JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 31 No. 5, pp. 95-104, May 2026
https://doi.org/10.9708/jksci.2026.31.05.095

A Simulation-Based VQA Dataset for Evaluating Intrinsic Physical Property Inference Capabilities of Vision-Language Models

DongJu Jang*, Yeong-In Lee**, Ha-Young Kim***

*Student, Dept. of Artificial Intelligence, Yonsei University, Seoul, Korea
**Student, Graduate School of Information, Yonsei University, Seoul, Korea
***Associate Professor, Graduate School of Information, Yonsei University, Seoul, Korea

[Abstract]

This study proposes a multi-view video question-answering benchmark and dataset to train and evaluate vision-language models on inferring intrinsic physical properties, such as mass and elasticity, through robot-object interactions beyond simple scene recognition. To this end, we collected data by designing cube pushing and sphere dropping tasks based on inverse kinematics control within a simulation environment, and analyzed the performance by fine-tuning state-of-the-art models. The experimental results demonstrated that although pre-trained models showed low accuracy, their performance improved significantly in the mass inference task where the final displacement remains static, successfully overcoming existing text response biases after fine-tuning. Conversely, in the elasticity inference task, which requires tracking a momentary dynamic trajectory, the performance improvement was limited and the models exhibited a limitation of regressing to linguistic biases. In conclusion, this dataset provides an environment to quantitatively evaluate the physical reasoning capabilities of the models, contributing to laying the foundation for efficient action planning and decision-making in real-world robots in the future.

▸ **Key words**: Vision-Language Model, Interactive Perception, Simulation, Robot Dataset, Physical property

[요 약]

본 연구는 시각-언어모델이 장면 인식을 넘어 로봇과 객체 간 상호작용을 통해 질량과 탄성 같은 내재적 물리 속성을 추론하도록 학습 및 평가하는 멀티뷰 비디오 질의응답 벤치마크와 데이터셋을 제안한다. 이를 위해 시뮬레이션 환경에서 역기구학 제어를 바탕으로 큐브 밀기와 구 낙하 태스크를 설계해 데이터를 수집하고, 최신 모델들을 미세조정하여 성능을 분석하였다. 실험 결과, 사전 학습 모델은 낮은 정답률을 보였으나 미세조정 후 최종 변위가 정적인 질량 추론 태스크에서는 성능이 크게 향상되며 기존의 텍스트 응답 편향을 극복하였다. 반면 찰나의 동적 궤적을 추적하는 탄성 추론 태스크에서는 성능 향상이 제한적이었고 언어적 편향성으로 회귀하는 한계를 보였다. 결론적으로 본 데이터셋은 모델의 물리 추론 능력을 정량적으로 평가할 수 있는 환경을 제공하여, 향후 실제 로봇의 효율적인 행동 계획과 의사결정 기반을 마련하는 데 기여한다.

▸ **주제어**: 시각-언어 모델, 상호작용 인지, 시뮬레이션, 로봇 데이터셋, 물리 속성

• First Author: DongJu Jang, Yeong-In Lee, Corresponding Author: Ha-Young Kim
*DongJu Jang (tygu1004@yonsei.ac.kr), Dept. of Artificial Intelligence, Yonsei University
**Yeong-In Lee (zeroin@yonsei.ac.kr), Graduate School of Information, Yonsei University
***Ha-Young Kim (hayoung.kim@yonsei.ac.kr), Graduate School of Information, Yonsei University
• Received: 2026. 03. 27, Revised: 2026. 05. 11, Accepted: 2026. 05. 13.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

최근 시각-언어 모델(Vision-Language Model, VLM)은 이미지 및 비디오 이해, 장면 묘사, 그리고 일반적인 상식 추론에서 뛰어난 성능을 보이며 빠르게 발전하고 있다[1-4]. 이러한 발전은 로보틱스 분야로도 확장되어, VLM이 단순한 환경 인식 도구를 넘어 로봇의 고차원적인 행동 계획(Action planning)을 지원하는 핵심 모듈로 활용되고 있다[5-9]. 특히 시각적 장면 이해와 언어 지시 해석을 바탕으로 고수준 의사결정을 수행하고, 나아가 행동 생성 및 제어 과정과 연계되는 연구가 활발히 이루어지고 있다[5-9].

그러나 이러한 발전에도 불구하고, 실제 로봇 조작 환경에서의 적용은 여전히 쉽지 않다. 기존 VLM 기반 접근은 주로 관찰된 장면으로부터 객체의 색상, 형태, 위치와 같은 명시적 시각 단서(Explicit visual cues)를 해석하고 이를 바탕으로 행동을 계획하는 데 초점을 두고 있다[1-4]. 하지만 실제 조작에서는 이러한 외형 정보만으로 충분하지 않으며, VLM이 실세계에서 안전하고 신뢰할 수 있는 행동 계획을 수립하기 위해서는 상호작용을 통해서만 파악할 수 있는 객체의 내재적 물리 속성(Intrinsic physical properties)을 이해하는 능력이 필수적이다[11-13]. 예를 들어, 사람이 외형이 동일한 여러 개의 상자를 안정적으로 쌓으려 할 때, 단순히 눈으로 관찰하는 것에 그치지 않고 상자를 살짝 밀어보거나 들어보는 상호작용을 수행한다. 이를 통해 질량과 무게 중심을 파악한 뒤, 무거운 상자를 아래로 배치하는 등 비교적 안정적인 궤적과 행동을 계획한다 [12,13]. 마찬가지로, 로봇 제어 시스템의 VLM 플래너 역시 대상 객체를 밀거나 떨어뜨리는 등의 능동적 상호작용(Interactive perception)을 통해 질량이나 탄성과 같은 숨겨진 물리 속성을 파악하고, 이를 반영하여 정교한 제어 계획을 세울 수 있어야 한다.

그러나, 현재 사용되는 비디오 질의응답 및 물리 추론 벤치마크[14-18]는 대부분 인터넷에서 수집된 일반 영상이나 정적인 질문에 기반하고 있으며[14-18], 로봇이 작업 공간 내에서 직접 물리적 피드백을 획득하는 과정을 충분히 반영하지 못한다[19,20]. 또한 우리가 아는 한, 로봇-객체의 직접적인 상호작용을 통해 질량이나 탄성과 같은 내재적 물리 속성이 드러나도록 설계된 데이터셋은 매우 제한적이다[14,17-20]. 즉, 기존 벤치마크와 실제 로봇 조작 환경 사이에는 근본적인 도메인 격차(Domain gap)가 존재한다.

