CoLR 참관기

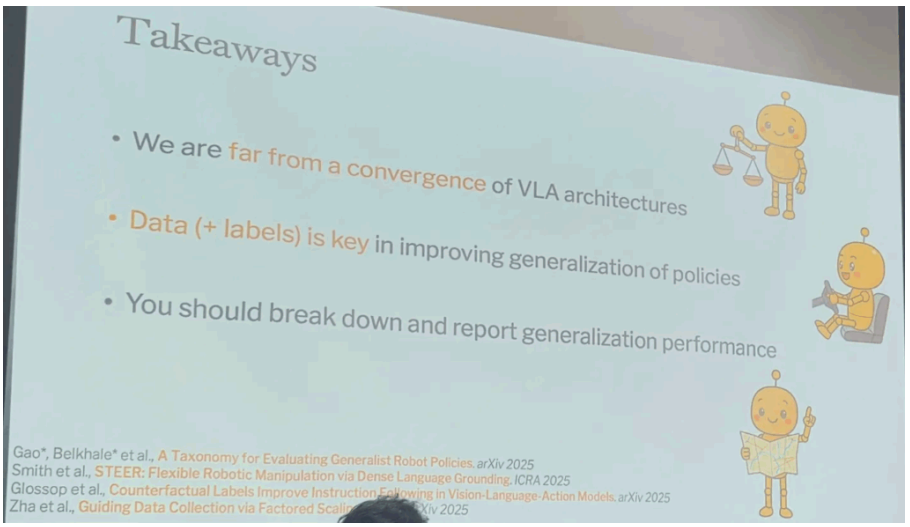
2025/10/13 석사과정 이재찬

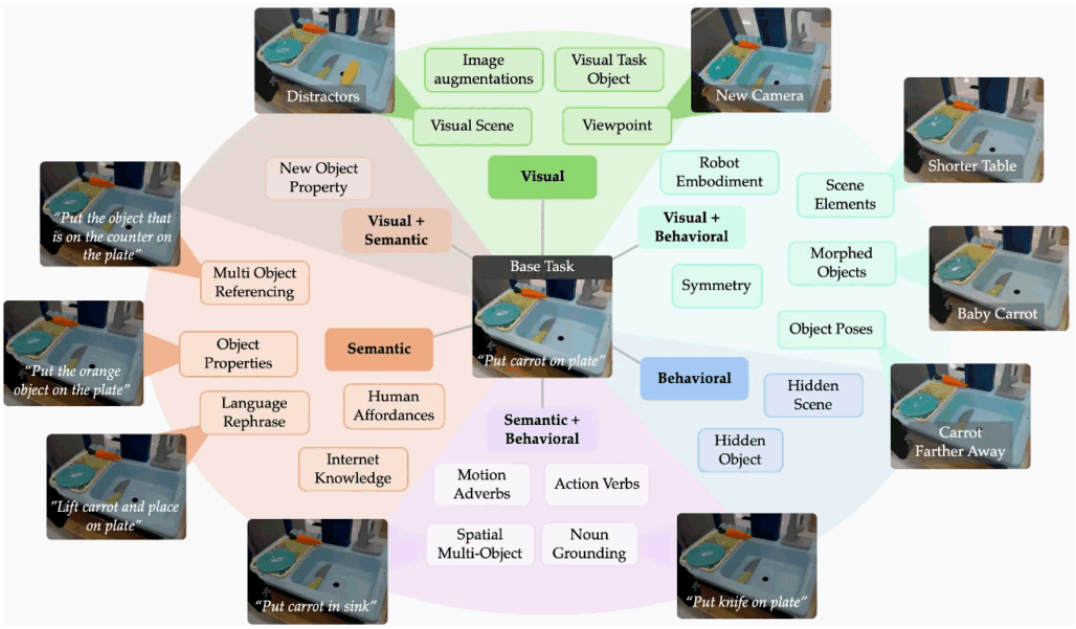
안녕하세요. CoRL 2025 참관기를 작성해보고자 합니다. 학회는 9/27(토)~9/30(화)까지 4일간 코엑스에서 진행되었습니다. 국제 탑티어 학회는 처음 참관하기도 하고 외국인 연구자들이 정말 바글바글해서 첫날에는 설레면서 긴장이 많이 되었던 것 같습니다. 정보량이 너무 많기도 하고 제가 이해하지 못한 컨셉도 너무 많지만 그럼에도 제가 듣고 경험한 것들을 토대로 학회 디테일들을 조금씩 정리해보도록 하겠습니다.

Workshop에 관하여

Workshop이 가장 먼저 첫날에 이루어졌는데 확실히 본 학회 행사를 위한 마중물 느낌이 강했습니다. 정해진 주제들로 각 세션마다 연사님들의 발표가 이어졌고, 해당 발표 내에는 다양한 workshop 논문들과 oral, spotlight paper들도 소개되곤 했습니다. 정말 여러가지 주제들이 시간대별로 촘촘히 나뉘어져 있었는데 다 들을 순 없으니 선택과 집중을 해야되는 게 좀 아쉬웠습니다. 고심 끝에 집중해서 들은 workshop 하나 중 기억에 남는 내용들을 적어보자면, 크게 4가지 입니다.

1. policy learning -> evaluation -> failure detection -> adapt to env -> deployment 라는 policy의 deployment 관점의 큰 생애주기
2. VLA의 test-time 시의 uncertainty기반 failure detection 이나 self-certainty 기반 deploy로 성능 향상 내용
3. generalization을 위해선 VLA 모델 아키텍쳐 개선보다 data-driven 방식인 data scaling도 중요한데, 이미 많은 정보들 중에서 data curating도 중요하다.
4. 위 모든 것들에도 불구하고 아직 VLA 아키텍쳐는 어느 하나가 정답이다라는 모델구조가 없어서, 학습 끝난 policy들 간의 eval하는 방법이나 benchmark도 효율적이고 더 세분화되고 신뢰할 만한 방식이 필요하다.





