Bridging LLM Agents to Physical World

👨‍🎓박사과정 김정환 | jhbale11@snu.ac.kr

2025.01.13

HCC LAB SNU

Background & Motivation

From Digital to Physical

Nvidia Live at CES 2026 in Las Vegas

(26.01.05) Jensen Huang says,

"AI is Leaving the screen to enter the physical world."

Photo: Patrick T. Fallon/AFP via Getty Images

Agenda

1 Model : NVIDIA GR00T

2 Platform : LeRobot

3 Workflow : NVIDIA Blueprint

4 Proposal

2025-3-28

GR00T N1: An Open Foundation Model for Generalist Humanoid Robots

NVIDIA[1]

HCC LAB SNU
HUMAN CENTERED COMPUTING LABORATORY
서울대학교 SEOUL NATIONAL UNIVERSITY

NVIDIA GR00T

Problem : Data Pyramid

Figure 1: **Data Pyramid for Robot Foundation Model Training.** GR00T N1's heterogeneous training corpora can be represented as a pyramid: data quantity decreases, and embodiment-specificity increases, moving from the bottom to the top.

- **Top (Real-World Data)**
  - 휴머노이드가 실제 Real world에서 구동하는 데이터
  - 가장 적음, 가장 비쌈
- **Middle (Synthetic/Sim Data)**
  - 물리 시뮬레이션 데이터
  - 아직은 많이 없음, 완성도 낮음
- **Base (Web & Human Videos)**
  - 가장 많음, Robot Action이 없음

HCC LAB SNU

NVIDIA GR00T

GR00T Overview - A Generalist Agent

- **Concept** : 다양한 로봇(Humanoid, Arm)을 제어할 수 있는 Foundation Model.
- **Dual-System Architecture** (인간 인지 모델 차용) :
  - System 2 (Reasoning): 느리지만 논리적. VLM (Vision-Language Model) 기반.
  - System 1 (Action): 빠르고 반사적. DiT (Diffusion Transformer) 기반.

HCC LAB SNU

NVIDIA GR00T

GR00T Overview - A Generalist Agent

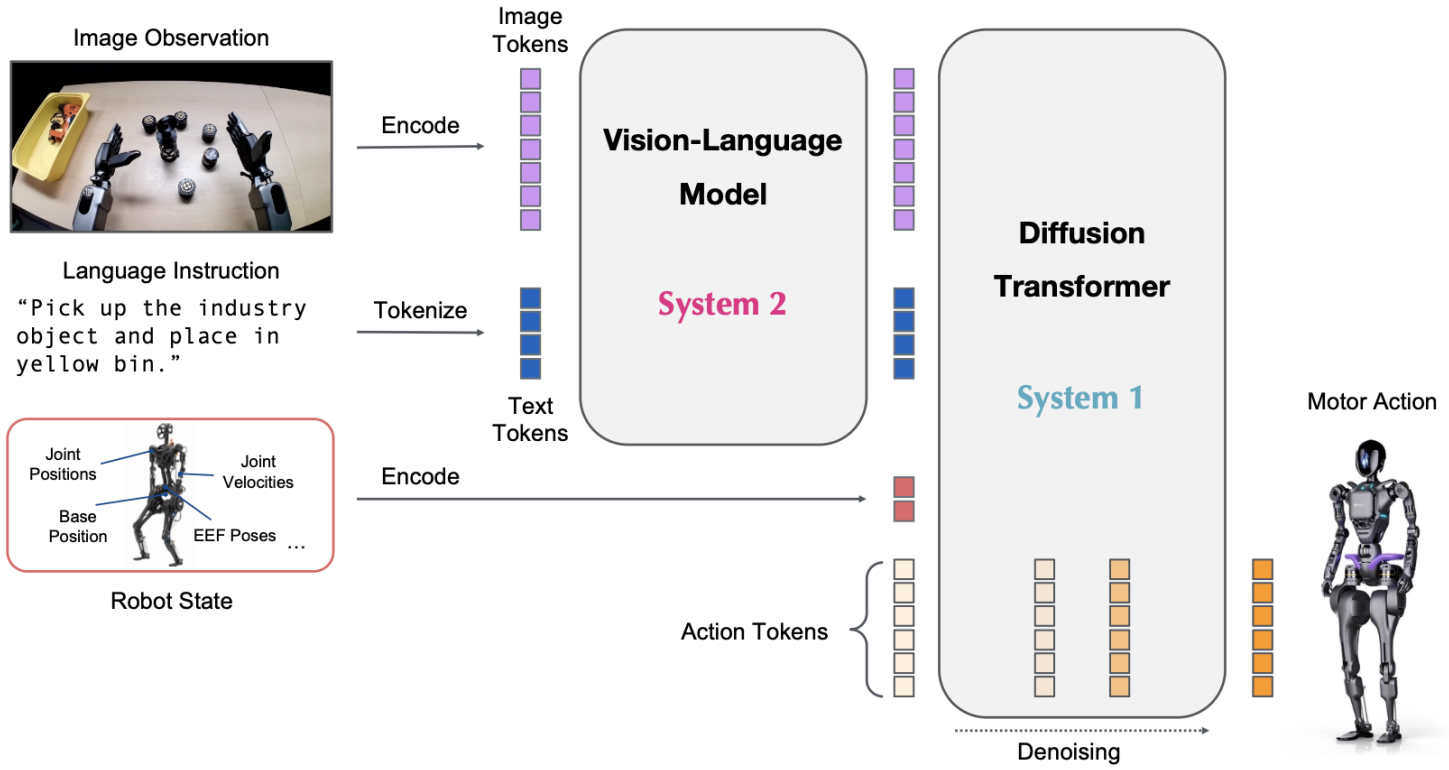
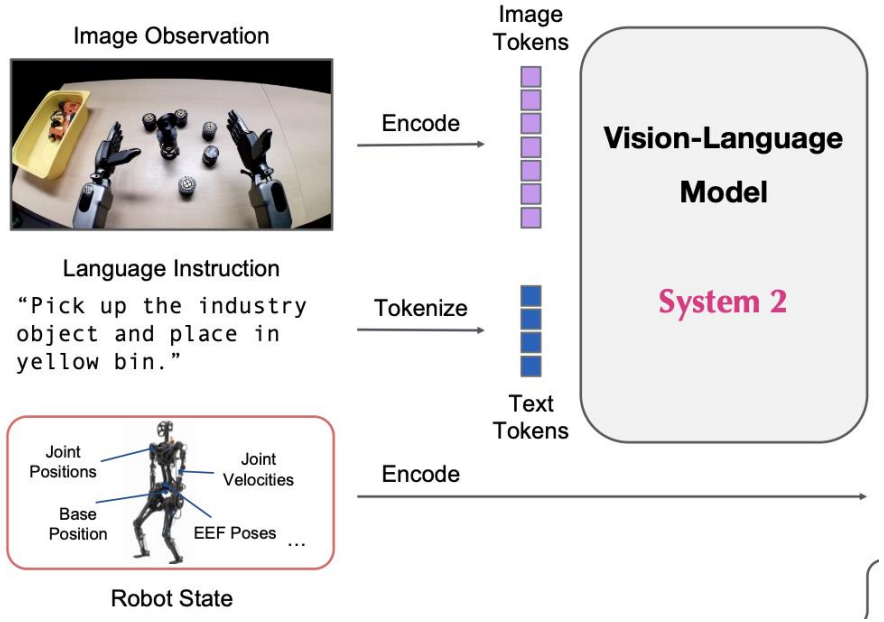


