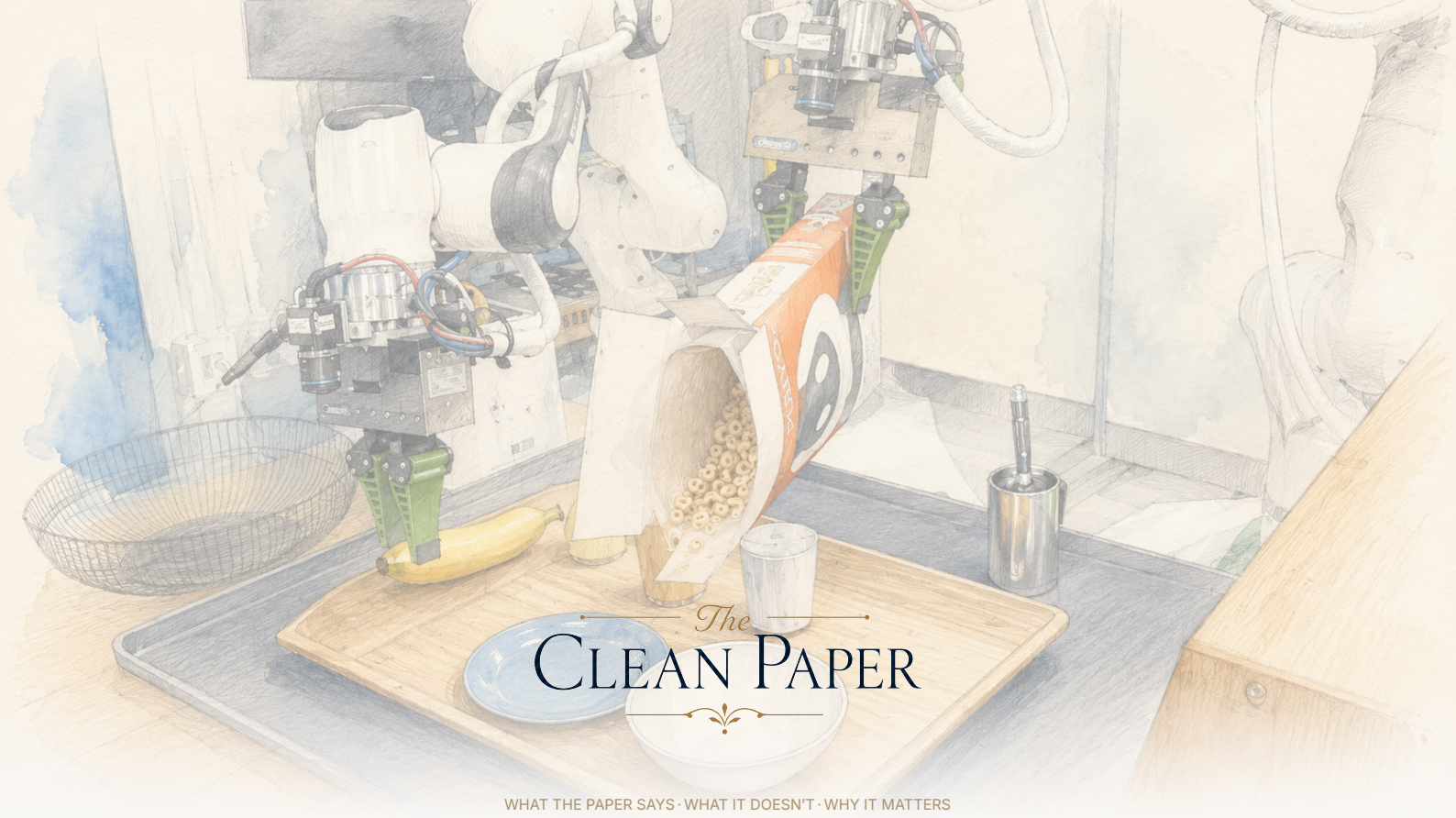


The Clean Paper

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

# Czy duże robotyczne „modele fundamentalne" naprawdę działają lepiej? Ostrożna odpowiedź — umiarkowanie tak, a większość badań nie potrafi tego stwierdzić

