The Clean Paper

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

Robot 'foundation model' ใหญ่ทำงานดีกว่าจริงไหม? คำตอบที่วัดดีคือ ใช่เล็กน้อย — และงานส่วนใหญ่แยกไม่ออก

Toyota Research Institute ฝึก 'large behavior models' — robot policy ที่ pretrain บนข้อมูล manipulation หลากหลาย ~1,700 ชั่วโมง — แล้วทดสอบกับ single-task policy from scratch ด้วยความเข้มงวดผิดปกติ: blind, randomized, sample ใหญ่ (~1,800 โลกจริง, 47,000+ simulation) พร้อมสถิติจริง หลัง finetune ต่อ task โมเดลใหญ่ทำดีกว่าโดยเฉลี่ย ใช้ข้อมูลเฉพาะ task น้อยกว่าราว 3–5× และ robust กว่าเมื่อเงื่อนไขเปลี่ยน; performance สูงขึ้นเรียบตามข้อมูล pretraining แต่เมื่อไม่ finetune มันไม่ได้ชนะ single-task model สม่ำเสมอ effect หลายอย่างเล็กจนต้อง sample ใหญ่จึงเห็น และตัวเลือก data-normalisation ธรรมดามีผลมากกว่า architecture นี่เป็นหลักฐานสนับสนุนทิศ robot-foundation-model แบบวัดได้ — ไม่ใช่หุ่นยนต์ทั่วไป ไม่ใช่ zero-shot generalist ไม่ใช่ 'emergent leap' — พร้อมคำเตือนว่างาน robotics จำนวนมากอาจวัด noise

---

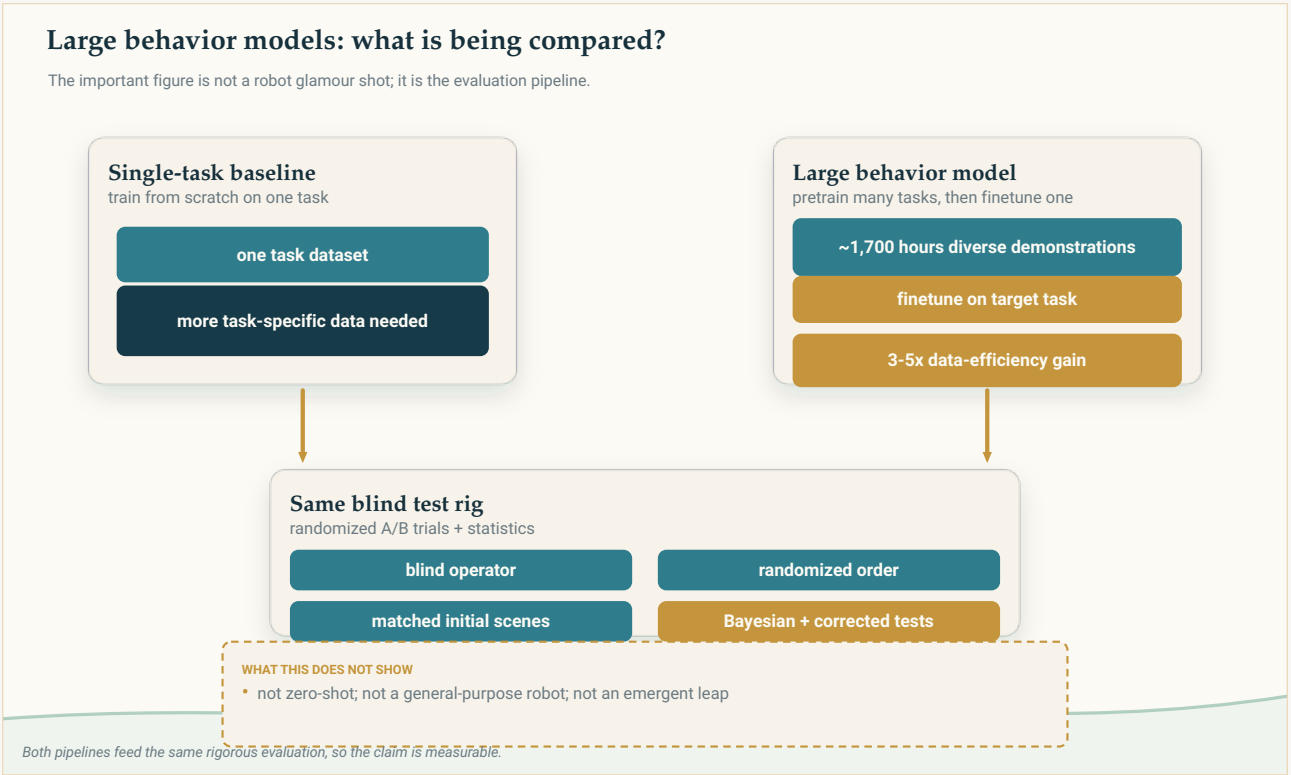
Written by Lucio Vaglio · Cross-checked by Laura Nesso · Edited by Michele Renda · Figures by Laura Nesso and Nova Vela · AI-assisted, human-reviewed

