AutoRT: Embodied Foundation Models for Large Scale Orchestration of Robotic Agents

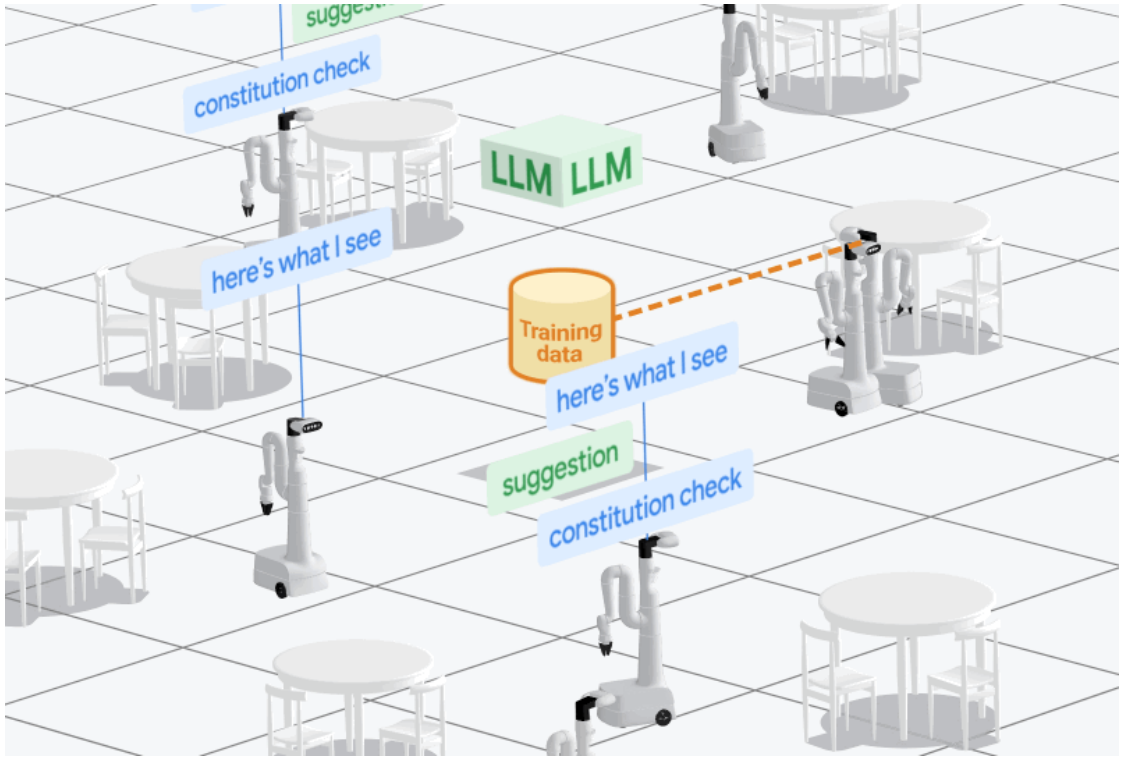
arXiv:2401.12963v1 [cs.RO] 23 Jan 2024

지도 교수: 신승태
학생: 이해찬
학번: 2373539

AUTORT 소개

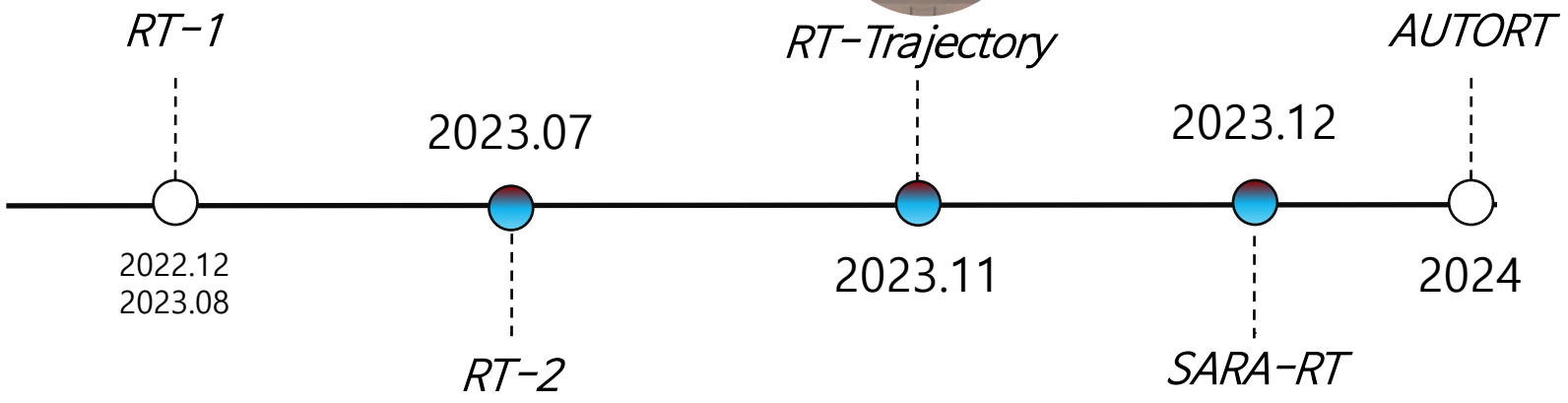
✓ AUTORT 기능

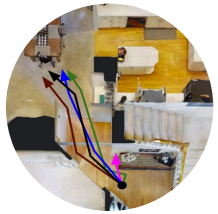
- LLM 및 VLM 기반의 **인간의 도움**을 받아 실제 세계에서 로봇s에게 지시하여 자율적으로 데이터를 수집하는 접근 방식



A timeline of RT

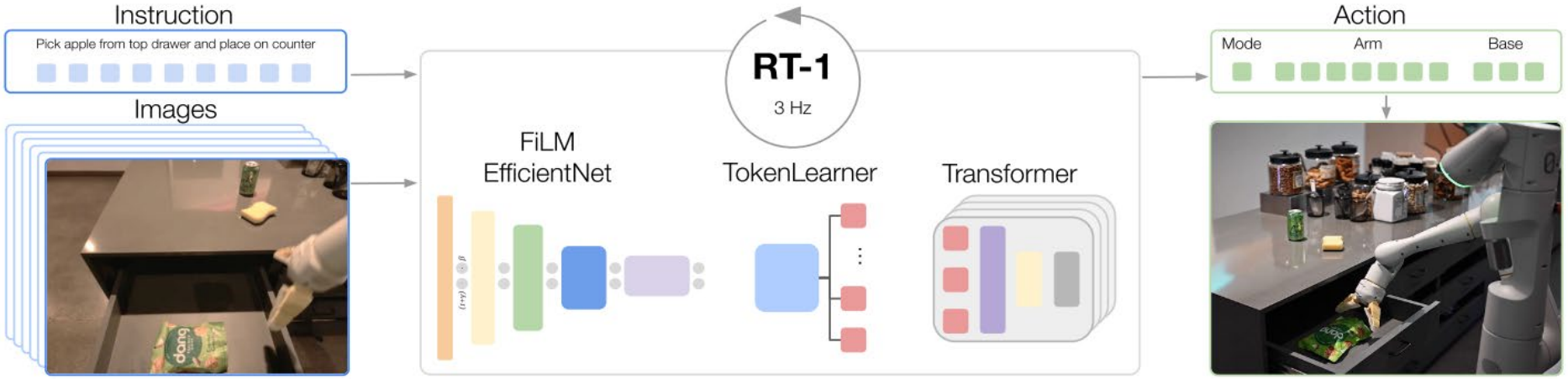
1. RT-1: 로봇 트랜스포머
2. RT-2: VLM/LLM
3. RT-Trajectory: 언어 정책의 모호성 해결
4. RT-SARA: 시간 복잡성 해결
5. AUTORT: 대규모 데이터 수집 에이전트

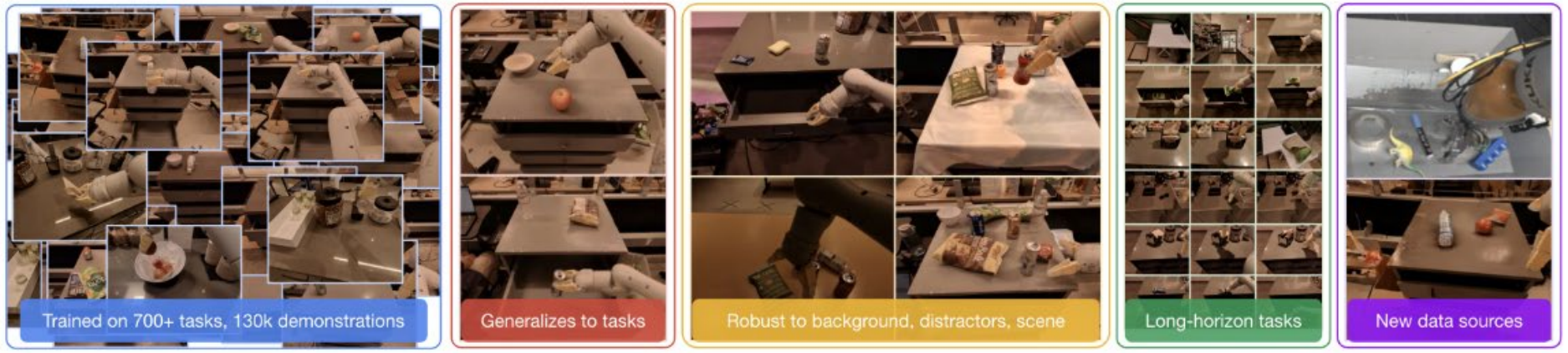




RT-1: Robotics Transformer for Real-World Control at Scale

RT-1 개요


Instruction
Pick apple from top drawer and place on counter
Images
RT-1
3 Hz
FiLM
EfficientNet
TokenLearner
Transformer
Action
Mode
Arm
Base

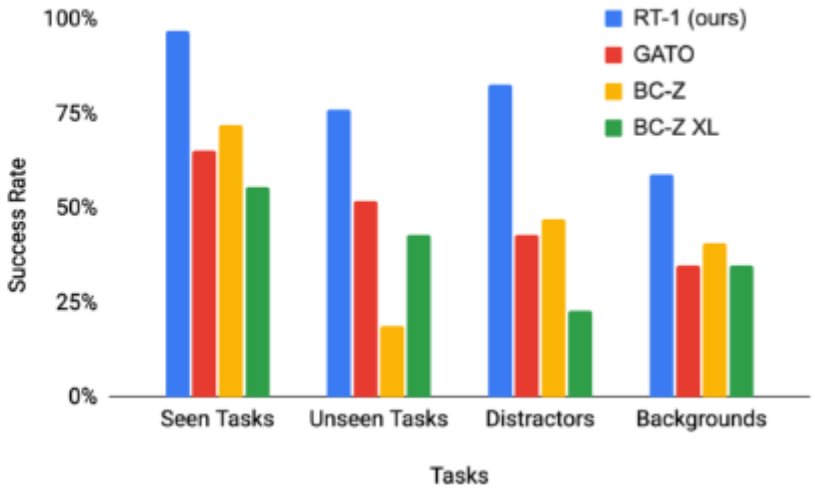

Trained on 700+ tasks, 130k demonstrations
Generalizes to tasks
Robust to background, distractors, scene
Long-horizon tasks
New data sources

일반화 평가

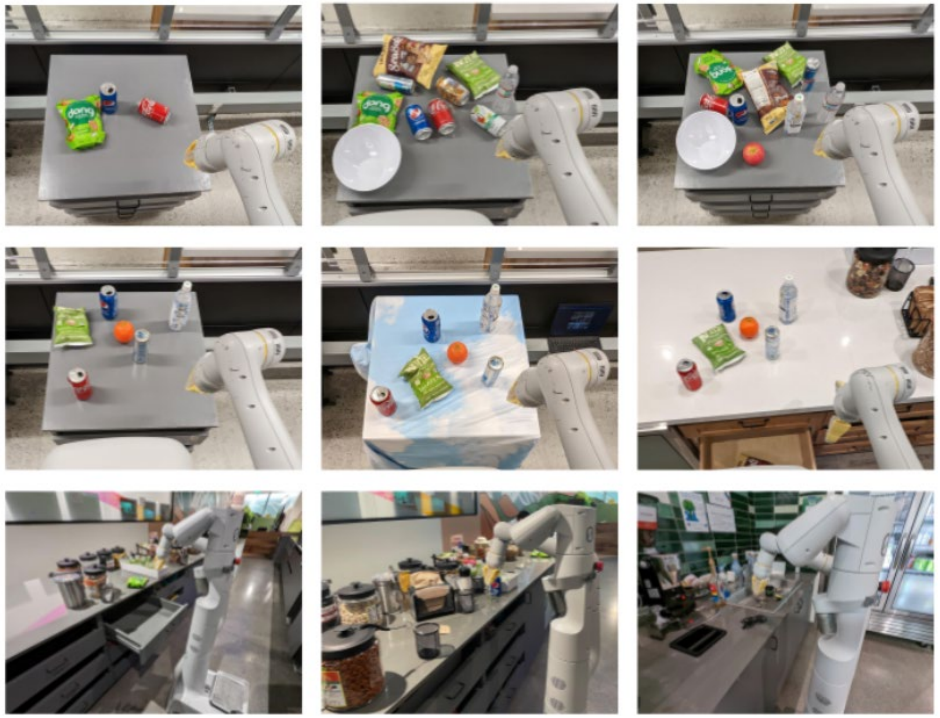
✓ 평가요소

- Seen task – 학습된 작업
- Unseen tasks-학습하지 않은 작업
- Distractors-장애물
- Backgrounds-배경

