Ego4D R3M VRB

실제 사람의 영상으로 학습하는 Physical AI

20240600 최지안

영상에서의 타깃의 상태변화



0.5 s



4 s

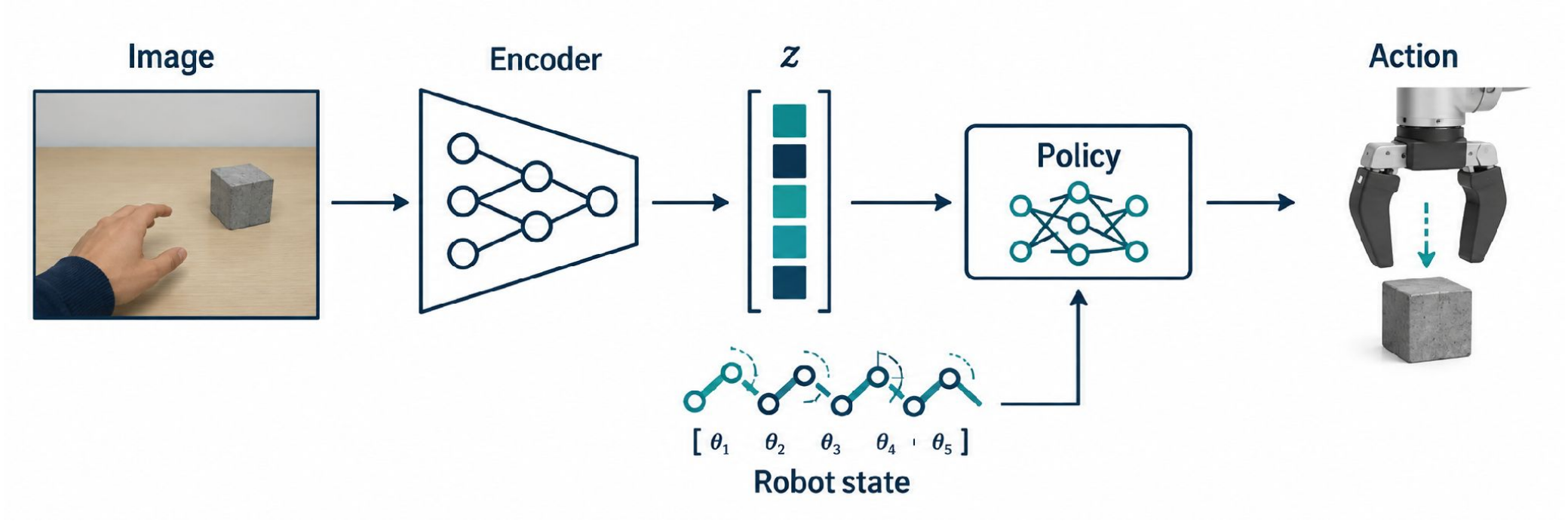


8 s

손의 접근, 물체의 이동, 이후 배치를 시간 순서로 관찰한다.