이에 본 연구는 로봇-객체 간 상호작용 과정에서 나타나는 동적 물리 단서를 바탕으로 내재적 물리 속성을 추론할 수 있도록 하는 데이터셋과 학습 구조를 제안한다. 구체적으로 본 연구의 기여는 다음과 같다.

첫째, 내재적 물리 속성 추론을 위한 물리 시뮬레이션 기반 합성 비디오 데이터 생성 프레임워크를 제안한다. 제안된 프레임워크는 시뮬레이션 환경에서 물리 파라미터를 통제한 상태로 로봇-객체 상호작용 장면을 생성하고, 이를 멀티뷰 비디오와 질의응답 학습 데이터로 자동 구성할 수 있도록 한다.

둘째, 질량과 탄성을 중심으로 한 로봇 상호작용 태스크 및 데이터셋을 구축한다. 구체적으로, 큐브 밀기(Cube pushing)과 구 낙하(Sphere dropping) 시나리오를 설계하여 모델이 동적 반응을 근거로 내재적 물리 속성을 추론하도록 유도한다.

셋째, 상대적 비교 기반의 멀티뷰 비디오 VQA(Visual Question Answering) 벤치마크를 제시한다. 이를 통해 VLM이 단순한 장면 인식 수준을 넘어, 로봇-객체 간 상호작용에서 나타나는 이동 거리, 반발 양상, 복원 궤적과 같은 물리적 단서를 활용하여 내재적 물리 속성을 추론할 수 있는지를 정량적으로 평가한다.

본 연구는 VLM이 단순한 시각 인식을 넘어 로봇-객체 간 상호작용에 기반한 내재 물성 추론 능력을 학습하고 평가할 수 있는 새로운 방향을 제시하며, 향후 실제 로봇 시스템에서 보다 안전하고 효율적인 의사결정과 행동 계획을 가능하게 하는 기반을 마련한다.

본 논문의 구성은 다음과 같다. II장에서 시각-언어 모델, 상호작용 기반 인지, 물리 추론을 위한 시뮬레이션 기반 데이터셋 관련 선행연구를 정리하고, III장에서 제안하는 물리 시뮬레이션 기반 데이터셋 구축 과정과 로봇-객체 상호작용 태스크 설계 방법을 설명한다. IV장에서는 구축한 데이터셋을 활용한 VLM 미세조정 및 성능 평가 결과를 분석하며, 마지막으로 V장에서 결론 및 향후 연구 방향을 제시한다.

실험 재현 가능성 확보를 위해 전체 소스 코드 및 데이터셋은 각각 GitHub(https://github.com/tygu1004/hidden_phys_vqa_project), HuggingFace(https://huggingface.co/datasets/mlcf-robot/hidden_phys_vqa_dataset s-v1)를 통해 공개하였다.

II. Related work

2.1 Vision-Language Models

시각-언어 모델은 대규모 이미지-텍스트 정렬 학습을 통해 시각 정보와 언어 표현을 통합하는 방향으로 발전해

왔다[1-4]. 초기의 대표적 연구인 CLIP[1]은 대규모 시각-언어 사전학습이 다양한 다운스트림 작업에 효과적임을 보여주었고, 이후 연구들은 단순한 정렬을 넘어 멀티모달 추론과 지시 이행 능력까지 확장되었다[2-4].

최근에는 비디오 이해와 장시간 시퀀스 정보 처리가 가능한 모델들[10,21-25]이 등장하면서, VLM은 정적 이미지 이해를 넘어 동적 장면 해석으로 빠르게 확장되고 있다. 대표적으로 QwenVL[10,21] 계열은 장시간 비디오 이해를 지원하며, InternVL[22-24] 계열은 멀티모달 사전학습과 복합 추론 성능을 크게 확장하였다. 또한, Cosmos-Reason[25] 계열은 일반적인 장면 이해를 넘어, 물리적 상식과 체화 추론(Embodied reasoning)을 중심으로 설계되어, 공간적·시간적 관계와 물리적 상호작용에 대한 추론 가능성을 보여주었다. 이는 최근 VLM 연구가 단순한 시각 인식을 넘어 물리적 세계 이해로 확장되고 있음을 보여준다.

그러나 이러한 최신 모델들의 발전에도 불구하고, 질량이나 탄성과 같은 내재적 물리 속성을 실제 상호작용 비디오를 통해 얼마나 안정적으로 추론할 수 있는지는 여전히 충분히 검증되지 않았다. 이는 VLM의 물리 추론 능력이 장면 이해를 넘어 내재적 물성 추론으로 확장될 수 있는지에 대한 추가적인 분석 필요성을 시사한다.

2.2 Interactive Perception

상호작용 기반 인지(Interactive Perception)는 로봇이 물체와의 목적 지향적 상호작용을 통해 질량, 마찰, 강성 등과 같은 내재된 물리 속성에 대한 단서를 획득하는 연구 흐름이다[12]. 이는 시각 관측을 수동적으로 수용하는 기존 접근에서 벗어나, 행동을 센싱 수단으로 활용하는 능동적 관점을 취한다. 예를 들어, 물체를 밀거나 떨어뜨리는 과정에서 발생하는 동역학적 반응은 외형만으로는 구분하기 어려운 물리 속성 차이를 드러내며, 이러한 상호작용은 잠재적인 물리 속성 추론에 중요한 단서를 제공한다.

그러나 기존 연구는 주로 특정 센서로 측정된 물성 정보를 입력으로 활용하거나[26], 비전 및 촉각 정보를 결합하여 물성을 추정, 식별하는 접근에 집중되어 왔다[27,28]. PhysBench[14]와 같은 연구는 VLM의 물리 이해 및 추론 능력을 평가하기 위한 벤치마크를 제시하였으나, 상호작용을 통해 드러나는 동적 물리 단서를 활용하여 내재된 물성을 추론하는 문제는 충분히 다루어지지 않았다. 또한, 기존 연구 방향은 물리적 가능성 판단이나 결과 예측에 초점을 맞추는 경우가 많아 제한적이었다.

이에 본 연구는 상호작용 기반 인지의 문제의식을 계승하되, 별도의 촉각 정보나 전용 물성 센서 없이 로봇-물체 상호작용으로부터 생성된 동적 비디오만을 입력으로 사용하여, VLM이 내재적 물리 속성을 VQA 형식으로 추론할 수 있는지를 평가한다.

