

講 題：科研整合與成果轉譯的關鍵挑戰——
制度、平台與產業鏈結

以台灣多模態大型語言模型開發為例
中央研究院資訊所廖弘源



- Gemma 4 E2B: everything is normal
- Ours: The person in the yellow shirt falls onto their back and then rolls onto their side on the trampoline.

- Gemma4 基本上是一個 black box 。
- 為了實現主權 AI，台灣應該盡量用自己研發的基礎模型來建構：
 - VLM (visual language model): TaiwanGBA，YOLOv7，YOLOv9，G-YOLO (generalist-YOLO)，Depth Recovery Model(?)
 - LLM (large language model): TAIDE
 - TTS (text to speech model) : 與 TAIDE 連結
 - ASR (automatic speech recognition model): 與 TAIDE 連結

台灣多模態大型語言模型之架構

上游:基礎模型

TAIDE
TaiwanGBA
YOLOv7 or
YOLOv9
Gemma4
G-YOLO

...

中游:大範圍應用

以行為辨識為基
礎之照護系統

多方位保全系統

公共場域之安全
系統

...

下游:特定範圍應用

家庭、醫院、浴室
之照護

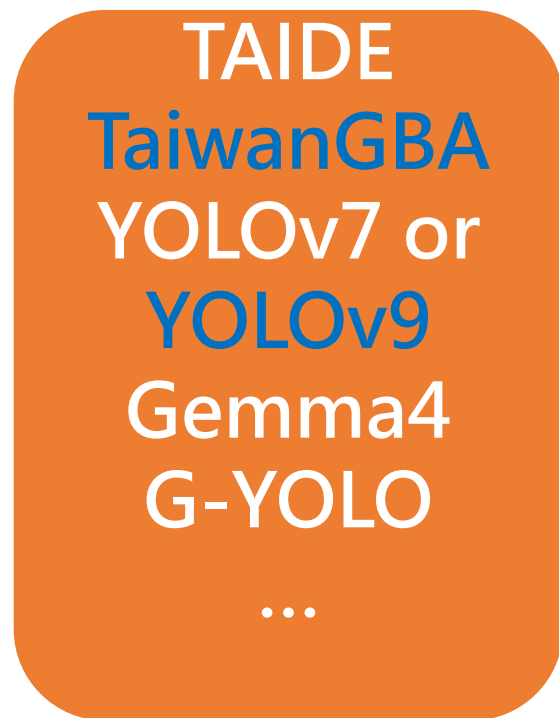
住家、公司、公家
機構之保全系統

戶外公共場域之危
險偵測

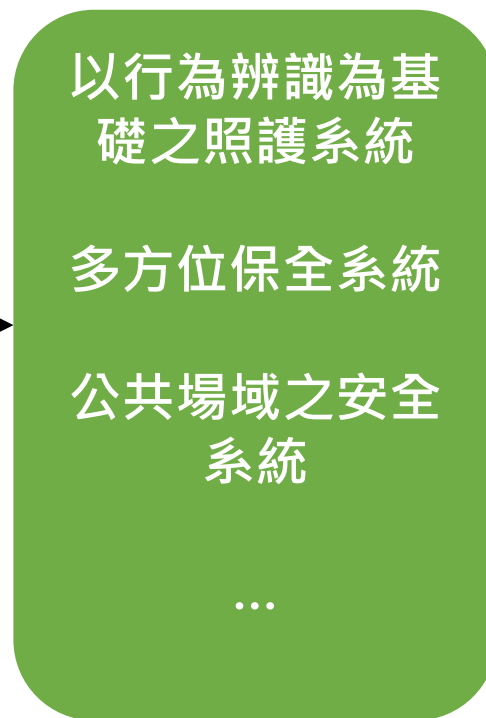
一個實際的應用例子

台灣多模態大型語言模型之架構

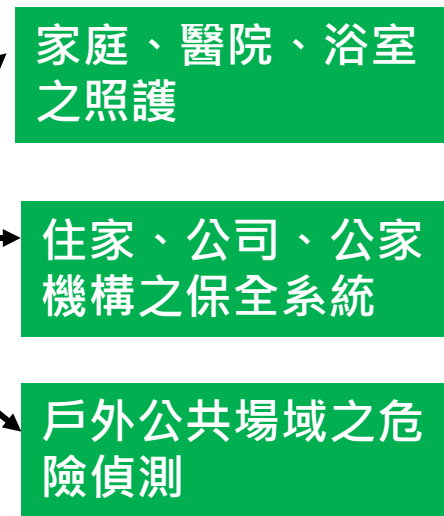