Figure 2: **GR00T N1 Model Overview.** Our model is a Vision-Language-Action (VLA) model that adopts a dual-system design. We convert the image observation and language instruction into a sequence of tokens to be processed by the Vision-Language Model (VLM) backbone. The VLM outputs, together with robot state and action encodings, are passed to the Diffusion Transformer module to generate motor actions.

HCC LAB SNU

NVIDIA GR00T

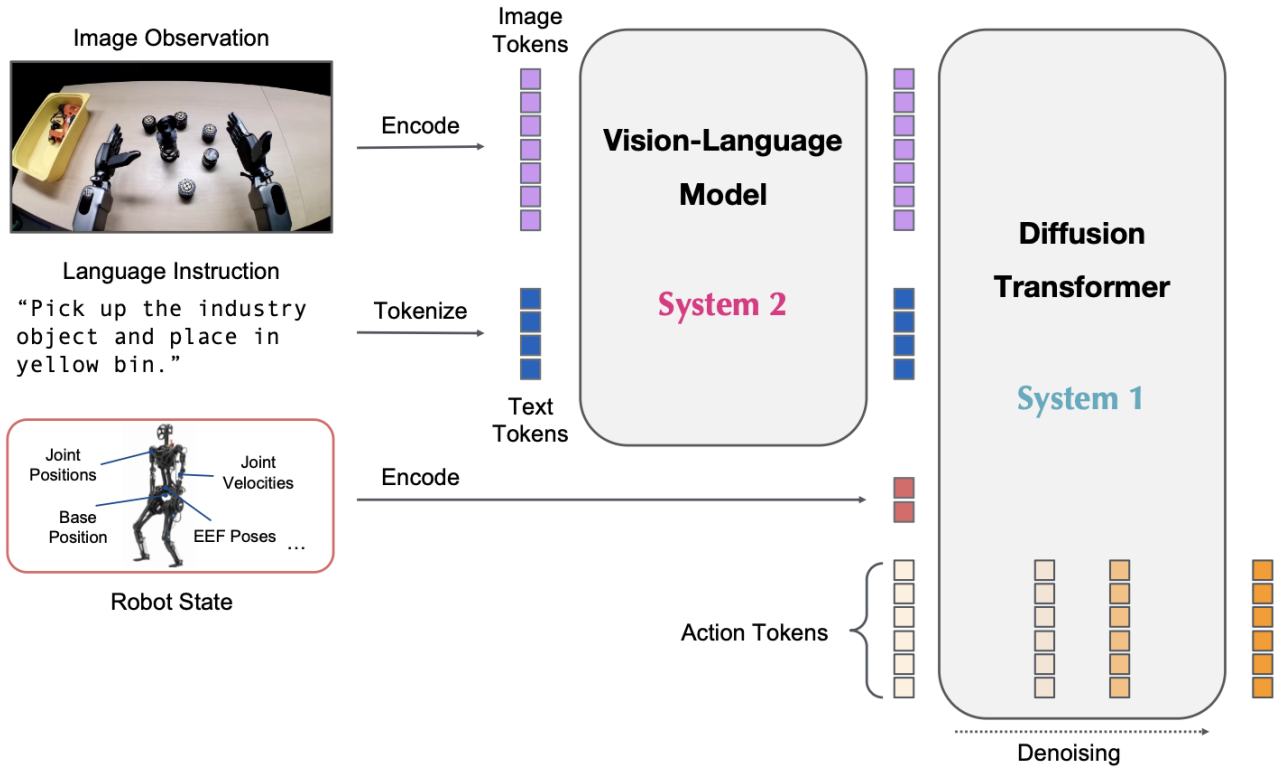
Methodology 1 - System 2 (VLM Backbone)



- **Architecture :** NVIDIA Eagle-2 (VLM) 활용.
- **Process :**
  - **이미지**와 **텍스트 명령(Instruction)**을 입력 받음.
  - VLM이 이를 처리하여 고차원적인 Vision-Language Token을 생성.
  - 이 토큰들은 직접적인 제어 신호가 아니라, System 1(Action Model)의 Conditioning Context로 사용됨.
- **Insight :** LLM Agent가 단순 텍스트 출력이 아닌, 행동 생성을 위한 '맥락'을 제공하는 역할로 진화.

