VLA 架構與端側智算系統演進

具身 AI 大爆發 機器人要

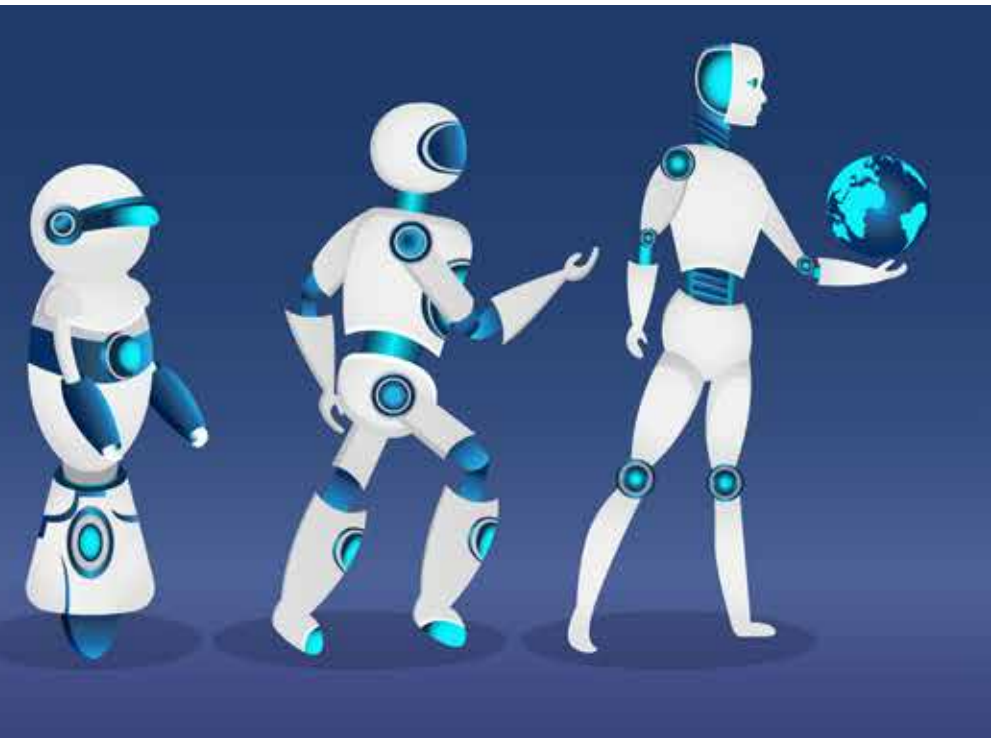
現在 AI 不只是會寫文案、畫圖，而是開始真正走進現實物理世界。

具身智慧大模型 (Embodied AI Model) 是指將人工智慧集成到物理系統中，使其具備與物理世界交互的能力。讓機器人等物理實體具備像人類一樣“感知、思考和行動”的核心大腦與技術工具。與主要處理和分析資料的資訊人工智慧不同，具身智慧將 AI 的功能擴展到物理系統，例如通用機器人、人形機器人、智慧汽車，甚至是工廠和倉庫設施。機器學習、感測器和電腦視覺的融合使這些系統能夠在現實環境中感知、推理和行動。

簡單的說，就是給機器人、人形機器人、智慧汽車裝上一顆會「看、會想、會動」的大腦，不再只處理網路上的數據，而是直接跟真實環境互動。

過去這一兩年，具身 AI 迎來一次大換代：從早期的大語言模型 (LLM) 與視覺語言模型 (VLA, Vision-Language-Action)，演進到 VLA 視覺語言動作端到端架構，代表作品像是 Physical Intelligence 團隊的 Pi 0.5。

過去依賴感知、規劃、控制分離的模組化系統，Pi-0.5 類型的 VLA 模型採用了高度整合的神經網路工作流：其前端透過 SigLIP 編碼器進行高解析度的視覺與多模態特徵提取，隨後無縫接入一個輕量級的 2B (約 20 億參數) 語言模型



