The Clean Paper

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

大型機器人「基礎模型」真的更好嗎？一個謹慎的回答——適度地是的，而大多數研究無法判斷

豐田研究所訓練了「大型行為模型」——在約 *1,700* 小時多樣化操作資料上預訓練的機器人策略——並以不尋常的嚴謹性將其與從零開始的單任務策略進行了對比：盲法、隨機化、大樣本量試驗（約 *1,800* 次真實世界，*47,000+* 次模擬）並使用真實統計。經過逐任務微調後，大模型平均表現更好，需要大約少 *3–5* 倍的任務特定資料，並且在條件變化時更穩健；效能隨預訓練資料量平穩上升。但在沒有微調的情況下，它們並未一致地擊敗單任務模型，若干效應小到只有大樣本量才能揭示，而且一個平凡的資料標準化選擇比架構影響更大。這是對機器人基礎模型方向的有節制支持——並非通用機器人，不是零樣本萬能者，也不是「湧現式的飛躍」——加上一條尖銳的警告：機器人學中的許多研究可能正在測量雜訊。

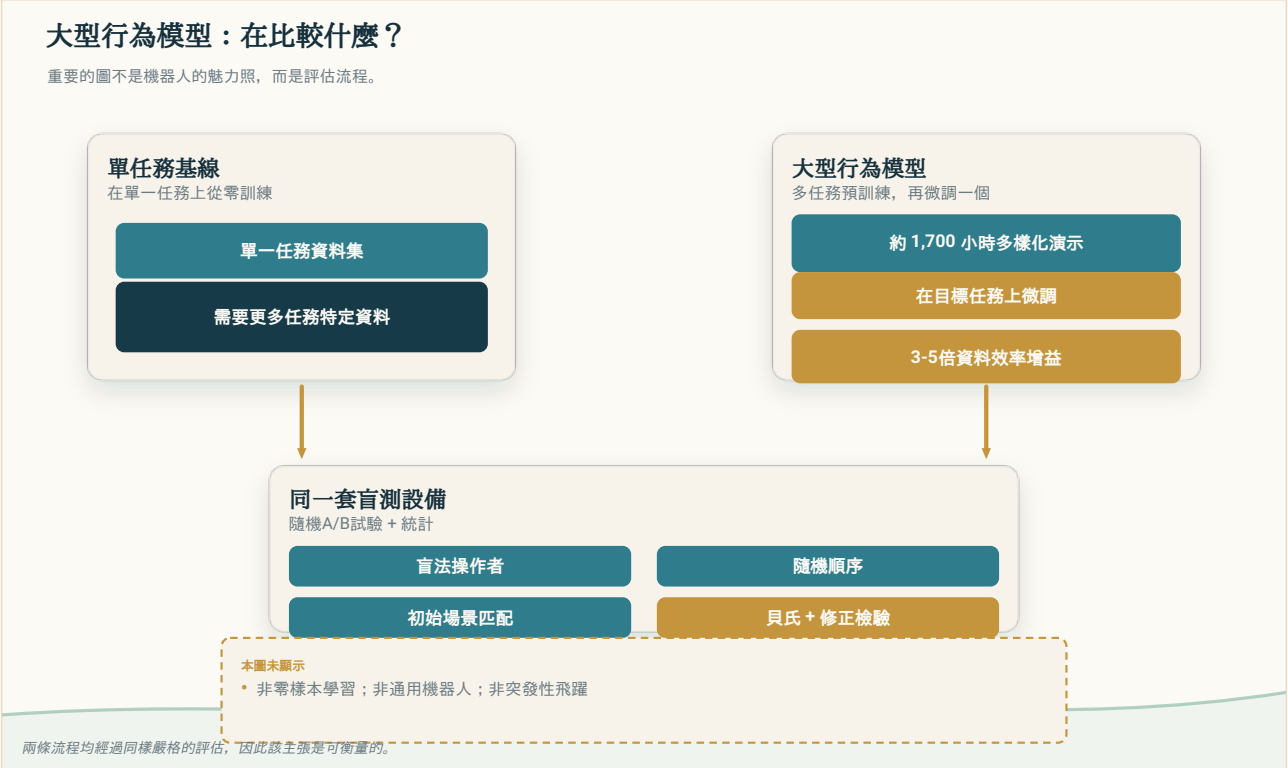
Written by Lucio Vaglio · Cross-checked by Laura Nesso · Edited by Michele Renda · Figures by Laura Nesso · AI-assisted, human-reviewed

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

一篇關於機器人的論文，其真正主題是誠實

機器人學正處於「基礎模型」熱潮之中。這個想法借鑑自語言和視覺 AI，非常誘人：與其一次為一個任務訓練一個機器人，不如在一個龐大、多樣化的演示資料集上訓練一個大型「**大型行為模型**」（**LBM**），從而得到一個能力廣泛且能快速適應的系統。熱情——以及投資——是巨大的。標題不言自明：通用機器人大腦來了。

