Chapter 01

로봇 동작 성능 평가를 위한 벤치마크 기술 동향

손영성_한국전자통신연구원 책임연구원
한효녕_한국전자통신연구원 책임연구원
이은서_한국전자통신연구원 책임연구원

본 고는 시각-언어-행동(VLA) 및 비디오 월드 모델의 발전에 따라 중요해진 로봇 동작 성능 평가와 최신 벤치마크 기술 동향을 분석한다. 최근 로봇 평가는 단순 이진 성공률을 넘어 실제-시뮬레이션 제어 격차(real-to-sim gap) 극복, 하위 물리 추론, 언어 강건성, 예측 영상의 물리적 실행 가능성을 다각도로 검증하는 방향으로 진화하고 있다. 이에 기존 벤치마크의 한계를 짚고, 고충실도 시뮬레이션 정렬을 이룬 REALM, 객관식(MCQ) 기반의 저비용 대규모 물리 추론 진단 도구 ManipBench, 생성 영상의 실행력을 검증하는 RoboWM-Bench, 명령어 의역(paraphrase) 취약성을 분석하는 LIBERO-Para, 통합 행동 언어 기반의 교차-형태(cross-embodiment) 일반화 평가 기술을 상호 비교한다. 결론적으로 향후 로봇 벤치마크는 초기 인지 진단부터 언어 과적합 점검, 실로봇 기반 품질 검증이 유기적으로 결합한 '다층 평가 스택' 구조로 발전할 것임을 전망한다.

I. 서론

최근 시각-언어-행동(Vision-Language-Action: VLA) 모델 및 비디오 월드 모델의 발전은 로봇이 자연어 지시를 이해하고 복잡한 물리적 행동을 수행할 수 있는 비약적인 성장을 이끌고 있다. 인터넷 규모의 방대한 데이터로 사전학습을 거친 로봇 파운데이션 모델들은 개념 이해와 장면 인식 면에서 과거와 비교할 수 없는 탁월한 성능을 보여주며 다양한 범용 조작의 가능성을 열고 있다[1]-[3].

* 본 내용은 손영성 책임연구원(☎ 042-860-4822, ysson@etri.re.kr)에게 문의하시기 바랍니다.
** 본 내용은 필자의 주관적인 의견이며 IITP의 공식적인 입장이 아님을 밝힙니다.
** 본 연구는 산업통상자원부 및 산업기술기획평가원(KEIT) 연구비 지원[P0028420, 고난도 로봇 조작이 필요한 다양한 산업 환경에서 범용적으로 활용 가능한 MMA(Multi-Modal Action) 파운데이션 모델 기술 개발], [RS-2024-00419641, 정밀 조립작업 대상 실환경 파라미터 반영된 로봇용 가상환경 플랫폼 개발]에 의한 연구임

이러한 기술적 진보에도 불구하고, 최신 로봇 AI 모델들을 실제 세계의 복잡한 환경에 곧바로 적용하는 것은 여전히 초기 단계에 머물러 있다. 많은 연구기관과 산업 현장에서는 “모델이 시각적으로 환경을 이해하는 것 같지만, 실제 물리적 제어 상황에서는 쉽게 무너진다”는 현실적 어려움을 호소한다. 특히, 실제 세계의 미세한 환경 변화나 조명, 지시문의 단순한 의역(paraphrase), 혹은 로봇 형태(embodiment)의 작은 차이 앞에서도 모델이 여전히 큰 성능 저하를 나타내어 신뢰성 있는 배포를 저해하고 있다.

이는 로봇의 지능을 평가하는 패러다임에 있어 단순히 특정 환경에서 단일 작업의 ‘성공 여부’만을 확인하는 개별 테스트 접근이 아니라, 모델의 인지적·물리적 한계를 해부하는 다층적이고 체계적인 평가 접근이 선행되어야 한다. 즉, 오늘날의 로봇 벤치마크는 단순한 모듈 평가를 넘어 다각도의 구조적 평가 체계가 필요하다. 그 이유는 다음과 같다.

첫째, 로봇이 직면하는 물리적 현실은 단일하지 않으며 수많은 교란 요인(perturbations)이 얽힌 복잡한 환경이다. 시각적(조명, 시점), 의미론적(지시어 변경), 행동적(물체 질량이나 크기 변화) 요인이 끊임없이 변하므로 모델은 환경 변화에 적응하는 일반화 능력을 증명해야 한다. 그러나 실제 환경에서 이를 대규모로 테스트하는 것은 비용과 재현성 문제로 한계가 있으므로, 시뮬레이션과 실제 세계의 제어(control alignment)를 정밀하게 정렬하여 시뮬레이션에서의 평가가 현실 성능의 신뢰할 수 있는 대리 지표가 되도록 하는 체계적인 접근이 필요하다.

둘째, 로봇 파운데이션 모델에 있어 조작은 “지시를 이해했다”는 표면적인 개념으로 끝나는 것이 아니다. 로봇은 환경과 상호작용하기 위해 파지 지점 선택, 물체 간의 물리적 역학 관계 인식, 유연체(deformable object)의 변형 예측 등 하위 수준의 물리적 추론(low-level reasoning)을 수행해야 한다. 이러한 정밀한 물리 세계 이해도를 직접 궤적을 실행하기 전 단계에서 선제적이고 저비용으로 진단할 수 있는 모듈형 평가 구조가 필요하다.

셋째, 로봇 평가 환경은 시각적 그럴듯함(perceptual realism)에 머물러서는 안 되며 물리적 실행 가능성(physical executability)을 철저히 검증해야 한다. 최근 영상 생성 기반 월드 모델이 로봇의 미래 궤적을 훌륭하게 예측해 내고 있지만, 시각적 충실도가 실제 물리 법칙이나 기하학적 제약과 항상 일치하지는 않는다. 따라서 생성된 계획이나 영상을 실제 동역학에 맞춰 실행해 보고, 단계별(step-level) 물리적 정합성을 확인하는 평가 루프가 필수적이다.

