



双锋猎鹰

——面向机载平台的智能目标检测系统

CONTENTS

汇报目录

01

作品背景

低空经济场景 · 边缘检测四大难题 · 总体方案与目标

02

系统设计

切割点选择 · 细腰与 CAST 传输 · 弹性接管 · Huntrix 与 LangGraph 智能体

03

运行效果

检测精度 · 平台性能对照 · 双机协同与任务闭环实测

04

创新总结与应用

六大创新点 · 应用场景 · 意义与价值

低空经济进入规模化落地期，机载实时感知成为刚需

政策牵引：低空经济上升为国家战略

- 低空经济先后写入 2024、2025 年政府工作报告，纳入国家战略性新兴产业布局
- 无人机由单纯飞行平台，向集低空组网、边缘计算、智能感知于一体的综合物联网载体转型
- 低空巡检、目标搜索与区域安全等场景，要求无人机具备机

载自主感知能力

传统模式之痛：“空中采集 — 数据回传 — 地面处理”难以为继

空中采集图像 / 视频

➡

链路回传地面平台

➡

地面集中检测识别

低空作业网络环境恶劣，链路带宽有限、质量波动明显：持续回传大量图像视频既占用通信资源，又增加处理时延，难以满足实时任务需求。

把检测能力部署到无人机边缘平台、实现机载自主感知，是低空智能化发展的必然方向 —— 这也正是“双锋猎鹰”的出发点。

场景落地：目标检测需求多样且时效严苛

-  **电力巡检** 自动发现绝缘子缺陷、导线异物等设备异常
-  **应急救援** 快速发现受困人员、车辆与临时避险点
-  **森林防火** 及时识别烟火、非法采伐与异常活动目标
-  **交通监测** 实时检测拥堵、事故车辆与违规行为
-  **边境巡护** 对人员、车辆、装备持续搜索识别与告警

面临的难题：低算力边缘平台目标检测的四重约束

01

算力受限：大模型难以在边缘实时运行

GPU 架构、内存带宽与多任务并行共同限制推理帧率。

实测：Jetson Nano CPU 推理 YOLO26n 仅约 1.3 FPS

02

通信受限：中间特征形成新的传输瓶颈

切分后中间特征通道数高，数据量远超压缩图像。

实测：P3/P4/P5 FP16 约 8.2 MB，是 JPEG 的 80.8 倍

03

结构受限：压缩不能破坏网络输入结构

Neck 依赖 FPN/PAN 多尺度融合，丢弃尺度即输入偏移。

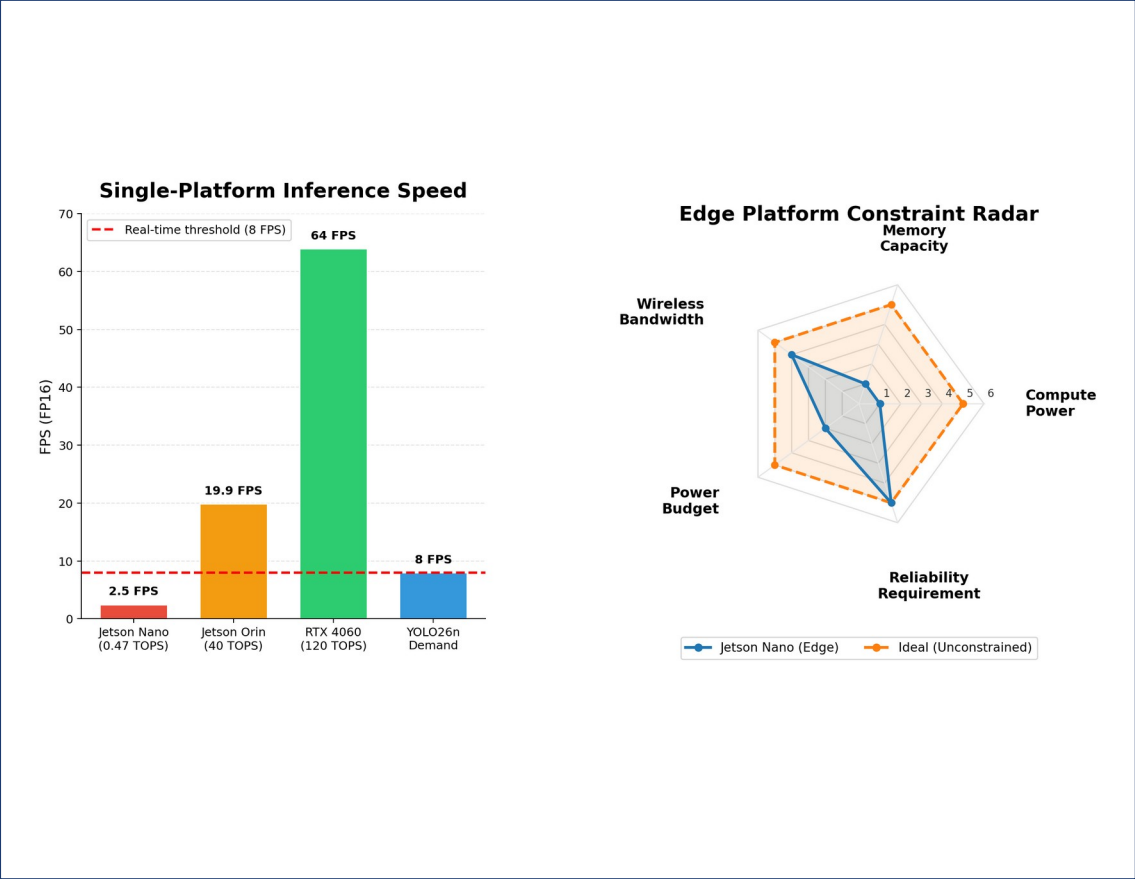
实测：仅传 P5 时置信度崩塌，甚至零检测

04

链路不稳：任务要求“退而不崩”

