논문 세미나

RILS LAB

Robot Intelligence Learning & System

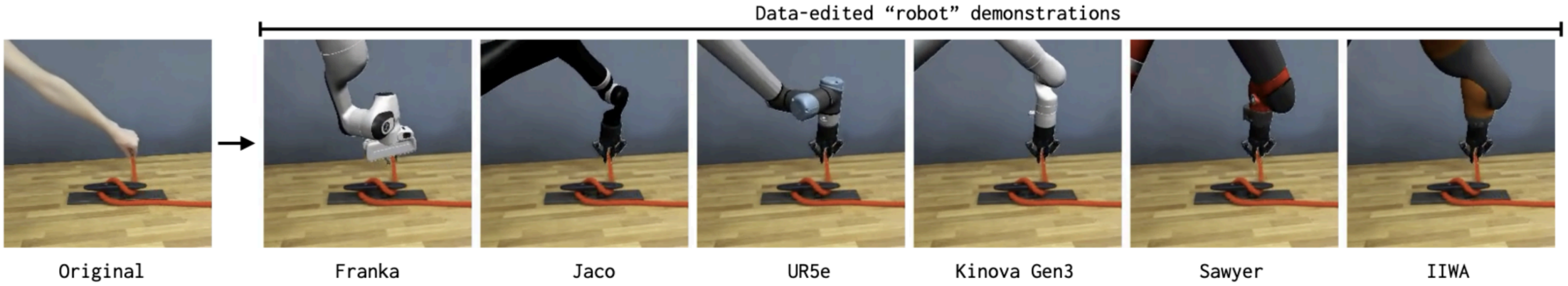
신승철

2026. 08. 28

연구 분야

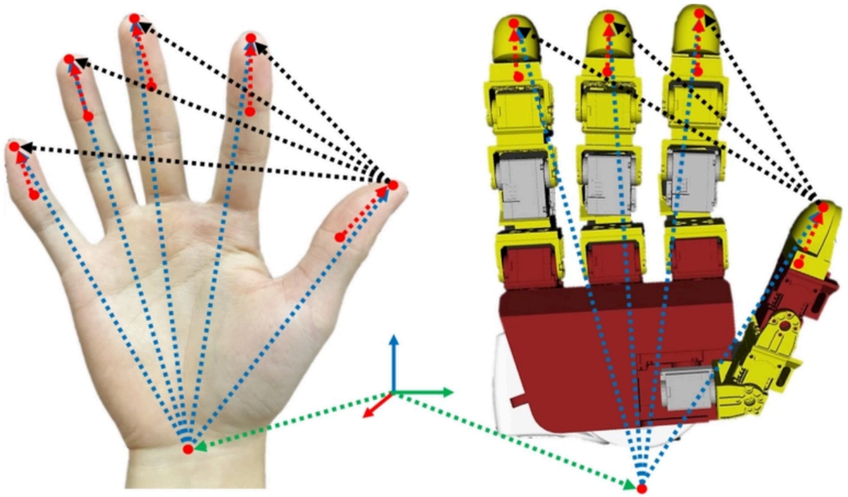
해결하고자 하는 문제 : robot 데이터 수집의 어려움 및 데이터 부족
아이디어 : 세상에 널린 human video를 robot data로 변환해서 사용하자

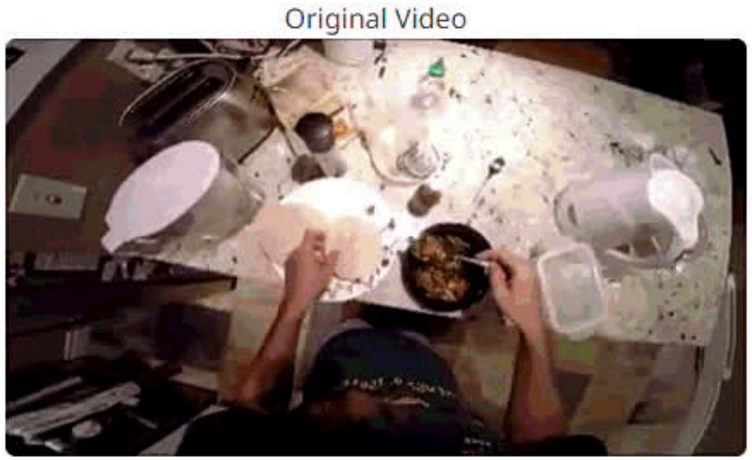
이를 통해 대규모 image, text 데이터셋 크기에 대응할만한 robot dataset을 만들자

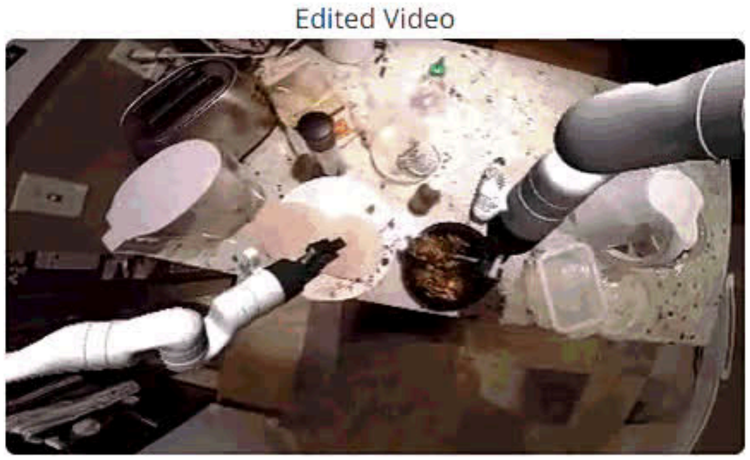


이러한 연구를 발전하기 위해 필요한 지식들이 뭐가 있을까??

- human hand → robot hand/gripper action trajectory 변환
- cross embodiment visual gap 해소
- zero/one shot transfer







human to robot
cross embodiment visual gap 해소하기 위해서
비디오를 편집하자!!

01

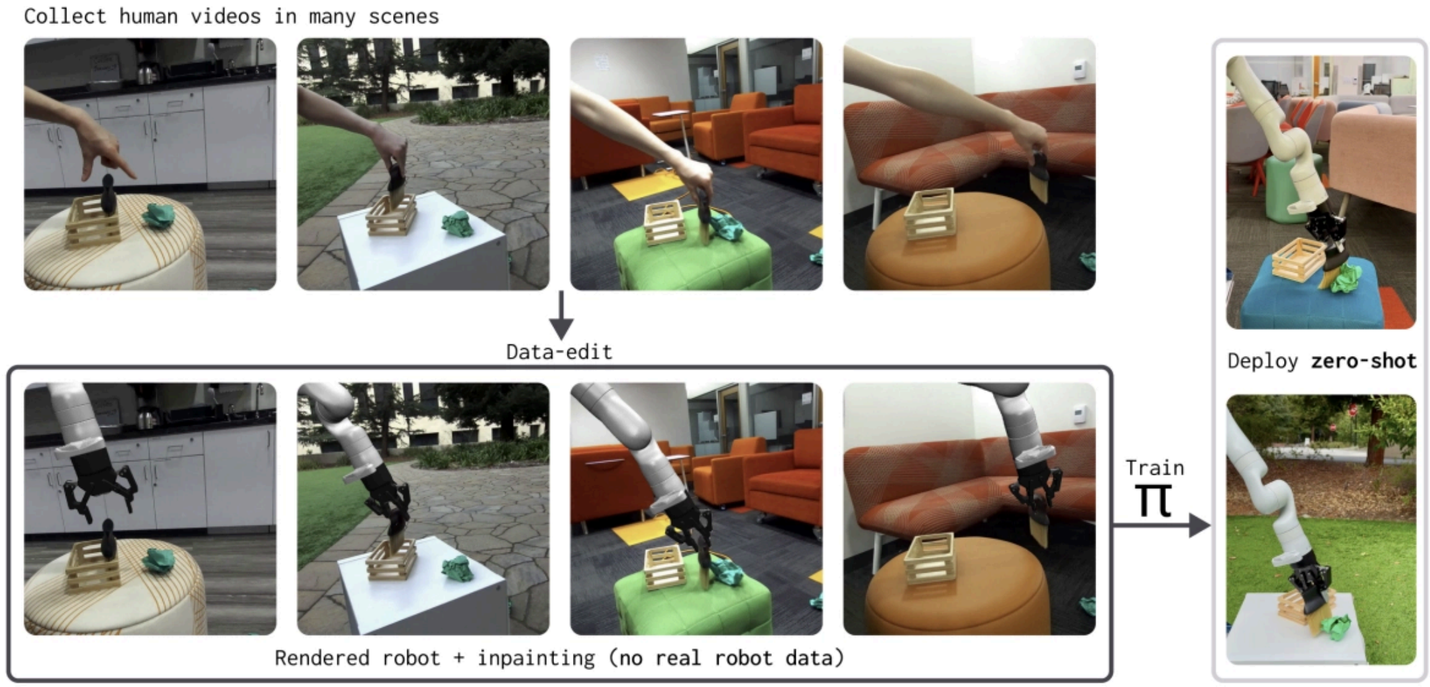
Phantom: Training Robots Without Robots Using Only Human Videos

Marion Lepert, Jiaying Fang, Jeannette Bohg

CoRL / 2025

문제 상황
train 시에 사람 손과 팔을 위주로 학습하다가
test시에 로봇 팔을 보면 학습한 것과 gap, 성능 하락

아이디어
data editing을 사용해서 human to robot tranfer을
진행해서 데이터를 변환하자

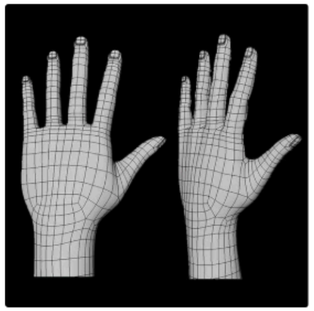


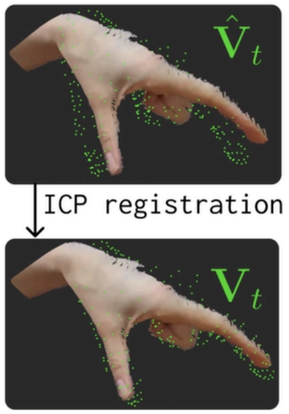
핵심 아이디어

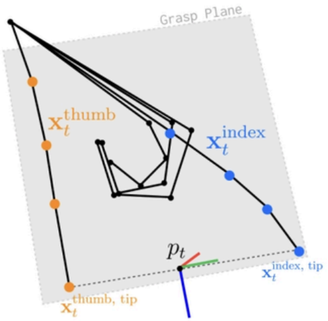
1. hand pose → robot action
3D pose와 mesh를 추정하는 hand reconstruction 모델인 HaMeR와 ICP방식으로 hand pose 추정해서 변환

