경량 시각-언어-행동 모델의 미세조정 과정에서 미래 이미지 예측 효과 분석

박태진, 김지원, 이재구*
국민대학교 컴퓨터공학과
*jaekoo@kookmin.ac.kr

요 약

최근 물리적 인공지능(Physical AI) 분야의 발전과 함께, 실제 환경에서 행동을 수행할 수 있는 로봇에 인공지능을 탑재하는 연구가 중요한 연구 방향으로 자리 잡고 있으며, 시각-언어 정보를 기반으로 다양한 로봇 과업을 수행하는 시각-언어-행동(Vision-Language-Action, VLA) 모델이 주목받고 있다. 이러한 모델에서 효율적인 구조를 위해 가벼운 모델을 사용하고, 계획 능력을 향상시키기 위해 미래 세계 지식을 예측하는 구조를 통합하려는 시도들이 증가하고 있다. 따라서 본 논문에서는 작은 크기의 시각-언어-행동 모델의 단기 학습 단계에서 미래 이미지 예측을 정책 학습 과정에 통합하였을 때 로봇 조작 과업에서 어떠한 영향을 미치는지를 분석한다. LIBERO 시뮬레이션 환경 실험 결과, 미래 이미지 예측은 공간적 관계 이해, 물체 중심적 조작, 목표 지향적 과업에서는 비교 모델 대비 평균 14.3% 성능 향상을 보였으나, 장기 조작이 과업에서는 오히려 성능 저하로 이어질 수 있음을 확인하였다. 이러한 결과는 미래 세계 지식 예측이 시각-언어-행동 모델에서 항상 효과적인 보조 학습 목표가 아님을 보여주며, 과업의 특성에 맞춘 통합이 필요함을 시사한다.

1. 서론

최근 로보틱스 분야에서는 시각-언어 모델(Vision Language Model, VLM)의 발전을 기반으로 시각-언어-행동(Vision-Language-Action, VLA) 모델[1-9] 연구가 활발히 진행되고 있다. VLA 는 VLM 이 보유한 대규모 지식을 활용하기 위해 VLM 을 백본(backbone)으로 설정하여 주어진 환경에서 적절한 행동을 예측하도록 학습된다.

이러한 VLA 아키텍처는 크게 두 가지로 구분된다. 가장 기본적인 구조는 그림 1. (a)와 같이 VLM 된다 [1,4,5]. 이 경우 입력은 VLM 과 같이 시각 이미지와 언어 지시를 입력으로 하며 필요에 따라 추가적인 환경 정보(로봇의 관절)를 활용한다. 출력으로는 자연어 문장 대신 로봇의 행동을 예측하도록 하여 VLA 를 학습시킨다. 다른 구조로는 그림 1. (b)와 같이 그림 1. (a)와 동일한 입력이지만, VLM 의 은닉 상태(Hidden State)를 정책(Policy)의 조건으로 활용하여, 정책이 시각-언어 정보를 바탕으로 행동을 예측하도록 한다 [2, 3, 6, 7].

특히 DreamVLA[7]와 VLA-adapter[3]는 효율적인 구조를 위해 작은 크기의 VLM 을 백본으로 사용하고, 학습 가능한 쿼리를 도입하여 이를 VLA 학습에 활용한다. DreamVLA[7]는 미래 예측 쿼리와 별도의 미래 예측 모델을 사용하여 행동 생성