距离、遮挡、姿态、干扰均可致链路波动甚至中断。

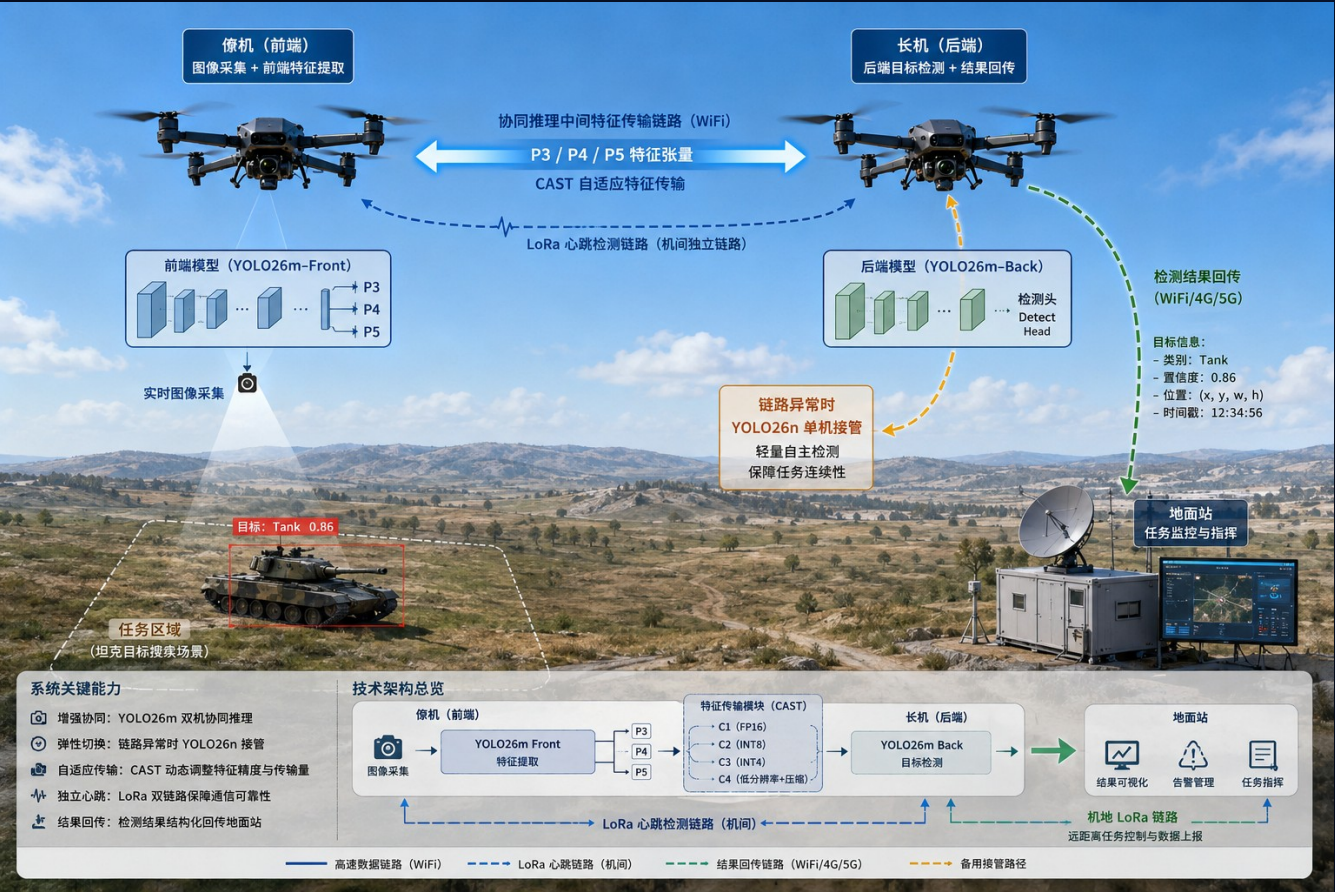
需求：短时精度下降可接受，完全无输出不可接受



图：低算力边缘平台的算力、带宽与连续性约束实测

四重约束相互关联、相互制约，需要“计算 × 通信 × 结构 × 可靠性”协同的系统设计。

总体方案：“增强协同 × 轻量自主”的弹性目标检测体系



图：双锋猎鹰系统示意图

核心思想：弹性推理

- 通信良好时：双机协同推理，发挥较大模型检测能力
- 通信受限时：细腰 + CAST 降载，维持协同运行
- 通信异常时：轻量模型单机兜底，任务不中断

作品目标与关键功能实测

- 高精度识别 自训练细腰 YOLO26m
mAP@0.5 = 0.9243，坦克样本 conf = 0.8026
- 协同推理 YOLO26m 层 10 前后端切分
前端 8.15M / 后端 11.16M 参数，负载均衡
- 异常接管 LoRa 心跳 + YOLO26n 兜底
1 Hz 心跳 / 连续 3 次超时判定策略
- 任务闭环 LLM 智能体 + 六阶段状态机
多层 fail-closed 安全闸门，328 项离线自动化测试通过

