JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 31 No. 11, pp. 167-180, November 2026
https://doi.org/10.9708/jksci.2026.31.11.167

# Layer-Selective PEFT of Pi-Zero VLA for Trajectory Prediction and Control on Cross-Embodiment Robot Arms

Yu-Been Ha*, In-Yeop Choi**

*Student, Dept. of Computer Science, Kangnam University, Gyeonggi-do, Korea
**Professor, Dept. of Computer Science, Kangnam University, Gyeonggi-do, Korea

## [Abstract]

Vision-Language-Action (VLA) models have demonstrated promising capabilities for general-purpose robot manipulation. However, transferring pretrained VLA models to new robot embodiments remains challenging due to differences in robot morphology, kinematics, and action spaces. To address this cross-embodiment problem, this paper proposes a Layer-Selective Parameter-Efficient Fine-Tuning (PEFT) method. The proposed method freezes the language model backbone and applies layer-selective Low-Rank Adaptation (LoRA) to the high-level semantic layers (Blocks 18–26) of the vision encoder and to all layers of the Action Expert. The robot-specific input/output layers, which directly correspond to the robot's physical action dimensions, are fully fine-tuned. This design preserves pretrained knowledge as much as possible and adapts only the embodiment-dependent components, enabling efficient adaptation to a new robot with a small amount of new data. All experiments were conducted on a single robot in a single environment. To validate the effectiveness of the proposed method, the adapted VLA model was deployed on an NVIDIA Jetson AGX Orin edge platform and evaluated on multiple tasks with a physical 6-DoF Magician E6 robot arm. In real-robot experiments, the model achieved success rates of 92.1% in the object recognition test involving pretrained (seen) objects and a newly added object, and 100% in the color discrimination test, without catastrophic forgetting. The pick-and-place test achieved a success rate of 90%. These results show that layer-selective PEFT can effectively achieve cross-embodiment adaptation with limited training data and only about 0.44% of the total parameters trained, while preserving pretrained visual-semantic knowledge.

▸ **Key words**: VLA Model, PEFT, Cross Embodiment, robotic manipulation, ROS 2, Pi-Zero

• First Author: Yu-Been Ha, Corresponding Author: In-Yeop Choi
*Yu-Been Ha (gkdbqls889@kangnam.ac.kr), Dept. of Computer Science, Kangnam University
**In-Yeop Choi (billychoi@kangnam.ac.kr), Dept. of Computer Science, Kangnam University
• Received: 2026. 09. 07, Revised: 2026. 10. 31, Accepted: 2026. 11. 03.

Copyright © 2026 The Korea Society of Computer and Information
http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945
www.kci.go.kr

**[요 약]**

비전-언어-행동(VLA) 모델은 범용 로봇 조작에서 유망한 역량을 보여주었다. 하지만 로봇의 형태, 운동학 및 행동 공간의 차이로 인해 사전 학습된 VLA 모델을 새로운 로봇 임바디먼트에 전이하는 것은 여전히 어려운 과제로 남아 있다. 본 논문은 이러한 크로스 임바디먼트(cross-embodiment) 문제를 해결하기 위해 계층 선택적 PEFT(Layer-Selective Parameter-Efficient Fine-Tuning) 방법을 제안한다. 제안 방법은 언어 모델 백본을 동결하고, 비전 인코더의 상위 블록(18-26)과 액션 전문가의 전 계층에 저랭크 적응(LoRA, Low-Rank Adaptation)을 적용하며, 로봇의 상태와 행동 벡터를 다루는 로봇 특화 입출력 층은 전체 파인튜닝한다. 이를 통해 사전 학습된 지식은 최대한 유지하면서 임바디먼트에 종속적인 부분만 조정하여, 적은 양의 신규 데이터로 효율적으로 적응할 수 있도록 한다. 제안 방법을 검증하기 위해 적응된 모델을 NVIDIA Jetson AGX Orin 엣지 플랫폼에 배포하고 실제 6-DoF Magician E6 로봇팔에서 평가하였다. 그 결과 사전 학습에 포함되고 파인튜닝 데이터에서는 제외한 객체와 새로 추가한 객체를 대상으로 한 객체 인식 테스트에서 92.1%, 색상 구별 테스트에서 100%, pick-and-place 테스트에서 90%의 성공률을 달성하였다. 이는 시험한 객체에 대해 치명적 망각이 발생하지 않았음을 뒷받침하나, 여러 개의 처음 보는 객체를 이름으로 구별하는 능력은 확인되지 않았다. 모든 실험은 단일 로봇과 단일 환경에서 수행하였다. 이러한 결과는 계층 선택적 PEFT가 적은 학습 데이터와 전체 파라미터의 약 0.44%에 해당하는 학습만으로 크로스 임바디먼트 적응을 효과적으로 수행할 수 있음을 보여준다.

▸ **주제어**: 비전-언어-행동 모델, 저랭크 적응, 크로스 임바디먼트, 로봇팔 조작, 로봇 운영체제, 파이제로

## I. Introduction

최근 인공지능 분야에서는 시각 정보 인지, 자연어 이해 그리고 물리적 행동 제어를 통합하는 VLA(Vision Language Action) 모델 연구가 가속화되고 있다. 기존의 로봇 제어 파이프라인은 객체 검출(Object Detection), 포즈 추정(Pose Estimation), 모션 플래닝(Motion Planning) 등의 개별 모듈로 분리되어 연산 지연과 에러 누적(Error Accumulation) 문제를 겪었다. 이와 달리 VLA 모델은 입력된 이미지와 자연어 명령을 바탕으로 엔드투엔드(End-to-End) 방식으로 로봇의 제어 토큰을 생성함으로써 높은 유연성과 지능적 작업 수행 능력을 보여준다. 특히 오픈소스 기반의 VLA 모델인 파이제로(Pi-Zero)[1]는 대규모 멀티모달 학습을 통해 높은 오픈 도메인 일반화(Generalization) 성능을 갖춘 것으로 알려져 있다. 파이제로는 사전 학습된 VLM(Vision-Language Model)에 플로우 매칭(flow matching) 기반 행동 생성(action generation) 구조[2]를 결합한 VLA 모델로, 시각-언어 의미 이해 능력과 연속 행동 생성 능력을 결합하였다. 다양한 로봇 플랫폼에서 언어 명령 기반 조작과 새로운 스킬(skill)에 대한 파인튜닝(Fine-Tuning) 가능성을 보였다. 이 모델은 다양한 시각적 환경과 텍스트 명령어 간의 상관관계를 깊이 있게 학습한다. 그 결과 학습 과정에서 직접 경험하지 못한 비정형 환경이나 유사 객체에 대해서도 강인한 제로샷(Zero-shot) 및 퓨샷(Few-shot) 적응력을 나타낸다.

그러나 이러한 뛰어난 시각-언어 일반화 성능에도 불구하고, 사전 학습 단계에서 고려되지 않은 서로 다른 하드웨어 구조를 가진 크로스 임바디먼트 로봇팔에 VLA 모델을 그대로 이식하는 데에는 여전히 기술적 어려움이 존재한다. 파이제로가 사전 학습(Pre-training) 시 학습한 로봇의 기하학적 메커니즘과 실제 현장에 도입하려는 로봇팔 간에는 링크 구조, 관절의 가동 범위, 운동학(Kinematics)적 특성에서 비롯되는 상당한 도메인 갭(Domain Gap)이 발생하기 때문이다. 모델의 모든 가중치를 업데이트하는 전체 파인튜닝(Full Fine-Tuning)을 시도할 경우, 막대한 연산 자원이 소모된다. 이는 컴퓨팅 리소스가 제한된 임베디드 엣지 플랫폼에서 메모리 부족(Out of Memory) 문제를 유발할 수 있다. 이뿐만 아니라 파이제로 모델이 본래 지니고 있던 강력한 시각-언어 일반화 지식 기반마저 파괴하는 치명적 망각(Catastrophic Forgetting)을 초래할 위험이 있다. 그러므로 파이제로 모

www.kci.go.kr

델을 실제 배포하기 위해서는 전체 모델을 재학습하는 방식보다 제한된 로컬 데이터와 하드웨어 자원 안에서 필요한 모듈만 선택적으로 조정하는 PEFT 전략이 필요하다. LoRA[3]는 PEFT의 여러 방법론 중 현재 가장 널리 쓰이고 효율성이 검증된 핵심 알고리즘이다. 이 알고리즘은 사전 학습 모델의 전체 가중치를 업데이트하지 않고 저랭크 행렬만 학습함으로써 학습 파라미터 수와 GPU 메모리 사용량을 크게 줄이는 기법이다.