NVIDIA GR00T

Methodology 2 - System 1 (Action Flow Matching)



- **Architecture** : Diffusion Transformer (DiT).
- **Mechanism (Flow Matching) :**
  - 기존 Diffusion보다 학습 안정성이 높은 Flow Matching 기법 사용.
  - 노이즈(Random Noise)로부터 시작해, System 2의 토큰을 조건(Cross-Attention)으로 받아 점진적으로 Action Chunk(일련의 동작)를 생성.
- **Action Chunking :** 한 번에 1스텝이 아닌, 미래 H 스텝의 동작을 한 번에 예측하여 동작의 부드러움(Smoothness) 확보.

NVIDIA GR00T

Methodology 2 - System 1 (Action Flow Matching)

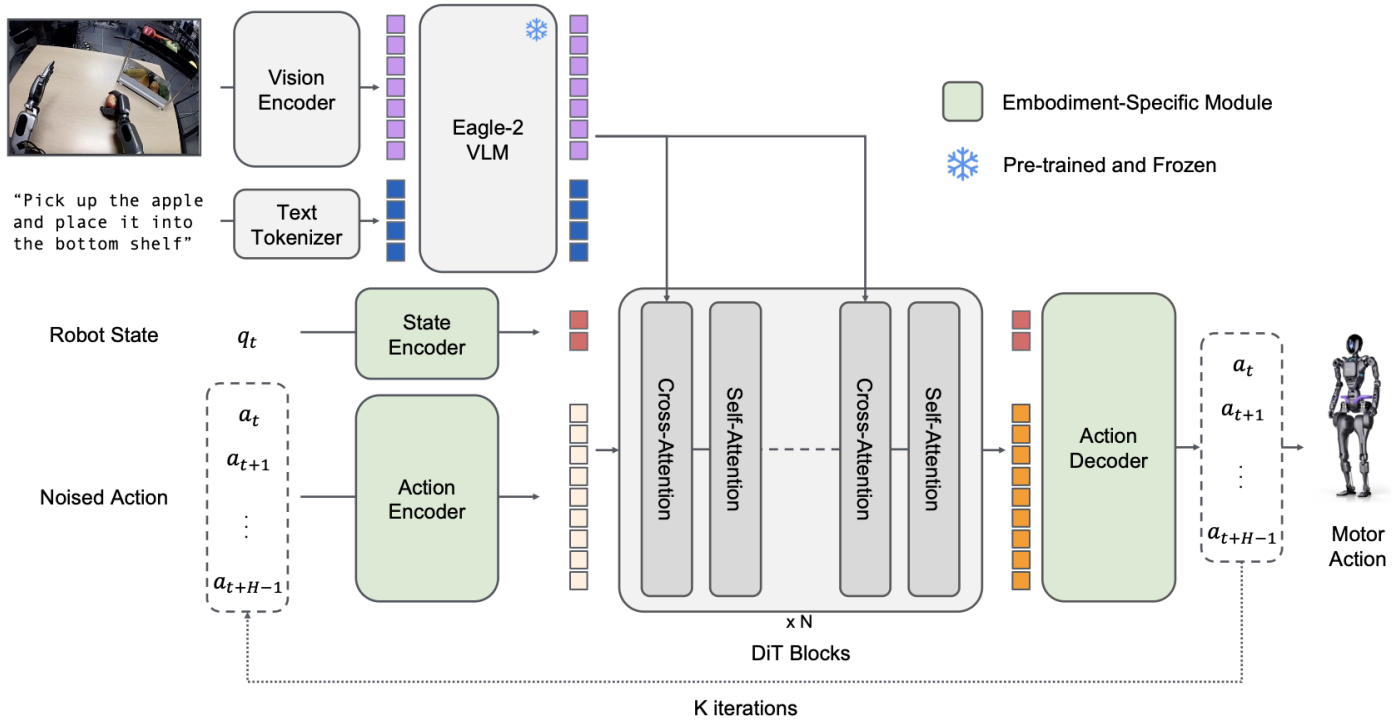


Figure 3: **GR00T N1 Model Architecture.** GR00T N1 is trained on a diverse set of embodiments ranging from single-arm robot arms to bimanual humanoid dexterous hands. To deal with different robot embodiment's state observation and action, we use DiT blocks with an embodiment-aware state and action encoder to embed the robot's state and action inputs. GR00T N1 model leverages latent embeddings of the Eagle-2 model to incorporate the robot's visual observation and language instructions. The vision language tokens will then be fed into the DiT blocks through cross-attention layers.

- **Architecture** : Diffusion Transformer (DiT).
- **Mechanism (Flow Matching) :**
  - 기존 Diffusion보다 학습 안정성이 높은 Flow Matching 기법 사용.
  - 노이즈(Random Noise)로부터 시작해, System 2의 토큰을 조건(Cross-Attention)으로 받아 점진적으로 Action Chunk(일련의 동작)를 생성.

- **Action Chunking :** 한 번에 1스텝이 아닌, 미래 H 스텝의 동작을 한 번에 예측하여 동작의 부드러움(Smoothness) 확보.