2.3 Simulation-based Datasets for Physical Reasoning

물리 추론 연구에서는 실제 환경에서 대규모 상호작용 데이터를 수집하기 어렵기 때문에, 합성 데이터와 시뮬레이션 환경이 중요한 역할을 해왔다. 초기 연구들은 비디오로부터 질량, 밀도, 반발계수와 같은 물리 속성을 학습하려는 문제를 제시하였고[11,15], 이후 CLEVRER[17]와 PHYRE[18]와 같은 벤치마크는 합성 환경에서 시간적·인과적 추론이나 물리적 가능성 판단 능력을 평가하는 방향으로 발전하였다. 최근에는 PhysBench[14]와 같이 VLM의 물리 이해 능력을 평가하기 위한 멀티모달 벤치마크도 제안되고 있으며, 로봇 조작 데이터를 기반으로 한 VQA 및 벤치마크[19,20]도 등장하였다.

그러나 이러한 연구들은 주로 사건 예측과 광범위한 물리 이해 평가에 초점을 두고 있으며, 외형만으로 직접 관찰하기 어려운 내재적 물리 속성을 상호작용을 통해 추론하는 문제를 직접적으로 다루지는 않는다. 또한 물리 파라미터를 통제한 상태에서 특정 물성의 영향만을 분리해 분석하거나, 이를 VLM 기반 질의응답 학습과 연결한 연구는 아직 제한적이다.

따라서 본 연구는 Isaac Sim[29] 환경에서 물리 파라미터를 체계적으로 제어하고, 로봇-객체 상호작용을 통해 질량과 탄성이 드러나는 비디오 데이터와 질의응답 쌍을 구축한다. 이를 통해 기존 시뮬레이션 기반 물리 추론 연구가 충분히 다루지 못한 내재적 물성 중심의 VQA 문제를 정식화하고, VLM이 동적 물리 단서를 바탕으로 내재적 물리 속성을 추론할 수 있는지를 평가하고자 한다.

III. Dataset Construction

본 연구에서는 VLM이 비디오를 통해 질량, 탄성과 같이 겉으로 드러나지 않는 객체의 내재적 물리 속성을 추론할 수 있도록, 물리 시뮬레이션 기반의 합성(Synthetic) VQA 데이터셋을 구축하는 파이프라인을 제안한다.

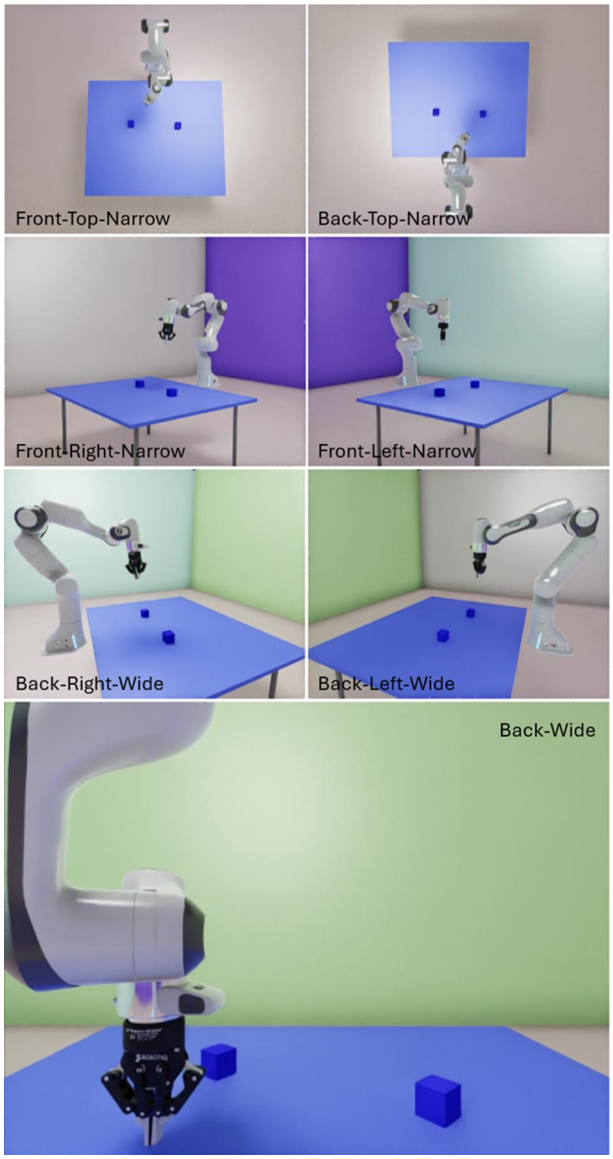

Fig. 1. Multi-view cameras for the cube pushing task

3.1 Parallel Data Collection Environment based on Isaac Sim

대규모 비디오 데이터를 단시간에 효율적으로 생성하기 위해, 고해상도 물리 렌더링을 지원하는 NVIDIA Isaac Sim과 로보틱스 시뮬레이션 프레임워크인 Isaac Lab을 활용하였다. 데이터 수집 파이프라인은 Isaac Lab의 매니저 기반 환경(ManagerBasedRLEnv)을 바탕으로 환경을 병렬화하여 구성하였다. 이를 통해 수백 개의 에피소드를 동시에 시뮬레이션함으로써 데이터 수집 속도를 극대화하였다.

시뮬레이션 환경 내에서 정밀한 동역학 모델링과 사실적인 렌더링이 적용된 Franka Emika Panda 로봇 팔과 Robotiq 2F-85 그리퍼(Gripper), 그리고 상호작용을 위한 작업 테이블 등의 에셋(Asset)을 배치하였다. 이를 통해 실제 물리 세계와 매우 유사한 수준의 마찰, 충돌, 중력 연산이 수행되는 고정밀 상호작용 환경을 제공한다. 다양성을 고려하여 구성한 시뮬레이션 환경의 예시는 Fig.1과 Fig.2에서 확인할 수 있다.

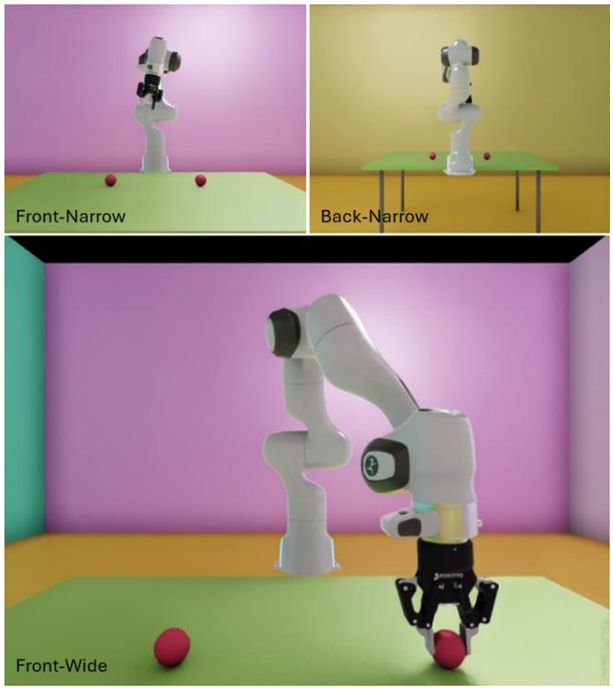

Fig. 2. Multi-view cameras for the sphere dropping task (including only the additional cameras from the cube pushing setup)

