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——答案是肯定的，但幅度有限，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）並採用正式統計檢驗。經過逐任務微調後，大模型平均表現更好，只需原來的約 *1/3* 至 *1/5* 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。

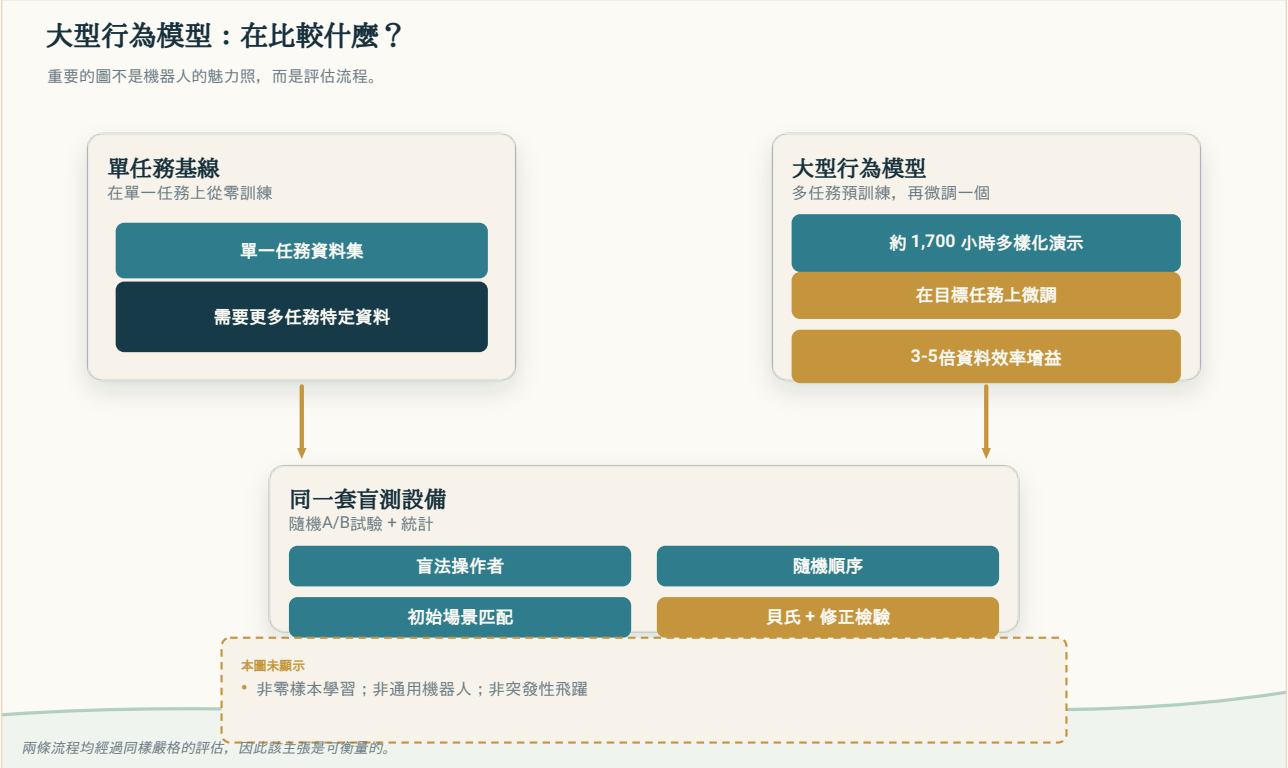
Written by Lucio Vaglio · Cross-checked by Laura Nesso · Edited by Michele Renda · Figures by Laura Nesso and Nova Vela · AI-assisted, human-reviewed

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

一篇真正主題是誠實的機器人論文

機器人學正處於一股「基礎模型」熱潮之中。這個借鑑自語言與影像 AI 的想法很吸引人：與其一次只為一項任務訓練機器人，不如利用大量且多樣的示範，訓練一個大型的「**大型行為模型**」（**LBM**），打造出能力廣泛、適應迅速的系統。熱情——以及投資——都十分高漲。頭條幾乎已經替自己寫好了：通用型機器人的大腦來了。