NVIDIA GR00T

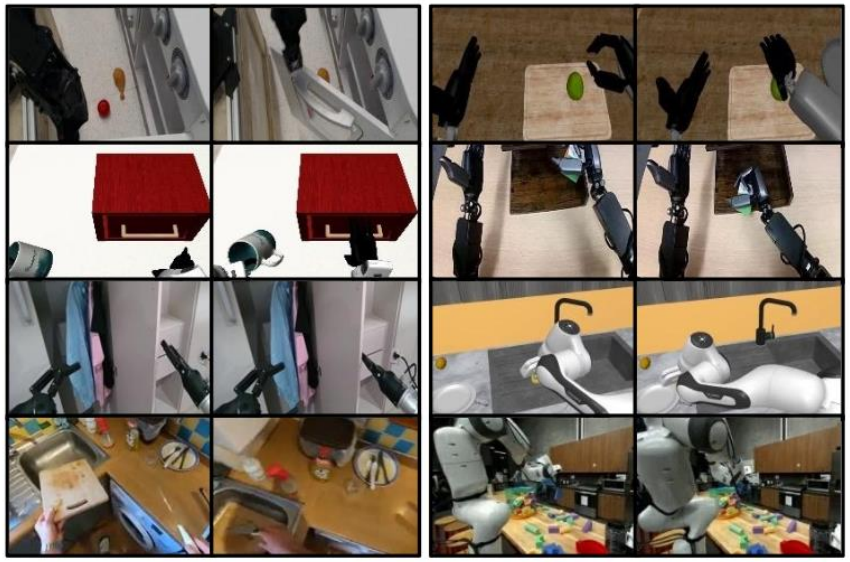
Methodology 3 - Training with "Action-less" Data

- **Problem:**
  - 로봇 데이터는 적고, 사람 비디오(YouTube 등)는 많다.
  - 하지만 사람 비디오엔 Action(Joint Angle) 정보가 없다.
- **Solution :** Latent Action Learning (VQ-VAE)
  - 비디오의 연속된 프레임 간의 변화를 압축하여 Latent Action(잠재 행동)으로 정의.
  - 실제 로봇의 물리적 Action이 없더라도, 영상 속의 '변화' 자체를 Action으로 간주하고 학습.
  - 이후 적은 양의 실제 로봇 데이터로 Fine-tuning.

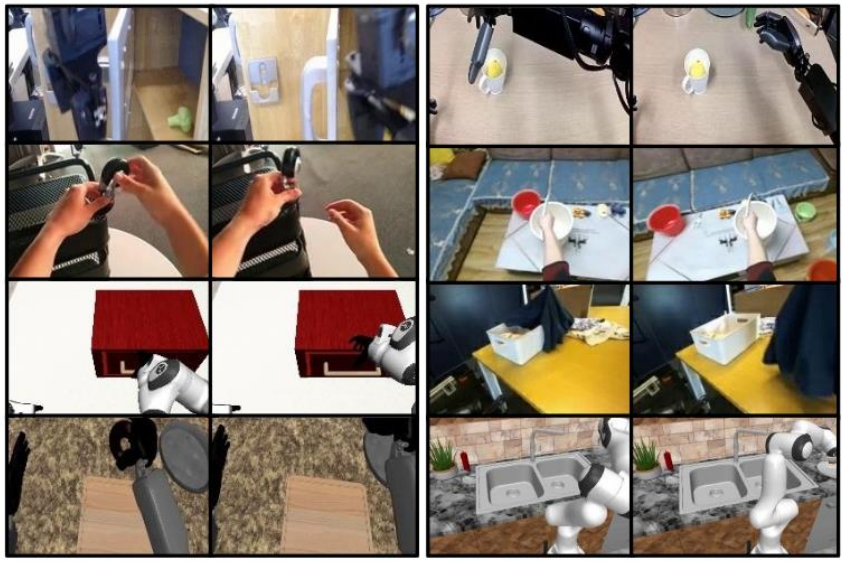
HCC LAB SNU

NVIDIA GR00T

Methodology 3 - Training with "Action-less" Data



Latent action 1: Move the right arm to the left



Latent action 2: Move the right arm to the right

Figure 4: **Latent Actions.** We retrieve similar latent embeddings across various embodiments. The left images illustrate the latent action that corresponds to moving the right arm (or hand) to the left, while the right images illustrate the latent action that corresponds to moving the right arm (or hand) to the right. Note that this general latent action is not only consistent in different robot embodiments, but also in human embodiment.

HCC LAB SNU

NVIDIA GR00T

GR00T's Performance & Implication

- **Generating Training Data - From Human to Robot**
  - **Latent Actions :** Human Video에는 Action(Joint 값)이 없음 → VQ-VAE로 Latent Action을 학습하여 대체.
  - **Neural Trajectories :** Video Generation Model(비디오 생성 AI)을 사용해 하나의 궤적을 다양한 시점/배경으로 증강(Augmentation).
  - **DexMimicGen :** 시뮬레이션 상에서 인간의 시연을 로봇 데이터로 대량 증식.

- **Implication :**
  - Simulation 및 Real-world(GR-1 Humanoid)에서의 우수한 성능.
  - 10%의 데이터만으로도 기존 SOTA(Diffusion Policy)를 능가.