특히 3번 관해서는 ★-Gen(arXiv 2025) 라는 좀 세밀한 벤치마크에 대한 언급이 있었어서 기억에 남습니다. 원래는 uncertainty 기반 failure detection 이라는 키워드가 있길래 끌려서 들은 워크샵이었지만 data-driven 방식으로 너무 급격하게 VLA 연구가 성장하느라 간과하고 있던, 학습을 위한 데이터 자체의 문제(data curating), 평가를 위한 기준 자체의 문제(benchmark, eval) 관점에서도 문제를 생각해볼 수 있는 좋은 시간이었습니다.

이외에도 세미나방을 나와 복도로 나오면 여러 포스터들이 붙어있었고 본 학회행사보다 규모가 작기에 포스터 저자에게 질문할 기회가 정말 많았지만 이 학회 자체와 코엑스라는 공간 자체에도 적응하느라 어색해서인지 이 날은 포스터 질문은 거의 영어 물꼬트기 정도로만 찔끔찔끔했던 것 같습니다. 예를 들면, can you explain? 같이 포스터 보고도 이해 잘 못했을 때 하는 국룰 멘트 등을 하고는, 저자의 설명을 겉핥기로만 듣고 깊게는 질의응답 못하는 그런 식이었습니다. 나름대로는 본 학회를 위한 준비운동을 해보려했던 것 같습니다.

본 학회에 앞서 관심있던 나의 연구 관심사

제 관심사는 일단 가장 크게는 'Long-horizon manipulation task(장기 작업, 혹은 복잡하고 어려운 작업)를 어떻게 로봇 스스로 잘하게 만들지'와 관련된 것이었습니다. 구체적으로는 그 동안 서칭해오고 고민하던 연구방향은 VLM(MLLM)의 reasoning 능력을 활용해서 high-level planning을 먼저 수행하고, 그렇게 각각의 sub-plan들에 대해 low-level policy(IK기반 motion planner, IL기반 action policy)를 어떻게 잘 결합할 것인가에 대한 큰 흐름이었습니다. 현재는 VLM 기반 sub-task decomposition 여기에만 집중하고 있긴 합니다만, 여기저기 기웃기웃하고 보고듣다보니 든 떠오르는 생각으로는 다음의 것들도 관심을 갖고 있었습니다.

1. 환경 내 물체들이 articulated, deformable object들이라면 조작 태스크가 더 어려워질텐데, 물체의 6d pose 혹은 물체 affordance 영역 내의 6d pose 같이 spatial한 환경정보 이해나, 그 외의 물리적 정보, 속성 정보 등 이런 다양한 요소들을 어떻게 잘 캐치해서 잘 잡을 수 있게 VLM의 reasoning에 녹여낼 수 있을지
2. 양팔 로봇이라면 어떻게 효율적으로 작업 분배를 진행할지(오른팔, 왼팔이 동기적으로/비동기적으로?)
3. 작업 중 failure detection은 어떻게 할거고, 그 실패탐지한 걸 토대로 어떻게 회복 동작을 계획해서 동작하게 할건지

그래서 대체로 저는 VLM planning, reasoning 관련된 색다른 방식이 없을 지 이번 학회에서 많이 기대하고 갔었는데요.

본 학회에서는 **VLA, Data Scaling, HW, Humanoids**에 굉장히 집중하고 있었다.

아쉽게도 생각보다 색다른 방법들은 많지 않았고 다 제가 평소 고민해오던 방식 그대로 활용하는 것들이 대부분이었습니다. 기본적으로 LLM/VLM 한테 planning 혹은 reasoning 맡기는 방식을 쓴다면 해당 LLM/VLM 모듈에서의 failure한 경우는 크게 고려하지 않고, 잘하겠지~라는 전제를 깔고 들어가는 것 같습니다. (어느 저자는 poster에서 제가 GPT 아직 Visual Grounding 못한다! 고 하니까, 절 굉장히 귀엽게 보며 눈 찡긋하고 말 빨라지면서 야야; 뭔소리냐 GPTs can do Visual Grounding haha~ 이러더라구요. 뭐 제가 영어 표현이 미숙해서 visual grounding 못하는 경우도 있다! 에 대한 의도가 잘 전달이 안된 것 같긴한데, 아무튼 이들은 결국은 자기들이 정한 태스크와 문제정의 안에서 로봇의 액션이 성공하냐 못하냐에 가장 큰 관심을 두고 있습니다.)그래서인지 LLM/VLM Planning에 관한 내용이 그렇게 contribution이 많지 않아 poster 시간 때 질의해본 몇 논문들만 세미나 시간에만 살짝 풀어볼까 하구요,, 맞게 정리한 지는 모르겠지만 이번 CoRL 학회에서 제가 느낀 핵심 키워드들은 정리해보자면 다음과 같습니다.

CoRL 2025 ➔ Main Keywords

- Robot Data Scaling
  - How to teleoperate efficiently
    - Various HW to teleoperate
    - Using Human Video
  - Sim2Real transfer
- Towards Humanoids
  - Dexterous Hands
  - Bi-manual manipulation
  - Locomotion
  - Tactile / Force sensor
- VLA Architecture
  - Latent Space
    - + World Model
    - + Video Generative Model
    - + Reasoning
- Long-horizon Tasks / Dexterous Tasks

역시나 VLA를 학습하기 위한 데이터 취득, VLA 학습 방법론, HW에 관한 관련된 내용이 가장 많았고, 이번 CoRL 학회는 특히 Humanoids 학회랑 겹쳤다보니 Humanoids의 내용도 정말 많았습니다. 오죽하면 마지막에 Best Paper Awards 도 toward humanoids 스러운 논문들이 상을 받았습니다.