3.2 IK-based Control for Robot-Object Interaction

로봇과 객체 간의 일관된 상호작용을 구현하여 모델이 물리적 움직임 그 자체에 집중할 수 있도록, 역기구학(Inverse Kinematics, IK) 컨트롤러 기반의 확정적 로봇 제어 방식을 채택하였다. 시뮬레이션 환경 내 대상 객체의 초기 위치(Root link position)를 기준으로 로봇의 엔드 이펙터(End-effector)가 거쳐야 할 접근(Approach), 파지(Grasp), 조작(Push/Drop), 해제(Release) 지점 등의 3차원 웨이포인트(Waypoints)를 사전에 계산하였다. 이후 IK 컨트롤러가 이 설정된 궤적을 따라 관절 각도를 연산하여 다음 두 가지 핵심 태스크를 수행한다.

▸ 질량 비교를 위한 큐브 밀기 태스크 (Cube Pushing Task): 로봇이 테이블 위에 놓인 두 개의 큐브를 각각 앞으로 밀어내는 동작을 수행한다. 로봇이 동일한 궤적과 힘으로 물체를 조작할 때, 대상의 질량에 따라 큐브가 밀리는 속도 및 가속도의 시각적 차이가 발생하도록 유도한다.

▸탄성 비교를 위한 구 낙하 태스크 (Sphere Dropping Task): 로봇이 두 개의 구(Sphere)를 동일한 높이로 들어

올린 후 낙하시키는 동작을 수행한다. 낙하 후 튕겨 오르는 궤적을 통해 물체의 탄성을 시각적으로 파악할 수 있도록 동작을 설계하였다.

3.3 Domain Randomization

상호작용의 시각적 관찰만으로 대상의 정확한 물리적 수치(절대치)를 도출하는 것에는 본질적인 한계가 있다. 따라서 본 연구는 모델이 사람의 인지 방식과 유사하게 시각적 단서를 통해 물리량의 상대적 차이를 구분할 수 있는지 평가하는 데 목적을 두었으며, 사람이 시각적으로 구분이 가능한 수준의 차이가 발생하도록 파라미터 범위를 설정하였다.

매 에피소드가 시작될 때마다 큐브 밀기 태스크에서는 두 큐브의 질량을 각각 0.5~1.0(가벼운 물체)과 3.0~5.0(무거운 물체) 사이에서 무작위로 샘플링하였다. 구 낙하 태스크에서는 두 구의 반발 계수(Restitution)를 0.0에서 1.0 사이로 설정하되, 시각적 구분을 위해 두 물체 간 최소 0.2의 반발 계수 차이(Min gap)를 두었다.

학습 모델이 특정 시야각이나 표면적 텍스처에 과적합되는 것을 방지하기 위해 광범위한 도메인 무작위화(Domain randomization) 또한 적용하였다. 매 에피소드마다 테이블, 바닥, 벽, 상호작용 물체의 색상과 조명 위치를 무작위로 변경하였으며, 대상 객체의 초기 위치에도 x, y축 기준 ±1cm 이내의 미세한 변화를 주었다. 더불어 특정 동작 순서에 따른 시간적 편향(Temporal bias)을 방지하기 위해, 로봇이 좌우 두 물체 중 어느 것과 먼저 상호작용할지에 대한 순서도 무작위로 결정하였다.

마지막으로 다양한 시점에서의 물리적 상호작용 과정을 담기 위해, 환경 내에 전후방, 광각 및 협각 등 서로 다른 시야각과 초점 거리를 가진 7개의 카메라 센서를 배치하여 RGB 비디오 클립을 자동 렌더링하고 저장하도록 데이터 수집을 다각화하였다. 서로 다른 카메라 시점 및 시야각이 적용된 큐브 밀기 태스크의 환경 예시는 Fig.1에서 확인할 수 있으며, Fig.2는 큐브 밀기 태스크와 다르게 설정된 구 낙하 태스크의 카메라 배치 및 시점 구성을 보여준다.

3.4 VQA Dataset for Intrinsic Physical Properties

시각적 관찰을 통한 물리 속성 파악은 정확한 수치 예측이 아닌 상대적 차이 비교만이 가능하므로, 본 연구의 VQA 프롬프트 역시 두 물체 간의 특성을 비교하는 질문으로 설계되었다. 수집된 7개 시점의 멀티뷰 비디오는 모델의 정량적 학습 및 평가를 위해 단답형 VQA 형식으로 가공되었으며, 시스템 프롬프트를 통해 오직 'left' 또는 'right'로만 응답하도록 제한하여 객관적인 성능 지표 산출이 가능하도록 구성하였다.

