© COPYRIGHT INNODISK CORPORATION

iQ-Foundry 課程講義

黃誌祥 | 智能周邊應用事業處 軟體研發部 副理 · 2026-07-30

innodisk

© COPYRIGHT INNODISK CORPORATION

AGENDA

- 實作里程碑
- iQ-Foundry 定位
- 實作流程說明
- 動手做

© COPYRIGHT INNODISK CORPORATION

iQ9 實作演練 · 下半場 · iQ-Foundry

上半場你跑的是官方模型。

現在，把你自己的 YOLO 26 跑上 NPU。

取得模型 → 優化轉換 → 部署 PoC → 驗證。 全程動手

Bring Your Own Model

© COPYRIGHT INNODISK CORPORATION

四個里程碑，全程動手

	里程碑	動手實作
1	取得模型	**高通模型下載** YOLO 26：Ultralytics 預訓練 / AI Hub / 自訓練 .pt。
2	優化轉換	**QC / API Export** 量化編譯成可在 NPU 跑的 INT8 模型。
3	部署 PoC	**test · 實機** 部署上機跑推論，送進 Streampipe。
4	驗證	**mAP** mAP@0.5 對比 FP vs INT8，自己驗精度。

© COPYRIGHT INNODISK CORPORATION

WHAT IS...
iQ-Foundry 定位

一條 BYO-Model 流水線。

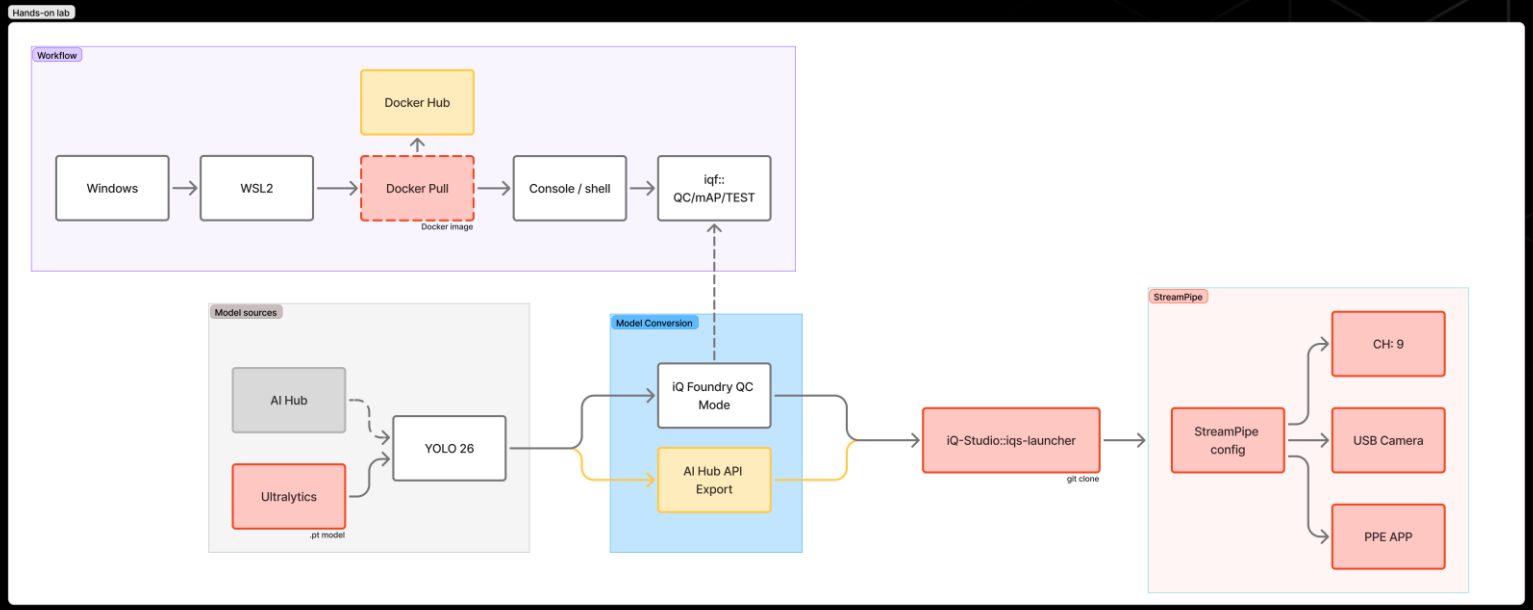
把你自己的模型，量化 → 驗證 → 部署自動化。模型自主權始終在你手上；Apache-2.0。

