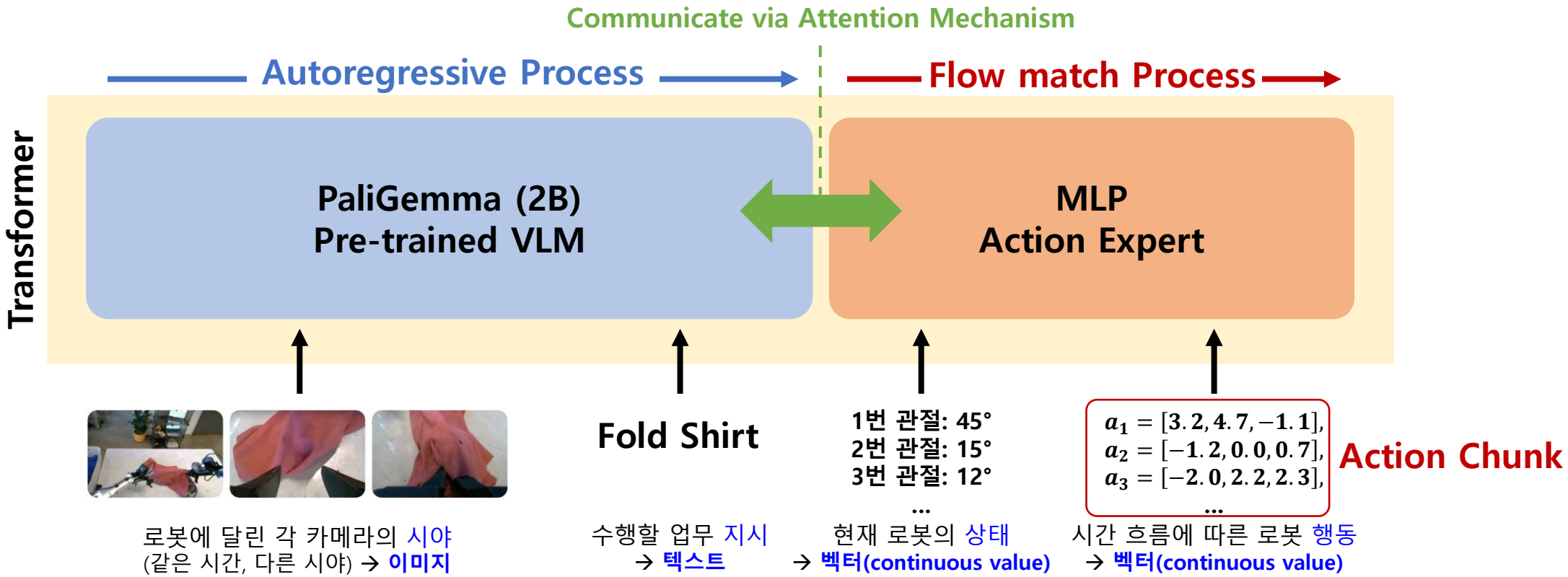
Brohan, Anthony, et al. "Rt-2: Vision-language-action models transfer web knowledge to robotic control." arXiv preprint arXiv:2307.15818 (2023).

Robot Foundation Models

$\pi 0$: A Vision-Language-Action Flow Model for General Robot Control

❖ $\pi 0$: A Vision-Language-Action Flow Model for General Robot Control (Physical Intelligence, 2024, 15회 인용)

- 추론 속도를 높이고 성능 향상을 위해서 최신 방법론을 robot learning에 알맞게 조합한 모델 (Transfusion + MoE + ALOHA + FlowMatch)
- 결과적으로 더 적은 파라미터를 사용하고도 복잡한 업무를 수행하고 동시에 빠른 추론 속도를 유지함



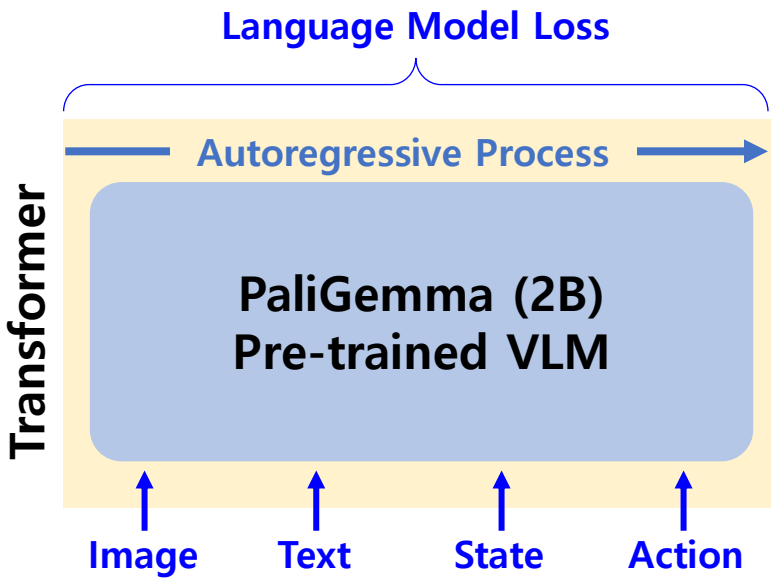
O'Neill, Abby, et al. "Open X-Embodiment: Robotic Learning Datasets and RT-X Models : Open X-Embodiment Collaboration0." IEEE International Conference on Robotics and Automation (ICRA). 2024.

