# 研究問題與動機：VTLA 模型之發展

觸覺視覺語言行動模型（Vision-Tactile-Language-Action Model）代表機器人感知與控制技術的重大突破。本計畫以精密製造業為核心應用場景，整合視覺、觸覺、語言理解與機械控制四大模態，建立智能自動化系統的新範式。透過觸覺感測器與高分辨率視覺感測，結合大型語言模型的推理能力，VTLA 模型將使工業機器人達到人類水準的靈巧度與適應性，在高精密製造、品質檢測、複雜組裝等領域創造30-40%以上的效能提升。

## 1.1 研究背景與問題缺口

現有工業機械手主要依賴視覺伺服與簡單力控，無法應對以下場景：

- 視覺遮擋下的盲目取料（如線纜插入、微小零件組裝）
- 物體特性的實時感知（柔度、表面紋理、材料特性識別）
- 接觸過程中的即時反饋與自適應調整
- 滑動檢測與過力保護

結果是機械手需要多次重複嘗試、重新夾取同一物體，造成周期時間延長、成品率下降。而在 peg-in-hole 任務中將面臨逐步升級的技術挑戰：

- 單孔組裝：首要克服多模態融合可能破壞 VLM 特徵的問題，力/觸覺模擬與真實環境的巨大落差。
- 多孔組裝：面臨操作過程中的視覺遮擋及序列任務中的累積誤差。
- 不同材質的插件（軟性排線）：因軟性材料模擬困難、接觸力微弱難辨，且操作不當易導致排線損壞。

## 1.2 解決方案：整合物理先驗與混合專家機制的 MVLA 系統

為解決從剛性孔軸到軟性 FFC 排線組裝過程中面臨的感知瓶頸、環境不確定性及高損壞風險，本計畫提出一套整合正在開發的 MVLA（Material-Vision-Language-Action Model）系統加上觸覺感知的閉環模型。本系統透過三大核心技術，實現從「被動反應」到「主動預測與適應」的技術跨越。針對組裝環境中的位姿偏差與感測器雜訊。

本計畫不只依賴視覺，更引入多模態架構進行特徵融合：

接觸前預測

- 利用視覺訊號在接觸前識別材質，並將觸覺、視覺、語言、材質送入語言模型(Fig.1)，以估算物體體積與物理屬性。
- 根據任務階段(如：接近、接觸、插入)動態調整視覺、語言與力覺專家的權重。在視覺受遮蔽時，自動提升力覺專家的決策比重。

時序增強推理與動態 Token 編碼

- 針對多孔組裝任務中對連續動作的高精度要求，引入時序增強機制解決通用 VLM 在處理高頻物理互動時存在的反應遲滯與非連續性問題。

- 賦予模型對動態過程的短時記憶與推理能力，確保機器人在操作時能實現精確、連貫且穩定的組裝動作。

安全導向的主動適應控制

採用雙重優化策略以確保操作安全性：

- 自適應強化學習力控：利用強化學習機制，鼓勵機器人在不確定環境中進行穩健探索。系統依據即時力覺回饋，以主動適應軟性材料的非線性變形與接觸力變化。
- 安全偏好對齊機制：引入偏好學習技術，將專家的安全操作經驗轉化為模型的監督訊號使機器人不僅學習路徑，更能掌握類似人類的接觸姿態，有效降低精密組件在操作過程中的損壞風險。