본 논문에서는 파이제로 VLA 모델의 우수한 범용 일반화 성능을 보존하면서 새로운 로봇 임바디먼트에 효율적으로 적응할 수 있는 계층 선택적 PEFT를 제안한다. 제안한 방법의 실효성을 검증하기 위해, 사전 학습 데이터에 포함되지 않은 로봇팔인 Dobot Magician E6를 대상으로 관절 궤적의 실시간 예측 및 제어를 수행하였다. 이를 위해 학습에 사용하지 않은 새로운 객체를 포함하여 총 180개의 에피소드로 구성된 시각-행동 데이터셋을 구축하고, 이를 이용해 파이제로 모델을 파인튜닝하였다. 파인튜닝된 VLA 모델은 임베디드 시스템인 NVIDIA Jetson AGX Orin(64GB)에 배포되었으며, 이 시스템은 유선 이더넷(Ethernet)을 통해 Magician E6와 연결된다. 런타임 시 모델은 엔드이펙터 포즈 6차원과 흡착 명령 1차원으로 이루어진 7차원 행동 청크를 생성한다. 엔드이펙터 포즈는 MoveIt 2의 역기구학을 통해 6개 관절각으로 변환되고, ROS 2 기반 통신을 통해 ServoJ 명령으로 로봇팔에 전달된다. 실제 태스크(task)로는 객체를 집거나(pick) pick-and-place 작업을 사용하였다. 이 테스트를 통해 두 가지를 검증하고자 하였다. 첫째는 사전 학습에 포함되어 있던 객체와 새롭게 추가한 객체에 대한 인지 및 조작 성능을 유지하면서 동시에 완전히 새로운 미학습 객체를 정확히 식별하는지를 확인하는 것이다. 둘째는 4가지 종류의 색상을 구별하는 색상 기반 조작 테스트를 성공적으로 수행하는지를 확인하는 것이다. 이로써 본 연구가 제안하는 계층 선택적 파인튜닝이 크로스 임바디먼트 로봇으로의 전이 학습에 효과적임을 입증하고자 하였다.

## II. Preliminaries

### 1. Related works

#### 1.1 Research trends and limitations of existing research

최근 로봇 조작 분야에서는 시각 관측, 자연어 명령, 행동 생성을 하나의 정책 모델로 통합하는 VLA 모델과 로봇 파운데이션 모델에 대한 연구가 활발히 진행되고 있다. RT-1[4]은 대규모 실제 로봇 궤적 데이터를 활용하여 트랜스포머 기반 정책이 다양한 작업, 객체, 환경에 대해 일반화될 수 있음을 보였다. RT-2[5]는 인터넷 규모의 VLM에서 획득한 의미 지식을 로봇 행동 토큰과 결합함으로써 미학습 객체와 언어 명령에 대한 일반화 성능을 향상시켰다. 이러한 연구들은 이미지와 언어 명령으로부터 직접 행동을 생성하는 엔드투엔드 제어 가능성을 제시했다는 점에서 중요한 의미를 가진다.

로봇 파운데이션 모델의 실용화를 위해서는 다양한 로봇 임바디먼트 간의 전이 가능성이 중요하다[6]. Open X-Embodiment 및 RT-X[7]는 여러 기관에서 수집한 다양한 로봇 데이터셋을 통합하여 cross-robot 정책 학습의 가능성을 보였다. Octo[8]는 이를 기반으로 새로운 센서 입력과 행동 공간에 대한 파인튜닝 가능성을 제시하였다. OpenVLA[9]는 대규모 로봇 시연과 VL(vision-language) 사전 학습을 결합하여 LoRA 및 양자화를 활용한 효율적 학습 가능성을 보였다. 또한 ACT/ALOHA[10] 계열 연구는 행동을 청크(chunk) 단위로 예측하여 장기 예측 구간(long horizon)에서의 누적 오차와 제어 불안정성을 완화하였고, Diffusion Policy[11]와 RDT[12]는 고차원 로봇 행동 분포와 궤적 생성을 다루는 방법을 제안하였다.

그러나 이러한 연구들은 대체로 대규모 데이터셋과 고성능 학습 인프라를 전제로 하거나 특정 벤치마크 환경에서 검증되었다. 소량의 로컬 데이터와 제한된 엣지 컴퓨팅 자원만으로 사전 학습 데이터에 포함되지 않은 로봇팔에 실시간 배포까지 엔드투엔드로 검증한 사례는 찾아보기 어렵다.

#### 1.2 The distinguishing features of this study

본 연구는 사전 학습 데이터에 포함되지 않은 6-DoF Magician E6 로봇팔을 대상으로 파이제로 VLA 모델을 로컬 환경에서 효율적으로 전이하는 것을 목표로 한다.

이를 위해 본 연구는 두 가지 측면에서 기존 연구와 차별화된다. 첫째, 계층 선택적 PEFT 방법을 제안하여, 소량의 로봇별 데이터만으로 파인튜닝을 수행하면서도 파이제로 모델이 사전 학습을 통해 획득한 시각-언어 학습 능력의 손상을 최소화하였다. 이는 대규모 데이터셋 없이도 새로운 임바디먼트에 안전하게 적응할 수 있는 방법을 제시함으로써, 데이터·자원 의존성이 높았던 기존 크로스 임바디먼트 연구의 한계를 보완한다.

둘째, ROS 2 기반 제어 구조와 액션 청킹(action chunking) 기반 실행 방식을 결합하였다. 추론 1회로 생

www.kci.go.kr

성한 16스텝 중 앞 8스텝을 실행한 뒤 재추론함으로써, 매 스텝마다 추론하는 방식에 비해 행동당 추론 횟수를 1/8로 줄였다. 기존 연구들은 주로 오프라인 벤치마크나 시뮬레이션 환경에서 정확도 중심으로 검증되었다. 이와 달리 본 연구는 엣지 컴퓨팅 환경에서의 실시간 물리 로봇 제어까지 포함하여 방법의 실용성을 검증했다는 점에서 구별된다.

이러한 두 가지 구성 요소를 통해, 본 연구는 제한된 로컬 데이터와 엣지 컴퓨팅 자원만으로도 사전 학습 데이터에 포함되지 않은 이종 로봇팔에 대해 VLA 기반 실시간 조작이 가능함을 실험적으로 검증하였다.

## III. The Proposed Scheme

본 장에서는 제안하는 계층 선택적 PEFT의 전체 구성을 설명한다. 1절에서는 파이제로 VLA 모델의 파인튜닝에 사용된 듀얼 카메라 기반 데이터셋의 수집 방식과 공간 표현 방법을 설명한다. 2절에서는 계층 선택적 PEFT의 구체적인 구성을 설명한다. 이는 각 모듈이 범용 지식을 담고 있는지, 혹은 특정 로봇에 고유한 정보를 다루는지에 따라 파인튜닝 범위를 차등적으로 적용하는 것을 원칙으로 한다. 구체적으로는 언어 모델 백본의 동결, 비전 인코더 상위 블록에 대한 계층 선택적 LoRA 적용 그리고 액션 전문가는 18개 전 계층에 LoRA를 적용하였다. 로봇 특화 입출력 층은 전체를 학습하였다. 3절에서는 이러한 학습 설정에 따른 손실 수렴 결과를 제시한다.

### 1. Dual camera-based dataset collection and spatial representation

#### 1.1 Environment settings for data collection

로봇팔의 궤적을 학습하기 위해 로봇의 작업 공간을 Fig. 1처럼 다각도에서 캡처할 수 있는 듀얼 카메라(Dual-camera)를 사용하였다. 이는 로봇팔 주위 환경에서 단일 시점(Single-view) 카메라 사용 시 발생할 수 있는 시야가 가려지는 문제를 해결하고 3차원 공간 기하학 정보를 포착하기 위해서이다. HIK 카메라는 작업 공간 전면을 조망하는 전역 카메라 역할로 대상 물체의 위치와 작업 공간의 기하 정보를 제공한다. ZED 카메라는 로봇팔 전체가 프레임에 포함되도록 배치하여, 작업 대상과 매니퓰레이터(manipulator)의 상대적 기하 관계 및 팔의 현재 자세를 관측하는 역할을 담당한다.