Ego4D 공개 cmu_1 MP4

Visual Representation과 Policy

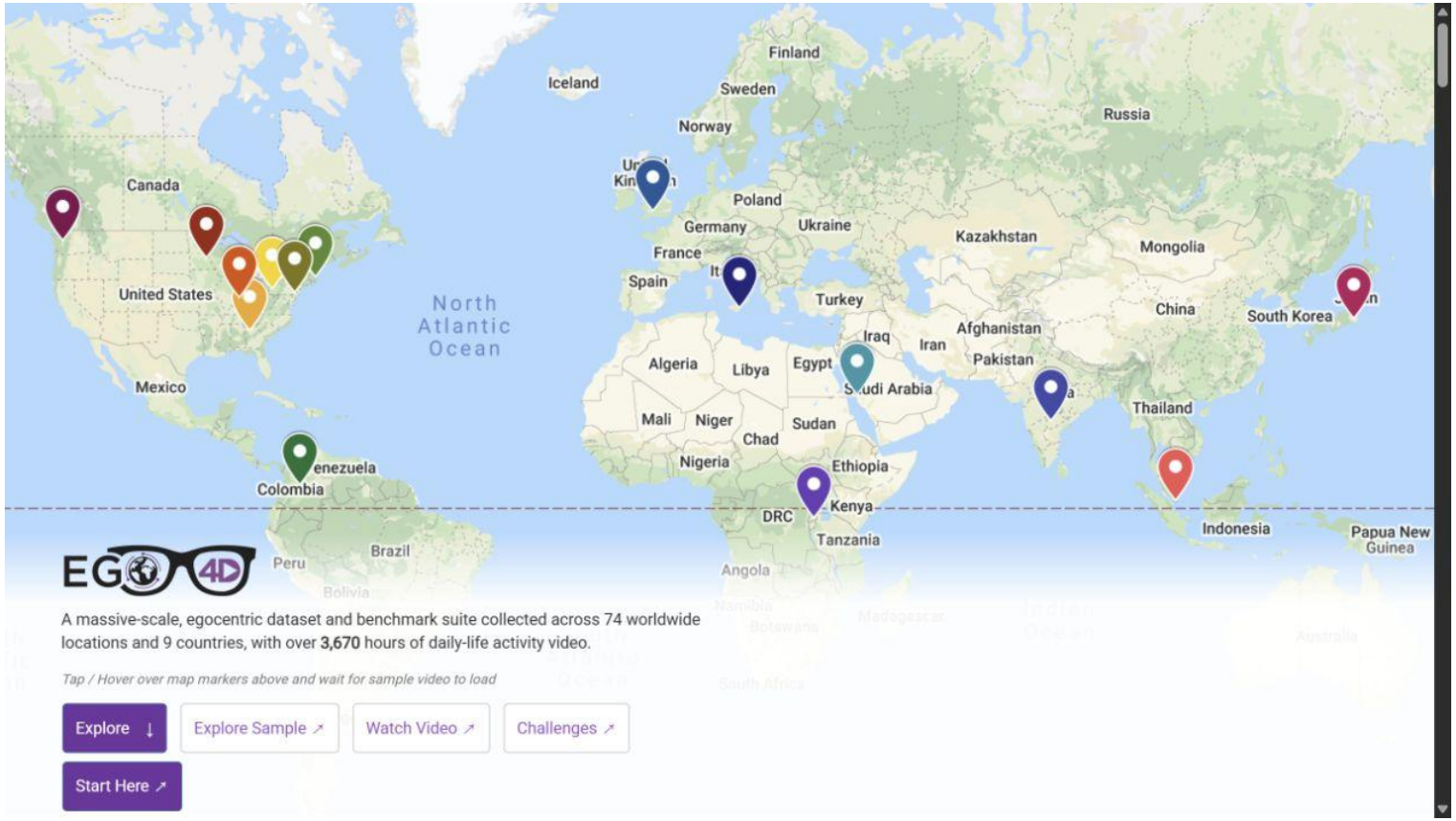


Visual Representation z는 장면의 특징을 담는다.

Policy는 그 특징과 로봇 상태를 받아 행동을 선택한다.

R3M의 역할 구분을 설명한 개념도 · 이미지 생성으로 제작

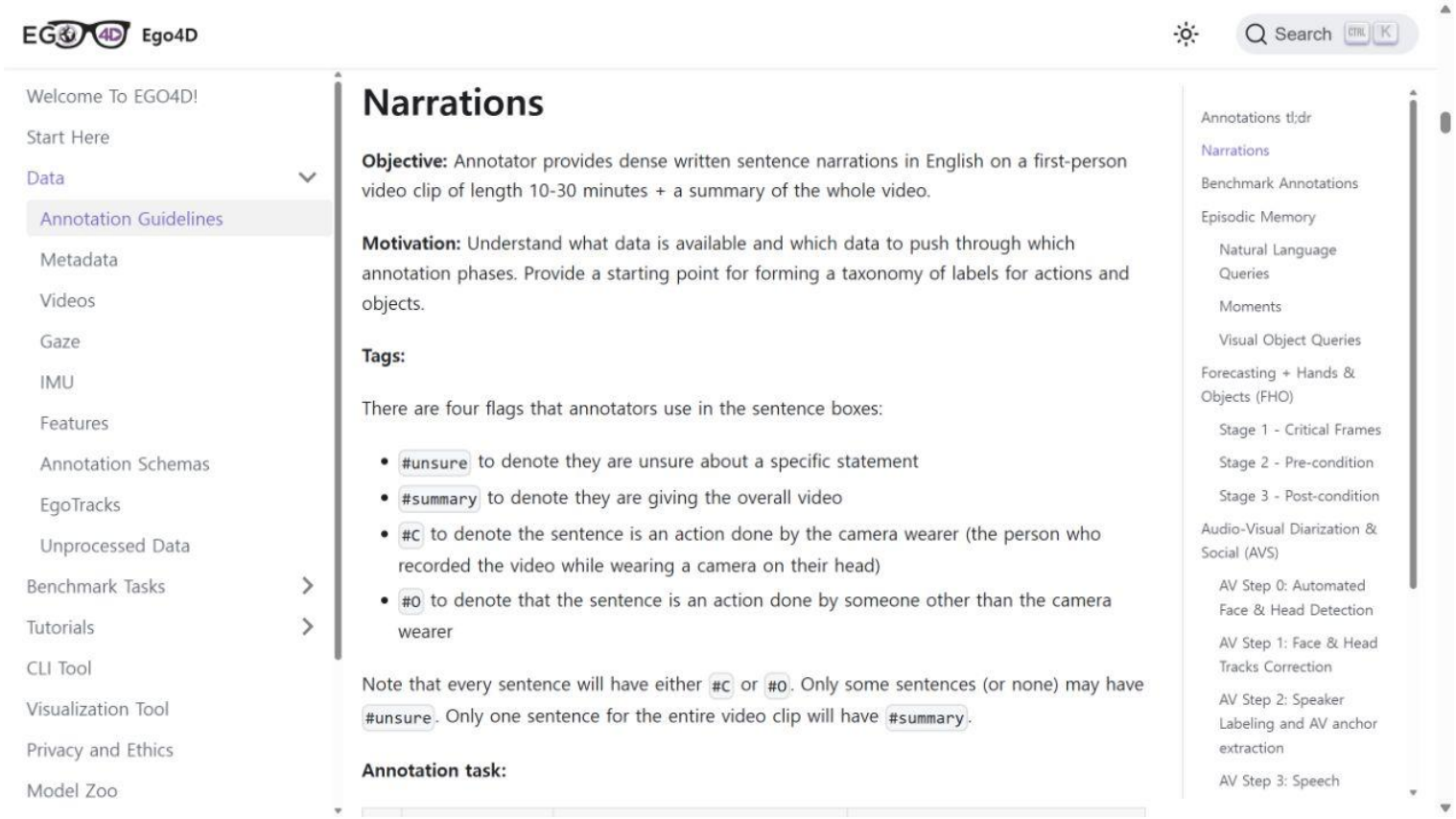
홈페이지에서 데이터의 수집범위



3,670시간의 일상 영상
9개국·74개 지역
일부 영상에 추가 센서

Ego4D 홈페이지 · 2026-09-15

Narrations: 관찰된 행동을 언어로 기록



#C: 카메라 착용자의 행동

#O: 다른 사람의 행동

#unsure: 불확실한 기술

Annotation Guidelines / Narrations · 2026-09-15

상태 변화의 단계: PRE, PNR, POST



State-change: Plant removed from ground

PNR은 상태 변화가 시작되는 프레임이다.

손의 움직임과 물체의 상태 변화는 구분해서 판단한다.