看懂眞實世界了

編輯部整理

來處理語義理解與空間推理，最終傳遞給動作生成專家網路 (Action Generation Expert Networks)，直接輸出機器人的關節扭矩與馬達控制指令。以前機器人是拆成好幾個模組：先感知環境、再做規劃、最後才控制動作，模組之間傳資料會有延遲。而 VLA 架構是一條路跑到底：先用編碼器解析畫面影像，交給輕量大模型做理解跟空間判斷，直接輸出機器人馬達、關節的控制訊號。好處很直白：延遲壓得很低，還具備零樣本泛化能力。就算是沒看過的廚房、房間，機器人也能自主完成十幾分鐘的家事、清掃這類複雜任務。

與此同時，為了解決機器人在複雜物理世界中的空間感知痛點，4D Occupancy (體積佔用) 網路在過去一年半內成為了非結構化環境避障的絕對主流標準。相較於傳統依賴 2D 邊界框 (Bounding Box) 或稀疏點雲的演算法，最新一代高達 4.5 億參數的 4D Occupancy 網路，能將機器人周遭的物理空間徹底「體素化 (Voxelized)」。透過引入時間維度 (4D)，該網路不僅能夠進行對象級的無障礙三維幾何建模，還能精準預測動態障礙物 (如走動的人類、移動的推車) 在未來幾秒內的運動軌跡。這種高密度的空間表徵機制，讓具身機器人得以在高度擁擠且隨機變化的工廠產線或家庭客廳中，實現毫秒級的實時碰撞閃避與路徑重規劃，徹底補足了純視覺方案在深度估計與幾何感知上的短板。

中美兩國在 AI 技術路線明顯不一樣。

在這波技術浪潮中，全球在 AI 路線上出現了顯著的分野與多樣化的戰略分化。資策會 MIC 指出，目前共有三大主要技術路徑並行發展：視覺語言動作模型 (VLA)、世界模型，以及可同步預測狀態與動作的「世界動作模型」。以美國為首的技術陣營，高度崇尚 AI-Native 原生模型與虛擬物理模擬的「暴力美學」。正如 NVIDIA 執行長黃仁勳多次強調的 Physical AI 願景，美國頂尖機構大規模利用 Cosmos 等超級虛擬物理模擬平台，並結合「建模雙生」技術，提供低成本、大規模的虛擬演練與學習環境。透過合成數據 (Synthetic Data) 在 Omniverse 中對大模型進行數十億次的試錯訓練，追求模型在開源世界中的極致

表 1：具身大模型架構演進對比

比較	傳統模組化架構	端到端 VLA 模型	具身世界模型 (World Model)
資訊處理流程	感知 → 語義推理 → 運動規劃 → 關節控制 (存在嚴重的資訊損耗與延遲)	視覺 / 語音特徵直接映射至低階馬達扭矩指令 (One-Stage 決策)	具備物理法則與因果推理能力，能在內部沙盒中預演多種動作結果後再執行
空間感知方式	2D 邊界框 (Bounding Box) 或稀疏點雲 (Point Cloud)	結合高解析度多模態編碼器與基礎深度估計	4D Occupancy (體素化) 網路，包含動態時間序列與物理幾何推演
訓練環境依賴	重度依賴人工標註的真實環境數據，長尾場景適應力極差	海量互聯網數據預訓練 + 機器人本體跨域微調 (Co-training)	極度依賴如 Cosmos 等虛擬物理模擬平台生成的物理合成數據
適應能力	僅能執行固定腳本與預設工位任務	具備跨環境 (Cross-domain) Zero-shot 任務執行能力，可處理未見過的廚房等場景	物理世界自主泛化，能在災難救援、極端工況下進行自主生存與任務創新
典型架構	ROS 1/2 框架下的傳統決策樹與 MPC 演算法	RT-X, Pi-0.5, GROOT	NVIDIA GR00T(融入世界模型概念)，Sora (視覺預測延伸)

泛化與 Zero-shot 能力，並使建模雙生逐步發展為可即時同步、自主運作的動態世界模型。相對而言，中國的技術路線則展現出極強的場景實用主義，傾向採用「雲端大腦 + 邊緣小腦」的分層混合架構。此外，針對非紅供應鏈商機，臺廠及具備特定產業優勢的業者宜避開高成本的通用大模型，轉向鎖定「專用世界模型」，透過「真實到模擬、再回到真實 (Real-to-Sim-to-Real, R2S2R)」建立資料飛輪，推動軟硬整合與實體 AI 的場域落地。