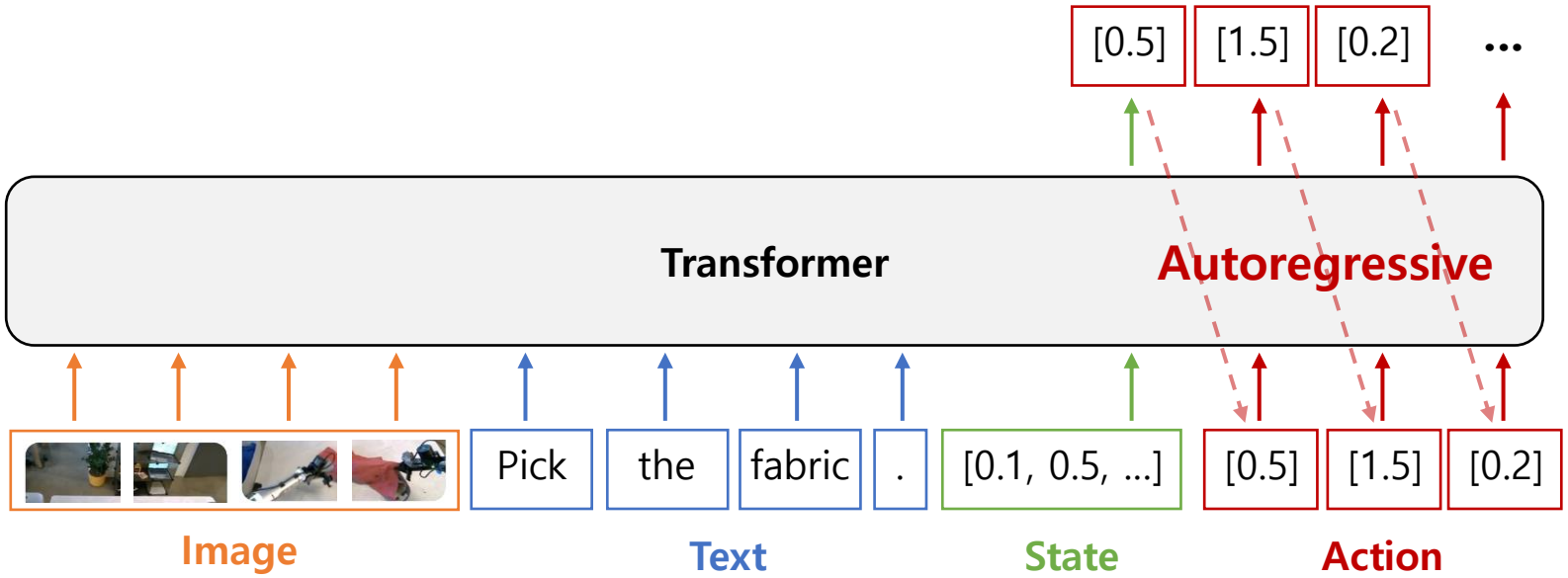
Robot Foundation Models

$\pi 0$: A Vision-Language-Action Flow Model for General Robot Control

❖ **How $\pi 0$ developed its architecture?: Step-by-step approach**

- 기존 robot foundation model은 공통적으로 vision-language-model의 autoregressive한 과정을 그대로 사용
- **$\pi 0$**는 이를 어떻게 변형했는지 차근차근 알아보자



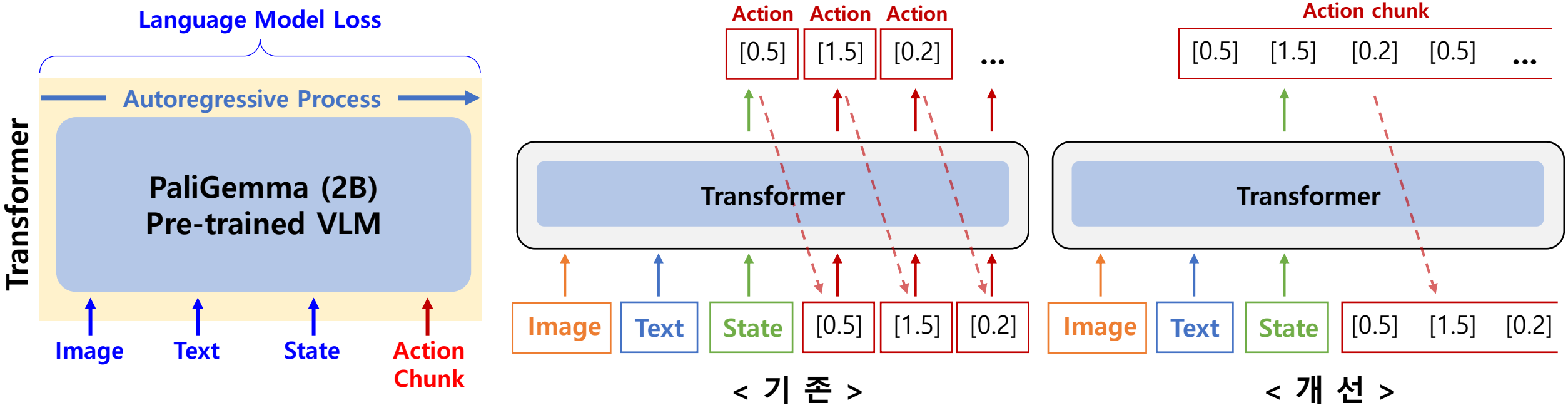


Robot Foundation Models

$\pi 0$: A Vision-Language-Action Flow Model for General Robot Control

❖ $\pi 0$: Action chunk

- $\pi 0$는 한 번에 하나의 행동을 예측하는 것이 아닌 여러 개의 행동(action chunk)을 동시에 예측함 (action chunk = $\{action_1, \ldots, action_{50}\}$)
- 해당 방법을 사용하여 작동까지 소요되는 추론 시간이 감소함