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

งานหุ่นยนต์ที่ประเด็นสำคัญจริง ๆ คือความซื่อตรงของการวัดผล

วงการหุ่นยนต์กำลังอยู่ในกระแส “โมเดลพื้นฐาน” แนวคิดที่ยืมมาจาก AI ด้านภาษาและภาพฟังดูน่าดึงดูดมาก: แทนที่จะฝึกหุ่นยนต์ทีละงาน ก็ฝึก **“โมเดลพฤติกรรมขนาดใหญ่” (large behavior model; LBM)** ตัวเดียวจากตัวอย่างการสาธิตจำนวนมหาศาลและหลากหลาย แล้วหวังว่าจะได้ระบบที่ทำงานได้กว้างและปรับตัวกับงานใหม่ได้เร็ว ความตื่นเต้น — และเงินลงทุน — จึงสูงมาก จนพาดหัวแทบเขียนตัวเองว่า *สมองหุ่นยนต์อเนกประสงค์มาถึงแล้ว*

งานจาก Toyota Research Institute ชิ้นนี้น่าสนใจตรงที่ปฏิเสธพาดหัวแบบนั้น ผลงานสำคัญไม่ได้อยู่ที่หุ่นยนต์ที่ดูหวือหวากว่าเดิม แต่อยู่ที่การถามคำถามเรียบง่ายซึ่งหลอกตัวเองได้ง่ายมากว่า **แนวทางใช้โมเดลใหญ่ดีกว่าจริงหรือไม่ และเราจะรู้ได้อย่างไรว่าดีกว่าจริง** ผู้เขียนตอบด้วยการออกแบบการประเมินที่รัดกุมทางสถิติอย่างผิดปกติสำหรับวงการ ผลลัพธ์คือคำตอบแบบ “ใช่ แต่มีเงื่อนไข” ที่มีประโยชน์จริง และคำเตือนว่างานหุ่นยนต์จำนวนไม่น้อยอาจกำลังวัดเพียงความผันผวนของข้อมูล



ทั้งสองแนวทางถูกนำเข้าสู่การประเมินแบบปกปิดและสุ่มด้วยวิธีเดียวกัน ข้ออ้างของงานไม่ใช่ว่าโมเดลพื้นฐานสำหรับหุ่นยนต์เป็นระบบอเนกประสงค์แบบ zero-shot แต่คือการฝึกล่วงหน้าบนข้อมูลจำนวนมากสามารถเพิ่มประสิทธิภาพการใช้ข้อมูลและความทนทานต่อสภาวะที่เปลี่ยนไปได้ หากวัดอย่างระมัดระวัง

Original The Clean Paper diagram CC BY 4.0

ผู้วิจัยทำอะไร

ผู้วิจัยสร้าง LBM ในความหมายที่เฉพาะเจาะจงมาก คือ **นโยบายควบคุมการเคลื่อนไหวจากภาพแบบ diffusion** โดยใช้ Diffusion Transformer ที่อ่านภาพจากกล้อง คำสั่งภาษาสั้น ๆ และตำแหน่งข้อต่อของหุ่นยนต์ แล้วส่งชุดคำสั่งมอเตอร์สั้น ๆ ที่