- **결론 : "양질의 데이터와 강력한 Base Model(VLM)이 있다면, 적은 양의 Real Data로도 가능하다."**

HCC LAB SNU

LEROBOT: AN OPEN-SOURCE LIBRARY FOR END-TO-END ROBOT LEARNING

Anonymous authors
Paper under double-blind review

HCC LAB SNU
HUMAN CENTERED COMPUTING LABORATORY
서울대학교 SEOUL NATIONAL UNIVERSITY

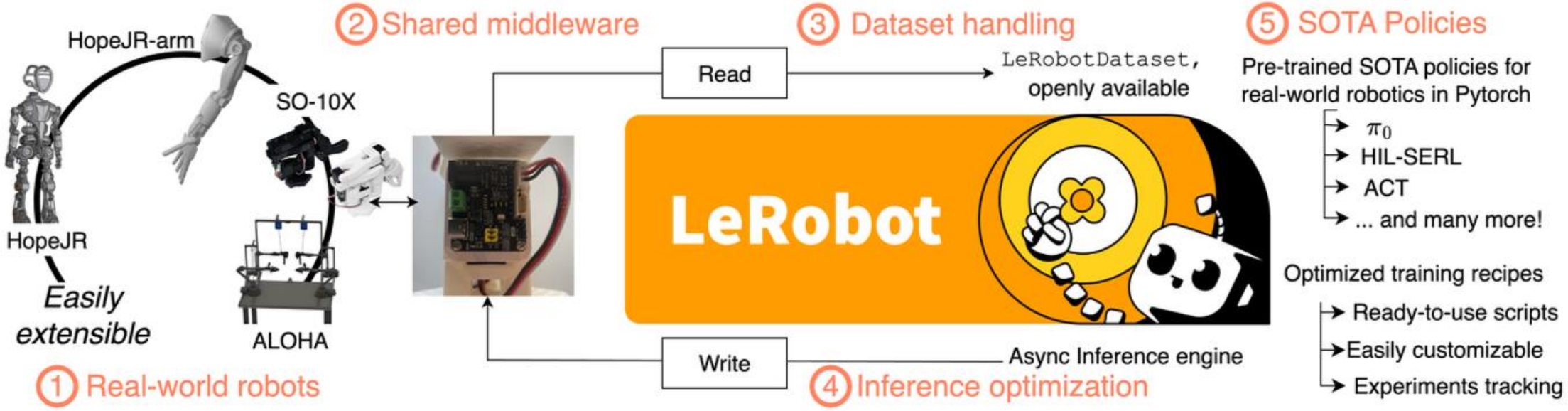
LeRobot

Problem

- **미들웨어의 파편화(Disaggregated Middleware)**
  - 로봇 하드웨어와 소프트웨어를 연결해주는 '미들웨어'가 로봇 제조사나 플랫폼마다 다름.
- **데이터 형식의 불일치(Datasets and Formats)**
  - 대규모 로봇 데이터셋들이 서로 다른 형식(TensorFlow, ROS bags, JSON)으로 공유됨.
- **학습 프레임워크의 재현성 부족(Learning Frameworks)**
  - 딥러닝 알고리즘의 아주 사소한 구현 차이나 데이터 처리 방식의 차이가 결과에 큰 영향.
  - 로봇은 하드웨어 특성까지 타기 때문에 더 심각함.

LeRobot

End-to-end Robot Learning with LeRobot



LeRobot

Why LeRobot? (The "Hugging Face" of Robotics)

- **Philosophy :**
  - 폐쇄적인 로봇 소프트웨어를 타파하고, PyTorch 기반의 접근성 높은 라이브러리 제공.
- **Integration :**
  - 우리 연구실이 익숙한 Hugging Face 생태계(데이터셋 허브, 모델 공유)와 완벽 호환.
- **Hardware Agnostic :**
  - 특정 로봇에 종속되지 않음 (나중에 로봇을 사도 코드 수정 최소화).

LeRobot

LeRobot Dataset Format (Standardization)

- **Structure :**
  - observation.images: 카메라 뷰 (MP4로 압축 저장)
  - observation.state: 로봇 관절 상태 (Joint position/velocity)
  - action: 제어 신호
- **Format :**
  - Parquet + MP4 조합 사용.

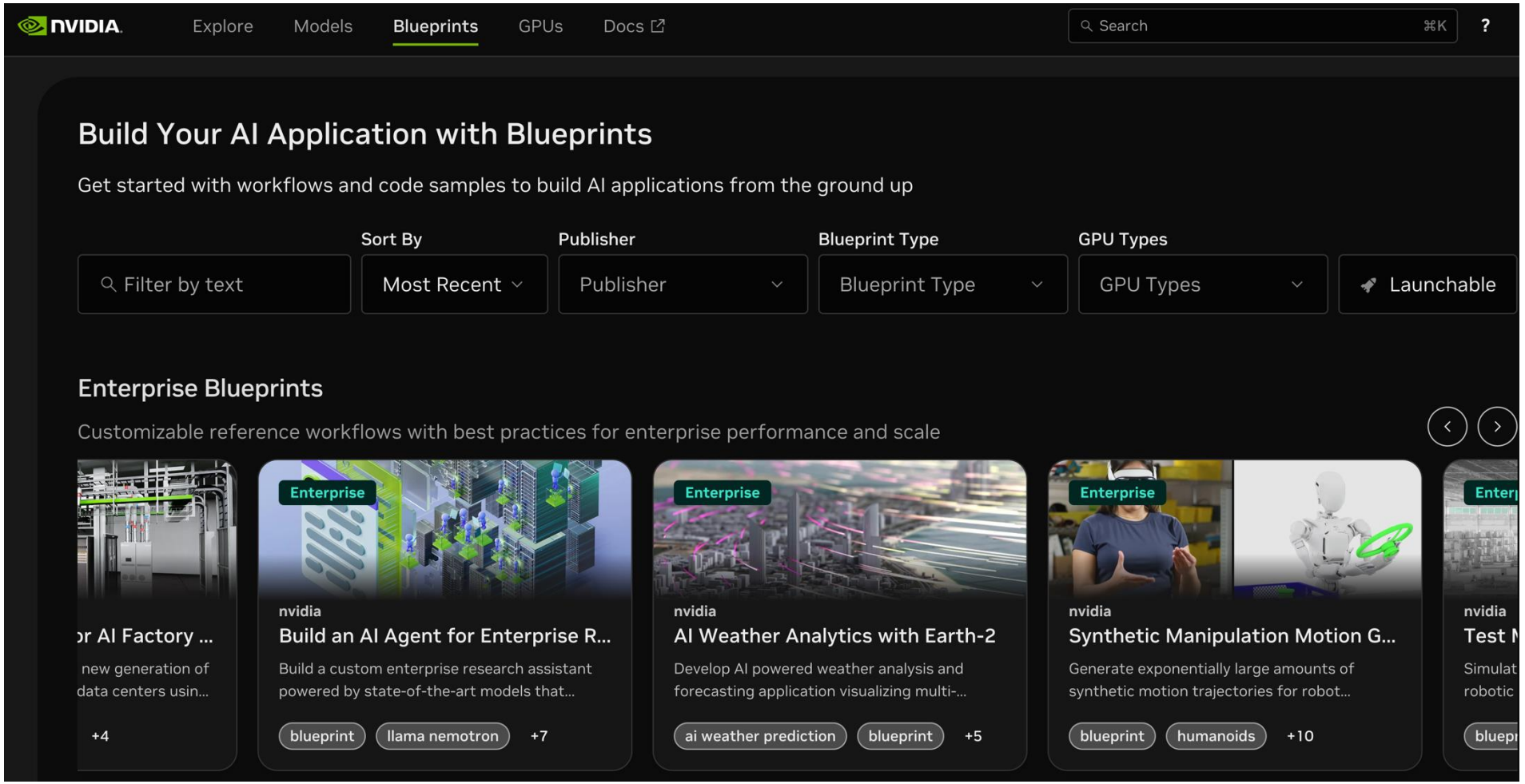
  (가볍고 빠름, Streaming 가능)