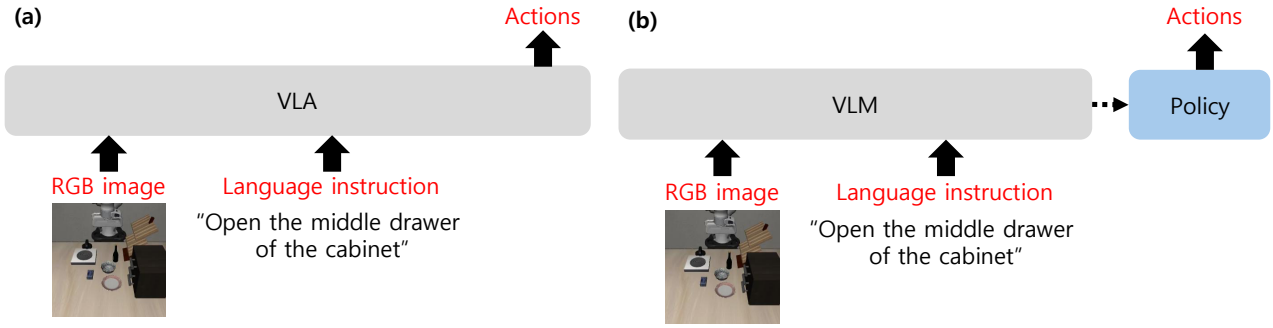

그림 1. 시각-언어-행동 모델 구조

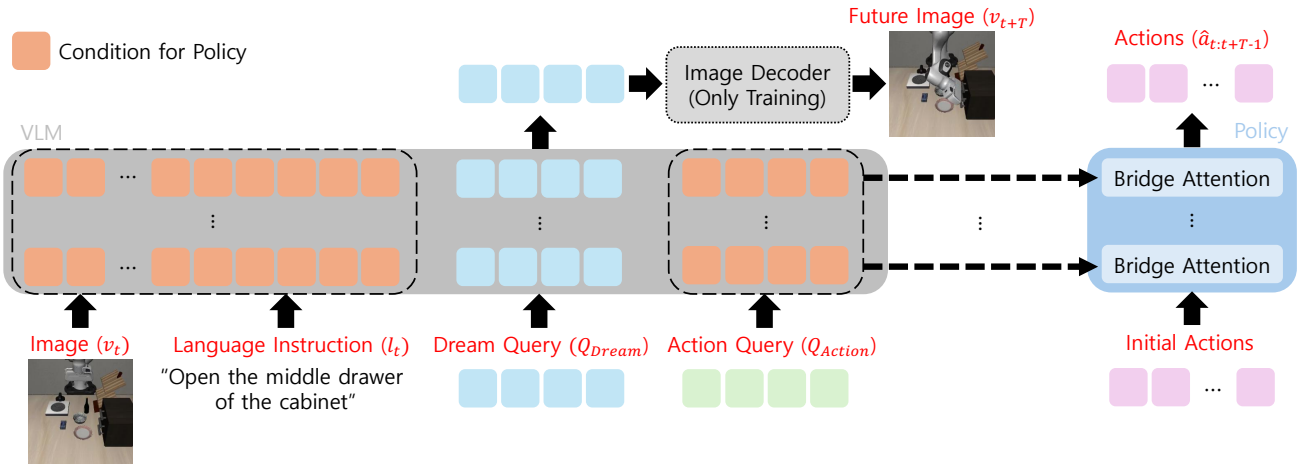

그림 3. 미래 이미지 예측 기반 VLA-adapter 구조

이전에 미래 세계의 지식(이미지, 동적 영역, 깊이, 의미론적 정보 등)을 예측한다. 미래 예측 이후에는 행동 쿼리를 정책에 활용하여 행동을 생성한다. 이 때, 미래 세계 지식 예측은 VLA 의 계획(Planning) 과정에 있어 유용한 단서를 제공하며 행동 예측 성능을 향상시킨다[7,8]. 반면 VLA-adapter[3]는 행동 쿼리를 사용하여 VLM 의 모든 레이어에서 추출된 은닉 상태를 정책 네트워크에 활용함으로써 성능을 향상시킨다.

그러나 DreamVLA[7]는 로봇 데이터집합 미세조정 과정 이전에 다양한 미래 상태 예측을 위한 별도의 모델들과 로봇 데이터집합에서의 사전학습 과정이 필요하다. 반면 VLA-adapter 는 기존 모델에서 로봇 데이터집합에서의 미세조정 과정만 수행하지만, 미래 예측 기능을 활용하지 않는다.