Grauman et al., Ego4D, Fig. 8 · Hand & Object Interactions

주석 형식

목표: making a fruit salad

계층	주석의 설명	시작 (s)	종료 (s)
4	Slice oranges	505.100	1083.061
4.1	Slice orange	508.326	530.501
4.2	Wear finger glove	530.635	536.245

Step 안에 substep이 들어가며, 같은 시간 축을 사용한다.

사람과 로봇은 몸과 시점이 다르다



사람의 손·착용 카메라

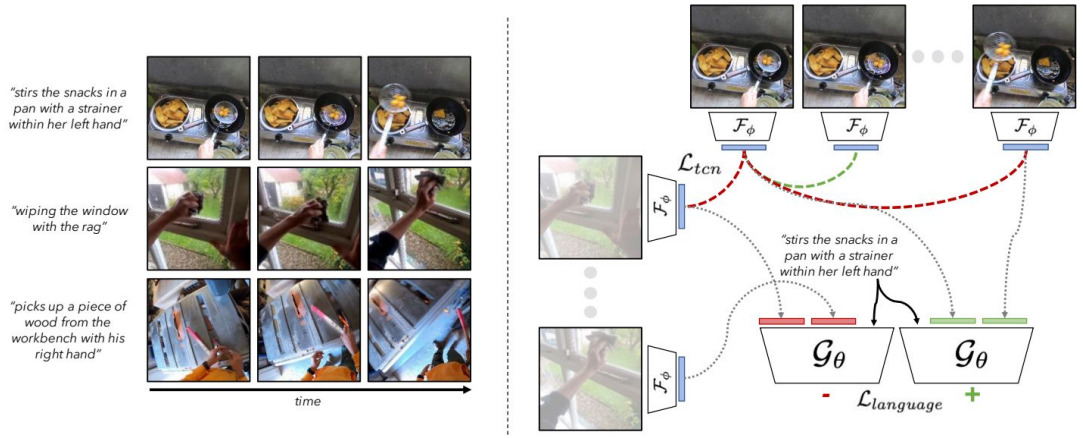


로봇 그리퍼·로봇 관찰

왼쪽 Ego4D 공개 샘플 / 오른쪽 R3M Fig. 5 · 서로 다른 작업 사례

R3M의 모델 구조와 학습 대상

R3M(Reusable Representations for Robotic Manipulation): 로봇 조작(Robot Manipulation) 작업을 위한 범용 시각 표현



이미지 I → 인코더 F_ϕ → 특징 z

L_{tcn} : 시간적 대조학습, 시점 간 특징 비교

$L_{language}$: 비디오–언어 정렬, 작업의 맥락을 읽을 수 있게 함

L1·L2 : 출력 특징의 크기 규제

$$\mathcal{L}(\phi,\theta) = \mathbb{E}_{I_0^{1:B}\,i\,i\,k\sim\mathcal{D}}\left[\lambda_1\mathcal{L}_{tcn} + \lambda_2\mathcal{L}_{language} + \lambda_3\|\mathcal{F}_\phi(I_i)\|_1 + \lambda_4\|\mathcal{F}_\phi(I_i)\|_2\right]$$

사전학습: 인코더 ϕ와 언어 점수 모듈 Θ를 갱신한다.

로봇학습: 인코더를 고정하고 별도 정책을 시연으로 학습한다.

Nair et al., R3M §3.3 · 원문 수식 재표기 및 수학 해설

로봇 시연의 관찰·행동 쌍으로 정책을 학습한다



$$z = F_{\phi}(I) \qquad \hat{a} = \pi_{\psi}(z, s)$$

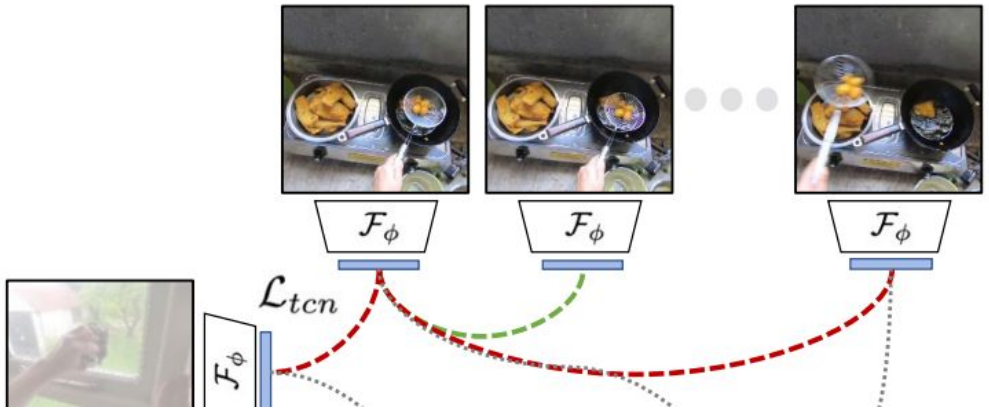
$$L_{BC} = E[\, \| \hat{a} - a_{demo} \|_2^{\,2} \,]$$

인코더 F_ϕ 고정
로봇 시연으로 정책 π_ψ 학습

관찰·로봇 상태 → 행동 예측 → 시연 행동과 비교 → 정책 갱신

R3M Fig. 1 / §4.1 · 식은 연속 행동 BC의 설명용 표기

Time Contrastive Learning의 InfoNCE 손실함수



$$\mathcal{L}_{tcn} = -\sum_{b \in B} \log \frac{e^{\mathcal{S}\left(z_i^b, z_j^b\right)}}{e^{\mathcal{S}\left(z_i^b, z_j^b\right)} + e^{\mathcal{S}\left(z_i^b, z_k^b\right)} + e^{\mathcal{S}\left(z_i^b, z_i^{\neq b}\right)}}$$