$\hat{X}_t \in \mathbb{R}^{21\times3}$

$\hat{V}_t \in \mathbb{R}^{778\times3}$



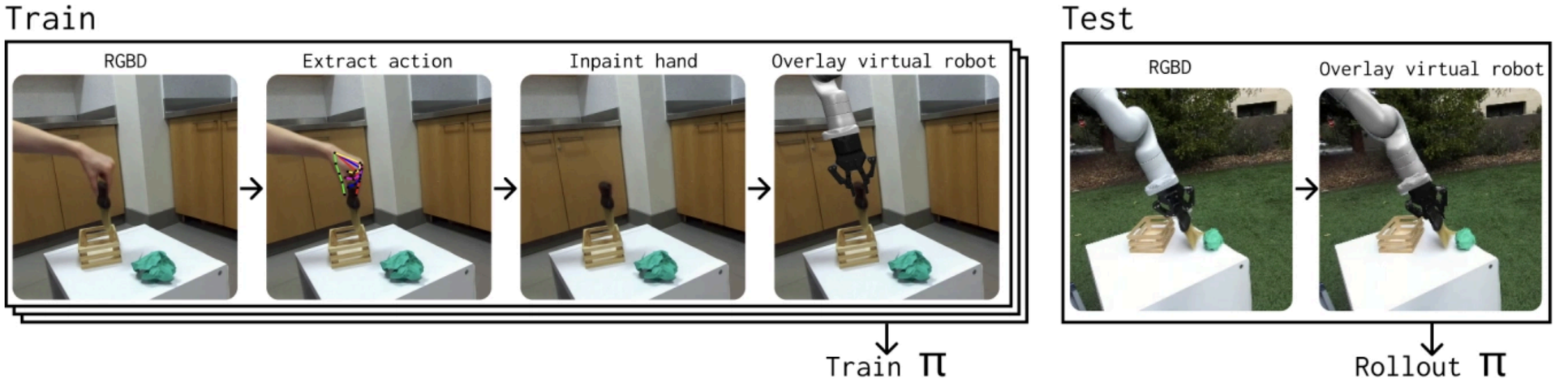




2. Data editing
- SAM2로 사람 손과 팔 영역을 segmentation
- E2FGVI video inpainting으로 사람 손/팔을 이미지에서 제거
- 앞에서 hand pose를 추정하며 얻은위치에 virtual robot을 렌더링하고 렌더링한 로봇을 원본 장면 위에 overlay

결과

데이터 파이프라인



Data editing 성능

	Pick/ Place Book	Stack Cups	Tie Rope	Rotate Box
Hand Inpaint	0.92	0.72	0.64	0.72
Hand Mask	0.92	0.52	0.60	0.76
Red Line	0.0	0.0	0.0	0.0
Vanilla	0.0	0.0	0.0	0.0

human video 사용 성능

# of demos	Robot only	Human only
25	0.16	—
50	0.52	0.44
100	0.88	0.64
300	—	0.84

02

Masquerade: Learning from In-the-wild Human Videos using Data-Editing

Marion Lepert, Jiaying Fang, Jeannette Bohg

ICRA / 2026

문제 상황
의도적으로 잘 촬영하고 고정된 카메라로 촬영하는
인간 demonstration은 현실성이 부족

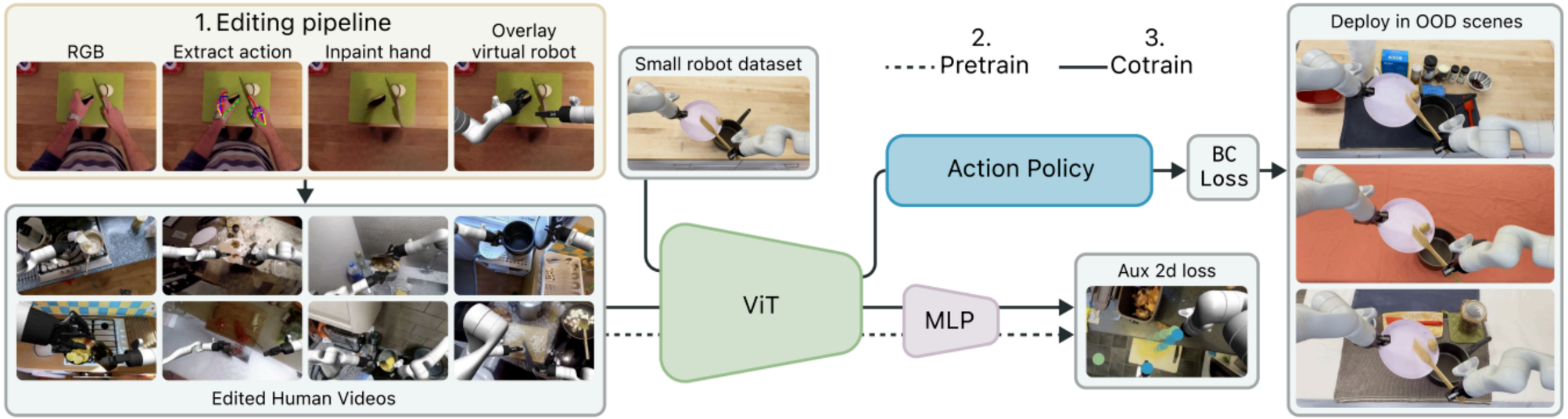
아이디어
실제 환경에서 자연스럽게 촬영된 in-the-wild 1인칭 인간
영상 데이터를 robot data로 변환하자

long horizon task/bimanual에도 가능하게 적용해보자

핵심 아이디어

Editing pipeline
In-the-wild egocentric human video에서 **2D hand pose를 추출**하고,
인간의 팔을 inpainting으로 제거한 뒤, 동일한 자세를 취하는 **렌더링된 양팔 로봇을 영상 위에 합성**한다 (phantom과 동일)

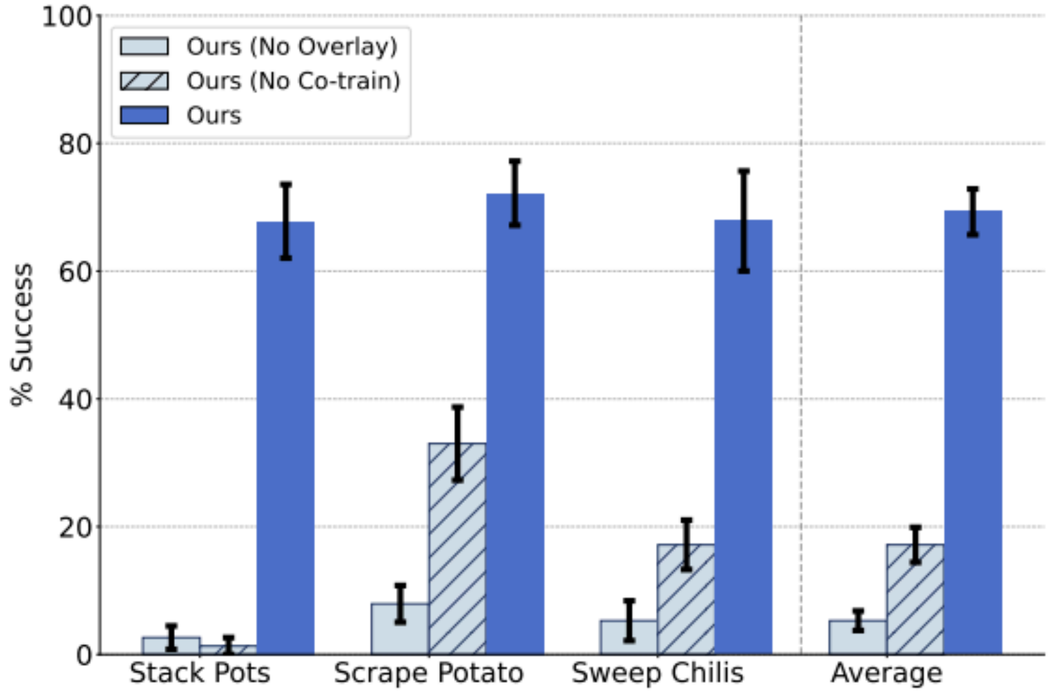
Co-training
edited human video에는 **auxiliary 2D loss를 적용**하고
실제 robot demonstration에는 **imitation loss를 적용**하여 훈련 진행



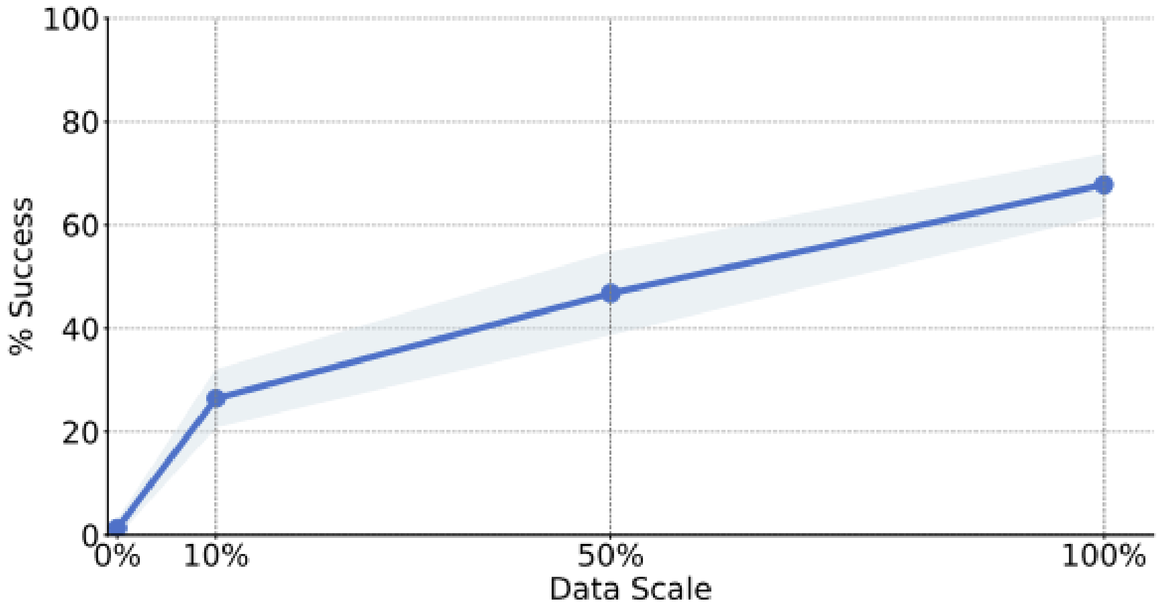
human video를 통해 → 화면에서 어디로 움직여야 하는지
real robot video를 통해 → 그러한 경로를 robot action으로 어떻게 변환할지

결과

co-train, overlay ablation 실험



Amount of In-the-wild Data 성능 실험



human video를 보고 훈련 없이
바로 로봇 액션을 수행할 수 있게 해보자!