따라서 본 논문에서는 로봇 데이터집합을 이용한 미세조정 과정만을 수행하는 VLA-adapter 에서도 단기적 학습 단계에서 미래 지식 예측이 과업 성공률 향상에 기여하는지를 검증한다. 이를 위해 VLA-adapter 에 미래 예측 쿼리와 미래 이미지 예측 모듈을 추가로 설계하여 학습하고, 시뮬레이션 환경에서 그 효과를 평가한다.

2. 미래 이미지 예측 기반 경량 시각-언어-행동 모델

본 논문은 미래 세계 지식 예측이 VLA-adapter 의 계획 능력을 강화하여 성능 향상에 도움을 줄 수 있는지 확인하기 위해, VLA-adapter 에 행동 예측을 수행하기 전 DreamVLA 의 구조와 같이 미래 지식 예측을 위한 쿼리(Dream Query)와 디코더를 추가하였다. 미래 세계 지식 중 보편적으로 활용되어왔던[7,8,9] 이미지를 예측하도록 모델을 구성 하였으며, 제안 방법의 개요는 그림 2 와 같다.

2.1 미래 이미지 예측

미래 이미지 예측 과정은 현재 시점 t에서 환경

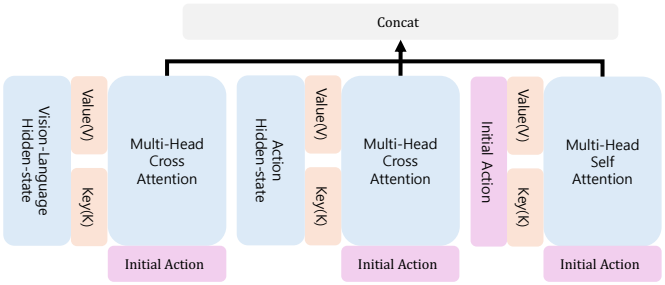

그림 2. Bridge Attention[3] 구조

정보인 이미지(Image, v_t)와 언어지시(Language Instruction, l_t), 학습가능한 미래 예측 쿼리(Dream Query, Q_{Dream}) 를 VLM 에 입력하고, Q_{Dream}에 해당하는 위치의 출력 값을 얻도록 진행된다. 다음으로 해당 출력 값을 이미지 디코더(Image Decoder)에 넣어 현재 시점 t를 기준으로 T 시점 이후의 미래 이미지(Future Image, v_{t+T})를 예측한다. 이 때, 손실 함수는 이미지 패치 단위의 교차 제곱 오차를 사용하며, 식 (1)과 같다.

$$L_{Recon} = \frac{1}{2}\sum_{n=1}^{N}\left(P_{t+T,n} - \hat{P}_{t+T,n}\right)^2 \quad (1)$$

이는 패치 수가 N 인 정답 미래 이미지 패치(P_{t+T})와 이미지 디코더를 통과하여 생성된 미래 이미지 패치($\hat{P}_{t+T}$) 사이의 평균 제곱 오차(Mean Square Error, MSE)를 사용하여 손실(L_{Recon})을 계산한다. 미래 이미지 예측 과정은 DreamVLA 와 동일하게 학습 시에만 활용되며, 추론 및 평가 시에는 학습된 Q_{Dream}만을 사용하여 추론 효율성을 증가하도록 하였다.