ความถี่ 10 Hz

โมเดลเหล่านี้ถูก **ฝึกล่วงหน้าด้วยข้อมูลการสาธิตการทำงานของหุ่นยนต์ประมาณ 1,700 ชั่วโมง** ครอบคลุมงานมากกว่า 500 แบบ ทั้งข้อมูลที่เก็บเองและชุดข้อมูลสาธารณะ จากนั้นจึง **ปรับจูนเพิ่มเติมสำหรับแต่ละงาน** แล้วเปรียบเทียบกับนโยบายควบคุมแบบงานเดียวที่ฝึกจากศูนย์ด้วยข้อมูลเฉพาะของงานนั้น

แต่หัวใจของงานอยู่ที่ **วิธีประเมิน** มากกว่าตัวโมเดล ผู้เขียนถือส่วนนี้เป็นผลลัพธ์สำคัญ เพื่อหลีกเลี่ยงการหลอกตัวเอง พวกเขาใช้:

- **การทดสอบ A/B แบบปกปิดและสุ่มในโลกจริง** — ผู้ควบคุมหุ่นยนต์ไม่รู้ว่านโยบายใดกำลังถูกทดสอบ และลำดับการทดลองถูกสุ่ม
- **สภาพเริ่มต้นที่ควบคุมและทำซ้ำได้** — ก่อนแต่ละรอบ ผู้ปฏิบัติงานจัดฉากให้ตรงกับภาพอ้างอิง
- **จำนวนรอบทดลองมาก** — 50 รอบในโลกจริงต่อหนึ่งงาน ต่อหนึ่งนโยบาย และต่อหนึ่งเงื่อนไข; 200 รอบต่อหนึ่งงานในการจำลอง รวมแล้วประมาณ **1,800 รอบในโลกจริงแบบปกปิด และมากกว่า 47,000 รอบในการจำลอง**
- **สถิติที่เหมาะสม** — ใช้การประมาณความน่าจะเป็นสำเร็จแบบเบย์ และการทดสอบสมมติฐานแบบเปรียบเทียบเป็นคู่พร้อมแก้ปัญหาการเปรียบเทียบหลายรายการ แทนการดูว่าแถบค่าคลาดเคลื่อนทับกันหรือไม่ด้วยตา นอกจากนี้ยังตรวจซ้ำหนึ่งในสี่ของการให้คะแนนโดยมนุษย์เพื่อวัดความคลาดเคลื่อนของผู้ประเมิน

[video pending]

เปรียบเทียบโมเดลสองแบบขณะจัดโต๊ะอาหารเช้า: ซ้ายคือโมเดลงานเดียวที่ฝึกจากศูนย์ ขวาคือ LBM ทั้งสองวิดีโอเล่นด้วยความเร็ว 1 เท่า นี่เป็นเพียงหนึ่งงานที่ใช้ประเมิน ไม่ใช่หลักฐานว่าหุ่นยนต์มีความสามารถอัตโนมัติอเนกประสงค์

ระบบวัดผลทั้งหมดนี้คือประเด็นของงาน ผู้เขียนกำลังโต้แย้งว่า หากไม่มีการประเมินที่รัดกุมพอ เราแทบแยกไม่ได้ว่าผลที่ดีขึ้นมาจากโมเดลจริงหรือจากโชคและความผันผวน

พบอะไร

โมเดลใหญ่ที่ปรับจูนแล้วทำได้ดีกว่าโมเดลงานเดียวที่ฝึกจากศูนย์ — โดยเฉลี่ย เมื่อรวมหลายงาน LBM ที่ผ่านการฝึกล่วงหน้าแล้วปรับจูนสำหรับงานเฉพาะ ทำผลงานดีกว่านโยบายควบคุมที่ฝึกจากศูนย์บนงานเดียวกันอย่างน่าเชื่อถือ ทั้งในการจำลองและโลกจริง ความแตกต่างโดยรวมมีนัยสำคัญทางสถิติ ในระดับงานแต่ละงาน LBM ทำได้อย่างน้อยเท่ากันหรือดีกว่าเกือบทุกกรณี: 3/3 งานในโลกจริง และ 15/16 งานในการจำลอง

