The Clean Paper

WHAT THE PAPER SAYS · WHAT IT DOESN'T · WHY IT MATTERS

큰 로봇'파운데이션 모델'은 정말 더 잘 작동할까? 신중한 답은'조금 그렇다'—그리고 대부분의 연구는 구분하기 어렵다

Toyota Research Institute 는 약 *1,700* 시간의 다양한 조작 데이터로 사전학습한'대규모 행동 모델'을 만들고, 처음부터 학습한 단일작업 정책과 이례적으로 엄격하게 비교했다. 블라인드·무작위·대규모 시험 *(*실제 약 *1,800* 회, 시뮬레이션 *47,000* 회 이상*)* 과 통계를 사용한 결과, 작업별 미세조정 뒤 큰 모델은 평균적으로 더 잘했고 약 *3~5* 배 적은 작업별 데이터가 필요했으며 조건 변화에도 더 강했다. 하지만 미세조정 없이는 일관된 우위가 없었고, 일부 효과는 큰 표본이 아니면 보이지 않을 만큼 작았으며, 평범한 데이터 정규화 선택이 아키텍처보다 중요하기도 했다. 로봇 파운데이션 모델 방향에 대한 절제된 지지이지 범용 로봇, 제로샷 일반화, '창발적 도약'은 아니다. 동시에 로봇공학의 상당 부분이 잡음을 측정하고 있을 수 있다는 경고다.

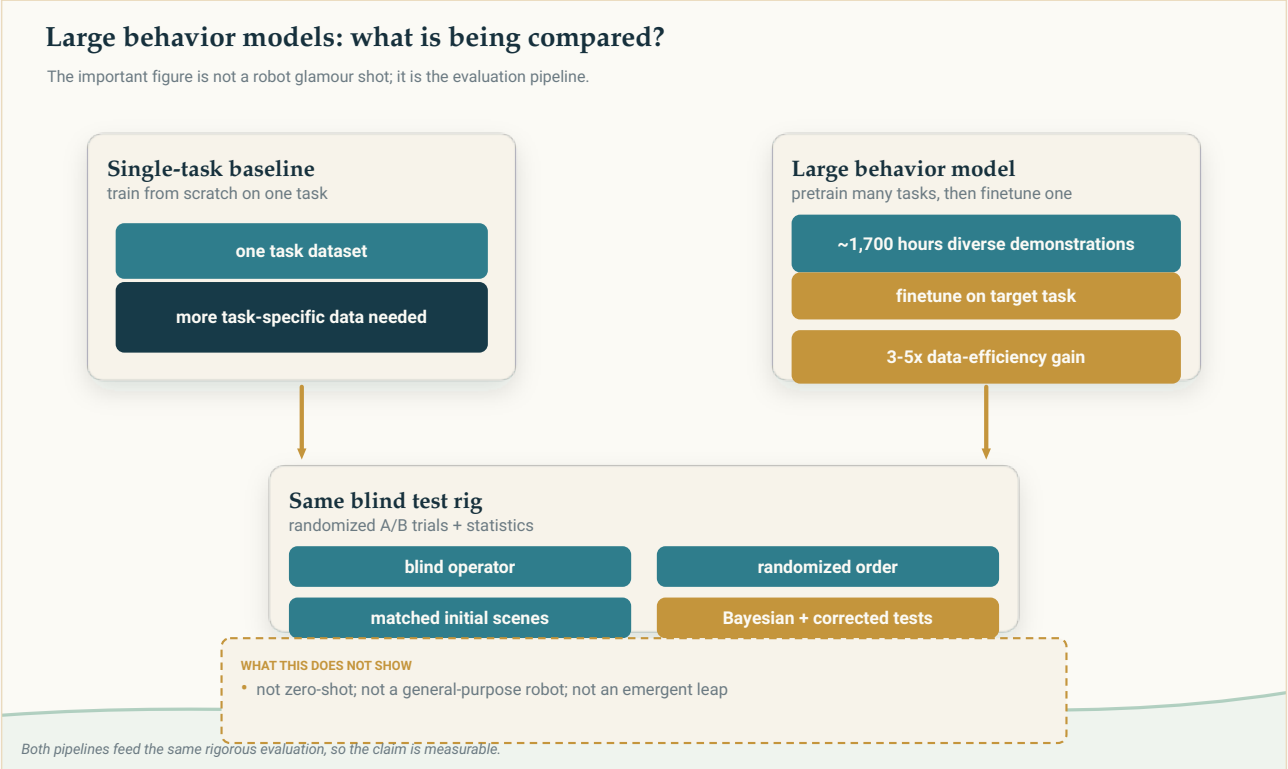
Written by Lucio Vaglio · Cross-checked by Laura Nesso · Edited by Michele Renda · Figures by Laura Nesso and Nova Vela · AI-assisted, human-reviewed

