电子科技大学（深圳）高等研究院
Shenzhen Institute for Advanced Study, UESTC

数据智能研究中心
Data Intelligence Group

边缘视觉智能研究进展

组会报告

汇报人：许新操

电子科技大学（深圳）高等研究院

2026年2月8日

汇报提纲

1 **个人简介**

2 研究背景

3 研究内容

4 后续计划

个人基本信息

基本信息

- 许新操（出生年月：1994.11）
- 重庆大学 工学博士（2023.06）
- 电子科大深研院 博士后 (2023.07)
- 电子科大深研院 副研究员 (2025.09)
- 研究方向：边缘视觉智能

主持省/部级项目

- ✓ 2025年广东省自然科学基金面上项目入选者
- ✓ 2023年中国博士后科学基金面上资助入选者

代表性学术成果

- **IEEE T-ITS 2024**：车路协同下异构信息实时融合与有效性评估，提升协同感知效率
- **IEEE TCE 2024**：数字孪生构建中的感知-上传-资源分配联合优化，提升数字孪生质量
- **ECAI 2025**：基于小波解耦与对比学习的跨模态交互表征，提升感知数据特征鲁棒性与泛化
- **JSA 2023**：NOMA环境下任务高效卸载与计算资源协同，提升复杂通信服务成功率
- **电子学报 2021**：边缘计算中的同信道干扰抑制与资源调度，提升任务完成率与信道利用率