ชัยชนะที่ชัดที่สุดคือใช้ข้อมูลเฉพาะงานน้อยลง LBM ที่ปรับจูนแล้วทำผลงานได้เทียบเท่าโมเดลที่ฝึกจากศูนย์โดยใช้ข้อมูลเฉพาะงาน **น้อยกว่าประมาณ 3–5 เท่า** ในงานโลกจริงหนึ่งงาน — การจัดโต๊ะอาหารเช้า — LBM ที่ใช้ข้อมูลสาธิตเพียง **15%** ยังทำได้ดีกว่าโมเดลงานเดียวที่ใช้ข้อมูล **100%**

การฝึกล่วงหน้าช่วยมากที่สุดเมื่อสภาพแวดล้อมเปลี่ยน เมื่อผู้วิจัยจงใจทำให้เงื่อนไขทดสอบต่างจากข้อมูลฝึก หรือเกิด **การเปลี่ยนแปลงการกระจายของข้อมูล (distribution shift)** ข้อได้เปรียบของ LBM ยิ่งชัดขึ้น ในชุดการจำลองหนึ่ง มันชนะโมเดลที่ฝึกจากศูนย์อย่างมีนัยสำคัญเพียง 3 จาก 16 งานในสภาพปกติ แต่เพิ่มเป็น 10 จาก 16 งานเมื่อเงื่อนไขเปลี่ยน เนื่องจากการใช้งานจริงย่อมค่อย ๆ เบี่ยงออกจากข้อมูลฝึก ความทนทานแบบนี้อาจเป็นผลที่สำคัญที่สุดในทางปฏิบัติ

ข้อมูลฝึกล่วงหน้ามากขึ้นช่วยอย่างค่อยเป็นค่อยไป ประสิทธิภาพเพิ่มขึ้นอย่างสม่ำเสมอเมื่อเพิ่มข้อมูลฝึกล่วงหน้า โดย **ไม่พบการกระโดดของความสามารถแบบฉับพลันหรือ "emergent leap"** ในช่วงขนาดข้อมูลที่ทดสอบ ผลดีมีอยู่จริง แต่เป็นไปอย่างต่อเนื่องและคาดเดาได้

แต่เรื่อง "โมเดลอเนกประสงค์ที่ไม่ต้องปรับจูน" ยังไม่เกิดขึ้น LBM ที่ผ่านการฝึกล่วงหน้าแล้วนำไปใช้แบบ **zero-shot** โดยไม่ปรับจูนเฉพาะงาน ไม่ได้ทำได้ดีกว่าโมเดลงานเดียวอย่างสม่ำเสมอ เครือข่ายเดียวทำหลายงานได้ แต่ความฝันแบบ "แค่บอกมันว่าต้องทำอะไร" ยังไม่ถูกยืนยันในงานนี้ ผู้เขียนคิดว่าส่วนหนึ่งอาจมาจากตัวเข้ารหัสภาษาที่ค่อนข้างเล็กและเปราะ

และผลดีหลายอย่างเล็กพอที่จะพลาดได้ง่าย — หรือดูเหมือนมีทั้งที่ไม่มี หลายผลลัพธ์เห็นชัดเพราะใช้ตัวอย่างมากกว่างานหุ่นยนต์ทั่วไปและทดสอบทางสถิติอย่างระมัดระวัง ผู้เขียนพูดตรง ๆ ว่า เมื่อเทียบขนาดผลกับความผันผวนของข้อมูล **มีความเสี่ยงจริงที่งานหุ่นยนต์จำนวนมากกำลังวัดเพียง noise ทางสถิติ** พวกเขายังพบว่าตัวเลือกธรรมดาอย่างวิธีทำ normalization ของข้อมูลมีผลต่อผลลัพธ์มากกว่าการเปลี่ยนสถาปัตยกรรม และพบข้อผิดพลาดด้าน normalization ในขั้นฝึกล่วงหน้าหลังการประเมินเสร็จแล้ว

สิ่งนี้น่าจะหมายความว่าอย่างไร

ข้อสรุปที่ป้องกันได้คือ การฝึกล่วงหน้าขนาดใหญ่บนข้อมูลหุ่นยนต์ที่หลากหลาย **มีประโยชน์จริง** ทำให้ต้องใช้ข้อมูลเฉพาะงานใหม่น้อยลง และทำให้นโยบายควบคุมทนต่อโลกที่ไม่ตรงกับข้อมูลฝึกได้ดีขึ้น ผลนี้สนับสนุนทิศทางที่วงการกำลังลงทุนอยู่จริง