03

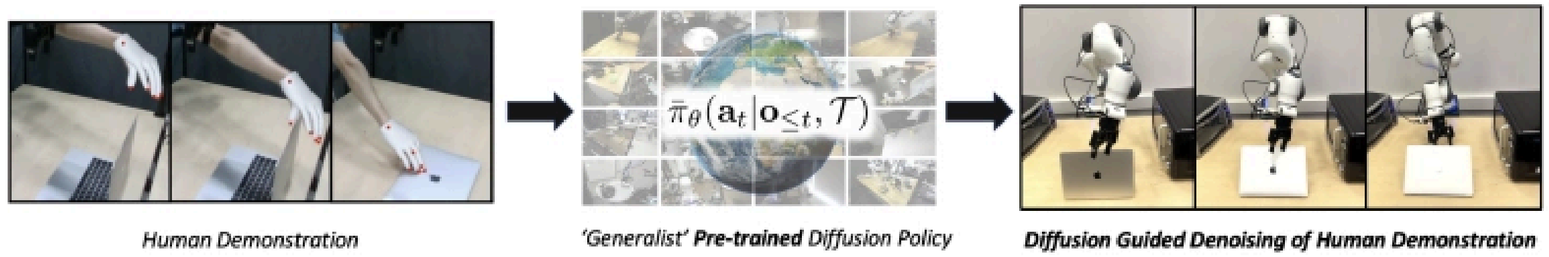
DemoDiffusion: One-Shot Human Imitation using pre-trained Diffusion Policy

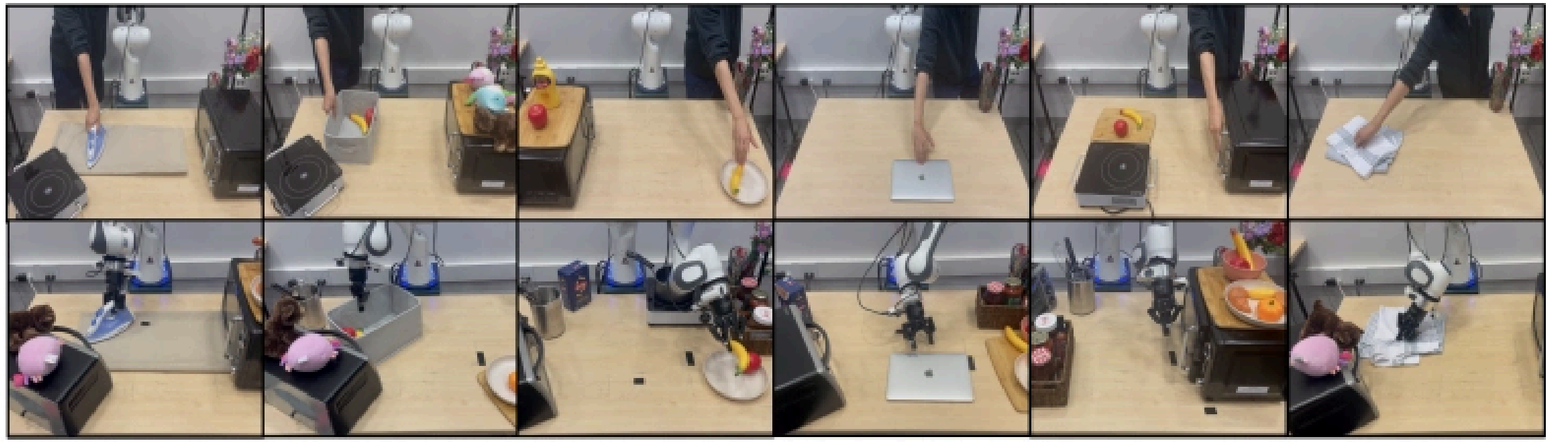
Sungjae Park, Homanga Bharadhwaj, Shubham Tulsiani

ICRA / 2026

문제 상황
새로운 환경에서 zero-shot 혹은
새로운 task에 굉장히 성능이 낮고, finetuning 불가피함.

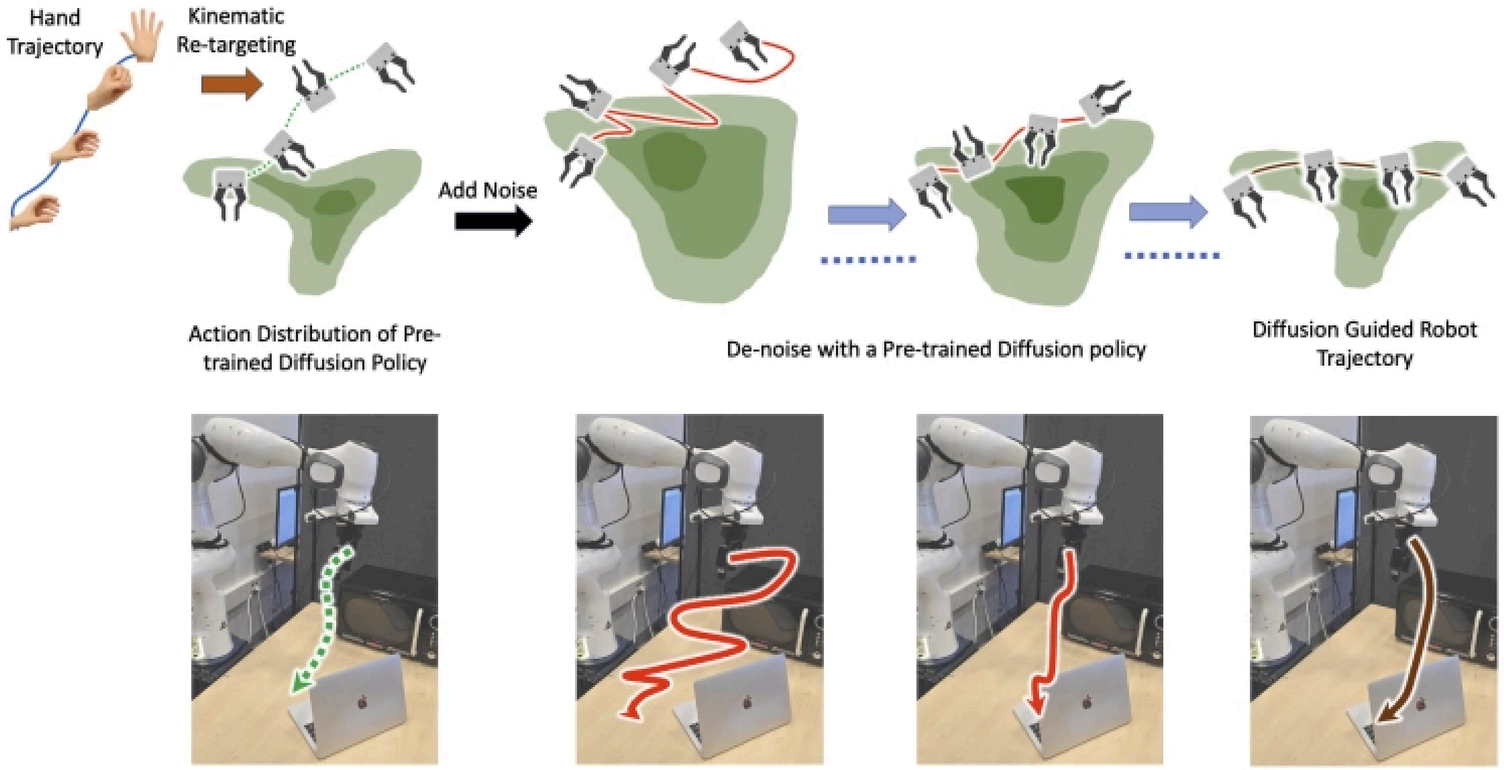
아이디어
generalist policy를 활용해서 single human demonstration을 모방해서 로봇이 새로운 task를 수행





핵심 아이디어

human trajectory의 high-level motion을 유지하면서
pretrained policy가 판단하는 plausible robot action distribution 안에 들어가는 robot action을 얻는 것



$$\tilde{\mathbf{a}}_t^{(s^*)} = \sqrt{\alpha_{s^*}}\hat{\mathbf{a}}_t + \sqrt{1-\alpha_{s^*}}\varepsilon_t, \quad \varepsilon_t \sim \mathbf{N}(0, I)$$

결과

Group	Robot Policy	Kinematic	*DemoDiffusion*
Small	25.4	2.6	**31.8**
Medium	26.6	0.7	**31.6**
Large	27.4	1.5	**29.6**
Average	26.5	1.6	**31.0**

Task	Pi-0	Kinematic	*DemoDiffusion*
Shut Down Laptop	20	10	**60**
Close Microwave	10	80	**90**
Drag Basket	10	**80**	**80**
Wipe Table	50	0	**100**
Iron Curtain	20	70	**90**
Pick up Bear	0	40	**60**
Pick Place Bowl	0	70	**100**
Pick Place Banana	0	70	**90**
Average	13.8	52.5	**83.8**

simulation과 real 모두에서 zero shot 성능이 매우 좋게 나오는 것을 확인할 수 있다.

04

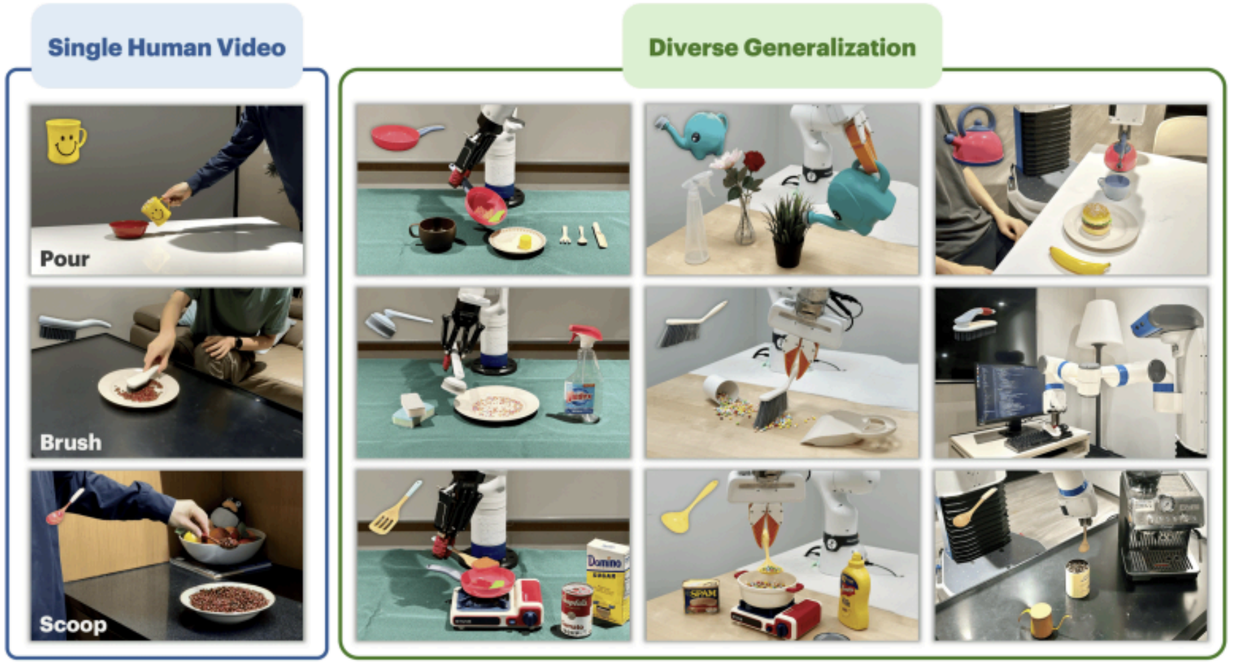
MimicFunc: Imitating Tool Manipulation from a Single Human Video via Functional Correspondence

Chao Tang, Anxing Xiao, Yuhong Deng, Tianrun Hu, Wenlong Dong, Hanbo Zhang, David Hsu, Hong Zhang