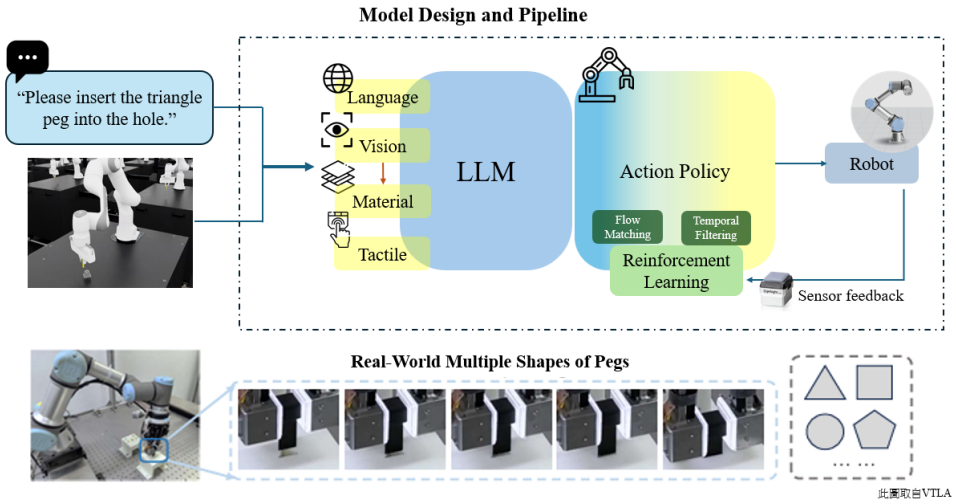


Fig.1 VTLA 模型架構圖

## 2. 研究內容、進行步驟及執行進度

本計畫擬分三年期，建構具備「材料感知」能力的觸覺-視覺-語言-動作（MVLA）機器人模型，由剛性物體組裝逐步拓展至柔性物體操作。

### 第一年：插銷組裝（Peg-in-Hole Assembly）

目標: 建立具備材質感知與時序推理能力的多模態 VLA 基礎模型

建構多模態 VLA 基礎模型

- 整合視覺（Vision）、觸覺/力覺（Tactile/Force）與語言（Language）與材質（Material）輸入。設計視覺引導的時序增強機制，將接觸過程中的連續時序資訊編碼為具備時間依賴性的特徵，以解決模型對動態過程反應遲鈍的問題。
- 引入 6 軸力矩感測器，轉換為與視覺語言模型兼容的嵌入向量，使力覺成為模型的第一類模態。

模擬環境與數據集建置

- 建立高保真度的機器人模擬環境，收集包含視覺影像、觸覺、、力覺及材質數據以及自然語言指令的同步組裝數據。
- 利用域隨機化技術（如隨機化摩擦係數、光照條件、相機視角），以縮小模擬與真實世界物理特性的差異，利於後續 Sim2Real 遷移。

模型訓練與驗證:Peg-in-Hole 基準測試

以不同公差（0.1-0.5mm）與材質形狀組合進行單孔組裝，驗證模型對材料屬性的預測與力量修正能力。

預期查核點

- 完成 MVLA 系統架構搭建與模擬環境(Isaac Sim)部署(如 Fig.2 所示)。
- 進行指令微調，使模型能根據自然語言指令（如「Insert the peg」）執行相應的動作控制。
- 單孔組裝任務：成功率 90%，接觸力過衝降低 15%以上。

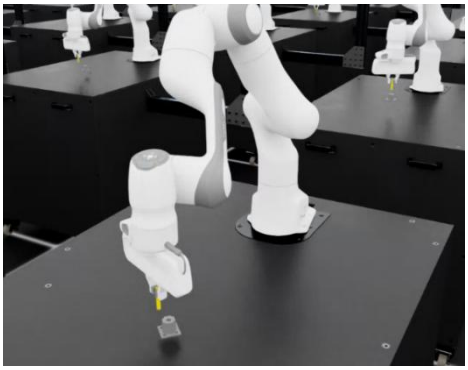

Fig.2 使用 Nvidia Isaac Sim 做為模擬平台

**第二年：多孔物件組裝（Assembly of Multiple-Hole Objects）**

**目標**：提升序列任務的幾何泛化能力、接觸穩健性與異常恢復機制。

多感知混合專家與分層規劃機制

- 動態路由與專家分工：開發動態路由機制，依據任務階段與視覺遮蔽程度，自適應切換視覺語意與精細力覺專家權重。

- 分層強化學習架構: 高層策略以已完成的孔位一為物理基準，利用相對座標規劃孔位二路徑；低層策略則重置狀態，專注於單孔的精細力回饋控制。此設計利用物理約束消除累積誤差，確保序列任務多孔組裝的精準度。

複雜接觸動態與魯棒性控制

- 抗遮蔽與盲插調整：在視覺受限的情況下，相機可能被機械手臂本體、夾爪或工件遮擋，導致無法實時獲取孔位的當前位置，導入「時序濾波機制」利用過去的軌跡狀態，預測當前不可見的狀態，推算當前孔位應該在哪裡。

- 多點觸覺融合與異常恢復：針對多孔同時接觸的複雜力學約束，開發多點觸覺資訊融合與異常檢測機制。一旦偵測到任一接觸點受力異常，即觸發回退或重規劃策略，賦予系統自我修正的能力。

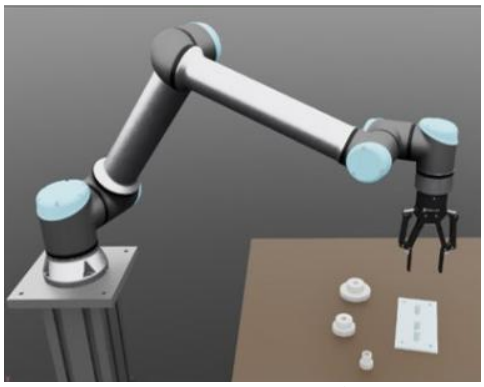

Fig.3 使用 Isaac Sim 模擬多孔物件場景

預期查核點

- 任務完成率：實現多孔位的連續組裝任務(場景示意如 Fig.3)，序列完成率達 80%以上。
- 錯誤恢復效率：建立異常檢測與自我修正能力的控制，平均錯誤恢復時間少於 3 秒。