แต่ประโยชน์ดังกล่าว **มีขนาดปานกลางและมีเงื่อนไข** ส่วนใหญ่เห็นหลังการปรับจูนเฉพาะงาน และชัดที่สุดเมื่อรวมหลายงานหรือทดสอบในสภาวะที่เบี่ยงจากข้อมูลฝึก นี่ยังไม่ใช่การมาถึงของหุ่นยนต์อเนกประสงค์ที่หยิบมาใช้กับงานใหม่ได้ทันที

ความหมายที่เงียบกว่าแต่สำคัญกว่าอยู่ด้านวิธีวิทยา งานนี้ทำหน้าที่เหมือนไม้บรรทัด แสดงว่าต้องใช้หลักฐานมากแค่ไหนจึงจะอ้างได้อย่างน่าเชื่อถือว่านโยบายควบคุมหุ่นยนต์หนึ่งดีกว่าอีกแบบ และชี้ว่าความตื่นเต้นในงานตีพิมพ์จำนวนมากอาจตั้งอยู่บนข้อมูลน้อยเกินไป นี่อาจเป็นการแก้ทิศทีวงการต้องการมากกว่าการเพิ่มโมเดลใหม่อีกหนึ่งตัว

สิ่งที่ผลนี้ *ยังพิสูจน์ไม่ได้*

- **นี่ไม่ใช่หุ่นยนต์อเนกประสงค์** ผลดีถูกแสดงในสถาปัตยกรรมเฉพาะ คือ diffusion policy ที่ต้องปรับจูนแยกตามงาน ในสภาพควบคุม และเรียนจากการสาธิตที่มนุษย์ควบคุม ไม่ใช่หุ่นยนต์อัตโนมัติที่รับงานใหม่อะไรก็ได้ตามคำสั่ง
- **ไม่ได้ยืนยันการใช้แบบ zero-shot** หากไม่ปรับจูน โมเดลใหญ่ไม่ได้ชนะโมเดลงานเดียวอย่างสม่ำเสมอ
- **ไม่ใช่หลักฐานของการกระโดดความสามารถแบบฉับพลัน** การเพิ่มขนาดข้อมูลทำให้ดีขึ้นอย่างราบรื่น ไม่มีจุดที่ความสามารถ "โผล่ขึ้นมา" ทันที
- **ตัวเลขยังผูกกับห้องปฏิบัติการและเป็นการเปรียบเทียบสัมพัทธ์** อัตราความสำเร็จโดยสัมบูรณ์ถูกปรับให้ใกล้ 50% เพื่อให้ไวต่อความแตกต่าง จึงไม่ใช่การวัดความน่าเชื่อถือของหุ่นยนต์ในโลกจริง
- **ยังไม่อธิบายว่าทำไมแต่ละงานสำเร็จหรือล้มเหลว** มีหลายงานที่โมเดลใหญ่ทำแย่กว่า แต่ไม่ได้มีคำอธิบายเชิงกลไก
- **ไม่บอกอะไรเรื่องความปลอดภัย ความเป็นอัตโนมัติ หรือการใช้งานนอกชุดประเมิน**

หลักฐานน่าเชื่อถือเพียงใด

สำหรับข้ออ้างเปรียบเทียบหลัก — LBM ที่ปรับจูนแล้วทำได้ดีกว่าโมเดลงานเดียวโดยรวม ใช้ข้อมูลน้อยกว่าหลายเท่า และทนต่อการเปลี่ยนสภาพได้ดีกว่า — หลักฐานค่อนข้างแข็งแรงและควบคุมดีผิดปกติ: มีการปกปิดข้อมูลกลุ่ม การสุ่ม ตัวอย่างจำนวนมาก การทดสอบทางสถิติ และการประกันคุณภาพการให้คะแนน นี่เป็นกรณีที่วิธีวิทยาดีพอให้รับข้อสรุปหลักได้ค่อนข้างตรงตาม