這篇來自豐田研究所的論文之所以有趣，恰恰是因為它拒絕寫出這樣的標題。它真正的貢獻不是一個更花俏的機器人。而是對一個看似簡單的問題——大模型方法真的更好嗎？我們又該怎麼知道？——的一次嚴厲審視，並以在這個領域據作者自述不同尋常的統計嚴謹性作出了回答。其結果是「是的，但是」——一條有用的結論，以及一則警告：機器人學中的許多研究可能在測量雜訊。



兩種方法都進入相同的盲法隨機化評估。論文的主張並非機器人基礎模型是零樣本萬能者，而是預訓練可以透過仔細的測量提高數據效率和穩健性。

作者做了什麼

他們以一種具體而明確的方式構建了 LBM：**基於擴散的視覺運動策略**（一種擴散 Transformer，讀取相機影像、簡短的語言指令和機器人自身的關節位置，並以 10 Hz 輸出短促的運動指令）。這些策略在**約 1,700 小時**的機器人演示資料上進行了**預訓練**——涵蓋內部收集的 500 多個不同任務以及公共資料集——然後對單個任務進行**微調**。全文的比較對象始終是一個**從零開始訓練的單任務策略**。

然而，論文的核心是評估，作者將其視為主要成果。為了避免自欺欺人，他們使用了：

- **真實世界中的盲法、隨機化 A/B 測試**——操作機器人的人不知道正在測試哪種策略，測試順序也被隨機化。
- **受控、可重複的初始條件**——操作員在每次試驗前將場景與影像疊加層對齊。
- **大樣本量**——每個任務、每種策略、每種條件下進行 50 次真實世界 rollout；模擬中每個任務進行 200 次。總計：約 **1,800 次盲法真實世界 rollout 和超過 47,000 次模擬 rollout**。
- **恰當的統計方法**——成功機率的貝氏估計，以及帶有多重比較校正的配對假設檢定，而非僅憑目測重疊的誤差條。他們甚至對四分之一的人工評分試驗進行了品質保證通行，以測量評分誤差。

[video pending]

設置早餐桌的模型並排對比：（左）單任務基線，（右）LBM。兩個影片均以 1 倍速度播放。這是一個被評估的任務，而非通用自主性的證明。

這套機制才是重點。整篇論文都在論證：沒有它，你無法將真正的改進與運氣區分開來。

他們發現了什麼

經微調的大型模型平均而言優於從零開始的單任務模型。在所有任務上聚合，經過預訓練然後微調的 LBM 在模擬和真實世界中均可靠地優於同一任務上從零開始訓練的策略，且這種差異具有統計顯著性。在單個任務上，微調 LBM 在幾乎所有情況下均統計上達到同等或更優（真實世界任務 3/3，模擬任務 15/16）。

最大且最清晰的收益在於數據效率。一個微調 LBM 使用大約**少 3–5 倍**的任務特定數據即可達到從零開始相當的效能。在一個真實世界任務（設置早餐桌）中，一個僅在 **15%** 演示數據上微調的 LBM 擊敗了在 **100%** 數據上訓練的單任務策略。

當條件發生變化時，預訓練幫助最大。當測試環境發生變化（例如，不同的照明或物體位置）時，微調 LBM 比從零開始的策略保持了更多的效能。

收益隨預訓練數據量平穩增長。將預訓練數據從幾小時擴展到全部 1,700 小時，獲得了持續而平滑的收益，沒有飽和的跡象。沒有突然的跳躍，沒有「湧現」的門檻——只是一條持續向上的曲線。

一個單調的資料選擇比其他任何因素都更重要。在撰寫本文期間，他們發現訓練數據中的一個標準化錯誤——一個影響輸入編號方式的錯誤——正在損害從零開始的效能，使其看起來比應有的更差。修復後，兩者之間的差距縮小了，儘管 LBM 仍然勝出。這一點之所以至關重要，不是因為結果被否定了，而是因為它凸顯了該領域測量的脆弱性：如果一個數據前處理的錯誤可以在信號中造成真正模型差異大小的起伏，那麼許多已發表的效果在沒有這類嚴格基礎設施支撐的情況下，很可能無法重現。

沒有微調就沒有優勢

一個關鍵的否定性結果：**在沒有微調的情況下，大模型並未一致地超越從零開始的模型**。一個僅在廣泛資料集上預訓練、沒有特定任務微調的 LBM，在真實世界中的表現並不比一個對該任務從零開始訓練的策略更好。這項工作中沒有出現零樣本萬能者。LBM 是一種更好的起點，而不是一個更好的成品。

這沒有證明什麼

- 它**沒有**展示一個通用機器人或零樣本自主性。無微調 = 無優勢。
- 它**沒有**評估安全性、可靠性或部署就緒性。成功是二元的且經過調校，使得區分度最大化（使成功/失敗接近 50/50），而非最大化絕對成功。
- 它**沒有**將系統描述為能執行人類無法處理的任務。
- 它**沒有**證明「湧現」或任何能力上的不連續跳躍。
- 它**沒有**提供超出所研究設置範圍的可泛化性證據——沒有工業部署、沒有家庭實驗、沒有長期自主性。