넷째, 자연어 지시를 따르는 로봇 AI는 인간과 매끄러운 협업을 지향한다. 그러나 동일한 의미의 명령어를 “스토브를 켜라”에서 “쿡탑을 켜라”로 바꾸었다고 해서 로봇이 작업 대상을 완전히 잃어버린다면 사용자는 로봇을 신뢰할 수 없다. 따라서 실패의 원인이 모터 제어의 오류인지, 객체 접지(object grounding)와 같은 계획 수준의 오해인지 그 병목 현상을 정확히 진단해 내는 언어적 강건성(paraphrase robustness) 평가 체계가 내재화되어야 한다.

마지막으로, 차세대 로봇 AI의 핵심이라 할 수 있는 파운데이션 모델들은 단일 하드웨어 플랫폼에 묶이지 않고 여러 이기종 로봇 간의 형태 일반화(cross-embodiment generalization) 능력을 요구받고 있다. 또한, 실제 산업 배포를 위해서는 단순히 컵을 옮겼다는 이진 성공률을 넘어, 얼마나 부드럽고 안전하게 이동했는지(smoothness & safety)를 따지는 품질 중심의 평가가 요구된다. 결과적으로 로봇 평가가 “다층적이고 체계적으로 접근해야 하는 이유”는, 실험실 내의 한정된 성공 시연을 넘어 실제 산업 환경의 변화에 대응하고 안전하게 배포될 수 있는 견고한 모델을 걸러내기 위함이다.

이러한 맥락에서 본 고의 II장에서는 로봇 동작 성능을 평가하는 기본 방법과 핵심 평가 지표의 진화 과정을 살펴본다. III장에서는 일반화, 하위 물리 추론, 영상 실행력, 언어 강건성을 세밀하게 측정하기 위해 최근 등장한 특화 벤치마크 기술(REALM, ManipBench, RoboWM-Bench, LIBERO-Para 등)의 특징을 소개하며, IV장에서는 이들 벤치마크 기술의 상호 비교를 통해 모델의 병목을 진단하는 입체적 평가 프레임워크를 분석하고, 마지막 장에서는 결론 및 향후 발전 방향을 제시한다.

II. 로봇 동작 성능 평가의 기본 방법

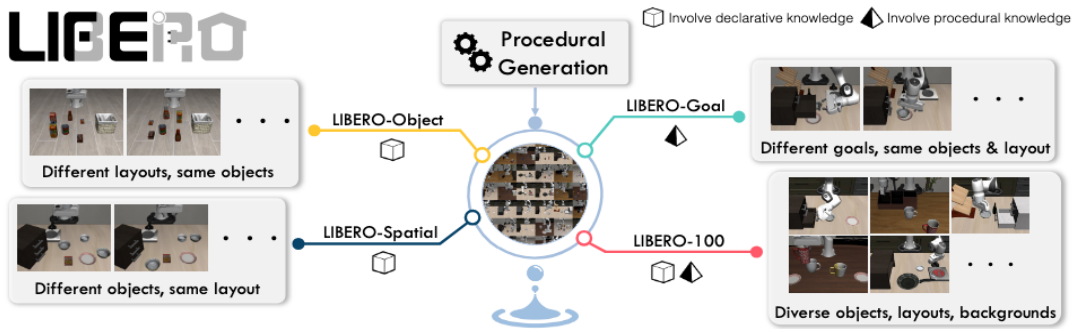
로봇 파운데이션 모델을 실제 환경에서 신뢰할 수 있게 배포하기 위해서는 모델의 동작 성능을 정확하게 측정하고 검증하는 평가 방법론이 필수적이다. 로봇 동작 성능 평가는 크게 실제 로봇을 활용한 실세계 평가(real-world evaluation)와 물리 엔진 기반의 시뮬레이션 평가(simulation-based evaluation)로 나뉘며, 평가 지표 역시 과거의 단순한 작업 완료 여부에서 점차 세부적으로 고도화되는 구조를 채택하고 있다. 이러한 평가 구조를 살펴보고, 기존 검증 과정에서 발생하는 병목 요인을 분석하는 것은 최신 벤치마크 기술을 이해하는 데 중요한 요소이다.

가장 궁극적이고 확실한 검증 방법은 실제 세계 기반 평가이다. 실제 로봇은 중력, 마찰, 조명 변화, 센서 잡음, 미세한 접촉 오차 같은 요소를 모두 겪기 때문에, 모델이 실제 환경에서 안전하고 안정적으로 동작하는지 확인하기 위해서는 실로봇 실험이 필요하다. 그러나 특정 환경에 국한되지 않은 모델의 범용성을 검증하기 위해 실제 환경에서 수백 번 이상의 반복 실험(rollout)을 수행하는 것은 비용과 시간이 기하급수적으로 소모되고 동일한 조건을 통제하고 재현하기 어렵다는 명확한 한계를 지닌다.