$i < j < k$ · 다른 영상의 특징 z^-

양성: $(z_i, z\square)$

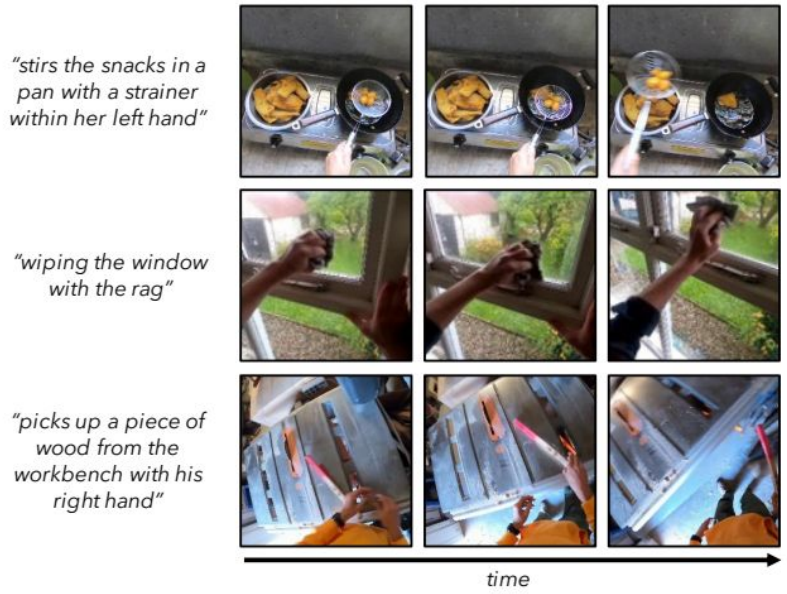
음성: $(z_i, z\square)$, (z_i, z^-)

$$S(x, y) = -\|x - y\|_2$$

한 영상의 손실을 표시 · 원문 Eq.1은 배치에 대해 합산

Nair et al., R3M §3.3 · 원문 수식 재표기 및 수학 해설

Video-Language Alignment의 수식



G_θ(초기 특징, 이후 특징, 언어) → 점수

g_late = $G_\theta(z_0, z\square, l)$

g_early = $G_\theta(z_0, z_i, l)$, $i < j$

g_wrong = $G_\theta(z_0^-, z\square^-, l)$

같은 영상의 더 이른 장면과,
언어에 맞지 않는 다른 영상을 비교한다.

$$\mathcal{L}_{language} = -\sum_{b \in B} \log \frac{e^{\mathcal{G}_\theta\left(z_0^b, z_{j>i}^b, l^b\right)}}{e^{\mathcal{G}_\theta\left(z_0^b, z_{j>i}^b, l^b\right)} + e^{\mathcal{G}_\theta\left(z_0^b, z_i^b, l^b\right)} + e^{\mathcal{G}_\theta\left(z_0'^b, z_{j>i}'^b, l^b\right)}}$$

올바르고 더 진행된 조합의 상대 점수를 높인다.

손실은 G_θ뿐 아니라 이미지 인코더 F_ϕ(=z)까지 역전파된다. (하나의 커다란 신경망)

Nair et al., R3M §3.3 · 원문 수식 재표기 및 수학 해설

L1·L2 Regularization의 역할

규제 대상은 출력 특징 $z = F_\phi(I)$ · 가중치 자체의 weight decay와 구분

	L1 norm	L2 norm
정의	$\|z\|_1 = \Sigma_r \|z_r\|$	$\|z\|_2 = \sqrt{(\Sigma_r z_r^2)}$
성분별 기울기	$\text{sign}(z_r)$ $(z_r \neq 0)$	$z_r / \|z\|_2$ $(z \neq 0)$

$$R(z) = \lambda_3 \|z\|_1 + \lambda_4 \|z\|_2$$

L1은 희소성을 유도하고, L2는 전체 크기를 억제한다.
시간·언어 손실과 함께 사용하며, 단위 길이로 정규화하지 않는다.