上游:基礎模型



中游:大範圍應用

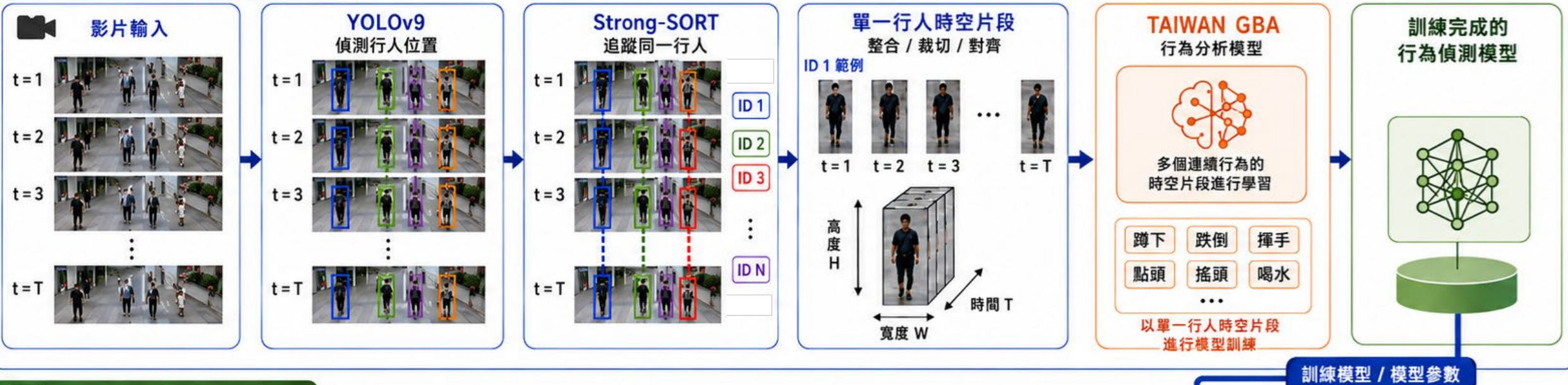


下游:特定範圍應用

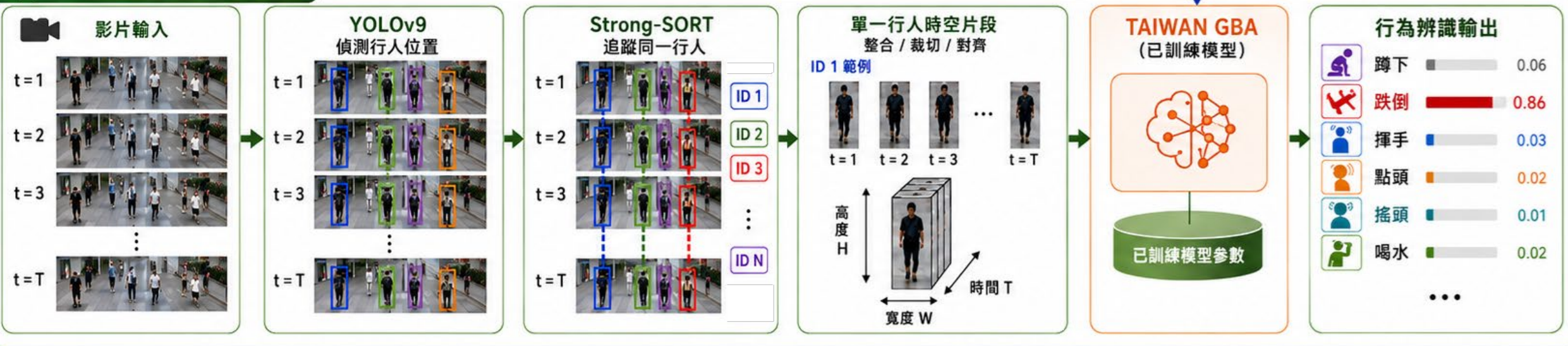


YOLOv9 + TAIWAN GBA 行為辨識流程示意圖

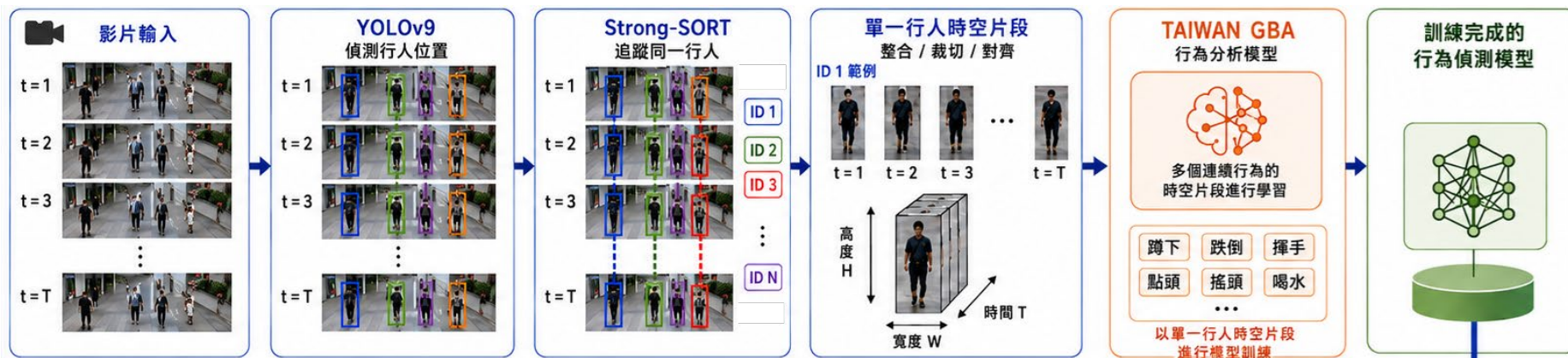
Off-line 端 (模型訓練)



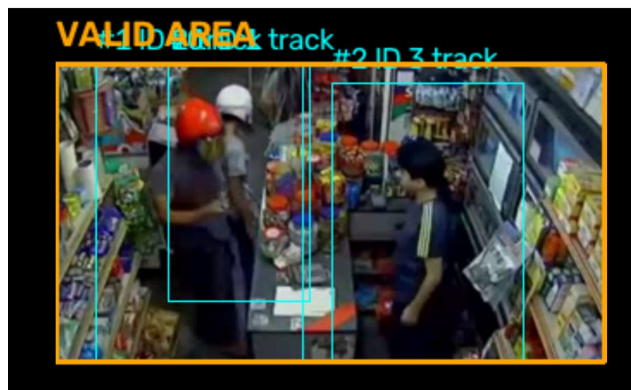
On-line 端 (即時辨識)



YOLOv9+TaiwanGBA模型訓練



(影片輸入)



(YOLOv9偵測行人+
Strong-SORT行人追蹤)

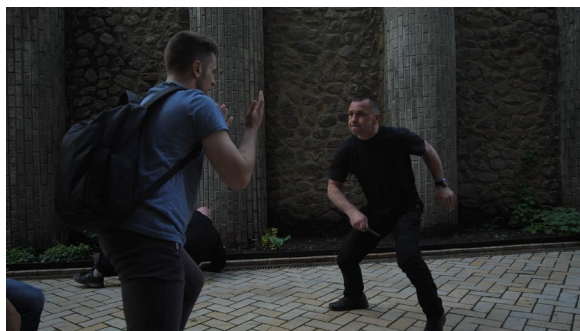
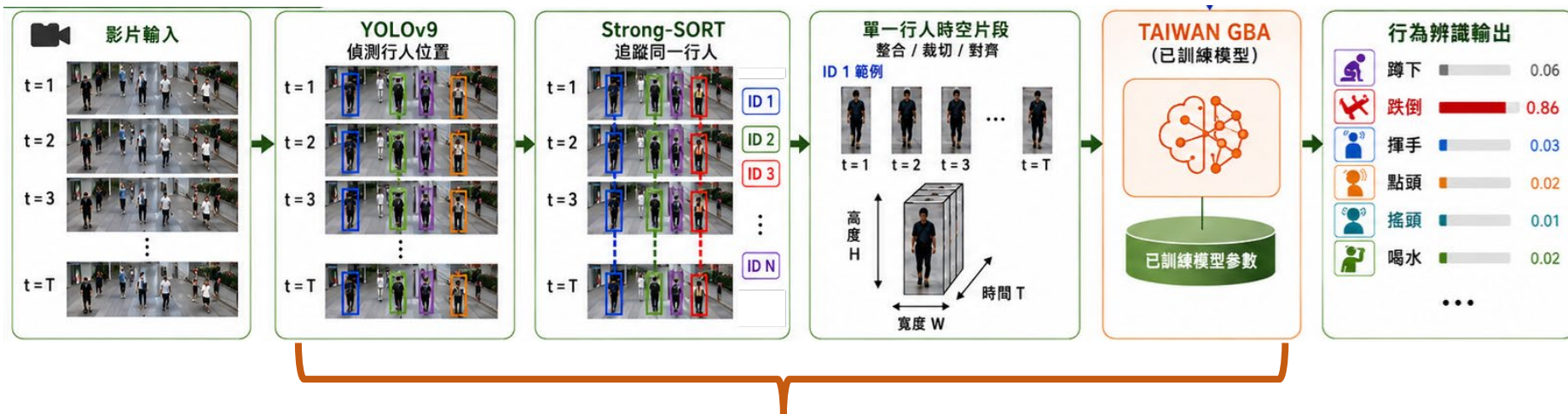