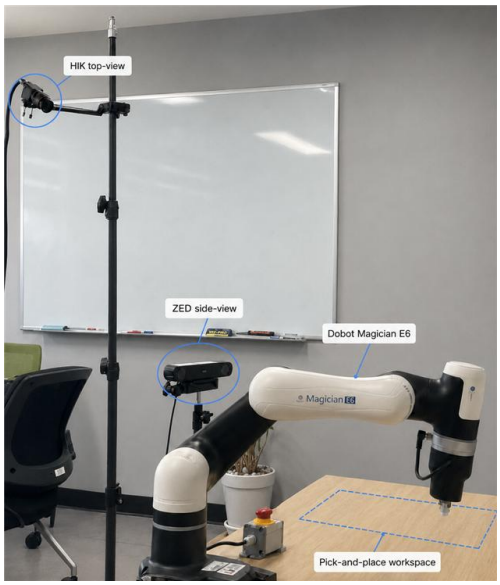


Fig. 1. Environment Setup for Data Collection

#### 1.2 Data collection

파이제로 VLA 모델에서 사용하지 않은 새로운 객체를 학습시키기 위해 각 400 프레임으로 구성된 총 180개의 에피소드(총 72,000 프레임)의 궤적 데이터를 수집하였다. 수집된 궤적 데이터는 LeRobot[13] 포맷 형식을 사용하여 로봇을 학습한다. 수집된 180개 에피소드는 궤적 데이터(Parquet), 카메라 영상(MP4), 메타데이터(JSON/JSONL)로 구성된다. 데이터의 각 프레임은 관측과 행동으로 구성된다. 관측은 현재 6개 관절각(deg), 2대 카메라 영상, 자연어 명령이며, 영상은 모델 입력 시 224×224로 변환된다. 행동은 7차원 벡터로, 앞 6개 차원은 로봇 베이스 좌표계 기준 엔드이펙터(위치 : X, Y, Z [mm], 방향: Rx, Ry, Rz [deg])이고, 마지막 1개 차원은 흡착 명령(0: 해제, 1: 흡착)이다. 즉 모델은 관절 공간의 상태를 입력받아 데카르트 좌표 공간의 행동을 출력하며, 행동의 포즈는 실행 시 역기구학을 통해 관절각으로 변환된다(5.2절 참조). LeRobot 데이터 형식은 Fig. 2와 같다.

| FIELD | DTYPE | SHAPE | 설 명 |
|---|---|---|---|
| exterior_image_1_left | image | [224, 224, 3] | HIK — 작업 공간 전면 (global view) |
| exterior_image_2_left | image | [224, 224, 3] | ZED — 로봇팔 포함 장면 (scene view) |
| state | float32 | [7] | 관절 상태 [j1…j6, gripper], 도 (°) |
| action | float32 | [7] | 관절 델타 6축 (deg/frame) + 그리퍼 절대값 (0/1) |
| task_index | int64 | [1] | 언어 조건 인덱스 → meta/tasks.jsonl의 10개 phase 프롬프트 |
| timestamp | float32 | [1] | 에피소드 내 경과 시간 (s) |
| frame_index | int64 | [1] | 에피소드 내 프레임 번호 |
| episode_index | int64 | [1] | 에피소드 번호 |
| index | int64 | [1] | 데이터셋 전역 프레임 번호 |
| next.done | bool | [1] | 에피소드 종료 플래그 |
| next.reward | float32 | [1] | 보상 (본 연구에서는 미사용) |

Fig. 2. LeRobot Format

www.kci.go.kr

## 2. Layer-selective LoRA fine-tuning of vision encoder high layer, and full-layer LoRA fine-tuning of Action Expert

본 연구의 파인튜닝 대상은 세 구성요소이며, 언어 모델 백본은 동결한다. 첫째, 비전 인코더의 상위 블록(18-26)에만 LoRA를 적용하는 계층 선택적(layer-selective) LoRA 파인튜닝이다. 둘째, 18층 트랜스포머로 구성된 액션 전문가의 전 계층에 LoRA를 적용한다. 셋째, 로봇 특화 입출력 층(상태 투영, 행동·시간 임베딩 MLP, 행동 출력 투영)은 전체 파라미터를 파인튜닝한다. 상태 투영과 행동·시간 임베딩 MLP는 액션 전문가가 받는 입력 토큰을 만들고, 행동 출력 투영은 액션 전문가의 출력을 속도장으로 변환한다. 세 구성요소는 하나의 플로우 매칭 손실 함수로 함께 최적화된다. 2.1절에서는 비전 인코더의 LoRA 구성을, 2.2절에서는 액션 전문가와 로봇 특화 입출력 층의 학습 방식 및 손실 함수를 설명한다.

### 2.1 Vision encoder high layers (18-26) layer-selective LoRA fine-tuning

파이제로 VLA 모델의 뛰어난 일반화 성능은 텍스트 명령과 시각적 특징을 고차원 공간에서 정렬하는 LLM(Large Language Model) 백본에 크게 기인하는 것으로 알려져 있다[14]. 이에 본 기법에서는 대규모 매개변수를 가진 LLM 백본을 완전히 동결(frozen)하여 연산 효율성을 확보하고, 사전 학습된 범용 지식을 보존하고자 하였다.

반면 새로운 로봇 임바디먼트인 Magician E6와 카메라 간의 상대적 정렬과 새로운 작업 공간의 고수준 의미론(high-level semantics)을 반영하기 위해서는 비전 인코더 내 일부 계층에 대한 조정이 필요하다. 비전 인코더의 하위 블록은 선, 면, 단순 색상과 같이 특정 로봇이나 환경에 종속되지 않는 저수준(low-level) 시각 특징을 추출하는 것으로 알려져 있으므로[15][16][17], 이 영역은 동결 상태를 유지하였다. 반면 객체의 형태적 정체성과 공간적 위치 관계를 통합적으로 처리하는 상위 계층은 새로운 작업 공간에 맞춰 조정될 필요가 있다고 판단하였다. 그래서 본 연구에서는 비전 인코더의 18번부터 26번 계층(이하 High Layer)을 선택적으로 지정하고, LoRA 랭크를 16으로 설정하여 LoRA 부분 파인튜닝을 수행하였다. LoRA 랭크 16과 계층 범위는 아래 두 실험을 근거로 결정하였다. LoRA 랭크 설정 실험은 오렌지 박스를 pick하는 실험으로 설정하였다. 실험은 랭크별로 10회씩 수행하였고, 수행 시간의 단위는 초이다.

Table 1. Test Results on LoRA Rank Settings

| LoRA Rank | Task | Time |
|---|---|---|
| 8 | Fail | - |
| 16 | Success | 28.1s |
| 32 | Success | 25.7s |

실험 결과, 랭크 8은 Table 1과 같이 목표물에 도달하지 못했다. 실험 10회 모두 목표물의 1~3 cm 부근에서 계속해서 맴돌면서 수행 시간이 2분을 넘겨서 동작을 중지시켰다. 랭크 32는 랭크 16보다 수행 완료 시간이 평균 약 2.4초 정도 빨랐다. 하지만 랭크 32는 랭크 16 대비 학습 파라미터가 약 2배이나 목표 도달에서 뚜렷한 차이가 없어, 충분한 성능을 보이는 랭크 16을 선택하였다. 랭크가 높을수록 소량 데이터에서 과적합 위험이 커질 수 있으나, 본 연구는 랭크별 과적합을 직접 평가하지는 않았다.

계층 선택적 PEFT의 성능을 측정하기 위해 pick-and-place 태스크를 각 구성별로 20번씩 수행하였다. 태스크 수행 완료 시간의 단위는 초이다. Table 2에서 액션 전문가의 전 계층 LoRA 파인튜닝은 A로, 로봇 특화 입출력 층의 전체 파인튜닝을 B로 표기하였다.

Table 2. Performance Comparison Test Results for Layer-Selective PEFT

| Composition | Task | Working Time |
|---|---|---|
| Vision Encoder frozen + A + B | Failed | - |
| Vision Encoder Full FT + A + B | Failed | - |
| Vision Encoder Full Layer(0-26) LoRA + A + B | Success | 65.78s ± 10.83s |
| Vision Encoder Low Layer(0-8) LoRA + A + B | Success | 125.39s ± 28.71s |
| Vision Encoder Middle Layer(9-17) LoRA + A + B | Success | 146.17s ± 29.13s |
| Vision Encoder High Layer(18-26) LoRA + A + B | Success | 65.16s ± 5.10s |

처음 실험에서는 액션 전문가의 전체 파인튜닝을 수행하였는데, 테스트 결과 여러 가지 오류가 관찰되었다. 특히 파이제로 모델에서 사전 학습한 계란, 아몬드와 충전기 같은 seen 객체에 대해서 치명적 망각 현상도 발견되었다. 이러한 이유로 액션 전문가의 전체 파인튜닝 설정 테스트는 더 이상 수행하지 않았다.

Table 2 결과처럼 비전 인코더와 액션 전문가에 LoRA

www.kci.go.kr

파인튜닝을 적용한 구성에서는 치명적 망각 현상이나 테스트 실패는 발견되지 않았다. 다만 태스크 수행 완료 시간에 큰 차이가 났다. 비전 인코더의 전체 계층 또는 상위 계층에 LoRA를 적용했을 때, 태스크 수행 완료 속도가 하위 계층에 비해 1.92배, 중위 계층에 비해 2.24배 빠르다. 모델 추론 시간은 2.4초로 모두 동일한데, 속도 차이가 나는 이유를 분석해 보았다.