*Toyota Research Institute wytrenował „duże modele zachowania" — polityki robotów poddane pretreningowi na ~1700 godzinach zróżnicowanych danych manipulacyjnych — i porównał je z politykami pojedynczych zadań trenowanymi od zera za pomocą wyjątkowo rygorystycznych, ślepych i randomizowanych prób o dużej liczebności (~1800 fizycznych, ponad 47 000 symulowanych) oraz właściwej statystyki. Po dostrojeniu duże modele średnio działały lepiej, potrzebowały około 3–5 razy mniej danych właściwych dla zadania i były odporniejsze na zmianę warunków; wynik rósł płynnie z ilością danych pretreningowych. Bez dostrajania nie przewyższały jednak konsekwentnie modeli jednego zadania, kilka efektów było bardzo małych, a zwykła decyzja o normalizacji danych znaczyła więcej niż architektura. To zmierzone wsparcie tego kierunku — nie robot ogólnego przeznaczenia, uniwersalny model zero-shot ani „wyłaniający się skok" — oraz ostrzeżenie, że znaczna część robotyki może mierzyć szum.*

---

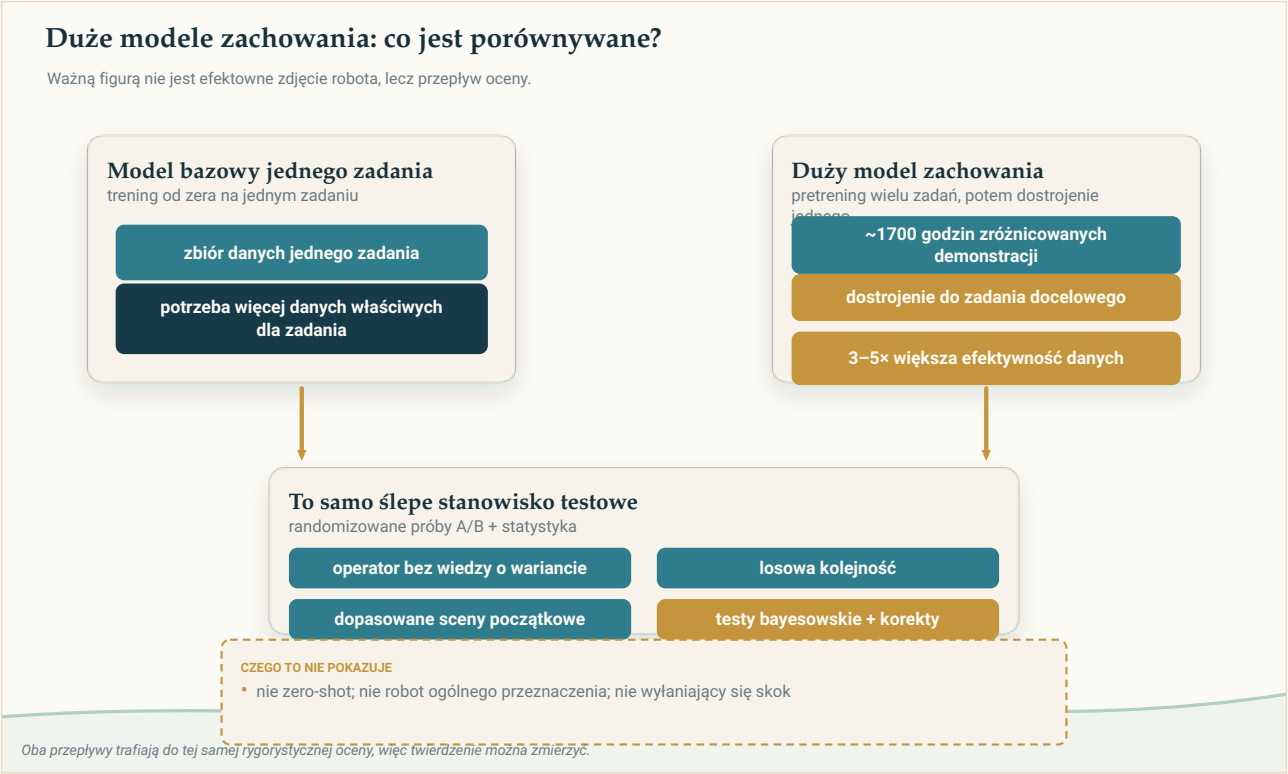
Written by Lucio Vaglio · Cross-checked by Laura Nesso · Edited by Michele Renda · Figures by Laura Nesso and Nova Vela · AI-assisted, human-reviewed