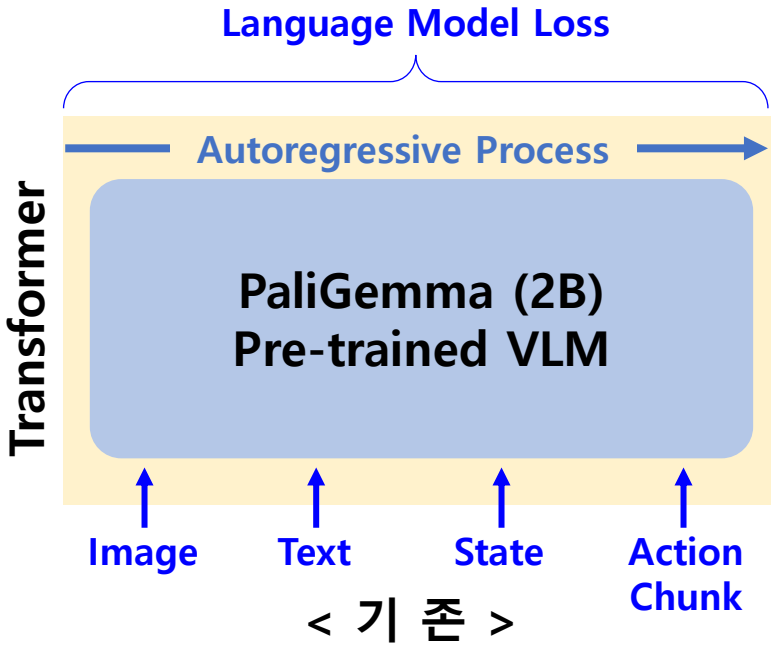


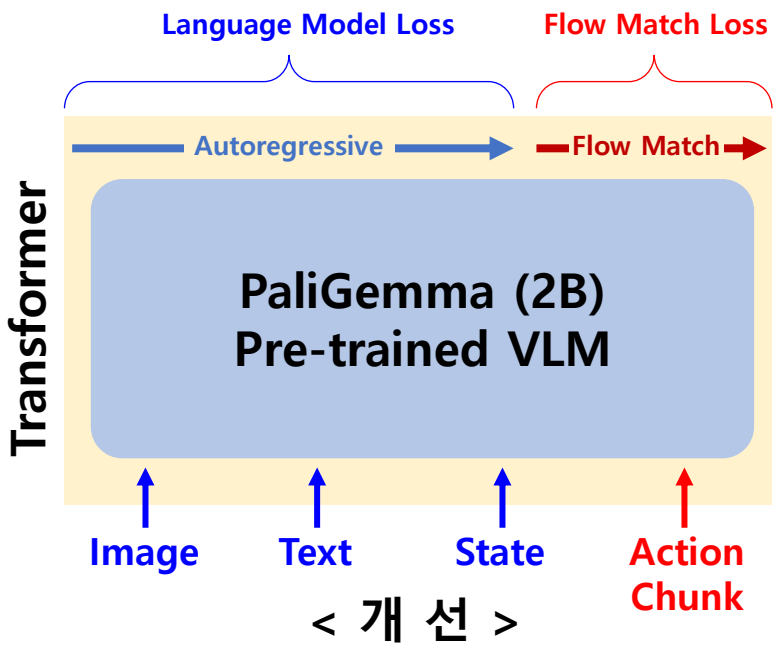
Robot Foundation Models

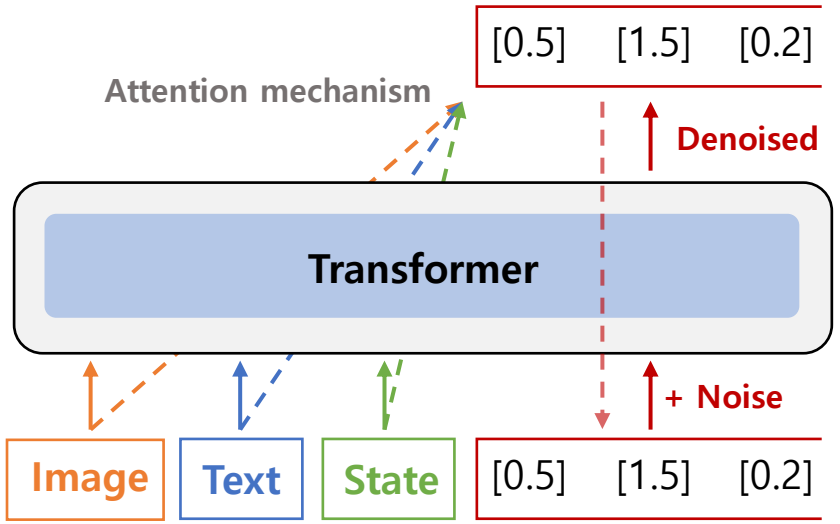
$\pi 0$: A Vision-Language-Action Flow Model for General Robot Control