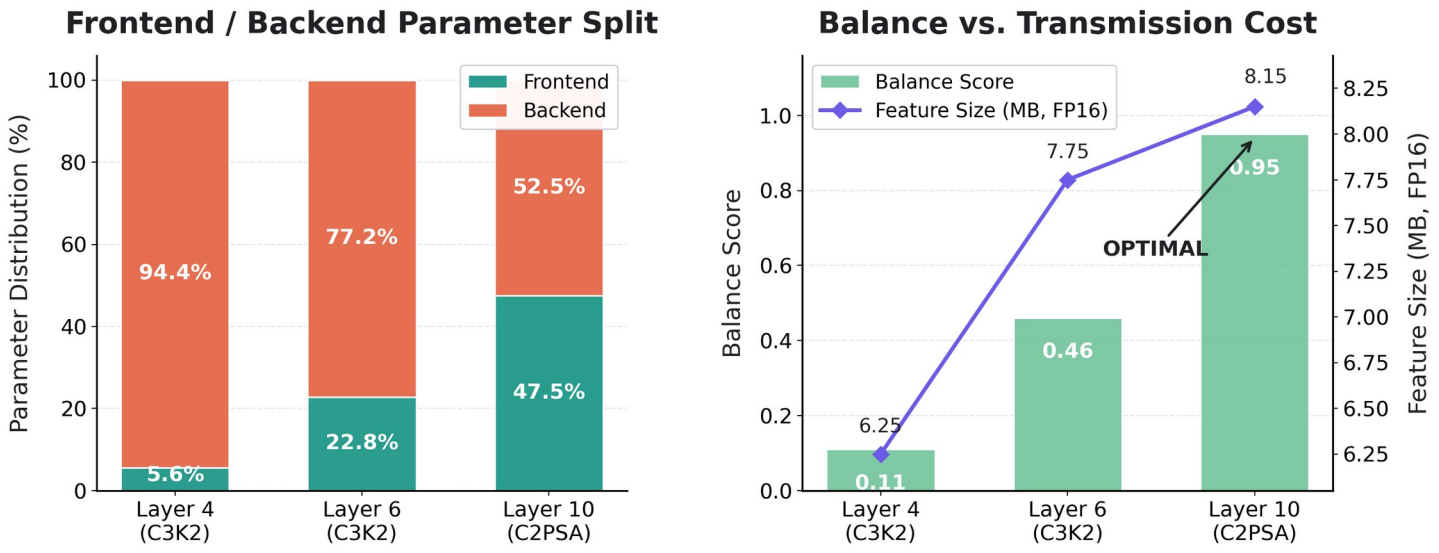
系统总览：三大感知层模块 + 一个任务层智能体模块



形成完整闭环： 能切分 → 能传输 → 能容错 → 能闭环

切割点选择提供协同边界；细腰结构与 CAST 解决传输瓶颈；弹性接管保障链路异常时检测不断流；Huntrix 把检测结果转化为安全、可审计的任务动作。

切割点选择算法：模型在哪里切，让结构数据说话



图：结构感知切割点选择的决策流程

切割点	传输内容	前端参数	后端参数	均衡度
层 4	P3	5.6%	94.4%	0.11
层 6	P3 + P4	22.8%	77.2%	0.46
层 10	P3+P4+P5	47.5%	52.5%	0.95

算法机制：结构优先的多目标评分

硬约束：后端跨层输入完整，P3/P4/P5 三尺度不缺失

软指标：计算均衡 35% · 结构 25% · 通信 20% · 流水线 10% · 复杂度 10%

规模：12 个 YOLO 家族模型，276 个候选点，输出 12 个可切割点位

YOLO26m 评估结果：推荐层 10 后切分

77.58

综合评分（候选点最高）

52.5/47.5

前后端 GMAC 占比 %

100

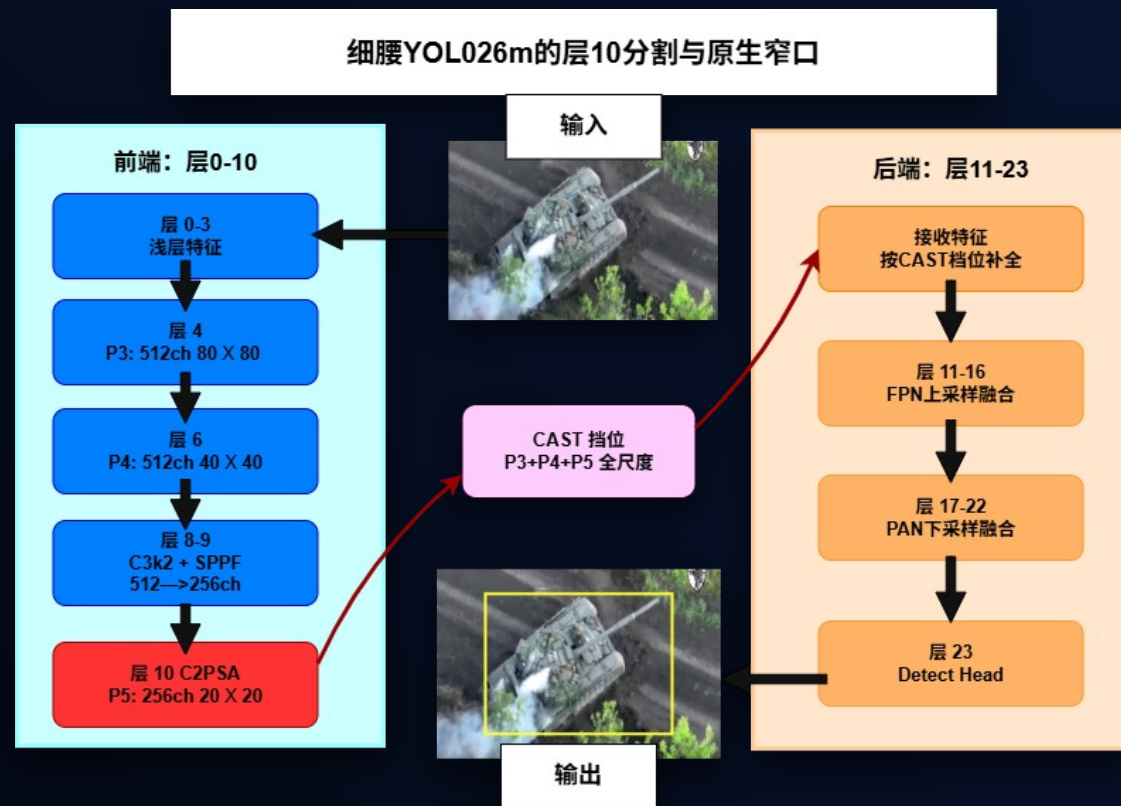
结构评分（满分）

8.20 MB

边界特征量（FP16）

层 10 恰位于 Backbone-Neck 自然边界：三尺度特征完整形成、前后端负载接近 —— 是协同推理的结构起点，也为细腰与 CAST 留下优化空间。

细腰化模型结构：让模型在切割边界天然适合传输



图：细腰 YOLO26m 的层 10 分割与原生窄口

设计思路：架构原生窄口，而非外加压缩模块

- BCM 外加压缩路线探索失败后，转向架构原生窄口：层 8/9/10 通道 512 → 256 缩窄
- 从训练阶段就让模型适配协同传输，不改变后端 Neck 拓扑，三尺度接口保持不变
- 无需额外 Encoder/Decoder，结构简洁、稳定、可解释