Paper: *A Careful Examination of Large Behavior Models for Multitask Dexterous Manipulation*, Toyota Research Institute Large Behavior Model Team —J. Barreiros, A. Beaulieu, et al.; senior authors incl. R. Ambrus, B. Burchfiel, S. Feng, H. Kress-Gazit (Cornell), R. Tedrake, Science Robotics (2026); preprint arXiv:2507.05331 · DOI: 10.1126/scirobotics.aea6201

사실 진짜 주제가' 정직함' 인 로봇 논문

로봇공학은 지금 **파운데이션 모델 (foundation model)** 열풍 한가운데 있다. 언어·이미지 AI 에서 빌려온 아이디어는 매혹적이다. 로봇에게 작업 하나씩 따로 가르치는 대신, 방대하고 다양한 시범 데이터로 하나의 큰 **"대규모 행동 모델 (large behavior model, LBM)"** 을 훈련해 폭넓은 능력과 빠른 적응성을 얻자는 것이다. 기대도 투자도 엄청나다. 그러면 헤드라인은 쉽게 나온다. 범용 로봇 두뇌가 왔다.

Toyota Research Institute 의 이 논문이 흥미로운 이유는 정확히 그 헤드라인을 거부하기 때문이다. 진짜 기여는 더 화려한 로봇이 아니다. 겉보기에는 단순한 질문을 매우 엄격하게 들여다본다. 큰 모델 방식이 실제로 더 잘 작동하는가? 그리고 그걸 대체 어떻게 알 수 있는가? 저자들 스스로 말하듯 이 분야에서는 보기 드문 수준의 통계적 주의를 기울여 답했다. 결과는 실제로 유용한 "그렇다, 하지만" 이고, 로봇공학 연구의 상당 부분이 잡음을 측정하고 있을 수 있다는 경고다.



두 접근법 모두 같은 눈가림·무작위 평가를 거친다. 논문의 주장은 로봇 파운데이션 모델이 제로샷 범용 지능이라는 것이 아니라, 사전학습이 신중하게 측정했을 때 데이터 효율과 강건성을 높일 수 있다는 것이다.

Original The Clean Paper diagram CC BY 4.0

연구진이 한 일

연구진은 LBM 을 구체적인 의미로 만들었다. **확산 기반 시각-운동 정책**으로, 카메라 영상과 짧은 언어 지시, 로봇 자신의 관절 위치를 읽고 10 Hz 로 짧은 운동 명령 묶음을 내놓는 Diffusion Transformer 다. 이 모델은 사내에서 모은 500 개 이상의 서로 다른 작업과 공개 데이터셋을 합쳐 약 **1,700 시간의 로봇 시범 데이터**로 **사전학습**한 뒤, 개별 작업에 맞게 **미세조정**했다. 모든 비교 대상은 오직 그 한 작업의 데이터만

으로 **처음부터 훈련한 단일작업 정책**이다.

하지만 논문의 중심은 모델보다 평가다. 저자들은 평가 자체를 핵심 결과처럼 다룬다. 스스로를 속이지 않기 위해 다음을 사용했다.

- 실제 환경에서 **눈가림 무작위 A/B 테스트**를 했다. 로봇을 운영하는 사람은 어떤 정책을 시험 중인지 몰랐고 순서도 무작위였다.
- **통제되고 반복 가능한 초기 조건**을 만들었다. 매 시험 전에 운영자가 화면에 겹쳐 보이는 기준 이미지에 맞춰 장면을 배치했다.
- **많은 시험 횟수**를 사용했다. 실제 환경에서는 작업·정책·조건당 50 회, 시뮬레이션에서는 작업당 200 회 실행했다. 총합은 약 **1,800 회의 눈가림 실제 로봇 실행과 47,000 회가 넘는 시뮬레이션 실행**이다.
- **제대로 된 통계**를 사용했다. 겹치는 오차 막대를 눈대중으로 비교하는 대신 성공 확률의 베이지안 추정과 다중비교 보정을 포함한 쌍별 가설검정을 했다. 사람이 성공 여부를 판정한 시험의 4 분의 1 에는 별도의 품질보증 검사를 해 채점 오류 자체도 측정했다.