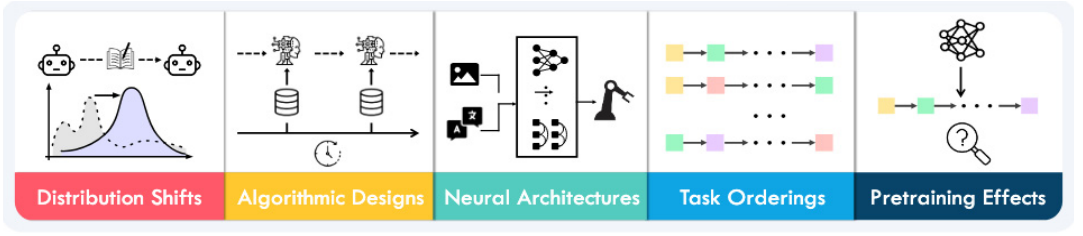
이러한 실세계 평가의 확장성과 재현성 문제를 극복하기 위해 물리 엔진에 기반한 시뮬레이션 평가가 대안이자 필수 요소로 자리 잡았다. 시뮬레이션 환경은 통제된 조건에서 방대한 환경 및 물체 조합을 저비용으로 테스트할 수 있다. 그러나 시뮬레이션 평가 역시 시각적 품질(visual fidelity)의 차이와 물리적 제어 매개변수의 불일치로 인해 발생하는, 이른바 '실세계-시뮬레이션 격차'(Real-to-Sim gap)라는 고질적인 병목이 존재하여 정책 성능이 왜곡될 수 있다.

초기 로봇 평가를 이끈 대표적인 기반 벤치마크 연구로는 RLBench, LIBERO, CALVIN 등이 있으며, 이들은 각기 다른 시뮬레이션 환경에서 평가 기준을 확립하는 구조를 가진다. 예를 들어, RLBench는 100개 이상의 다양한 조작 과제를 제공하며 모션 플래너 기반의 데모 생성을 지원해 모조 학습(imitation learning) 연구의 확고한 기반이 되었다[4]. 이와 함께 LIBERO는 절차적 지식과 선언적 지식의 전이를 분석하는 평생 학습(lifelong learning)에 초점을 맞춰 130개의 과제를 제공하였으며, CALVIN은 언어 지시에 따라 여러 기술을 연쇄적으로 구성하는 장기 지평(long-horizon) 조작 과제를 평가하는 방법을 제시하였다[5][6]. 이러한 시뮬레이션 벤치마크들과 더불어, 실제 환경(In-the-wild) 기반의 대규모 로봇 조작 데이터셋인 DROID 연구는 방대한 환경과 시각적 다양성을 갖춘 데모를 제공하여 로봇 파운데이션 모델 학습은 물론 평가 시 주요 하드웨어 플랫폼(embodiment)이자 필수적인 데이터 기반으로 널리 활용되고 있다[7]. 이들은 특정 조건과 학습 패러다임에서 강력한 기준점이 되어 왔지만, 현대 시각-언어-행동(VLA) 모델의 다양한 오류를 해부하기에는 평가 차원의 세분성이 부족한 한계가 있다.

이와 같이 평가 방식의 진화와 더불어 동작 성능을 측정하는 지표 구조도 고도화되고 있다. 전통적인 평가는 작업의 최종 완수 여부만을 따지는 이진 성공률(binary success rate)에 크게 의존해 왔다. 그러나 입력되는 자연어 지시나 이미지 패치의 변화에 따라 모델이



(a) 공간 · 물체 · 목표 · 혼합 지식 전이를 검증하는 4종의 절차적 작업 검증 기능



(b) LLDM 관련 5가지 핵심 연구 토픽으로 구성

〈자료〉 B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, et al., "LIBERO: Benchmarking Knowledge Transfer for Lifelong Robot Learning", Advances in Neural Information Processing Systems(NeurIPS), Vol.36, 2023, pp.44776-44791.

[그림 1] LIBERO 벤치마크 구성

어떤 과정에서 실패했는지 알기 어렵다. 이에 따라 최근에는 과제를 '접근, 파지, 들어올리기, 이동, 배치' 등의 세부 단계로 쪼개어 달성도를 0과 1 사이로 정량화하는 단계별 진행률(tiered progression) 방식이 설계되었다. 나아가 행동의 부드러움(smoothness)과 안전성, 효율성을 함께 평가하는 품질 중심의 지표가 도입되며 단순한 성공을 넘어서는 평가가 이루어지고 있다.

결과적으로, 기존의 벤치마크 구조와 이진 성공률 위주의 추론 평가 과정에서는 최신 파운데이션 모델이 지닌 세밀한 인지적, 물리적 병목 현상을 정확히 짚어내는 데 한계가 발생한다[그림 1]. 모델이 지시어의 언어적 변형에 취약한 것인지, 미세한 역학을 추론하지 못하는 것인지, 아니면 시뮬레이션과 실세계의 제어 격차 때문인지를 단순히 '실패'로만 처리하게 된다. 이러한 병목은 모델 규모와 작업 복잡도가 증가할수록 더욱 두드러지며, 실시간성과 신뢰성이 요구되는 로봇 응용에서 중요한 제약 요소가 된다.

따라서 범용 로봇 검증을 위해서는 모터 제어의 물리적 실행력을 확인하는 시뮬레이션 평가 효율을 개선하는 기술과 궤적 실행 이전의 인지적 추론 및 언어 과적합을 최적화하여

측정하는 특화 기술이 함께 고려되어야 하며, 이를 다층적으로 결합한 새로운 벤치마크 연구가 활발히 진행되고 있다.

III. 최신 벤치마크 기술 종류와 특징

본 고에서 소개하는 차세대 벤치마크 기술들은 크게 물리적 추론 능력을 측정하는 시스템, 실제 작업 실행력을 검증하는 구조 그리고 언어적 강건성을 정밀 타격하는 프레임워크로 분류된다. 이들 최신 벤치마크는 단순한 연산 최적화 평가를 넘어, 인지 · 물리 · 언어적 차원에서 모델의 세부 병목 구간을 다각도로 진단할 수 있는 정교한 평가 구조를 갖춘 것이 특징이다.