(分割出不同ID行人
動作影片)

(訓練TaiwanGBA)

使用YOLOv9+TaiwanGBA系統辨識行為

TaiwanGBA行為辨識系統執行流程

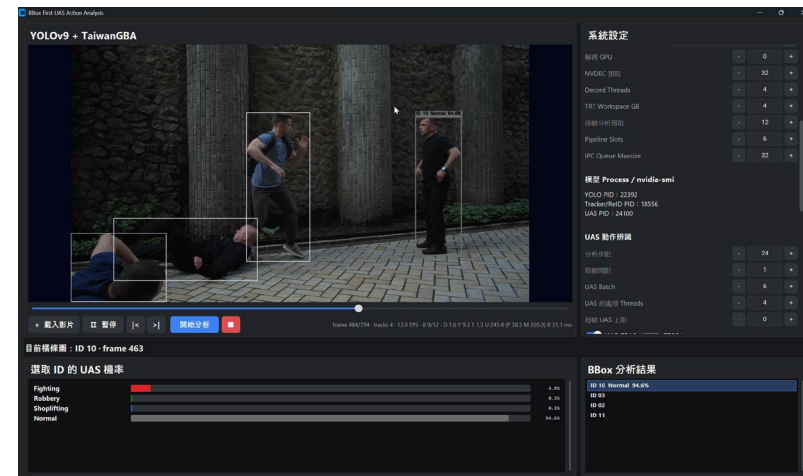


(影片輸入)



Asus ROG Strix SCAR 18電競筆電

TaiwanGBA
系統



TaiwanGBA系統

- TaiwanGBA辨識系統特點：
 - Edge AI: 全系統可獨立在筆電上執行，不需連接伺服器。
 - 高效率動作辨識：在電競筆電辨識速度可達12FPS。

最好能建多一些屬於本土的基礎模型

台灣多模態大型語言模型之架構

上游:基礎模型

TAIDE
TaiwanGBA
YOLOv7 or
YOLOv9
Gemma4
G-YOLO

...

中游:大範圍應用

以行為辨識為基
礎之照護系統

多方位保全系統

公共場域之安全
系統

...