證據有多強？

對平均改進，證據是強有力的。真實世界和模擬中的聚合效應量在統計上顯著且方向一致。嚴格的評估協議給機器人學領域帶來了不尋常的可信度。

對單個任務，即使在 50 次試驗中，效應也較小，一些信賴區間重疊，且在某些案例中，從零開始的策略在數值上略勝一籌。論文對此是透明的；這與聚合收益的穩健性並不矛盾。

對數據效率，收益是清晰且可操作的。3–5 倍的減少是一個實質性的實用收穫——少寫数据——即使模型能力本身只是適度提高。

對可擴充性，證據是平滑且單調的，但僅基於一個模型的規模（從約 200 小時擴展到 1,700 小時預訓練數據）。沒有驗證「越大越好」是否會在更大尺度上繼續成立。

對標準化效應，教訓是既謙卑又重要的：論文最大的效應量之一被追溯至一個數據錯誤，修復後該效應縮小了。這不是方法論上的失敗——這正是那種在事後應被稱讚的透明度——但它確實表明，該領域許多已發表的結果可能無法在這種程度的審查下倖存。

為什麼重要

「機器人基礎模型」是一個為誇大宣稱而生的短語，而像這樣的研究很容易在任一方向上被誤讀——作為「它奏效了 !」的凱旋，或是「它被過度炒作了」的輕蔑。準確認知比兩者都更有用：在多樣化數據

上進行預訓練能帶來真實、可測量但適度的收益——主要是每任務需要更少的數據，以及更強的穩健性——而且這條路徑隨規模預測性地改善。

它更深層的原因在於這篇論文將其嚴謹性對準了它自己的領域。透過展示真實的效應量小到在不嚴謹的評估下會消失，並且像數據標準化這樣一個乏味的選擇可能壓倒一個聰明的新架構，它提出了一個理由：機器人學習進展中的很大一部分需要更紮實的測量才能被相信。一篇花掉自己的可信度來警戒線區分結果與期望的論文，正在做一些比在排行榜上登頂更罕見、也更有價值的事情。

清潔摘要

豐田研究所的研究人員訓練了「大型行為模型」——基於擴散的機器人策略，在約 1,700 小時多樣化操作數據上進行了預訓練——並使用一套異常嚴格的協議將其與從零開始的單任務策略進行了對比：盲法、隨機化、大樣本量（約 1,800 次真實世界和 47,000+ 次模擬試驗），並進行真實統計。經過逐任務微調後，大模型在聚合中可靠地表現更好，需要大約**少 3–5 倍**的任務特定數據，並且在**條件變化時更穩健**，效能隨預訓練數據的增加而平滑提高。但**在沒有微調的情況下**，它們並未一致地擊敗單任務模型，若干效應量小到只有大樣本量才能揭示，而且一個平凡的數據標準化選擇比架構影響更大。這是對機器人基礎模型方向堅實而有節制的支持——**並非**一個通用機器人，不是零樣本萬能者，也不是一次「湧現的飛躍」——加上一條尖銳的警告：機器人學中的許多研究可能正在測量雜訊。

No-BS 檢查

論文所展示的：透過一次嚴謹、盲法、統計有力的評估（約 1,800 次真實世界 + 47,000+ 次模擬 rollout），經過多任務預訓練然後微調的擴散策略（LBM）在聚合中優於從零開始的單任務策略，使用大約少 3–5 倍的任務特定數據即可達到同等效能，並且在分佈偏移下更穩健；效能隨預訓練數據平滑增長。

合理但未經證實的：這些收益可遷移到更大的視覺-語言-動作模型（他們的語言編碼器很小）；平穩的增長趨勢繼續超出所測試的數據範圍。

它沒有展示的：通用或零樣本機器人（無微調 → 無一致優勢）；任何「湧現」式的能力跳躍；對特定任務層面失敗的解釋；絕對意義上的真實世界可靠性（成功率被調校至接近 50% 以獲得靈敏度）；關於安全性或自主部署的任何內容。

主要侷限：統計捕捉了評估隨機性，但未捕捉訓練執行的隨機性；每個任務 50 次真實世界試驗可能遺漏小效應；單一架構和單一實驗室；適中的語言編碼器；在評估後發現了一個數據標準化錯誤；分析基於預印本（未檢查已發表版本）。

普通讀者應該抱有多大信心？高度信心——多任務預訓練加微調能帶來真實、適度的收益——特別是在數據效率和穩健性方面——並且這些收益得到了異常仔細的測量。高度信心——這**不是**一個通用或零樣本機器人，也**不是**一次湧現式的飛躍。中等信心——這些收益在多大規模上可擴展至更大的模型。值得認真對待：作者自己的警告，該領域的效應量已經小到統計力度不足的研究可能正在報告雜訊。恰當的立場：對該方法持有節制的樂觀，以及對缺乏這類統計基礎的機器人 AI 結果保持健康的

懷疑。

來源

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/