Award Finalists

Paper ID	Title	
1	Learning a Unified Policy for Position and Force Control in Legged Loco-Manipulation	best paper
2	LocoFormer: Generalist Locomotion via Long-context Adaptation	
3	Visual Imitation Enables Contextual Humanoid Control	best student paper
4	Fabrica: Dual-Arm Assembly of General Multi-Part Objects via Integrated Planning and Learning	best paper
5	DexUMI: Using Human Hand as the Universal Manipulation Interface for Dexterous Manipulation	
6	The Sound of Simulation: Learning Multimodal Sim-to-Real Robot Policies with Generative Audio	
7	Pi 0.5: a Vision-Language-Action Model with Open-World Generalization	
8	Steering Your Diffusion Policy with Latent Space Reinforcement Learning	

VLA와 Data Scaling 관련해서는 영규형이 정말 많이 정리를 하고 있었던 것 같아서 저는 이 중에 저의 연구에 직접적으로 관련있는 것은 아니었지만 많은 인사이트를 얻었던, 저한테는 인상깊었던 Tactile 센싱과 Humanoids 관련 얘기를 아래부터 해볼까 합니다.

Tactile 센서와 관련한 기존의 생각

저도 사실 막연하게는 앞으로의 로봇 연구에서 tactile 데이터라는 것이 정밀하고 세밀한 조작을 위해서는 무조건 필요하다고 생각이 들었습니다. 실제로 논문 이리저리 뒤적뒤적 하던 와중에, tactile 이나 sounds 등의 다른 modality 데이터를 같이 학습한 VLA 와 관련된 fei-fei li 교수님의 논문을 봤어서, 다른 센서

modality 와 관련된 연구는 money power, man power만 있다면 데이터 구축과 학습이 또 근 몇 년 안의 시간문제겠구나 싶었습니다. 저희 연구실 같은 경우는 tactile을 당장은 다룰 일이 없어보이기도 하고, 다루게 된다면 dexterous hands 쪽으로 넘어가게 돼서 manipulation 전반이 아니라 dex hands HW쪽에도 굉장히 많은 집중이 필요해지게 되고, 해당 분야 followup을 하기까지 또 한번 성장할 시간이 필요하기에,,, 막연하기만 했습니다.

근데 이번 CoRL에서 얼리키노트부터 humanoids에 대한 패널톡 얘기를 들어보니, 앞으로 tactile 센싱과 force 센싱에 대한 연구가 굉장히 박차가 가해질 것 같다는 생각이 들었습니다. 해당 연구가 타 방법론들과 자연스레 융합될 정도로 성숙된 건 아닌듯해 아직은 모르겠지만 활용해먹을 수 있을 정도로 유연해진다면 활용하는 것이 맞는 방향이지 않을까 싶습니다. 이 생각을 들게 한 얼리키노트와 패널톡 얘기를 해보겠습니다.

Early Career Keynotes(Tactile)

먼저 simulation based tactile sensing과 perception 관련된 발표였습니다. 특이하게 무슨 연유때문인지 기억은 안 나는데 온라인으로 발표를 진행해주셨습니다.

일단 발표의 요약부터 하면 '시뮬레이션 기반 촉각 데이터 생성' 파이프라인과 그걸 활용해서 'tactile sensor-invariant learning' 으로 transfer까지 하는 모델 아키텍쳐를 소개해주셨습니다. 영규형 연구가

많이 생각이 났는데요. 즉 sim 기반의 tactile 이미지데이터 생성 자동화 파이프라인, 그걸 가지고 데이터셋구축, 또 그걸 가지고 모델 학습까지,, tactile 데이터를 가지고 거의 딥러닝 생애주기를 쏵 훑는 본인의 연구를 소개해주셨습니다. 결론적으로 다양한 로봇 촉각 센서 sensor-agnostic하게 전이 가능성을 확장하는 촉각 인지 연구까지 진행했다고 결론이 났는데, 발표 듣기 전엔 early career keynotes에서 early career가 뭔 뜻인지 모르다가, 발표 다 듣고 나선 떠오르는 매우 핫한 신진 연구자를 일컫는 말이구나 싶었습니다.

발표는 핵심 question부터 던지면서 시작되었는데, “로봇이 real world의 촉각을 어떻게 이해할 수 있을까?” 에 초점을 두셨습니다. 이를 위해선 ‘감각의 표현’, 즉 데이터의 표현을 잘 설계해야 된다고 합니다. 물체의 contact geometry 즉 접촉표면 형상을 이미지로 변환하는 센싱방식으로 GelSight라는 tactile sensor가 있는데, 근데 문제는 이런 이미지 기반 촉각 데이터 자체가 high-resolution 이다보니 데이터 수집과 학습 측면에서 좀 챌린징했다고 합니다. 데이터 수집은 사실 누구나 느끼듯이 사람이 직접 로봇의 grasping 동작을 관찰하고 여러 센서들 HW들 연결 직접하고 이런 과정이 반복되어야했기 때문에 노동집약적이고 비효율적이죠. 요 교수님도 이 점에 집중해서 Taxim이라는 매우 적은 실제 샘플 데이터만 인풋에 넣고도 real 센서의 동작을 근사하는 시뮬레이션 모델을 개발했다고 합니다. 이게 말씀해주시기로는 굉장히 계산이 빠르고 calibration이 쉬워서, 한 10~20개 남짓한 샘플 데이터로도 CPU 머신에서도 돌아가는 대략 18FPS 정도되는 속도로, 실제 데이터 센싱이랑 거의 근사한 데이터를 보였다고 하여 좀 오호~ 했던 기억이 있습니다.