「大腦」跟「小腦」架構在晶片硬體上的物理實現

演算法想法再漂亮，最後還是要靠硬體跑起來。具身智能從軟體演算法走向真實物理世界的過程中，計算架構面臨了前所未有的工程挑戰。為了解決複雜語義推理與超低延遲運動控制之間的矛盾，全球產業界在最近兩年，全面轉向「大腦決策層」與「小腦運控層」異構解耦的硬體物理實現方案。

「大腦」要跑 VLA 這類大模型，關鍵就是算力跟記憶體頻寬。NVIDIA DRIVE Thor 本來是給自駕車設計的晶片，現在很多人形機器人拿它當主大腦。FP8 稀疏算力高達 2000 TFLOPS，同時處理多台高解析度攝影機與 LiDAR 的 4D 占

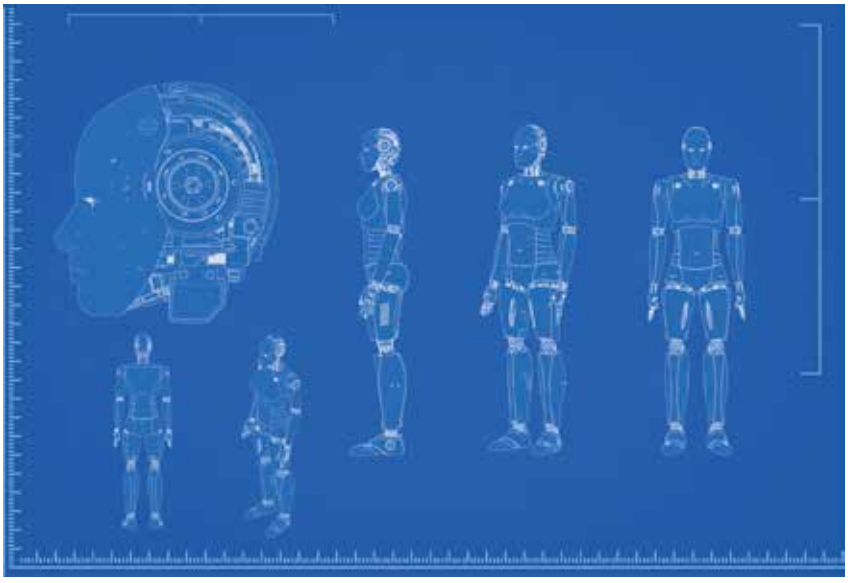
圖說：NVIDIA DRIVE Thor 用於汽車平臺的智慧大腦系統

資料來源：NVIDIA DRIVE Platform Official

用 數據流的空間數據，還要高速跑大模型推論。但高算力伴隨一個大麻煩：功耗接近千瓦 (kW) 等級。機器人體積很小，必須塞進散熱板甚至晶片級液冷，才不會因為過熱降頻宕機。

相對的，小腦不追求超強大算力，但講求極低、確定的延遲。維持機器人站穩、關節伺服、觸覺回饋，都需要微秒級反應，不能等主大腦慢慢算。近兩年邊緣 AI 微控制器 MCU 迎來大升級，各家半導體廠都在拚這塊。

相對於大腦專注於宏觀決策，小腦不追求超強大算力，但講求極低、確定的延遲。機器人的「小腦」運控層則掌管著維持雙足平衡、關節伺服控制與高頻觸覺回饋的命脈。在這個層面，毫秒甚至微秒級的絕對確定性 (Determinism) 比單純的算力更為致命。過去兩年間，邊緣 AI 微控制器 (MCU) 領域迎來了規格大升級的全面爆發，各家半導體廠都在拚這塊。意法半導體 (ST) 推出的 STM32N6 系列 MCU 堪稱此世代的經典之作，其內部首度內嵌了 ST 自主研發的 Neural-ART 邊緣人工智慧加速器。STM32N6 能夠在不喚醒主大腦的情況下，以極低的功耗在邊緣端獨立完成運動學反解與局部視覺 / 觸覺特徵的預處理，