❖ $\pi 0$: Using different process to different modality

- $\pi 0$는 Transfusion에서 제안한 multimodal 구조를 활용하여 생성할 때 autoregressive 방식과 flow match 방식(diffusion의 변형)을 함께 수행
- 각 modality에 적합한 방식을 적용하여 모델의 성능을 향상 (최근 continuous action 학습에서 diffusion 기반의 모델이 보다 뛰어난 성능을 보임)





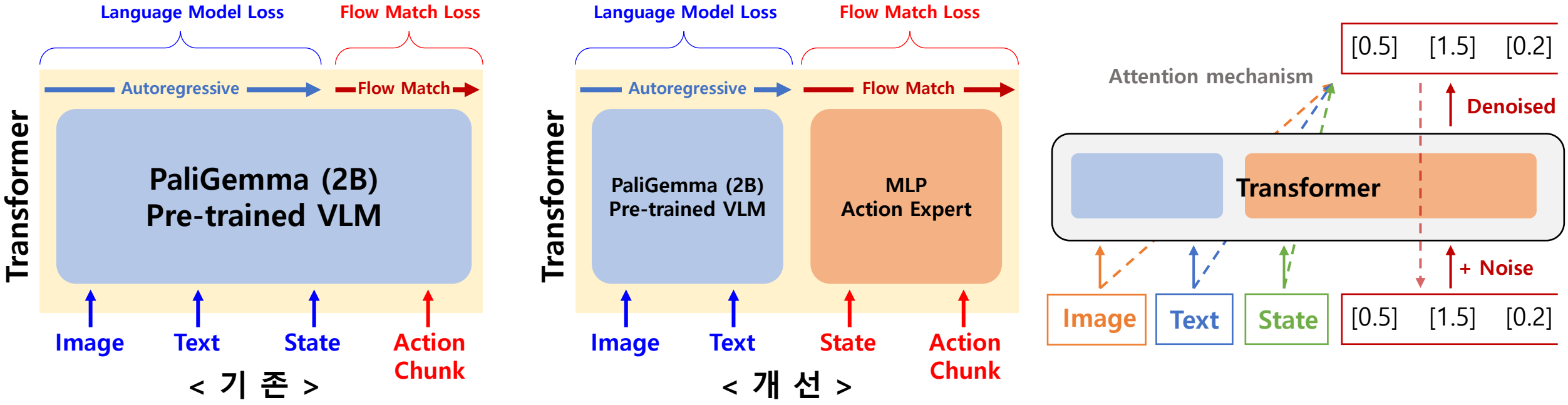


Robot Foundation Models

$\pi 0$: A Vision-Language-Action Flow Model for General Robot Control

❖ $\pi 0$: Mixture-of-Expert

- $\pi 0$는 Mixture-of-Expert에서 제안한 한 구조 내 여러 가중치 집합(expert)을 활용하여 각 가중치가 특화된 역할을 학습함
- VLM은 사전학습에 쓰인 modality인 이미지와 텍스트만을 처리하며 새롭게 추가한 action expert는 로봇의 관절 상태와 행동 값을 처리함
  - VLM에서 발생할 수 있는 distribution shift 문제를 완화 → 모델의 성능 향상
  - Action expert는 VLM보다 크기가 작음 → 모델의 연산 효율성 증가