這篇來自豐田研究院的論文之所以有趣，正是因為它拒絕寫出那樣的頭條。它真正的貢獻不是打造更炫的機器人，而是嚴肅檢視一個看似簡單的問題——大型模型的方法真的表現更好嗎？我們又要如何知道？——並以作者自稱在這個領域相當少見的統計嚴謹程度作答。結果是一個確實有用的「是，但要看情況」，也是一項警告：機器人學的許多研究可能測量到的只是噪聲。



兩種方法都接受相同的盲測隨機評估。論文的主張不是機器人基礎模型是零樣本通才，而是經過仔細測量後，預訓練能提升資料效率與穩健性。

作者做了什麼

他們所建立的 LBM 有一個具體而明確的定義：**基於擴散的視覺運動策略**（一個 Diffusion Transformer，讀取攝影機影像、簡短的語言指令，以及機器人自身的關節位置，並以 10 Hz 輸出一小段馬達指令）。這些模型先用 **約 1,700 小時**的機器人示範進行**預訓練**——包括內部收集的 500 多項不同任務，以及公開資料集——之後再針對個別任務進行**微調**。全篇的比較對象，是只用該項任務的資料**從頭訓練的單任務策略**。

不過，論文的核心其實是評估，作者將其視為主要成果。為了避免自欺，他們採用了：

- **真實世界中的盲測隨機 A/B 測試**——操作機器人的人不知道當時測試的是哪個策略，測試順序也經過隨機化。
- **受控且可重複的初始條件**——每次試驗前，操作人員都先比對影像疊加畫面，將場景調整至相符。
- **大量試驗次數**——每個任務、每個策略、每種條件在真實世界各進行 50 次 rollout；模擬中每個任務進行 200 次。總計約有 **1,800 次盲測真實世界 rollout，以及超過 47,000 次模擬 rollout**。
- **適當的統計方法**——使用成功機率的貝葉斯估計，以及經過多重比較校正的成對假設檢定，而不是用肉眼查看彼此重疊的誤差棒。他們甚至對四分之一由人類評分的試驗進行品質保證複核，以測量評分誤差。

[video pending]

模型擺設早餐餐桌的並排比較：（左）單任務基線，（右）LBM。兩段影片都以 1x 速度播放。這是一項經過評估的任務，並不能證明通用型自主能力。

這套機制才是重點。整篇論文都在論證：沒有這些做法，你就無法分辨真正的改進和運氣。

他們發現了什麼

經過微調的大型模型平均勝過從頭訓練的單任務模型。將各項任務合併計算後，先經過預訓練、再針對任務微調的 LBM，在模擬環境和真實世界中都可靠地勝過了在同一任務資料上從頭訓練的策略，而且差異具有統計顯著性。在個別任務上，經過微調的 LBM 幾乎每次都在統計上達到同等或更好的表現（真實世界任務為 3/3，模擬任務為 15/16）。

最大、最清楚的優勢是資料效率。經過微調的 LBM 只用約為從頭訓練所需任務特定資料的 **1/3 至 1/5**，就達到相當的表現。在一項真實世界任務（擺設早餐餐桌）中，只用 **15%** 示範進行微調的 LBM，勝過使用**全部 100%** 示範資料訓練的從頭訓練策略。

當條件發生變化時，預訓練最有幫助。當測試環境被刻意擾動，偏離訓練條件（「分布偏移」）時，微調 LBM 的優勢擴大了。在一組模擬實驗中，正常條件下它在 16 項任務中的 3 項於統計上勝過從頭訓練的策略；在分布偏移下則是 16 項中的 10 項。由於真實部署的環境總會偏離訓練條件，這種穩健性可以說是最具實際重要性的發現。