圖說：STM32 家族第一款具有 AI 加速功能的 MCU

資料來源：STMicroelectronics STM32N6 Series

表 2：全球機器人小腦運控與邊緣 AI MCU 旗艦規格對標表

廠商型號	核心與時脈	邊緣 AI 加速單元	主要特色	機器人應用場景
ST STM32N6	高達 1 GHz (Cortex-M 系列進階架構)	Neural-ART accelerator	大幅降低機器視覺預處理延遲，減少大腦喚醒頻率，提升續航力	端側視覺預處理、局部避障決策與靈巧手觸覺感知網關
TI MSPM0G5187	Cortex-M0+ 高效能調優架構	TinyEngine NPU	將複雜馬達控制演算法的邊緣計算延遲降低近 90 倍	高精度關節伺服馬達驅動、分散式多自由度 (DOF) 同步控制
Renesas RA8T1	Arm Cortex-M85 (480MHz)	Helium 向量擴展指令集	相較 M7 核心 DSP/ML 效能提升 4 倍，支援實時預測性維護 (PHM)	全身動態平衡運算、高階 AC 伺服控制與底盤步態解算
NXP MCX N 系列	Cortex-M33 雙核心 (高達 150MHz)	eNPU 神經網路加速單元	提升 30 倍以上的機器學習推論效能，確保運控閉環極致安全	高安全級別姿態控制、運動學防碰撞干預與異常狀態隔離

的頻寬壓力，更將局部避障延遲壓縮至極致。

除此之外，其他半導體巨頭也紛紛在小腦運控晶片上推動了即時控制技術的內捲式躍升。德州儀器 (TI) 發布了搭載 TinyEngine NPU 的 MSPM0G5187 系列，透過硬體級神經網路加速，將傳統複雜馬達控制演算法的運算延遲降低了驚人的 90 倍，使得單一晶片得以同時駕馭更多自由度 (DOF) 的靈巧手關節。瑞薩電子 (Renesas) 則推出了 RA8T1 電機控制旗艦晶片，憑藉高達 480MHz 運作時脈的 Cortex-M85 核心與 Helium 向量擴展指令集，直接針對多軸交流伺服馬達進行高效能解算；其不僅具備強大的 DSP 與機器學習實測指標，還能即時執行預測性維護 (Predictive Maintenance) 的 AI 演算法。恩智浦 (NXP) 的 MCX N 系列微控制器同樣透過 eNPU 與邊緣運控閉環，為整車的動態平衡與微觀姿態調整提供了安全可靠的硬體底座。這些小腦 MCU 的共同演進，標誌著人形機器人的伺服驅動已全面從傳統的「純數學公式解算」轉向了「神經網路輔助 (AI-Native)」的全新紀元。

機器人內部通訊革命：不再瘋狂傳原始數據

大腦、小腦晶片都到位，還有一道坎：遍布全身的感測器、馬達要快速溝通。

在具身智能大模型從實驗室的虛擬沙盒走向真實物理工廠與家庭環境的進程中，算力與演算法的突破僅僅是完成了「大腦」的構建，而如何讓這顆大腦與遍布全身的高精密感測器、關節伺服馬達進行微秒級的無縫對話，則成為了決定機器人能否真正在物理世界自主泛化的終極工程瓶頸。

圖說：恩智浦 (NXP) MCX N 系列透過 eIQ Neutron 神經網路處理單元 (NPU)，以支援機器學習應用。

資料來源：NXP

長久以來，人形機器人的機身內部網路架構與物聯網底層邏輯，始終受制於 1948 年確立的香農定理 (Shannon Theory)，

表 3：機器人機身內部高頻數據流傳輸延遲與通訊協議拓撲表

通訊技術	核心技術	效果表現	應用場景
語義通訊 (Semantic Communications)	從比特傳輸轉向意圖 Token 提取，依賴端側 NPU 在射頻物理層 (SC-PHY) 執行特	原始感測流傳輸量降低 80% 以上，實測物理帶寬需求縮減高達 10~50 倍。	靈巧手觸覺異常回報、關節預測性維護警報、免喚醒語音意圖微秒級轉換。
AI 原生射頻 (Neural RF) 與異構協同	天線物理與微型神經網路在硬體晶圓層面直連，具備空間電磁感知本能，主動調整波束成形。	微秒級自適應干擾閃避，消除多徑金屬干擾與長尾延遲，實現移動中的零感卡頓。	克服機器人機身金屬骨架自遮擋與複雜車間環境下的無縫異構通訊。
傳輸預占機制 (TXOP Preemption)	賦予對延遲極度敏感的運動控制數據包「強行插隊」特權，可強制中斷並掛起後台大文件傳輸。	將偶發性的網路排隊延遲從毫秒級強制壓縮至微秒級，確保絕對確定性。	高動態雙足平衡數據流與頭部高頻空間計算 (XR) 渲染指令的極速下發。
空中計算 (AirComp)	利用物理電磁波在空氣中自然干涉的疊加原理，直接在類比域完成模型權重的加權平均計算。	幾毫秒內完成海量節點參數同步，物理層面絕對零排隊，無帶寬擠佔。	智慧工廠內數百台具身機器人的大模型權重實時協同更新與群體強化學習。