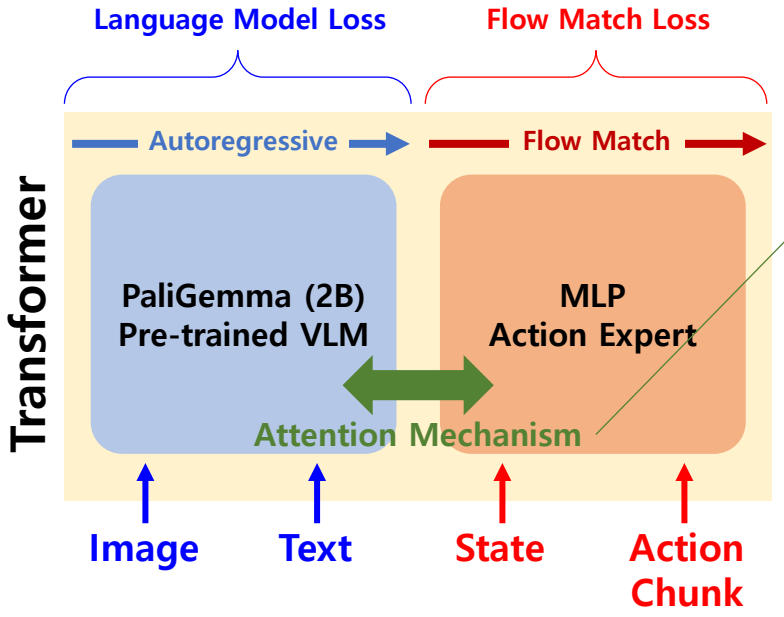


Robot Foundation Models

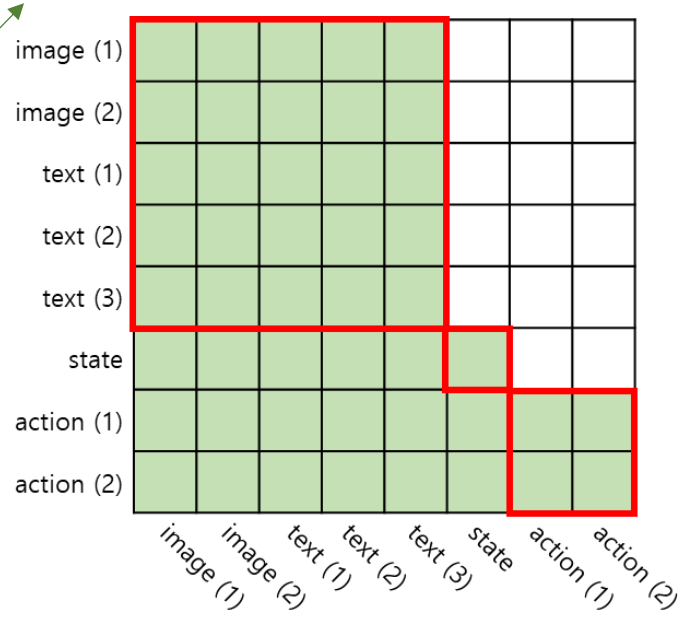
$\pi 0$: A Vision-Language-Action Flow Model for General Robot Control

❖ $\pi 0$: Blockwise causal attention mask

- $\pi 0$는 blockwise causal attention mask를 사용하여 attention을 수행
- 같은 블록 내에서는 bidirectional self-attention이 이루어지며 블록 간에는 causal self-attention이 수행됨
- 이를 통해서 VLM이 image-text 데이터를 표현할 때 다른 modality로부터 방해 받지 않고 기존의 사전학습된 지식을 온전히 활용할 수 있음
- 반면 action을 생성할 때에는 모든 정보를 참고할 수 있음



<Blockwise Causal Attention Mask>



Robot Foundation Models

π0: A Vision-Language-Action Flow Model for General Robot Control

❖ Experiment Settings (dataset)

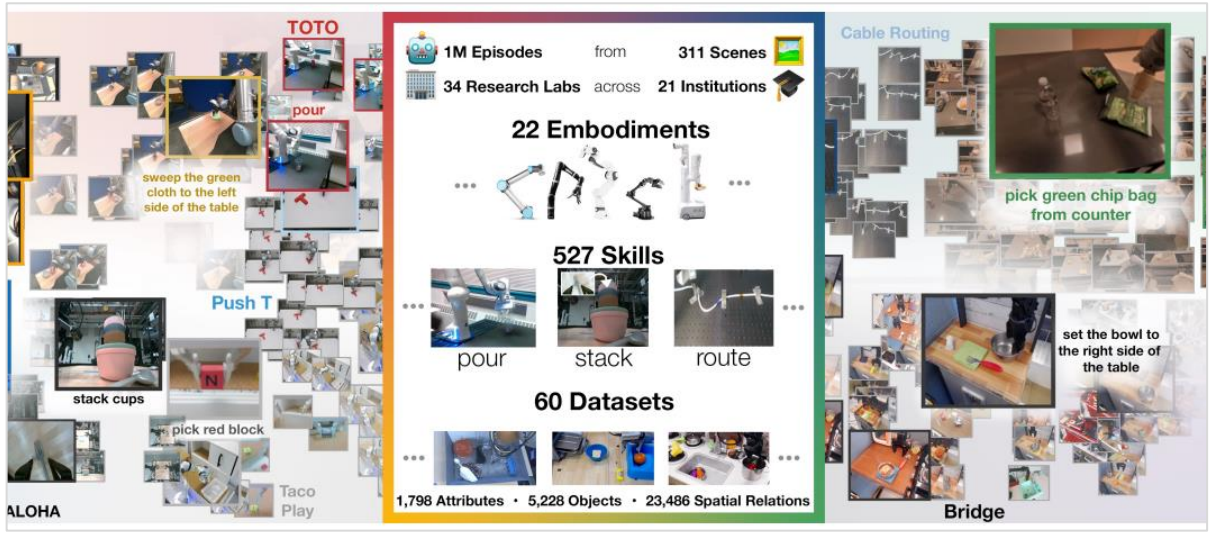
- Pre-training은 모델이 제어 능력을 배우고 일반화 능력을 갖추는 것을 목표로 함

  → OXE를 일부 추출하여 만든 OXE Magic Soup 데이터셋과 π Cross-Embodiment 데이터셋을 사용하여 다양한 작업과 로봇을 경험