Table 3. Number of model inferences and output change for each layer

| Layers | Number of Inference | Output Change |
|---|---|---|
| Low (0-8) | 52.3 | 31.3% |
| Middle (9-17) | 61.2 | 31.0% |
| High (18-26) | 27.5 | 53.0% |
| Full (0-26) | 27.4 | 50.7% |

속도 차이는 Table 3에서 보듯이 하위와 중위 계층에서 모델 추론을 2배 이상 많이 수행한 결과임을 볼 수 있다. 모델 추론에서 생성한 궤적이 정확하지 않으므로 목표물에 도착하기까지 추론을 더 많이 수행하여 속도가 느려졌음을 알 수 있다. 추론 결과의 성능이 계층마다 다른 이유는 출력 변화에서 볼 수 있다. 출력 변화는 한 이미지를 비전 인코더 27층을 통과시킨 후 계산한 다음에 바뀐 양의 크기를 잰 것이다. 출력 변화는 같은 크기의 가중치 변화가 뒤쪽 층(상위 계층)에서 훨씬 큰 변화를 만드는 레버리지 차이를 보여준다. 하위 및 중위 계층의 출력 변화는 31%로 아웃풋에 가까운 상위 계층의 53%에 비해 현저히 낮음을 볼 수 있다. 이러한 결과는 계층 구성에 따라 로봇의 목표물 접근 효율성이 달라지며, 특히 하위 및 중위 계층 구성에서 접근 성능이 저하되었음을 보여준다. 이러한 차이는 하위 및 중위 계층의 파인튜닝이 뒤쪽 층인 상위 계층을 통과하면서 희석되어 행동 예측과 궤적 생성의 효율성을 저하시켰을 가능성을 시사한다. 다만 시각 표현과 행동 예측의 변화를 직접 분석하지 않았으므로, 이러한 메커니즘의 인과관계는 추가 검증이 필요하다.

결론적으로 두 가지 테스트 결과를 근거로 본 논문에서는 가장 효율적인 LoRA 랭크 16 설정과 계층 선택적 LoRA 파인튜닝을 제안한다.

### 2.2 Training of the Action Expert and robot-specific input/output layers

파이제로의 액션 전문가는 Gemma 2B와 같은 구조에서 폭(1024)과 MLP 차원(4096)을 줄인 약 300M 규모의 18층 트랜스포머이다[1]. 파이제로는 VLM 백본과 액션 전문가가 서로 다른 가중치를 가지며 self-attention 층에서만 상호작용하는 구조이다. 이미지·언어 토큰은 VLM이 처리하고, 로봇 특화 토큰(로봇 상태와 행동)은 액션 전문가가 처리한다. 로봇 특화 토큰은 로봇 특화 입출력 층을 거친다. 상태 투영은 로봇 상태를 상태 토큰 1개로 변환하고, 행동·시간 임베딩 MLP는 노이즈가 섞인 행동 청크(16스텝)에 시간 정보를 융합하여 행동 토큰 16개로 변환한다. 액션 전문가는 이 토큰들을 처리하면서 매 층에서 VLM이 처리한 이미지·언어 토큰의 키·값을 참조하고, 행동 토큰 위치의 출력은 행동 출력 투영을 거쳐 벡터장(vector field)이 된다. 추론 시에는 이 벡터장을 수 회 적분하여 연속 행동 청크를 생성한다.

본 연구에서 학습하는 대상은 두 부분이다. 첫째, 액션 전문가는 18개 계층 전체의 질의(Query), 키(Key), 값(Value), 출력(Output) 투영과 MLP에 랭크 16의 LoRA 어댑터를 부착하고 원본 가중치는 동결하였다(11,059,200개, 액션 전문가의 약 3.7%). 사전 학습으로 습득한 행동 생성 능력을 보존하면서 적은 파라미터로 새로운 임바디먼트에 적응하기 위해 LoRA를 선택하였다. 둘째, 로봇 특화 입출력 층은 전체 파라미터를 학습하였다(총 3,248,160개). 이 층은 상태 투영(33,792개), 행동·시간 임베딩 MLP(3,181,568개)과 행동 출력 투영(32,800개)으로 이루어져 있다. 전체 파라미터를 학습한 이유는 벡터의 차원별 의미(상태는 관절각, 행동은 엔드이펙터의 위치·방향 성분, 그리고 단위와 정규화)가 로봇마다 달라 파이제로 VLA의 사전 학습 가중치를 그대로 쓰기 어려운 층이기 때문이다. 이 논문에서 훈련시킨 전체 파라미터 수는 Table 4와 같이 14,970,912개로 3.37B의 0.444% 수준이다.

Table 4. Configuration of Training Parameters

| Components | Learning Method | No. of Params | Percent (Total 3.37B) |
|---|---|---|---|
| Vision Encoder Layer(18-26) | LoRA r=16 | 663,552 | 0.020% |
| Action Expert Full Layer | LoRA r=16 | 11,059,200 | 0.328% |
| Robot-spec I/O Layers | Full Fine-tuning | 3,248,160 | 0.096% |
| Total | | 14,970,912 | 0.444% |

학습은 플로우 매칭으로 이루어진다. 행동을 이산 토큰이 아닌 연속 벡터로 다루며, 학습 분할(144개 에피소드)의 정답 행동 A와 가우시안 노이즈 ε를 선형 보간한 $x_t = tA + (1-t)\varepsilon$를 입력으로 하여 정답 벡터장 $A - \varepsilon$를 예측

www.kci.go.kr

하도록 학습한다(t=0은 노이즈, t=1은 정답 행동). 손실은 예측 벡터장과 정답 벡터장의 L2 오차이며, 비전 인코더 LoRA, 액션 전문가 LoRA, 로봇 특화 입출력 층은 이 하나의 손실로 함께 최적화된다. 손실의 그래디언트는 어텐션을 통해 비전 인코더 LoRA까지 전달된다. 추론 시에는 VLM 토큰의 키·값을 한 번 계산해 캐시하고 적분 스텝마다 재사용한다.

### 2.3 Loss function optimization

비전 인코더(LoRA), 액션 전문가(LoRA), 로봇 특화 입출력 층(Full)은 개별적으로 최적화되지 않고, 단일 플로우 매칭 손실 함수 하나로 함께 최적화된다. 구체적으로는 액션 전문가가 예측한 벡터장과 수학적으로 정의된 정답 벡터장 간의 차이를 L2 손실(loss)로 계산한다. 계산된 오차를 최소화하는 방향으로 비전 인코더의 LoRA 어댑터, 액션 전문가의 LoRA 어댑터, 로봇 특화 입출력 층의 파라미터가 동시에 갱신된다. 따라서 학습 과정에서 세 구성 요소를 구분하는 개별 손실은 존재하지 않으며, 관측되는 학습 손실은 세 요소가 공동으로 최적화된 단일 스칼라 값이다.

이때 단일 타임스텝이 아니라, 미래의 연속된 행동 구간을 하나의 단위로 묶은 액션 청크(Action Chunk, 16개 타임스텝)를 통째로 예측하도록 학습한다. 이는 매 타임스텝마다 추론하는 방식에 비해 행동당 추론 횟수를 줄이는 동시에, 청크 내 행동 간의 시간적 일관성을 확보하여 궤적을 보다 부드럽고 안정적으로 생성하기 위함이다.

### 2.4 Vector field inference with Action Expert

학습 시 액션 전문가는 정답 행동으로부터 얻어진 노이즈-행동 쌍을 이용해 벡터장을 학습한다. 하지만 실제 추론(inference) 시점에는 정답 행동을 알 수 없으므로 순수한 가우시안 노이즈에서 출발하여 반복적으로 행동을 생성해야 한다. 이를 위해 액션 전문가는 임의의 시간 $t$에서의 노이즈가 섞인 행동 후보와 VLM으로부터 전달된 시각-언어 문맥 정보 그리고 로봇의 현재 관절 상태(proprioception)를 함께 입력받는다. 액션 전문가는 이 입력을 바탕으로 "현재 상태에서 정답 행동에 도달하려면 어느 방향과 속도(velocity)로 이동해야 하는가"에 해당하는 벡터장을 추론한다. 이 벡터장을 따라 노이즈 상태를 t = 0 에서 t = 1까지 여러 스텝에 걸쳐 점진적으로 갱신함으로써, 최종적으로 실행 가능한 7차원 액션 청크가 생성된다.

## 3. Training Result

파이제로 VLA 모델은 연속적인 7차원 행동 값을 생성하므로, 이산적 분류 정확도 대신 손실(loss) 그래프를 통해 학습 결과를 평가하였다.

본 실험의 학습 설정은 다음과 같다. 비전 인코더는 SigLIP의 상위 블록 18번부터 26번 블록만 LoRA로 파인튜닝하였고, 언어 모델 백본과 비전 인코더 하위 블록은 동결하였다. LoRA 랭크는 16으로 설정하였다. 학습 데이터는 앞서 구축한 180개 에피소드, 총 72,000프레임의 시각-행동 데이터 셋을 사용하였다. 180개의 에피소드 중에서 training으로 144개 에피소드, 57,600 프레임을 사용하였다. validation용으로는 18개 에피소드, 7,200 프레임을 사용하였다. 나머지 18개 에피소드, 7,200 프레임은 Test를 위해 남겨 두었다. 1 에폭(epoch)은 4,250 스텝이고, RTX A5000 단일 GPU 환경에서 20,000 스텝 동안 학습하였다. 손실 함수는 플로우 매칭 목적함수에 따라 예측 벡터장과 정답 벡터장 사이의 L2 오차를 최소화하도록 정의하였다. 최적화에는 AdamW를 사용하였고, 학습 초기의 불안정성을 줄이기 위해 학습률 워밍업(learning rate warm-up)과 그래디언트 클리핑(gradient clipping)을 적용하였다.