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

# Artykuł o robotach, którego prawdziwym tematem jest uczciwość

Robotyka przeżywa gorączkę „modeli fundamentalnych”. Zaprawiona w AI językowej i obrazowej idea jest kusząca: zamiast uczyć robota każdego zadania osobno, wytrenować jeden wielki **„duży model zachowania” (LBM)** na ogromnym, zróżnicowanym zbiorze demonstracji i uzyskać system o szerokich zdolnościach, szybko przystosowujący się do nowych zadań. Entuzjazm — i inwestycje — są olbrzymie. Nagłówek pisze się sam: *mózgi robotów ogólnego przeznaczenia już tu są*.

Praca Toyota Research Institute jest interesująca właśnie dlatego, że odmawia napisania takiego nagłówka. Jej głównym wkładem nie jest efektowniejszy robot, lecz wymagająca analiza pozornie prostego pytania — *czy podejście z dużym modelem rzeczywiście działa lepiej i skąd mielibyśmy to wiedzieć?* — przeprowadzona z troską o statystykę, która według samych autorów jest w tej dziedzinie nietypowa. Wynikiem jest naprawdę użyteczne „tak, ale” oraz ostrzeżenie, że znaczna część robotyki może mierzyć szum.




Oba podejścia przechodzą tę samą ślepą, randomizowaną ocenę. Praca nie twierdzi, że robotyczne modele fundamentalne są uniwersalnymi systemami zero-shot, lecz że starannie zmierzony pretrening może poprawić efektywność danych i odporność.

# Co zrobili autorzy

Zbudowali LBM w konkretnym znaczeniu: **wizyjno-ruchowe polityki oparte na dyfuzji** (Diffusion Transformer odczytujący obrazy z kamer, krótką instrukcję językową i położenie własnych stawów robota, a następnie generujący krótkie serie poleceń ruchu z częstotliwością 10 Hz). Modele te poddano **pretreningowi na około 1700 godzinach** demonstracji robotycznych — ponad 500 różnych zadań zebranych wewnętrznie oraz publicznych zbiorach danych — a następnie **dostrajano** je do poszczególnych zadań. Punktem odniesienia była zawsze **polityka jednego zadania trenowana od zera** wyłącznie na danych tego zadania.

Sednem pracy jest jednak *ocena*, którą autorzy traktują jako główny wynik. Aby nie oszukiwać samych siebie, zastosowali:

- **Ślepe, randomizowane testy A/B** w świecie fizycznym — operator robota nie wiedział, którą politykę sprawdza, a kolejność była losowa.
- **Kontrolowane, powtarzalne warunki początkowe** — przed każdą próbą operatorzy dopasowywali scenę do nałożonego obrazu wzorcowego.
- **Dużą liczbę prób** — 50 rzeczywistych wykonań na zadanie, politykę i warunek oraz 200 na zadanie w symulacji. Łącznie około **1800 ślepych wykonań w świecie fizycznym i ponad 47 000 w symulacji**.
- **Właściwą statystykę** — bayesowskie oszacowania prawdopodobieństwa sukcesu oraz parami testowane hipotezy z korektą na wielokrotne porównania, zamiast oceny nakładających się słupków błędu na oko. Dodatkowo kontrolnie sprawdzili jedną czwartą prób ocenianych przez ludzi, aby zmierzyć błąd punktacji.

[video pending]

Porównanie modeli ustawiających stół śniadaniowy: po lewej bazowy model jednego zadania, po prawej LBM. Oba filmy są odtwarzane w normalnym tempie. To jedno oceniane zadanie, a nie dowód autonomii ogólnego przeznaczenia.