[video pending]

아침 식탁을 차리는 모델의 나란한 비교. 왼쪽은 처음부터 학습한 단일작업 기준 모델이고, 오른쪽은 LBM 이다. 두 영상 모두 1 배속이다. 평가한 작업 하나를 보여주는 것이지 범용 자율성을 증명하는 영상은 아니다.

바로 이 평가 장치가 핵심이다. 논문 전체는 이런 장치 없이는 진짜 개선과 운을 구분할 수 없다는 주장이다.

무엇을 발견했나

미세조정한 큰 모델은 처음부터 학습한 단일작업 모델보다 평균적으로 더 잘했다. 여러 작업을 묶어 보면, 사전학습한 뒤 해당 작업에 맞게 미세조정한 LBM 이 같은 작업 데이터로 처음부터 학습한 정책보다 시뮬레이션과 실제 환경 모두에서 일관되게 더 높은 성능을 냈고 차이는 통계적으로 유의했다. 개별 작업에서도 미세조정 LBM 은 거의 모든 경우 처음부터 학습한 모델과 통계적으로 동등하거나 더 좋았다. 실제 환경 3/3 작업, 시뮬레이션 15/16 작업이다.

가장 크고 분명한 이점은 데이터 효율이다. 미세조정 LBM 은 작업별 데이터가 대략 **3~5 배 적어도** 처음부터 학습한 모델과 비슷한 성능에 도달했다. 실제 작업 하나인 아침 식탁 차리기에서는 전체 시범의 **15%** 만으로 미세조정한 LBM 이 **100%** 데이터를 사용해 처음부터 학습한 정책보다 더 잘했다.

조건이 달라질수록 사전학습의 도움이 커졌다. 시험 환경을 의도적으로 훈련 조건에서 벗어나게 만든 **분포 이동 (distribution shift)** 에서 미세조정 LBM 의 우위가 더 커졌다. 한 시뮬레이션 세트에서는 정상 조건에서 16 개 작업 중 3 개에서만 처음부터 학습한 모델을 통계적으로 이겼지만, 분포 이동에서는 16 개 중 10 개에서 이겼다. 실제 배포 환경은 늘 훈련 조건에서 조금씩 벗어나므로, 이 강건성이 실용적으로는 가장 중요한 결과일 수 있다.

사전학습 데이터가 많을수록 성능은 매끄럽게 좋아졌다. 데이터를 늘릴수록 성능은 꾸준히 상승했고, 시험한 규모에서는 **갑작스러운 도약이나 "창발적" 점프가 없었다.** 유용하고 예측 가능하며 드라마틱하지 않은 결과다.

하지만 미세조정 없는 범용 모델 이야기는 성립하지 않았다. 사전학습 LBM 을 작업별 미세조정 없이 제로샷으로 사용했을 때 단일작업 정책을 일관되게 이기지 못했다. 하나의 네트워크가 여러 작업을 수행할 수는 있었지만 "그냥 말로 지시하면 된다" 는 꿈은 여기서는 확인되지 않았다. 저자들은 비교적 작은 언어 인코더의 취약성도 일부 원인으로 본다.

그리고 이득은 놓치기 쉬울 만큼 작았다 — 혹은 가짜로 만들어내기 쉬울 만큼. 여러 효과는 평소보다 큰 표본과 신중한 통계 검정을 사용해야만 보였다. 저자들은 효과 크기와 잡음을 고려할 때 **많은 로봇공학 논문이 통계적 잡음을 측정하고 있을 위험이 상당하다**고 직접 쓴다. 또 평범해 보이는 선택인 데이터 정규화 방식이 아키텍처 변경보다 결과에 더 큰 영향을 줬고, 사전학습의 정규화 **버그**는 평가가 모두 끝난 뒤에야 발견됐다.

이것이 아마 뜻하는 것

방어 가능한 해석은 이렇다. 다양하고 대규모인 로봇 데이터로 사전학습하는 것은 실제로 가치 있는 구성요소다. 새 작업마다 필요한 데이터를 줄이고, 현실이 훈련 환경과 정확히 일치하지 않을 때 정책을 더 튼튼하게 만든다. 이 분야가 투자하고 있는 방향을 진짜로 지지하는 결과다. 그러나 이득은 중간 규모이고 조건부다. 주로 미세조정 뒤에 나타나며, 여러 작업을 합쳐 보거나 스트레스 조건에서 가장 분명하다. 즉시 투입 가능한 범용 로봇의 등장은 아니다.