更多預訓練資料帶來平穩的提升。隨著加入更多預訓練資料，表現穩定上升——在測試的規模內，**沒有突然跳升或「湧現式」飛躍**。有用、可預測，而且並不戲劇化。

但不經微調的通才故事並不成立。對預訓練 LBM 進行零樣本使用——不做任務特定的微調——並沒有**持續地**勝過單任務策略。單一網路確實能同時完成許多任務，但「只要給它提示」的夢想在這裡沒有得到證實；作者將部分原因歸於他們的小型語言編碼器不夠穩健。

而且這些增益小到很容易被漏掉——或被製造出來。許多效果只有在比通常更大的樣本數和仔細的檢定下才看得見。作者直白地表示，考慮到效果的大小和噪聲，**機器人學的許多論文正在測量統計噪聲的風險很高**。他們也發現，一個平凡的選擇——資料如何正規化——對結果的影響大於架構變更；而且預訓練中的一個正規化**錯誤**，直到評估結束之後才浮現。

這可能意味著什麼

最站得住腳的解讀是：在多樣的機器人資料上進行大規模預訓練，是一個真實且值得採用的要素——它能減少每項新任務所需的資料，也讓策略在現實環境不符合訓練條件時更加穩健。這確實支持了這個領域押注的方向。但這些增益是有限且有條件的（主要在微調後出現，也最清楚地體現在整體結果和壓力測試下），並不代表一個可直接使用的通用機器人已經問世。

更低調、也更重要的意義在方法論上。這篇論文實際上是一把測量尺：它展示了要對機器人策略提出可信主張，究竟需要多少證據，也暗示許多已發表的振奮結果建立在過少的證據上。比起又一個模型，這才是這個領域更需要的校正。

這不能證明什麼

- 這**不是通用型機器人**。這些優勢是在特定架構（擴散策略）上、於受控環境中、根據遠端操作示範逐任務微調後展示的——不是一個接到指令就能自主完成任意新工作的機器人。
- 它**不能**驗證**零樣本**使用。不進行微調時，大型模型沒有持續勝過單任務基線。
- 這**不是**「**湧現式飛躍**」的證據。擴展規模帶來的是平穩改善；這裡沒有任何不連續性，可以支持「然後它突然變得有能力」這類敘事。
- 這些數字是**相對的，而且受實驗室條件限制**。絕對成功率被刻意調整到約 50%，以提高比較的敏感度；它們不是對真實世界可靠性的衡量，而這項工作也只涉及一間實驗室的一種架構。
- 它**沒有**解釋任何策略為何成功或失敗，而且論文報告了幾項大型模型表現**較差**的具體任務，卻沒有解釋原因。
- 它對評估裝置以外的**安全性、自主性或部署**沒有任何說明。

證據有多強？

就其核心比較主張——經過微調的 LBM 整體勝過從頭訓練的基線、所需資料少數倍，並且在分布偏移下更加穩健——而言，證據既充分，控制**也**異常嚴謹：盲測、隨機化、大樣本、統計檢定，還有評分的品質保證複核。這是少見的情況：方法論足夠扎實，可以近乎直接地接受其標題式結論。

作者自己也坦率提出了一些保留意見。他們的誤差棒涵蓋了評估的隨機性，卻**沒有**涵蓋訓練的隨機性——同一個模型訓練兩次，可能得到有實質差異的策略，而這種變異並未納入統計。真實世界的任務各進行 50 次試驗，足以捕捉中等效果，但可能漏掉小效果。語言條件控制使用了能力有限的編碼器，因此關於「只要告訴機器人該做什麼」的主張，在更大型的系統上可能會有不同結果。另外，他們也坦率披露了一個事後才發現的正規化錯誤。這些問題都不足以推翻主要發現，但正是這類事情，論文認為這個領域通常會置之不理。