下游:特定範圍應用

家庭、醫院、浴室
之照護

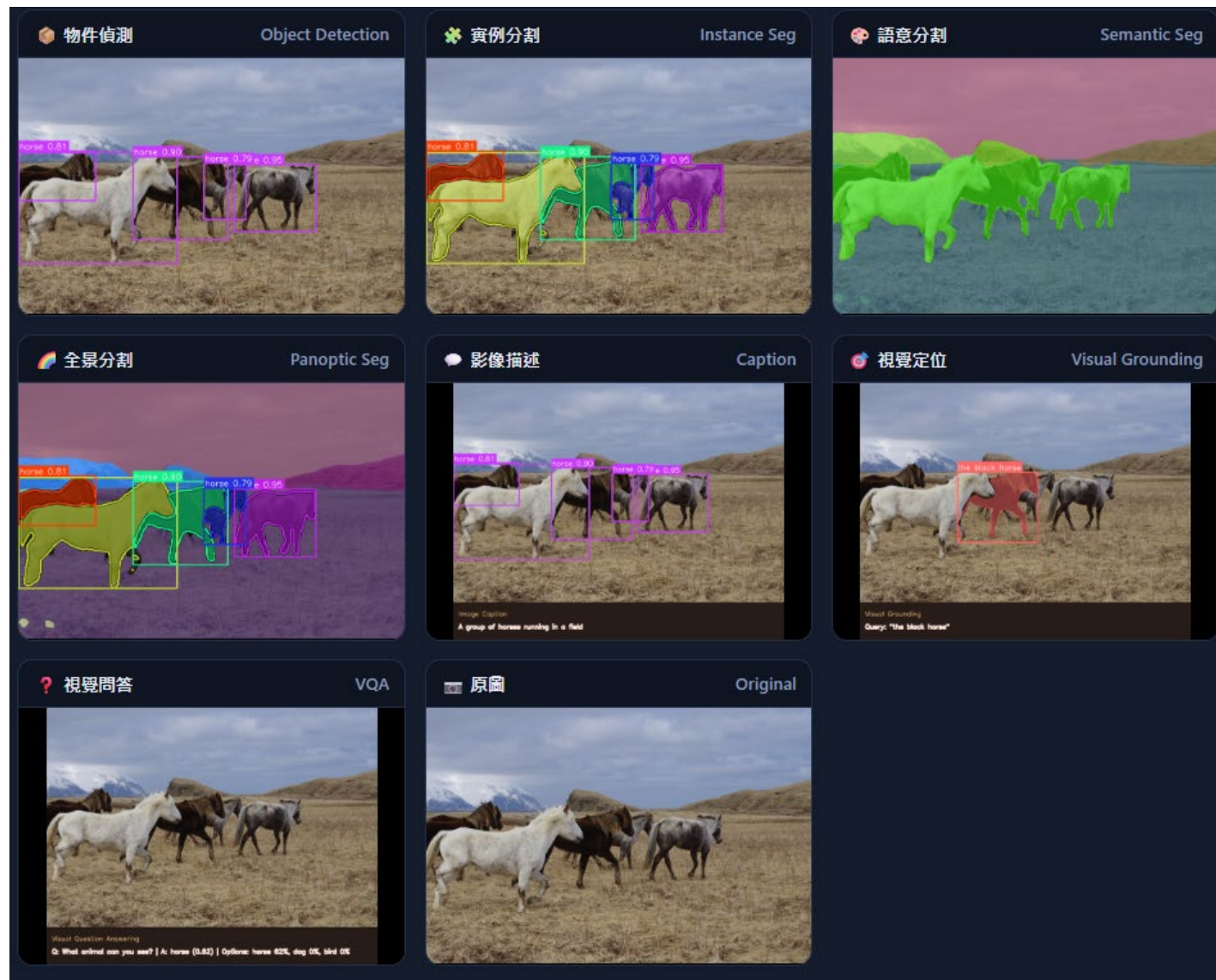
住家、公司、公家
機構之保全系統

戶外公共場域之危
險偵測

GYOLO Foundation Model

7 CV Tasks for common images

1. Object Detection
2. Instance Segmentation
3. Semantic Segmentation
4. Panoptic Segmentation
5. Image Captioning
6. Visual Grounding
7. Visual QA