더 조용하지만 중요한 의미는 방법론에 있다. 이 논문은 사실상 평가 기준을 다시 세우는 역할을 한다. 로봇 정책에 대해 신뢰할 수 있는 주장을 하려면 실제로 어느 정도의 증거가 필요한지 보여주고, 출판된 연구의 과도한 기대 상당 부분이 너무 적은 자료 위에 서 있을 수 있음을 암시한다. 이 분야에는 또 하나의 모델보다 이런 교정이 더 필요할지 모른다.

이것이 증명하지 않는 것

- **범용 로봇이 아니다.** 특정 아키텍처인 확산 정책을 작업마다 미세조정해, 원격조작 시범으로 학습하고 통제된 환경에서 보여준 결과다. 아무 새 작업이나 명령만 받으면 자율적으로 수행하는 로봇이 아니다.
- **제로샷 사용을 검증하지 않는다.** 미세조정이 없으면 큰 모델은 단일작업 기준 모델을 일관되게 이기지 못했다.
- **"창발적 도약" 의 증거가 아니다.** 규모를 키우자 성능이 매끄럽게 개선됐다. 어느 순간 갑자기 능력이 생겼다는 서사를 뒷받침할 불연속은 없다.
- 수치는 **상대적이고 실험실에 묶여 있다.** 절대 성공률은 비교 민감도를 높이기 위해 일부러 약 50% 근처가 되도록 작업 난이도를 맞췄다. 실제 환경 신뢰성의 측정치가 아니며, 한 연구실의 한 아키텍처에 대한 결과다.
- 어떤 정책이 왜 성공하거나 실패하는지 **설명하지 않는다**. 큰 모델이 오히려 더 못한 몇몇 작업도 보고됐지만 이유는 밝혀지지 않았다.
- 평가 장치 밖의 **안전, 자율성, 배포에 대해서는 아무것도 말하지 않는다.**

근거는 얼마나 탄탄한가?

중앙 비교 주장 — 미세조정 LBM 이 집계 결과에서 처음부터 학습한 기준 모델을 이기고, 몇 배 적은 데이터가 필요하며, 분포 이동에서 더 강건하다는 것 — 에 대한 증거는 강하고 이례적으로 잘 통제되어 있다. 눈가림, 무작위, 큰 표본, 통계 검정, 채점 품질보증까지 갖췄다. 방법론이 충분히 탄탄해 핵심 헤드라인은 거의 그대로 받아들여도 되는 드문 경우다.

정직한 단서들은 저자들 자신이 제시한다. 오차 범위에는 평가의 무작위성은 들어가지만 **훈련 자체의 무작위성**은 들어가지 않는다. 같은 모델을 두 번 훈련해도 의미 있게 다른 정책이 나올 수 있고 그 변동은 통계에 반영되지 않았다. 실제 환경 작업당 50 회 시험은 중간 크기 효과를 잡기에는 충분하지만 작은 효과를 놓칠 수 있다. 언어 조건부 입력에는 비교적 소형 인코더를 사용했기 때문에 "그냥 로봇에게 말하면 된다" 는 문제는 더 큰 시스템에서 달라질 수 있다. 그리고 평가 뒤에 발견된 정규화 버그도 솔직하게 공개됐다. 어느 것도 핵심 결과를 무너뜨리지는 않지만, 바로 이런 문제를 분야가 흔히 밀어놓는다고 논문은 주장한다.

출처에 대해서도 같은 정신으로 한 가지 적는다. 이 해설은 저자들의 프리프린트를 바탕으로 했다. 저널 출판본을 확보하지 못했으므로 프리프린트와 최종 출판 텍스트 사이에 변경이 있었는지는 확인하지 못했다.

왜 중요한가

"로봇 파운데이션 모델" 은 과장을 부르기 좋은 표현이고, 이런 연구는 어느 방향으로든 쉽게 오독할 수 있다. 승리의 "됐다!" 로도, 냉소적인 "과장이다" 로도 읽을 수 있다. 더 정확한 해석이 둘보다 유용하다. 다양한 데이터의 사전학습은 실제로 측정 가능한 중간 정도의 이득, 특히 작업별 데이터 절감과 강건성 향상을 주며, 규모를 키울수록 예측 가능한 방식으로 좋아진다.

더 깊은 이유는 논문이 자기 분야 자체에 엄격함을 들이댄다는 데 있다. 진짜 효과가 허술한 평가에서는 사라질 만큼 작다는 점, 그리고 데이터 정규화 같은 지루한 선택이 영리한 새 아키텍처보다 더 큰 영향을 줄 수 있다는 점을 보여줌으로써, 로봇 학습의 많은 진전이 믿을 만해지려면 측정부터 더 튼튼해야 한다고 말한다. 결과와 기대를 엄격히 구분하려는 논문은 순위표 1 위를 노리는 논문보다 드물고, 더 가치 있을 때가 있다.