CoRL / 2025

문제 상황
기존 단일 인간 영상 기반 imitation은 형상·외형 유사성에 의존
모양이 크게 다른 새로운 도구로는 일반화가 어려움

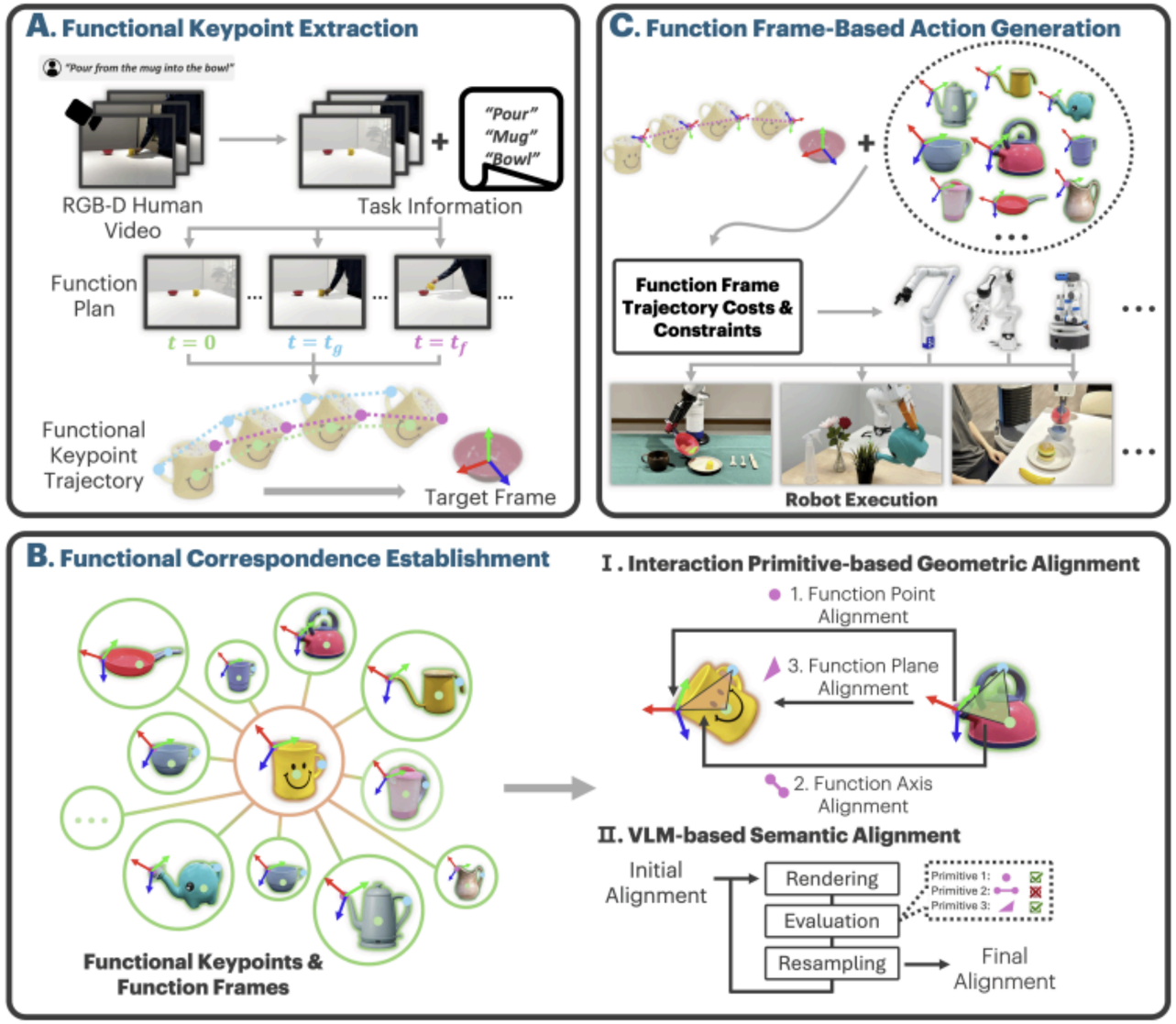
아이디어
도구의 생김새가 아니라 “어디를 잡고, 어디가 실제 기능을 수행하며, 어떤 방향으로 사용되는가”에 집중해서 action 생성



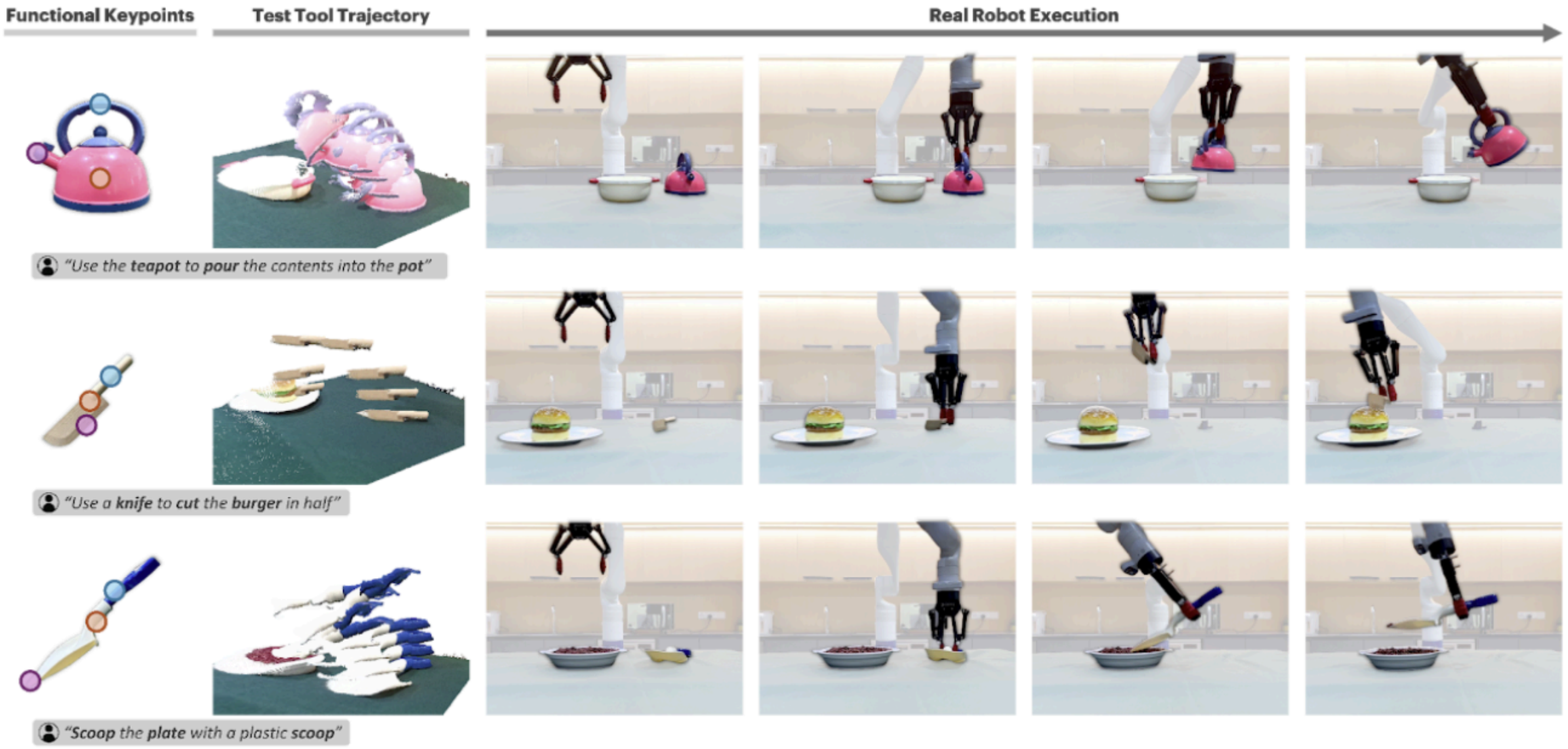
핵심 아이디어

1. Functional Keypoint 추출
2. 새로운 Tool과 Functional Correspondence 생성
   - VLM + dense correspondence로 새 도구의 대응 keypoint 탐색
   - 세 점으로 Function Frame을 만들고
     point / axis / plane 기준으로 human tool과 정렬.

→ 이를 통해 최적화된 robot action 생성



결과



functional keypoint와 새로운 tool에 최적화된 경로를 생성해서 one-shot 수행이 가능해짐

05

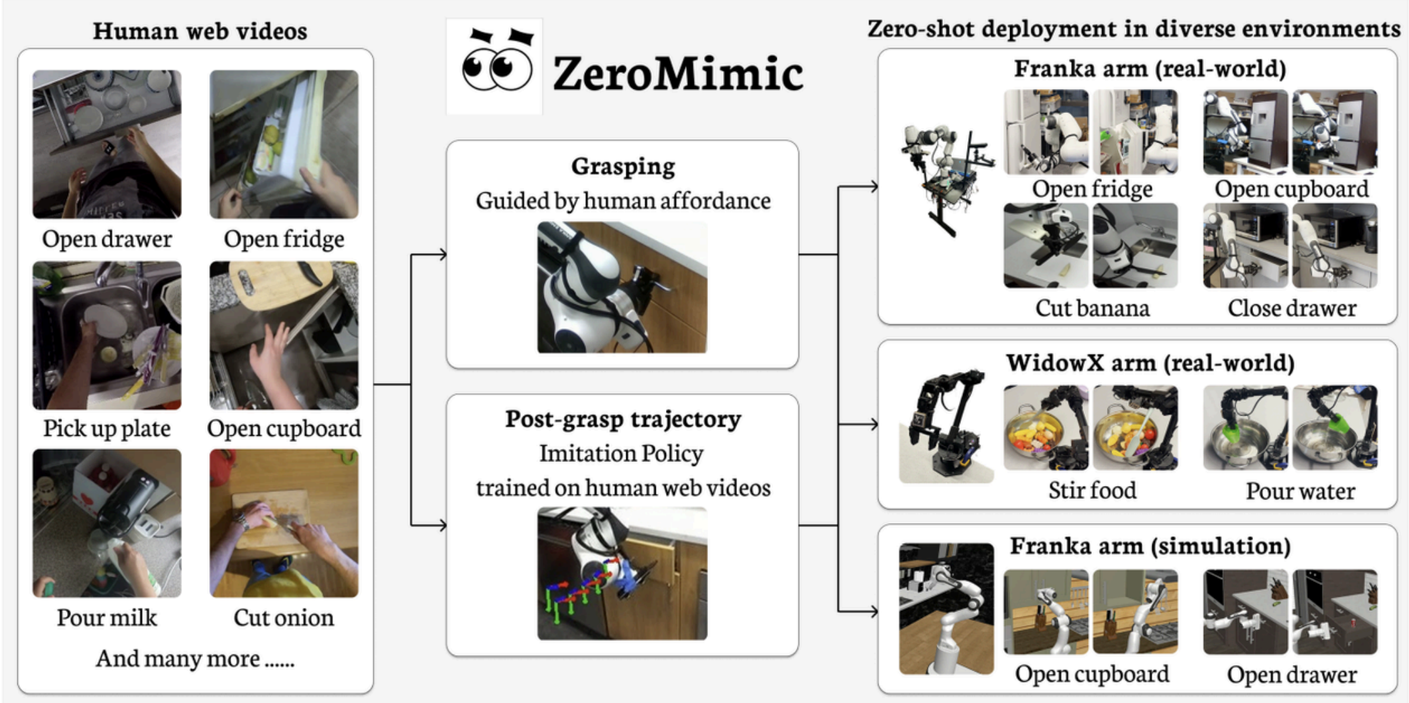
ZeroMimic: Distilling Robotic Manipulation Skills from Web Videos

Junyao Shi, Zhuolun Zhao, Tianyou Wang, Ian Pedroza, Amy Luo, Jie Wang, Jason Ma, Dinesh Jayaraman

ICRA / 2025

문제 상황
Web human video는 많지만 human-robot embodiment 차이와 흔들리는 카메라 때문에 그대로 쓰기 어려움

아이디어
그 task를 수행하기 위해서
“어디를 잡을지”와 “잡은 뒤 어떻게 움직일지”를 분리해서 학습해보자!