汇报提纲

1 个人简介

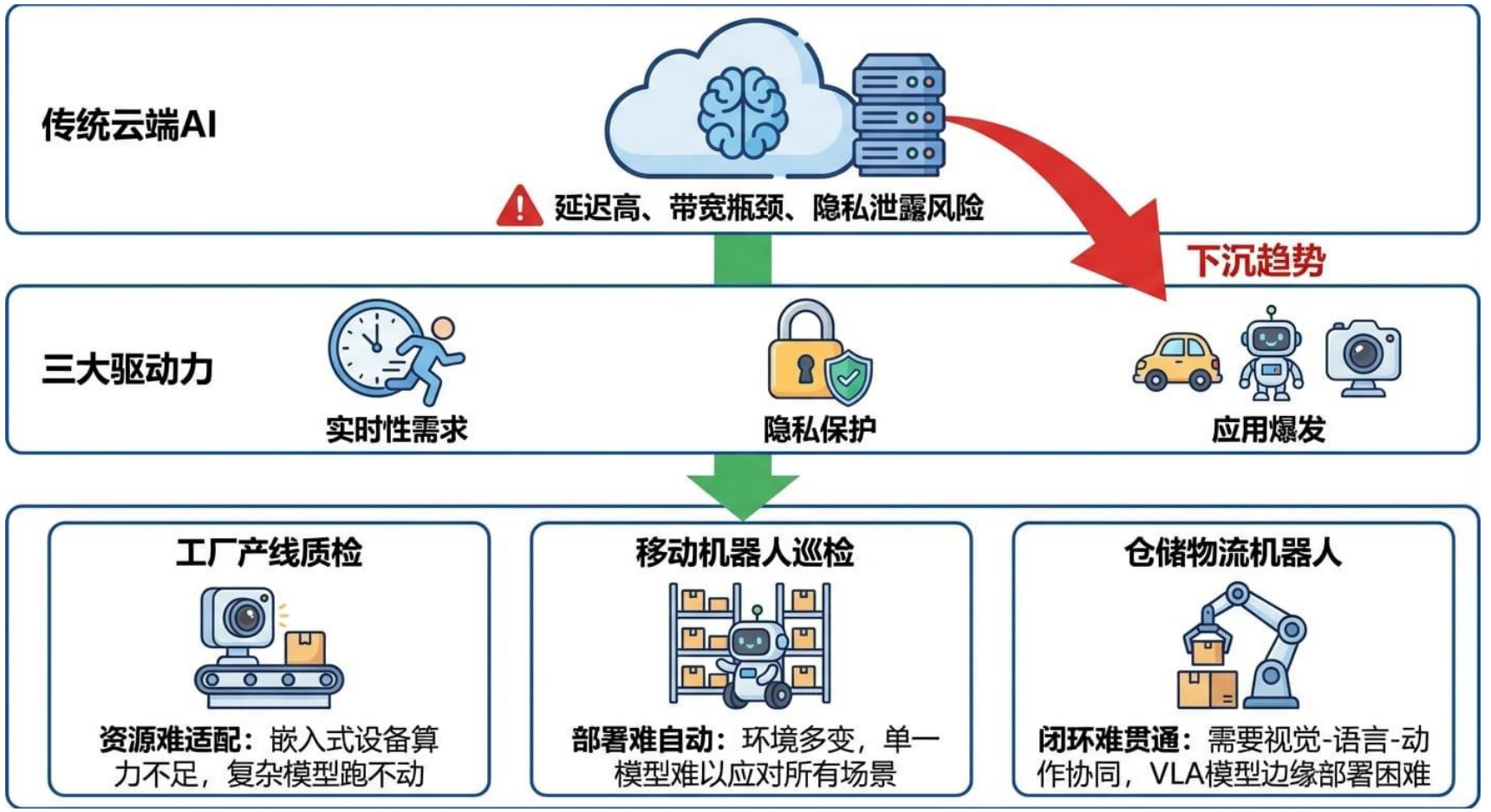
2 **研究背景**

3 研究内容

4 后续计划

研究背景：边缘视觉智能趋势与挑战

- 趋势一：视觉AI从云端下沉到边缘（实时性、隐私保护需求）
- 趋势二：应用场景爆发（自动驾驶、工业机器人、智能安防）
- 趋势三：设备形态多样（从服务器到MCU）
- 面临挑战
  - 资源难适配
  - 部署难自动
  - 闭环难贯通



边缘视觉智能算力适配、模型部署、系统闭环三重挑战亟待突破

研究背景：边缘视觉智能

挑战	资源难适配	部署难自动	闭环难贯通
关键问题	科学问题 异构资源约束下模型-设备协同优化机理	科学问题 LLM驱动的架构搜索与反馈演化机制	科学问题 边缘VLA模型压缩与快速适应耦合优化机理
研究内容	协同推理 异构边缘设备的重叠部署与动态执行调度机制	模型定制 LLM驱动端侧模型自动化生成与硬件感知优化	系统实现 资源受限条件下VLA的压缩-适应协同优化
目标	在资源受限、动态异构的边缘环境下，实现从算法→部署→应用的闭环		

研究背景： 研究定位


视觉推理
工厂产线质检
自动驾驶
移动机器人巡检
智能安防监控
仓储物流分拣
异构边缘协同推理
资源难适配
跨设备调度
模型分区
重叠部署
LLM驱动的端侧模型定制
LLM
部署难自动
自然语言解析
NAS搜索
多策略压缩
边缘高效VLA
闭环难贯通
视觉-语言-动作协同
具身智能