https://huggingface.co/spaces/lerobot/visualize_dataset

LeRobot

Practical Implementation Plan

- **Simulation Environment**
  - LeRobot은 MuJoCo 기반의 Gym 환경을 기본 제공 (Aloha, PushT)
- **Training Loop**
  - 데이터셋 로드 → 전처리 → Policy 학습 → Sim 평가 (Python으로 가능)
- **Plan**
  - 하드웨어 도입 전까지 LeRobot의 Sim 환경에서 모델 학습 및 데이터 파이프라인 최적화



NVIDIA Blueprint

HCC LAB SNU

VERI LUX TAS MEA

NVIDIA Blueprint

What is NVIDIA Blueprint for Humanoid?

- **Definition**
  - 로봇 개발의 전 과정(데이터 수집 → 학습 → 배포)을 가속화하기 위해 NVIDIA가 제공하는 Reference Workflow 패키지
- **Core Components for Us**
  - **AI-based Teleoperation** : Apple Vision Pro / VR 기기를 활용한 인간 행동 캡처.
  - **Isaac Lab** (Simulation) : Python 기반의 로봇 학습 및 시뮬레이션 프레임워크.
  - **MimicGen** : 소수의 인간 데모를 대량의 합성 데이터(Synthetic Data)로 증폭.
- **Why Blueprint?**
  - "맨땅에 헤딩"하지 않고, NVIDIA가 검증한 SOTA 파이프라인(GR00T)을 그대로 차용 가능.

HCC LAB SNU

NVIDIA Blueprint

What is NVIDIA Blueprint for Humanoid?

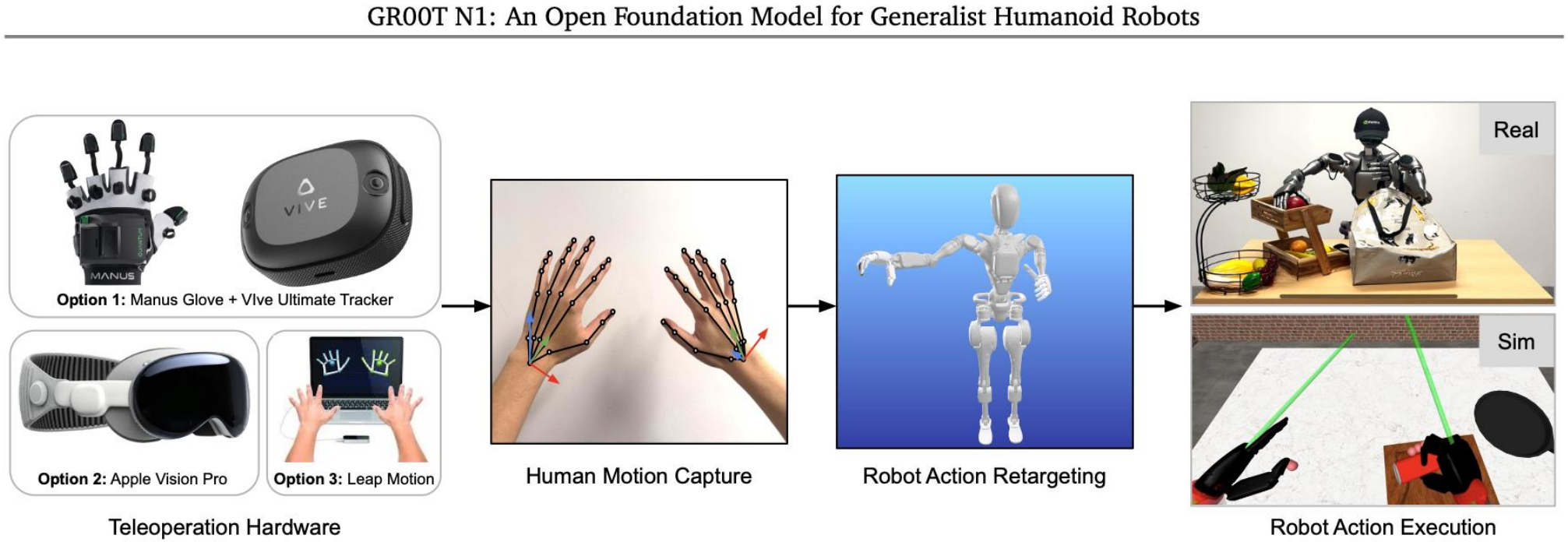


Figure 6: **Data Collection via Teleoperation.** Our teleoperation infrastructure supports multiple devices to capture human hand motion, including 6-DoF wrist poses and hand skeletons. Robot actions are produced through retargeting and executed on robots in real and simulation environments.