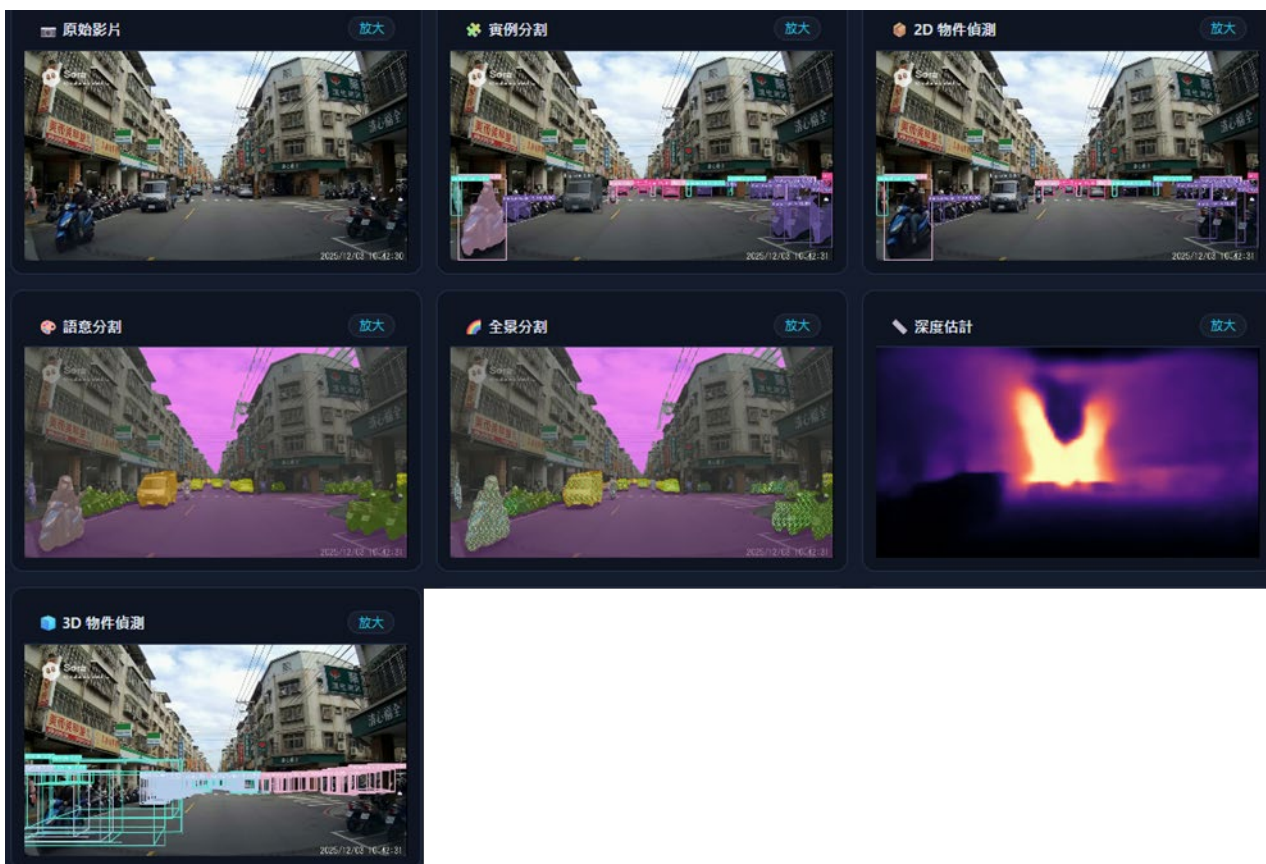


原本 YOLOv7 只能做 Real-time Object Detection, Instance Segmentation, 及 Pose Estimation。

功能擴充後的 Generalist YOLO (簡稱 G-YOLO) 具 Object Detection
、 Instance Segmentation 、 Semantic Segmentation 、 Panoptic Segmentation 、 Image Captioning 、 Visual Grounding 、 Visual QA 等七種功能，它更能夠支援中下游資服業者開發百工百業所需的系統