汇报提纲

1 个人简介

2 研究背景

3 **研究内容**

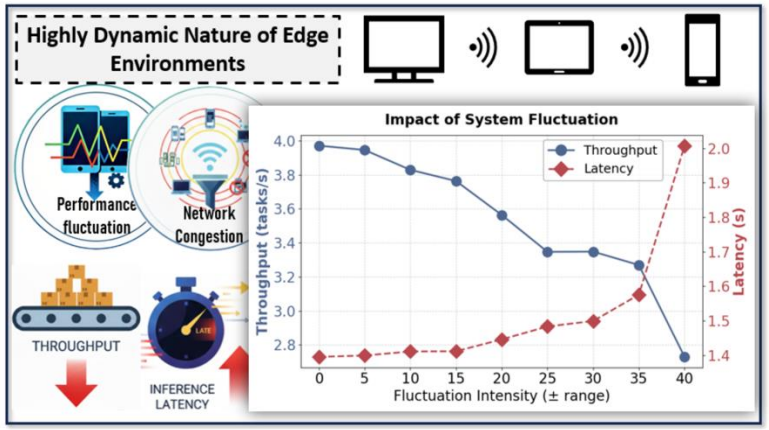
4 后续计划

研究内容一：异构边缘协同推理

背景

边缘流水线推理：资源波动/设备故障，重部署方案开销大，只优化吞吐量

研究内容



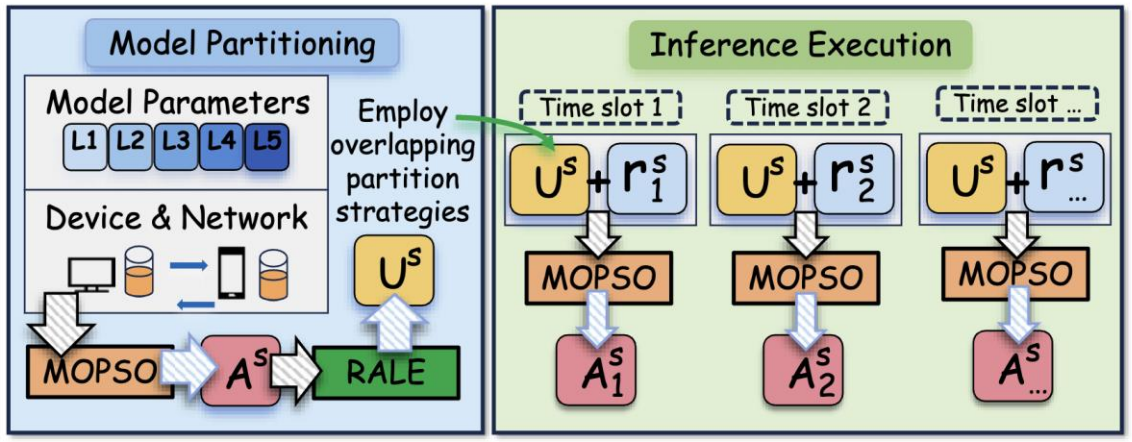
- **系统建模**：提出边缘流水线的重叠部署与动态执行系统模型，建立可靠性-吞吐量联合优化模型
- **调度框架**：构建部署-执行解耦的两阶段优化流程，支持无需重部署的实时自适应推理调度
- **算法设计**：提出 MOPSO 多目标分区算法与 RALE 冗余扩展算法，成功率最高↑8.7 倍，吞吐量↑1.4 倍

相关成果

- 已投稿 **1 篇论文**：投至 IEEE/ACM TON (CCF A 类国际期刊)
- 下一步工作：研究Transformer架构下Prefill与Decode阶段的异构多设备并行调度优化