이렇게 자기들이 개발한 시뮬레이션 활용해서 데이터 수집의 자동화를 어느정도 만들었나 싶었는데, 문제가 하나 더 있었습니다. 시뮬레이터 환경에서 데이터 증강을 하고자 하는 영규형의 연구 얘기를 몇번 들었던 지라 결국엔 모델에 학습하고 나면 sim과 real간의 domain gap 으로 인한 성능 저하 문제도 생길 수 밖에 없지 않나? 하고 들으면서 생각했었는데, 살짝은 다른 늬앙스로 domain shift 문제를 표현하더라고요. 바로 sensor domain shift 였습니다. 같은 접촉 상황이라도 기본적으로 tactile 데이터가 기존 다른 모달리티들에 비해 공간해상도랑 센서정밀도, 노이즈 이런거에 민감해서 데이터 처리나 share에 관한 일반화가 어렵다는 문제 있었습니다.

이것도 결국 물리 기반 optical simulation 내에서 풀어냈다고 합니다. GelSight와 같은 광학식 촉각 센서를 sim 내부에서 물리적, 광학적으로 요리조리 센서 내부의 반사 굴절 경로를 모두 추적하면서, 실제 센서처럼 sim의 센서를 모델링 해서 real같은 sim GelSight 촉각 이미지를 생성했다고 합니다. 촉각 이미지 데이터 자동 생성하는 파이프라인도 생겼겠다~, 이 교수님은 결국 이거 가지고 다양한 물체의 3D mesh를, 다양한 포즈, 접촉 위치 다 고려해서 10k의 대규모 합성 촉각 데이터셋을 구축했다고 합니다. 아래 논문(arxiv 2025)이 해당 연구였던 것 같습니다.

SENSOR-INVARIANT TACTILE REPRESENTATION

Harsh Gupta* **Yuchen Mo*** **Shengmiao Jin** **Wenzhen Yuan**
University of Illinois Urbana-Champaign
{hgupt3,yuchenm7}@illinois.edu

ABSTRACT

High-resolution tactile sensors have become critical for embodied perception and robotic manipulation. However, a key challenge in the field is the lack of transferability between sensors due to design and manufacturing variations, which result in significant differences in tactile signals. This limitation hinders the ability to transfer models or knowledge learned from one sensor to another. To address this, we introduce a novel method to extract Sensor-Invariant Tactile Representations (SITR), enabling zero-shot transfer across optical tactile sensors. Our approach utilizes a transformer-based architecture trained on a diverse dataset of simulated sensor designs, allowing generalizability to new sensors in the real world with minimal calibration. Experimental results demonstrate our method's effectiveness across various tactile sensing applications, facilitating data and model transferability for future advancements in the field.

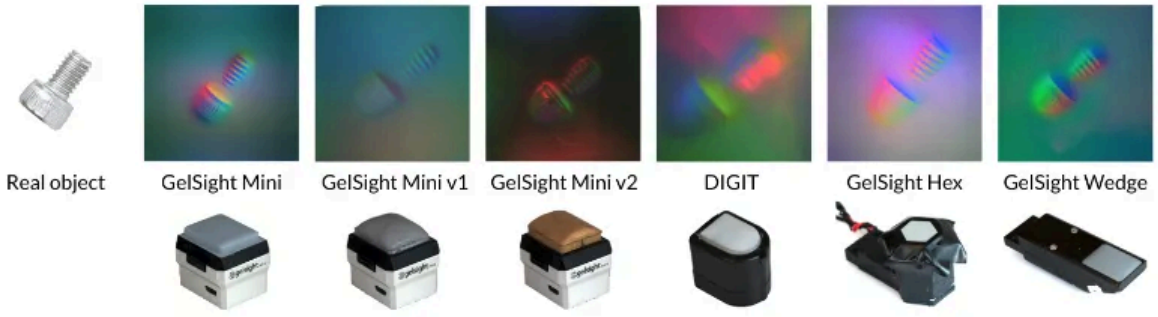


Figure 1: Vision-based tactile sensors vary in both optical design and physical properties. Even with the same contact object, a screw, the tactile images produced by each sensor differ significantly. These variations highlight the challenge of transferring models from one sensor to another.

그래서 이제 이 교수님은 대규모 합성 촉각 이미지 데이터 구축했으니, 이런 데이터에서의 sensor-agnostic representation 을 학습하기 위한 모델 아키텍쳐도 개발합니다. 특정 센서에 의존하지 않게, 접촉 형상 정보만을 표현하도록 contrastive learning이랑 어찌저찌 학습했다고 하는데 학습 과정은 사실 잘 모르겠습니다. 결론은 물체 형상복원, 물체 분류, 접촉 위치 추정 등과 같은 다운스트림 태스크에서, sim으로만 학습한 모델이 real 센서 데이터 인풋으로도 괜찮은 zero-shot 전이 성능을 보였습니다.