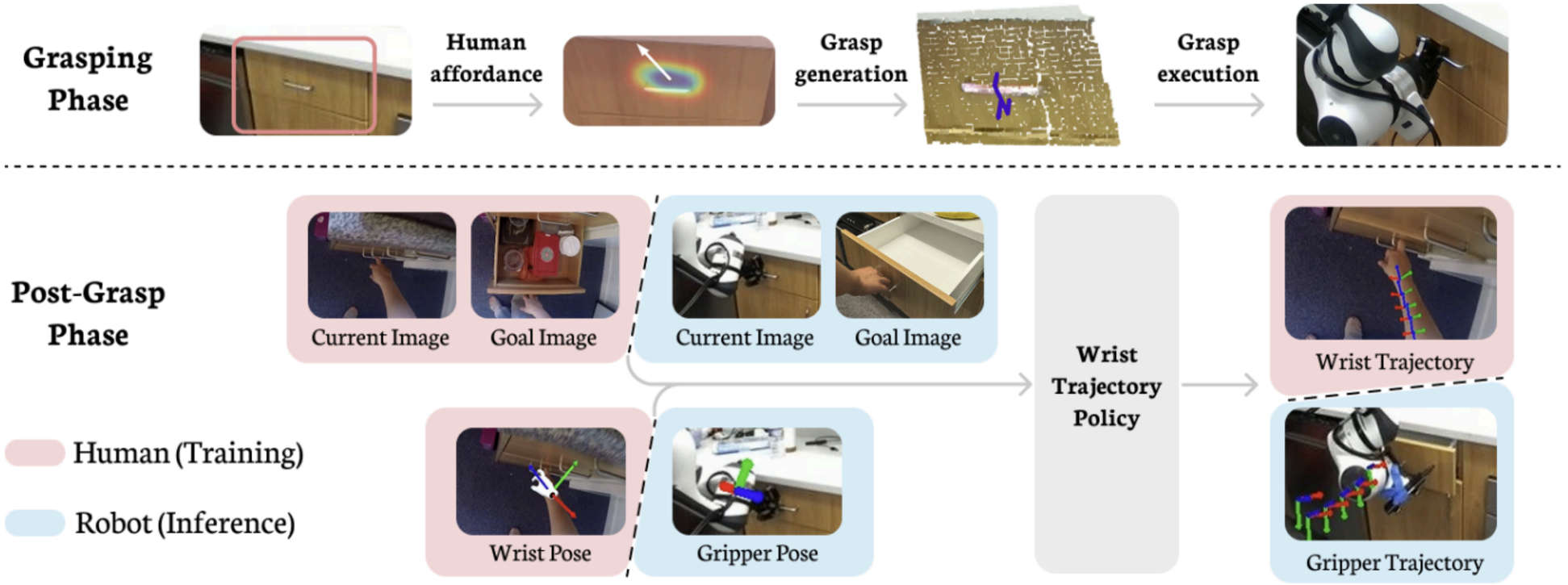


핵심 아이디어

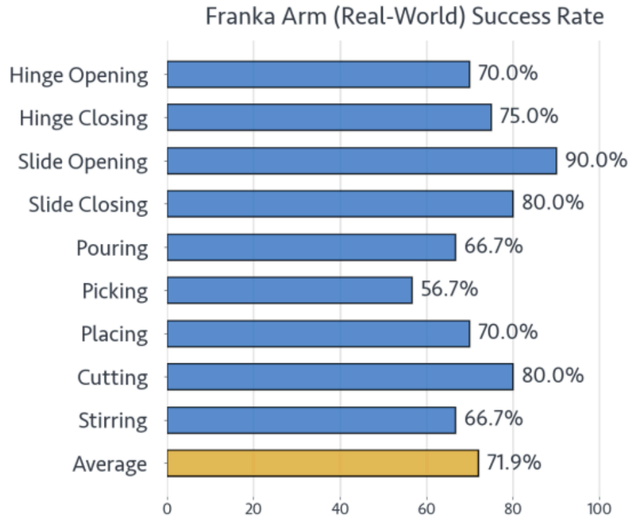
Web Human Video → Grasp Affordance + Post-Grasp Motion → Robot Execution

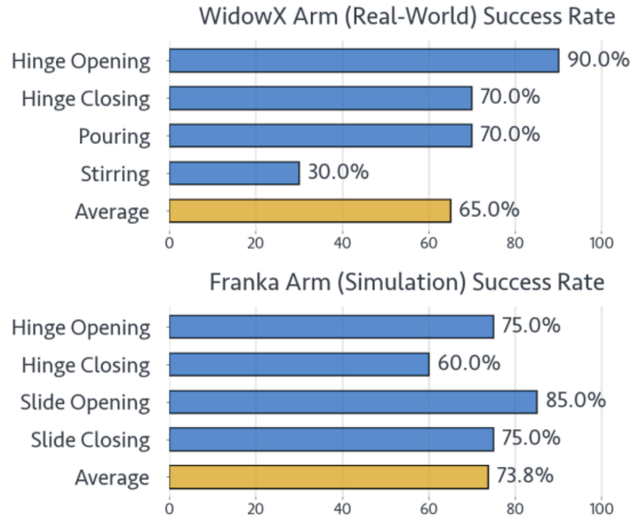
Grasping : VRB로 human video에서 task-relevant contact/affordance 위치 예측후 robot gripper에 맞는 실제 grasp pose 선택
Post-Grasp Motion : HaMeR + COLMAP으로 human wrist motion을 world-frame 6D trajectory로 복원

이러한 변환을 통해 추가 robot fine-tuning 없이 바로 실행



결과


Franka Arm (Real-World) Success Rate
Hinge Opening 70.0%
Hinge Closing 75.0%
Slide Opening 90.0%
Slide Closing 80.0%
Pouring 66.7%
Picking 56.7%
Placing 70.0%
Cutting 80.0%
Stirring 66.7%
Average 71.9%
0 20 40 60 80 100


WidowX Arm (Real-World) Success Rate
Hinge Opening 90.0%
Hinge Closing 70.0%
Pouring 70.0%
Stirring 30.0%
Average 65.0%
0 20 40 60 80 100
Franka Arm (Simulation) Success Rate
Hinge Opening 75.0%
Hinge Closing 60.0%
Slide Opening 85.0%
Slide Closing 75.0%
Average 73.8%
0 20 40 60 80 100


Autonomous
2x speed


Autonomous
5x speed

human hand pose의 표현 공간과
robot hand/gripper의 표현 공간을 정렬해서 변화하는 alignment를 사용해보자!

06

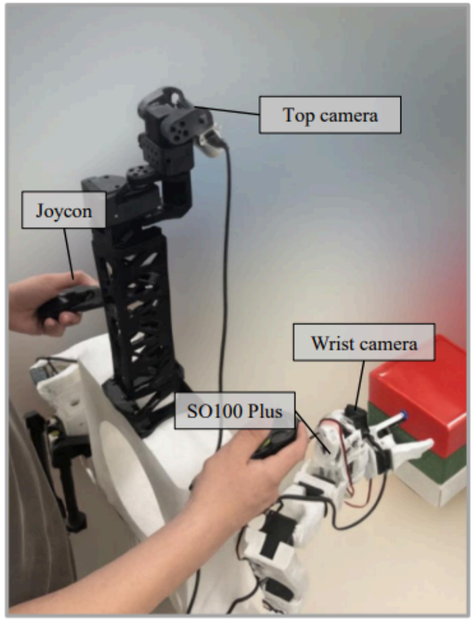
EasyMimic: A Low-Cost Framework for Robot Imitation Learning from Human Videos

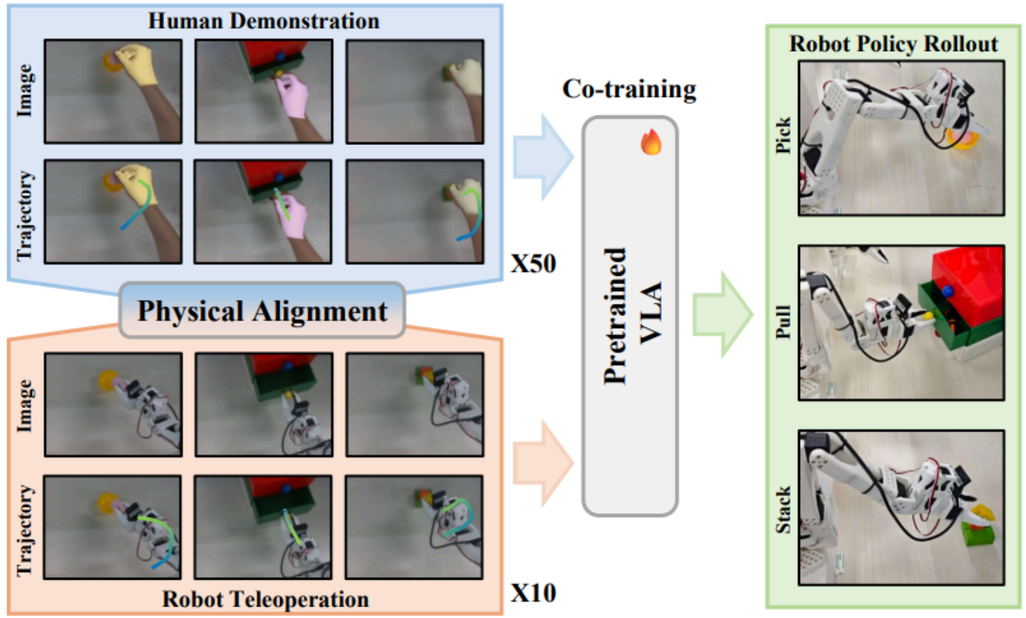
Chao Tang, Anxing Xiao, Yuhong Deng, Tianrun Hu, Wenlong Dong, Hanbo Zhang, David Hsu, Hong Zhang

ICRA / 2026

문제 상황
사람 손과 로봇 gripper의 외형 차이(Visual Gap)와 kinematics/action space 차이(Action Gap) 때문에 바로 학습에 사용하기 어려움

아이디어
Human video와 robot video의 Action-space + Visual-space를 맞춰주자!





핵심 아이디어

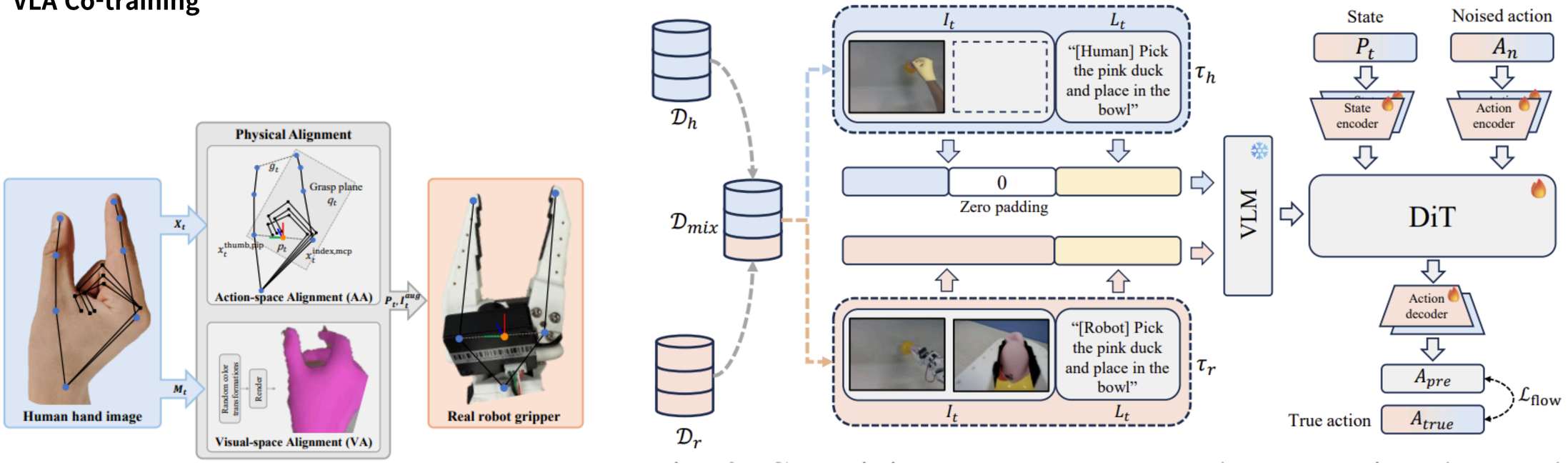
- **Action Space Alignment**

human hand의 3D keypoint/trajectory 추출.
손의 위치·방향·손가락 간 거리를 각각 robot의 EE position / orientation / gripper state로 변환.