ตัวอักษร

ข้อจำกัดที่สำคัญเป็นสิ่งที่ผู้เขียนระบุเอง แถบค่าคลาดเคลื่อนสะท้อนความสุ่มของ **การประเมิน** แต่ไม่ได้รวมความสุ่มของ **การฝึกโมเดล** หากฝึกโมเดลเดิมสองครั้งอาจได้ผลต่างกันมาก และความแปรปรวนนี้ไม่อยู่ในสถิติ งานโลกจริงมี 50 รอบต่อหนึ่งงาน พอจับผลขนาดกลางแต่ยังอาจพลาดผลเล็ก ตัวเข้ารหัสภาษามีขนาดไม่ใหญ่ จึงไม่รู้ว่าระบบที่ใหญ่กว่าจะให้ผลอย่างไร และยังมีการเปิดเผยตรง ๆ ว่าพบข้อผิดพลาดด้าน normalization หลังการประเมินเสร็จแล้ว

อีกข้อสังเกตด้านแหล่งข้อมูล: บทความอธิบายนี้อิงจาก **บทความก่อนตีพิมพ์ของผู้เขียน** เราไม่สามารถดึงฉบับวารสารที่ตีพิมพ์แล้วมาเปรียบเทียบ จึงยังไม่ได้ตรวจว่ามีการเปลี่ยนแปลงระหว่างสองฉบับหรือไม่

ทำไมจึงสำคัญ

คำว่า “robot foundation model” ถูกใช้เกินจริงได้ง่าย งานแบบนี้จึงถูกอ่านผิดได้ทั้งสองทิศ — เป็นชัยชนะว่า *“มันใช้ได้แล้ว!”* หรือถูกปัดว่า *“ก็แค่กระแสเกินจริง”* การอ่านที่แม่นยำมีประโยชน์กว่าทั้งสองแบบ: การฝึกล่วงหน้าบนข้อมูลหลากหลายให้ประโยชน์จริง วัดได้ แต่มีขนาดปานกลาง โดยเฉพาะการลดข้อมูลที่ต้องใช้ต่อหนึ่งงานและเพิ่มความทนทานเมื่อสภาพแวดล้อมเปลี่ยน

เหตุผลที่ลึกกว่านั้นคือ งานนี้หันความเข้มงวดกลับมาที่วงการของตัวเอง ด้วยการแสดงว่าผลจริงเล็กพอจะหายไปหากประเมินอย่างหยาบ และตัวเลือกธรรมดาอย่าง normalization ของข้อมูลอาจมีผลมากกว่าสถาปัตยกรรมใหม่ที่ดูฉลาด งานจึงกำลังบอกว่า ความก้าวหน้าหลายอย่างใน robot learning ต้องผ่านการวัดที่แข็งแรงกว่านี้ก่อนจึงควรเชื่อ งานที่ใช้ความน่าเชื่อถือของตัวเองเพื่อแบ่งเส้นระหว่าง “ผลลัพธ์” กับ “ความหวัง” มีค่ามากกว่าการขึ้นอันดับหนึ่งบนตารางคะแนนเสียอีก

สรุป

นักวิจัยจาก Toyota Research Institute ฝึก **large behavior models** ซึ่งเป็นนโยบายควบคุมหุ่นยนต์แบบ diffusion ที่ผ่านการฝึกล่วงหน้าด้วยข้อมูลการหยิบจับและเคลื่อนย้ายหลากหลายประมาณ 1,700 ชั่วโมง แล้วเปรียบเทียบกับโมเดลงานเดียวที่ฝึกจากศูนย์ด้วยวิธีประเมินที่เข้มงวดผิดปกติ: ปกปิดข้อมูลกลุ่ม สุ่ม ใช้ตัวอย่างมาก (ประมาณ 1,800 รอบในโลกจริง และมากกว่า 47,000 รอบในการจำลอง) และใช้สถิติจริง หลังปรับจูนแยกตามงาน โมเดลใหญ่ทำได้ดีกว่าโดยรวมอย่างน่าเชื่อถือ ใช้ข้อมูลเฉพาะงาน **น้อยกว่าประมาณ 3–5 เท่า** และ **ทนต่อสภาวะที่เปลี่ยนไปได้ดีกว่า** โดยประสิทธิภาพเพิ่มขึ้นอย่างราบรื่นเมื่อเพิ่มข้อมูลฝึกล่วงหน้า แต่เมื่อใช้ **โดยไม่ปรับจูน** มันไม่ได้ชนะโมเดลงานเดียวอย่างสม่ำเสมอ ผลหลายอย่างเล็กจนต้องใช้ตัวอย่างจำนวนมากจึงจะเห็น และวิธี normalization ของข้อมูลธรรมดา ๆ มีผลมากกว่าการเปลี่ยนสถาปัตยกรรม นี่เป็นหลักฐานที่แข็งแรงและวัดอย่างระมัดระวังเพื่อสนับสนุนแนวทาง robot foundation model — **ไม่ใช่** หุ่นยนต์อเนกประสงค์ ไม่ใช่ zero-shot generalist และไม่ใช่ “emergent leap” — พร้อมคำเตือนชัดเจนว่างานหุ่นยนต์จำนวนไม่น้อยอาจกำลังวัดเพียง noise