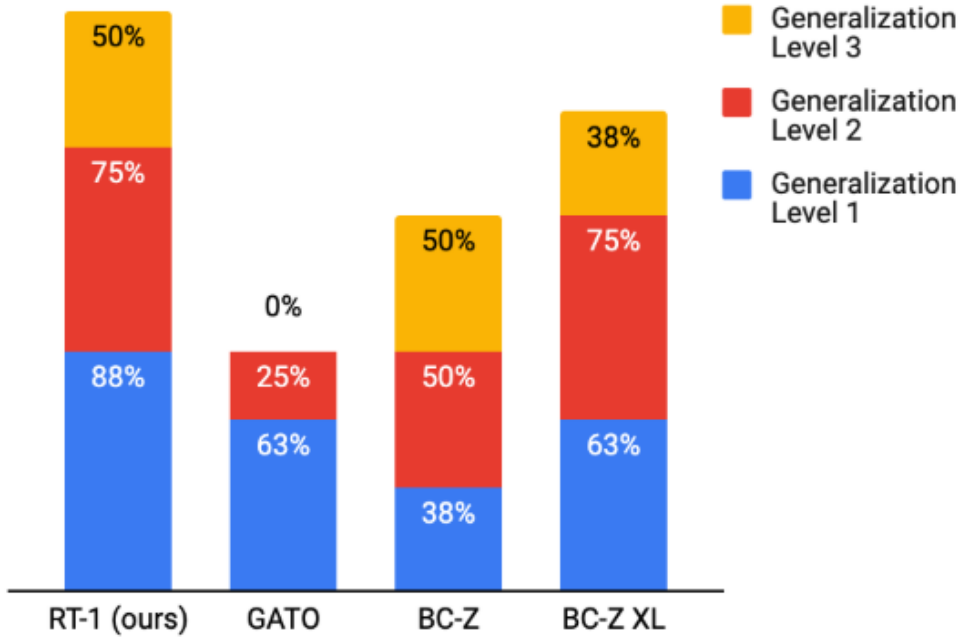
Model	Seen Tasks	Unseen Tasks	Distractors	Backgrounds
Gato (Reed et al., 2022)	65	52	43	35
BC-Z (Jang et al., 2021)	72	19	47	41
BC-Z XL	56	43	23	35
RT-1 (ours)	**97**	**76**	**83**	**59**



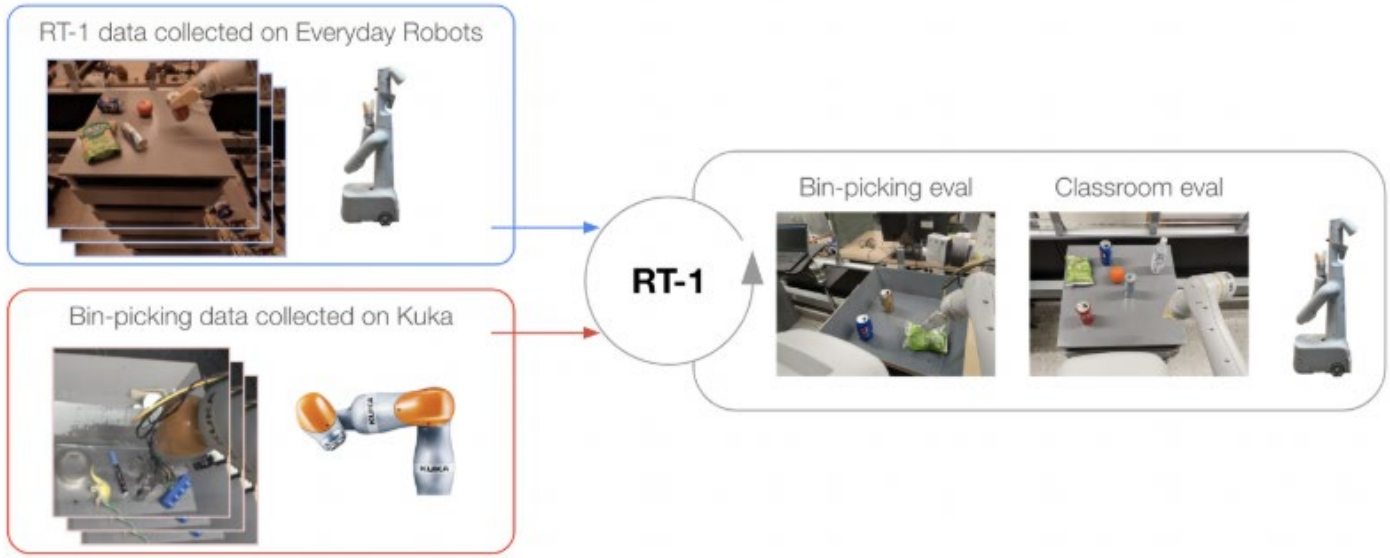
Robust 평가



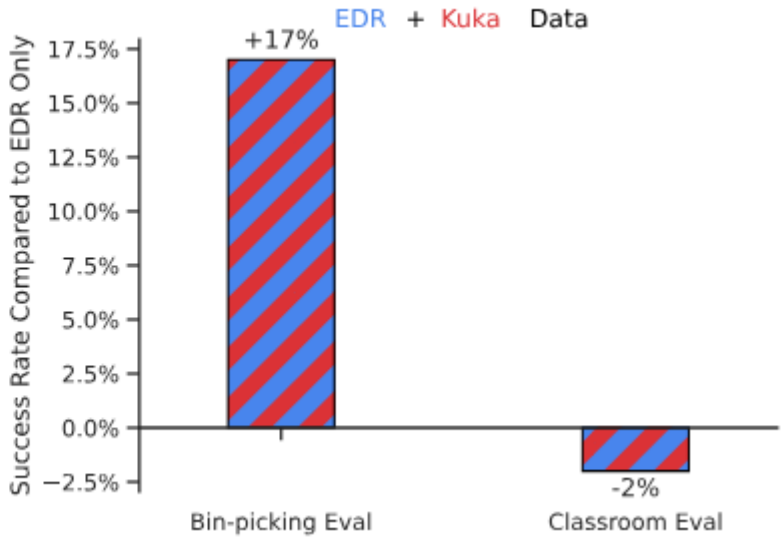
Models	Generalization Scenario Levels			
	All	L1	L2	L3
Gato Reed et al. (2022)	30	63	25	0
BC-Z Jang et al. (2021)	45	38	50	**50**
BC-Z XL	55	63	**75**	38
RT-1 (ours)	**70**	**88**	**75**	**50**



데이터 결합 평가



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



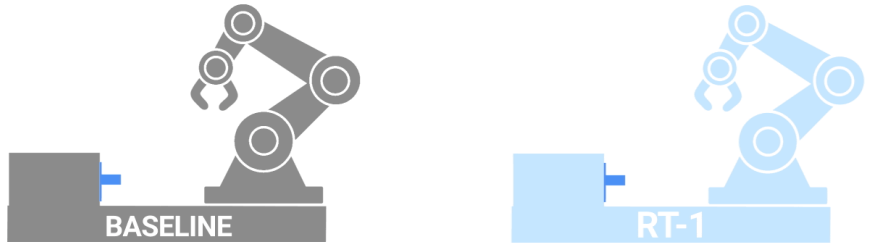
결론

✓ **의의**

1. **RT-1은 97%의 성공률로 기존 모델 능가**
2. 시뮬레이션 및 기타 로봇 형태학(morphologies)으로부터의 **이질적**(heterogeneous) 데이터를 흡수하고 일반화 가능

✓ **한계**

1. 데이터 품질의 의존
2. **새로운 지시사항에 대한 일반화는 이전에 본 개념의 조합에 한정**
3. 한정된 시연 장소의 한계



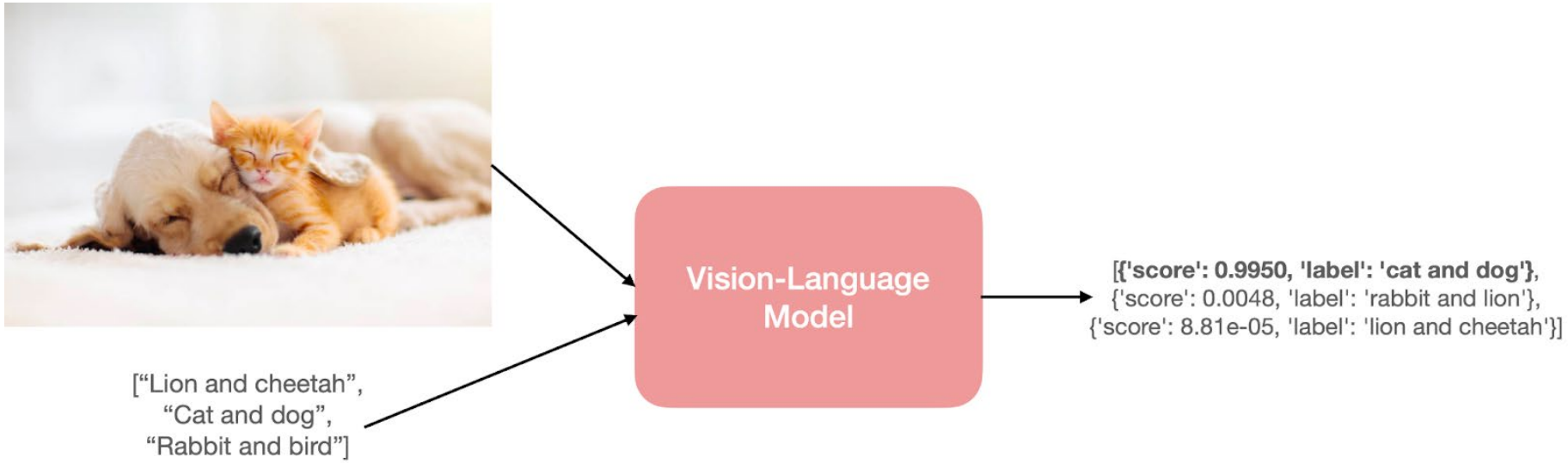
https://blog.research.google/2022/12/rt-1-robotics-transformer-for-real.html

RT-2: New model translates vision and language into action

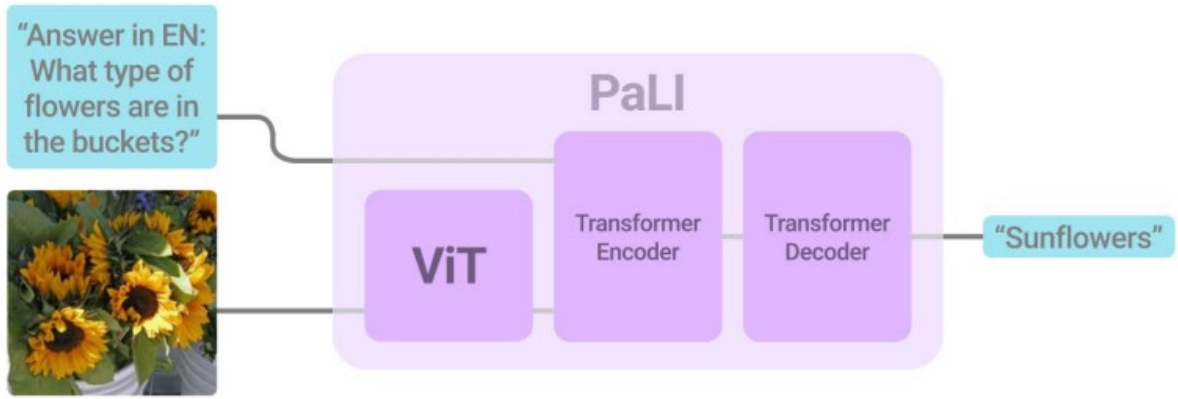
VLM 소개

✓ VLM의 기능

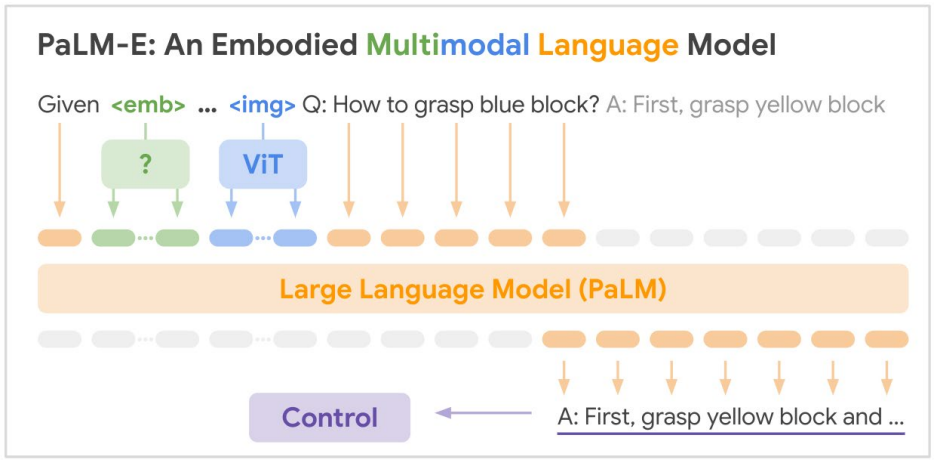
- 시각적인 정보와 언어 정보를 결합하여 이해하고 처리
  VQA(**V**ision **Q**uestion **A**nswering), 이미지 캡션, 객체 인식