Nair et al., R3M §3.3 · 원문 수식 재표기 및 수학 해설

R3M의 실제 로봇 실험

Franka Panda · 과제당 로봇 시연 20개 · 과제당 평가 10회

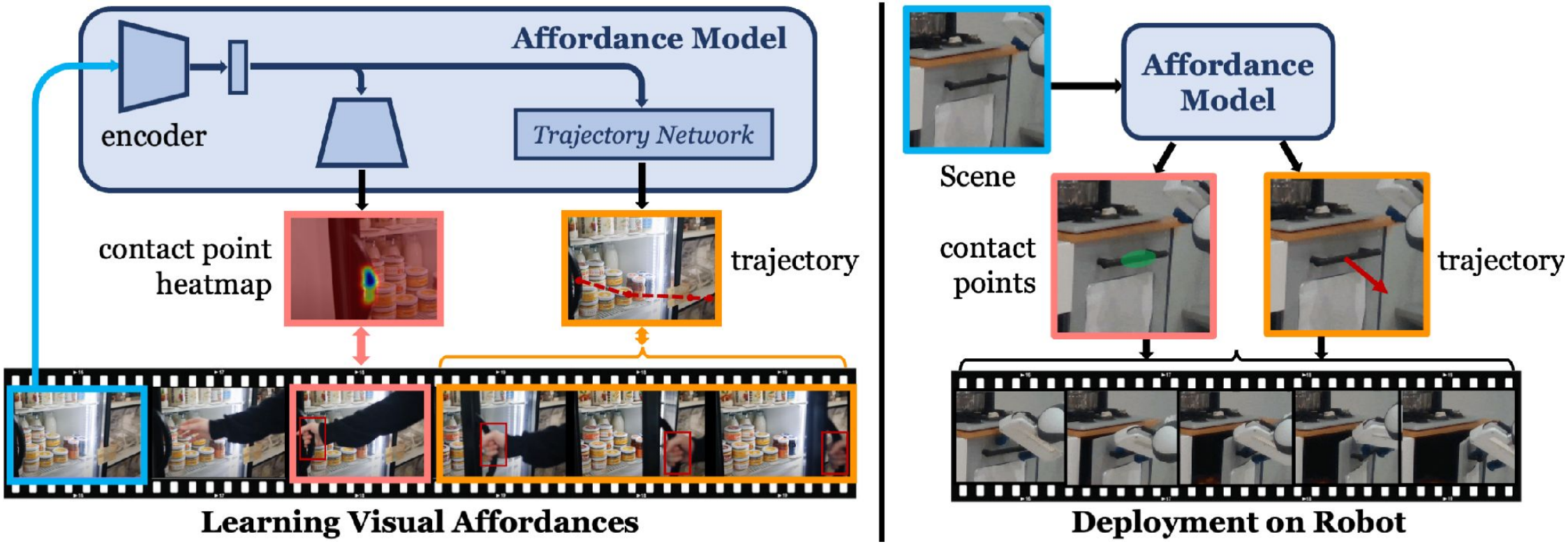


5개 과제 평균은 R3M 56%, CLIP 24%였다.
다른 환경과 학습 방식으로의 일반화는 별도 검증이 필요하다.

Nair et al., R3M, Table 3 · 단위: 성공률 (%)

VRB: Vision–Robotics Bridge

사람 영상에서 접촉 위치와 접촉 후 2D 손목 궤적을 배운다



모델 입력: 장면 이미지 → 예측: 접촉 위치·궤적 → 활용: 로봇 학습

Bahl et al., CVPR 2023 · Affordances from Human Videos · 원논문 개요 그림

VRB는 접촉 위치와 접촉 후 손목 궤적을 추출한다



Frame w/ contact



Post contact traj



First frame w/ human

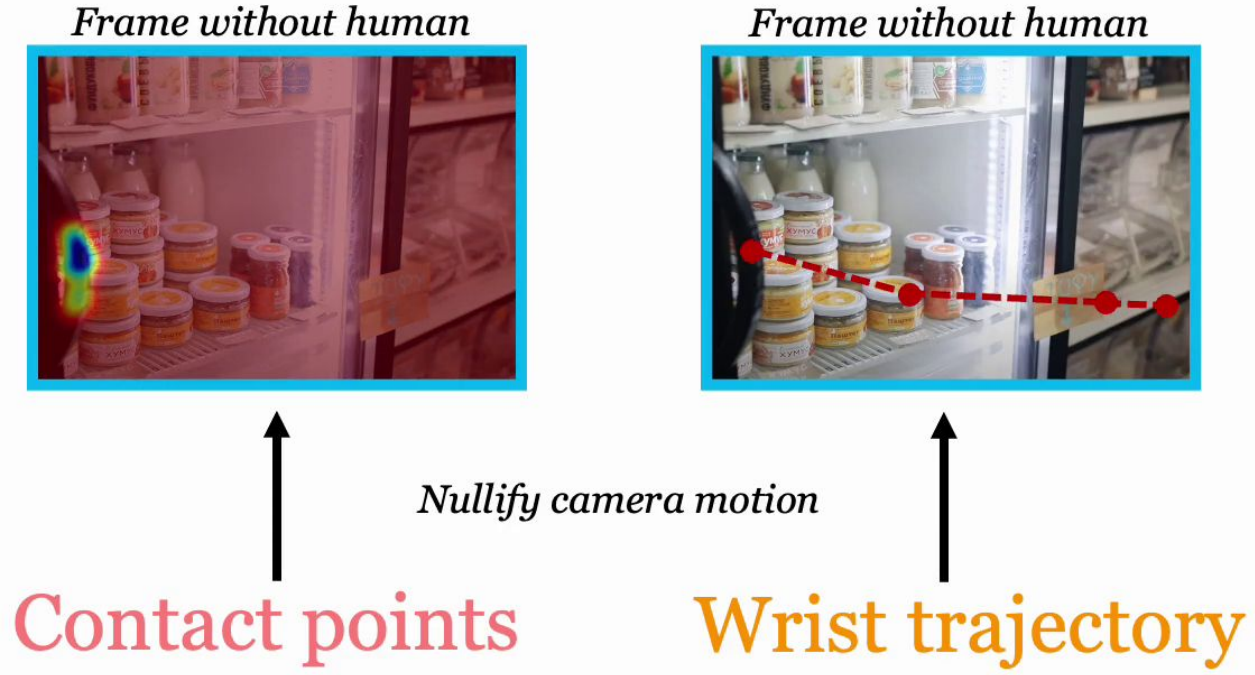
접촉을 탐지하고 2D 손목 궤적을 추적해 학습 라벨을 만든다.
카메라 움직임을 보상하고 기준 프레임에 정렬한다.