Cała ta aparatura jest sednem. Praca argumentuje, że bez niej nie da się odróżnić rzeczywistej poprawy od szczęśliwego przypadku.

# Co odkryli

**Dostrojone duże modele pokonały trenowane od zera modele jednego zadania — średnio.** Po zsumowaniu wyników zadań LBM poddany pretreningowi, a następnie dostrojony do zadania, niezawodnie przewyższał politykę trenowaną od zera na danych tego samego zadania — zarówno w symulacji, jak i w świecie fizycznym — a różnica była statystycznie istotna. W niemal każdym pojedynczym zadaniu dostrojony LBM był statystycznie równie dobry lub lepszy od modelu trenowanego od zera (3/3 zadań fizycznych, 15/16 symulowanych).

**Największą i najwyraźniejszą korzyścią jest efektywność danych.** Dostrojony LBM osiągał wynik

modelu trenowanego od zera przy użyciu około **3–5 razy mniejszej ilości danych właściwych dla zadania**. W jednym zadaniu fizycznym (nakrywaniu stołu śniadaniowego) LBM dostrojony na zaledwie **15%** demonstracji pokonał politykę trenowaną od zera na **100%** danych.

**Pretrening pomaga najbardziej, gdy warunki się zmieniają.** Kiedy celowo odsunięto środowisko testowe od warunków treningowych („przesunięcie rozkładu"), przewaga dostrojonego LBM *wzrosła*. W jednym zestawie symulacji model statystycznie pokonał wariant od zera w 3 z 16 zadań w zwykłych warunkach, ale w 10 z 16 przy przesunięciu rozkładu. Ponieważ rzeczywiste wdrożenia zawsze odbiegają od treningu, ta odporność jest prawdopodobnie najważniejszym praktycznie wynikiem.

**Więcej danych pretreningowych pomagało — płynnie.** Wyniki rosły równomiernie wraz z ilością danych, **bez nagłego skoku ani „wyłaniającego się" przełomu** w badanej skali. Użytecznie, przewidywalnie, bez dramatu.

**Nie potwierdziła się jednak opowieść o uniwersalnym modelu bez dostrajania.** LBM używany *zero-shot* — bez dostrajania do zadania — nie przewyższał konsekwentnie modeli jednego zadania. Jedna sieć potrafiła wykonywać wiele zadań, lecz marzenie „po prostu wydaj mu polecenie" nie znalazło tu potwierdzenia; autorzy częściowo przypisują to kruchości małego kodera językowego.

**Korzyści były na tyle małe, że łatwo je przeoczyć — albo sfabrykować.** Wiele efektów stało się widocznych dopiero przy większych niż zwykle próbach i starannych testach. Autorzy piszą wprost, że wobec wielkości efektów i szumu **istnieje znaczne ryzyko, iż wiele prac robotycznych mierzy szum statystyczny**. Odkryli także, że przyziemna decyzja — sposób *normalizacji* danych — wpływała na wyniki bardziej niż zmiany architektury, a **błąd** normalizacji pretreningu ujawniono dopiero *po* zakończeniu ocen.

# Co to prawdopodobnie oznacza

Uzasadniona interpretacja: pretrening na dużej ilości zróżnicowanych danych robotycznych jest rzeczywiście wartościowym składnikiem — zmniejsza ilość danych potrzebnych dla nowego zadania i uodparnia polityki na niedopasowanie świata do treningu. Naprawdę wspiera to kierunek, na który stawia dziedzina. Korzyści są jednak *umiarkowane i warunkowe* (ujawniają się głównie po dostrojeniu, najwyraźniej w wynikach zbiorczych i pod obciążeniem), a nie oznaczają pojawienia się uniwersalnego robota gotowego do użycia.

Mniej widowiskowy, ale ważniejszy wniosek jest metodologiczny. Praca stanowi w praktyce wzorzec pomiarowy: pokazuje, ile dowodów naprawdę potrzeba do wiarygodnego twierdzenia o polityce robota, i sugeruje, że wiele publikowanego entuzjazmu opiera się na zbyt skąpych danych. Dziedzinie taka korekta jest potrzebna bardziej niż kolejny model.