HOST	**x86** Ubuntu 22.04 / WSL2，建議 ≥16GB RAM。
TARGET	**On-device** EXEC-Q911（QCS9075），推論在 Hexagon NPU。
兩條轉換路徑	**兩者都教** iQ Foundry QC Mode / AI Hub API Export。

© COPYRIGHT INNODISK CORPORATION

WORKFLOW: 今天下半場這條路

從你的模型，到 NPU 上的 PoC。



主機環境（WSL2/Docker）→ 取得模型（Ultralytics / AI Hub → YOLO 26）→ Model Conversion（QC Mode / AI Hub API Export）→ StreamPipe（CH:9 / USB Camera / PPE APP）。

© COPYRIGHT INNODISK CORPORATION

Host · 主機環境 ● 動手

WSL2 + Docker，裝好 iQ-Foundry。

```
$ # PowerShell(管理員): cd ~ ; git clone …/iQ-Foundry.git ; cd iQ-Foundry
$ powershell -ExecutionPolicy Bypass -File .\setup-windows-wsl.ps1
$ # ↓ 以下在 WSL 終端機（Docker 指令不在 PowerShell）
$ bash ./docker_install.sh
$ docker pull innodiskorg/iqf:latest
$ ./qaihub_login.sh --key <YOUR_QAI_HUB_API_KEY>
```

每人需自備 Qualcomm AI Hub 帳號與 API token（qc 雲端編譯需要）。

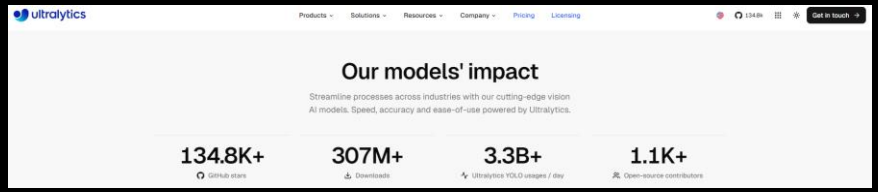
© COPYRIGHT INNODISK CORPORATION

Model · 取得模型（高通模型下載）

● 動手

YOLO 26，三種來源。

Ultralytics	下載官方預訓練權重 .pt。	預訓練
Qualcomm AI Hub	從 AI Hub 取得模型。	高通
自訓練	你自己訓練的 .pt 直接帶進來。	.pt



截取時間: 2026-07-28

© COPYRIGHT INNODISK CORPORATION

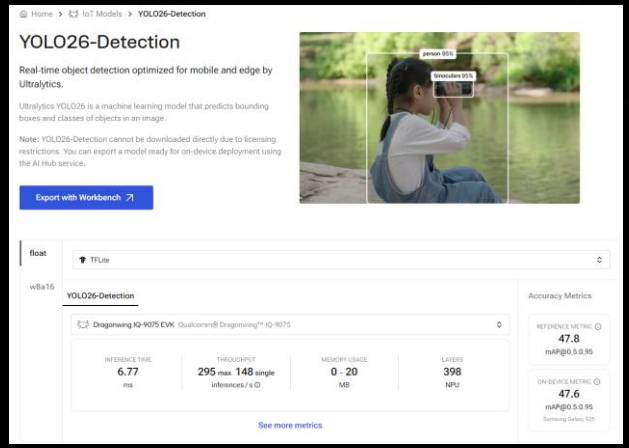
Convert · 路徑 A：AI Hub Export

● 動手

.tflite(FP32) → FP32 Running on NPU

```
$ python -m qai_hub_models.models.yolo26_det.export \
  --target-runtime tflite --precision float \
  --device "Dragonwing IQ-9075 EVK" --skip-profiling --skip-inferencing \
  --output-dir /ws/out_fp32
```