Bahl et al., VRB, §3.2.1 · 공식 프로젝트의 추출 과정 그림

카메라 움직임을 보상해 기준 프레임에 정렬한다



Human video

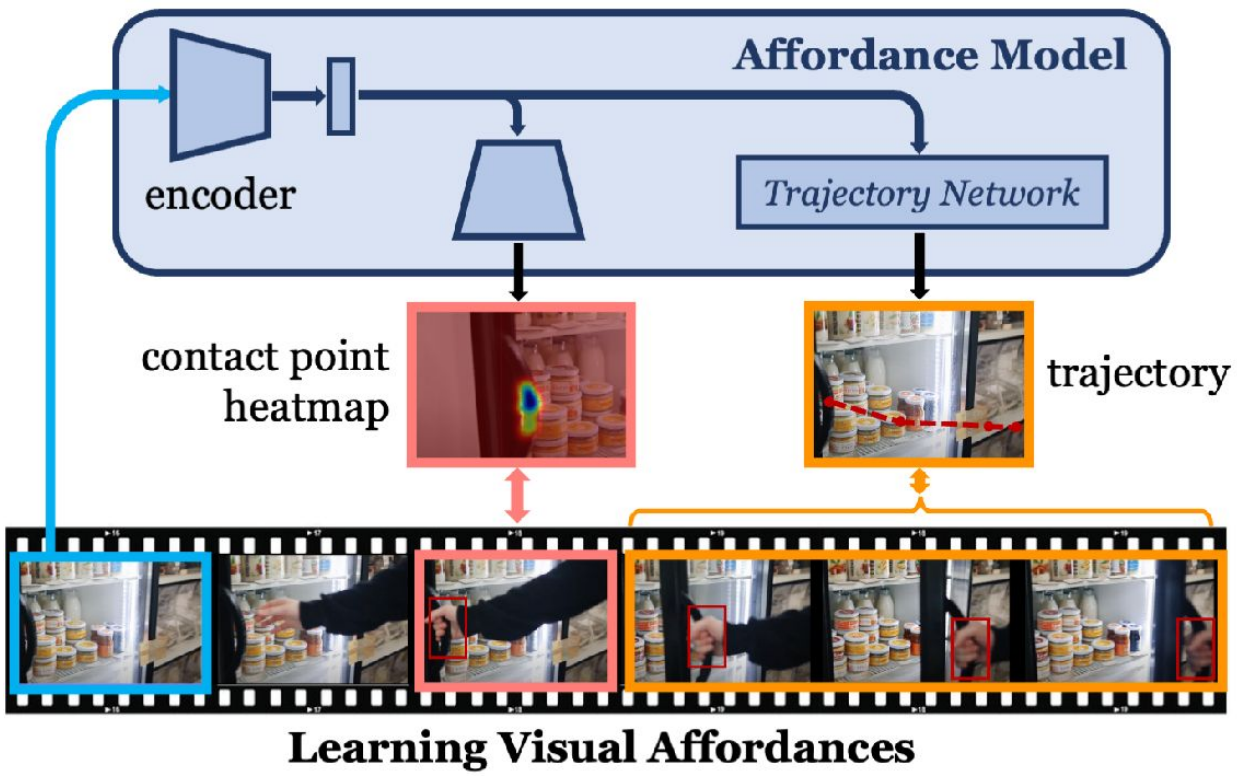


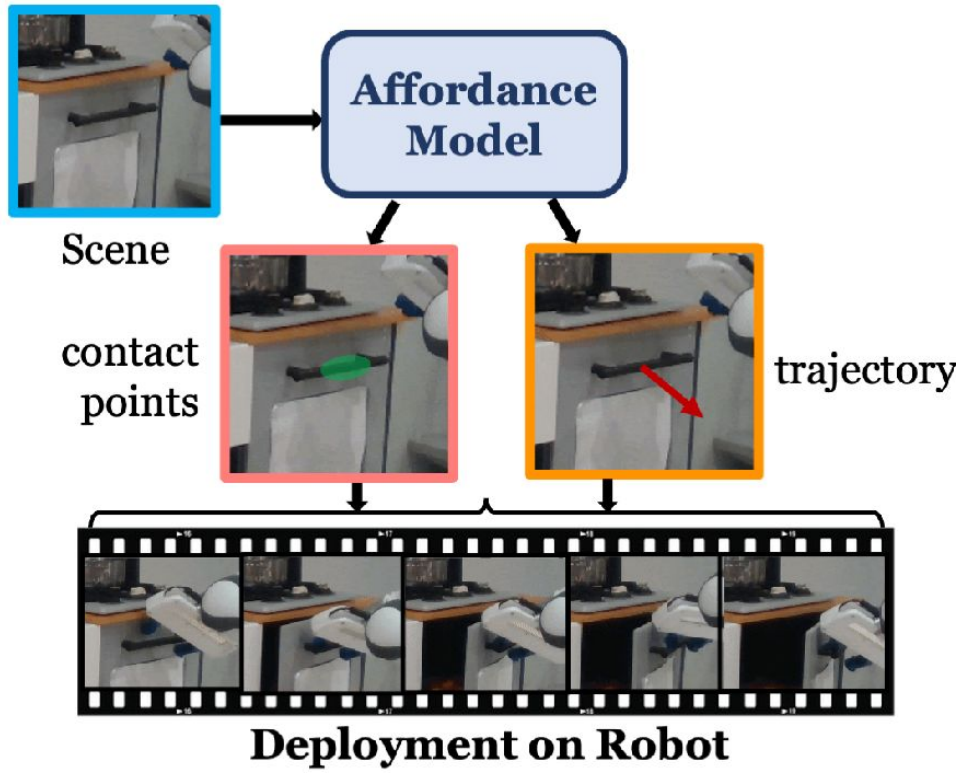
손목의 화면 이동에는 카메라 움직임도 섞인다.

Homography로 접촉·궤적을 기준 프레임에 정렬한다.