학습 결과, 모델은 과적합(overfitting) 징후 없이 안정적으로 수렴하였다. 손실 그래프를 보면 4.7 에폭에서 Train loss는 0.017, Validation loss는 0.014이다.

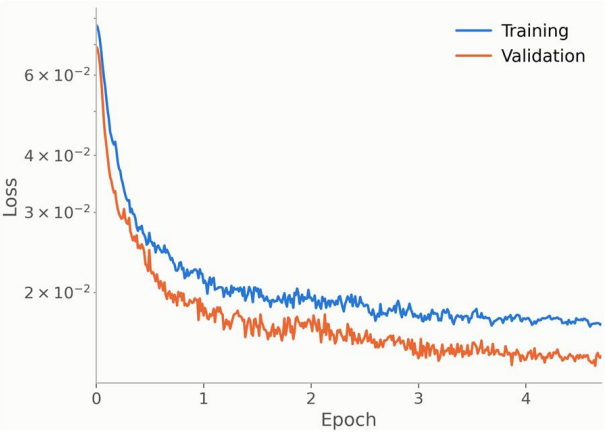


Fig. 3. Flow-Matching Training Loss of the Proposed PEFT Configuration

두 손실 값 사이의 격차가 매우 작은 점은 본 모델이 미학습 데이터에 대해서도 우수한 일반화 성능을 확보하였음을 시사한다. 모델의 손실은 학습 초기 0.087에서 급격히 감소하여 약 2,000 스텝 이후 완만한 하강 구간으로 진

www.kci.go.kr

입하였다. 그리고 10,000 스텝 이후에는 Train loss는 0.017, Validation loss는 0.014 수준에서 평탄화되었다. 전 구간에서 발산이나 진동 없이 안정적으로 감소하였다. 학습 손실과 함께 학습에 사용되지 않은 데이터에서 산출한 검증 손실(validation loss)을 모니터링하여 학습 과정의 수렴과 과적합 여부를 확인하였다.

다만 플로우 매칭 손실은 벡터장의 L2 오차이므로 최종 생성된 궤적의 정밀도를 직접 나타내지는 않는다. 이에 학습에 사용되지 않은 held-out 프레임에 대해 생성된 궤적을 정답과 비교하여 ADE(Average Displacement Error)와 FDE(Final Displacement Error)를 측정하였다(4절 참조).

## 4. Trajectory Prediction Accuracy Evaluation

최종 모델의 궤적 예측 정밀도를 측정하기 위해, 테스트용 18개의 에피소드에서 held-out 200개 프레임에 대해 ADE와 FDE를 산출하여 관절별로 분석하였다.

### 4.1 Trajectory prediction accuracy evaluation of the layer-selective PEFT model

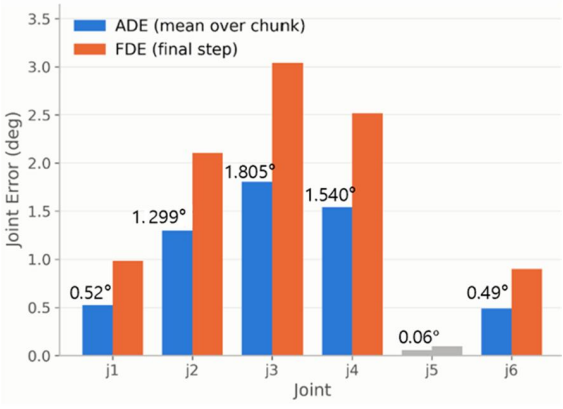


Fig. 4. Per-Joint ADE/FDE on 200 Held-out Frames.

Fig. 4는 학습에 사용되지 않은 200개 프레임에 대해 측정한 관절별 도착 오차를 나타낸다. 관절별 분석을 위해 예측 포즈를 MoveIt 2의 역기구학 솔버로 관절각으로 변환하여 비교하였다. 전체 평균 ADE는 0.952°, FDE는 1.607°이다. 관절 공간 오차는 엔드이펙터 포즈 오차가 역기구학을 통해 각 관절에 배분된 결과로 j2(1.299°), j3(1.805°), j4(1.540°)에 집중되었다. 이 세 관절은 로봇 팔을 수직 평면에서 펴고 구부려 엔드이펙터(End-Effector)의 위치 변화를 대부분 결정하는 축이다. 베이스 회전축인 j1(0.52°)과 손목 회전축인 j6(0.49°)의 오차는 상대적으로 작았다. j5의 오차는 0.06°로 가장 작았다. 가장 작은 이유는 Magician E6에서 j5가 약 −88°에 고정되어 정답 궤적의 변화량 자체가 작기 때문이며, 예측 정확도가 높음을 의미하지는 않는다.

엔드이펙터 위치 오차는 로봇의 반복 정밀도(±0.1 mm)보다 크므로 오차는 하드웨어가 아니라 모델 예측에서 비롯된다. 측정 결과 EEF(End-Effector) ADE는 8.84 mm, EEF FDE는 15.10 mm였다. 관절 공간에서의 ADE 평균이 0.952° 수준이던 오차가 말단에서는 8.84 mm의 위치 오차로 나타남을 확인하였다. 다만 이 값은 시연 궤적 대비 오차이며, 호출마다 시각 입력으로 보정되는 폐루프 제어의 최종 도달 오차와는 구별된다. 따라서 본 평가는 held-out 프레임에서 모델이 시연 궤적을 재현함을 보이는 것이며, 작업 수준의 성능은 IV장의 실로봇 시험에서 평가한다.

## 5. ROS-based Real-time Cross-Embodiment Trajectory Control Pipeline

### 5.1 Hardware configuration for operating a robot arm

실험 시스템은 엣지 컴퓨팅 플랫폼인 NVIDIA Jetson AGX Orin과 제어 대상인 Dobot Magician E6 로봇팔로 구성된다. Fig. 5는 전체 하드웨어 구성과 각 모듈 간의 데이터 흐름을 나타낸다.

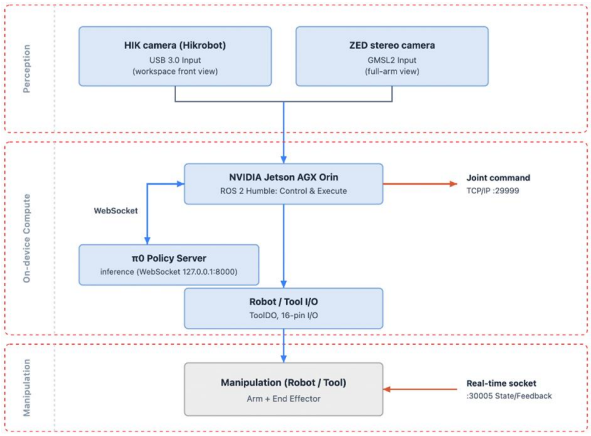


Fig. 5. Hardware Configuration and Data Flow for Robot Arm Control

① 시각 입력 모듈: Jetson AGX Orin에는 두 대의 카메라가 USB 3.0 인터페이스를 통해 연결된다. 하나는 HIK 카메라로 작업 공간 전면을 조망하고, 다른 하나는 ZED 스테레오 카메라로 로봇팔 전체가 프레임에 포함되도록 하여 팔의 기하 관계 및 현재 자세를 관측하는 역할을 담

www.kci.go.kr

당한다.

② 추론 및 제어 소프트웨어 스택: Jetson AGX Orin에는 ROS 2[18] Humble이 설치되어 있다. 이 스택은 상태 관측(state observation), 안전 감독(safety supervisor), 행동 디코딩(action decoder) 등의 기능을 수행한다. 이 스택은 청크 실행 중 약 62.5 ms(16 Hz) 간격으로 관절 명령을 전송한다. 파이제로 정책 서버는 1회 추론에 약 2.4초가 소요되며, 생성한 16스텝 청크 중 앞 8스텝을 실행한 뒤 다시 추론한다. ROS 2 스택과 정책 서버 간의 통신은 로컬 웹소켓(WebSocket) (127.0.0.1:8000)을 통해 이루어진다.

③ 로봇팔과의 통신: Jetson AGX Orin과 Dobot Magician E6는 이더넷(TCP/IP) 기반으로 연결된다. IP 주소 192.168.6.1과 포트 29999를 통해 제어 명령이 전송된다. Magician E6는 온보드 컨트롤러(on-board controller)를 내장하고 있으며, 16핀 I/O 포트를 통해 외부 장치와 연동된다. 엔드이펙터로는 흡착 방식의 진공 그리퍼(vacuum gripper)가 장착되어 있다. 로봇팔로부터의 상태 피드백은 실시간 소켓(real-time socket) 통신을 통해 ROS 2 스택으로 전달되어 폐루프(closed-loop) 제어를 구성한다.