**第三年：柔性扁平電纜（FFC）組裝（Flexible Flat Cables Assembly）**

**目標**：基於 MVLA 主動式力控，解決可變形物件操作與高精度接觸動態。

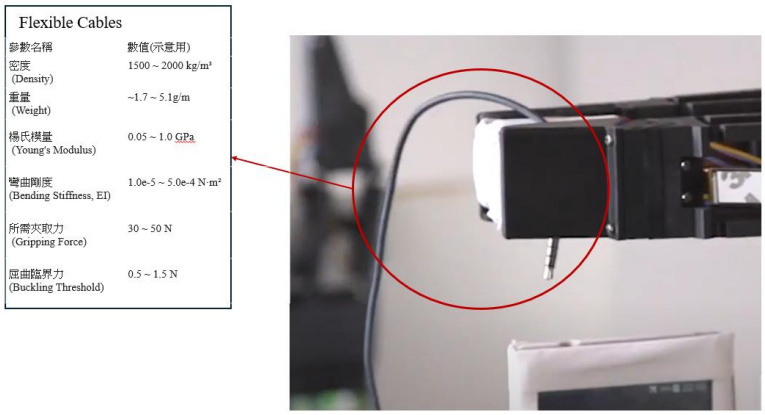


Fig.4 基於 MVLA 預測物理性質加強力控

導入 MVLA 偏好導向的安全力控

針對軟性排線（FFC）的高精度與低容錯需求，MVLA 系統超越單純的軌跡複製，轉向基於偏好學習的力控策略學習。

- 建立安全力覺邊界：收集包含力覺特徵的數據集(如 Fig.4)，標註人類專家在操作軟性材料時的安全邊界。利用微調模型，使 MVLA 不僅學習完成任務，更能掌握符合專家偏好的接觸力道與姿態，主動避免因施力過當導致精密排線損壞。

基於 Flow matching 的預測式柔順動作生成

- 材料感知的軌跡規劃：結合 MVLA 的材料識別能力，針對 FFC 易挫曲的物理特性，使用 Conditional Flow Matching 來學習動作的流場。當機器人觀察到目標，模型會生成一個從當前姿態平滑過渡到目標插入姿態的軌跡。
- 連續力/位混合生成：模型不僅生成位置軌跡，更同步生成連續且平滑的預期力矩軌跡。這能消除傳統控制中常見的瞬間加速度突波，確保機器人在插入過程中保持動作的連續性與柔順性，防止軟性線材因受力突變而變形。

混合專家導向的即時阻抗適應

- 主動式接觸動態調節：針對插座晃動或排線回彈等不確定性，MVLA 利用混合專家架構動態融合視覺與高頻力覺訊號。
- 動態阻抗控制：根據實時接觸狀態，系統能即時調整控制器的阻抗/順應參數。當偵測到異常接觸力時，模型會主動切換至高順應模式，實現對複雜接觸動態的即時適應，確保在非結構化環境下的操作安全性。

**預期查核點**

- 完成 FFC 插入或佈線任務之原型系統驗證。
- 實現無損操作：在操作過程中材料永久形變（損壞）< 10%。
- 開源包含剛性與柔性特徵的 MVLA 訓練數據集與模型權重。

## 3. 預期完成工作項目及成果

第一年：MVLA 基礎模型與單孔組裝驗證

- 完成整合視覺、力覺與材質感知的 MVLA 模型架構，並於 Isaac Sim 部署高保真模擬環境。
- 實現 Sim-to-Real 遷移，單孔插銷組裝成功率達 90% 以上。

第二年：多感知與序列控制

- 提出多感知混合專家機制與分層規劃策略，解決多孔位連續組裝之幾何泛化問題。
- 具備異常檢測與自我修正能力，多孔連續組裝成功率達 80%，平均錯誤恢復時間 < 3 秒。

第三年：柔性物件安全力控與開源貢獻

- 結合流匹配（Flow Matching）與偏好學習，實現柔性扁平電纜（FFC）之高精度無損組裝。
- 柔性材料操作損壞率< 10%；開源釋出包含力覺特徵的 MVLA 數據集與模型權重。

參考資料:

[1] C. Zhang et al., "VTLA: Vision-tactile-language-action model with preference learning for insertion manipulation," arXiv preprint arXiv:2505.09577, 2025.

[2] J. Yu et al., "ForceVLA: Enhancing VLA models with a force-aware MoE for contact-rich manipulation," arXiv preprint arXiv:2505.22159, 2025.

[3] Y. She et al., "Cable manipulation with a tactile-reactive gripper," arXiv preprint arXiv:1910.02860, 2020.

[4] S. Wang et al., "FlowRAM: Grounding flow matching policy with region-aware mamba framework for robotic manipulation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025.