NVIDIA Blueprint

Workflow Step 1 - Teleoperation & Retargeting

- **Definition**
  - 로봇 개발의 전 과정(데이터 수집 → 학습 → 배포)을 가속화하기 위해 NVIDIA가 제공하는 Reference Workflow 패키지
- **Core Components for Us**
  - **AI-based Teleoperation** : Apple Vision Pro / VR 기기를 활용한 인간 행동 캡처.
  - **Isaac Lab** (Simulation) : Python 기반의 로봇 학습 및 시뮬레이션 프레임워크.
  - **MimicGen** : 소수의 인간 데모를 대량의 합성 데이터(Synthetic Data)로 증폭.
- **Why Blueprint?**
  - "맨땅에 헤딩"하지 않고, NVIDIA가 검증한 SOTA 파이프라인(GR00T)을 그대로 차용 가능.

HCC LAB SNU

NVIDIA Blueprint

Workflow Step 2 - Scaling via MimicGen

- **Problem :** VR로 1,000번 시연하는 것은 노동 집약적임.
- **Solution : MimicGen** (Automatic Data Generation)
  - **원리:** 사람이 수행한 1개의 성공적인 궤적(Source Demonstration)을 1,000개의 다양한 상황으로 변형.
  - **구체적 방법:**
    - Object Pose Randomization: 컵의 위치를 5cm씩 바꿈.
    - Lighting/Texture Variation: 조명, 식탁 색상을 바꿈 (Visual Robustness).
    - Motion Interpolation: 시작점과 끝점을 알고 있으므로, 중간 경로를 물리 엔진상에서 안전하게 재생성.
- **Output** : 이 과정을 거치면 하드웨어 없이도 수만 개의 'Labeling이 완료된' 고품질 로봇 데이터 확보 가능.

HCC LAB SNU

NVIDIA Blueprint

Simulation Environment - NVIDIA Isaac Lab

- **Platform** : Isaac Lab (기존 Isaac Gym/Orbit의 후속작).
- **특징 (Why Good for Our Lab?):**
  - Python First: PyTorch와 완벽 호환.
  - Fast: GPU 기반 물리 엔진(PhysX 5)으로 수천 개의 환경을 병렬 시뮬레이션.
  - Asset Library: RoboCasa (부엌 환경), Google Scanned Objects 등 고품질 3D 에셋을 바로 import 가능.
- **Sample Code Location :**
  - GitHub: nvidia-omniverse/isaaclab
- **Key Scripts :**
  - scripts/teleop_se3_agent.py: 키보드/VR로 로봇 팔 조작하는 예제.
  - scripts/rsl_rl/train.py: 수집된 데이터로 강화학습/모방학습 돌리는 예제.

https://github.com/isaac-sim/IsaacLab?tab=readme-ov-file

Research Proposal

Practical Blueprint for Our Lab

- **Scenario :** "커피 타기(Making Coffee)" Task 데이터셋 구축.
- **Implementation Steps:**

  1) Environment Setup : Isaac Lab에서 Kitchen 환경 로드 (RoboCasa 에셋 활용).

  2) Human Demo : 연구원이 Vision Pro를 쓰고 가상의 부엌에서 커피를 타는 동작 수행 (약 50회 녹화).

  3) Data Generation : MimicGen 스크립트를 돌려 50회 → 50,000회 데이터로 증강 (밤샘 배치 작업).

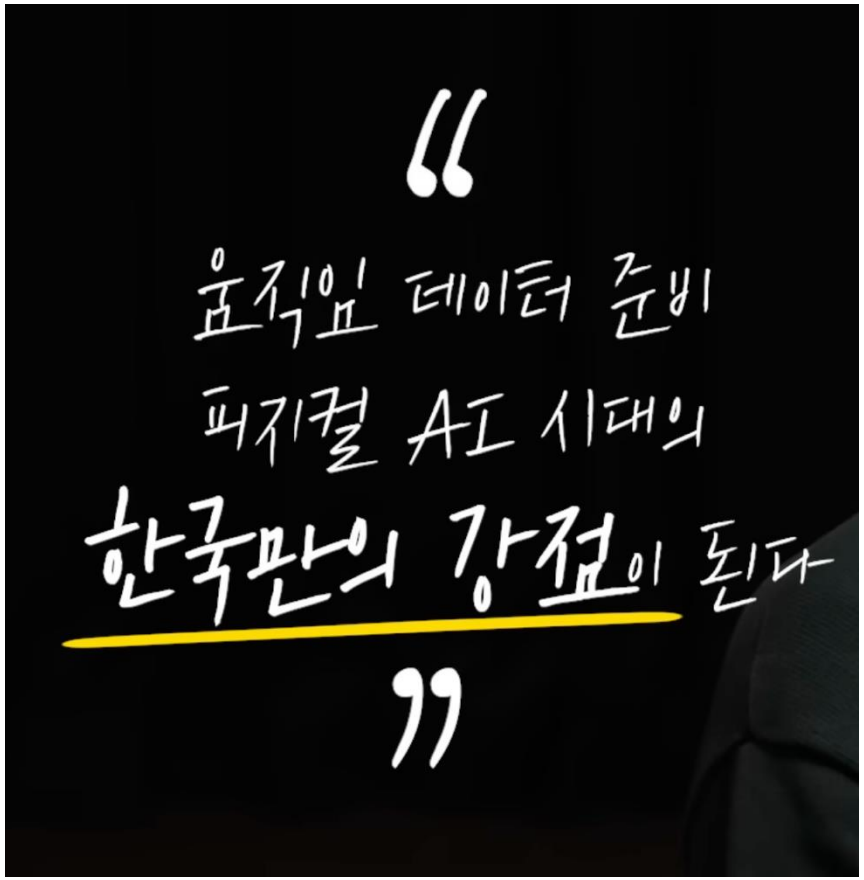
  4) Export : 데이터를 LeRobot 포맷(Parquet+MP4)으로 변환하여 저장.