2.2 시각-언어-행동 모델 학습

미래 이미지 예측 이후, 이미지, 언어 지시, Q_{Dream}, 행동 쿼리(Action Query, Q_{Action})를 VLM

표 1. LIBERO-Spatial

	1	2	3	4	5	6	7	8	9	10	Avg
Open VLA [1]	92%	92%	84%	96%	66%	40%	92%	82%	74%	74%	79.2%
Ours	**100%**	**100%**	**100%**	**100%**	**98%**	**94%**	**98%**	**92%**	**96%**	**100%**	**97.8%**

표 2. LIBERO-Object

	1	2	3	4	5	6	7	8	9	10	Avg
Open VLA [1]	66%	70%	76%	56%	88%	68%	46%	72%	48%	70%	66.0%
Ours	**100%**	**98%**	**100%**	**92%**	**100%**	**100%**	**100%**	**98%**	**100%**	**100%**	**98.8%**

표 3. LIBERO-Goal

	1	2	3	4	5	6	7	8	9	10	Avg
Open VLA [1]	62%	92%	90%	56%	92%	82%	76%	96%	78%	68%	79.2%
Ours	**88%**	**94%**	**98%**	**88%**	**100%**	**96%**	**92%**	**100%**	**98%**	**78%**	**93.2%**

표 4. LIBERO-10

	1	2	3	4	5	6	7	8	9	10	Avg
Open VLA [1]	52%	70%	58%	42%	**46%**	**70%**	**40%**	60%	**32%**	**44%**	**51.4%**
Ours	**64%**	**82%**	**94%**	**46%**	0%	18%	4%	**94%**	0%	3%	43.2%

에 통과시킨다. 이 때, VLM 에서 Q_{Dream}을 제외한 나머지 위치에 해당하는 모든 레이어로부터 은닉 상태(Hidden State)를 추출한다. 추출된 은닉 상태와 초기 행동(Initial Action)은 정책(Policy) 네트워크에서 동일한 레이어 위치에 해당하는 Bridge Attention 의 키(Key, K)와 밸류(Value, V)로 전달된다.

Bridge Attention[3]은 그림 3 과 같이 세 종류의 attention 이 존재한다. 세 가지 모두 attention 에서 쿼리는 행동으로 고정되며, 시각 모두 attention 에서 쿼리는 행 리 은닉 상태, 시각-언어와 행동 쿼리를 결합한 은닉 상태가 K 와 V 로 사용된다. 이렇게 Bridge Attention 을 모두 통과하게 되면 T길이의 행동 시퀀스 $\hat{a}_{t:t+T-1}$를 얻는다. $\hat{a}_{t:t+T-1}$는 t 시점부터 $t+T-1$ 시점까지 로봇이 수행해야 할 연속된 행동 값이며, 하나의 행동은 정답 행동 $a_{t:t+T}$과 함께 식 (2)와 같이 평균 절대 오차(Mean Absolute Error, MAE)로 손실을 계산하는 데 사용된다.

$$L_{\text{Action}} = \sum |a_{t:t+T} - \hat{a}_{t:t+T}| \quad (2)$$

최종적으로 시각-언어-행동 모델의 학습은 미래 이미지 예측과 행동 예측을 동시에 수행하도록 하며, 식 (3)과 같이 손실과 행동 손실을 더하여 이미지 예측과 행동 예측을 함께 학습할 수 있도록 사용한다.

$$L_{\text{total}} = \alpha * L_{Recon} + L_{Action} \quad (3)$$

이 때, α는 이미지 예측 손실을 얼마나 반영할지에 대한 가중치이다.

3. 실험

비교 모델은 시각-언어-행동 모델의 대표적인 모델인 OpenVLA[1]로 설정하였고, 비교 모델과 제안 모델에 대한 평가는 로봇 팔 조작 과업에서 대표적인 시뮬레이션 환경인 LIBERO[10]에서 수행하였다. 제안 모델의 학습은 LIBERO 에서 하위 환경 단위로 제공되는 데이터집합을 사용하였으며, 각 하위 환경에 대해 개별적으로 평가를 진행하였다.

LIBERO 는 로봇 조작 과업의 특성에 따라 LIBERO-Spatial, LIBERO-Object, LIBERO-Goal, LIBERO-10 의 네 가지 하위 환경으로 구성되며, 각 하위 환경은 10 개의 언어 지시로 이루어진 과업들로 구성되어 있다.