GYOLO Midstream Model

Finetune GYOLO to specific domain

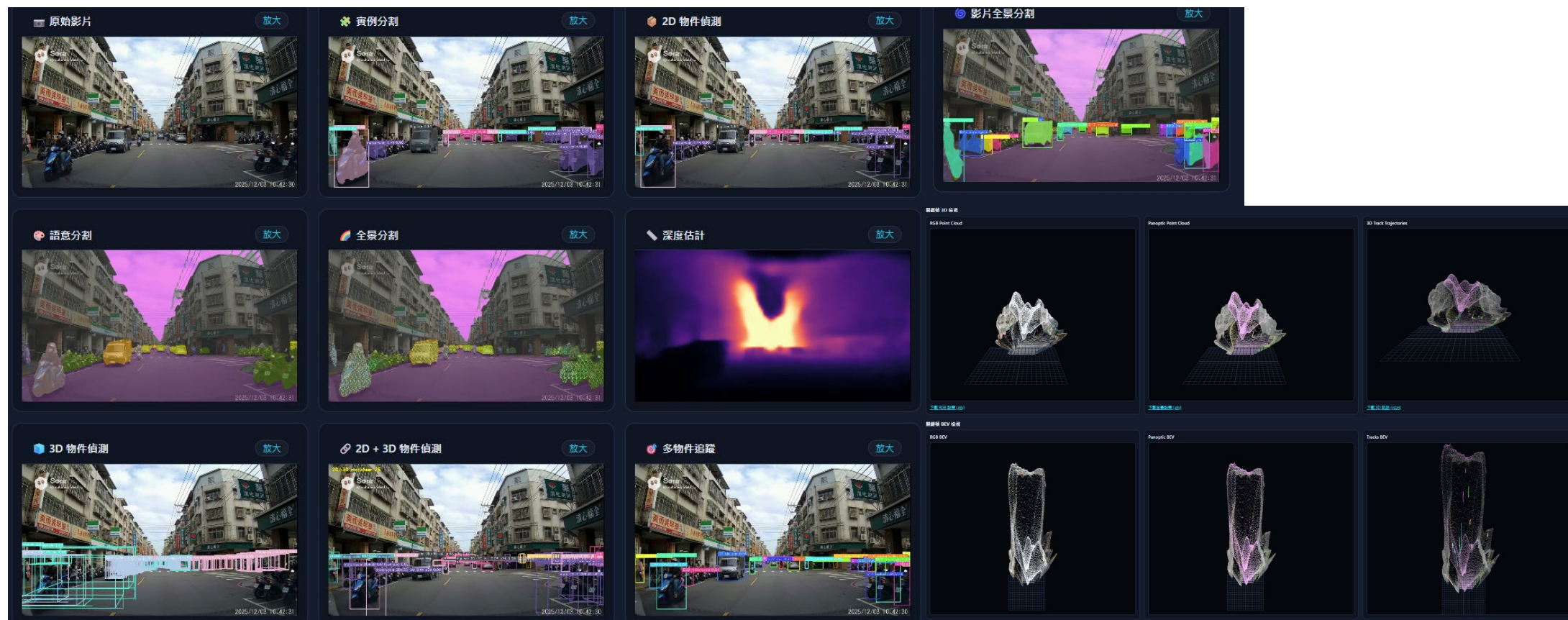


**Without
Visual Language
Understanding**

**With
Visual Language
Understanding**

GYOLO Downstream Model

Combine Finetuned GYOLO, MOT, and so on



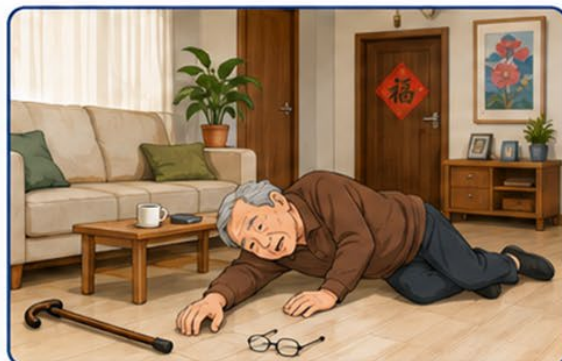
如何整合多模態以形成有效的語言模型
例子

一個應用情境

以下例子需要的基礎模型：用TaiwanGBA 執行行為分析、TTS（華文轉語音）、ASR（聽懂華文）、LLM（求救）



1 影片理解與事件判斷



臺灣居家老人跌倒影片



VLM
視覺語言模型



判斷結果：跌倒危險事件

2 文字轉語音主動詢問



系統主動詢問長者是否需要幫助



TTS
文字轉語音