≈200 KB

P5 FP16 特征量（原生窄口）

0.9243

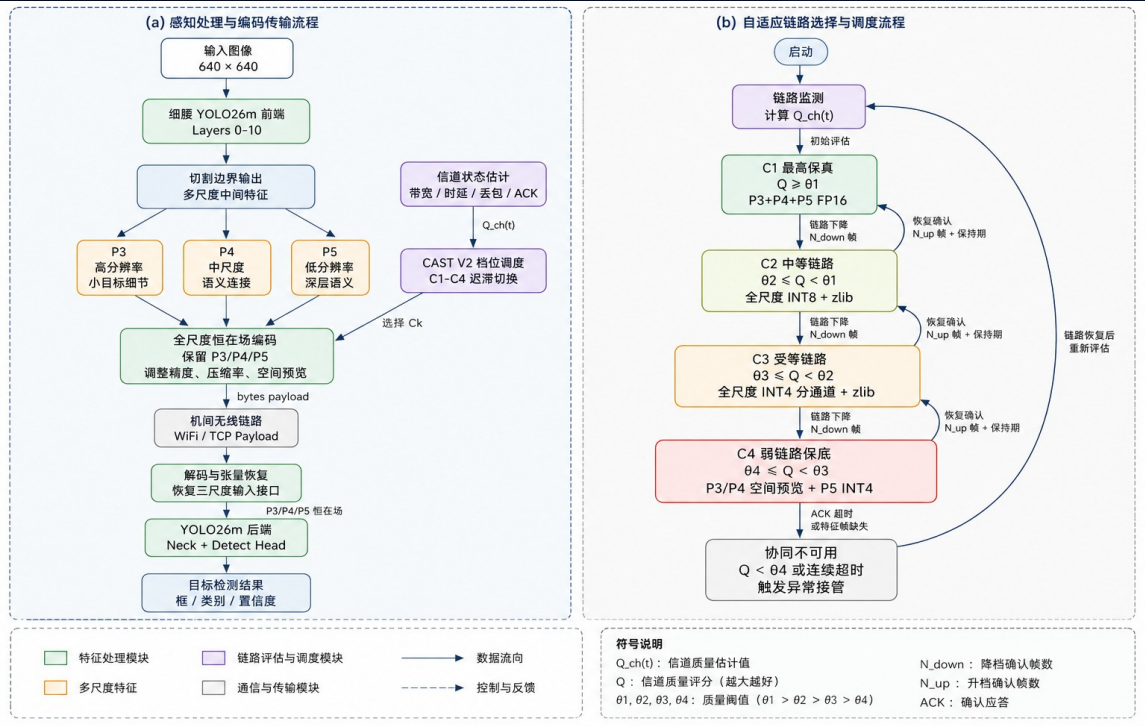
细腰模型
mAP@0.5（自训练）

0.5732

mAP@0.5:0.95

双模型体系：细腰 YOLO26m 负责通信良好时的**增强协同**；轻量 YOLO26n 负责通信异常时的**自主兜底**。FP16 在该实验边界下无 mAP@0.5 损失（0.9118 = 0.9118）。