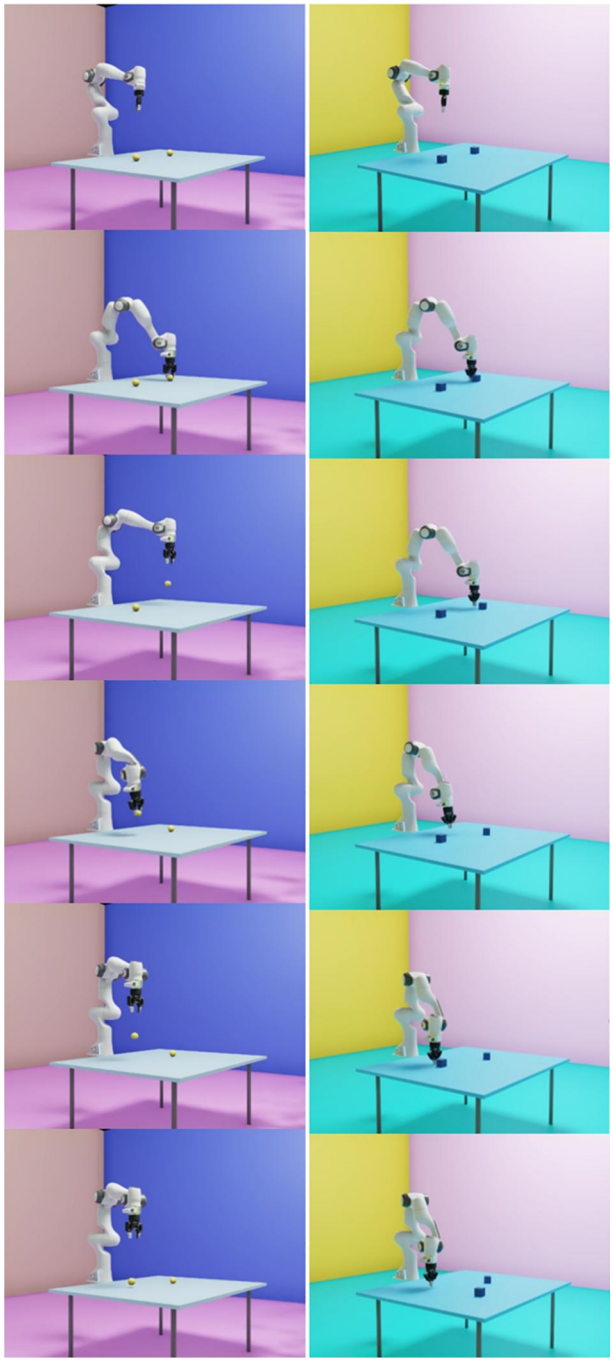

Fig. 3. Sample Episode (L:Drop/R:Push)

▸ 질량 추론 프롬프트

"From the perspective of the robot arm facing forward, which cube looks heavier, the left one or the right one?"

▸ 탄성 추론 프롬프트

“From the perspective of the robot arm facing

forward, which sphere looks more elastic, the left or the right?"

Table 1. Dataset Information

Dataset		Mass	Elasticity	Total
Episodes	Train	123	100	223
	Eval	31	25	56
	Total	154	125	**279**
VQA Pairs	Train	861	700	1,561
	Eval	217	175	392
	Total	1,078	875	**1,953**

전체 VQA 데이터셋은 다음과 같이 구축되었다. 질량 추론 태스크는 학습용 123개, 평가용 31개의 에피소드를 통해 총 1,078개(학습 861개, 평가 217개)의 VQA 쌍을 확보하였다. 탄성 추론 태스크는 학습용 100개, 평가용 25개의 에피소드로 총 875개(학습 700개, 평가 175개)의 VQA 쌍을 구축하였다. 결과적으로 총 253개의 상호작용 에피소드로부터 1,953개의 VQA 데이터 쌍이 최종 완성되었다. 데이터셋 샘플은 Fig.3에서 확인할 수 있다.

IV. Experiments

본 장에서는 앞서 구축한 물리적 상호작용 기반의 멀티뷰 VQA 데이터셋을 활용하여, 최신 대형 VLM들이 객체의 내재적 물리 속성을 얼마나 정확하게 추론할 수 있는지 정량적으로 검증한 결과를 제시한다. 먼저 본 실험의 상세 목적과 비교 모델 선정 기준, 그리고 미세조정(Fine-tuning) 환경을 서술한다. 이어서 질량 및 탄성 추론 태스크에 대한 정답률 변화를 통해 VLM의 물리 속성 추론 능력을 평가하고, 프롬프트 내 선택지 순서 변경 실험 등을 통해 사전 학습 모델에 내재된 텍스트 응답 편향(Response Bias)과 미세조정의 한계를 심층적으로 분석한다. 본 연구는 VLM의 물리적 추론 한계를 진단하는 것이 목적이므로, 사전 학습된 VLM의 추론 결과를 성능 비교의 기점으로 삼았다.

4.1 Experimental Purpose and Model Selection

본 실험의 주된 목적은 최신 VLM이 로봇과의 상호작용 비디오를 통해 대상 객체의 내재적 물리 속성(질량, 탄성 등)을 얼마나 정확하게 파악하고 추론할 수 있는지 그 능력을 검증하는 것이다. 이를 위해 본 연구에서는 시각 정보 처리 방식과 아키텍처의 설계 목적이 뚜렷하게 구분되는 세 가지 최신 오픈소스 VLM을 실험군으로 선정하여 성능을 교차 검증하였다.

특히, 로봇 시스템에서의 실용적인 적용 가능성과 연산 효율성을 고려하여 8~9B 파라미터 규모를 기준으로 세 모델의 체급을 유사하게 통제하였다. 이를 통해 단순한 파라미터 크기 차이로 인한 성능 왜곡을 방지하고, 각 아키텍처가 가진 본연의 물리 추론 능력을 공정하게 비교하고자 하였다.

- Qwen 3.5 (9B): 텍스트와 이미지/비디오 전반에서 범용적으로 뛰어난 멀티모달 지시 이행 능력을 입증한 대표적인 최신 VLM이다. 일반적인 대형 언어 모델 기반 VLM이 비디오 내 동역학적 변화를 얼마나 잘 이해할 수 있는지 가늠하기 위한 표준 기준점으로서 선정하였다.
- InternVL 3.5 (9B): 동적 고해상도 패칭(Patching) 기술과 강력한 시각 인코더(Vision encoder)를 탑재하여 세밀한 시각적 특징 추출에 강점을 지닌 모델이다. 물체가 밀리는 거리나 튕겨 오르는 높이 등, 시각적 민감도가 요구되는 본 실험 환경에서 시각 중심 아키텍처의 성능을 확인하기 위해 선정하였다.
- Cosmos Reason 2 (8B): 시각적 데이터로부터 물리적 세계의 동역학(Dynamics)을 학습하는 데 특화된 최신 월드 모델(World model) 아키텍처이다. 모델의 설계 목적 자체가 물리적 인과관계와 시공간적 변화의 이해에 있는 만큼, 본 연구의 '상호작용 기반 물리 속성 추론' 태스크에 가장 직접적으로 부합하는 모델로서 그 효용성을 검증하기 위해 선정하였다.

4.2 Experimental Setup for VLM Fine-tuning

위 세 모델의 내재적 물리 속성 추론 능력을 공정하게 비교하기 위해, 사전 학습(Pre-trained) 모델에 대한 추론 평가를 진행한 후 아래의 Table 2과 같이 동일한 학습 하이퍼파라미터를 적용하여 미세조정을 수행하였다. 파라미터 업데이트 연산 효율성을 위해 LoRA(Low-Rank Adaptation) 기법을 도입하였으며, 주요 어텐션 및 피드포워드 모듈을 학습 타겟으로 설정하여 시각적 동역학 특징이 모델 내부에 효과적으로 전이되도록 구성하였다.