### 5.2 Real-time trajectory generation and robot control pipeline

VLA 모델의 VLM은 듀얼 카메라 영상과 자연어 명령(예: “Pick the orange box”)을 처리하고, 액션 전문가는 매 층에서 VLM의 키·값을 참조하며 현재 관절 상태를 함께 반영하여 16스텝 행동 청크(16×7)를 생성한다. 각 행동은 로봇 베이스 좌표계 기준 엔드이펙터 포즈 6차원과 흡착 명령 1차원으로 구성된다. 엔드이펙터 포즈는 MoveIt 2의 역기구학을 통해 관절각으로 변환된 뒤 ServoJ 명령으로 로봇에 전달된다. 전체 파이프라인은 Fig. 6과 같으며, 단계별 동작은 ①~⑤와 같다. Magician E6 제어에는 Dobot이 공식 제공하는 ROS 2 패키지를 사용하였다.

① 행동 청크 수신: ROS 2 노드가 16스텝 행동 청크를 수신하여 앞 8스텝을 실행 대상으로 선택하고, 각 스텝을 엔드이펙터 포즈 6차원과 흡착 명령으로 분리한다. 흡착 명령은 이진화하여 별도의 그리퍼 제어 토픽으로 발행한다.

② 역기구학(IK): 각 목표 포즈를 MoveIt 2의 역기구학 솔버(Inverse Kinematics Solver, 역기구학 솔버)에 전달한다. 솔버는 Magician E6의 URDF 모델을 기반으로 6개 관절각을 계산하며, 현재 관절각을 초기값으로 사용하여 직전 자세와 연속적인 해를 선택한다.

③ 안전 확인: 안전 감독 노드가 관절 가동 범위와 스텝당 관절각 변화 한계를 확인한다.

④ ServoJ 전송: dobot_bringup_v4가 관절각을 ServoJ 명령으로 변환하여 약 62.5 ms 간격으로 이더넷(TCP/IP)을 통해 Magician E6 컨트롤러에 전송한다. 흡착 신호는 ToolDO로 전송된다.

⑤ 로봇 구동: Magician E6 컨트롤러는 수신한 목표 관절각에 맞춰 6개 축의 서보 모터를 구동한다.

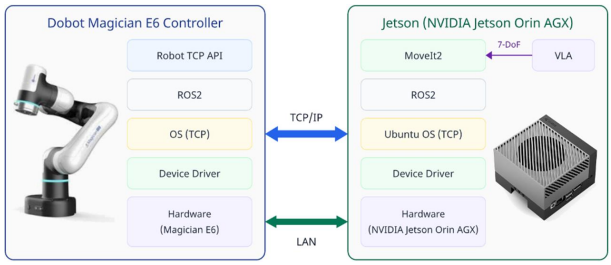


Fig. 6. Robot Control Pipeline

### 5.3 Runtime action sequence

파이제로 VLA 모델을 학습하고, 그 결과로 얻은 학습된 정책 체크포인트인 v23.pt 파일을 Jetson AGX Orin 보드에 배포한다. 런타임 시에 동작 순서는 Fig. 7과 같다. 런타임 시에 파이제로 정책 서버가 v23.pt 파일을 로드하고, 로컬 웹소켓(WebSocket) 프로토콜을 기반으로 127.0.0.1:8000 엔드포인트를 개방하고 대기한다. ROS 2 시스템과는 상시 양방향 채널을 확립하여 약 2.4초마다 VLA 모델의 행동 청크를 생성한다.

Camera-State Acquisition Node : 듀얼 카메라의 실시간 이미지 스트림과 dobot_bringup_v4가 읽어온 Magician E6의 현재 6축 관절 값(/joint_states)을 실시간으로 동기화하여 토픽으로 발행한다.

Task-Phase Node: 현재 로봇이 도달한 공정 단계를 트리거하여 VLA 모델에 입력될 텍스트 지시(예: "Pick up the blue cube") 토픽을 발행한다.

Policy Inference Bridge Node: 발행된 [듀얼 카메라 영상 + 현재 관절 상태] 토픽과 [텍스트 지시] 토픽을 실시간으로 구독(Subscribe)한다. 구독한 데이터들을 VLA 모델이 요구하는 텐서(Tensor) 형식의 단일 입력 구조체로 조립(assemble)한다. 조립된 입력 데이터를 파이제로 정책 서버로 전송한다.

파이제로 정책 서버는 수신된 이미지, 텍스트와 로봇 상태를 VLA 모델의 입력으로 넣는다. VLA 모델은 16스텝 행동 청크(16×7)를 생성하고, 파이제로 정책 서버는 이를 Bridge 노드로 반환한다. 청크를 수신한 Policy

www.kci.go.kr

Inference Bridge Node는 이를 ROS 2 커스텀 메시지 형태로 패키징하여 시스템 전체에 발행한다.

MoveIt 2는 실행 대상 8스텝의 엔드이펙터 포즈를 역기구학으로 관절각으로 변환하고, dobot_bringup_v4는 이를 ServoJ 명령으로 변환하여 이더넷을 통해 Magician E6 컨트롤러로 전송함으로써 모터를 구동한다.

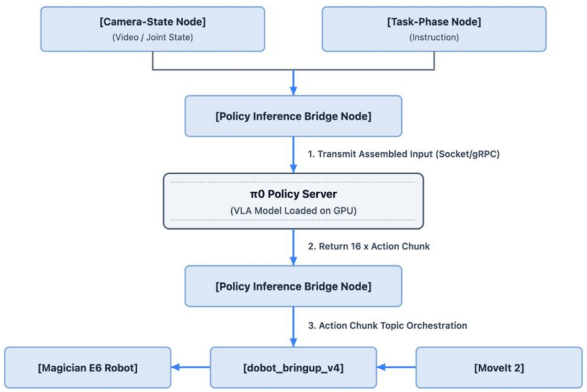


Fig. 7. Runtime Action Sequence

## IV. Experimental Results

본 논문에서 제안된 방법이 VLM의 일반적인 능력을 유지하면서 태스크를 연속적으로 수행하는지를 확인하기 위해 테스트를 진행하였다. 학습 모델을 테스트하기 위한 구성은 다음과 같다.

### 1. Object recognition test

본 실험에서는 제안된 계층 선택적 PEFT 방법이 사전 학습된 시각-의미론적 분류 능력을 유지하면서, 동시에 파이제로 학습 데이터에 없던 새로운 객체에 대해서도 정확한 인식 및 위치 추정이 가능한지를 검증하고자 하였다. 이를 위해 Fig. 8과 같이 형태적 특성이 서로 뚜렷하게 구분되는 6종의 테스트 객체를 선정하였다. 계란, 아몬드, 충전기는 파이제로의 로봇 사전 학습 데이터에 포함된 객체로 이후로는 seen 객체라 한다. 소형 박스(오렌지색 상자)는 파이제로 로봇 사전 학습 데이터에 없으며, 본 연구의 파인튜닝 데이터에 새로 추가한 객체이다. 이 객체를 이후로는 신규 객체(novel object)라 한다. 배터리와 칫솔은 사전 학습 데이터와 파인튜닝을 위한 신규 데이터에도 포함되지 않은 객체들로 이후로 unseen 객체라 한다.



Fig. 8. Objects Used in the Test

본 테스트는 객체 인식 및 엔드이펙터의 위치 정렬을 평가 시에 물체를 들어 올리는(lift) 과정 없이 로봇팔의 흡착 그리퍼가 목표 객체 표면에 조금이라도 접촉했을 때 성공으로 판정하였다.

#### 1.1 Single target test results

각 물체를 단일 목표(single target)로 책상 위에 배치하고, 각 물체당 20회씩 반복 시행하였다. Table 5에 나타난 바와 같이, seen 객체(계란, 아몬드, 충전기)뿐 아니라 신규 객체인 소형 박스에 대해서도 20회 시행하여 100%의 성공률을 달성하였다.

Table 5. Test Result of Object Recognition

| Target Object | Experimental Condition | Trials | Success Rate (%) |
|---|---|---|---|
| Egg | Single Target | 20 | 100 |
| Almond | Single Target | 20 | 100 |
| Charger | Single Target | 20 | 100 |
| Small Box | Single Target | 20 | 100 |

Table 5의 결과는 두 가지 측면에서 의미를 가진다. 첫째, 형태가 상이한 다양한 객체에 대해 모두 정확한 인식과 위치 정렬이 가능하였다. 둘째, 신규 객체인 소형 박스에 대해서도 20회 시행에서 동일한 성공률을 보였다. 이 점은 본 연구가 제안한 파인튜닝 방법이 신규 객체에 대한 인식 및 위치 추정 학습 능력을 갖추었음을 시사한다. 다만 단일 목표 시험은 객체에 대한 인식 및 위치 추정의 능력이 있음을 보이지만, 객체에 대한 의미론적 지식을 가지고 있다는 증거가 되지는 않는다. 왜냐하면 책상에 물체가 하나뿐이면, 무엇이든 그 물체로 가는 정책도 성공하기 때문이다.