VRB 공식 프로젝트 영상 프레임 · human_vrb / cropped_without_human

VRB는 접촉 위치와 2D 손목 궤적을 예측한다



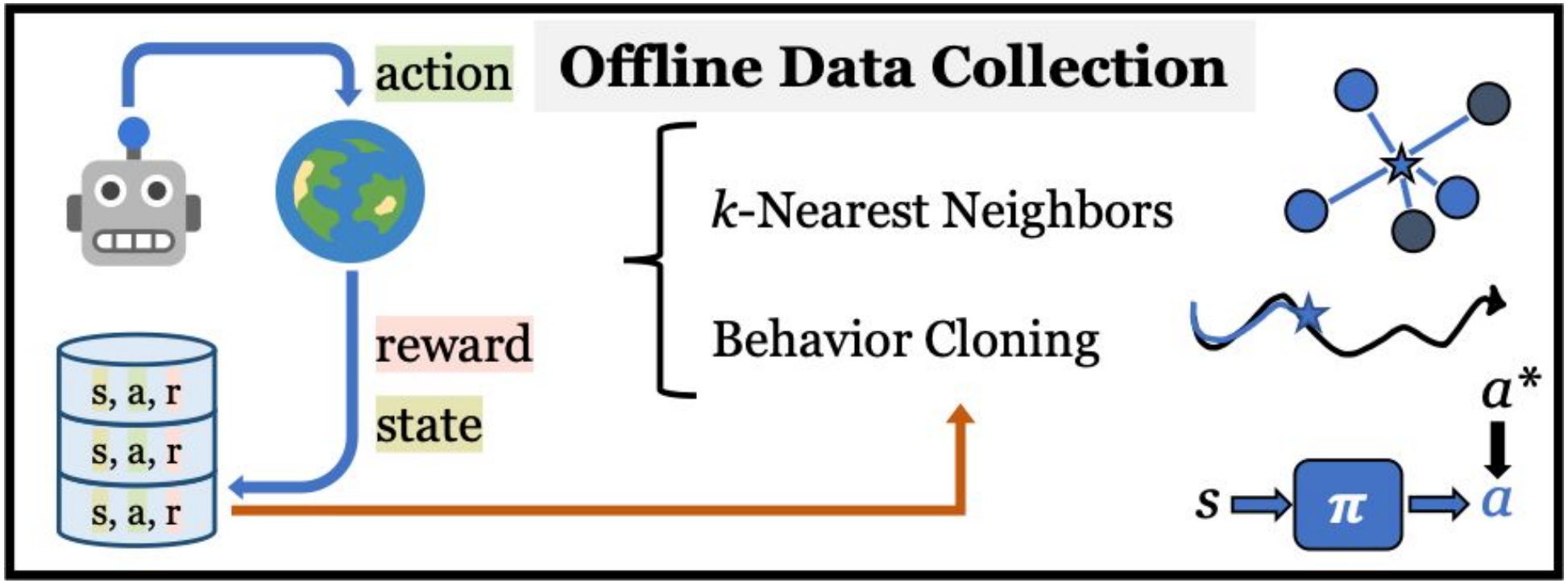


$$L_{contact} = \|\mu - \hat{c}(I)\|_2 \qquad L_{traj} = \|\tau - T\theta(z)\|_2$$