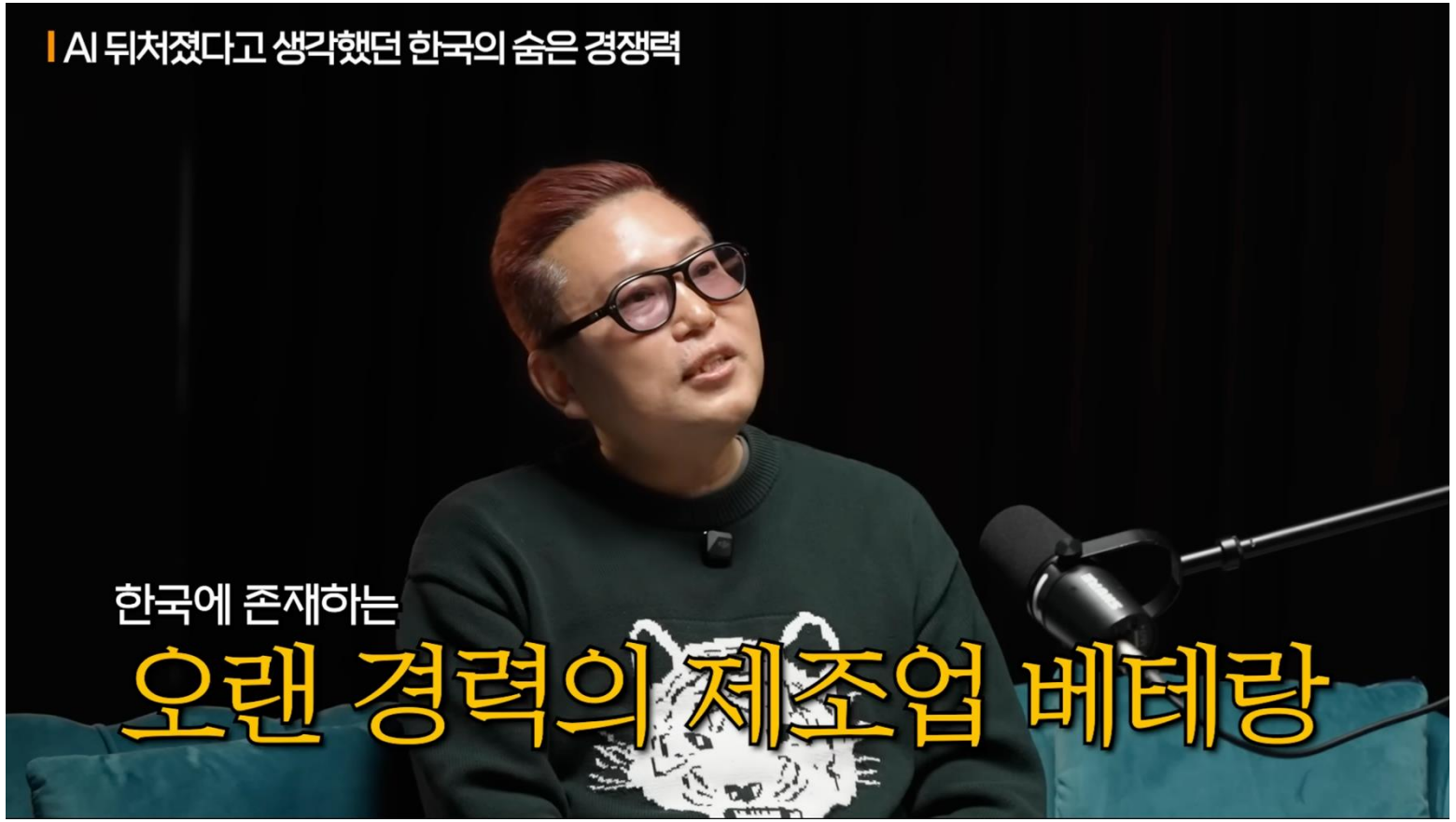
  5) Training : 이 데이터를 사용해 GR00T N1 혹은 ACT 모델 학습.

- **이 모든 과정에 실물 로봇은 필요 없음.**

HCC LAB SNU

Research Proposal





https://www.youtube.com/watch?v=3oIL71er4YI

Research Proposal

Practical Blueprint for Our Lab

- **Scenario :** 제조업 숙련공/장인들의 Task 데이터셋 구축.
- **Implementation Steps:**
  1) Environment Setup : Isaac Lab에서 필요 환경 구축
     💡) 환경 구축 과정부터 숙련공에 대한 인터뷰 통해 실제 작업하는 환경과 유사하게 구축
  2) Human Demo : 숙련공이 Vision Pro를 쓰고 가상의 환경에서 작업 동작 수행 (약 50회 녹화).
     💡) 왜 이렇게 행동했는지에 대한 실시간 annotation 방법론
  3) Data Generation : MimicGen 스크립트를 돌려 50회 → 50,000회 데이터로 증강 (밤샘 배치 작업).
     💡) Generated Data에 대한 LLM Agent 검수 또는 Agent 기반의 인간의 검수 과정 보조
  4) Export : 데이터를 LeRobot 포맷(Parquet+MP4)으로 변환하여 저장.
  5) Training : 이 데이터를 사용해 GR00T N1 혹은 ACT 모델 학습.

HCC LAB SNU

감사합니다

Thank You

HUMAN
CENTERED
COMPUTING
LABORATORY