CAST V2 自适应传输：全尺度恒在场，随链路平滑降级



图：CAST V2 全尺度保留与多挡位压缩策略

从 V1 失败到 V2 的结构修正

V1“链路越差、传得越少”：仅传 P5 虽大幅降低传输量，却破坏 Neck 多尺度输入假设，导致置信度崩塌甚至零检测。

V2 转向“保结构、调质量”：任何挡位都传输 P3/P4/P5 三尺度，只在精度、压缩率与空间细节上分级降级，并以迟滞机制防跳挡。

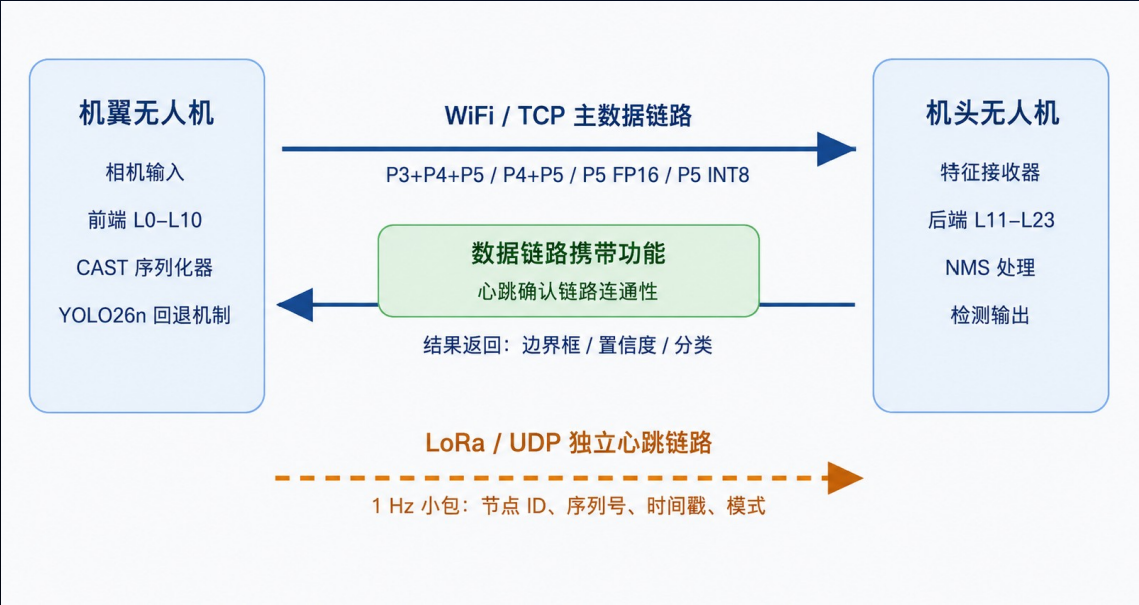
C1-C4 四挡编码策略

- C1** 全尺度 FP16 · 最高保真，适用高带宽近距离链路
- C2** 全尺度 INT8 对称量化 + zlib，适用中等链路
- C3** 全尺度 INT4 分通道量化 + zlib，适用受限链路
- C4** P3/P4 空间预览 + P5 INT4，弱链路保底可检出

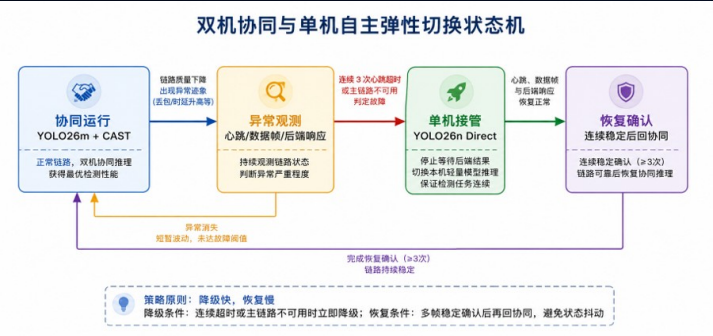
坦克图复测（同图同权重）	C1	C2	C3	C4
Payload / 置信度 / 匹配	8200.05 KB / 0.8026 / 1:1	2108.72 KB / 0.8028 / 1:1	802.21 KB / 0.7886 / 1:1	280.74 KB / 0.7332 / 1:1

数据量降至 C1 的约 1/29，四挡均保持目标检出 —— 传输量大幅下降而检测能力平滑退化（完整模型基线 conf=0.8026）。

异常检测与弹性接管：链路断了，检测不断



图：WiFi 主链路与 LoRa 独立心跳分离式通信架构



图：异常检测与弹性接管状态机

LoRa 独立心跳：职责分离的存活判定

- 主链路传中间特征，LoRa 只传轻量状态包，避免多链路同时失效
- 1 Hz 心跳、连续 3 次超时判定故障；心跳 + 数据帧双信号刷新存活时间
- 小包、低频、独立、可解释，适配 LoRa 信道负载

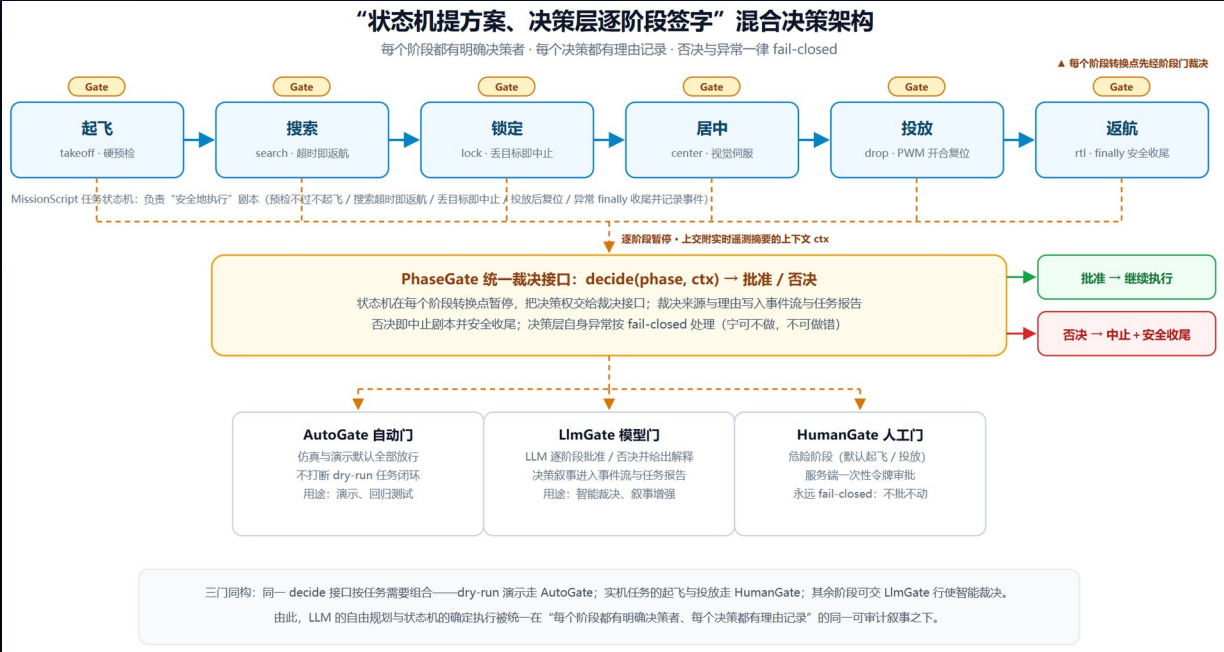
弹性接管状态机：降级快、恢复慢

- 协同运行 → 异常观测（CAST 降挡优先）→ 单机接管（YOLO26n 本机推理）→ 恢复确认（多帧稳定后再回协同）
- 故障发现时间与模型切换开销分开设计，避免阈值附近频繁抖动