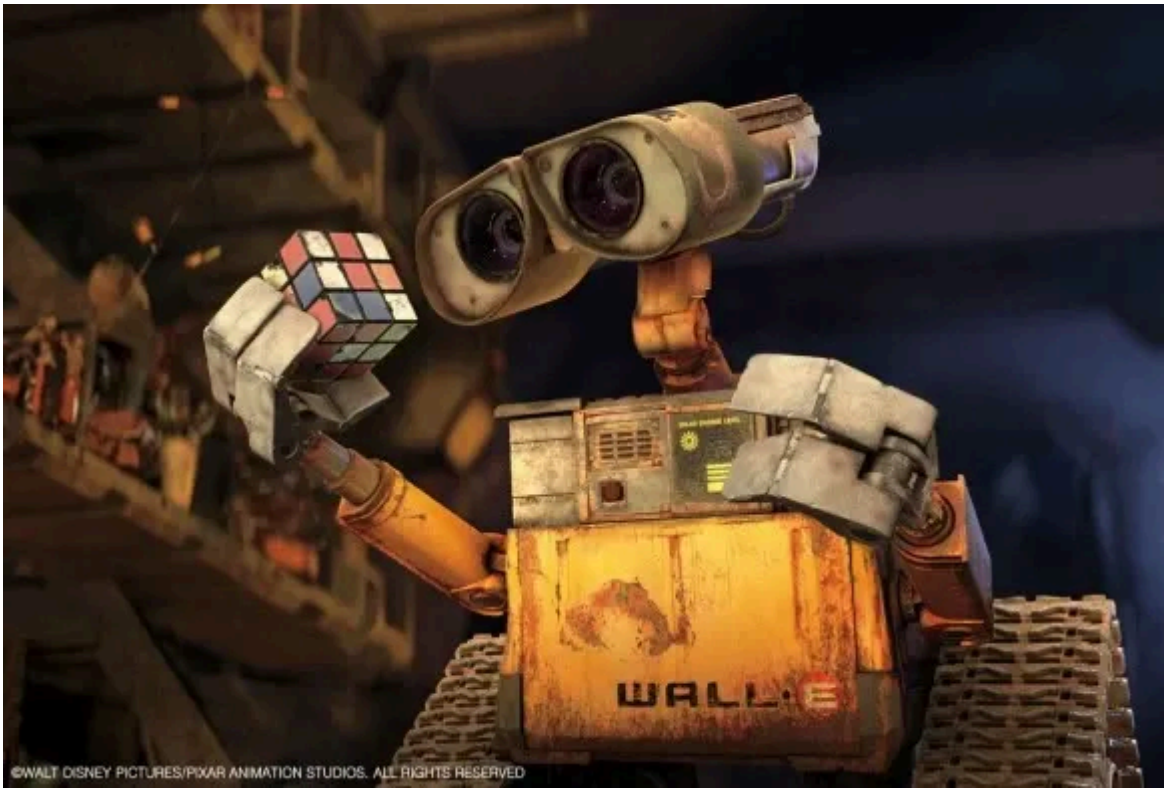
마지막으로는 현재 연구가 물체에 접촉을 반드시 해야만 하는 센서 위주의 연구였으나, 앞으로는 force sensor, 압력분포센서, magnetic 센서 등 더 다양한 센서 간 전이 문제 확장으로도 연구 계획이 있다고 했었습니다. 근데 아직 뭐 센서 간 표현 차이가 너무 크기에 완전한 전이가 가능한 것은 아니어서 새로운 데이터 수집 방식이 필요하다고 했습니다.

메모해놨던 질의응답에서 기억에 남았던 건, 시뮬레이터에서 센서 파라미터를 optimizing하는 방식으로 디자인하는 방향은 어떻게 생각하냐?는 질문에 발표하신 교수님이 놀라면서 실제로 해당 주제 publication 준비 중이라고 말했던 기억이 납니다. 역시 세계적인 석학들은 발표만 듣고도 뭔가 딱 캐치하고 서로 통하는 게 남다릅니다.

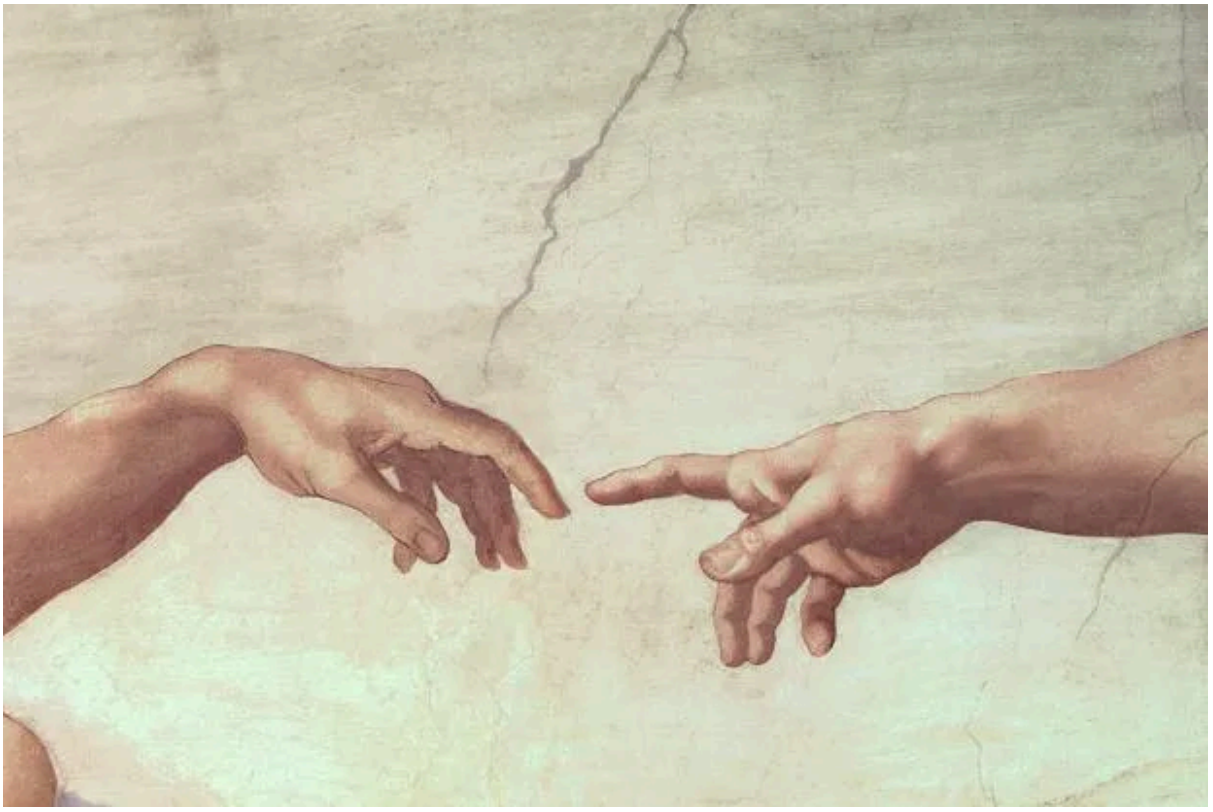
다음은 미시간대학교 Nima 교수님입니다.