#### 1.2 Multi-target test results

본 시험은 여러 객체를 책상 위에 함께 배치한 다중 목표 환경에서, 언어 명령으로 지정한 객체를 정확히 선택하

www.kci.go.kr

는지를 평가한다. 비교는 두 가지로 구성하였다. 첫째, seen 객체끼리 함께 놓은 장면에서 위치와 목표를 바꾸어 가며 seen 객체 간 구별 능력을 확인하였다. 둘째, seen 객체와 신규 객체를 함께 놓은 장면에서, 신규 데이터로 학습한 이후에도 seen 객체를 지시대로 선택하는지 확인하였다. 이는 계층 선택적 파인튜닝 학습 이후 seen 객체의 망각 여부를 점검하고, 신규 객체의 성공적인 지식 획득을 보기 위함이다. 각 조건(장면과 목표의 조합)은 각각 20회 시행하였다.

Table 6. Test Result of Multi Object

| Scene Configuration | Target Object | Success Rate (%) |
|---|---|---|
| Egg + Almond | Egg | 100 |
| Egg + Almond | Almond | 5 |
| Egg+Almond+Charger | Egg | 100 |
| Egg+Almond+Charger | Charger | 100 |
| Egg+Small Box+Charger | Egg | 100 |
| Egg+Small Box+Charger | Charger | 100 |
| Egg+Small Box+Charger | Small Box | 100 |

Table 6의 결과와 같이 seen 객체인 계란과 아몬드 두 물체만 놓인 환경에서 아몬드를 목표로 지시한 경우 성공률이 5%에 그쳐, 단일 목표 테스트에 비해 크게 저하되는 결과를 보였다. 아몬드의 크기는 1.6×2.5cm로 모델 입력 해상도(224×224)에서 패치 크기(14×14화소) 1개 보다 작다. 단일 목표 시험에서는 아몬드를 100% 접촉하였으므로, 물체의 검출보다는 다른 물체와 함께 있을 때의 형태 식별이 입력 해상도의 제약을 받았을 가능성이 있다. 이것은 입력 해상도의 제약으로 작은 객체 식별 성능이 낮아질 수 있음을 말한다. 이 결과는 객체의 크기와 형태적 특징에 따라 다중 목표 환경에서의 인식 성능에 편차가 발생할 수 있음을 시사한다.

이 테스트의 목적은 계층 선택적 PEFT가 seen 객체에 대한 지식이 보존됨을 보기 위함이므로 성능 편차를 야기할 수 있는 아몬드 선택을 배제하고 테스트를 진행하였다. Egg + Almond + Charger 세 객체가 함께 놓인 환경에서 아몬드를 배제하고, 계란과 충전기를 목표로 지정한 경우 각각 20회 테스트 시에 모두 100%의 성공률을 보였다. 또한 seen 객체인 계란과 충전기, 그리고 신규 객체인 소형 박스를 한 장면에 함께 배치하고 각 객체를 목표로 지시하였다. 그 결과 지시한 객체 각각 20회, 총 60회 모두 성공적으로 지시한 객체를 선택하였다.

이러한 결과는 본 연구가 제안한 계층 선택적 PEFT 방법이 기존 사전 학습된 객체에 대해 사전 학습 지식을 유지하면서, 신규 객체를 학습할 수 있음을 보여준다. 다만 이 테스트는 사전 학습된 객체에 대해 치명적 망각이 발생하지 않았음을 뒷받침할 수 있지만, 사전 학습 지식이 완전히 보존되었음을 입증하지는 못한다.

### 1.3 Unseen object recognition test

본 테스트는 seen 객체도 아니고, 신규 객체도 아닌 처음 보는 unseen 객체에 대한 테스트를 진행하였다. 테스트 목적은 첫 번째는 모르는 객체에 대해서 로봇팔이 어떻게 반응하는지를 보는 것이고, 두 번째는 모르는 객체가 있는 환경에서도 seen 객체를 지시대로 선택하여 치명적 망각이 발생하지 않는지를 보기 위함이다. 테스트는 지시한 목표 객체별로 20회씩 수행하였다.

Table 7. Test Result of Unseen Object

| Scene Configuration | Target Object | Success Rate (%) |
|---|---|---|
| Egg+Toothbrush+Battery | Toothbrush | check |
| Egg+Toothbrush+Battery | Battery | check |
| Egg+Toothbrush+Battery | Egg | 100 |

실험 결과는 칫솔과 배터리를 목표로 지시할 때, Table 7과 같이 로봇팔이 40회 모두 칫솔과 배터리 중간 지점을 가리켰다. 계란으로는 이동하지 않았으나 두 객체 중 하나를 고르지 못하고 두 객체의 중간으로 이동한 것으로 보인다. 이 결과로는 처음 보는 객체에 대해서 이름으로 식별하는 능력을 가지고 있음은 확인되지 않았다.

목표를 계란으로 지시했을 경우는 20회 모두 성공하였다. 파인튜닝 후에 seen 객체가 방해물(unseen 객체들)이 있는 환경에서도 지시대로 정확하게 선택하는 이 결과는 치명적 망각이 발생하지 않았음을 뒷받침한다.

다만 처음 보는 객체를 이름으로 식별하는 능력은 본 논문이 제안하는 계층 선택적 PEFT의 검증 대상이 아니다. 따라서 unseen 객체(칫솔, 배터리)를 목표로 한 40회는 전체 성공률 산정에서 제외하였고, unseen 객체를 방해물로 두고 seen 객체(계란)를 목표로 한 20회는 망각 여부를 점검하는 시험이므로 포함하였다. unseen 객체를 목표로 한 결과는 Table 7에 'check'로 표시하고 본 연구의 한계로 다룬다.

## 2. Color discrimination test

본 테스트는 제안된 계층 선택적 PEFT 방법이 형태적 단서 없이 순수하게 색상 속성만으로 목표 객체를 구별하고 실제로 파지(grasping)할 수 있는지를 검증하기 위해

www.kci.go.kr

수행되었다. 1절의 Object recognition test는 형태가 서로 다른 객체를 구분하는 능력을 검증하였다. 이와 달리 이 테스트는 Fig. 9와 같이 형태와 크기가 동일한 정육면체 4개(빨강, 파랑, 노랑, 초록)를 특정 영역 안에 함께 배치하여 오직 색상만이 객체를 구별하는 유일한 단서가 되도록 구성하였다.

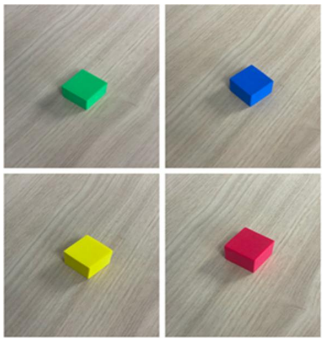

Fig. 9. Objects Used in Color Testing

성공 판정 기준 또한 접촉이 아닌 물체를 들어 올리는 실제 파지 성공 여부로 하였다. 사용된 네 가지 색상은 모두 사전 학습 데이터에 포함되어 있던 색상이다. 테스트는 언어 명령으로 특정 색상을 지정하면, 로봇팔이 해당 색상의 객체를 정확히 인식하여 파지에 성공하는 경우를 성공으로 간주하였다. 테스트는 색상별로 각각 20회씩 반복 시행하였다.

실험 결과는 Table 8에 나타난 바와 같이 네 가지 색상 모두에서 100% 성공률을 달성하였다. 이 결과는 사전 학습을 통해 습득한 색상 구별 능력이 새로운 임바디먼트 조건에서도 치명적 망각 없이 유지되었음을 보여준다. 특히 형태가 동일한 조건에서 순수하게 색상이라는 저수준 시각 속성만으로 목표 객체를 정확히 특정하고 파지까지 성공적으로 수행하였다. 이 점은 본 연구가 제안한 파인튜닝 방법이 사전 학습된 시각-의미론적 표현을 치명적 망각 없이 보존함을 뒷받침한다.

Table 8. Test Result of Color

| Target Object | Condition | Success Rate (%) |
|---|---|---|
| Red | Red Target | 100 |
| Blue | Blue Target | 100 |
| Yellow | Yellow Target | 100 |
| Green | Green Target | 100 |

### 3. Pick-and-place test

이 테스트는 인식, 파지, 이동, 배치(place)에 이르는 전체 조작 과정을 하나로 통합하여 실제 작업 수행 능력을 종합적으로 평가하기 위해 수행되었다. 일정 영역 안에 오렌지색 소형 박스를 무작위 위치에 배치한다. 이후에 로봇팔이 해당 객체를 인식하여 파지하고, 이를 지정된 목적지로 이동시켜 내려놓는 데 성공하면 성공으로 간주하였다.

총 20회의 시행 중 18회 성공하여 90%의 성공률을 기록하였다. 실패한 2회는 소형 박스를 집었다가 목적지로 옮기는 과정에서 원래 있던 자리로 돌아옴으로 인해 작업이 완료되지 못한 경우였다.