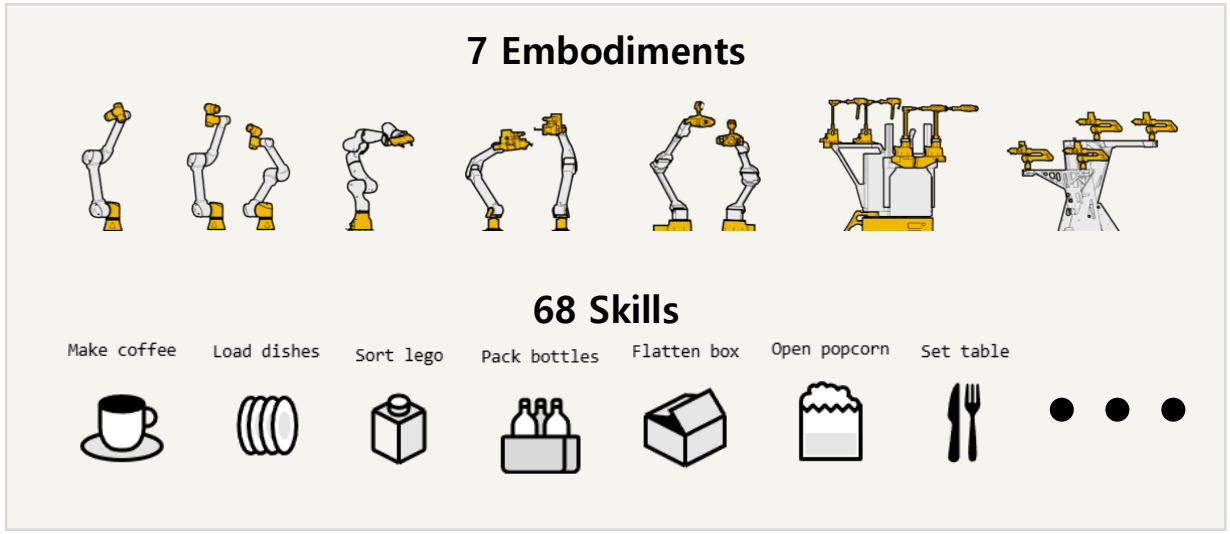
  → 다양한 환경과 작업에 적응할 수 있는 능력을 키워 이후 새로운 로봇이나 작업에 더 쉽게 적용할 수 있게 함

- Post-training은 특정 작업에 모델을 특화하는 것을 목표로 함 (=fine-tuning)

  → 사람이 직접 선별한 소량의 고품질의 데이터를 사용하여 진행



Open X-Embodiment (OXE) Dataset



π Cross-Embodiment Robot Dataset

https://www.physicalintelligence.company/blog/pi0

O'Neill, Abby, et al. "Open X-Embodiment: Robotic Learning Datasets and RT-X Models : Open X-Embodiment Collaboration0." IEEE International Conference on Robotics and Automation (ICRA). 2024.

Robot Foundation Models

$\pi 0$: A Vision-Language-Action Flow Model for General Robot Control

❖ Experiment Settings (robots)

- 모델의 성능 평가를 위해서 π Cross-Embodiment 데이터셋 수집에 사용했던 7가지 로봇들을 사용

Variety of Robot Types

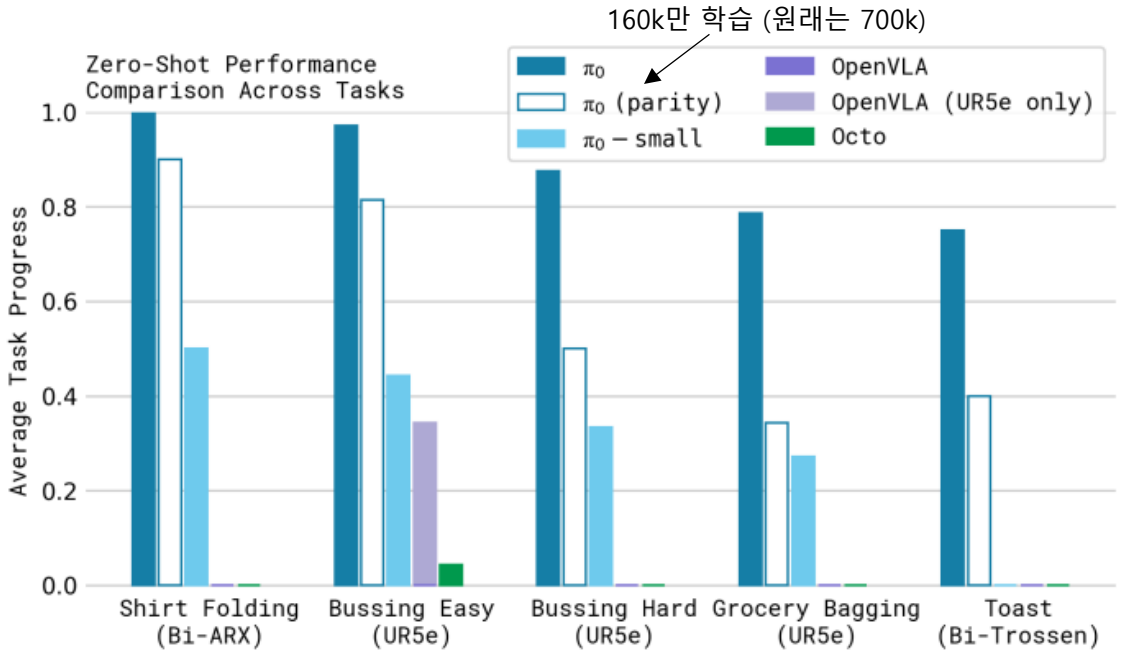
Black, Kevin, et al. "$₩pi_0 $: A Vision-Language-Action Flow Model for General Robot Control." arXiv preprint arXiv:2410.24164 (2024).