1. REALM: 실제-시뮬레이션 정렬 기반 범용성 평가 벤치마크

REALM(Real-to-Sim Validated Benchmark)[8]은 기존 시뮬레이션 환경들이 고충실도 시각 자료와 정밀한 제어 정렬을 제공하지 못해 시뮬레이션에서의 평가 결과가 실제 로봇의 성능을 왜곡한다는 문제에 주목하여, 실제 세계와의 상관관계를 극대화한 고충실도 시뮬레이션 평가 구조를 제안한다. 기존 방식에서는 범용 VLA 모델이 낯선 환경이나 물체에 대해 가지는 일반화(generalization) 능력을 검증하려면 실제 환경에서 수백 번 이상의 반복 실험이 필요하여 막대한 비용과 시간 낭비가 발생했다.

이는 3,500개 이상의 방대한 물체와 7가지 조작 기술을 지원하는 물리 엔진 환경을 구축하고, 실제 로봇 제어 장치와 동역학적 매개변수를 정밀하게 정렬(aligned control)하는 구조를 사용한다. 특히, 시각(visual), 의미(semantic), 행동(behavioral)의 3가지 카테고리에 걸쳐 '조명 변화, 물체 질량 변경, "바나나"를 "껍질을 벗길 수 있는 물체"로 지시어 변경' 등 15가지 교란 요인(perturbations)을 체계적으로 도입하였다. 더불어 단순한 이진 성공률을 넘어 작업을 '접근 → 파지 → 들어올리기 → 이동' 등으로 세분화해 0.0에서 1.0 사이로 정량화하는 단계별 진행률(tiered progression) 지표를 도입했다.

이러한 구조는 평가의 해상도를 높이고, 시뮬레이션의 실세계 대리(proxy) 신뢰도를 비약적으로 높인다. 실험 결과, 약 800쌍의 실제 및 시뮬레이션 궤적을 비교했을 때 강력한 피어슨

상관관계(r>0.88)를 달성하여 실제 성능을 높은 정확도로 예측했다. 이 결과는 평가 환경의 시각적, 물리적 정렬 관리가 대규모 평가의 유효성을 결정짓는 핵심 요소임을 보여준다. REALM은 시뮬레이터를 통한 스트레스 테스트와 다차원 교란을 결합함으로써, 고비용의 실세계 실험 없이도 VLA 모델의 물리적 범용성을 극대화하여 진단하는 시스템 수준 검증 기법이다.

2. ManipBench: VLM 기반 하위 수준 물리적 추론 평가

ManipBench[9]는 최근 GPT-4V나 Gemini와 같은 범용 비전-언어 모델(Vision-Language Model: VLM)을 로봇의 하위 수준 계획기(low-level planner)로 활용하려는 시도가 많음에도, 궤적(trajectory) 기반 시뮬레이션 실행 비용이 너무 많이 들어 대규모 모델 비교가 어렵다는 문제에 주목하여 객관식(Multiple Choice Question: MCQ) 기반의 진단형 평가 구조를 제안한다. 기존 방식에서는 궤적 실행(rollout)을 위해 매 스텝 무거운 물리 엔진 연산을 거쳐야 하므로 평가 시간이 기하급수적으로 증가했다.

이는 로봇이 환경과 상호작용을 하기 위해 파악해야 할 접촉 지점이나 이동 방향을 시뮬레이터에서 직접 실행하는 대신, 이미지 상에 시각적 마커(keypoint)나 그리드를 표시(mark-based prompting)한 뒤 4지선다형 질문으로 하는 구조를 사용한다. 기존의 단순한 집기/놓기를 넘어, 로봇이 다루기 까다로운 관절 물체(서랍 등), 유연 물체(천 펴기/접기), 도구 사용, 동적 조작(공 던지기)까지 5가지 넓은 태스크 카테고리에 걸쳐 12,617개의 방대한 시나리오를 구성했다.

이러한 구조는 물리 시뮬레이션 루프를 제거하여 평가 속도를 압도적으로 높이면서도 모델의 근본적인 물리적 이해도를 측정한다. 실험 결과, 무거운 궤적 실행 없이도 33개에 달하는 VLM 모델을 고속으로 평가할 수 있었으며, MCQ 정답률이 실제 실세계 조작 성공률과 강력한 상관관계(피어슨 0.889)를 맺음을 입증했다. 이 결과는 모터 제어 실행 이전 단계의 시각적·공간적 추론 능력을 고립시켜 진단하는 것이 매우 효율적인 평가 방식임을 보여준다. ManipBench는 복잡한 물리 연산을 배제하고 이미지 프롬프트를 활용함으로써, 방대한 최신 모델들이 로봇 제어에 필요한 '하위 수준 지능'을 갖추었는지 대규모로 스크리닝하는 고효율 사전 진단 기법이다.

3. RoboWM-Bench: 비디오 월드 모델의 물리적 실행 가능성 평가

RoboWM-Bench[10]는 최근 Sora, Veo, Wan 등 비디오 생성 월드 모델이 예측한 미래 영상이 시각적으로는 매우 현실적일지라도, 로봇의 기하학적 왜곡이나 접촉 오류로 인해 실제 물리 엔진에서는 실행 불가능(physical executability)한 경우가 많다는 환각(hallucination) 문제에 주목하여 실행 기반 검증 구조를 제안한다. 기존 벤치마크들은 생성된 영상의 프레임 간 일관성이나 화질 등 '시각적 그럴듯함'(perceptual plausibility)만을 점수화하여, 그것이 로봇 제어로 전이될 수 있는지를 측정하지 못했다.

이는 월드 모델이 생성한 인간 손 조작이나 로봇 조작 영상을 역동역학 모델(Inverse Dynamics Model: IDM)과 3D 리타겟팅 기술을 이용해 로봇의 관절 궤적(action sequence)으로 변환한 뒤, 실제 세계를 4D 가우시안 등으로 재구성한 고충실도 시뮬레이터(Real2Sim) 환경 내에서 직접 실행해 보는 폐루프(Closed-loop) 검증 구조를 사용한다. 단순히 최종 성공만 보는 것이 아니라, 접근, 파지 등의 각 단계가 동역학적으로 가능한지 점검하는 단계별 지표를 결합했다.