- **Visual Space Alignment**

Human hand mesh를 추출하고 랜덤 색상으로 rendering/augmentation을 적용해 human–robot visual gap 감소

- **VLA Co-training**



결과

Strategy	Pick	Pull	Stack	Avg.
EasyMimic	**1.00**	**0.90**	**0.70**	**0.87**
EasyMimic-AA	0.50	0.80	0.50	0.60
EasyMimic-VA	0.40	0.40	0.40	0.40

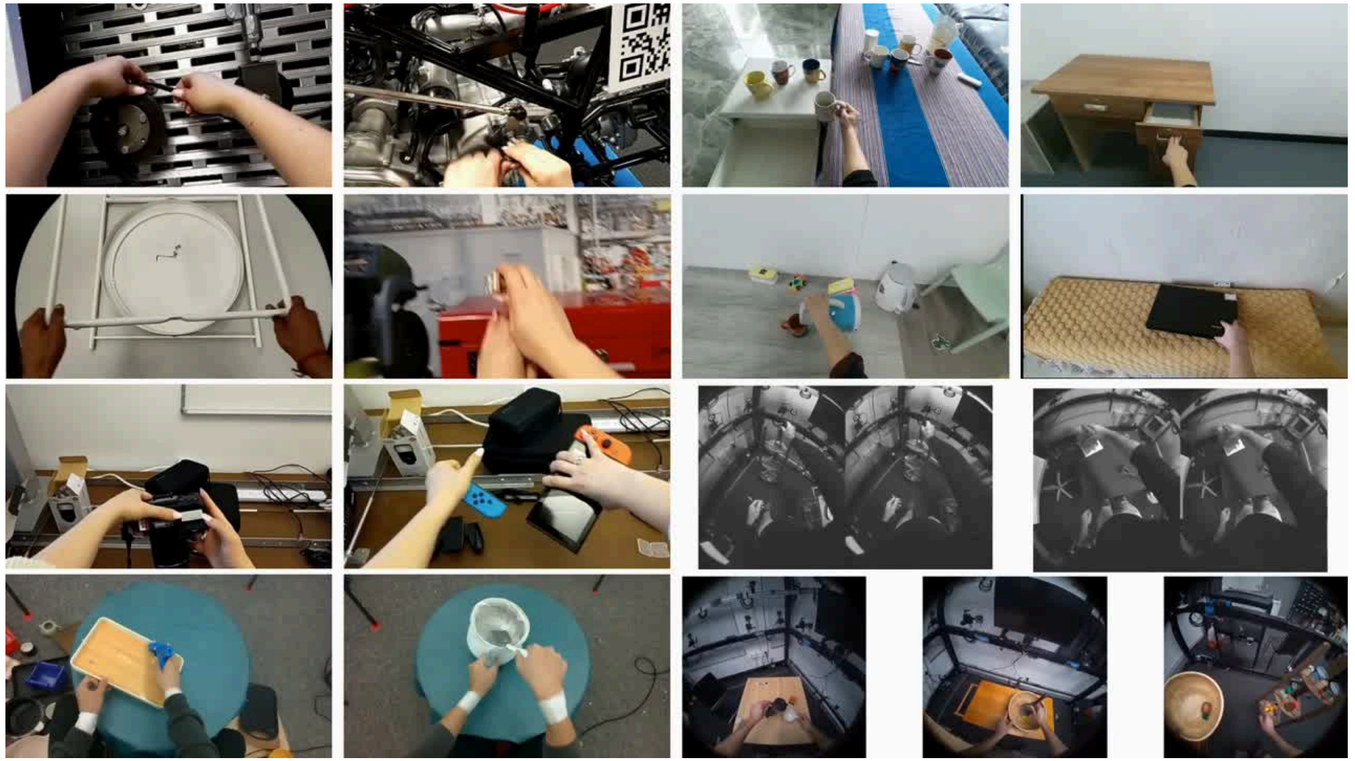
Strategy	Pick	Pull	Stack	LC	Avg.
Robot-Only (10 traj.)	0.30	0.30	0.15	0.30	0.26
Robot-Only (20 traj.)	0.60	0.70	0.35	0.40	0.51
Pretrain-Finetune	0.80	**0.90**	0.50	0.80	0.75
EasyMimic	**1.00**	**0.90**	**0.70**	**0.90**	**0.88**

07

EgoVLA: Learning Vision-Language-Action Models from Egocentric Human Videos

Ruihan Yang, Qinxi Yu, Yecheng Wu, Rui Yan, Borui Li

IROS / 2026



문제 상황
Human video는 많지만 human-robot action space와 외형 차이 때문에 바로 사용할 수 없다.

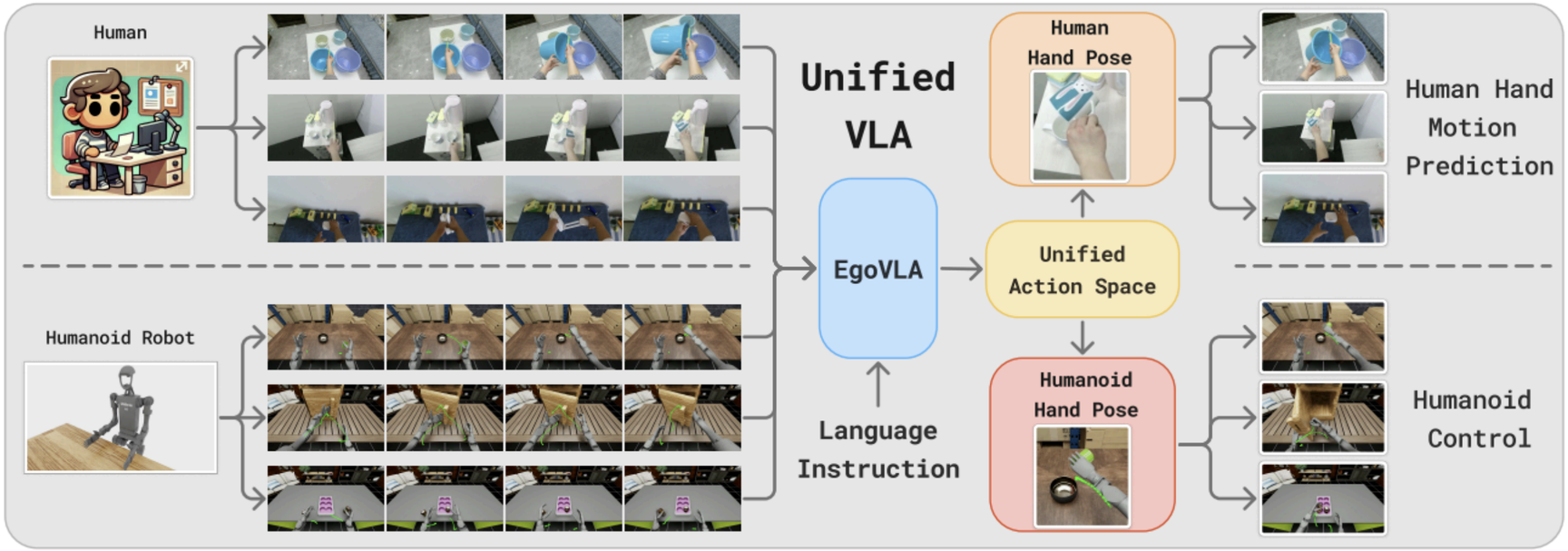
아이디어
human/robot의 action을 Unified Action Space로 정렬하고 소량의 robot data로 adaptation하자!

핵심 아이디어

- **Human VLA Pretraining**

- **Unified Action Space**
  Human과 robot 모두 wrist pose + MANO hand representation으로 표현
  Robot hand도 fingertip 위치를 기준으로 MANO representation에 맞춰 human/robot action space를 통일

- **Robot Action 변환 + Fine-tuning**
  이후 소량의 robot demonstrations로 EgoVLA를 fine-tuning하여 실제 robot policy로 adaptation



결과

Table 1: **Evaluation on Short-Horizon Tasks on Seen/Unseen Visual Configurations**

Method	Stack-Can		Push-Box		Open-Drawer		Close-Drawer		Flip-Mug		Pour-Balls		Open-Laptop		Mean	
	SR ↑	PSR ↑	SR ↑	PSR ↑	SR ↑	PSR ↑	SR ↑	PSR ↑	SR ↑	PSR ↑	SR ↑	PSR ↑	SR ↑	PSR ↑	SR ↑	PSR ↑
Seen Visual Configuration																
ACT [5]	22.22	22.22	11.11	85.19	18.52	66.67	48.15	96.30	7.40	7.40	3.70	77.78	62.96	62.96	24.87	59.79
EgoVLA-NoPretrain [38]	55.56	55.56	51.85	75.93	59.26	77.78	100.00	100.00	3.70	3.70	85.19	93.83	96.30	96.30	64.55	71.87
EgoVLA (50%)	44.44	44.44	33.33	66.67	22.22	61.73	100.00	100.00	22.22	22.22	77.78	92.59	37.04	44.44	48.15	61.73
EgoVLA	77.78	77.78	70.37	83.33	59.26	81.48	100.00	100.00	59.26	59.26	77.78	92.59	100.00	100.00	77.78	84.92
Unseen Visual Configuration																
ACT [5]	9.09	10.61	18.18	87.88	24.24	71.97	63.64	96.97	0.00	0.00	6.06	59.09	53.03	53.03	24.89	54.22
EgoVLA-NoPretrain [38]	57.58	57.58	69.70	82.58	46.15	71.28	86.36	89.90	4.69	4.69	46.03	79.37	48.48	53.03	51.28	62.63
EgoVLA	62.12	62.12	75.76	84.85	50.00	75.25	98.48	98.48	30.77	30.77	83.33	94.44	83.33	87.88	69.11	76.26

Table 2: **Evaluation on Long-Horizon Tasks on Seen/Unseen Visual Configurations.**

Method	Insert-And-Unload-Cans		Stack-Can-Into-Drawer		Sort-Cans		Unload-Cans		Insert-Cans		Mean	
	SR ↑	PSR ↑	SR ↑	PSR ↑	SR ↑	PSR ↑	SR ↑	PSR ↑	SR ↑	PSR ↑	SR ↑	PSR ↑
Seen Visual Configuration												
ACT [5]	0.00	11.11	11.11	37.04	0.00	28.63	0.00	35.80	0.00	19.75	2.22	26.47
EgoVLA-NoPretrain [38]	7.41	45.93	0.00	18.52	51.85	79.63	62.96	75.00	11.11	55.56	26.67	54.93
EgoVLA (50%)	0.00	28.15	29.63	57.41	0.00	33.33	0.00	30.56	7.41	49.07	7.41	39.70
EgoVLA	44.44	77.04	40.74	75.93	55.56	88.89	66.67	83.33	22.22	78.70	45.93	80.78
Unseen Visual Configuration												
ACT [5]	0.00	7.95	1.52	33.84	0.00	20.71	1.52	33.83	0.00	21.21	0.61	23.51
EgoVLA-NoPretrain [38]	0.00	29.09	0.00	20.08	15.15	45.83	34.85	55.68	6.06	30.30	11.21	36.20
EgoVLA	31.82	76.97	28.79	60.98	18.18	68.94	50.00	75.76	15.15	62.88	28.79	69.11