# Czego to *nie* dowodzi

- To **nie jest robot ogólnego przeznaczenia**. Korzyści wykazano dla konkretnej architektury (polityk dyfuzyjnych), dostrajanej osobno do zadań, w kontrolowanych warunkach i na demonstracjach teleoperowanych — nie dla autonomicznego robota wykonującego dowolne nowe polecenia.
- Wyniki **nie** potwierdzają działania **zero-shot**. Bez dostrajania duży model nie przewyższał konsekwentnie modeli bazowych jednego zadania.
- To **nie** jest dowód **„wyłaniającego się skoku”**. Skalowanie poprawiało wyniki płynnie; nie ma tu nieciągłości wspierającej opowieść „i nagle stał się zdolny”.
- Liczby są **względne i laboratoryjne**. Bezwzględne odsetki sukcesu celowo ustawiono w pobliżu ~50%, by uwrażliwić porównania; nie mierzą rzeczywistej niezawodności, a praca dotyczy jednej architektury z jednego laboratorium.
- Badanie **nie** rozstrzyga, **dlaczego** poszczególne polityki odnoszą sukces lub zawodzą; opisano kilka zadań, w których duży model był *gorszy*, ale ich nie wyjaśniono.
- Nie mówi **nic o bezpieczeństwie, autonomii ani wdrożeniu** poza stanowiskiem testowym.

# Jak mocne są dowody?

Dowody na główne twierdzenia porównawcze — że dostrojone LBM zbiorczo pokonują modele od zera, potrzebują kilkukrotnie mniej danych i lepiej znoszą przesunięcie rozkładu — są mocne i wyjątkowo dobrze kontrolowane: ślepe, randomizowane, o dużej liczebności, testowane statystycznie i z kontrolą jakości punktacji. To rzadki przypadek metodologii wystarczająco solidnej, by główne wnioski przyjąć niemal dosłownie.

Uczciwe zastrzeżenia podnoszą sami autorzy. Przedziały błędu obejmują losowość *oceny*, ale **nie** losowość *treningu* — dwukrotne wytrenowanie tego samego modelu może dać istotnie różne polityki, czego statystyki nie uwzględniają. Pięćdziesiąt prób na zadanie w świecie fizycznym wystarcza do wykrycia średnich efektów, lecz może przeoczyć małe. Warunkowanie językowe korzystało ze skromnego kodera, więc wyniki dotyczące „powiedz robotowi, co ma robić” mogą wyglądać inaczej w większych systemach. Dochodzi szczere ujawnienie błędu normalizacji po fakcie. Żaden z tych problemów nie obala głównych wyników, lecz są dokładnie tym, co według pracy dziedzina zwykle zamiata pod dywan.

Uwaga o źródle: omówienie opiera się na preprincie autorów. Nie udało nam się uzyskać wersji opublikowanej w czasopiśmie, więc nie sprawdziliśmy zmian między preprintem a ostatecznym tekstem.

# Dlaczego to ważne

„Robotyczne modele fundamentalne” to określenie stworzone do przesady, a tę pracę łatwo błędnie odczytać w obie strony — triumfalnie jako *„to działa!”* albo lekceważąco jako *„to przereklamowane”*. Trafne ujęcie jest użyteczniejsze od obu: pretrening na zróżnicowanych danych daje rzeczywiste,

mierzalne, lecz umiarkowane korzyści — głównie mniej danych na zadanie i większą odporność — a rezultaty poprawiają się przewidywalnie wraz ze skalą.

Głębsze znaczenie polega na skierowaniu rygoru na własną dziedzinę. Wykazując, że prawdziwe efekty są tak małe, iż znikają przy niedbałej ocenie, a nudna decyzja o normalizacji danych może znaczyć więcej niż pomysłowa architektura, autorzy dowodzą, że znaczna część postępu w uczeniu robotów potrzebuje solidniejszych pomiarów, zanim w niego uwierzymy. Publikacja, która wykorzystuje wiarygodność do pilnowania granicy między wynikiem a życzeniem, robi coś rzadszego i cenniejszego niż zdobycie pierwszego miejsca w rankingu.