Why is Manipulation so Damn hard? 이게 첫 장이었나 그랬던 것 같은데 뭔가 시작도 어그로를 확실히 끌어서 기억에 남는데, 해결방법보단 왜 dexterous manipulation이 어려운 지에 대해 고찰한 내용이 많았던 것 같습니다.

먼저 영화 Wall-E의 귀여운 로봇을 보여주시며 이목을 끌었는데요. 이 로봇이 영화에서 보면 완전 호기심 가득한 채로 막 돌아다니고 뭐든 보고 만지고 배우고 그러는데, 사실 우리들이 원하는 로봇의 이상적인 모습이 바로 이런 거라고 하면서



manipulation의 본질은 'contact' 이라고 강조합니다. 이게 manipulation이란 게 단순히 의도를 가진 'action'인 게 아니라, 의도를 가진 'contact'로 '인지'를 하는 능력이라고 보더라구요. 'manipulation이란 개념이 항상 접촉을 해야만 의미있는 거였나? 아까 그 전 발표한 교수님은 접촉 안하는 센서 모달리티도 고려하고 있다던데..' 하고 순간 갸우뚱했지만 바로 납득되었습니다. 사람도 뭔갈 피부로 접촉하면서 촉각정보로써 동작과 주변 물체들을 이해하던 것이 조작의 본질적인 행위니까요.



발표에서 미켈란젤로의 천지창조 그림 예시도 나오는데 손가락끼리 touch 하는 부분을 보여주면서, 인간의 촉각은 시각, 운동보다 먼저 진화한 근본적인 모달리티인데, 로봇 manipulation의 핵심도 접촉의 인식과 힘, geometry 등에 대한 이해가 근간이 되어야된다고 합니다.

근데 dex manipulation이란게 물체 종류, 접촉 위치, 접촉 시간 조합의 경우의 수가 exponential complexity를 갖고 있고, 항상 partially observed하며, contact에 대한 실제 physics가 미분불가해서 일단 이런 physics적 제약이 쉽지 않다고 합니다. 또 못 박으려면 망치말고도 여러가지 물체(프라이팬, 아령, 물병 등)로도 못 박을 수 있지 않느냐고 하며 semantic한 제약도 있다고 합니다. 이 두가지 제약을 말하면서, dex manipulation이란 게 physics, semantic 모두 어려운 복잡한 태스크라고 하는데, 결국엔 로봇 조작의 진정한 언어는 텍스트나 언어가 아니라 geometry와 force라고 강조하고 되게 문제 고찰을 열심히 해주시고 해결책의 경우는 열린 결말로만 남기시더라구요. 제어나 물리모델 기반의 온라인 MPC, 사람 시연 데이터 기반의 IL, sim에서의 오프라인 RL 등 세 접근법의 결합이 Foundation model들 위주로 아마도 필요하게 될 것이며, 이 때 tactile HW가 산업계에서 빠르게 경제적 가치를 인정받으면 tactile-driven한 방식들이 manipulation을 해결할 핵심이 되면서 tactile 연구가 아마 완전 뜰거다 라고 말했습니다.

Panel Talk(Humanoids)

우선 Google Deepmind에서 로봇 학습의 시대 변화에 대해서 서두를 던집니다.

- 2018 single arm RL 대두.
- 2020 mobile manipulator 대두.
- 2023 diffusion policy 기반 IL 대두.
- 2025 휴머노이드 대두.

그러면서 자기네들 Gemini Robotics 데모랑 연구결과 잘 보고 들었지? 자기네들이 충분히 이 휴머노이드 흐름에 근접해가고 있다고 어필 한번 합니다.

그 다음 Nvidia에서도 나와서 Gr00tN1.6 향후 공개할거다 라고 어필 한번 해줍니다. 특히 인상깊었던 게 human video data로부터 지식을 흡수하는 implicit world modeling 방법과 video world model을 통한 일반화를 보여주는 DreamGen 연구를 많이 강조했습니다. 저도 학회기간동안 속으로 월드 모델이 LLM과 VLM을 대체할 가능성이 있다면 결국 인간의 상상 역할을 대체하는 역할을 할 것이고 그 영향력이 매우 커질 것으로 생각했었는데, Nvidia가 확신을 갖고 말해주니 월드 모델도 LLM/VLM 처럼 엄청난 FM 패러다임으로 자리매김할 것 같은 막연한 생각이 들었습니다. 앞으로는 월드 모델에 대해서도 공부해보고 가능하다면 그러한 월드 모델의 latent한 feature 또한 활용해보고 싶다는 생각도 들었습니다.

그 다음은 Agility Robotics에서 상업적, 실제 안전 관점으로 humanoids 의 역할에 대해 소개했는데, 무거운 거 들고, 밸런스 잘 잡아야되고, 굉장히 다양한 툴 다뤄야하고 등등... 단순한 작업을 통한 점진적인 generality 확보가 휴머노이드 개발의 가장 빠르고 유일한 경로라고 주장했었습니다. Logistics -> Manufactoring -> Grocery/Retail -> Home 순으로 휴머노이드는 점차 넘어오게 될 것이라고 보더라구요.

