國立陽明交通大學
機械工程學系
徐顥瑄
專題研究: Material-Vision-Language-Action (MVLA)

研究計畫:

以世界模型為基礎，整合多模態VLA模型, 建立人機協同框架

研究動機與背景

大型語言模型（LLM）賦予機器「理解」語言的能力，而VLA（Vision-Language-Action）模型則進一步讓機器具備「行動」的能力。VLA的應用已從傳統機械手臂拓展至智慧車輛[1]（如BMW、Tesla）、人形機器人（如NVIDIA GR00T、Helix），並逐步延伸至零售業與家庭服務等多元場域。

然而，現有VLA模型存在著關鍵限制:

- 將動作視為單向輸出，缺乏對動作本身及其對環境影響的深度理解; 在長時序任務中易出現誤差累積
- 多針對單一機器人平台設計，缺乏跨平台部署能力，限制異構機器人環境的應用。

1. 世界模型(World Model)似乎提供了潛在的解方，能預測未來狀態，模擬環境動態，實現前瞻性決策。WorldVLA[2]展示了VLA與世界模型整合的潛力，透過雙向增強機制，提升整體性能。
2. 人機協同: 製造環境中,機器人要自主執行任務並與人安全、高效協作。現有研究[3,4]提出了階層式認知架構與多模態互動方法，但仍缺乏整合世界模型預測能力與VLA語義理解能力的統一框架。

核心動機是建立一個**人機協同架構**，使機器人系統能夠理解人類意圖, 預測環境演變，實現安全、靈活、可泛化的智慧製造應用。

本研究旨在開發一個創新的人機協同架構，具體目的包括：

理論/技術目標:

- 融合世界模型與VLA模型的統一架構，強化動作理解與環境預測能力
- 建立多模態感知融合機制，整合視覺、語言、觸覺與本體感覺資訊
- 基於開源VLA模型(如OpenVLA[5])建立可微調框架，降低訓練成本
- 開發人機協作安全機制，結合預測模擬與多模態意圖識別
- 在製造場景中測試複雜任務執行能力(如組裝、物料搬運, 老年看護等)
- 評估系統在人機協作中的安全性與效率

學術貢獻:

- 產出碩士學位論文，完整記錄架構設計、實作細節與實驗結果
- 發表1-2篇國際會議論文與1篇期刊論文
- 提供開源實作程式碼與資料集

研究範圍:

核心技術範圍	■ 世界模型：預測未來視覺狀態與環境動態 ■ VLA模型：以小型開源模型為基礎進行微調，支援多模態理解 ■ 人機協同：實作多模態意圖識別(語音、手勢、視線)與安全評估機制
平台範圍	■ 實作平台：輪型機器人(如UR系列+AMR、Franka系列+AMR；達明輪型機器人) ■ 模擬環境：使用LIBERO、RoboSuite等標準化模擬平台
任務範圍	■ 簡單任務：物體抓取、放置、推動(驗證基礎功能) ■ 複雜任務：組裝操作、協同搬運、動態障礙物迴避 ■ 人機協作：人類指令理解、協同任務分配、安全監控

研究困難/對應方法:

預期困難	可能解決方案
世界模型與VLA模型的訓練目標不同 如何在單一架構中實現有效整合	■ 採用**WorldVLA的聯合訓練框架**，設計共享詞彙空間 ■ **分階段訓練**：先訓練世界模型，再整合VLA，最後聯合微調
真實機器人資料收集成本高，標註困難	■ 使用開源資料集(Open X-Embodiment, LIBERO等) ■ 採用**Sim2Real**技術生成訓練資料
人機協作環境中需確保人類安全	■ 實作預測性安全評估，使用世界模型預測潛在碰撞 ■ 多層次安全架構：感知層(障礙物檢測)、規劃層(路徑優化)、執行層(力, 力矩控制)
整合模型推理速度慢，難以實現即時控制	■ 使用**模型量化技術**減少記憶體需求 ■ 整合部署: 平均分配運算比例

預期目標:

短期目標:

- 接續大學專題完成 MVLA 系統實驗，應用Sim2Real 技術擴增資料集，提升模型泛化能力, 同時透過實體機械手臂進行性能測試與資料蒐集
- 投稿至 2026 IROS