이러한 구조는 시각적 자연스러움의 이면에 가려진 물리적 허점을 명확히 파헤친다. 평가 결과, 기존 품질 평가 지표(PAI-Bench)로는 모든 모델이 포화된 높은 점수를 기록했으나, 실제 RoboWM-Bench로 검증했을 때는 장기 지평(Long-horizon)이나 유연체 과제에서 실행 성공률이 급격히 저하됨을 입증했다. 이 결과는 단순한 비디오 생성 품질이 아니라 동역학 구조의 보존이 모델의 물리적 유효성을 결정짓는 핵심 요소임을 보여준다. RoboWM-Bench는 행동 변환과 물리 엔진 실행을 결합함으로써, 월드 모델이 다운스트림 로봇 제어를 위한 진정한 데이터 엔진으로 기능할 수 있는지를 확인하는 실행 중심 평가 기법이다.

4. LIBERO-Para: 언어적(의역) 강건성 진단

LIBERO-Para[11]는 자연어 명령을 따르는 VLA 모델이 훈련 데이터에 존재하는 소수의 명령어 형식에만 과적합(overfitting)되어 동일한 의미가 있는 다른 표현(패러프레이즈)이 주어지면 작업에 완전히 실패하는 문제에 주목하여 언어적 의역 강건성 진단 구조를 제안한다. 기존 로봇 평가는 새로운 물체나 조명 등 시각적 일반화에만 집중하여 언어 변화에 따른 오류 원인을 진단하기 어려웠다.

이는 10가지 작업에 대해 어휘적(동의어), 구조적(구문 변화), 맥락적(간접 화법) 등 행동 축과 객체 축을 변형시키는 4,092개의 방대한 의역 지시문을 시뮬레이션 평가에 도입하는 구조를 사용한다. 나아가 모델이 단순히 원래 지시문과 텍스트가 유사한 문장에서만 성공하여 강건성이 과대평가되는 것을 막기 위해, 의역의 난이도와 언어적 구조 거리에 가중치를 부여하는 'PRIDE'(로봇 지시 편차의 패러프레이즈 견고성 지수)라는 정량적 지표를 고안했다.

이러한 구조는 로봇 모델이 명령어를 표면적으로 암기했는지, 진정으로 이해했는지를 세밀하게 분해한다. 실험 결과, 수십억 파라미터의 VLA 모델조차 "스토브를 켜라"를 "쿡탑을 켜라"로 객체 단어만 바꾸었을 때 성공률이 22.8~51.9%포인트 범위로 급락했다. 실패 궤적 분석 결과, 이러한 실패 중에서 79.5~95.5%가 모터 제어 문제가 아니라 대상을 완전히 착각하는 '의미론적 계획(Far-GT) 오류'임이 밝혀졌다. 이 결과는 단순한 데이터 크기 확장이 아니라, 언어 공간과 물리 공간을 연결하는 의미론적 객체 접지(object grounding)가 모델 신뢰성을 결정짓는 핵심 병목임을 보여준다. LIBERO-Para는 의역 난이도 가중치와 궤적 분석을 결합함으로써, VLA 모델의 표면적 키워드 매칭 한계를 정교하게 진단하는 언어-행동 평가 기법이다.

5. Being-H0.5(교차-형태 일반화 평가)

Being-H0.5[12] 평가는 로봇 파운데이션 모델이 특정 기구학(kinematics)이나 센서 구성, 하드웨어에 종속되어 여러 로봇 간의 전이가 어렵다는 문제에 주목하여 평가 패러다임 변화의 관점에서 교차-형태(cross-embodiment) 일반화 검증 구조를 제안한다. 기존 방식에서는 단일 로봇 플랫폼 위주로 평가가 이루어져, 모델을 다른 로봇에 배포할 때마다 개별적인 엔지니어링이나 해킹(hacks)이 수반되는 비효율이 발생했다. 이는 단일 로봇의 제어에 의존하지 않고, 하나의 통합된 행동 언어 공간(unified action language)을 통해 서로 다른 로봇 형태 사이에서도 모델이 일관되게 동작할 수 있는지를 교차 검증하는 구조를 사용한다. 평가 프로토콜은 공간 이동, 장기 지평(long-horizon), 양팔 협업(bimanual) 등을 포함하는 10가지 실제 실로봇 작업과 시뮬레이션(LIBERO, RoboCasa 등)을 오가며 단일 제너럴리스트(Generalist) 체크포인트의 성능을 측정한다.

이러한 구조는 특정 하드웨어에 대한 과적합을 방지하고, 행동 인터페이스의 범용성을 엄격하게 시험한다. 실험 결과, 통합 기반 모델은 훈련에서 직접 짝지어 본 적 없는 태스크와

로봇 형태(task-embodiment pairs)의 조합에서도 적절한 다단계 행동 구조를 초기화해 내며, 시뮬레이션(LIBERO 기준 98.9%)과 5개의 상이한 실제 로봇 환경 양쪽에서 전문가(specialist) 모델에 근접하는 높은 성공률을 유지했다. 이 결과는 단순한 시각-언어 데이터의 확장을 넘어, 행동 제어의 통일된 추상화 관리가 하드웨어 제약을 뛰어넘는 모델 성능을 결정짓는 핵심 요소임을 보여준다. Being-H0.5의 평가 방식은 단일 도메인 한계를 허물고, 범용 로봇 지능의 배포 안정성을 평가하는 교차 배포 최적화 프토로콜이다.