YOLO26n 轻量兜底模型（知识蒸馏）

- 细腰 YOLO26m 冻结作教师，30 轮直训 + 20 轮蒸馏微调（T=4.0）
- mAP@0.5 = 0.9759，mAP@0.5:0.95 = 0.6785
- 与协同链路同输入输出接口，接管后上层告警、跟踪无感

Huntrix 智能体：自然语言驱动的安全受控任务闭环



图：“状态机提方案、决策层逐阶段签字”混合决策架构

328 项

离线自动化测试全部通过
安全闸门经真实链路实证：投放未批准则舵机指令不下发；电量 0% 预检拒绝
起飞并返回可读原因。

120 s

审批超时自动过期、指令不下发

权责边界：模型规划、系统执行

大语言模型不直接接触无人机链路，动作空间约束在 16 个语义化受控工具内；基于 LangGraph 构建工具调用闭环，决策轨迹可导出审计，支持云端 / 本地模型切换。

混合决策：六阶段剧本 + 阶段门裁决

任务固化为起飞—搜索—锁定—居中—投放—返航六阶段剧本；每个阶段转换点由自动门 / 模型门 / 人工门裁决，否决即安全收尾。

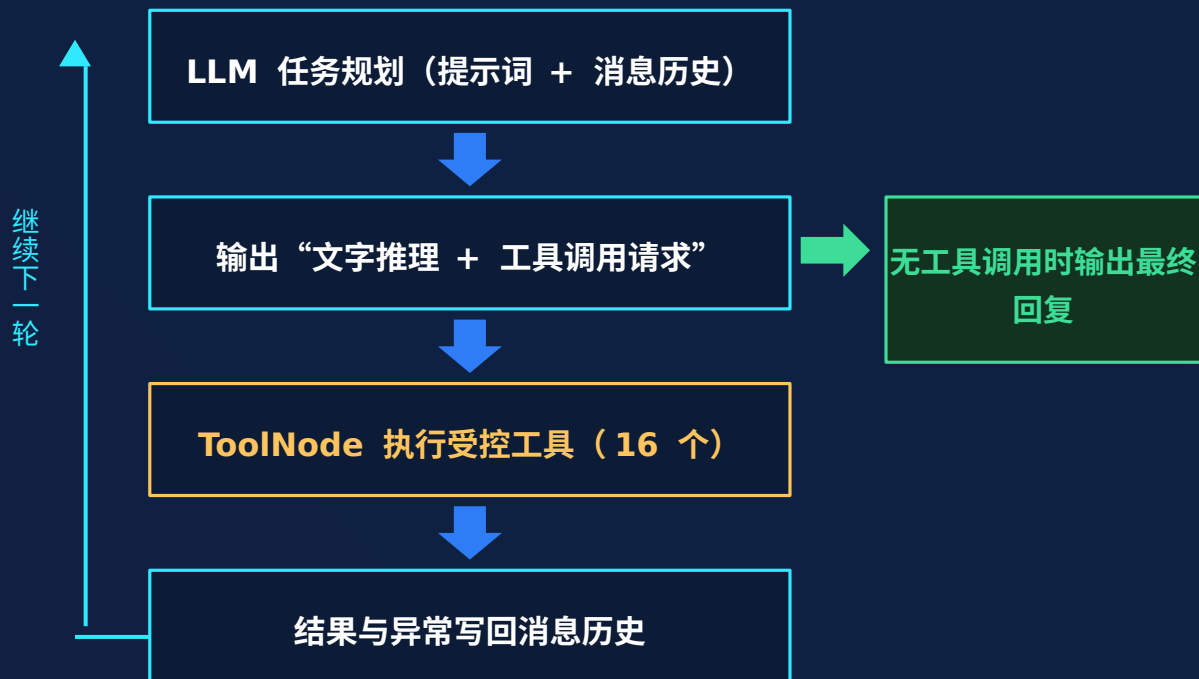
五层失效关闭（fail-closed）安全强制体系

- 配置层 live 双开关 + 投放总开关，默认关闭
- 校验层 指令本地范围校验，越界即拦截
- 预检层 起飞前硬预检，不达标直接拒绝
- 策略层 返航电量、地理围栏、摇杆冷却
- 在环层 人工审批一次性令牌 + 遥测看门狗

一切不确定情形都向“不做”一侧收敛；投放等不可逆动作必须人在回路。

LangGraph 智能体架构：可控、可审计的工具调用闭环

工具调用闭环：每一轮都在掌控之中



Checkpoint 持久化：同一任务多轮指令共享记忆；完整决策轨迹（模型推理、工具调用与返回）可导出 JSON，供逐轮复盘审计。

为什么是 LangGraph：明确的工程权衡

重型框架 X 工具循环封装在黑盒内，状态持久化、人工中断审批、过程可视化都要反向修改框架

手写循环 X 消息历史、工具路由、状态持久化全部自行维护

LangGraph ✓ StateGraph + ToolNode + 条件路由 + Checkpointer，显式原语恰好匹配“可控、可审计”，不引入多余抽象

双模型接入：断网仍保有规划能力

- 走 OpenAI 兼容协议，云端模型与地面站本地模型仅改环境变量即可切换
- 本地模型定位断网兜底；模型端点故障自动降级本地关键词模式并如实标注

受控工具集（16 个）：能力即边界

状态读取 · 目标查询 · 起飞 / 降落 / 返航 / 悬停 · 模式切换 · 八方向摇杆 · 云台指向 / 变焦 · 拍照 / 录像 · 投放 / 复位 · 推流控制