Table 2. Hyper-parameters for Vision-Language Model fine-tuning

Hyper-parameter		Value
Optimizer		adamw
Learning Rate		0.00005
LR Scheduler		Linear
Random Seed		42
Batch Size		4(A100), 2(RTX6000ADA)
Train Epochs		3
Video	FPS	2
	Resolution	224*224 (qwen3.5/cosmos-reason2) 448*448 (internvl3.5)
LoRA	Rank(r)	16
	Alpha	32
	Dropout	0.1
	Target Modules	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj

4.3 Evaluation of Intrinsic Physical Property Inference Performance

Table 3. Accuracy of VLMs on physical property inference tasks: Pre-trained vs. Fine-tuned

Task	큐브 밀기 (Push)		
Model	Pre-trained	Fine-Tuned	Δ
Qwen 3.5	0.4977	**0.9908**	+0.4931
InternVL 3.5	0.5300	**0.9908**	+0.4608
Cosmos-R2	0.5346	**0.9816**	+0.4470
Task	구 낙하 (Drop)		
Model	Pre-trained	Fine-Tuned	Δ
Qwen 3.5	0.4514	**0.7371**	+0.2857
InternVL 3.5	0.6057	**0.5771**	-0.0286
Cosmos-R2	0.5429	**0.6171**	+0.0742

모델의 평가는 VLM의 응답이 실제 물리 속성의 대소 관계 정답과 일치하는 비율인 정답률을 지표로 사용하였다. 미세조정 전후의 태스크별 전체 정답률 비교 결과는 Table 3과 같다.

실험 결과, 미세조정을 거치지 않은 사전 학습 모델들은 두 태스크 모두에서 약 0.45~0.53의 정답률을 기록하여 사실상 무작위 추측 수준에 머물렀다. 이는 기존 VLM들이 비디오 내 객체의 동적 움직임만을 관찰하여 내재적 물리 속성을 파악하는 능력을 자체적으로 갖추지 못했음을 시사한다.

그러나 제안된 데이터셋으로 미세조정을 수행한 후, 질량 비교를 위한 큐브 밀기 태스크에서는 세 모델 모두 0.98~0.99에 달하는 비약적인 성능 향상을 달성하였다. 이는 로봇 팔이 물체를 밀어낸 후 발생하는 최종 이동 거리의 차이가 비디오 후반부에 뚜렷한 정적 상태로 남기 때문으로 분석된다. VLM의 미세조정을 통해 명확한 시각적 변위 차이를 질량이라는 내재적 속성과 연관 짓는 능력을 성공적으로 학습한 것으로 볼 수 있다.

반면, 탄성 비교를 위한 구 낙하 태스크의 결과는 다소 상이한 양상을 보였다. Qwen3.5와 Cosmos-Reason2는 성능 향상을 보였으나 큐브 밀기 태스크의 정답률에는 크게 미치지 못했으며, InternVL3.5는 미세조정 후 오히려 정답률이 소폭 하락하는 현상을 보였다. 탄성 추론은 단순히 최종 상태를 비교하는 것이 아니라, 두 물체가 바닥에 부딪히고 튕겨 오르는 과정에서의 최고 도달 고도라는 찰나의 동적 궤적(Dynamic trajectory)을 동시에 추적해야 한다. 본 결과는 이러한 고주파적(High-frequency) 시공간적 변화를 포착하고 물리 법칙을 추론하는 것이 현재의 VLM 아키텍처에 있어 구조적으로 매우 난이도가 높은 과제임을 실증적으로 보여준다.

4.4 Analysis of Response Bias and Limitations in Pre-trained VLMs

사전 학습된 VLM이 물리적 직관이 부족한 상태에서 어떠한 응답 경향성을 띠는지, 그리고 태스크의 시공간적 난이도에 따라 미세조정의 편향 교정 효과가 어떻게 달라지는지 검증하기 위해, 편향성이 가장 뚜렷하게 관찰된 InternVL 3.5 모델을 중심으로 정답 위치에 따른 예측 정확도를 분석하였다(Table 4).

Table 4. Comparison of task accuracy by ground truth position for InternVL 3.5

Pos	Left		Right	
Task	Pre-train	Fine-T	Pre-train	Fine-T
Push	1.0000	1.0000	0.0286	0.9810
Drop	0.9714	0.8571	0.0571	0.1571

미세조정 전 결과를 살펴보면, 큐브 밀기와 구 낙하 두 태스크 모두에서 정답이 왼쪽일 때는 100%에 가까운 정확도를 보이지만 오른쪽일 때는 2.8% 정답률을 보이며 극단적인 응답 편향이 관찰된다.

추가적으로 프롬프트 내 선택지 순서를 변경하는 절제 연구를 수행하였으나, InternVL 3.5 모델은 여전히 'left'를 정답으로 출력하는 강한 응답 편향을 보였다. 특히 구 낙하 태스크의 경우 미세조정 후에도 이러한 편향성이 해결되지 않았으며, 이는 모델이 동역학 정보를 제대로 추론하지 못할 때 텍스트 편향으로 강하게 회귀함을 시사한다[30].

제안된 데이터셋으로 미세조정을 거친 후, 태스크의 물리적 난이도에 따라 편향 교정 효과가 극명하게 갈리는 현상이 확인되었다. 이는 모델이 텍스트 확률에 의존하는 것을 멈추고 시각적 변위 단서와 질량 간의 인과관계를 성공적으로 학습했음을 의미한다.

반면, 찰나의 동적 궤적을 추적해야 하는 구 낙하 태스크의 경우, 미세조정을 수행한 이후에도 여전히 편향이 해결되지 않았다(Table 4). 이는 VLM이 비디오 내의 고주파적 시공간적 변화를 추론하기 어려울 때, 새로 학습한 물리적 지식보다 기존에 내재된 언어적 편향성으로 다시 회귀하려는 경향이 있음을 시사한다.

결론적으로 본 실험은 VLM이 단순한 물리적 상호작용은 미세조정을 통해 성공적으로 학습하고 편향을 극복할 수 있으나, 복잡한 동역학 추론을 완벽히 수행하기 위해서는 단순한 미세조정을 넘어 아키텍처의 근본적인 개선이 필요함을 실증적으로 제시한다.

V. Conclusion

본 연구는 로봇-객체 간 상호작용 과정에서 드러나는 동적 물리 단서를 기반으로, 시각-언어 모델이 질량과 탄성과 같은 내재적 물리 속성을 추론할 수 있는지를 체계적으로 검증하고자 하였다. 이를 위해 물리 시뮬레이션 기반의 합성 멀티뷰 비디오 데이터 생성 프레임워크를 구축하고, 질량 비교를 위한 큐브 밀기 태스크와 탄성 비교를 위한 구 낙하 태스크를 설계하였으며, 상대적 비교 기반의 비디오 VQA 벤치마크를 제안하였다. 또한 제안된 데이터셋을 활용하여 최신 오픈소스 VLM들을 미세조정하고, 상호작용 기반 물성 추론 능력을 정량적으로 평가하였다.