同樣本著這種精神，還有一點來源說明：這篇解說是根據作者的預印本撰寫。我們無法取得期刊發表版本，因此沒有檢查預印本與發表文本之間是否有任何變更。

為什麼重要

「機器人基礎模型」是一個天生容易過度宣稱的詞，而像這樣的研究很容易被往兩個方向誤讀——得意地解讀成「它有效！」，或輕蔑地解讀成「這只是炒作。」準確的看法比兩者都更有用：在多樣資料上進行預訓練，確實帶來可測量但中等程度的好處——主要是每項任務所需資料更少，以及更高的穩健性——而且這條路徑會隨規模可預測地改善。

它重要的更深層原因，是這篇論文把嚴謹性用來檢視自己的領域。它展示真正的效果小到足以在草率的評估中消失，也展示像資料正規化這種平淡無奇的選擇，影響可能超過巧妙的新架構；因此它主張，機器人學的許多進展在獲得信任之前，都需要更可靠的測量。一篇用自身的可信度嚴格把關、區分結果與願望的論文，所做的事情比登上排行榜頂端更罕見，也更有價值。

重點摘要

豐田研究院的研究人員訓練了「大型行為模型」——在約 1,700 小時多樣化操作資料上預訓練的基於擴散的機器人策略——並以異常嚴謹的流程，將它們與從頭訓練的單任務策略比較：盲測、隨機化、大樣本（約 1,800 次真實世界與 47,000 多次模擬試驗），並使用正式統計。逐任務微調後，大型模型在整體上可靠地表現更好，需要的任務特定資料約為原來的 **1/3 至 1/5**，而且在條件變化時**更加穩健**；隨著預訓練資料增加，表現也平穩提升。但在**不經微調**的情況下，它們沒有持續勝過單任務模型；有些效果小到只有大樣本數才能揭示；而一項平凡的資料正規化選擇，比架構更重要。這是對機器人基礎模型方向扎實而克制的支持——**不是**通用型機器人，不是零樣本通才，也不是「湧現式飛躍」——同時也是一項尖銳警告：機器人學的許多研究可能測量到的只是噪聲。

務實檢視

論文展示了什麼：透過嚴謹、盲測且具統計效力的評估（約 1,800 次真實世界 + 47,000 多次模擬 rollout），經多任務預訓練後再微調的擴散策略（LBM）在整體上勝過從頭訓練的單任務策略，以約為原來 **1/3 至 1/5 的任務特定資料**達到相同表現，並在分布偏移下更加穩健；表現隨預訓練資料增加而平穩提升。

合理但尚未證實的內容：這些好處能延伸到大得多的視覺–語言–行動模型（他們的語言編碼器很小）；平穩擴展能在經過測試的資料範圍之外繼續。

它沒有展示什麼：通用型或零樣本機器人（不微調 → 沒有持續優勢）；任何「湧現式」能力跳躍；對特定任務失敗的解釋；絕對意義上的真實世界可靠性（成功率為提高敏感度而調整到接近 50%）；任何關於安全或自主部署的內容。

主要限制：統計涵蓋評估的隨機性，卻沒有涵蓋不同訓練執行之間的隨機性；每項任務 50 次真實世界試驗可能漏掉小效果；一種架構和一間實驗室；能力有限的語言編碼器；評估後才發現資料正規化錯誤；分析根據預印本（未檢查發表版本）。

一般讀者應有多大信心？可以高度確信，多任務預訓練加上微調確實帶來中等程度的好處——尤其是

資料效率與穩健性——而且這些好處經過異常仔細的測量。也可以高度確信，這**不是**通用型或零樣本機器人，**不是**湧現式飛躍。至於這些增益能在多大程度上擴展到更大型的模型，信心應屬中等。同時值得認真看待的是：作者自己警告，這個領域的效果小到統計效力不足的研究可能回報的只是噪聲。合適的態度是：對這種方法保持審慎樂觀，對缺乏這類統計支持的機器人 AI 結果保持合理的懷疑。

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/