LIBERO-Spatial 은 동일한 객체를 사용하되, 객체의 공간적인 배치를 변화시켜 물체 간 공간 관계 이해를 평가하기 위한 환경이며, LIBERO-Object 는 공간 배치는 고정한 채 객체를 변화시켜가며 객체 자체에 대한 이해를 평가하기 위한 환경이다.

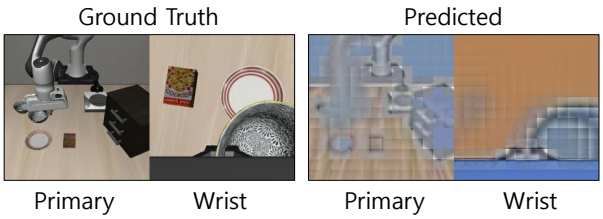

그림 4. 미래 이미지 예측 결과 시각화

LIBERO-Goal 은 객체와 기타 주변 환경을 고정한 채로 목표하는 언어 지시를 변경해가며 모델의 목표 지향적 행동에 대한 지식을 평가하기 위한 환경이고, 마지막으로 LIBERO-10 은 다양한 객체, 주변 환경과 함께 장기 조작을 평가하기 위한 환경이다.

3.1 실험 설정

제안 방법은 RTX3090 GPU 4 장을 사용하여 20,000 단계 동안 정책 학습을 진행하였다. 입력 시 사용한 시각 이미지는 작업 공간 전체를 포괄적으로 관찰할 수 있는 고정 카메라 시점과 로봇의 손목 시점을 사용하였다. 예측할 행동 시퀀스 길이 T는 8 로 설정하였으며, VLM 은 LoRA[11]를 적용하여 미세조정 하였다. 미래 이미지 예측을 위한 가중치 α는 0.1 로 설정하였다.

비교 모델과 제안 방법의 평가는 LIBERO 의 네 가지 하위 환경(LIBERO-Spatial, LIBERO-Object, LIBERO-Goal, LIBERO-10)별로 포함된 10 개의 과업에 대해 과업당 50 회의 에피소드를 실행하여 총 500 회의 평가를 진행하였다. 비교 모델은 하위 환경에서 학습하여 제공된 모델을 사용하여 평가하였으며, 실험 결과는 하위 환경에 대해 과업 성공 횟수를 기반으로 평균 성공률을 표 1-4 에 기록하였고, 표에서 우수한 성능을 보인 결과는 굵게 표시하였다.

3.2 시뮬레이션 환경에서의 성능 평가

평가 결과, 기존의 OpenVLA[1]는 LIBERO 전체 평균 68.95%의 성공률을 보였다. 이에 비해 미래 이미지 예측을 도입한 방법(Ours)은 LIBERO 전체 평균 83.25%로, OpenVLA 대비 14.3%의 성능 향상을 달성하였다. 구체적으로는 LIBERO-Spatial 에서 18.6%, LIBERO-Object 에서 32.8%, LIBERO-Goal 에서 14.0%의 성능 향상을 보였다. 이는 공간적 이해 또는 물체 중심, 목표 중심의 비교적 단기 조작 과업에서 미래 이미지 예측이 정책의 행동 학습에 보조적인 시각적 신호로 작용 될 수 있음을 시사한다.

반면 장기 조작 과업인 LIBERO-10 에서는 성공률이 51.4%에서 43.2%로 감소하였다. 이러한 성능 저하는 미래 이미지 예측이 장기적인 행동 예측으로 이어질 때, 단기적인 학습 단계에서 미세조정 된 경량 시각-언어-행동 모델에서는 일관되게 전체적인 성능 향상을 보장하지 않으며, 장기 조작 과업에서는 상반된 영향을 미칠 수 있음을 보여준다.