3 語音回覆辨識需求



長者語音回覆



ASR
自動語音辨識



辨識需求：
需要協助、通知家人



ASR 準確辨識長者回覆

4 依需求通知家人



LLM
大型語言模型



SMS

傳簡訊給老人的家人



影片輸入



VLM
視覺語言模型



TTS
文字轉語音



ASR
自動語音辨識



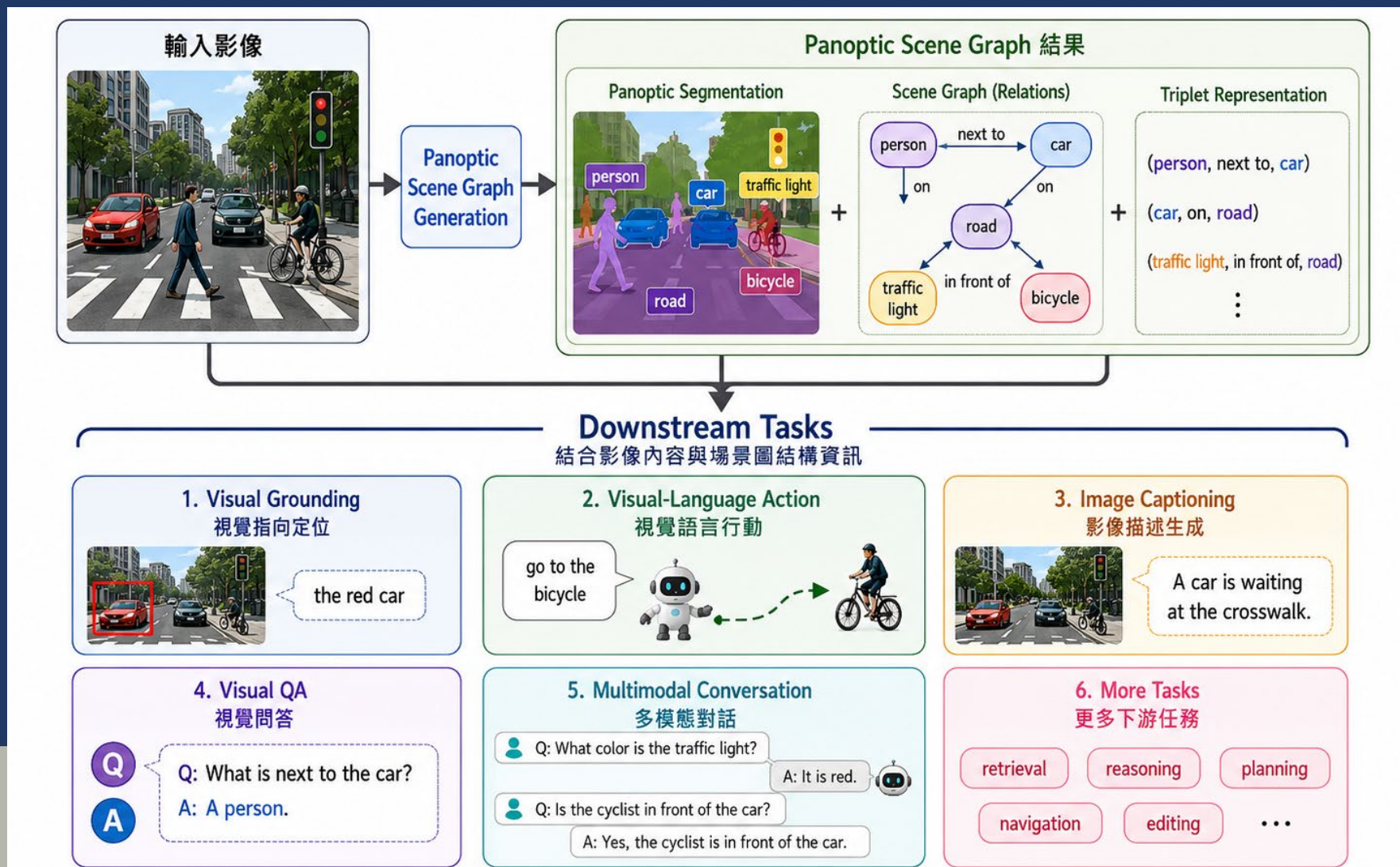
LLM
大型語言模型



通知家人

要進入具更複雜功能的 edge 端多模態語言模型系統，必須要
與影像或視訊能‘對話’
對話的基礎是必須更清楚、完整的描述影像/視訊內容

什麼是與影像對話？



未來:能與影像/視訊對話

- 多模態大型語言系統，能在 edge 端同時處理音訊、語言、影像、視訊，並融合起來一起達成任務。
- 困難:
 - Edge 端運算、記憶體、及各種條件非常嚴苛
 - 模型要夠小、但速度及準度必須維持
 - 多模態協作模型大小不一，應用也不盡相同，系統化的訓練方式有一定難度

Thanks for listening