핵심 요약

Toyota Research Institute 연구진은 약 1,700 시간의 다양한 조작 데이터로 사전학습한 확산 기반 로봇 정책인 "대규모 행동 모델" 을 만들고, 처음부터 학습한 단일작업 정책과 비교했다. 평가는 이례적으로 엄격했다. 눈가림·무작위·대규모 시험 (실제 환경 약 1,800 회, 시뮬레이션 47,000 회 이상) 과 제대로 된 통계를 사용했다. 작업별 미세조정 뒤 큰 모델은 집계상 안정적으로 더 잘했고, 작업별 데이터가 **약 3~5 배 적어도** 비슷한 성능을 냈으며, **조건이 바뀔 때 더 강건했다**. 사전학습 데이터를 늘리면 성능은 매끄럽게 상승했다. 하지만 **미세조정 없이** 사용하면 단일작업 모델을 일관되게 이기지 못했고, 일부 효과는 큰 표본이 아니면 보이지 않을 정도로 작았으며, 평범한 데이터 정규화 선택이 아키텍처보다 더 큰 영향을 주기도 했다. 로봇 파운데이션 모델 방향을 신중하게 지지하는 결과이지 **범용 로봇도, 제로샷 일반화 모델도, "창발적 도**

약" 도 아니다. 동시에 로봇공학 연구의 상당 부분이 잡음을 측정하고 있을 수 있다는 날카로운 경고다.

과장 없이 보면

논문이 보여주는 것: 약 1,800 회의 실제 로봇 실행과 47,000 회 이상의 시뮬레이션 실행을 포함한 엄격한 눈가림·통계적 평가에서, 다중작업 사전학습 뒤 미세조정한 확산 정책 (LBM) 이 집계상 처음부터 학습한 단일작업 정책보다 더 잘하고, 약 3~5 배 적은 작업별 데이터로 동등한 성능에 도달하며, 분포 이동에 더 강건했다. 사전학습 데이터가 늘수록 성능은 매끄럽게 증가했다.

그럴듯하지만 입증되지 않은 것: 이런 이점이 훨씬 큰 비전-언어-행동 모델에도 그대로 옮겨갈 것인지. 여기서는 언어 인코더가 작았다. 또 시험한 데이터 범위를 넘어도 매끄러운 스케일링이 계속될지 여부.

보여주지 않는 것: 범용 또는 제로샷 로봇 (미세조정 없이는 일관된 우위가 없음), 어떤 "창발적" 능력 점프, 특정 작업 실패의 원인, 절대적인 실제 환경 신뢰성 (비교 민감도를 위해 성공률을 약 50% 로 맞춤), 안전이나 자율 배포에 관한 어떤 결과.

주요 한계: 통계가 평가 무작위성은 반영하지만 훈련 실행 간 무작위성은 반영하지 않음. 실제 작업당 50 회로 작은 효과를 놓칠 수 있음. 한 연구실의 한 아키텍처. 비교적 작은 언어 인코더. 평가 후 데이터 정규화 버그 발견. 분석은 프리프린트 기반이며 출판본은 확인하지 못함.

일반 독자는 어느 정도 확신해도 되는가? 다중작업 사전학습 + 미세조정이 실제로 중간 정도의 이점, 특히 데이터 효율과 강건성 향상을 주며 이를 이례적으로 신중하게 측정했다는 데에는 높은 확신을 가져도 된다. 이것이 **범용 또는 제로샷 로봇도 아니고 창발적 도약도 아니라는 것**에도 높은 확신이 필요하다. 더 큰 모델로 얼마나 확장될지는 중간 정도의 확신만 가능하다. 그리고 표본이 부족한 연구가 잡음을 결과로 보고할 수 있다는 저자들의 경고는 진지하게 받아들일 가치가 있다. 적절한 태도는 접근법에는 절제된 낙관을, 이런 수준의 통계적 뒷받침이 없는 로봇 AI 결과에는 건강한 회의를 갖는 것이다.

출처

arXiv:2507.05331
https://arxiv.org/abs/2507.05331

Peer-reviewed version (not retrieved) —Science Robotics, 10.1126/scirobotics.aea6201
https://doi.org/10.1126/scirobotics.aea6201

Project page
https://toyotaresearchinstitute.github.io/lbm1/