μ·τ: 추출한 접촉 위치·2D 손목 궤적 / ĉ·Tθ: 예측값 / z: 이미지 특징

VRB §3.2.2, Eq. (2) · ĉ(I)는 접촉 위치 예측 함수를 줄여 쓴 표기

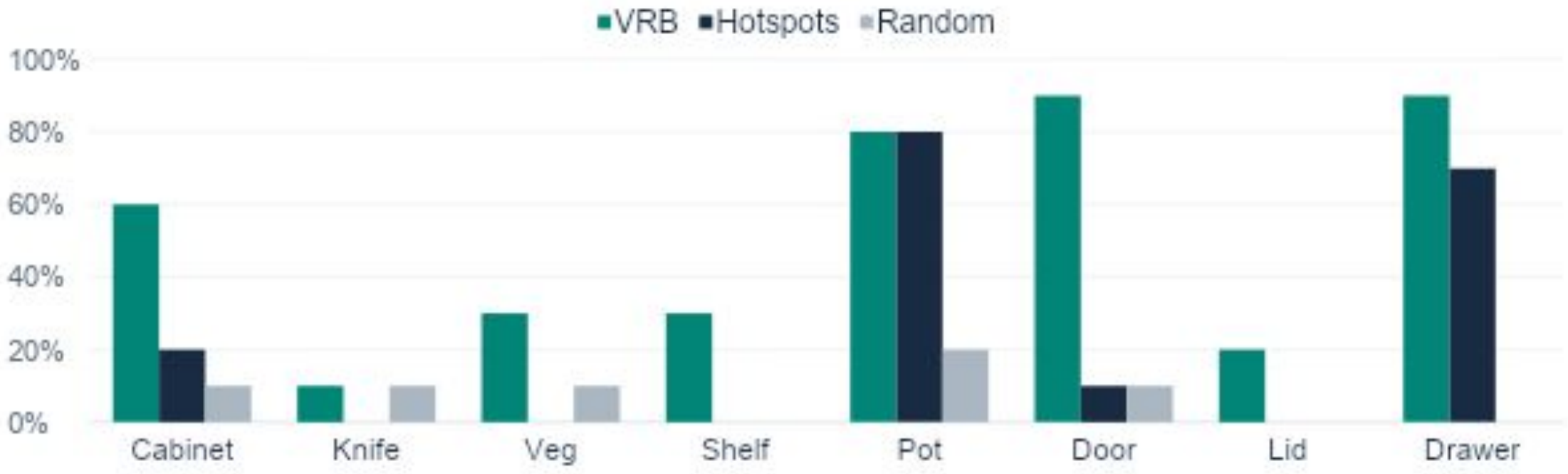
Affordance로 데이터를 모은 뒤 정책을 학습한다



접촉·2D 궤적 예측 → 로봇 데이터 수집 → 목표 관련 데이터 → 정책 학습

VRB의 데이터 수집 효과를 후속 정책의 성공률로 평가한다

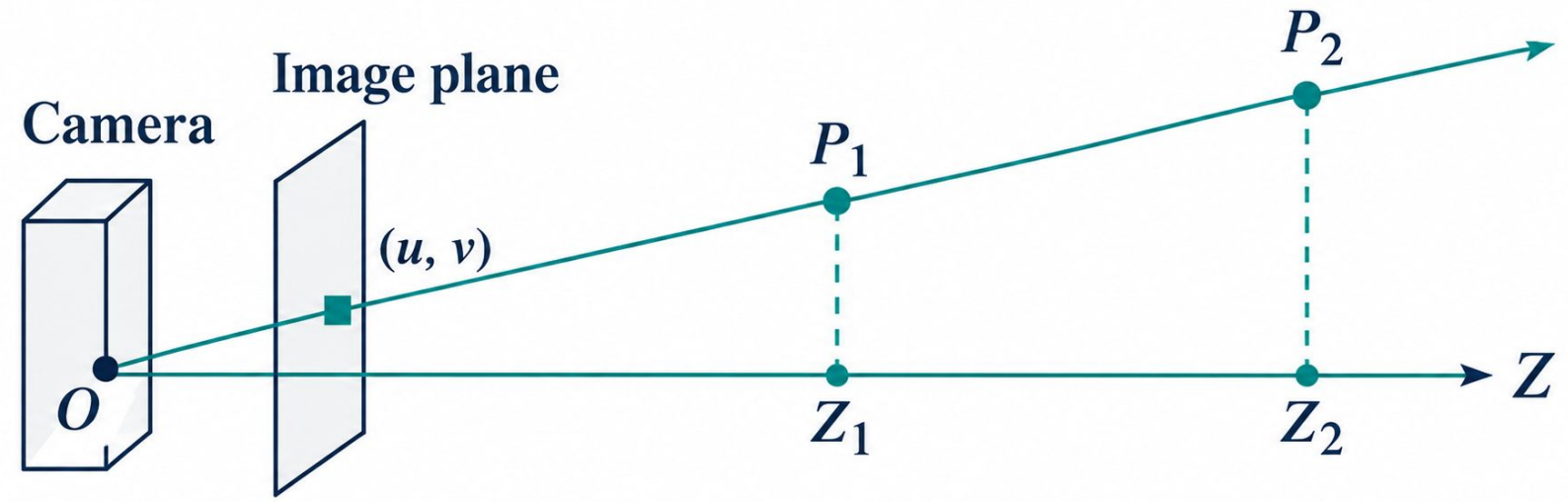
Behavior Cloning · 수집 30–150 에피소드 · 과제당 평가 10회



Affordance는 로봇이 데이터를 모으는 방향을 제공한다.
수집된 데이터로 정책을 학습하는 단계가 뒤따른다.

Bahl et al., VRB, Table 1의 BC 결과 일부 · 단위: 성공률 (%)

같은 픽셀에 서로 다른 3D 위치가 대응한다



$X = (u - c_x) Z / f_x \qquad Y = (v - c_\gamma) Z / f_\gamma$

Z: 광축 방향 깊이 / c_x, c_γ: 주점의 픽셀 좌표

보정된 핀홀 카메라를 가정한 개념도 · 이미지 생성으로 제작 / Colab E1

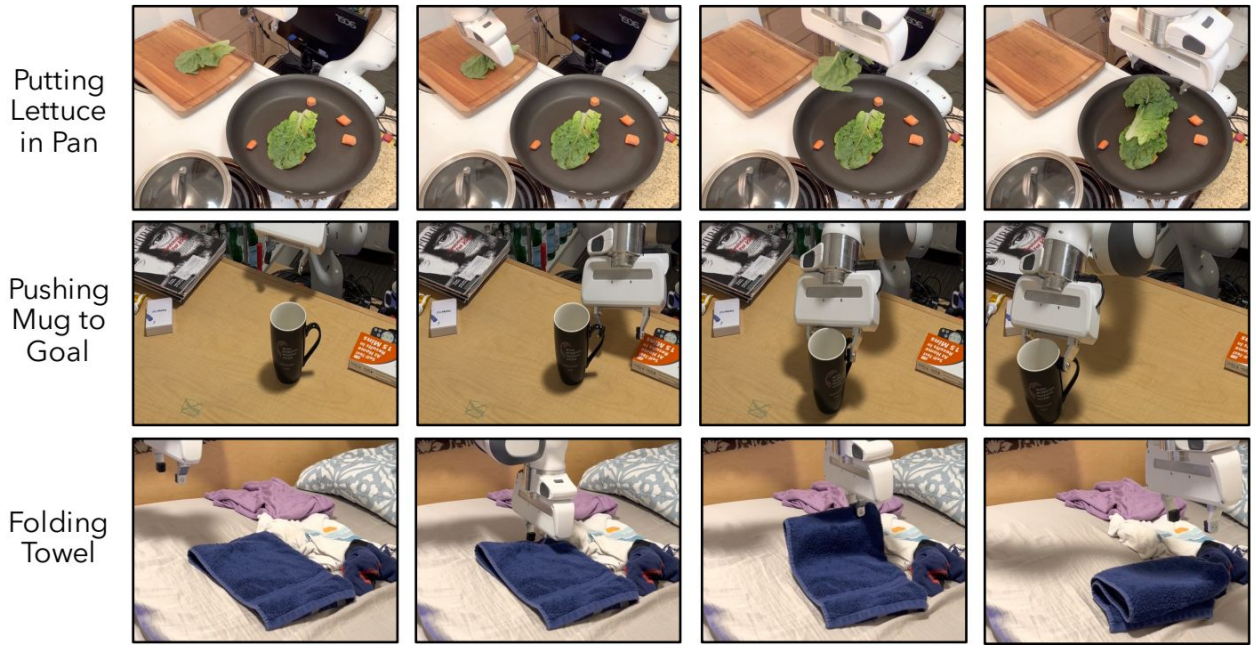
다른 데이터는 시점·기하·로봇 행동 정보를 보완한다

데이터	중심 정보	보완하는 부분
Ego4D	일상 1인칭 영상·주석	행동과 상태 변화의 맥락
Ego-Exo4D	동기화된 다중 시점·포즈	시점·기하·숙련 활동
Open X-Embodiment	로봇 관찰·행동 에피소드	실제 로봇 행동 데이터

영상 시간만 비교하기보다, 필요한 정보와 제공 범위를 확인해야 한다.

Ego4D · Ego-Exo4D 공식 문서 · Open X-Embodiment 저장소

영상에서 배운 정보는 로봇 실행으로 검증해야 한다



영상의 접촉 추정 ≠ 힘 측정이다.
예측 성능과 작업 성공률을
로봇 환경에서 각각 평가한다.

R3M, Fig. 5 및 §5 · VRB §4–5

Human video는 로봇 학습에 유용한 학습 신호를 준다

영상은 행동의 맥락과 물체의 변화를 제공한다.
표현 또는 접촉·이동 정보로 로봇 학습에 연결할 수 있다.
그 효과는 추가 데이터와 실제 제어 조건 안에서 평가해야 한다.