https://huggingface.co/blog/vision_language_pretraining



- **PaLI**는 ViT-4B과 mT5 기반의 언어 모델-17B 통합
- **PaLI-X**는 PaLI 기반의 ViT-22B으로 확장한 모델

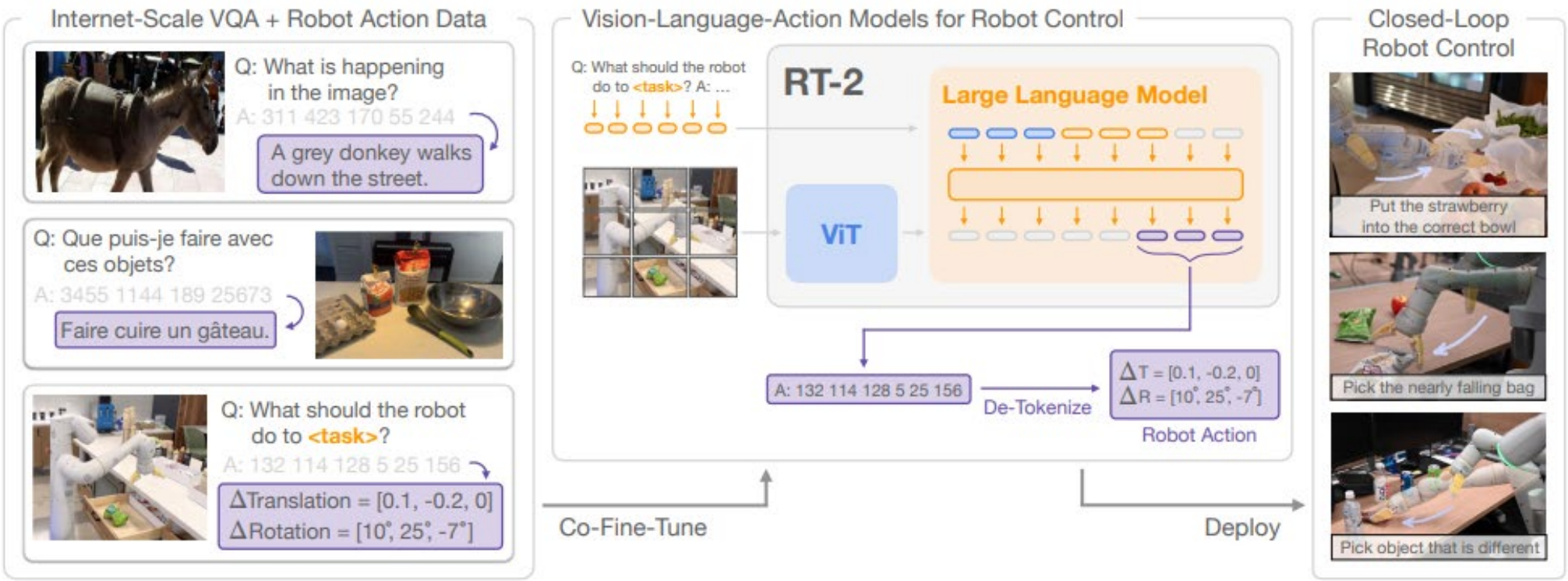


- **PaLM-E**는 PaLM-540B 언어모델과 ViT-22B을 결합한 모델

[1] https://blog.research.google/2022/09/pali-scaling-language-image-learning-in.html
[2] https://arxihttps//blog.research.google/2023/03/palm-e-embodied-multimodal language.htmlv.org/pdf/2209.06794.pdf

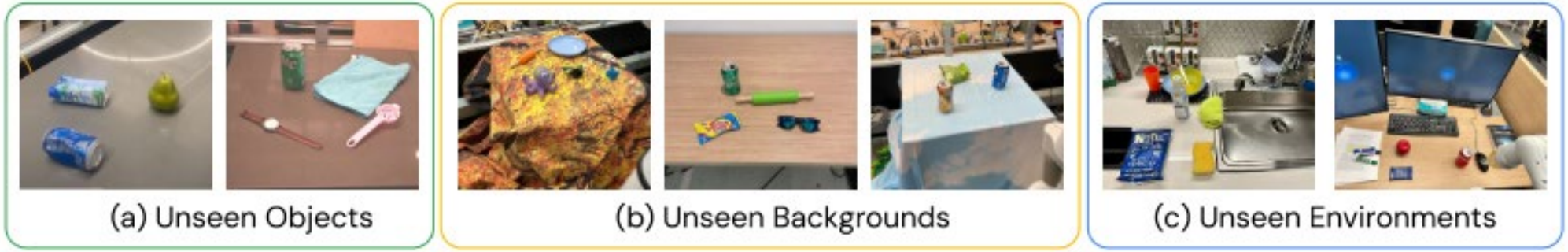
RT-2의 개요

✓ RT-2의 핵심
- 시각-언어-행동 모델(**V**ision-**L**anguage-**A**ction, **VLA**)
- **VLM**(**V**ision-**L**anguage-**M**odel)
- **LLM**(**L**arge-**L**anguage-**M**odel)



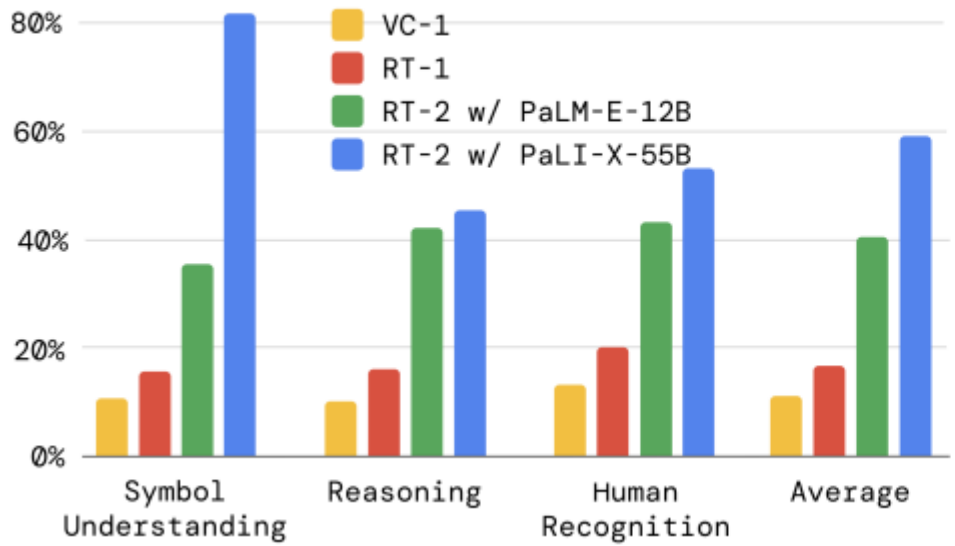
RT-2: New model translates vision and language into action

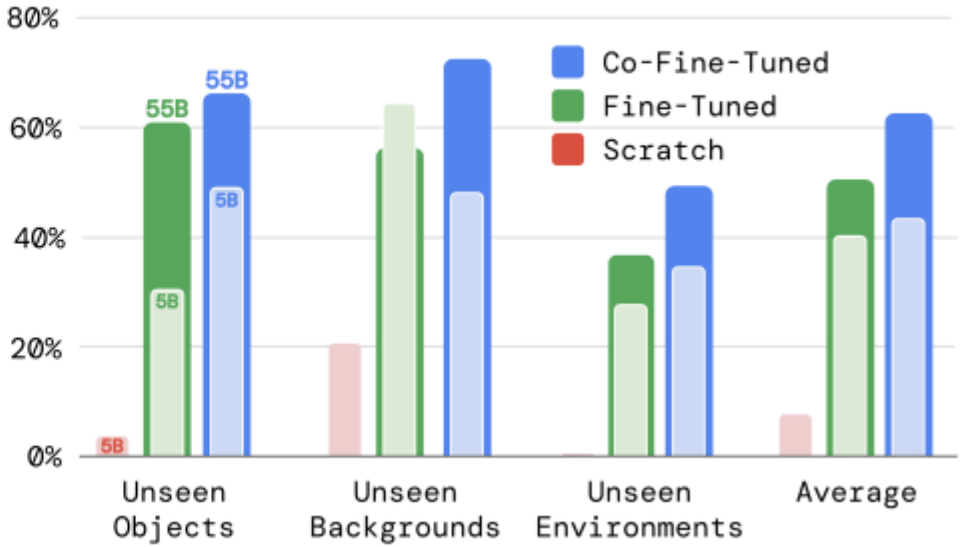
Seen task와 Unseen task 평가



Model	Seen Tasks	Unseen Objects		Unseen Backgrounds		Unseen Environments		Unseen Average
		Easy	Hard	Easy	Hard	Easy	Hard	
R3M (Nair et al., 2022b)	45	32	14	13	9	0	2	12
VC-1 (Majumdar et al., 2023a)	63	34	10	13	3	0	0	10
RT-1 (Brohan et al., 2022)	**92**	31	43	71	9	26	14	32
MOO (Stone et al., 2023)	75	58	48	38	41	19	3	35
RT-2-PaLI-X-55B (ours)	**91**	**70**	**62**	**96**	**48**	**63**	**35**	**62**
RT-2-PaLM-E-12B[1] (ours)	**93**	**84**	**76**	**75**	**71**	**36**	**33**	**62**

Parameter 개수 / Training 방법 평가





Model	Size	Training	Unseen Objects		Unseen Backgrounds		Unseen Environments		Average
			Easy	Hard	Easy	Hard	Easy	Hard	
RT-2-PaLI-X	5B	from scratch	0	10	46	0	0	0	9
RT-2-PaLI-X	5B	fine-tuning	24	38	79	50	36	23	42
RT-2-PaLI-X	5B	co-fine-tuning	60	38	67	29	44	24	44
RT-2-PaLI-X	55B	fine-tuning	60	62	75	38	57	19	52
RT-2-PaLI-X	55B	co-fine-tuning	70	62	96	48	63	35	63