其核心目標是無差錯地傳輸每一個比特 (Bit) 數據。在這種傳統的「比特驅動」範式下，機器人末端的感測節點——例如指尖的高解析度觸覺陣列、關節處的 6 軸慣性測量單元 (IMU) 以及頭部的多路 4K 視覺鏡頭——往往扮演著極度「愚蠢的搬運工」角色。它們將海量的原始視頻串流、高頻微波震動波形與六維力矩信號，未經任何過濾與處理，直接透過機身內部的有線匯流排或傳統無線射頻，瘋狂地灌入機器人的主大腦或雲端智算中心。這種「先傳輸、後計算」的粗暴模式，在千億參數模型實時推論的龐大需求面前，不僅造成了災難性的通訊匯流排頻寬擠兌，更極大地消耗了機身寶貴的電池電能，導致關節控制神經網絡出現致命的長尾延遲。

為解決信息擁塞問題，邊緣人工智慧 (Edge AI) 下沉至無線射頻的基帶 (Baseband) 與媒體存取控制 (MAC) 底層，推動從比特 (Bit) 傳輸全面邁向語義 (Semantic) 通訊的技術進步。

現在技術方向改變，叫做語義通訊：把 AI 下沉到感測器、射頻晶片底層。感測器不再傳送一長串原始數據，本地微型 NPU 先做分析，只送出濃縮過的關鍵資訊。舉個例子：手撞到東西，不傳完整震動波形，直接送出一句濃縮資訊：「手腕位置發生撞擊，軸承故障機率 99%」。根據 IEEE 通訊學會發布的權威白皮書實測數據顯示，傳輸資料量可以縮減 10-50 倍，頻寬壓力、耗電同時大幅下降，上百個感測節點也能常駐運作。

另外工廠環境充滿金屬機台、馬達電磁干擾，無線訊號很容易被擋住。傳統通訊遇到干擾只能重傳，對於需要即時控制的機器人來說很危險。AI 原生射頻 Neural RF 就解決這個問題：神經網路直接跟射頻晶片綁在一起，即時偵測周遭電磁環境，預判干擾，主動調整訊號發送方向。再搭配 Wi Fi8 的傳輸搶占機制，關鍵運動指令可以優先插隊傳送，把延遲鎖死在微秒等級，就算環境干擾嚴重，動作也能順暢不卡。

除了頻寬與能耗的挑戰，人形機器人在真實的工業智造車間或複雜的家庭環境中，還面臨著極為嚴苛的電磁干擾考驗。未來工廠不會只有一台機器人，而是幾百台一起協同工作。大家要互相分享現場學到的經驗、更新模型。如果每台機器人都把完整模型丟上雲端，網路直接會被塞爆。這就帶出「空中計算 AirComp」的構想：多台機器人在同一時間、同一頻段同時發送更新參數，利用電磁波在空中自然疊加，接收端拿到的訊號就等於多台機器人參數的平均值。等於直接在空氣裡完成模型聚合計算，不用排隊，幾毫秒就完成多機協同訓練，也不會洩露單台機器人的資料。通訊網路不再只是傳輸線，變成機器人的智慧神經。

市場空間與安全防護

在過去一兩年的產業實踐中，工程界與學術界深刻地認識到，端到端大模型雖然在處理非結構化環境時展現出驚人的直覺，但其本質依然是一個由數百億個神經元權重交織而成的龐大「黑盒」。