ประเมินแบบไม่เกินจริง

งานวิจัยแสดงอะไร: เมื่อใช้การประเมินแบบปกปิดและสุ่มที่มีตัวอย่างมากและมีกำลังทางสถิติเพียงพอ นโยบายควบคุมแบบ diffusion ที่ผ่านการฝึกล่วงหน้าหลายงานแล้วปรับจูนเฉพาะงาน ทำได้ดีกว่าโมเดลงานเดียวที่ฝึกจากศูนย์โดยรวม ใช้ข้อมูล

เฉพาะงานน้อยกว่าประมาณ 3–5 เท่าเพื่อให้ได้ผลงานเทียบเท่า และทนต่อการเปลี่ยนแปลงการกระจายของข้อมูลได้ดีกว่า

อะไรที่เป็นไปได้แต่ยังไม่พิสูจน์: ประโยชน์เหล่านี้อาจถ่ายทอดไปยังโมเดล vision-language-action ที่ใหญ่กว่ามาก และแนวโน้มที่ดีขึ้นอย่างราบรื่นอาจดำเนินต่อเกินช่วงข้อมูลที่ทดสอบ

มันไม่ได้แสดงอะไร: ไม่ได้แสดงหุ่นยนต์อเนกประสงค์หรือระบบ zero-shot ที่ได้เปรียบโดยไม่ปรับจูน ไม่ได้แสดงการกระโดดความสามารถแบบ emergent ไม่ได้อธิบายความล้มเหลวในแต่ละงาน ไม่ได้วัดความน่าเชื่อถือโลกจริงแบบสัมบูรณ์ และไม่ได้บอกอะไรเรื่องความปลอดภัยหรือการใช้งานอัตโนมัติจริง

ข้อจำกัดหลัก: สถิติรวมความสุ่มจากการประเมิน แต่ไม่รวมความแปรปรวนจากการฝึกโมเดลแต่ละครั้ง งานโลกจริงมี 50 รอบต่อหนึ่งงานจึงอาจพลาดผลเล็ก ใช้สถาปัตยกรรมหนึ่งแบบจากห้องปฏิบัติการหนึ่งแห่ง ตัวเข้ารหัสภาษามีขนาดไม่ใหญ่ พบข้อผิดพลาดด้าน normalization หลังการประเมิน และการวิเคราะห์นี้อิงจากบทความก่อนตีพิมพ์ที่ยังไม่ได้เทียบกับฉบับวารสาร

ผู้อ่านทั่วไปควรมั่นใจแค่ไหน? มั่นใจสูงว่าการฝึกล่วงหน้าหลายงานแล้วปรับจูนเฉพาะงานให้ประโยชน์จริงระดับปานกลาง โดยเฉพาะด้านการใช้ข้อมูลและความทนทาน และงานวัดผลเหล่านี้อย่างรัดกุมผิดปกติ มั่นใจสูงเช่นกันว่านี่ **ยังไม่ใช่** หุ่นยนต์อเนกประสงค์หรือ zero-shot และ **ไม่ใช่** การกระโดดความสามารถแบบ emergent ส่วนคำถามว่าประโยชน์จะขยายไปไกลแค่ไหนเมื่อใช้โมเดลที่ใหญ่ขึ้นยังมีความมั่นใจปานกลาง

แหล่งข้อมูล

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/