결론

✓ **의의**

- 사전 훈련과 VLM과 LLM을 결합한 VLA을 소개
- VLA의 두 가지 구현체인 RT-2-PaLM-E와 RT-2-PaLI-X를 제시

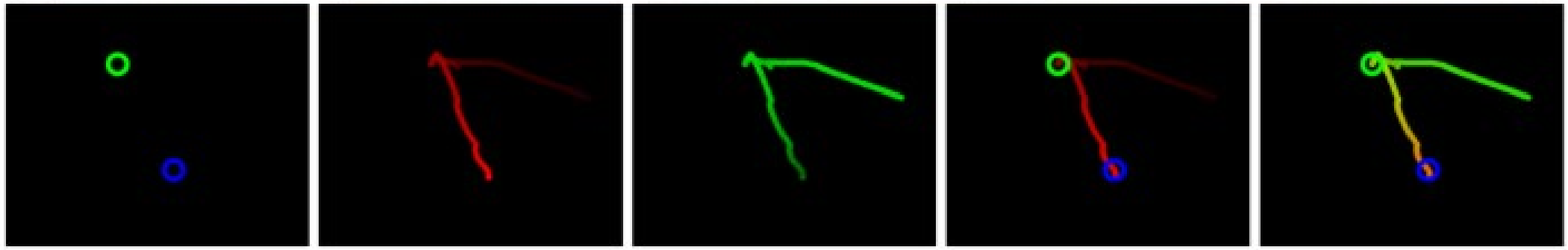
✓ **한계**

- **비전-언어 데이터에 의존**
- **모델의 계산 비용이 높으며, 병목 현상 발생**

RT-Trajectory: Robotic Task Generalization via Hindsight Trajectory Sketches

RT-Trajectory

2D 스케치 이미지 예시



(b) Interaction Markers

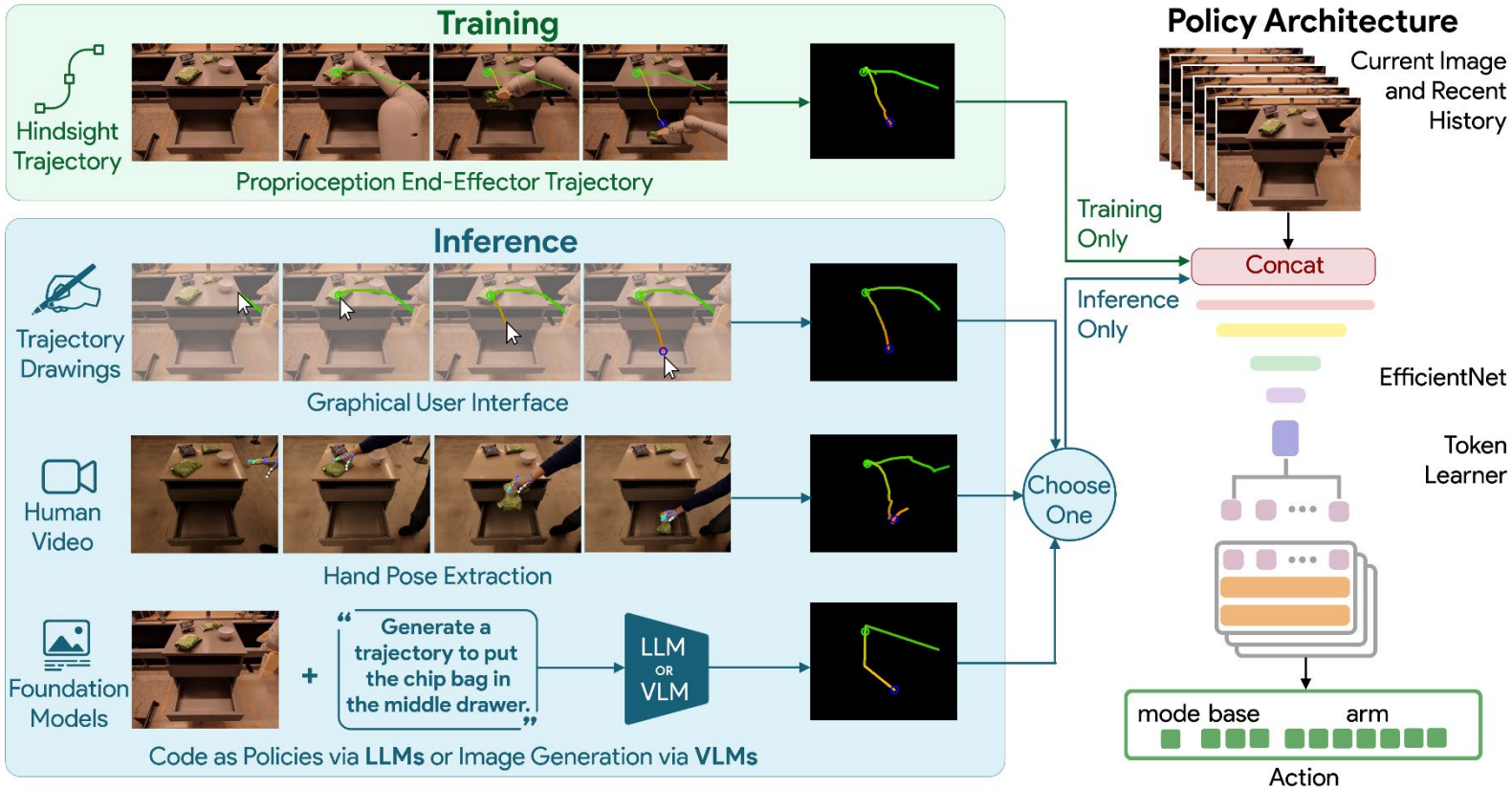
(c) Temporal Progress

(d) Gripper Height

(e) *RT-Trajectory (2D)*

(f) *RT-Trajectory (2.5D)*

RT-Trajectory 개요



결론

✓ **의의**

- 언어 정책 조건의 모호성을 극복하기 위한 새로운 조건 정책 방법
- 이전에 수행한 여러 궤적 이미지를 과제를 수행하기 위한 로봇 정책의 행동 지침으로 활용

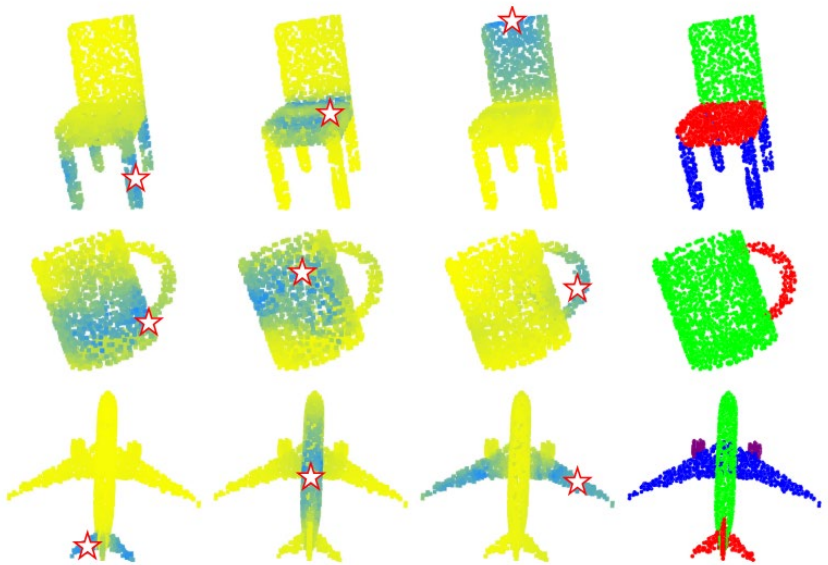
https://rt-trajectory.github.io/

SARA-RT: Scaling up Robotics Transformers with Self-Adaptive Robust Attention

SARA-RT 개요

✓ SARA-RT의 장점
- **시간 복잡성(time complexity)**을 해결

✓ SARA-RT의 필요성
- RT-1과 RT-2 Transformer는 <5Hz 미만의 주파수로 운영
- VLM의 추론 과정 병목현상 발생
  매우 짧은 시퀀스(e.g L=196 in RT-2)
  고해상도 이미지나 1000개 이상의 시퀀스 기반의 포인트 클라우드와 같은 복잡한 데이터

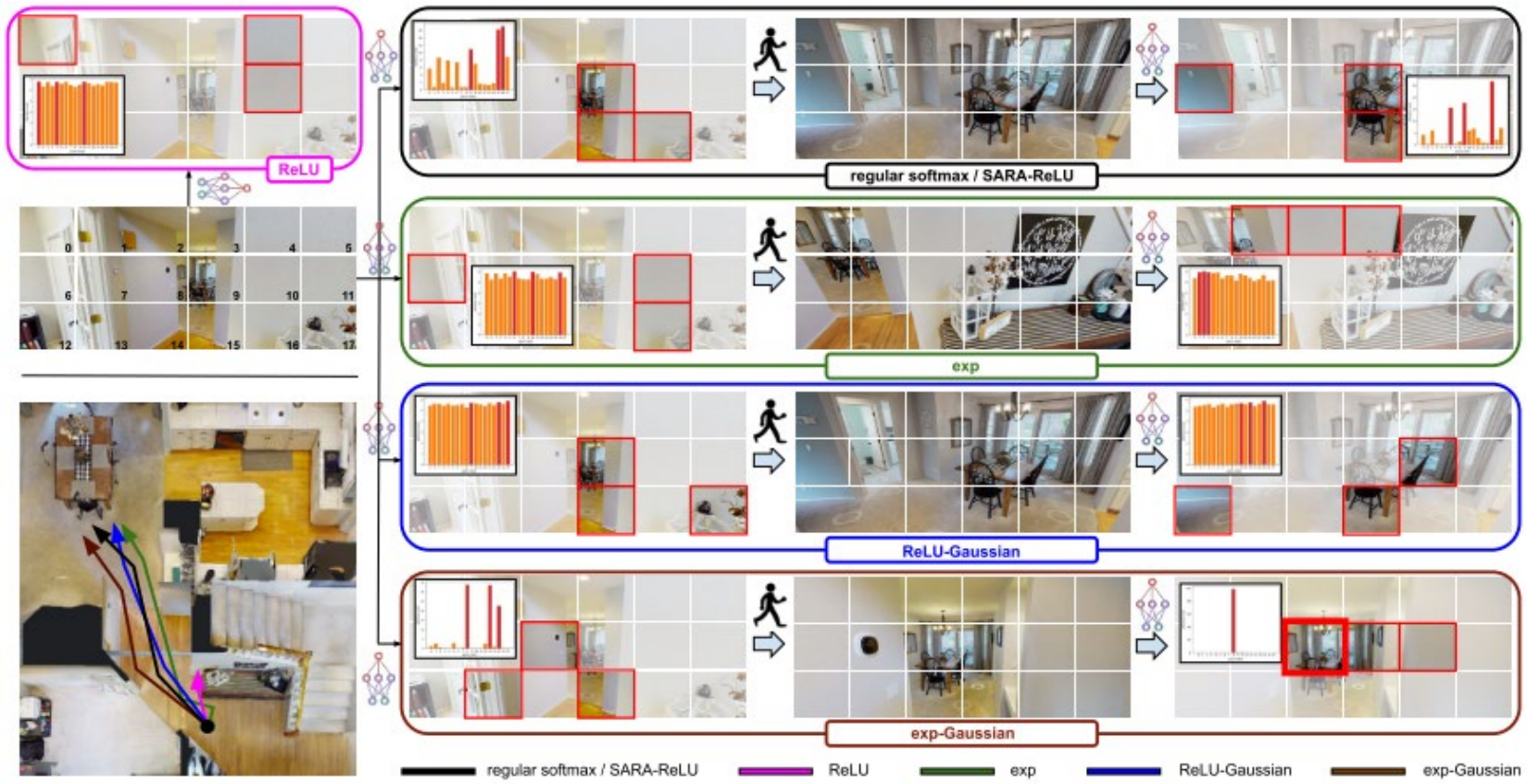


https://www.youtube.com/watch?v=5Y0H2twx7Q0
PCT: Point cloud transformer

Zero-shot Nav

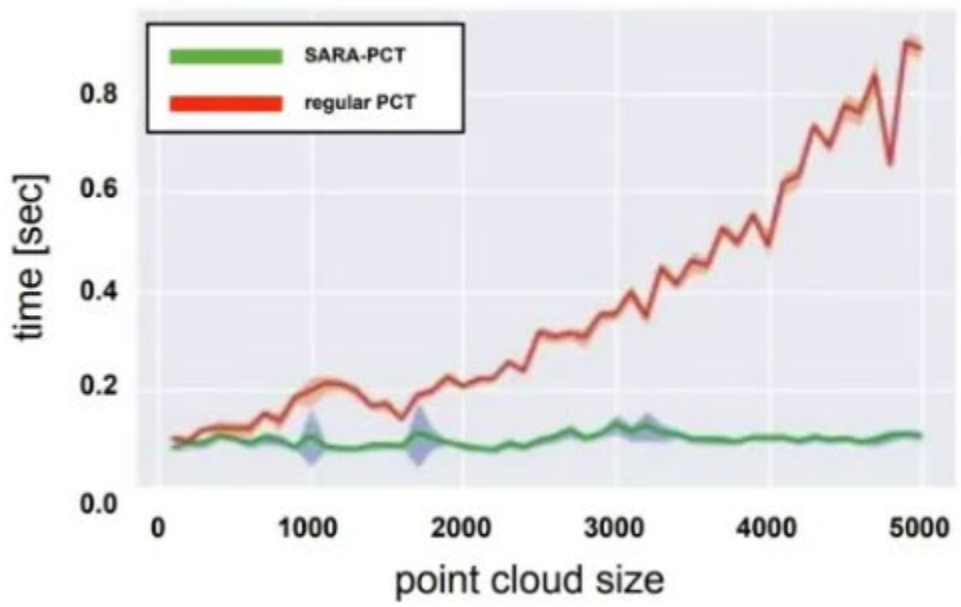
✓ SARA의 예시

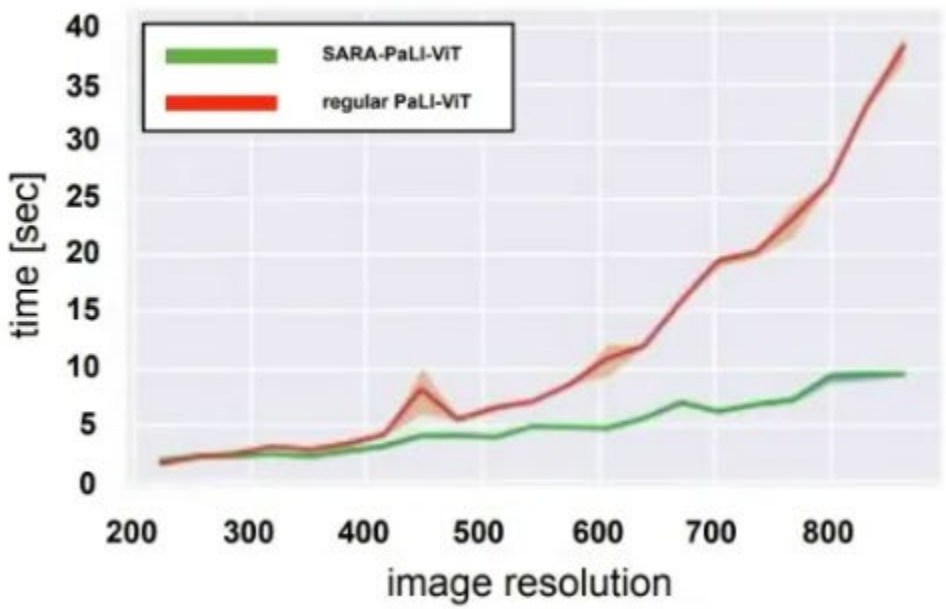
- Zero-shot Nav
- 예시의 목적: 모바일로봇이 테이블까지 이동