未暴露的能力（如直接发布原始报文）对模型而言不存在。

效果（一）：检测精度与边缘平台性能验证



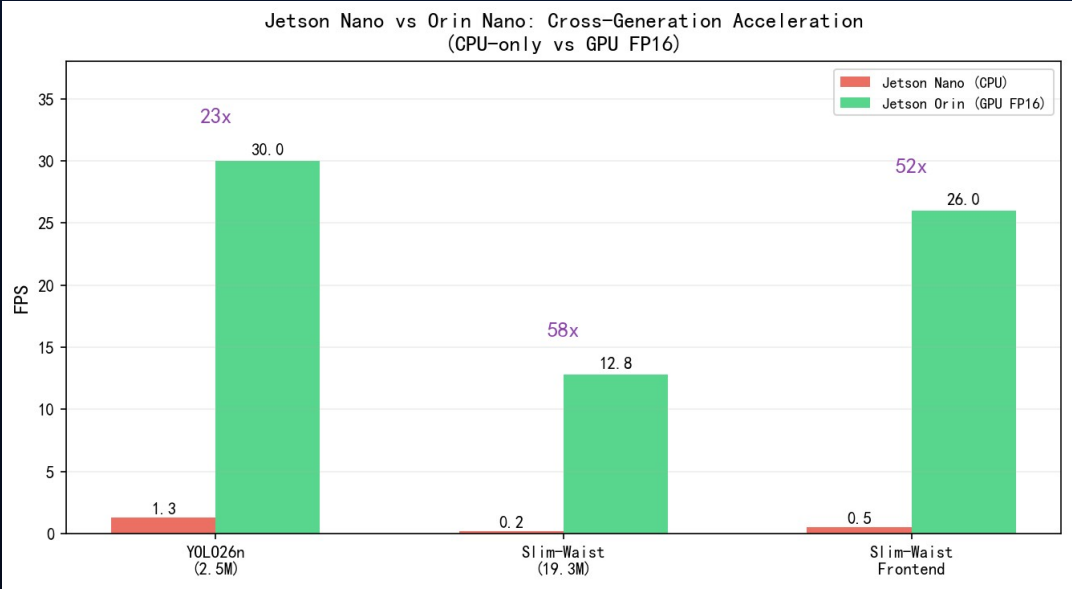
图：项目坦克样本及自训练细腰模型检测结果（conf = 0.8026）

0.9243

细腰 YOLO26m mAP@0.5

0.9759

YOLO26n mAP@0.5（兜底）

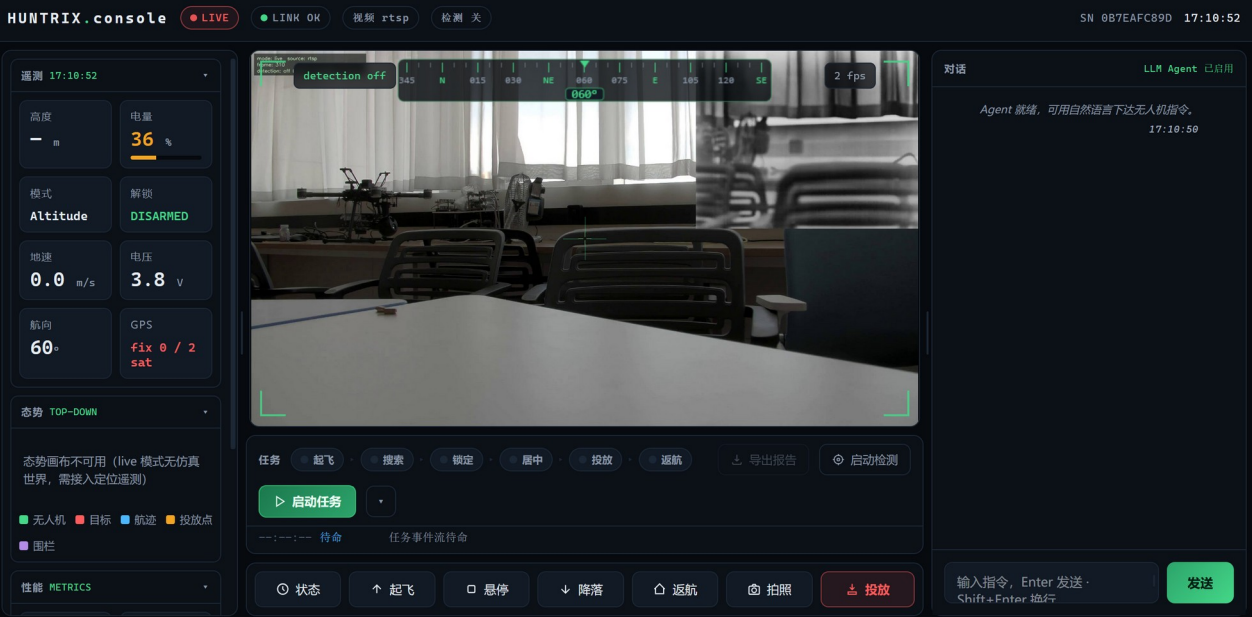


图：Nano 与 Orin 平台性能差异

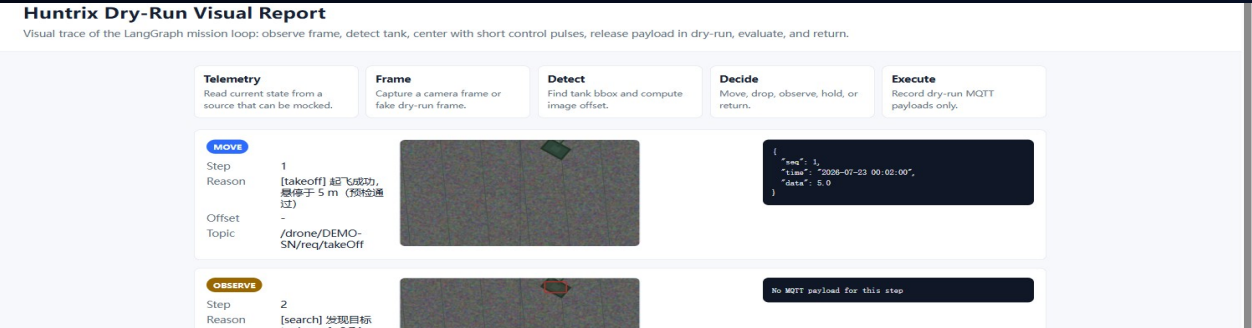
模型	Nano CPU-only	Orin GPU 实测	结论
YOLO26n	约 1.3 FPS	约 19.9–30 FPS	Orin 可实时
细腰模型	约 0.2 FPS	约 12.8–17.8 FPS	Nano 难以承载
分割链路	—	本地 L1 约 10 FPS	序列化待优化