碩士班第一年:

- 整合開源VLA預訓練模型
- 將世界模型與VLA連接為統一架構, 實作基本的動作生成與預測功能
- 整合深度相機、力矩感測器、開發多模態融合演算法(早期/晚期/混合融合)

碩士班第二年:

- 建構多模態融合架構，部署於家居照護型機器人
- 實作語音/手勢指令識別, 開發預測性安全評估系統
- 論文撰寫與投稿

長期目標(博士階段)

- 深化世界模型為基礎，建立跨平台人機協同架構
- 實現多機器人間的訊息共享與協作行動能力
- 貢獻: 深入投入 AGI 與人機互動領域，產出具社會影響力的研究成果

參考文獻

[1] A Survey on Vision-Language-Action Models for Autonomous Driving** (Sicong Jiang et al., 2025)

[2] WorldVLA: Towards Autoregressive Action World Model** (Jun Cen et al., 2025)

[3] HMCF: A Human-in-the-loop Multi-Robot Collaboration Framework Based on LLMs** (Zhaoxing Li et al., 2024)

[4] Collaborative Conversation in Safe Multimodal Human-Robot Collaboration** (Davide Ferrari et al., 2024)

[5] OpenVLA: An Open-Source Vision-Language-Action Model** (Moo Jin Kim et al., 2024)

讀書(學習)計畫

我從小對機器人就很有興趣,因此希望能進入學程學習各位老師專長, 使我能深入探討多模態融合、模型部署與行為評估等議題，並具備獨立規劃實驗與分析結果的能力

- 王傑智教授：在自駕車與機器人領域豐富的研究經驗，可以引導我將VLA部屬在多樣化載具
- 楊古洋教授：專精於基於舒適度的運動引導與力控制，具備多重感測融合與VR/機器人整合經驗
- 林顯易教授：擅長適應性力控制，指導我實現根據材質特性調整抓取力道的技術

除了學習老師們的專長外, 為了完成整合VLA研究計畫, 我制定了以下讀書計畫:

A. 短程階段：大學四年級 → 碩士班入學前

目標鞏固基礎、完成大學專題、投稿國際會議為研究打下扎實技術基礎。
理論學習與工作項目如下表:

主題	學習內容	實踐項目
完成MVLA系統實驗	模型訓練與效能驗證	實體手臂進行材料分類與損傷辨識
Sim2Real資料擴增	平台功能學習	模擬環境 → 實體手臂資料轉換
平台功能學習	Nvidia Isaac Lab平台操作	建立模擬場景、導出資料

B. 碩士班第一年：奠定研究基礎, 初步架構整合與核心功能實現

課程/研習內容	實踐項目
機器人學	手臂運動控制、感測器整合、ROS模組建置
深度學習含實驗	CNN模型訓練、影像辨識、分類任務
自主駕駛車技術、自走式機器人	路徑規劃、物件偵測、模擬實驗、提前學習面對未來部屬不同載具

自學資源

- **論文精讀**：WorldVLA、GAIA-1、OpenVLA、RT-2、ChatVLA-2, etc.
- **模型技術**：世界模型-VLA聯合訓練架構, 多模態融合演算法開發, 多模態意圖識別, 預測性安全評估系統
- **開發框架**：PyTorch、JAX、Hugging Face、ROS 2、NVIDIA IsaacLab等系列
- **資料集**：Open X-Embodiment、LIBERO、BridgeData、語音/手勢公開資

C. 碩士班第二年：加深研究, 優化模型並與國內外頂尖實驗室交流

課程/研習內容	實踐項目
感測與智慧系統	整合多模態感測器（影像、力覺、溫度等）於機器人平台
深度學習	訓練CNN/RNN/Transformer模型進行分類、預測、生成任務
論文撰寫	撰寫論文、整理研究成果、投稿至研討會（如**IROS**、**ICRA**）、尋找國際合作等機會
	註：已於大三修畢「感測器原理與量測系統」

進度檢查:

- 與指導教授每月進行進度報告，動態調整研究方向
- 持續追蹤VLA/世界模型領域的最新進展，及時納入研究設計