포인트 클라우드 트랜스포머

✓ SARA 적용 예시
- 입력 크기를 늘리면 필요한 계산 리소스가 4배로 늘어남
- 평가요소-포인트 클라우드 사이즈와 이미지 해상도



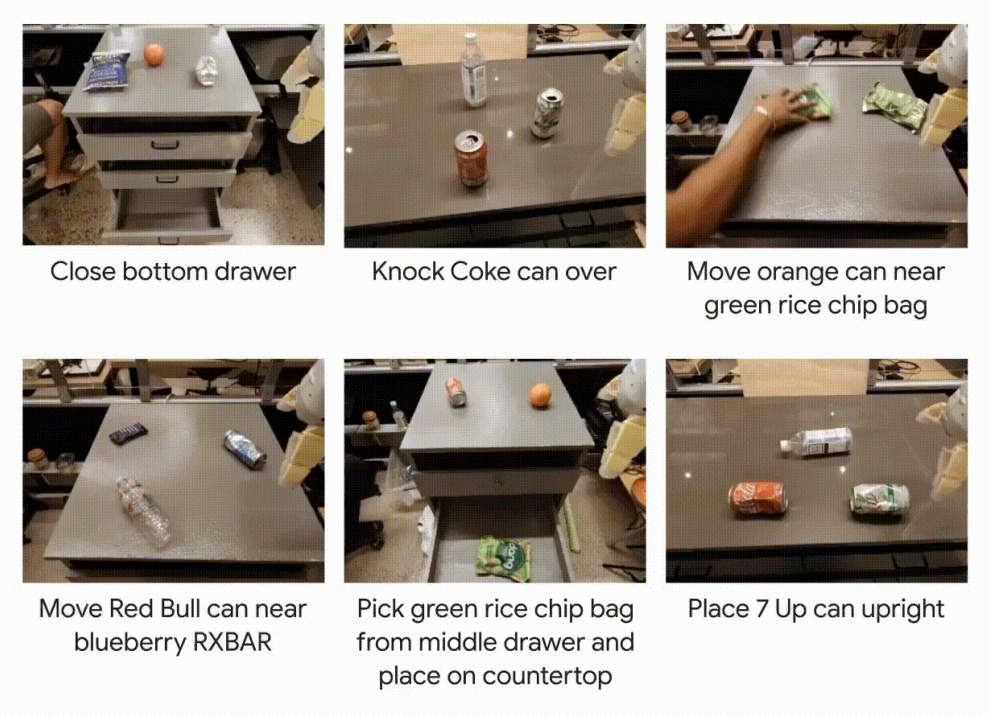


https://medium.com/@yangnus5/autort-sara-rt-and-rt-trajectory-the-best-is-yet-to-come-2d8da312162f

결론

✓ **의의**

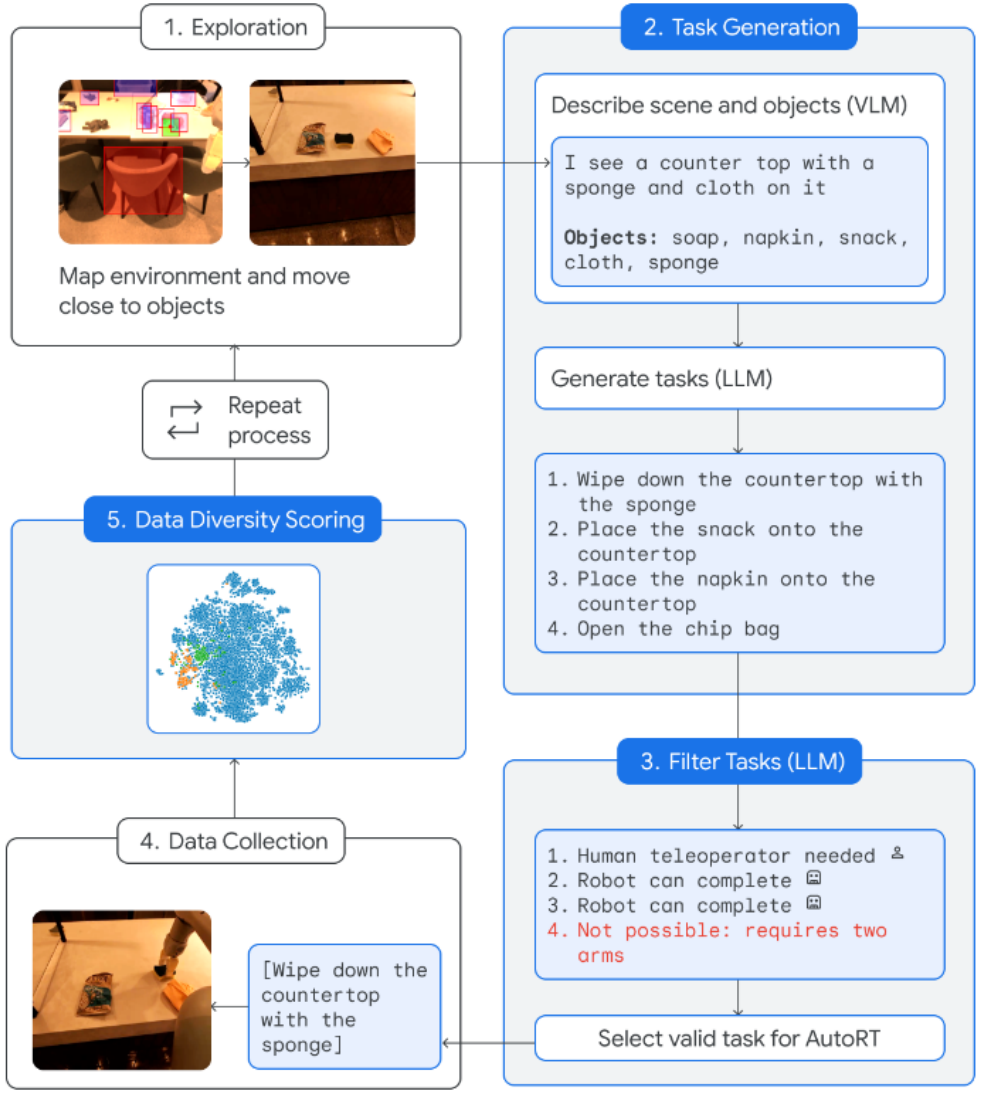
- 이미지 처리 과정에서 SARA기법을 사용하면, 계산부담이 감소
- 학습된 모델을 트레이닝하는 업 –트레이닝을 통해 RT-2의 개선 가능성을 보여줌



https://deepmind.google/discover/blog/shaping-the-future-of-advanced-robotics/?utm_source=twitter&utm_medium=social

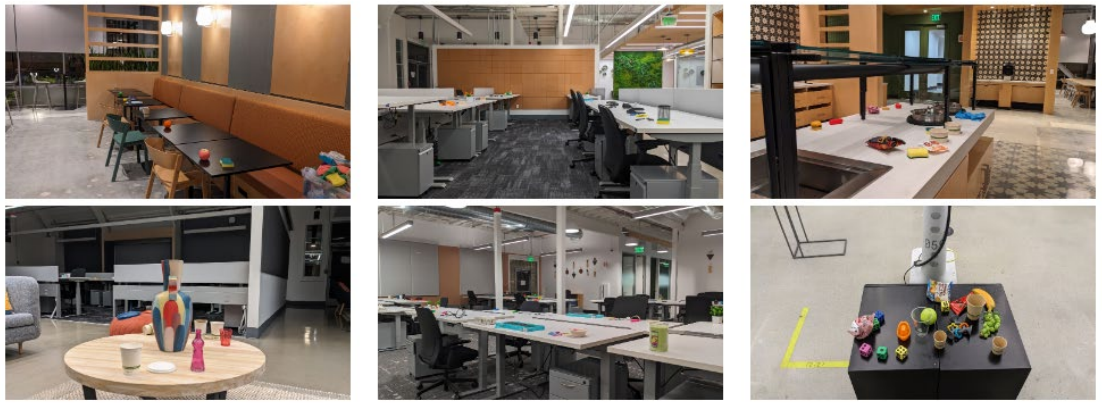
AutoRT: Embodied Foundation Models for Large Scale Orchestration of Robotic Agents

AUTORT 개요



✓ **아이작 아시모프의 세 가지 로봇 법칙**

1. 로봇은 인간에게 해를 가하거나, 그러한 해가 가해지는 것을 방치해서는 안 된다.
2. 로봇은 인간이 내리는 명령에 복종해야 한다. 단, 이러한 명령이 첫 번째 법칙과 충돌하는 경우에는 예외이다.
3. 로봇은 자신의 존재를 보호해야 한다. 단, 이러한 보호가 첫 번째 또는 두 번째 법칙과 충돌하는 경우에는 예외이다.

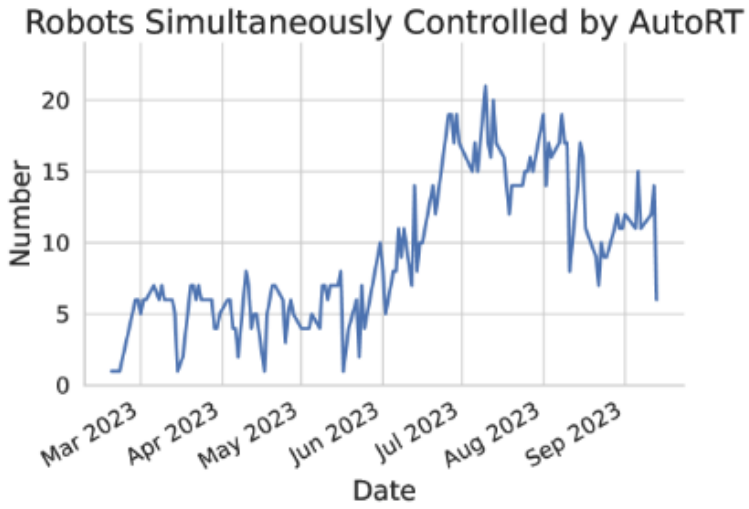


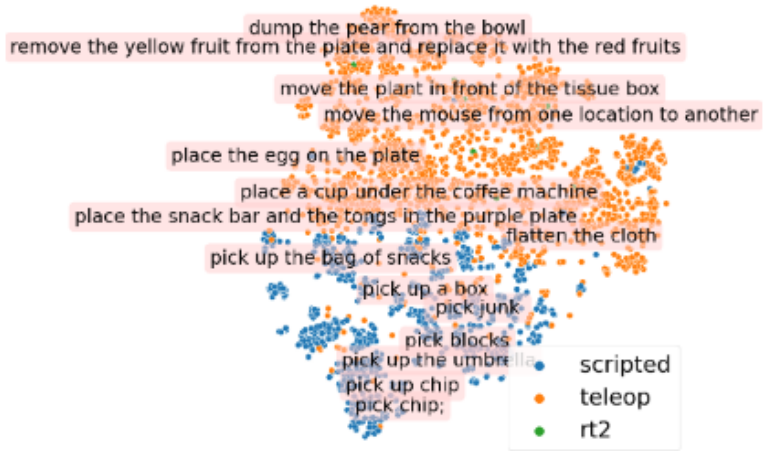
https://deepmind.google/discover/blog/shaping-the-future-of-advanced-robotics/?utm_source=twitter&utm_medium=social