실험 결과, 사전 학습된 VLM들은 두 태스크 모두에서 대체로 무작위 추측 수준의 성능에 머물렀으며, 이는 일반적인 비디오 이해 능력만으로는 내재적 물리 속성 추론이 충분히 이루어지지 않음을 보여주었다. 반면 제안된 데이터셋으로 미세조정을 수행한 후, 질량 비교를 위한 큐브 밀기 태스크에서는 세 모델 모두 0.98~0.99 수준의 높은 정확도를 달성하여, 명확한 시각적 변위 차이를 질량이라는 내재적 속성과 연결하는 능력을 효과적으로 학습할 수 있음을 확인하였다. 그러나 탄성 비교를 위한 구 낙하 태스크에서는 성능 향상이 제한적이었으며, 일부 모델에서는 오히려 정확도가 하락하는 현상도 관찰되었다. 이는 순간적인 반발 궤적과 같은 고주파 시공간 정보를 추적하고 물리 법칙과 연결하는 문제가 현재 VLM 아키텍처에 여전히 어려운 과제임을 시사한다.

또한 응답 편향 분석을 통해, 사전 학습된 모델은 물리적 직관이 부족할 경우 특정 텍스트 응답에 강하게 편향되는 경향을 보였으며, 미세조정 이후에도 태스크의 물리적 난이도에 따라 편향 교정 효과가 다르게 나타남을 확인하였다. 특히 질량 비교 태스크에서는 편향이 크게 완화된 반면, 탄성 비교 태스크에서는 편향이 여전히 남아 있었다. 이는 단순한 미세조정만으로는 복잡한 동역학 추론 문제를 충분히 해결하기 어렵고, 물리적 인과관계를 보다 정교하게 포착할 수 있는 모델 구조의 개선이 필요함을 보여준다.

본 연구는 VLM이 단순한 시각 인식을 넘어 로봇-객체 간 상호작용으로부터 내재적 물리 속성을 추론하도록 학습하고 평가할 수 있는 새로운 연구 방향을 제시한다는 점에서 의의를 가진다. 특히 기존 인터넷 비디오 기반 벤치마크와 실제 로봇 조작 환경 사이의 도메인 격차를 완화하고, 로봇 조작에 필요한 잠재 물성 추론 능력을 정량적으로 분석할 수 있는 기반을 마련하였다. 향후 연구에서는 마찰, 강성, 무게 중심 등 더 다양한 물성으로 태스크를 확장하고, 실제 로봇 환경으로의 전이 실험을 수행할 필요가 있다. 또한 복잡한 동역학 추론과 응답 편향 문제를 해결하기 위해, 시간적 변화와 물리적 인과관계를 보다 효과적으로 반영할 수 있는 VLM 아키텍처 및 학습 전략에 대한 추가적인 연구가 요구된다.

ACKNOWLEDGEMENT

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No.RS-2025-02303870, Software Technology for Efficient Multimodal Visual Information Processing in High-Speed Spatial Interactions)

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2023R1A2C200337911)

REFERENCES

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," Proceedings of the 38th International Conference on Machine Learning (ICML), pp. 8748-8763, July 2021. DOI: 10.48550/arXiv.2103.00020

[2] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Bińkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan, "Flamingo: a Visual Language Model for Few-Shot Learning," Advances in Neural Information Processing Systems (NeurIPS), Vol. 35, pp. 23716-23736, December 2022. DOI: 10.48550/arXiv.2204.14198

[3] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual Instruction Tuning," Advances in neural information processing systems, 36, 34892-34916, April 2023. DOI: 10.48550/arXiv.2304.08485

[4] J. Li, D. Li, S. Savarese, and S. C. H. Hoi, "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models," Proceedings of the 40th International Conference on Machine Learning (ICML), pp. 19730-19742, July 2023. DOI: 10.48550/arXiv.2301.12597

[5] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, and C. Finn, "OpenVLA: An Open-Source Vision-Language-Action Model," Proceedings of The 8th Conference on Robot Learning (CoRL), pp. 2679-2713, November 2024. DOI: 10.48550/arXiv.2406.09246

[6] B. Zitkovich, A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. M. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, P. Florence, C. Fu, M. Gonzalez Arenas, K. Gopalakrishnan, K. Han, K. Hausman, A. Herzog, J. Hsu, B. Ichter, A. Irpan, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, L. Lee, T.-W. E. Lee, S. Levine, Y. Lu, H. Michalewski, I. Mordatch, K. Pertsch, K. Rao, K. Reymann, M. Ryoo, G. Salazar, P. Sanketi, P. Sermanet, J. Singh, A. Singh, R. Soricut, H. Tran, V. Vanhoucke, Q. Vuong, A. Wahid, S. Welker, P. Wohlhart, J. Wu, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and A. Brohan, "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control," Proceedings of The 7th Conference on Robot Learning (CoRL), pp. 2165-2183, November 2023. DOI: 10.48550/arXiv.2307.15818

[7] M. U. Din, W. Akram, L. S. Saoud, J. Rosell, and I. Hussain, "Vision Language Action Models in Robotic Manipulation: A Systematic Review," arXiv preprint arXiv:2507.10672, July 2025. DOI: 10.48550/arXiv.2507.10672

[8] Z. Lou, K. Xu, Z. Zhou, and R. Xiong, "ExploreVLM: Closed-Loop Robot Exploration Task Planning with Vision-Language Models," arXiv preprint arXiv:2508.11918, August 2025. DOI: 10.48550/arXiv.2508.11918

[9] H. Wang, S. Karnik, B. Lim, and S. Bansal, "Using Vision Language Models as Closed-Loop Symbolic Planners for Robotic Applications: A Control-Theoretic Perspective," arXiv preprint arXiv:2511.07410, November 2025. DOI: 10.48550/arXiv.2511.07410

[10] S. Bai, X. Wang, Q. Gao, Y. Huo, X. Chen, C. Li, J. Liu, Z. Wu, and X. Li, "Qwen2.5-VL Technical Report," arXiv preprint arXiv:2502.13923, February 2025. DOI: 10.48550/arXiv.2502.13923

[11] J. Wu, J. J. Lim, H. Zhang, J. B. Tenenbaum, and W. T. Freeman, "Physics 101: Learning Physical Object Properties from Unlabeled Videos," Proceedings of the British Machine Vision Conference (BMVC), September 2016. DOI: 10.5244/C.30.39