驗收 產出 /ws/out_fp32/yolov26.tflite（經 QAI Hub 雲端編譯）。



© COPYRIGHT INNODISK CORPORATION

Convert · 路徑 B：iQ Foundry QC Mode

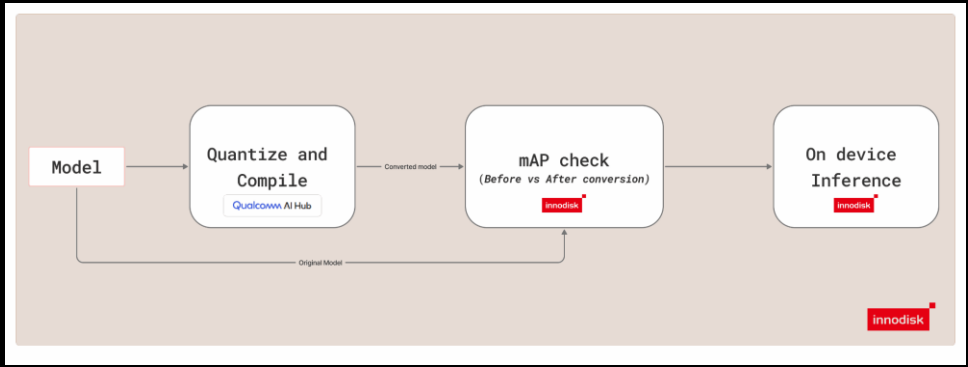
● 動手

.pt(FP32) → 全 INT8 .tflite。

```
$ ./docker/iqf run qc --type yolov26 --model yolov26n.pt --calib_dir calib/
```

選項：--max_calib(200) / --qc-quant-scheme mse|minmax / --qc-head one2many|one2one

驗收 產出 out/model/yolov26/yolov26_int8_<ts>.tflite（經 QAI Hub 雲端編譯）。



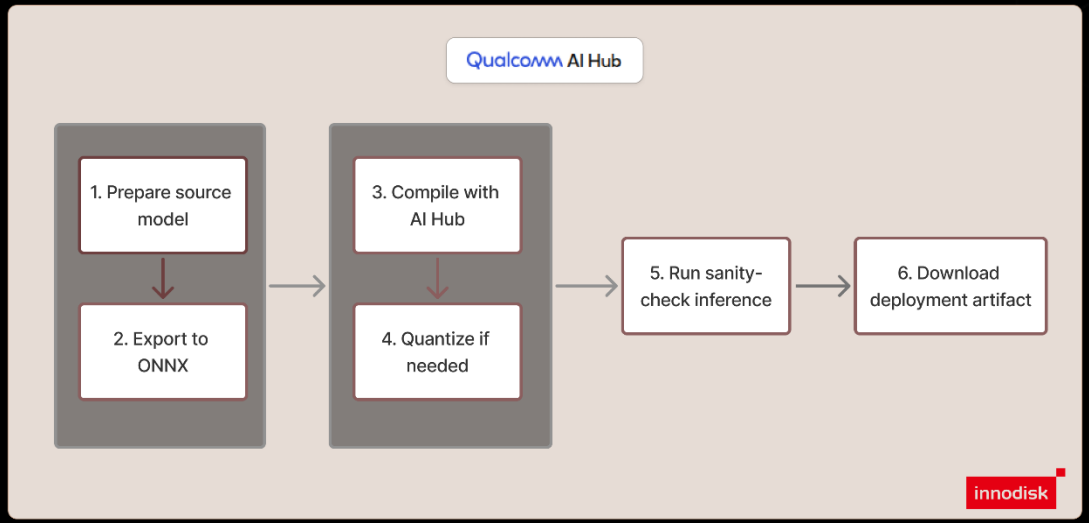
© COPYRIGHT INNODISK CORPORATION

Convert · 路徑 C：Non-YOLO AI Hub API Export

非支援模型，也能上 NPU。

不在 yolov10/11/26 內的視覺模型：自走 ONNX → Qualcomm AI Hub Workbench（compile / quantize / inference）。