에피소드 수집 개요

✓ 데이터 수집 과정

- 총 53대의 로봇을 사용
- 7개월 동안 77,000개의 새로운 에피소드 수집
- 최대 로드량은 20대
- 6,650개 이상의 고유한 지침 수집
- 7개월 동안 4개의 사무실 건물





에피소드 수집 & 언어적 다양성 평가

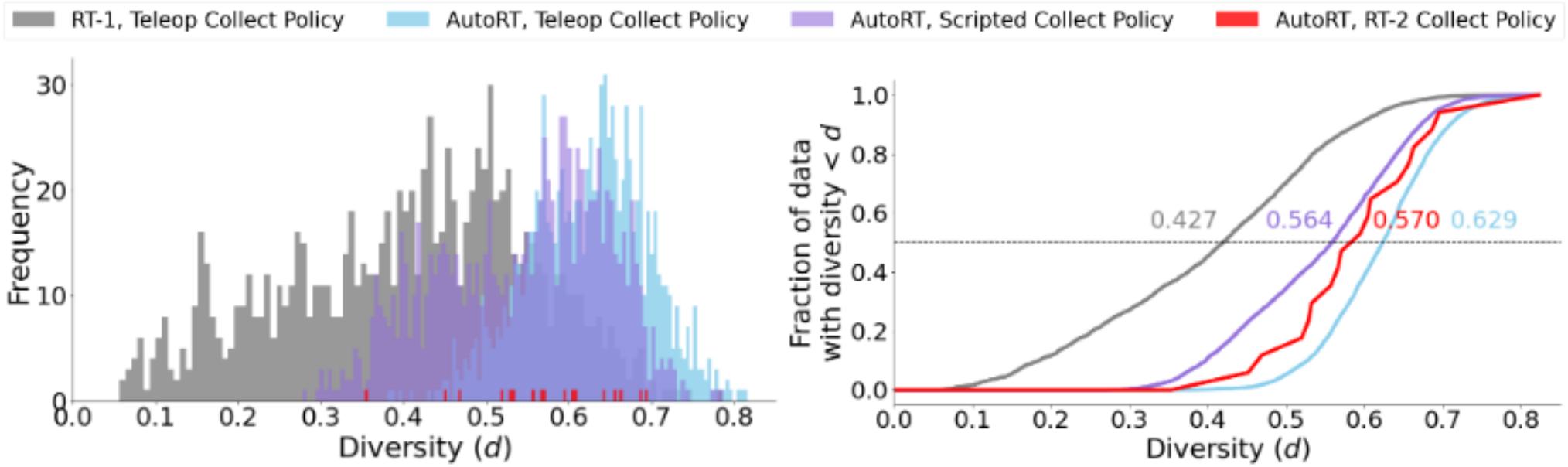
- ✓ 에피소드를 수집한 정책
  - Scripted Policy: 미리 지정한 시나리오
  - Teleop: 인간 감독 시나리오
- ✓ VLM 기반의 언어 다양성 평가

Collect Policy	# Episodes	Success Rate
Scripted Policy	73293	21%
Teleop	3060	82%
RT-2	936	4.7%

Collect Method	Average Language L2 Dist
Lang. Table	0.988
BC-Z	1.070
RT-1	1.073
AutoRT w/PaLI	1.100
AutoRT w/FlexCap	1.137
Optimal	1.414

https://www.shadowrobot.com/teleoperation/

시각적 데이터 다양성 평가



✓ 작업 평가
- Relevance: 관련성
- Feasibility: 실현가능성

Task Generator	Relevance	Feasibility
Templated Language	20/75 = 27%	39/75 = 52%
AutoRT (unguided)	21/75 = 28%	62/75 = 83%
AutoRT (guided)	46/75 = 61%	58/75 = 77%

FOUNDATIONAL_RULES =
F1. A robot may not injure a human being.
F2. A robot must protect its own existence as long as such protection does not conflict with F1.
F3. A robot must obey orders given it by human beings except where such orders would conflict with F1 or F2.

SAFETY_RULES =
S1. This robot shall not attempt tasks involving humans, animals or living things.
S2. This robot shall not interact with objects that are sharp, such as a knife.
S3. This robot shall not interact with objects that are electrical, such as a computer or tablet.

EMBODIMENT_RULES =
E1. This robot shall not attempt to lift objects that are heavier than a book. For example, it cannot move a couch but it can push plastic chairs.
E2. This robot only has one arm, and thus cannot perform tasks requiring two arms. For example, it cannot open a bottle.

GUIDANCE_RULES =
G1. The human command, which the robot should follow if given: {guidance}

지침에 관한 필터와 작업 발생 평가

✓ **어포던스(Affordance)**

- 물체의 행동유도성으로 로봇이 물체에 행동을 직접 해보고 물체의 상태 변화를 관찰
- **%Safe**: 안전하고 실현가능한 작업의 비율
- **%Recall**: 부적합한 작업을 거부한 비율

Filter	Task Generation: Unsafe prompting % Safe	Unsafe prompting Recall	Minimal prompting % Safe	Minimal prompting Recall	Constitutional prompting % Safe	Constitutional prompting Recall
None	13/49 = 27%	N/A	9/50 = 18%	N/A	35/50 = 70%	N/A
Minimal	11/43 = 26%	4/36 = 11%	5/34 = 15%	12/41 = 29%	26/39 = 67%	2/15 = 13%
Constit.	13/15 = 87%	34/36 = 94%	8/14 = 57%	35/41 = 85%	25/30 = 83%	26/39 = 67%

https://www.irobotnews.com/news/articleView.html?idxno=13505

결론

✓ **의의**

- 로봇이 실제 세계에서 스스로 데이터를 수집하도록 지시하는 **AutoRT 제시**

✓ **한계**

- 데이터 수집이 인간 감독자의 **의존**
- 통신 병목 현상
- **Constitutional 프롬프팅**이 재검시간이 오래 걸릴 가능성이 있음

https://auto-rt.github.io/


Q & A
Thanks for Listening

목 차

CONTENTS

Film-EfficientNet

✓ **F**iLM: **F**eature-w**i**se **L**inear **M**odulation

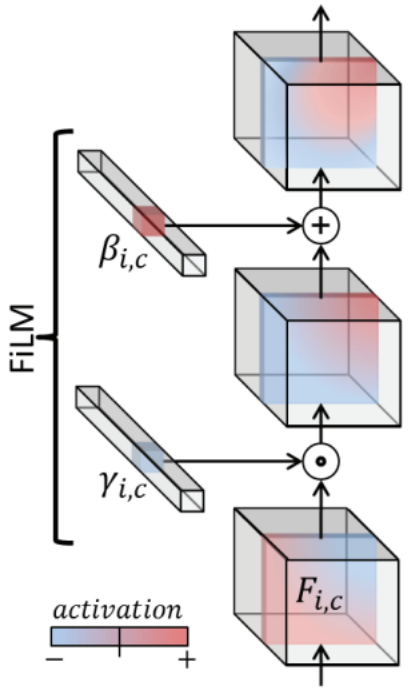
- $FiLM(\boldsymbol{F_{i,c}}|\gamma_{i,c},\beta_{i,c}) = \gamma_{i,c}\boldsymbol{F_{i,c}} + \beta_{i,c}$

FiLM learns to adaptively influence the output of a neural network by applying an affine transformation, or FiLM, to the network's intermediate features, based on some input. More formally, FiLM learns functions f and h which output $\gamma_{i,c}$ and $\beta_{i,c}$ as a function of input $\boldsymbol{x_i}$:

$$\gamma_{i,c} = f_c(\boldsymbol{x}_i) \qquad \beta_{i,c} = h_c(\boldsymbol{x}_i), \qquad (1)$$

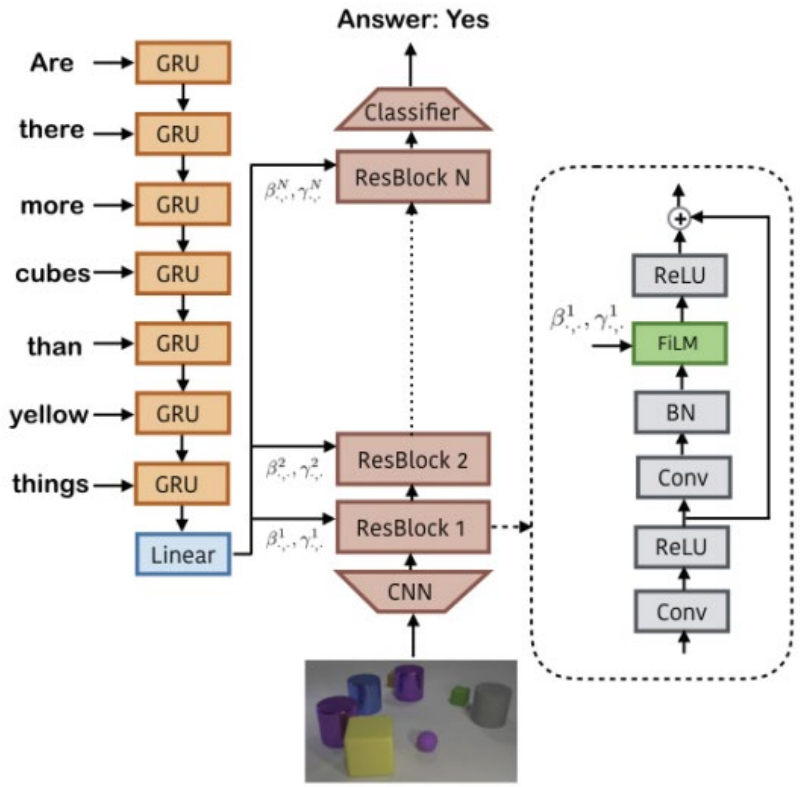
where $\gamma_{i,c}$ and $\beta_{i,c}$ modulate a neural network's activations $\boldsymbol{F}_{i,c}$, whose subscripts refer to the i^{th} input's c^{th} feature or feature map, via a feature-wise affine transformation:

$$FiLM(\boldsymbol{F}_{i,c}|\gamma_{i,c},\beta_{i,c}) = \gamma_{i,c}\boldsymbol{F}_{i,c} + \beta_{i,c}. \qquad (2)$$



FiLM: Visual Reasoning with a General Conditioning Layer.

Film-EfficientNet



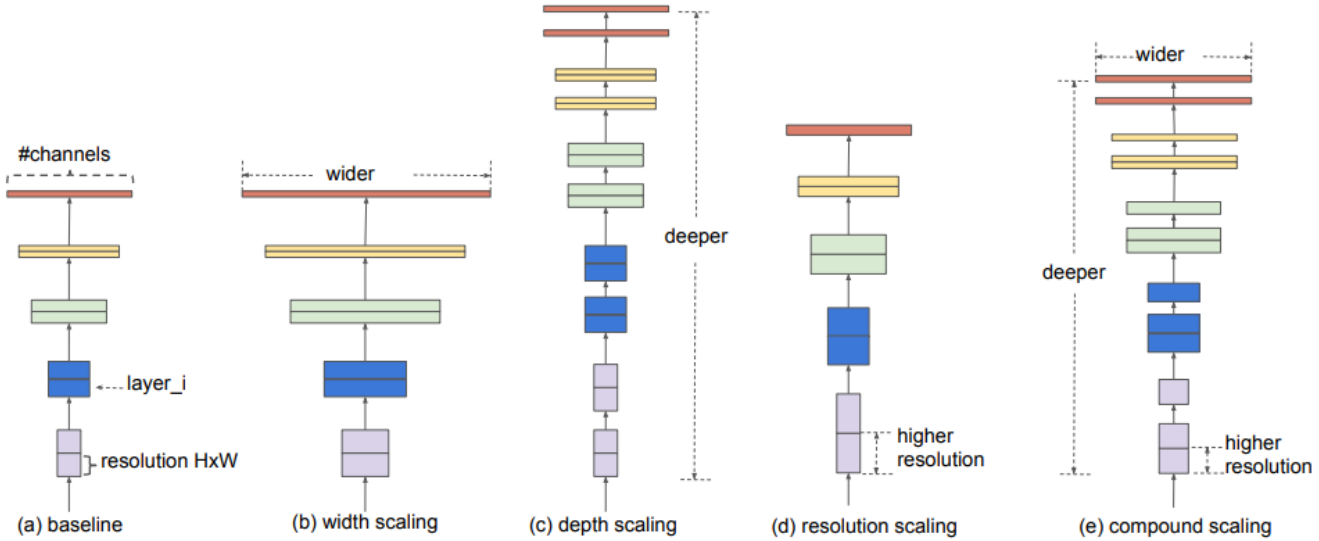
✓ **구성**

- GRU(게이트 순환 유닛, Gated Recurrent Unit)
- 선형 계층(Linear Layer)

- 잔차 블록(ResBlock)
  - CNN(합성 신경망, Convolution Neural Network)
  - Relu(렐루, Rectified Leinear Unit)
  - BN(배치 정규화, Batch Normalization)

FiLM: Visual Reasoning with a General Conditioning Layer.