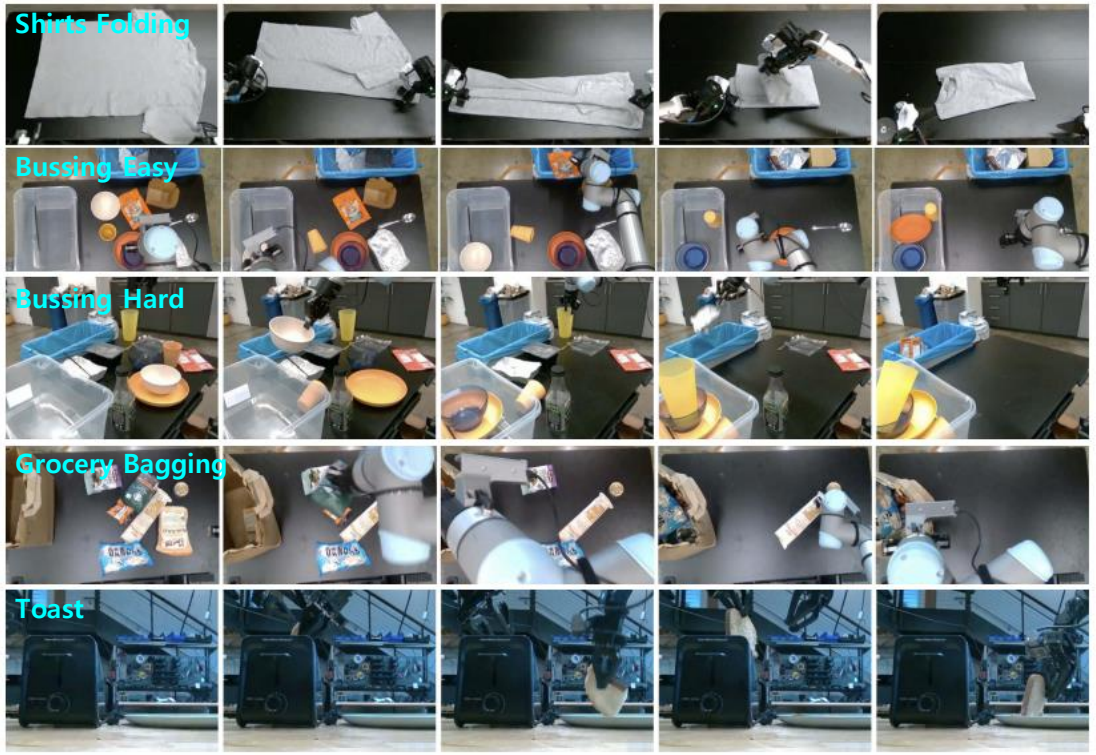
Robot Foundation Models

$\pi 0$: A Vision-Language-Action Flow Model for General Robot Control

❖ Experiment Results

- 주어진 다섯 가지 과제에 대해서 pre-training을 진행한 뒤 각각에 대한 post-training 없이 zero-shot 검증 성능을 측정
- 특히 Parity 모델은 베이스라인 모델인 OpenVLA, Octo와 동일한 학습 스텝을 설정하여 학습한 경우지만 여전히 보다 높은 성능을 달성함





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

Robot Foundation Models

$\pi 0$: A Vision-Language-Action Flow Model for General Robot Control

❖ Experiment Results

- Post-training (=fine-tuning) 성능에 대한 비교 실험을 수행
- 사전학습을 수행한 모델 뿐만 아니라 처음부터 학습하는 모델도 베이스라인 보다 나은 성능을 보여줌