IV. 최근 벤치마크 기술 상호 비교

로봇 파운데이션 모델이 직면한 다양한 추론 및 실행 병목을 정확히 파악하기 위해, 최근 학계에서는 평가 대상과 측정 방식이 특화된 벤치마크 기술들을 연이어 선보이고 있다.

[표 1] 로봇 벤치마크 기술 상호 비교

비교 항목	REALM	ManipBench	RoboWM-Bench	LIBERO-Para
핵심평가 대상	VLA 모델의 시각/물리/의미적 일반화 능력	VLM의 하위 수준(low-level) 물리 및 동작 추론 능력	생성형 비디오 월드 모델의 실제 물리적 실행 가능성	VLA 모델의 언어적(의역) 강건성과 과적합 진단
평가 방식	15가지 교란 환경이 정렬된 고충실도 시뮬레이터 내 직접 궤적 실행	이미지상 마커(Keypoint) 기반 객관식 질문(MCQ) 형식으로 궤적 없이 평가	생성 비디오를 역동역학(IDM) 등을 통해 행동으로 변환 후 시뮬레이션 실행	4,000+ 개의 의역 지시문을 부여하여 시뮬레이션 내 작업 궤적 실행
해결하려는 기존 한계	낮은 시각/물리 정합성(Real-to-Sim Gap) 및 실세계 반복 평가 비용 문제	대규모 정책 실행(Rollout)에 소모되는 엄청난 연산 비용과 검증 시간 문제	시각적으로만 그럴듯해(Perceptual Plausibility) 보이는 월드 모델의 환각 현상	제한된 지시어에만 동작하는 언어 과적합 및 표면적 키워드 매칭 문제
특화된 데이터 영역	3,500개 객체, 조명/질량 변화 등 15가지 스트레스 테스트 조건	유연체(천), 관절 물체, 동적 조작 등 5개 범주의 12,600+ 시나리오	인간 손 영상 리타겟팅 및 양팔, 유연체 조작을 포함한 다양한 로봇 영상	동의어, 간접 화법 등 4,000개 이상의 세분화된 언어 변형 세트
핵심평가 지표	단계별 진행률(Tiered Progression), 실제 성능과의 상관계수	객관식 정답률(Accuracy), 인간 기준선 대비 성능 비교	단계/과제 수준 물리적 실행 성공률(Execution Success)	의역 견고성 지수(PRIDE), 동적 시간 왜곡(DTW) 기반 계획 vs 실행 오류 분석

〈자료〉 M. Sedlacek, P. Yefanov, G. Ponimatkin, J. Bardhan, et al., "REALM: A Real-to-Sim Validated Benchmark for Generalization in Robotic Manipulation", arXiv preprint arXiv:2512. 19562, 2025.
E. Zhao, V. Raval, H. Zhang, J. Mao, et al., "ManipBench: Benchmarking Vision-Language Models for Low-Level Robot Manipulation", Conference on Robot Learning(CoRL), 2025
F. Jiang, Y. Chen, K. Xu, Y. Liu, H. Wang, et al., "RoboWM-Bench: A Benchmark for Evaluating World Models in Robotic Manipulation", arXiv preprint arXiv:2604.19092, 2026.
C. Kim, M. Kim, M. Kang, H. Kim, D. Jung, "LIBERO-Para: A Diagnostic Benchmark and Metrics for Paraphrase Robustness in VLA Models", arXiv preprint arXiv:2603.28301, 2026. 재구성

[표 1]은 앞서 III장에서 다룬 기술 중 엄밀한 의미의 최신 '전용 벤치마크 스위트' 4가지(REALM, ManipBench, RoboWM-Bench, LIBERO-Para)만을 중심으로 모델의 진단 및 평가 효율성을 비교하기 위해 구성되었다. 참고로 III장에서 함께 소개된 Being-H0.5의 경우, 독립적으로 규격화된 전용 벤치마크라기보다는 이기종 로봇 간의 일반화(cross-embodiment) 프로토콜을 보여주는 새로운 평가 패러다임 및 접근법으로서의 성격이 강하므로 본 비교표에서는 제외하였다. 각 벤치마크의 특징 및 상호 비교 분석은 다음과 같다.

1. 실행 비용 및 대규모 평가 효율성

ManipBench는 대규모 로봇 파운데이션 모델 평가 시 수백~수천 번의 궤적 실행을 요구하여 GPU 연산 비용과 검증 시간이 기하급수적으로 증가한다는 문제에 주목하여, 객관식 질문(MCQ) 기반의 진단형 평가 구조를 제안한다. 기존 방식에서는 로봇이 매 스텝마다 환경과 상호작용하며 물리 엔진의 연산을 거쳐야 하므로, 모델 구조나 파라미터가 다를 때마다 막대한 평가 오버헤드가 발생하여 동시 비교가 가능한 모델 수를 제한하는 주요 요인으로 작용한다.

이는 복잡한 시뮬레이터 연동이나 긴 궤적 실행 없이 로봇 관측 이미지 위에 시각적 마커를 오버레이한 뒤 모델에 이미지와 MCQ 질문만을 던져 파지점이나 이동 방향 정답을 유추하게 하는 구조를 사용한다. 이를 통해 각 모델은 무거운 폐루프(closed-loop) 제어 없이 정적 이미지 처리 한 번만으로 로봇의 물리적 상호작용에 대한 이해도를 평가받게 된다.