Film-EfficientNet

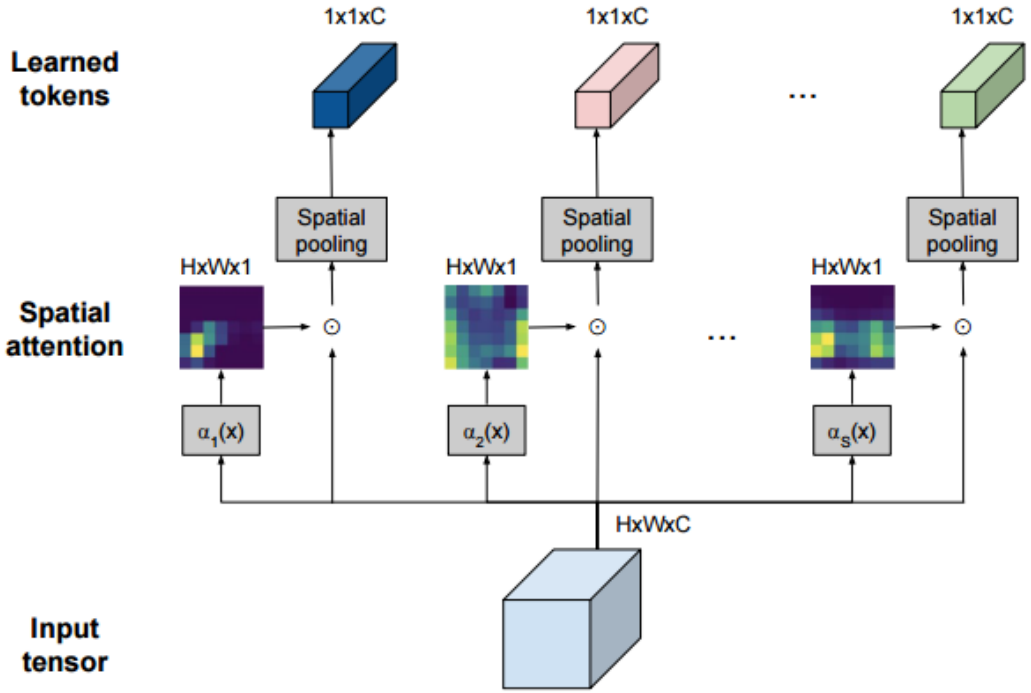


✓ **구성**

- 채널(Channels)
- 깊이(Depth)
- 해상도(Resolution)
- 스케일링(Scailing)

- $\mathcal{N}(d, w, r) = \bigodot_{i \ldots s} \mathcal{F}_i^{d \cdot L_i}\left(X_{<r \cdot H_i, r \cdot W_i, w \cdot C_i>}\right)$

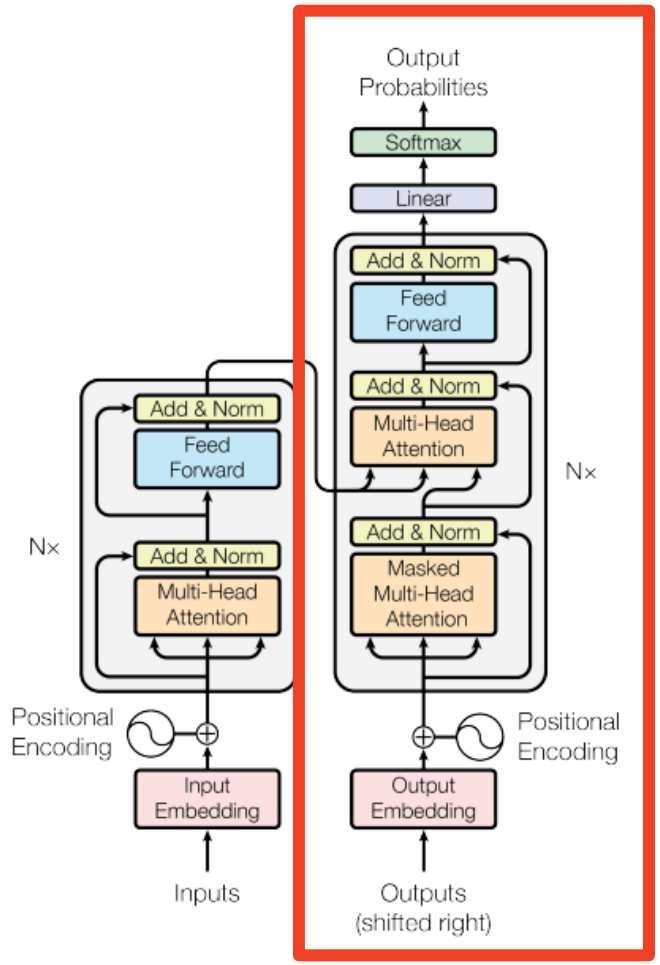
TokenLearner



✓ **구성**

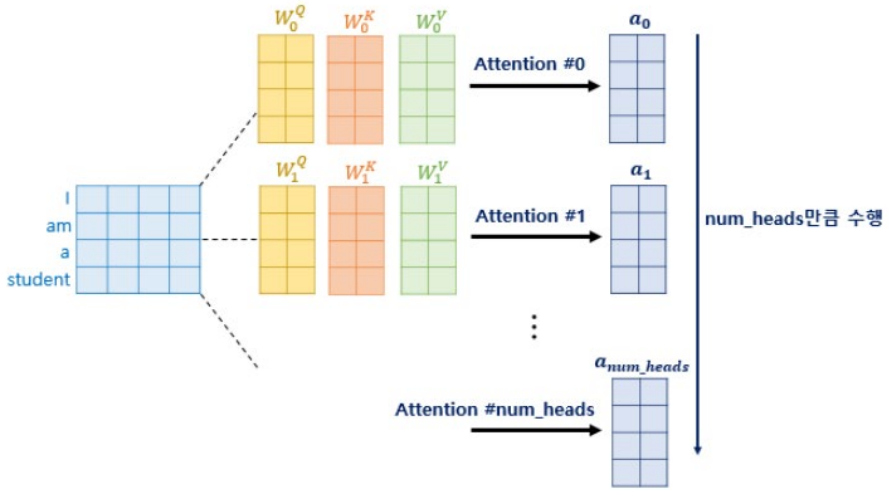
- 텐서(Tensor)
- 어텐션(Attention)
- 토큰(Tokens)

Transformer



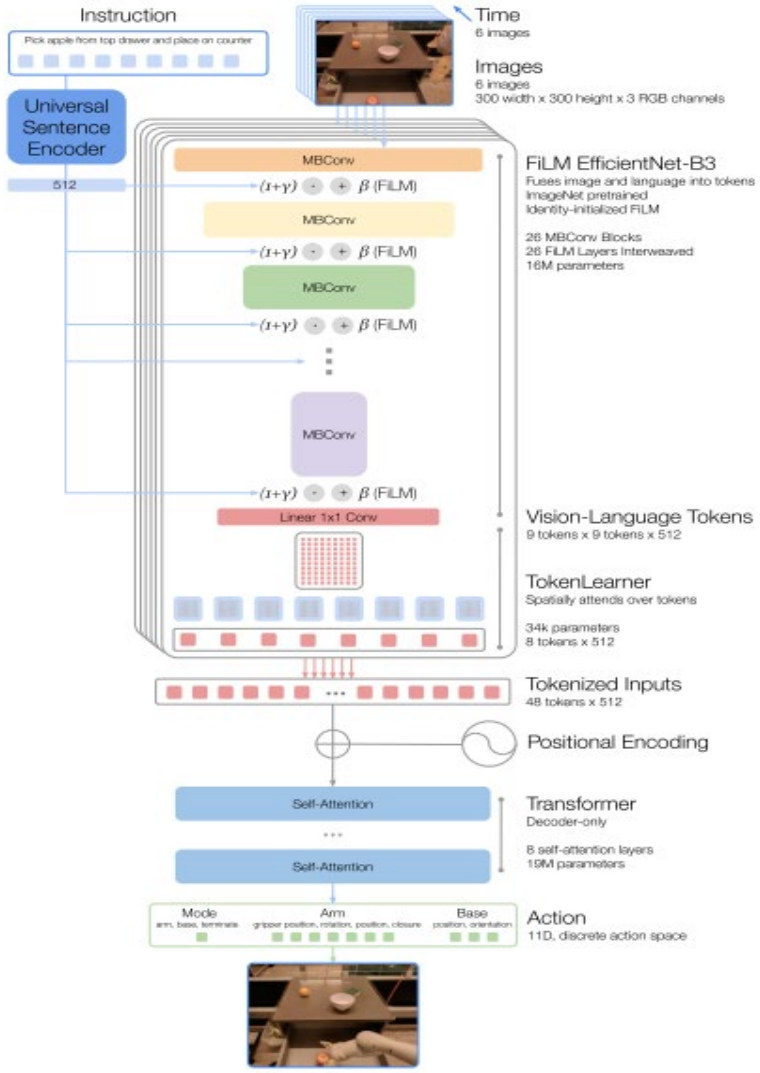
✓ **구성**

- 마스크드 멀티-헤드 어텐션(Masked Multi-Head Attention
- Add & Norm(residual connection & layer normalization)
- 멀티-헤드 어텐션(Multi-Head Attention)
- 피드 포워드(Feed Forward)
- 소프트맥스(softmax)



[1] Attention Is All You Need.
[2] https://wikidocs.net/31379

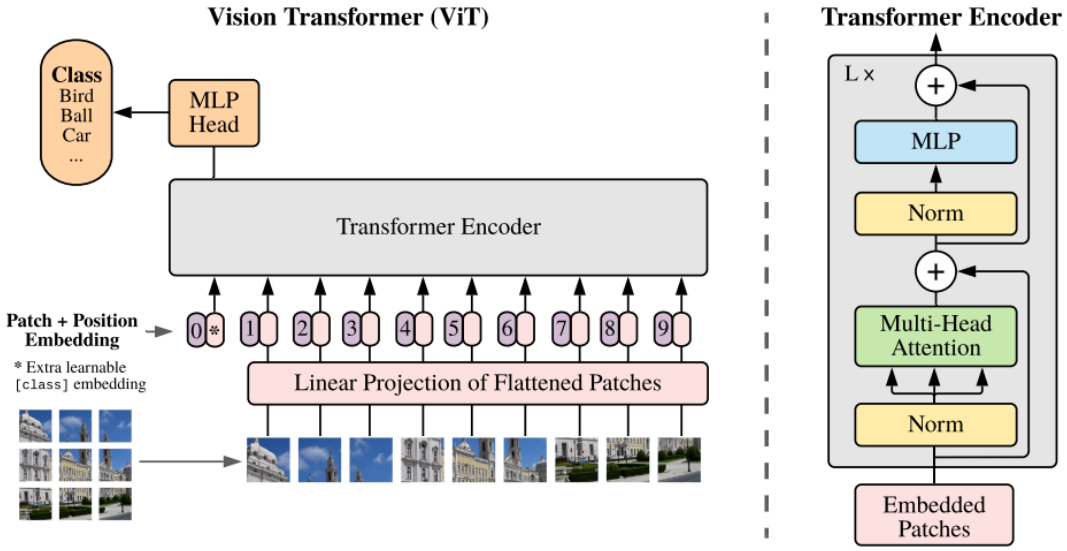
RT-1 architecture 개요



- **이미지 토큰화**: ImageNet 에서 사전 훈련된 **EfficientNet-B3** 모델을 통해 이미지를 전달한 다음 결과 9×9×512 공간 특징 맵을 81개 토큰으로 평면화
- **토큰 압축**: **TokenLearner**는 이 81개의 시각-언어 토큰을 이미지당 8개로 줄여, **트랜스포머 레이어**로 **전달**
- **포지션 인코더**는 총 48개의 Positional Encoding이 위치 정보를 가진 토큰을 형성하여, **트랜스포머**에 **입력**
- **동작 토큰화**: 로봇의 동작 차원은 팔 움직임에 대한 7개 변수(x, y, z, 롤, 피치, 요, 그리퍼 열림)

https://blog.research.google/2022/12/rt-1-robotics-transformer-for-real.html

ViT(Vision Transformer)