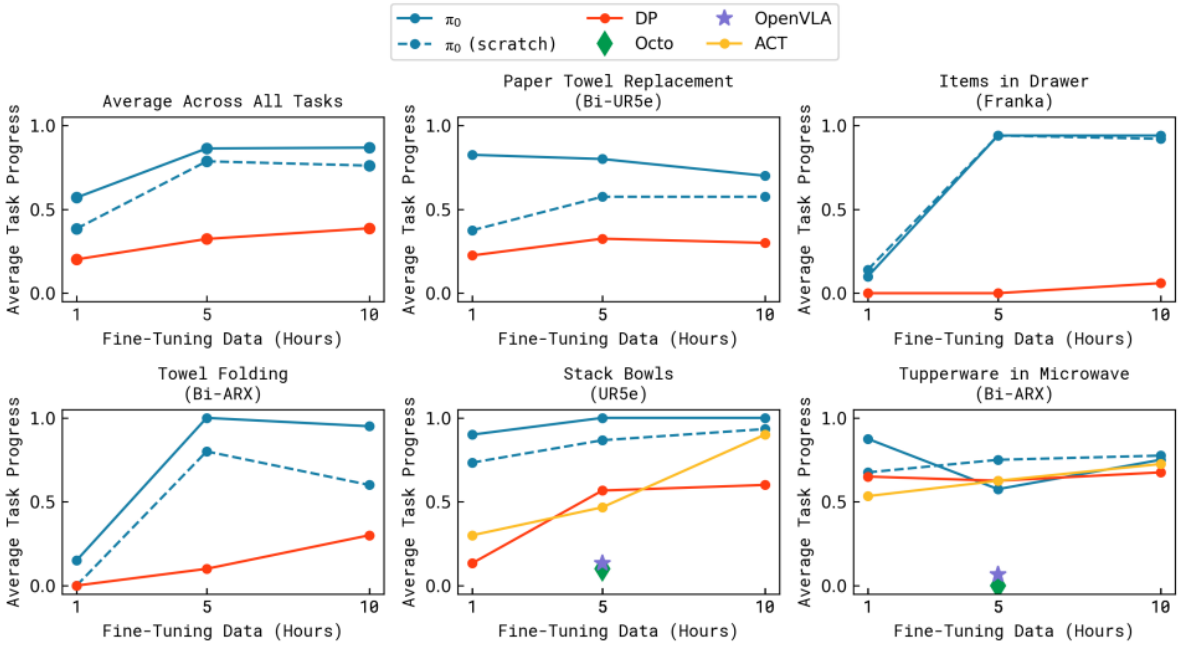
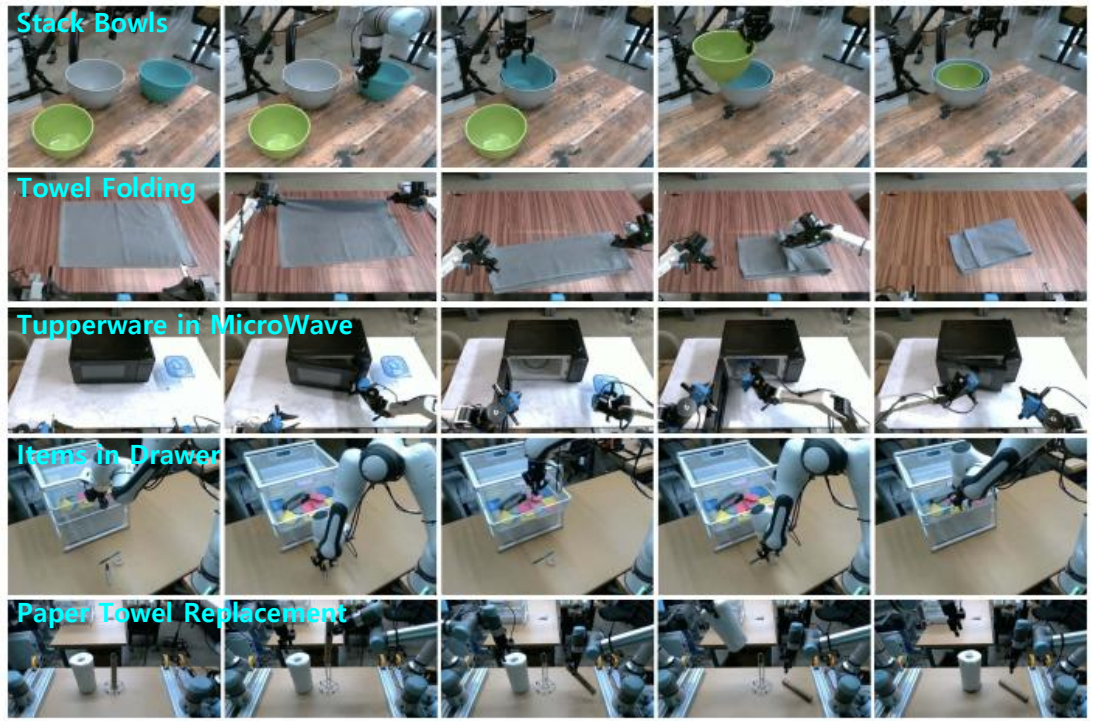


Fig. 11: **Fine-tuning with varying amounts of data.** π_0 can learn some easier tasks even with smaller amounts of data, and the pre-trained model often attains a larger improvement over the model trained from scratch.



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

Summary

❖ Training generalist robot policy using vision-language model

- 로봇 제어 모델의 핵심은 일반화 성능과 추론 속도
- RT-1은 대량의 데이터셋과 큰 규모의 모델을 사용하면 제어 모델이 좋은 일반화 성능을 확보함을 확인,
  또한 이종 로봇 간의 데이터셋도 학습 효과적임을 확인함
- PaLM-E는 multimodal large language model을 활용해서 제어 관련 데이터가 거대하지 않더라도
  제어 모델의 일반화 성능을 대폭 향상할 수 있음을 확인
- RT-2는 PaLM-E과 달리 vison-language model이 직접 제어모델로 작동하도록 방법론을 제안
- $\pi 0$는 사전학습된 vision-language model의 지식을 보다 잘 활용할 수 있도록 제어 모델에 적합하게 개선한 아키텍처를 제시

Citation

- Brohan, Anthony, et al. "Rt-1: Robotics transformer for real-world control at scale." *arXiv preprint arXiv:2212.06817* (2022).
- Driess, Danny, et al. "PaLM-E: an embodied multimodal language model." *Proceedings of the 40th International Conference on Machine Learning*. 2023.
- Brohan, Anthony, et al. "Rt-2: Vision-language-action models transfer web knowledge to robotic control." *arXiv preprint arXiv:2307.15818* (2023).
- Black, Kevin, et al. "$\pi_0 $: A Vision-Language-Action Flow Model for General Robot Control." *arXiv preprint arXiv:2410.24164* (2024).
- Tony Z. Zhao AND Vikash Kumar AND Sergey Levine AND Chelsea Finn, . "Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware." Proceedings of Robotics: Science and Systems. 2023.
- Zhou, Chunting, et al. "Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model." *CoRR* (2024).
- Shazeer, Noam, et al. "Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer." *International Conference on Learning Representations*. 2016.
- Perez, Ethan, et al. "Film: Visual reasoning with a general conditioning layer." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. No. 1. 2018.
- O'Neill, Abby, et al. "Open X-Embodiment: Robotic Learning Datasets and RT-X Models : Open X-Embodiment Collaboration0." *IEEE International Conference on Robotics and Automation (ICRA).* 2024.