또 Galaxia Dynamics 같은 경우는, 공장의 달인분들처럼 빠른 일처리 속도를 현재 로봇은 못 보이지 않냐. 로봇은 이제는 show용이 아니라 실제 real labor force로써 써야한다. robotics에서의 AGI는

Artificial General Intelligence가 아니라 Agility, Generalization, Infallibility가 되어야 한다. IL의 느린 deploy 속도를 개선하기 위해, RL이 로봇의 속도 측면의 문제 해소를 불러올거라고 보기에 → DemoSpeedUP(deploy 속도 개선)과 같은 연구가 가치가 있다 라고 언급했었습니다.

기억에 남는 토의주제는 다음과 같았습니다.

Q1: 전통적인 모델링 제어기법 **vs. policy** 러닝기법 중에서**,** 휴머노이드에는 뭐가 더 최적인가**?**

대부분 완벽한 physics 모델링이 어렵기 때문에 MPC는 어렵다고 보고, 대부분 RL, e2e IL 같은 러닝 기법이 맞다고 보고있었습니다.(MIT 김상배 교수님 한분만이 그건 태스크마다 다른 거다. 지능이 필요하지 않은 직관적이고 결정론적인 상황(예; 핑퐁)에선 MPC의 필요성이 있다. 결국 모델링과 러닝 간의 혼합 방식이 적절하다고 어필하셨습니다.)

Q2: 로봇 러닝 **for** 휴머노이드에서 특별한 점은 무엇이며**,** 다른 로봇과 어떻게 유사하고 다른가**?**

휴머노이드 연구는 인간의 형태를 모방하여 인간 중심의 데이터같이 **non-humanoid robot data**를 직접 활용해서 **policy learning**에 **transfer**할 가능성이 있다. 하지만, 자유도가 너무 높고 복잡한 시스템이라는 문제가 있어서 “Robot Learning 으로는 State Space Exploding 문제에 직면해 있다.” 가 공통적으로 들리는 표현이었습니다. 또한 느린 움직임에서는 인간과 유사해 보일지라도 역학적 측면에서는 인간을 모방하는 것이 올바른 방향이 아닐 수 있다고 보기도 하고**,** 오히려 **human relevance** 때문에 인간의 생활 환경에 가장 적합한 형태라는 주장도 있었습니다. 그리고 가장 중요한 게 균형과 locomotion에 대한 문제해결이 필요하면서도 다른 로봇과 마찬가지로 manipulation을 해결하는 게 가장 어렵다고 했습니다. 또 누군가는 우리 할머니한테 프랑카보여줬더니, 그냥 팔이란다. 일반 대중들에겐 로봇==휴머노이드로 받아들여지는 인식이 있다고도 말하길래 듣고 피식했던 기억이 납니다.

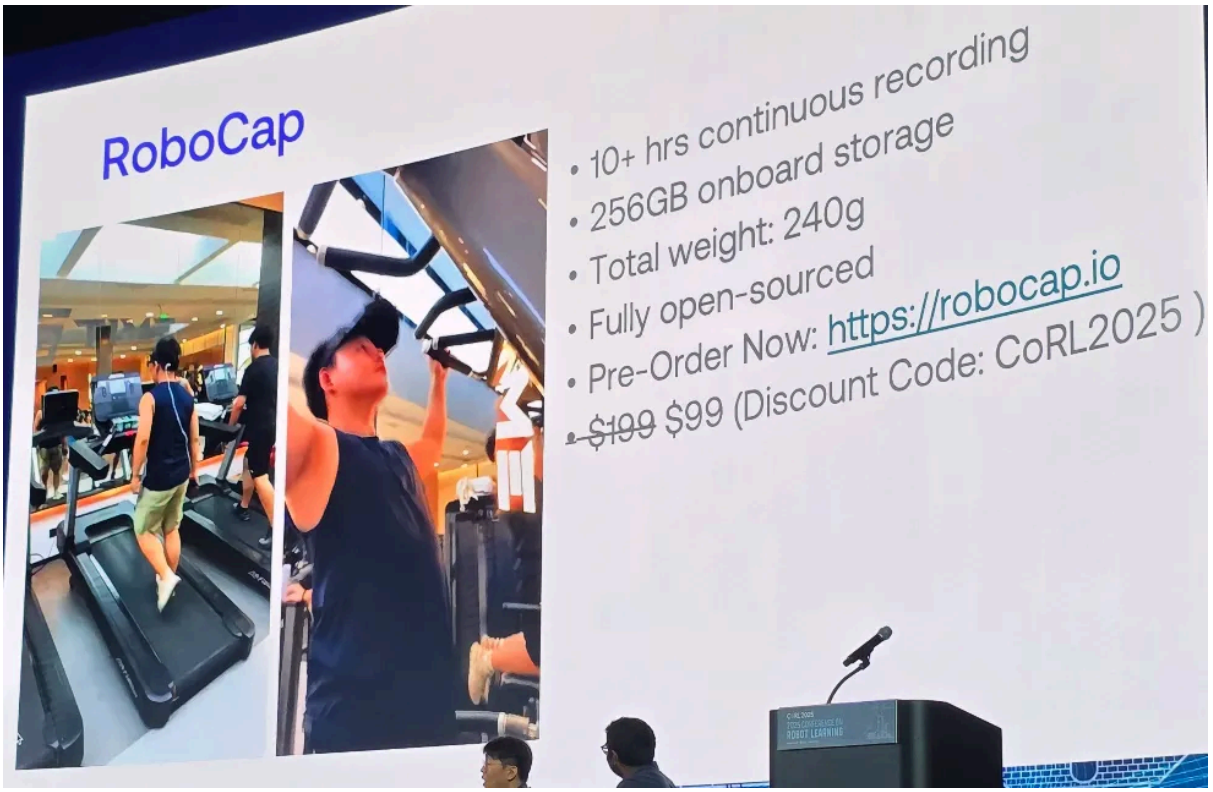
Q3: 로봇 러닝에서 우리 정말 **scaling data is all you need** 하냐**?** 휴먼데이터랑 얼마나 연관지을 수 있느냐**?**