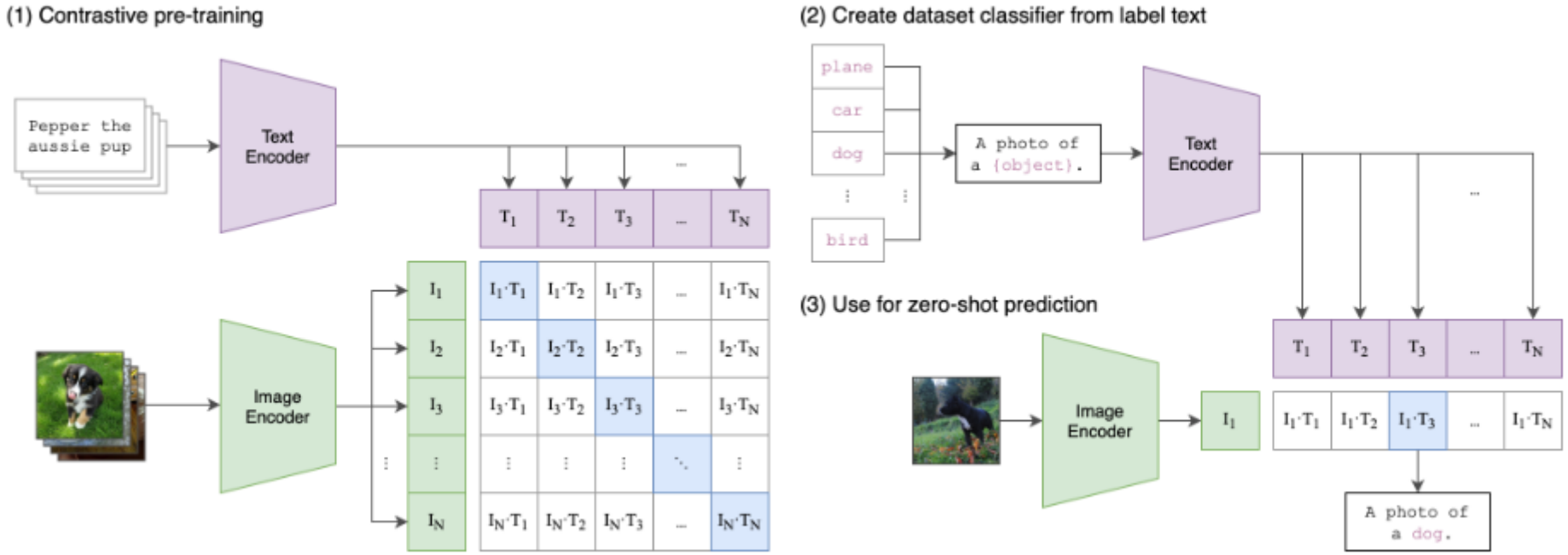
✓ **과정**

1. **패치 추출 및 평탄화(Flattening)**
   - ViT는 우선 이미지를 작은 패치로 나누고, 각 패치는 평탄화되어 1차원 배열로 변환
2. **선형 투영(Linear Projection)**
   - 평탄화된 패치는 선형 투영을 거쳐 고정된 크기의 벡터로 변환
3. **위치 임베딩(Position Embedding)**
4. **트랜스포머 인코더(Transformer Encoder)**
   - 다중 퍼센트론(Multi-Layer Perceptron, MLP)
   - MSP(Multi-Head Self Attention Layer)
   - LN(Layer Norm)
5. **MLP Head**
   - 인코더를 통과한 후, [class] 토큰의 출력은 MLP Head로 전달

[1] https://wikidocs.net/137253
[2] 강의명: 생성형 인공지능 입문, 제공 기관: K-MOOC, URL: www.kmooc.kr

CLIP (Contrastive Language-Image Pre-Training)

- 각 인코더를 통과해서 나온 N개의 이미지, 텍스트 특징 벡터들 사이의 올바른 (텍스트, 이미지) 쌍을 학습
- 데이터셋을 분류를 위한 클래스 출력

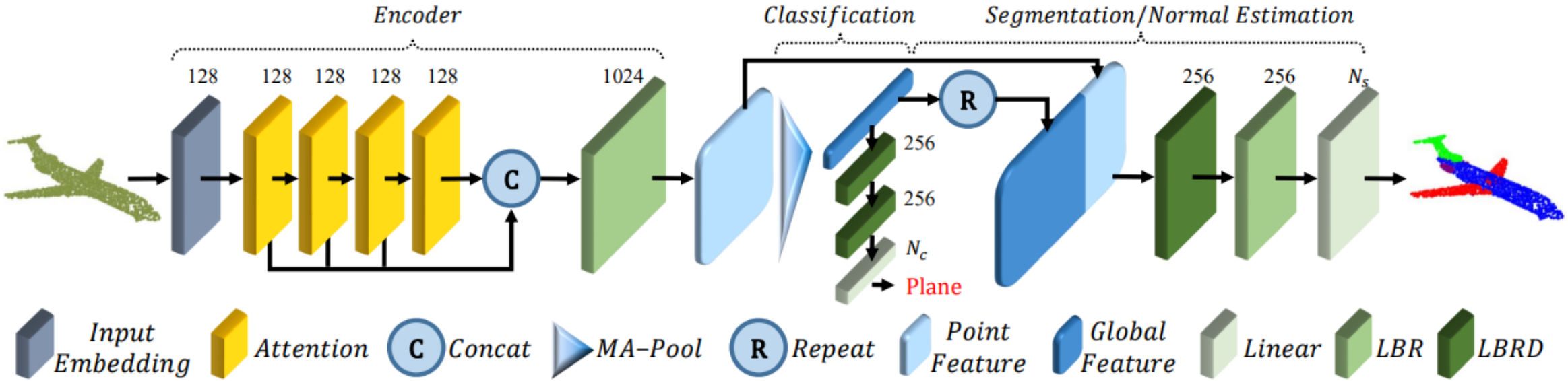


https://github.com/openai/CLIP

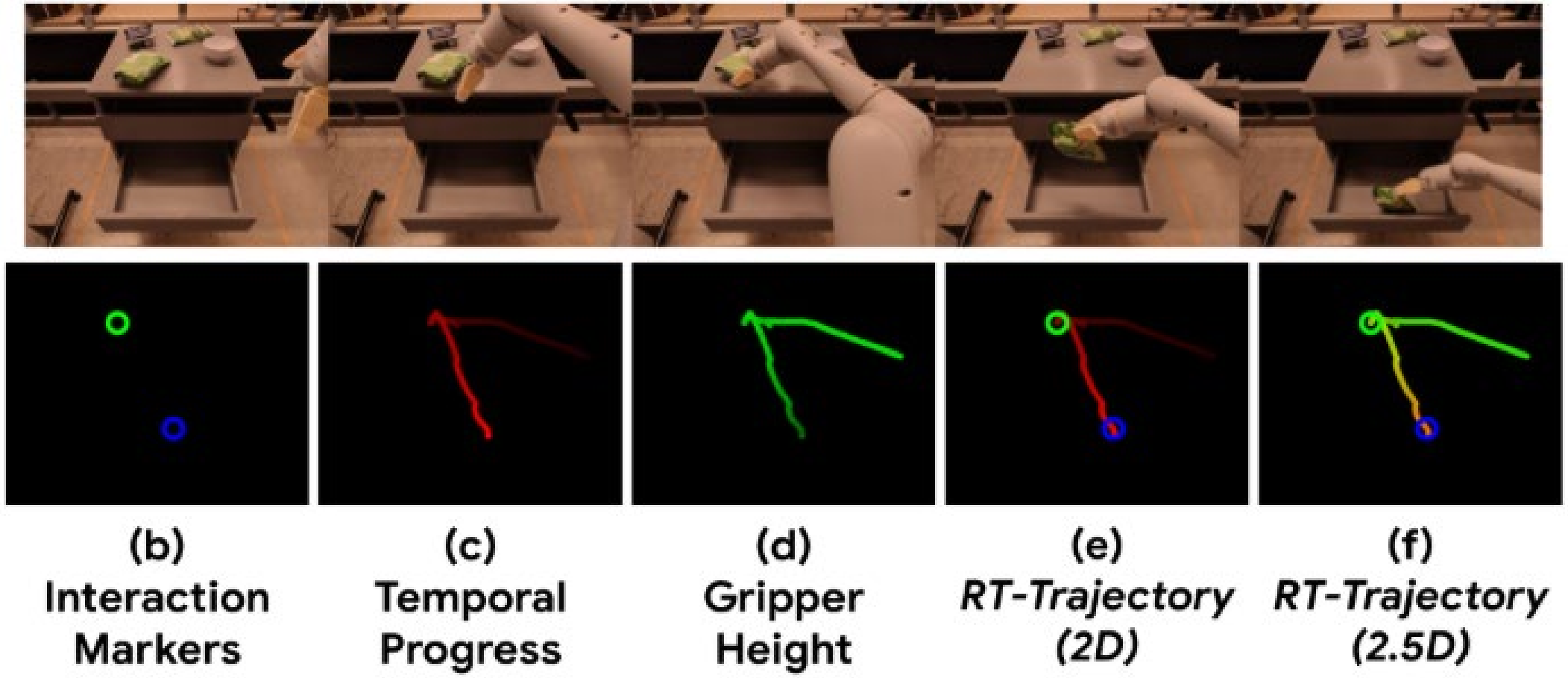
PCT: Point Cloud Transformer

✓ 구성

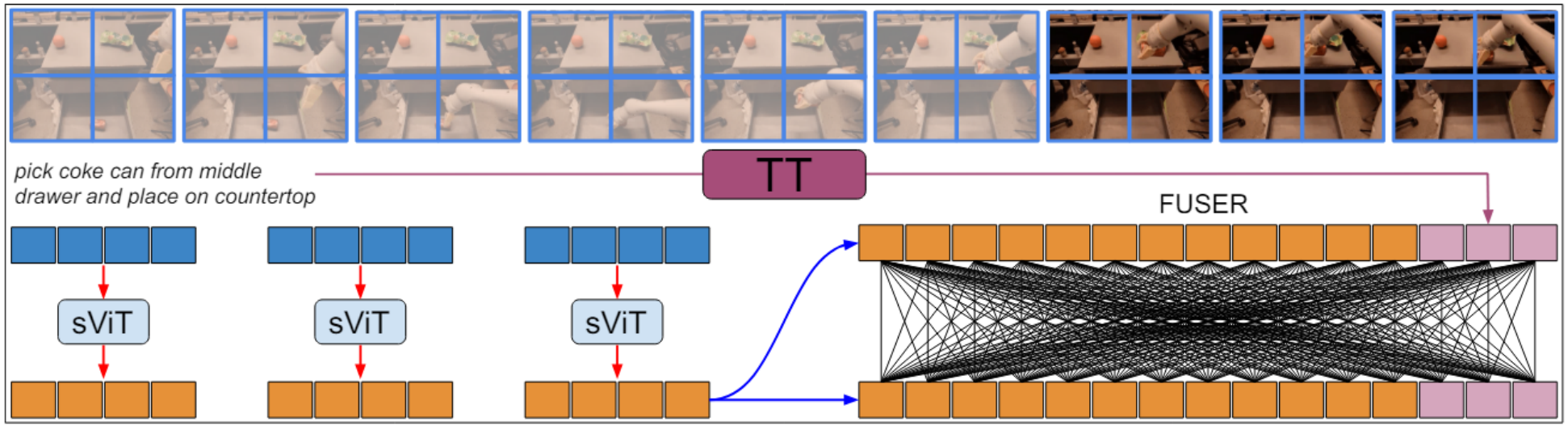
- **MA-Pool**
  최대 풀링(Max Pooling)과 평균 풀링(Average Pooling)의 연산을 결합
- **LBR**
  Linear, 배치정규화(Batch Normalization, BN), ReLU layer 결합
- **LBRD**
  LBR에 Dropout layer 결합



2D 스케치 이미지 예시


(b) Interaction Markers
(c) Temporal Progress
(d) Gripper Height
(e) RT-Trajectory (2D)
(f) RT-Trajectory (2.5D)

SARA-RT



Theorem 3.3: Consider normalized Transformer's attention layer with queries $(\mathbf{q}_i)_{i=1}^M$ and keys $(\mathbf{k}_j)_{j=1}^N$ of length r. Denote: $\tau = \min_{i,j} \mathrm{K}(\mathbf{q}_i, \mathbf{k}_j)$, $\rho = \max_{i,j} \mathrm{K}(\mathbf{q}_i, \mathbf{k}_j)$.

Take $m = \lceil \frac{2\rho^2}{\delta^2 \tau^2} \log(2MN) \exp(-\frac{r^2}{A}) \rceil + 1$ for $A < 0$, $\delta > 0$. Then there exist $\mathbf{v} \in \mathbb{R}^m, \mathbf{G}_1, \mathbf{G}_2 \in \mathbb{R}^{m \times d}, f : \mathbb{R} \to \mathbb{R}$ such that the approximate attention matrix $\widehat{\mathbf{A}}$ (implicitly) given by the mappings $\phi_{f,u}^{\mathrm{SARA}}$ satisfies:

$$\|\mathbf{A} - \widehat{\mathbf{A}}\|_\infty \leq \delta \tag{8}$$