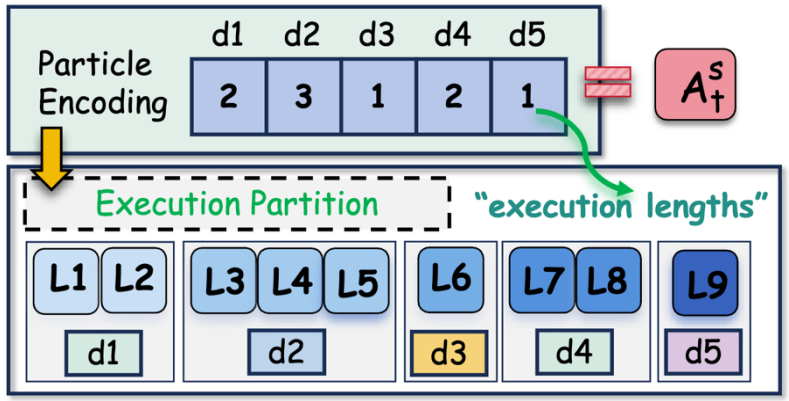
重叠部署 + 动态执行，通过冗余存储实现故障容忍，无需频繁重部署

研究内容一： 两阶段优化框架



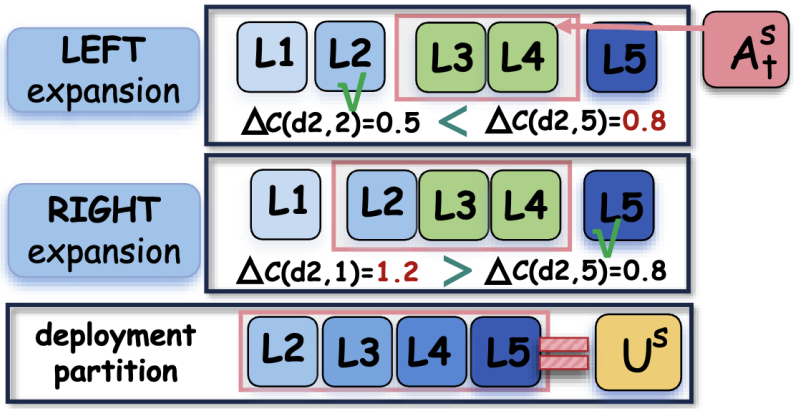
阶段	算法	作用
部署阶段	MOPSO + RALE	确定重叠部署方案
执行阶段	MOPSO	动态选择执行分区