Arm 뿐만 아니라, 이동형 장치에도 사용할 수 있게 해보자

VLM을 이용해서 plan을 세워서 action을 예측하는 방식을 만들어보자

robot으로 수집한 데이터를 증강해서 사용하자

08

HoMMI: Learning Whole-Body Mobile Manipulation from Human Demonstrations

Xiaomeng Xu, Jisang Park, Han Zhang, Eric Cousineau, Aditya Bhat, Jose Barreiros, Dian Wang, Jeannette Bohg, Shuran Song

RSS / 2026

아이디어
로봇을 사용하지 않고 human demonstration만으로 whole-body mobile manipulation을 학습할 수 있는 데이터 수집 및 정책 학습 프레임워크를 제안

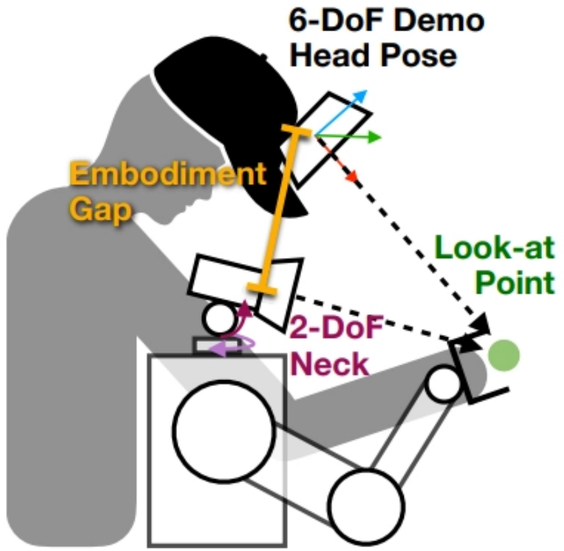
핵심 아이디어

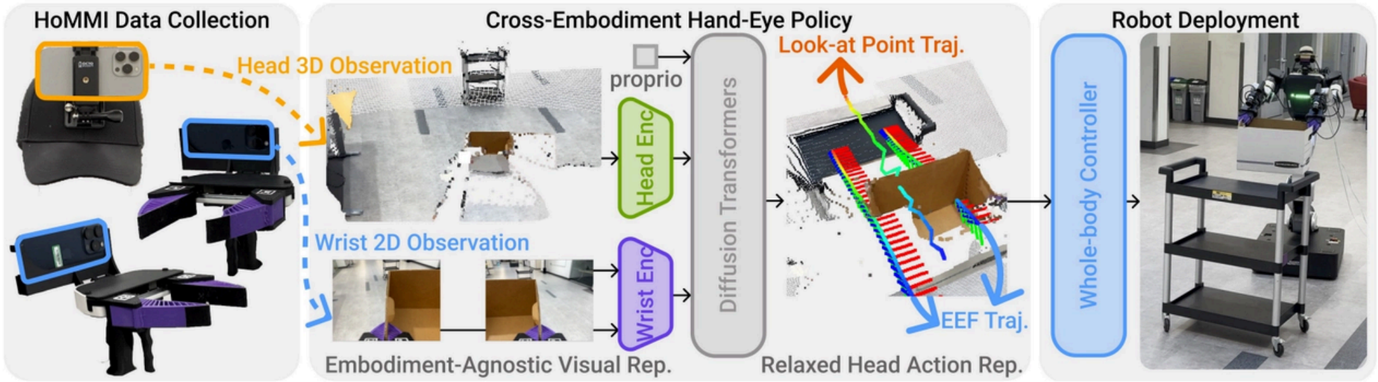
1. **UMI를 추가해서 데이터 수집**
   iphone 3대 + UMI gripper 2개를 이용

2. **visual gap을 줄이기 위한 visual representation 사용**
   RGB → 3D visual representation으로 변환
   human hand mask

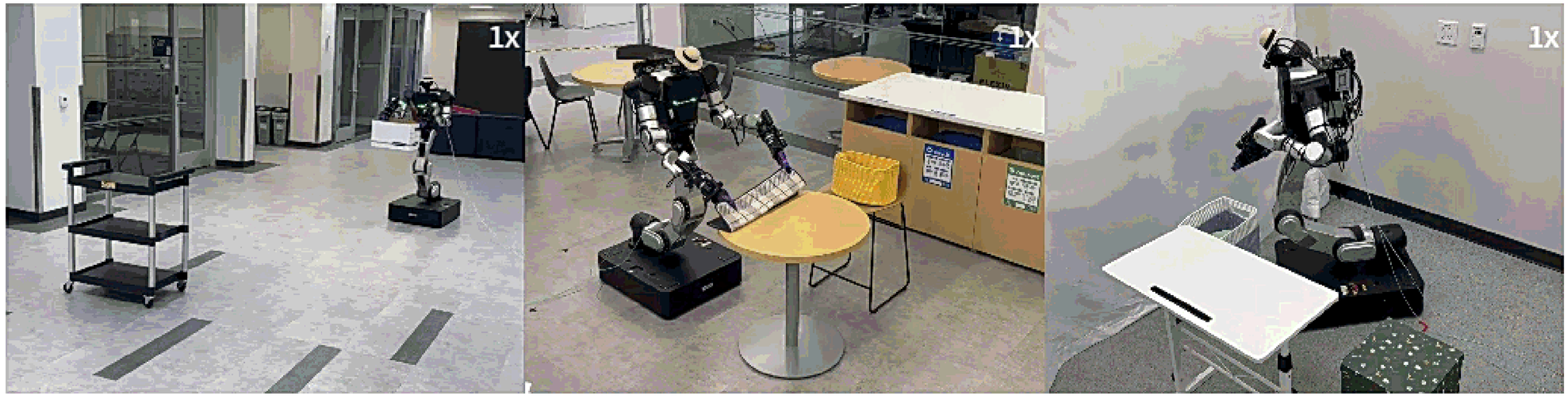
3. **kinematic gap을 줄이기 위해 Head Action Representation 사용**
   camera ray 와 pointmap을 통해 look-at-point 지정
   머리를 정확히 어떤 pose로 지정하는게 아니라 이 3D 위치를 보라고 명령

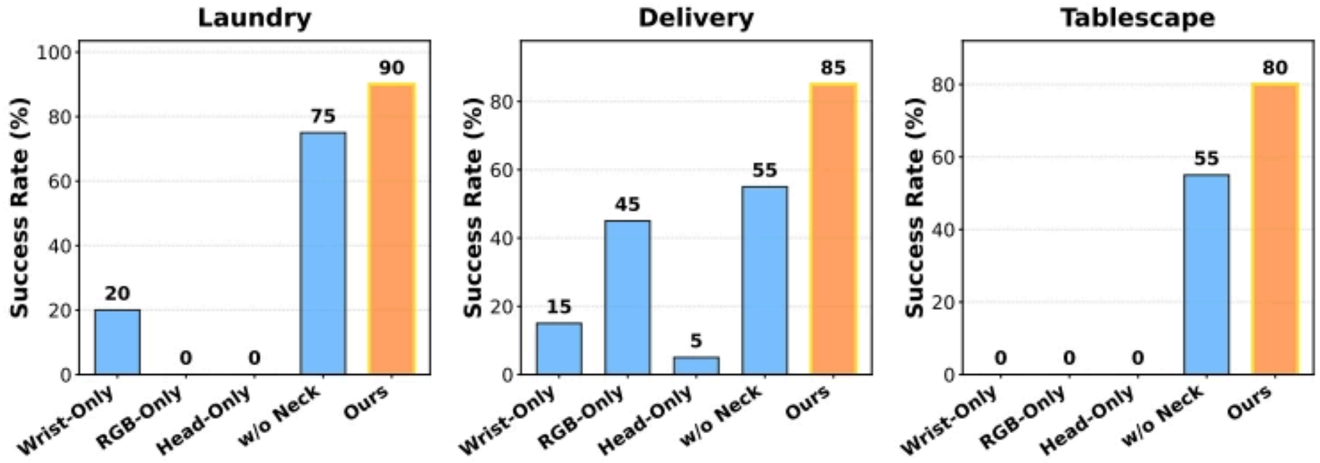
4. **whole body 제어를 위한 controller 사용**





결과





이동형 robot 장치에서도
human video를 통해서도
학습하고 실행할 수 있다는 것을 보여줌.

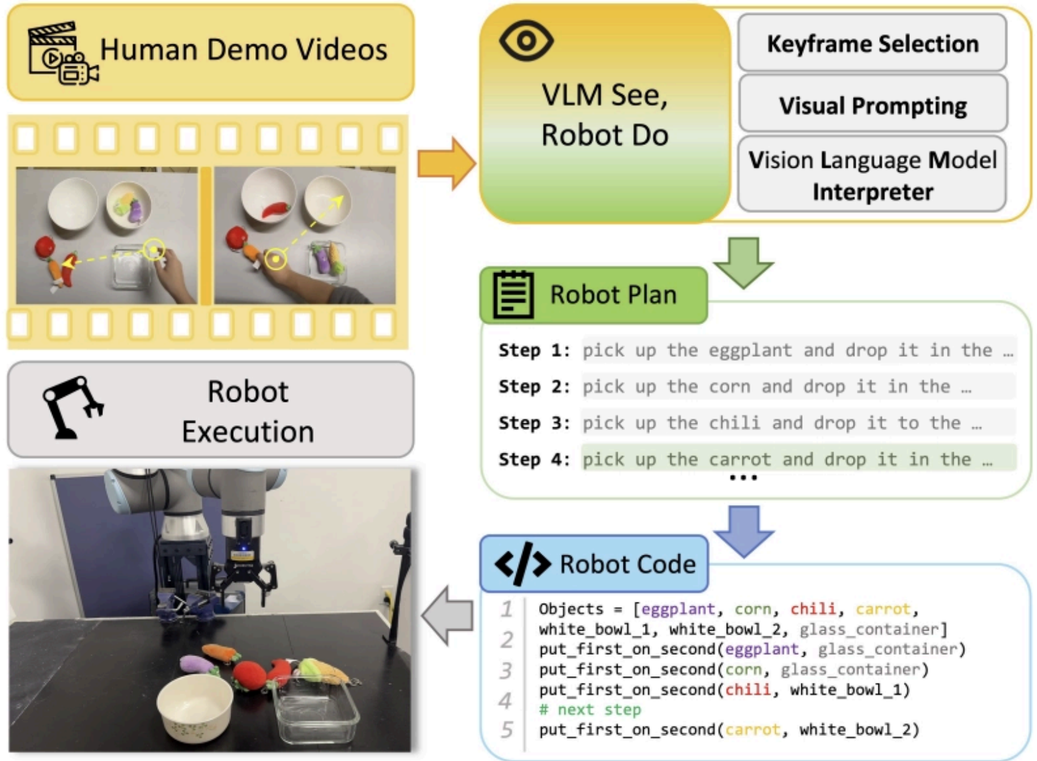
09

VLM See, Robot Do: Human Demo Video to Robot Action Plan via Vision Language Model