이 결과는 1절과 2절에서 개별적으로 확인한 인식 및 파지 능력이 이동과 배치까지 포함하는 완전한 pick-and-place 파이프라인으로 확장되었을 때에도 대체로 안정적으로 유지됨을 보여준다. 다만 단일 단계 테스트(1절 단일 목표 100%, 2절 100%)에 비해 성공률이 다소 낮아진 것은, 여러 단계(인식-파지-이동-배치)가 순차적으로 연쇄되는 과정에서 각 단계의 오차가 누적될 수 있음을 시사한다. 이는 4.1절에서 제시한 궤적 예측 오차(관절 공간에서의 ADE 평균이 0.952°, 엔드이펙터 기준 약 8.84 mm)가 단일 시점에서는 크지 않더라도, 파지에서 배치에 이르는 전체 궤적에 걸쳐 누적될 경우 최종 배치 위치의 정밀도에 영향을 줄 수 있음을 뒷받침하는 결과로 해석된다.

본 절을 포함하여 객체 인식, 색상 구별, pick-and-place 테스트를 수행하였다. seen 객체와 신규 객체의 시각-의미론적 분류 능력을 보는 객체 인식 테스트에서는 치명적 망각 현상 없이 92.1%의 성공률을 달성하였다. 객체 인식 테스트의 성공률 92.1%는 단일 목표 80회(Table 5), 다중 목표 140회(Table 6), unseen 객체를 방해물로 둔 환경에서 계란을 목표로 한 20회(Table 7)를 합한 240회 중 221회 성공으로 산정하였다. 색상을 분류하는 색상 구별 테스트에서도 치명적 망각 현상 없이 100%의 성공률을 달성하였다. pick-and-place는 90%의 성공률을 보였다. 이는 본 논문에서 제시하는 방법이 이기종 로봇팔 환경에서도 새로운 기하학적 형태적 특징을 매우 효과적으로 사상(Mapping)해내고 있음을 보인다.

## V. Conclusions

본 연구는 사전 학습된 파이제로 VLA 모델을 사전 학습 데이터에 포함되지 않은 6자유도 이종 로봇팔

www.kci.go.kr

(Magician E6)로 전이하기 위한 계층 선택적 PEFT 방법을 제안하고, 이를 엣지 컴퓨팅 환경의 실제 로봇에서 검증하였다. 본 연구의 기여는 다음과 같다.

첫째, 모듈별로 학습 범위를 차등화하는 계층 선택적 PEFT 방법을 제안하였다. 언어 모델 백본은 완전히 동결하고, 비전 인코더는 상위 블록(18-26)에만 LoRA로 부분 학습하였으며, 행동 생성을 담당하는 액션 전문가는 18개 전 계층에 LoRA로 부분 학습하였다. 로봇 특화 입출력 층은 전체를 학습하였다. 이는 범용 시각-언어 지식을 담은 부분은 최소한으로 조정하고 로봇 고유의 정보를 다루는 부분은 충분히 학습한다는 설계 원칙에 따른 것이다. 세 구성요소의 학습 파라미터는 전체의 약 0.44%이다.

둘째, 제안한 방법을 시뮬레이션이 아닌 실제 하드웨어에서 검증하였다. 학습된 모델을 NVIDIA Jetson AGX Orin 엣지 플랫폼에서 구동하고, ROS 2 기반 통신으로 Magician E6 로봇팔을 제어하였다.

셋째, 소량의 로봇별 데이터와 제한된 학습 파라미터만으로 새로운 로봇에 적응할 수 있음을 보였다. IV장의 인식·파지 시험 결과는 비전 인코더의 상위 블록만 조정하는 설계가 시각 구별 능력을 크게 해치지 않았음을 시사한다.

이러한 결과는 계층 선택적 PEFT 방법이 데이터와 연산 자원이 제한된 조건에서 사전 학습된 VLA 모델을 이종 로봇으로 전이하는 실용적인 접근이 될 수 있음을 시사한다. 대규모 재학습이나 고성능 인프라 없이 VLA 모델을 다른 로봇으로 확장할 수 있는 가능성을 보였다는 점에서 의미가 있다.

다만 본 연구에는 다음과 같은 한계가 있다. 검증은 단일 로봇과 단일 작업 환경에 한정되었고, 각 구성은 한 번의 학습 결과여서 시드 간 변동을 확인하지 못하였다. 또한 작은 객체(아몬드)는 다른 객체와 함께 놓였을 때 지시 대상으로 선택된 비율이 5%에 그쳤는데, 이는 입력 해상도(224×224)의 제약 때문일 가능성이 있다. 아울러 unseen 객체가 둘 이상 있을 때는 이름으로 식별하는 능력은 확인되지 않았다. 향후 다양한 로봇으로의 확장, 고해상도 입력, 시드 반복 실험으로 이러한 한계를 보완할 계획이다.

## REFERENCES


[1] K. Black et al., "π0: A Vision-Language-Action Flow Model for General Robot Control," in Proceedings of Robotics: Science and Systems, 2025, DOI: 10.15607/RSS.2025.XXI.010.

[2] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow Matching for Generative Modeling," in Proceedings of the International Conference on Learning Representations, 2023.

[3] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," in Proceedings of the International Conference on Learning Representations, 2022.

[4] A. Brohan et al., "RT-1: Robotics Transformer for Real-World Control at Scale," in Proceedings of Robotics: Science and Systems, 2023.

[5] B. Zitkovich et al., "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control," in Proceedings of the 7th Conference on Robot Learning, Proceedings of Machine Learning Research, Vol. 229, pp. 2165–2183, 2023.

[6] J. H. Yang, D. Sadigh, and C. Finn, "Polybot: Training One Policy Across Robots While Embracing Variability," in Proceedings of the 7th Conference on Robot Learning, Proceedings of Machine Learning Research, Vol. 229, pp. 2955–2974, 2023.

[7] Open X-Embodiment Collaboration et al., "Open X-Embodiment: Robotic Learning Datasets and RT-X Models," in Proceedings of the IEEE International Conference on Robotics and Automation, pp. 6892–6903, 2024, DOI: 10.1109/ICRA57147.2024.10611477.

[8] Octo Model Team et al., "Octo: An Open-Source Generalist Robot Policy," in Proceedings of Robotics: Science and Systems, 2024, DOI: 10.15607/RSS.2024.XX.090.

[9] M. J. Kim et al., "OpenVLA: An Open-Source Vision-Language-Action Model," in Proceedings of the 8th Conference on Robot Learning, Proceedings of Machine Learning Research, Vol. 270, pp. 2679–2713, 2025.

[10] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware," in Proceedings of Robotics: Science and Systems, 2023.

[11] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion Policy: Visuomotor Policy Learning via Action Diffusion," in Proceedings of Robotics: Science and Systems, 2023, DOI: 10.15607/RSS.2023.XIX.026.

[12] S. Liu et al., "RDT-1B: A Diffusion Foundation Model for Bimanual Manipulation," in Proceedings of the International Conference on Learning Representations, 2025.

[13] R. Cadene et al., "LeRobot: An Open-Source Library for End-to-End Robot Learning," in Proceedings of the International Conference on Learning Representations, 2026.

[14] A. Vaswani et al., "Attention Is All You Need," in Advances in Neural Information Processing Systems, Vol. 30, 2017.

[15] A. Dosovitskiy et al., "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," in Proceedings of the International Conference on Learning Representations, 2021.

[16] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid


www.kci.go.kr

Loss for Language Image Pre-Training," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11975-11986, 2023.

[17] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do Vision Transformers See Like Convolutional Neural Networks?," in Advances in Neural Information Processing Systems, Vol. 34, 2021.

[18] S. Macenski, T. Foote, B. Gerkey, C. Lalancette, and W. Woodall, "Robot Operating System 2: Design, Architecture, and Uses in the Wild," Science Robotics, Vol. 7, No. 66, eabm6074, 2022, DOI: 10.1126/scirobotics.abm6074.

[19] D. Coleman, I. A. Sucan, S. Chitta, and N. Correll, "Reducing the Barrier to Entry of Complex Robotic Software: a MoveIt! Case Study," Journal of Software Engineering for Robotics, Vol. 5, No. 1, pp. 3-16, 2014, DOI: 10.6092/JOSER_2014_05_01_p3.

## Authors



Yu-Been Ha is currently pursuing a B.S. degree with a double major in Software Engineering and Data Science at Kangnam University. He is expected to graduate in 2027.

Yu-Been Ha's research interests include robot learning, vision-language-action (VLA) models, and robotic manipulation. His recent work focuses on parameter-efficient fine-tuning and cross-embodiment adaptation of VLA models for real-world robot control, as well as the deployment of learned policies on embedded platforms using ROS 2.



In-Yeop Choi received the B.S. degree in Computer Science from Hanyang University in 2001 and the Integrated Master's-PhD degree in Computer Science from Korea University, Korea, in 2023.

Dr. Choi joined the faculty of the Department of Computer Science at Kangnam University, Gyeonggi-do, Korea, in 2023. He developed embedded systems at a company for 15 years. He is currently a Professor in the Department of Computer Science, Kangnam University. He is interested in Physical AI, Embodied AI, VLA(Vision Language Action), and Embedded System.

www.kci.go.kr