이러한 구조는 평가의 시간적·비용적 파편화를 크게 줄이고, 실질적인 대규모 벤치마킹을 가능하게 한다. 실험 결과, 33개에 달하는 다양한 VLM 모델 계열을 12,600개 이상의 시나리오에 대해 압도적인 속도로 평가할 수 있었다. 반면, 실제 모델의 동작 끝단까지 물리적 가능성을 확인해야 하는 REALM이나 RoboWM-Bench는 역동역학 모델(IDM) 변환 및 고충실도 시뮬레이션 기반의 궤적 실행 연산이 필수적으로 실시간 처리 가능 수준이다[1]. 이 결과는 단순한 성공률 최적화뿐 아니라, 평가 환경의 궤적 실행 여부가 시스템 전체의 평가 스케일과 속도를 결정짓는 핵심 요소임을 보여준다. ManipBench는 복잡한 물리 연산을 배제하고 이미지 프롬프팅을 결합함으로써, 방대한 최신 모델들을 실제 배포 전에 압도적인 속도로 스크리닝하는 고효율 초기 진단 기법이다. 반면, 실행 기반 벤치마크는 최종 배포

1) 예를 들어, RoboWM-Bench의 IDM 추론은 A800 GPU 1대로 5초 분량 1280x720, 30fps 비디오 처리 시 약 1.5초가 소모된다.

직전에 선택된 모델을 꼼꼼히 검증하는 용도로 사용되어 두 평가 시스템은 상호 보완적인 역할을 한다.

2. 물리 세계와의 일치도

REALM과 RoboWM-Bench는 시뮬레이션 평가 환경이 실제 물리 세계(real-to-sim)의 시각적 복잡성과 제어 응답성을 제대로 반영하지 못해 '시뮬레이션에서 성공이 실제의 성공을 보장하지 못한다'는 시뮬레이션-실세계 격차(real-to-sim gap) 문제에 주목하여 고도의 정렬 기법을 도입한 시뮬레이터 구조를 제안한다. 기존 방식에서는 단순히 가상 자산(에셋) 위주로 구성된 환경을 사용하므로, 실제 로봇 카메라로 들어오는 조명 변화나 로봇 제어기의 미세한 마찰력 차이로 인해 정책의 성능이 크게 왜곡되거나 과대평가된다.

REALM은 3,500개 이상의 현실적인 3D 물체를 활용해 고충실도(high-fidelity) 시각 자료를 렌더링하면서 실제 로봇의 동역학 매개변수와 제어기를 정밀하게 정렬(aligned control)하는 구조를 취한다. RoboWM-Bench 역시 단순 시뮬레이션 에셋이 아닌, 실제 세계 장면을 4D 가우시안 등의 기술로 가상 환경에 역재구성(scene reconstruction)하고 이를 통해 생성 비디오의 물리적 실행 가능성을 시뮬레이터와 교차 검증한다. 이를 통해 평가 모델은 연속된 가상 공간에서도 실제와 동일한 마찰, 충돌, 기하학적 제약을 경험하게 된다.

이러한 정밀한 물리 정렬 구조는 평가 결과의 왜곡을 방지하고 시뮬레이션 결과의 실세계 전이 신뢰도를 극대화한다. 실험 결과, REALM은 시뮬레이션에서의 단계별 진행률 평가 결과가 실제 로봇의 동작 결과와 강력한 피어슨 상관관계($r>0.88$)를 맺도록 물리 엔진 최적화에 성공하여 '시뮬레이션이 실제를 신뢰성 있게 대체할 수 있음'을 입증했다. RoboWM-Bench 또한 재구성된 시뮬레이션 환경에서 실제 성공 및 실패 궤적을 실행했을 때 100%의 결과 일관성을 보였다.

이 결과는 단순 3D 에셋의 확장을 넘어, 시각적 렌더링과 물리적 제어 매개변수의 동기화가 가상 평가 환경의 무결성을 결정짓는 핵심 요소임을 보여준다. REALM과 RoboWM-Bench는 실세계 재구성 기술, 제어 매개변수 최적화, 객체 동역학을 결합함으로써, 물리 세계에서 감수해야 할 값비싼 수천 번의 실험을 신뢰도 하락 없이 안전하게 대체하는 시스템 수준의 인프라 기술이다.

3. 지능의 병목(bottleneck) 식별

LIBERO-Para와 ManipBench는 로봇 파운데이션 모델 평가 시, 작업 실패라는 결과값(성공률 0%) 안에 제어 역학(실행)의 실패와 자연어 명령 오해(인지/계획)의 실패가 혼재되어 있어 지능의 진정한 병목을 진단하기 어렵다는 문제에 주목하여 인지적 추론 과정을 해부하는 진단 평가 구조를 제안한다. 기존 이진 성공률 지표는 작업이 중단된 경우 로봇이 대상 물체의 질량을 잘못 계산해서 놓친 것인지, 아니면 동의어로 바뀐 지시어를 아예 다른 물체로 착각한 것인지 파악하지 못해 모델 구조 개선에 한계를 갖는다.

LIBERO-Para는 동일한 작업을 지시하되 4,000개 이상의 방대한 의역(paraphrase) 지시문을 도입하고, 실패한 궤적을 동적 시간 왜곡(Dynamic Time Warping: DTW) 알고리즘으로 분석하여 실패가 목표 궤적 근처에서 발생했는지(Near-GT: 실행 오류) 아예 엉뚱한 곳으로 향했는지(Far-GT: 계획 오류)를 분류하는 구조를 취한다. ManipBench는 모터 제어를 아예 배제하고 객관식 질문(MCQ) 형태만 취하여, VLM이 공간적, 물리적 역학 관계(예; 유연체 변형이나 도구의 접촉점) 자체를 이해하고 있는지만 독립적으로 분리하여 묻는다.