teleop으로 얻는 trajectory 데이터로는 휴머노이드가 수행하려는 많은 일에 대한 데이터를 쉽게 표현할 수 없으며 결국 합성 데이터, 시뮬레이션, RL 등으로 generate해야 한다고 보는 시각이 많았습니다. 더구나 dexterous manipulation 도 수행해야하기에 tactile contact sensing도 필수다, 이를 위한 teleop 방식과 HW 개선도 필수다 라고 보았습니다. 제 개인적인 생각으로는 생각해보니 LLM도 현재 지구상의 거의 모든 text를 학습했다고 하지만 아직도 hallucination 문제가 남아있는데, humanoid policy도 나중엔 결국 action hallucination이 해결하기 어려운 큰 난제로 남을 것 같다는 생각도 들었습니다.

Q4: 현재 휴머노이드 연구에서 가장 큰 **bottleneck**이 뭐라고 생각하냐**?**

- multifinger dexterity.
- VLA 등 러닝기법의 아키텍처를 어떻게 구축할지.
- 시뮬레이션 및 학습을 위한 HW
- 빠르고 높은 신뢰도의 video model.
- HW 구축과 시뮬에서의 데이터
- reliable data.

마지막으로 요 정도가 있었습니다.



이 질문에서는 데이터 취득을 위한 HW 얘기가 좀 많았는데, 정말 많은 다양한 데이터취득을 위한 HW 구성 방식들이 쏟아져 나오고 있는 것 같았습니다. 기억에 남는 특이한 거 하나가 패널톡 시간 때는 아니었지만 diamond sponsor talk 세션에서의 bitrobot 에서 RoboCap이라는 모자가지고 egocentric 데이터 취득하는 것이었습니다. human video 로부터 latent action을 transfer 하는 방법론들이 점차 등장하고 그런 흐름으로 대부분 실제 로봇 HW를 최소화한 방식으로 데이터 취득하는 걸 목표로 다들 삼고있다보니, 언급은 안 했지만 학회기간동안 이런 저런 효율적인 HW들을 개발하고 쓴 논문들과 데모들이 쏙쏙 보이던 게 생각납니다.

Q6: 휴머노이드는 정확히 어디서부터 휴머노이드가 아니게 되는가**? (Agility**의 **DIGIT**처럼 무릎이 뒤로 꺾인 형태**, Galaxia**처럼 바퀴가 달린 형태 등**)**

사람과 휴머노이드 간에 draw the line을 어떻게 할거냐의 관점이었습니다. 대체로 다음과 같았습니다.

- "human-centric robot. not human-like robot". 인간스럽게 생긴게 문제가 아니라 인간 중심의 태스크를 수행해줄 줄 아는 로봇이 휴머노이드라는 주장.
- 사람의 말처럼 확률적**(stochastic)**인 개념이기에 무엇이 휴머노이드인지를 명확히 구분(delineate)하기는 어렵다는 주장.
- 태스크에 따라 달라진다는 주장**.** 바닥이 평평한 공장에선 이족 보행 휴머노이드가 아닌 모바일 매니퓰레이터 형태로도 휴머노이드 취급할 수 있다란 주장.

등이 있었습니다.

마지막으로 청중 과의 질의 중 저랑 똑같은 궁금증을 가진 사람이 있었는데, Nvidia 연사에게 World Model과 VLA의 통합에 대한 생각을 물어봤었습니다. 답변은 예상대로 Nvidia에서도 긍정적으로 보는 연구라고 하며, Video Model 혹은 World Model이 픽셀 수준으로 디테일을 예측하는 능력만 더 올라가면 행동 예측을 위한 어떤 대리 경로를 주는 역할을 하기에 좋을 것 같다하면서도, 아직 속도문제, prepriotive, tactile 과의 통합이 챌린징한 요소라고 답변했던 기억이 납니다.

느낀점 한 단락

커피 타임 때 주어진 커피를 정말 많이 마시면서 어떻게든 뭐라도 더 들어볼까 하면서 버티듯이 돌아다니곤 했습니다. 역시나 피곤을 당연히 감내할 수 있을 정도로 정말 많은 인사이트와 정말 많은 경험을 얻을 수 있는 좋은 기회였습니다. 더구나 영어는 다들 해외 학회 참석하시면 아마 공통적으로 드는 생각이실 것 같지만, 이 바닥은 영어는 잘하면 무조건 플러스 알파, 아니 +알파+베타+감마가 맞는 것 같습니다. 학회는 네트워킹하는 장소라는 것을 체감했던 게 정말 많은 사람들이 Oral이나 Keynotes 이런 발표 세션들에서 집중하고 치열하게 고찰하고 이러는 게 아니더라구요. 다들 Poster 세션에서 저자와 허심탄회하고 솔직하게 얘기하면서 인사이트를 얻고, Poster 세션 아니어도 그냥 복도에서 다른 나라의 친분 있는 사람과 만나서 이래저래 연구 고민 얘기하고, 친분 없어도 맘에 드는 논문의 저자가 있으면 스스럼없이 링크드인을 교환하고, 본인의 흥미와 관심과 궁금증을 막힘없이 자연스레 툭툭 내뱉는 게 정말 부러웠습니다. 그 중 가장 가까운 사람이 영규형이었구요. 영규 형한테 전해듣는 인사이트도 굉장히 많아서 더 그런 생각이 들었던 것 같네요. 저도 한 걸음씩 성장할 수 있기를 바랍니다.

마지막으로 좋은 기회를 제공해주신 교수님 그리고 학석사연계사업(AI로봇연구교육과정)을 지원해주신 모든 분들께 감사하다는 말씀 드리면서 마치도록 하겠습니다.