Beichen Wang, Juexiao Zhan, Shuwen Dong Irving Fang, Chen Feng

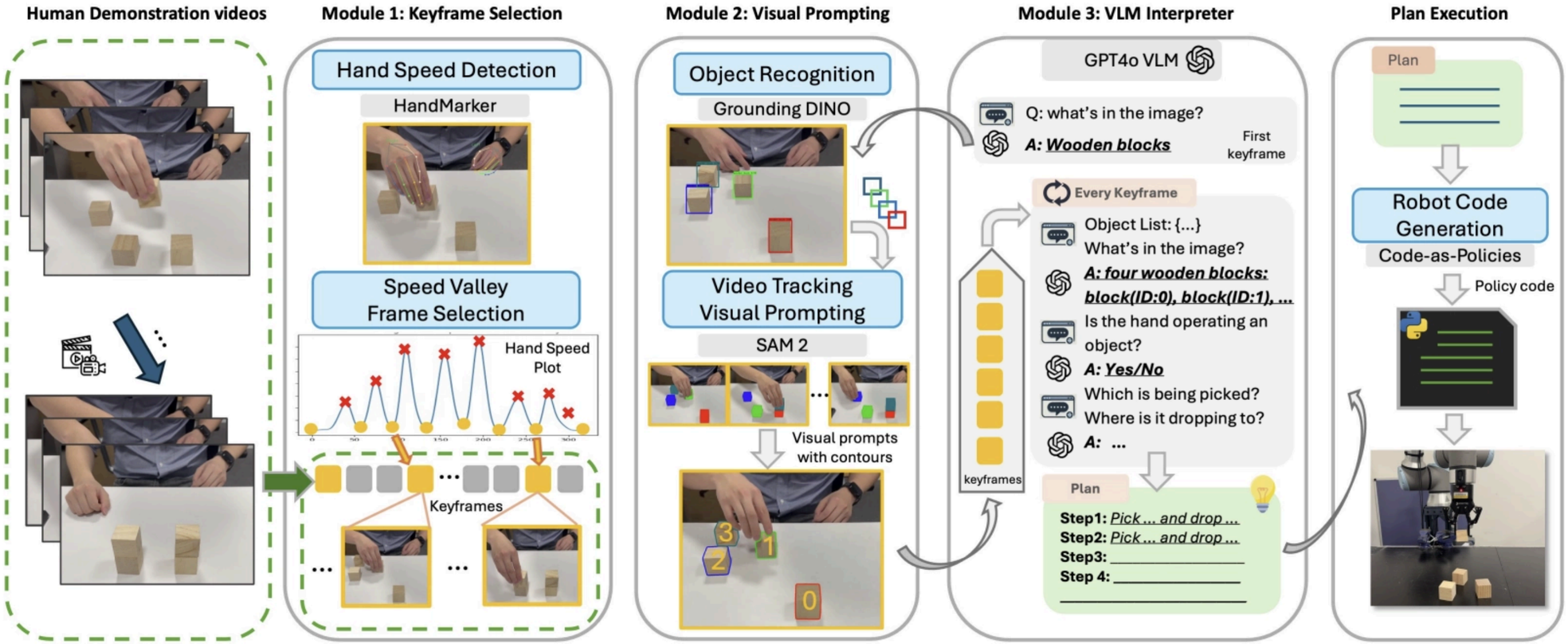
IROS / 2025

아이디어
VLM을 이용하여 인간의 시연 영상을 해석하고,
이를 바탕으로 로봇의 작업 계획을 생성한다면
더 높은 성공률의 액션을 만들어낼 수 있지 않을까??



핵심 아이디어

SEEDO agent 제안 (+ 3가지 module)
3가지 모듈(Keyframe Selection, Visual Prompting, VLM Interpreter)을 결합
human demonstration video를 해석하고 robot task plan을 생성하는 VLM 기반 agent



결과

VLM + 논문의 3가지 아이디어를 통해
좋은 성능을 보이는 것을 확인할 수 있음

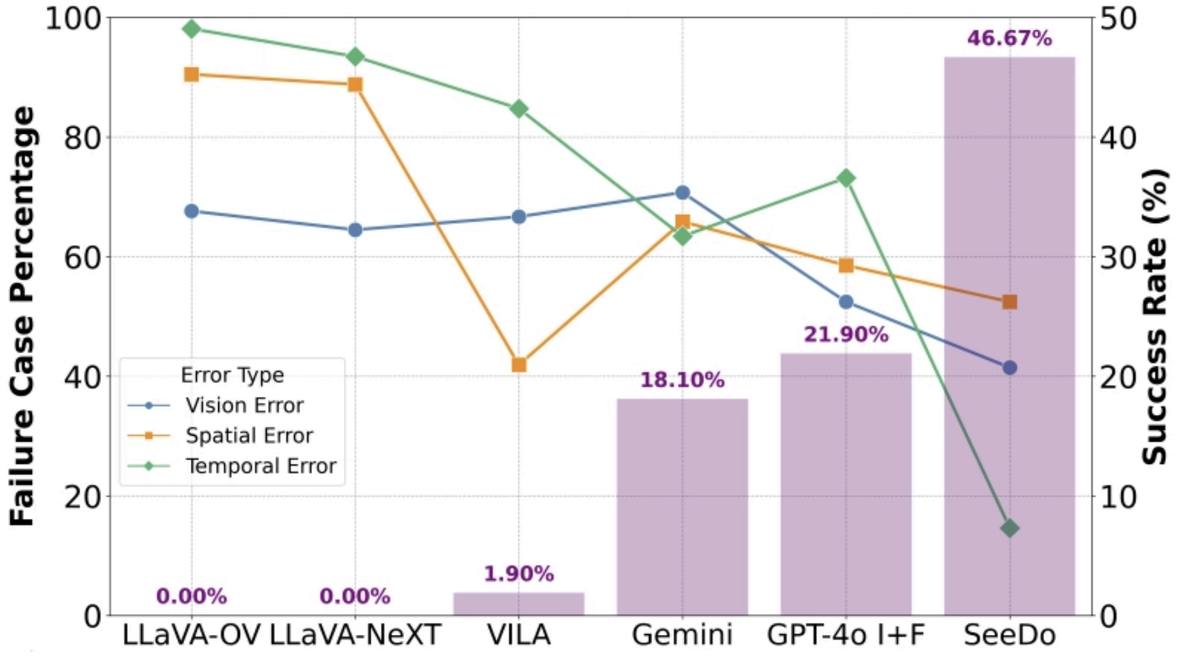


TABLE I: Model Success Rate Across Different Tasks

Model	Vegetable Organization			Garment Organization			Wooden Block Stacking		
	TSR	FSR	SSR	TSR	FSR	SSR	TSR	FSR	SSR
LLaVA-OneVision [10]	0.00	0.00	2.75	0.00	0.00	31.61	0.00	0.00	2.56
LLaVA-NeXT-Video-7B [47]	0.00	0.00	9.03	0.00	0.00	4.00	0.00	0.00	3.85
VILA1.5-8B [46]	5.26	5.26	14.38	0.00	0.00	0.90	0.00	0.00	3.41
Gemini 1.5 Pro [11]	39.47	39.47	70.00	16.67	16.67	57.22	0.00	0.00	13.80
GPT-4o [18] Init + Final	39.47	39.47	54.43	13.33	**30.00**	31.61	10.26	15.38	32.69
SeeDo	**60.53**	**60.53**	**80.40**	**26.67**	26.67	**66.50**	**21.62**	**21.62**	**52.48**

10

CRAFT: Video Diffusion for Bimanual Robot Data Generation

Jason Chen, I-Chun Arthur Liu, Gaurav Sukhatme, Daniel Seita

IROS / 2026

문제 상황
로봇으로 직접 수집된 데이터도
object pose·camera view·background·lighting·robot embodiment 다양성이 부족

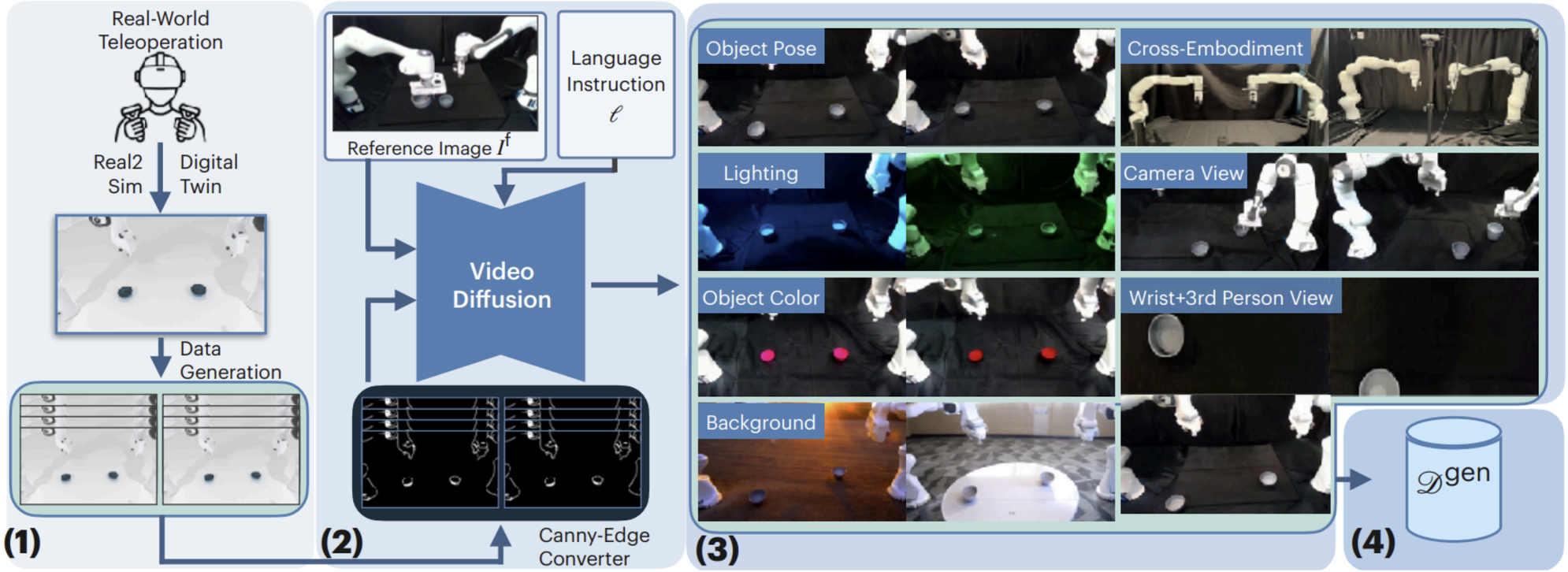
아이디어
Video Diffusion을 이용해서 다양한 환경의 video를 생성해보자!

핵심 아이디어

Trajectory Expansion
simulator의 digital twin으로 옮겨서
Object pose 등을 바꾸면서
trajectory를 재생하고 성공한 rollout만 유지

Video Generation
Simulation trajectory에서
Canny-edge video 추출
Video Diffusion에 넣어 photorealistic 영상 생성

다양한 Data Augmentation
Object pose/Object color
Lighting/Background
Camera viewpoint
Wrist + 3rd-person multi-view
Cross-embodiment



결과

Method	Lighting			Background			Camera View			Object Color			Wrist + 3rd Person		
	LR	PC	SB	LR	PC	SB	LR	PC	SB	LR	PC	SB	LR	PC	SB
ACT w/o Aug	3 / 20	1 / 20	0 / 20	4 / 20	0 / 20	0 / 20	6 / 20	3 / 20	2 / 20	2 / 20	0 / 20	1 / 20	15 / 20	11 / 20	13 / 20
CRAFT Pose-Only	5 / 20	3 / 20	2 / 20	7 / 20	2 / 20	3 / 20	13 / 20	5 / 20	7 / 20	5 / 20	2 / 20	3 / 20	13 / 20	8 / 20	10 / 20
ACT w/ Baseline Aug	13 / 20	9 / 20	8 / 20	4 / 20	5 / 20	6 / 20	14 / 20	8 / 20	6 / 20	15 / 20	9 / 20	11 / 20	N/A†	N/A†	N/A†
CRAFT (Ours)	**17 / 20**	**14 / 20**	**12 / 20**	**18 / 20**	**15 / 20**	**10 / 20**	**19 / 20**	**18 / 20**	**18 / 20**	**18 / 20**	**18 / 20**	**17 / 20**	**20 / 20**	**19 / 20**	**20 / 20**

Thank you

Question & Disscusion

RILS LAB