이러한 구조는 단일한 '실패' 결과를 세밀하게 분해하여 모델의 숨겨진 지능 결함을 명확히 노출시킨다. 실험 결과, LIBERO-Para는 조작 실패의 79.5~95.5%가 모터 제어 오류가 아닌 지시어 변경(예; 스토브 → 쿡탑)에 따른 '의미론적 객체 접지(object grounding) 실패' 즉 계획 수준의 오류임을 세밀하게 진단했다. ManipBench 또한 모델의 '하위 수준 물리 지능'(low-level reasoning)이 유연체 및 동적 조작에서 아직 인간 기준선에 크게 못 미친다는 것을 명확한 객관식 정답률 차이(예; 유연체 상호작용 추론 한계)로 수치화하여 보여준다.

이 결과는 단순 제어 성공률 측정 최적화뿐 아니라, 언어적 의미망과 하위 물리 역학의 이해도를 분리하여 측정하는 것이 차세대 파운데이션 모델의 아키텍처 개선 방향을 결정짓는 핵심 요소임을 보여준다. LIBERO-Para와 ManipBench는 의역 난이도 가중치 부여, 실패 궤적 분석, 텍스트-물리 매칭을 결합함으로써, 모델이 단순히 데이터를 암기(과적합)했는지 아니면 진정으로 범용적 물리 지능을 획득했는지를 해부하는 세부 인지 진단 기법이다.

V. 결론

본 고에서는 로봇 파운데이션 모델의 신뢰성 있는 성능 검증을 위해 기존 평가 방법의 한계를 짚어보고, 인지적 · 물리적 병목 요인을 진단하는 최신 벤치마크 기술을 분석하였다. 기존 벤치마크 구조에서 이진 성공률 지표는 작업 실패의 구체적 원인을 파악하기 어려우며, 시뮬레이션 환경은 시각적 품질과 물리적 제어 매개변수의 불일치(real-to-sim gap)로 인해 실제 성능을 왜곡하는 한계가 있다. 이러한 문제를 해결하기 위해 다각적인 측면을 진단하는 특화 벤치마크 기술이 제안되고 있다. 물리적 실행력 및 일반화 측면에서는 REALM, Robo WM-Bench 등이 15가지 교란 요인 기반의 고충실도 시뮬레이션 정렬과 역동역학 모델(IDM) 기반 행동 변환을 통해 실제 물리 법칙과의 일치도를 높이는 방향으로 발전하고 있다. 인지 및 추론 진단 측면에서는 ManipBench, LIBERO-Para 등이 객관식 문항(MCQ)과 방대한 의역(paraphrase) 지시문을 도입하여, 궤적 실행 이전에 모델의 하위 물리 추론 한계와 언어적 과적합 오류를 저비용으로 식별하는 접근을 보인다. 전반적으로 신뢰할 수 있는 로봇 지능 평가를 위해서는 모델의 언어 접지, 물리 추론, 실행 제어 능력을 통합적으로 고려하는 다층 평가 스택(multi-layered evaluation stack)을 구성하는 것이 중요하다. 특히, 시각-언어-행동(VLA) 및 비디오 월드 모델이 고도화될수록 생성된 영상의 시각적 그럴듯함(perceptual plausibility)과 실제 물리적 실행 가능성 사이의 간극으로 인한 병목이 더욱 심화될 가능성이 크다. 따라서 향후에는 초기 저비용 인지 진단부터 고충실도 일반화 검증, 더 나아가 교차-형태(cross-embodiment) 배포와 행동 품질 검증을 함께 반영한 입체적인 표준 평가 프레임워크 구축이 핵심 연구 방향이 될 것으로 전망한다.

참고문헌

[1] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, et al., “RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control”, Conference on Robot Learning(CoRL), 2023.

[2] M. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, et al., “OpenVLA: An Open-Source Vision-Language-Action Model”, Conference on Robot Learning(CoRL), 2024.

[3] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, et al., “π0: A Vision-Language-Action Flow Model for General Robot Control”, arXiv preprint arXiv:2410.24164, 2024.

[4] S. James, Z. Ma, D. R. Arrojo, A. J. Davison, “RLBench: The Robot Learning Benchmark and Learning Environment”, IEEE Robotics and Automation Letters(RA-L), 2020.

[5] B. Liu, Y. Zhu, C. Gao, et al., “LIBERO: Benchmarking Knowledge Transfer for Lifelong Robot Learning”, Advances in Neural Information Processing Systems(NeurIPS), Vol.36, 2023.
[6] O. Mees, L. Hermann, E. Rosete-Beas, W. Burgard, “CALVIN: A Benchmark for Language-Conditioned Policy Learning for Long-Horizon Robot Manipulation Tasks”, IEEE Robotics and Automation Letters(RA-L), Vol.7, No.3, 2022, pp.7327-7334.
[7] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, et al., “DROID: A Large-Scale In-The-Wild Robot Manipulation Dataset”, Robotics: Science and Systems(RSS), 2024.
[8] M. Sedlacek, P. Yefanov, G. Ponimatkin, J. Bardhan, et al., “REALM: A Real-to-Sim Validated Benchmark for Generalization in Robotic Manipulation”, arXiv preprint arXiv:2512. 19562, 2025.
[9] E. Zhao, V. Raval, H. Zhang, J. Mao, et al., “ManipBench: Benchmarking Vision-Language Models for Low-Level Robot Manipulation”, Conference on Robot Learning(CoRL), 2025.
[10] F. Jiang, Y. Chen, K. Xu, Y. Liu, H. Wang, et al., “RoboWM-Bench: A Benchmark for Evaluating World Models in Robotic Manipulation”, arXiv preprint arXiv:2604.19092, 2026.
[11] C. Kim, M. Kim, M. Kang, H. Kim, D. Jung, “LIBERO-Para: A Diagnostic Benchmark and Metrics for Paraphrase Robustness in VLA Models”, arXiv preprint arXiv:2603.28301, 2026.
[12] BeingBeyond Team, “Being-H0.5: Scaling Human-Centric Robot Learning for Cross-Embodiment Generalization”, BeingBeyond, 2026.