根據 IEEE 權威技術報告的深入剖析，這種黑盒架構面臨著致命的「缺乏可解釋性 (Explainability)」危機。與傳統基於規則 (Rule-based) 的 C++ 程式碼不同，當端到端模型輸出一個錯誤的關節扭矩指令時，人類工程師幾乎無法逆向追蹤出到底是哪一層神經網路、哪一個特徵維度導致了這次災難性的決策。

更為嚴峻的是，具身大模型極易遭受惡意對抗樣本 (Adversarial Attacks) 的干擾——例如，僅僅在一個搬運紙箱上貼上一張經過特殊設計的對抗噪聲貼紙，就可能誘發模型產生嚴重的「物理因果錯覺」(Causal Illusions)，導致機器人誤判物體的重量與深度，進而施加足以捏碎物體的破壞性力量，甚至引發機器人偏離既定軌跡、衝撞人類作業員的幽靈失控事件。這種由演算法黑盒引發的偶發性失控，不僅為後續的事故責任認定與保

險理賠帶來法律挑戰，更在社會學層面引發了公眾對與實體機器人「人機共存」的深度信任危機。

為了解決這一致命的黑盒效應與人機交互錯配帶來的安全隱患，全球機器人產業在軟體架構層面需要一場深刻的底層重構，全面引入基於確定性物理法則的「安全防護欄 (Guardrails)」隔離架構。這意味著，在未來的具身智算系統中，大模型的開放性創新與實體機械的破壞力之間，被強制設置了一道不可逾越的物理防火牆。

具體而言，當端到端的 VLA 模型生成了一系列複雜的連續動作指令矩陣後，這些指令並不會被立刻下發給底層的馬達驅動器；相反，它們必須先輸入至一個由傳統模型預測控制與嚴格逆運動學 (IK) 方程構成的「安全沙盒」中進行高頻校驗。這個防護欄系統會以微秒級的速度，實時計算大模型給出的扭矩向量是否會導致機器人的零矩點 (ZMP) 移出支撐多邊形 (即引發摔倒)、關節角速度是否超出了機械應力的極限撕裂閾值，以及本體的運動軌跡是否突破了與周圍人類的安全隔離半徑。一旦防護欄系統在虛擬沙盤中推演出任何潛在的物理崩潰風險，這套確定性的隔離架構將瞬間接管最高控制權，否決大模型的概率性決策，並強制將機器人切換回安全的降級運行模式 (例如啓動電子阻尼進行原地鎖死或緩慢復位)。

如果順利，在透過硬核的安全隔離機制成功克服了初期的落地陣痛後，具身機器人軟體層的潛在商業價值在資本市場迎來快速增長。傳統的工業機器人產業長期深陷於單純比拼金屬機構件製造成本的「低價內捲」泥淖，而新一代人形機器人的核心利潤池，已向上游的作業系統授權與大模型演算法訂閱轉移。

首先，隨著硬體形態的高度標準化與模組化，機器人作業系統 (Robot OS) 的商業授權將成為少數基礎軟體寡頭極為穩定收入來源。針對高階無塵室、生化醫療或危險特種作業等垂直場景，經過嚴格功能安全認證 (如 ISO 13849) 的閉源商業版 Robot OS，其單台授權費率將為軟體供應商貢獻深厚的利潤護城河。

更具顛覆性的是「硬體即服務 (HaaS, Hardware as a Service)」與具身基礎大模型訂閱制 (SaaS) 的強勢崛起。在此創新商業模式下，企業客戶在引入人形機器人勞動力時，無需再承擔高達數萬美元的整機與軟體一次性買斷成本 (CAPEX)。相反，硬體製造商甚至可能採取「硬體零利潤」乃至「極低押金」的激進補貼策略將實體機器人鋪入工廠，轉而透過按月或按年向客戶收取全棧服務的訂閱費 (OPEX)。這筆訂閱費不僅涵蓋了機器人的日常維保，更核心的是購買了雲端大模型持續的「智力升級」。

透過不斷攝取全球部署節點回傳的物理世界邊緣數據，雲端智算中心能夠高頻迭代模型的神經網路權重，並為終端機器人解鎖新的技能包 (如從簡單搬運升級為精密焊接) 與特定場景的微調優化服務。當一家企業能夠透過軟體定義的持續升級，讓實體機器人的資產價值不降反升時，它便打開了千億美元級別的軟體服務與數位勞動力市場的空間。 CTA