非 YOLO 的第二種轉法，是 QC 接不到時的逃生路徑。



© COPYRIGHT INNODISK CORPORATION

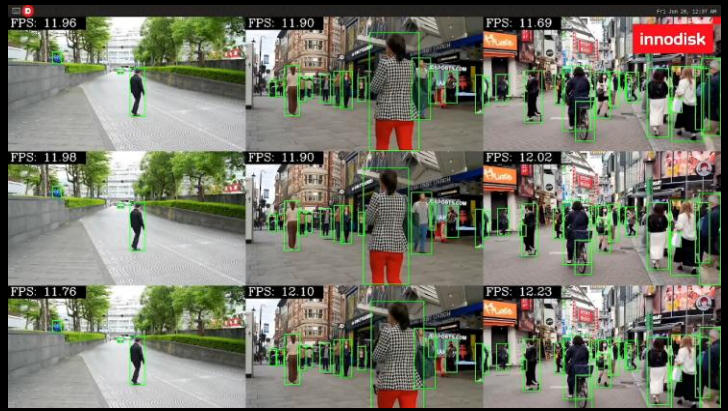
Deploy · 模型轉換與部署 AI PoC

● 動手

轉好的模型，不管是INT8或是FP32，上機跑。

```
$ iqs-launcher --autotag iqs-streampipe --other “-c yolo26_int8.config”    # 多路串流
```

轉好模型 → StreamPipe config → CH:9 · USB · PPE





驗收　模型在 EXMP-Q911 上跑出推論（TFLite + QNN/HTP → Hexagon NPU）。

© COPYRIGHT INNODISK CORPORATION

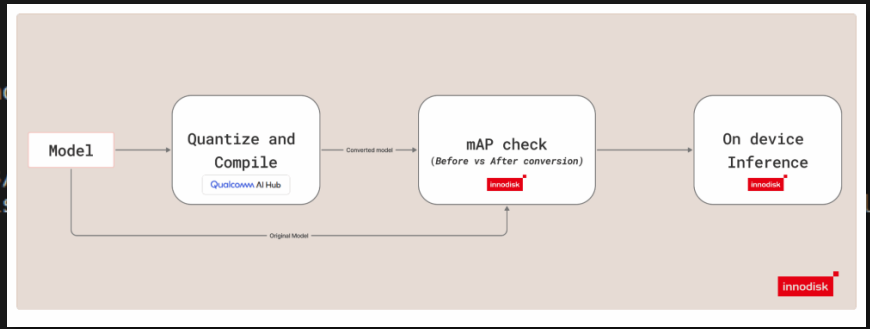
Validate · 驗證（mAP@0.5）

● 動手

量化掉多少精度？你自己驗。

```
$ ./docker/iqf run mAP --type yolov26 --fp-model yolov26n.pt \
  --int-model int8.tflite --annotations val.json --images val --max-images 5
```

```
adb shell 'rm -rf /data/local/tmp/yolo_map_eval/map_run_183661_1774235774598'
FP vs INT mAP@0.5 Pair Evaluation
model_type: yolov26
fp_head: one2many
annotations: /home/sanoop/Projects/yolo11_det/coco/annotations/instanc
images: /home/sanoop/Projects/yolo11_det/coco/val2017
num_images: 300
fp_model: /home/sanoop/Projects/qai_hub_conv_pipeline/mAP_eval_models/
int_model: /home/sanoop/Projects/qai_hub_conv_pipeline/mAP_eval_models
fp_map50: 0.523173
int_map50: 0.511969
abs_delta_int_minus_fp: -0.011204
pct_delta_vs_fp: -2.14%
trend: decrease
[ok] wrote: /home/sanoop/workspace/IQ-Studio/tutorials/model-pipeline/cv/yolo/out/mAP_results/yolov26/yolov26_mAP_result_20260323_111612.txt
```



© COPYRIGHT INNODISK CORPORATION

The Relay · 看到 → 擁有

取模 → 轉換 → 部署 → 驗證

一條可重複、可驗證的路。模型自主權，回到你手上。

iQ-Studio → **iQ-Foundry**

看到 · 系統能跑　　擁有 · 跑你自己的

Simplify the Workflow, Accelerate Deployment.

© COPYRIGHT INNODISK CORPORATION

innodisk

ARCHITECT INTELLIGENCE

Qualcomm AI Hub GenieX

© COPYRIGHT INNODISK CORPORATION