3.3 미래 이미지 예측 결과 시각화

본 실험에서 실제 미래 이미지 예측이 성공적으로 이루어졌는지 확인하기 위해, 평가 시 미래 이미지에 대해 시각화를 진행하였다. 이미지 디코더를 거친 시각화 결과는 고정 카메라 시점(Primary)과 손목 카메라 시점(Wrist)의 미래 이미지를 예측하며, 그림 4 와 같았다. 그림 4 우측의 예측 결과에서 보이듯이, 이미지 디코더는 로봇과 객체들에 대해 성공적으로 예측하는 것을 확인할 수 있었다. 이는 비교적 짧은 학습 단계 내에서도 이미지 디코더가 현재 관측으로부터 단기적인 시각적 변화 양상을 효과적으로 학습하였음을 의미한다.

LIBERO-10 에서도 미래 이미지 예측은 잘 수행되는 것을 확인하였으나, 오히려 단기 학습 단계에서는 성능 저하가 발생하였다. 이는 이미지 디코더는 잘 학습이 되어도 장기 조작 과업에서 정책 학습이 더디어지는 요인으로 적용됨을 확인할 수 있었다.

4. 결론

본 논문에서는 학습 효율적인 시각-언어-행동 모델을 대상으로 미래 세계 지식 예측이 정책 학습 성능에 미치는 영향을 확인하였다.

실험 결과, LIBERO 시뮬레이션 환경에서 공간적 이해, 물체 중심의 조작, 목표 이해가 필요한 과업에서는 미래 이미지 예측이 정책 학습에 도움을 줄 수 있음을 확인하였다. 반면, 장기적인 조작이 요구되는 과업에서는 이미지 디코더가 행동 학습을 방해하여 성능 저하로 이어질 수 있음을 보였다. 이러한 결과는 시각-언어-행동 모델에서 미래 지식 예측이 항상 효과적인 보조 학습 목표가 아님을 보여주며, 효과적인 통합을 위해 장기 조작 과업에 알맞는 미래 지식 예측 모듈 설계가 필요함을 시사한다.

감사의 글

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 디지털분야글로벌연구지원(IITP-2024-RS-2024-00417958)과 정보통신기획평가원의 지원(No.RS-2024-00357879, 지능형 개인 맞춤 만성질환 관리를 위한 AI 기반 생체데이터 융합 및 생성 기술)을 받아 수행된 연구 결과임.

참고문헌

[1] Kim, Moo Jin, et al. "Openvla: An open-source vision-language-action model." arXiv preprint arXiv:2406.09246 (2024).

[2] Kim, Moo Jin, et al. "Fine-tuning vision-language-action models: Optimizing speed and success." arXiv preprint arXiv:2502.19645 (2025).
[3] Wang, Yihao, et al. "Vla-adapter: An effective paradigm for tiny-scale vision-language-action model." arXiv preprint arXiv:2509.09372 (2025).
[4] Brohan, Anthony, et al. "Rt-1: Robotics transformer for real-world control at scale." arXiv preprint arXiv:2212.06817 (2022).
[5] Zitkovich, Brianna, et al. "Rt-2: Vision-language-action models transfer web knowledge to robotic control." Conference on Robot Learning. PMLR, (2023).
[6] Bjorck, Johan, et al. "Gr00t n1: An open foundation model for generalist humanoid robots." arXiv preprint arXiv:2503.14734 (2025).
[7] Zhang, Wenyao, et al. "Dreamvla: a vision-language-action model dreamed with comprehensive world knowledge." arXiv preprint arXiv:2507.04447 (2025).
[8] Black, Kevin, et al. "Zero-shot robotic manipulation with pretrained image-editing diffusion models." *arXiv preprint arXiv:2310.10639* (2023).
[9] Zhao, Qingqing, et al. "Cot-vla: Visual chain-of-thought reasoning for vision-language-action models." *Proceedings of the Computer Vision and Pattern Recognition Conference*. (2025).
[10] Liu, Bo, et al. "Libero: Benchmarking knowledge transfer for lifelong robot learning." Advances in Neural Information Processing Systems 36 (2023).
[11] Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." ICLR 1.2 (2022).