[12] J. Bohg, K. Hausman, B. Sankaran, O. Brock, J. Kragic, S. Schaal, and G. S. Sukhatme, "Interactive Perception: Leveraging Action in Perception and Perception in Action," IEEE Transactions on Robotics, Vol. 33, No. 6, pp. 1273-1291, December 2017. DOI: 10.1109/TRO.2017.2721939

[13] O. Kroemer, S. Niekum, and G. Konidaris, "A Review of Robot Learning for Manipulation: Challenges, Representations, and Algorithms," Journal of Machine Learning Research, Vol. 22, No. 30, pp. 1-82, 2021. https://www.jmlr.org/papers/v22/19-804.html

[14] W. Chow, J. Mao, B. Li, D. Seita, V. Guizilini, and Y. Wang, "PhysBench: Benchmarking and Enhancing Vision-Language Models for Physical World Understanding," arXiv preprint arXiv:2501.16411, January 2025. DOI: 10.48550/arXiv.2501.16411

[15] R. Riochet, M. Y. Castro, M. Bernard, A. Lerer, R. Fergus, V. Izard, and E. Dupoux, "IntPhys: A Framework and Benchmark for Visual Intuitive Physics Reasoning," arXiv preprint arXiv:1803.07616, March 2018. DOI: 10.48550/arXiv.1803.07616

[16] P. J. Jeshmol and B. C. Kovoor, "Video Question Answering: A Survey of the State-of-the-Art," Journal of Visual Communication and Image Representation, Vol. 105, no. 104320, December, 2024. DOI: 10.1016/j.jvcir.2024.104320

[17] K. Yi, C. Gan, Y. Li, P. Kohli, J. B. Tenenbaum, and A. Torralba, "CLEVRER: Collision Events for Video Representation and Reasoning," The International Conference on Learning Representations (ICLR), 2020. DOI: 10.48550/arXiv.1910.01442

[18] A. Bakhtin, L. van der Maaten, J. Johnson, L. Gustafson, and R. Girshick, "PHYRE: A New Benchmark for Physical Reasoning," Advances in Neural Information Processing Systems (NeurIPS), Vol. 32, December 2019. DOI:10.48550/arXiv.1908.05656

[19] K. Chen, S. Xie, Z. Ma, and K. Goldberg, "Robo2VLM: Visual

Question Answering from Large-Scale In-the-Wild Robot Manipulation Datasets," Advances in Neural Information Processing Systems (NeurIPS), 2025. DOI: 10.48550/arXiv.2505.15517

[20] E. Zhao, V. Raval, H. Zhang, J. Mao, Z. Shangguan, S. Nikolaidis, Y. Wang, and D. Seita, "ManipBench: Benchmarking Vision-Language Models for Low-Level Robot Manipulation," arXiv preprint arXiv:2505.09698, May 2025. DOI: 10.48550/arXiv.2505.09698

[21] S. Bai, X. Wang, Q. Gao, Y. Huo, X. Chen, C. Li, J. Liu, Z. Wu, and X. Li, "Qwen3-VL Technical Report," arXiv preprint arXiv:2511.21631, November 2025. DOI: 10.48550/arXiv.2511.21631

[22] Z. Chen, Z. Wang, Y. Ding, J. Zhang, Y. Li, and L. Wang, "Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling," arXiv preprint arXiv:2412.05271, December 2024. DOI: 10.48550/arXiv.2412.05271

[23] J. Zhu, Y. Xu, Z. Wu, H. Lin, Y. Li, Y. Zhang, and J. Wang, "InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models," arXiv preprint arXiv:2504.10479, April 2025. DOI: 10.48550/arXiv.2504.10479

[24] W. Wang, Z. Wu, Y. Li, H. Lin, J. Zhu, and Y. Xu, "InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency," arXiv preprint arXiv:2508.18265, August 2025. DOI: 10.48550/arXiv.2508.18265

[25] A. Azzolini, C. De Souza, F. Khatami, M. G. Porfiri, A. Rosasco, and D. Jayaraman, "Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning," arXiv preprint arXiv:2503.15558, March 2025. DOI: 10.48550/arXiv.2503.15558

[26] E. Kerr, T. M. McGinnity, and S. A. Coleman, "Material Recognition Using Tactile Sensing," Expert Systems with Applications, Vol. 94, pp. 94-111, March 2018. DOI: 10.1016/j.eswa.2017.10.045

[27] W. Yuan, S. Wang, S. Dong, and E. H. Adelson, "Connecting Look and Feel: Associating the Visual and Tactile Properties of Physical Materials," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 4494-4502, July 2017. DOI: 10.1109/CVPR.2017.478

[28] K. Takahashi and J. Tan, "Deep Visuo-Tactile Learning: Estimation of Tactile Properties from Images," Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI), pp. 4780-4784, January 2021. DOI: 10.24963/ijcai.2020/665

[29] Z. Zhou, J. Song, X. Xie, Z. Shu, L. Ma, D. Liu, J. Yin, and S. See, "Towards Building AI-CPS with NVIDIA Isaac Sim: An Industrial Benchmark and Case Study for Robotics Manipulation," Proceedings of the 46th International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP '24), pp. 263-274, April 2024. DOI: 10.1145/3639477.3639740

[30] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, Feng Zhao. "Are We on the Right Way for Evaluating Large Vision-Language Models?," Advances in Neural Information Processing Systems (NeurIPS), 2024. DOI:10.52202/079017-0850

Authors

DongJu Jang received the B.S. degree in Computer Science from Yonsei University, Seoul, South Korea, in 2021. He is currently pursuing the M.S. degree in the Department of Artificial Intelligence at Yonsei University, Seoul, South Korea, since 2025. His research interests include physical AI, robotics, MM-LLM, and autonomous driving.

Yeong-In Lee received the B.S. degree in the Department of Business Administration from Sookmyung Women's University, Seoul, Korea, in 2018. She is currently pursuing an integrated M.S.-Ph.D. degree at the Graduate School of Information, Yonsei University, Seoul, Korea since 2023. Her current research interests include robot learning, computer vision and deep learning.

Ha-Young Kim is an Associate Professor at Graduate School of Information, Yonsei University, Korea. She received her Ph.D. degree at department of Mathematics, Purdue University, USA.

From 2011 to 2016, she was a research staff member in Samsung Advanced Institute of Technology (SAIT) of Samsung Electronics, Korea, working on various recognition systems with deep learning. Her primary research areas are deep learning and computational Finance. She has published in leading journals, including Information Fusion, Applied Soft Computing, Expert Systems with Applications, Stochastic and Dynamics, Computers in Biology and Medicine, PLoS ONE, automation in construction, Journal of Computing in Civil Engineering and Annals of Finance. She is the the inventor of 8 patents and 13 patent applications.