# Zwięzłe podsumowanie

Badacze Toyota Research Institute wytrenowali „duże modele zachowania” — dyfuzyjne polityki robotów poddane pretreningowi na ~1700 godzinach zróżnicowanych danych manipulacyjnych — i porównali je z politykami pojedynczych zadań trenowanymi od zera za pomocą wyjątkowo rygorystycznego protokołu: ślepych, randomizowanych prób o dużej liczebności (≈1800 w świecie fizycznym i ponad 47 000 w symulacji) z prawdziwą statystyką. Po dostrojeniu do zadań duże modele niezawodnie wypadały lepiej zbiorczo, potrzebowały około **3–5 razy mniej danych właściwych dla zadania** i były **odporniejsze na zmianę warunków**, a wyniki rosły płynnie wraz z ilością danych pretreningowych. Używane **bez dostrajania** nie przewyższały jednak konsekwentnie modeli jednego zadania, niektóre efekty były widoczne dopiero w dużych próbach, a przyziemna normalizacja danych znaczyła więcej niż architektura. To solidne, wyważone wsparcie kierunku robotycznych modeli fundamentalnych — **nie** robot ogólnego przeznaczenia, uniwersalny model zero-shot ani „wyłaniający się skok” — oraz dobitne ostrzeżenie, że znaczna część robotyki może mierzyć szum.

# Kontrola bez upiększeń

**Co pokazuje praca:** Przy rygorystycznej, ślepej ocenie o mocy statystycznej (≈1800 wykonań fizycznych + ponad 47 000 symulowanych) wielozadaniowo pretrenowane, a następnie dostrajane polityki dyfuzyjne (LBM) zbiorczo przewyższają polityki pojedynczych zadań trenowane od zera, osiągają równoważne wyniki przy ~3–5 razy mniejszej ilości danych właściwych dla zadania i lepiej znoszą przesunięcie rozkładu; wynik rośnie płynnie z ilością danych pretreningowych.

**Co jest prawdopodobne, ale niedowiedzione:** Przeniesienie korzyści na znacznie większe modele wizyjno-językowo-ruchowe (ich koder językowy był mały); dalsze utrzymanie płynnego skalowania poza badanym zakresem danych.

**Czego praca nie pokazuje:** Robota ogólnego przeznaczenia ani zero-shot (brak dostrajania → brak stałej przewagi); żadnego „wyłaniającego się” skoku zdolności; wyjaśnień konkretnych porażek w zadaniach; bezwzględnej niezawodności w świecie rzeczywistym (odsetki sukcesu ustawiono blisko 50% dla czułości); niczego o bezpieczeństwie ani autonomicznym wdrożeniu.

**Główne ograniczenia:** Statystyki obejmują losowość oceny, lecz nie kolejnych treningów; 50 rzeczywistych prób na zadanie może przeoczyć małe efekty; jedna architektura i jedno laboratorium; skromny koder językowy; błąd normalizacji danych znaleziony po ocenach; analiza oparta na preprincie (wersji opublikowanej nie sprawdzono).

**Jak duże zaufanie powinien mieć czytelnik?** Duże do wniosku, że wielozadaniowy pretrening połączony z dostrajaniem daje rzeczywiste, umiarkowane korzyści — szczególnie w efektywności danych i odporności — i że zmierzono je niezwykle starannie. Duże także do tego, że **nie** jest to robot ogólnego przeznaczenia lub zero-shot ani wyłaniający się przełom. Umiarkowane co do skalowania korzyści na większe modele. Warto też poważnie potraktować ostrzeżenie autorów, że efekty w tej dziedzinie są tak małe, iż badania o niedostatecznej mocy mogą raportować szum. Właściwa postawa: wyważony optymizm wobec podejścia i zdrowy sceptycyzm wobec wyników robotycznej AI pozbawionych takiego zaplecza statystycznego.

---

## Źródła

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/