硬件对照确立部署边界：Nano 证明极低算力瓶颈真实存在，Orin 证明同一算法在合适平台上具备实时基础。

效果（二）：真实链路任务闭环实测



图：Huntrix 三轨 HUD 地面站控制台（live 链路实测画面）



图：六阶段任务闭环 HTML 报告（含各阶段真实画面快照）

A460 仿真机链路（4G 公网）实测指标

指标	实测值
YOLO26n 检测耗时	平均 16–19 ms , P95 23–31 ms
机载检测服务	约 19.9 FPS , 结果轮询约 10 Hz
视频端到端帧率	约 9 FPS （拉流 + 检测 + HUD ）
MQTT 回执延迟	百毫秒级（单次约 225 ms ）
LLM 单轮对话时延	约 4 s （云端模型，含推理）
模型降级行为	端点故障降级本地命令模式

闭环与安全实证

- 六阶段剧本仿真全部完成，视觉伺服进入 ± 32 px 死区收敛
- 投放动作 120 s 未批准自动过期，舵机指令未下发
- 起飞、移动、云台、拍照等在仿真机链路全链路按预期执行

面向真实无人机链路的双机协同原型与应用价值



图：外场双机实飞



Prometheus 科研无人机平台



双光球形光电吊舱

真实链路系统集成

- Jetson Orin Nano 机载计算
- WiFi 主链路 + LoRa 心跳双链路
- MQTT 遥测控制 / RTSP 视频 / ZMQ 结果
- RTK 高精定位、光电吊舱、投弹吊舱
- 模型、协议、硬件与异常机制联合验证

应用价值：方法不依赖单一目标

以坦克检测完成验证，核心方法可迁移：更换训练数据与模型，即可覆盖人员、车辆、火点、设备缺陷等目标。

服务于低空巡检、应急救援、区域安防、边境巡护、交通监测等远距离、弱通信、有限算力场景。

自然语言任务入口降低操作门槛，推动低空作业从“专业飞手操控”走向任务级指挥。

创新与特色：六项可复核、可解释、可迁移的贡献

01 结构感知的模型切割点选择方法

把切割点选择从经验判断推进为结构感知分析：网络拓扑为首要约束，12 个模型 276 个候选点验证，方案可复核、可解释、可迁移。

02 面向协同推理的细腰化模型结构

训练阶段即兼顾传输需求，层 8/9/10 通道缩窄形成原生窄口；与轻量 YOLO26n 构成“增强协同 + 自主兜底”双模型体系。

03 链路质量感知的 CAST 自适应传输

“保结构、调质量”：C1-C4 全尺度恒在场，链路波动转化为特征质量平滑调节，实现通信与 AI 推理的双向协同。

04 LoRa 心跳驱动异常检测与弹性接管

高速数据链路和低速状态链路冗余分离；小幅波动 CAST 降挡、断链轻量模型接管、恢复多帧确认，把运行稳定性提升到与检测精度同等高度。

05 面向真实无人机链路的双机协同原型

整合 MQTT 遥测、RTSP 视频、ZMQ 结果、Jetson 部署、WiFi 数据链与 LoRa 心跳，在完整工程链路中联合验证，而非停留在离线训练。

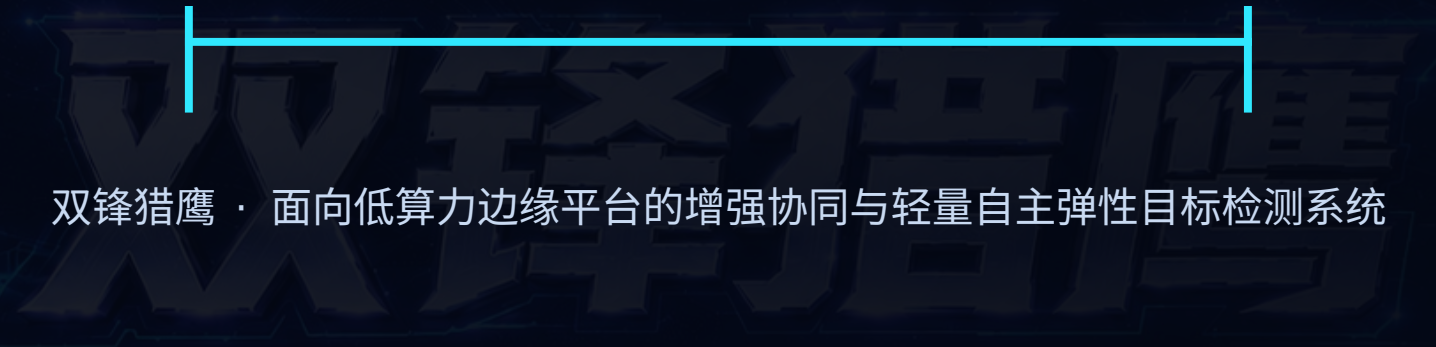
06 自然语言驱动的安全受控智能体闭环

大模型被纳入严格工程权责边界：受控工具动作空间、代码级 fail-closed 闸门、逐阶段签字裁决，为“AI 控机”提供可证明的安全范式。



请各位专家老师批评指正

THANK YOU FOR YOUR GUIDANCE



双锋猎鹰 · 面向低算力边缘平台的增强协同与轻量自主弹性目标检测系统

——面向机载平台的智能目标检测系统