MOPSO：多目标粒子群优化



- 粒子编码：每维表示设备执行的层数
- 双目标：最大化吞吐量 + 可靠性

RALE：冗余感知层扩展

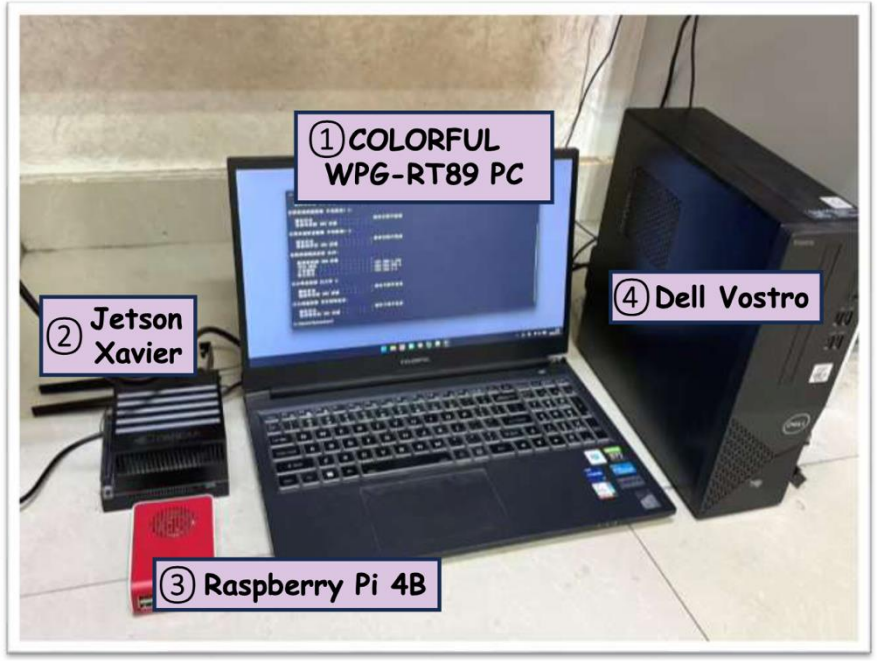


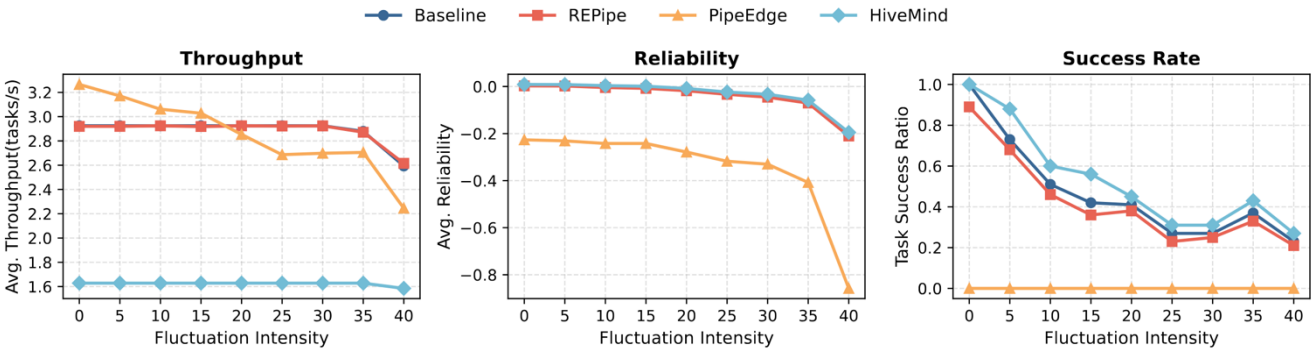
- 从执行分区出发，贪心扩展存储范围
- 优先扩展增量负载小的方向（左/右）

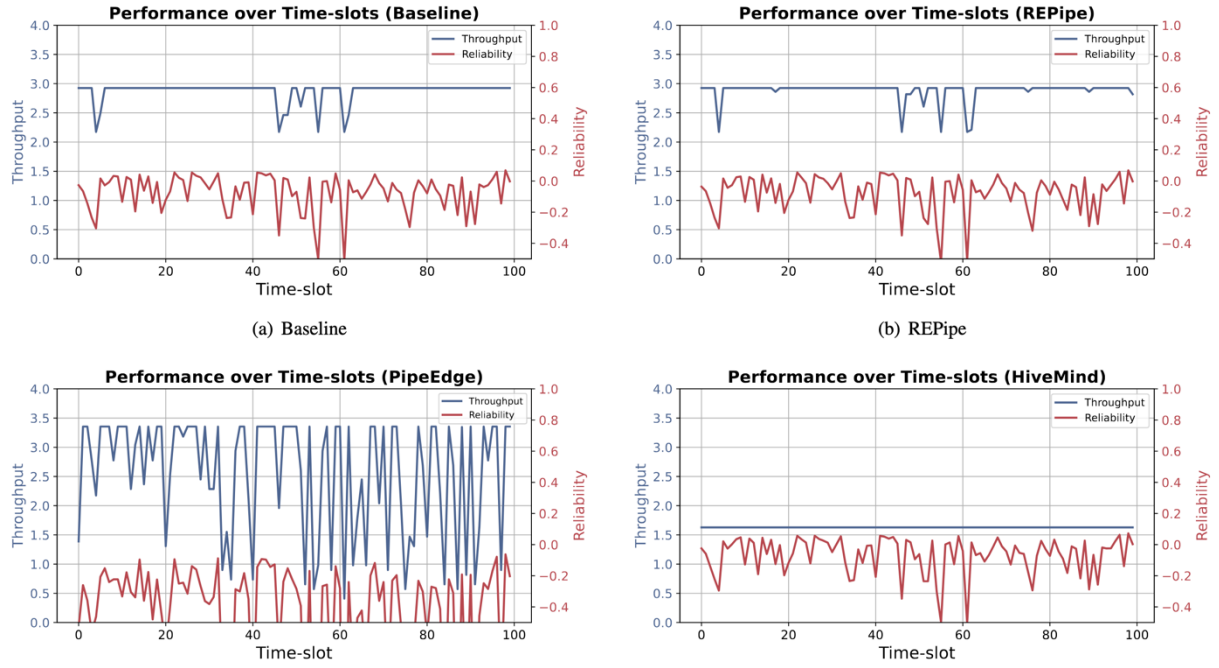
研究内容一：实验结果

实验配置

- 仿真：100个执行周期，资源动态波动
- 真实测试床：4台异构设备（PC/Jetson/树莓派）
- 模型：VGG19、YOLONet、ResNet50







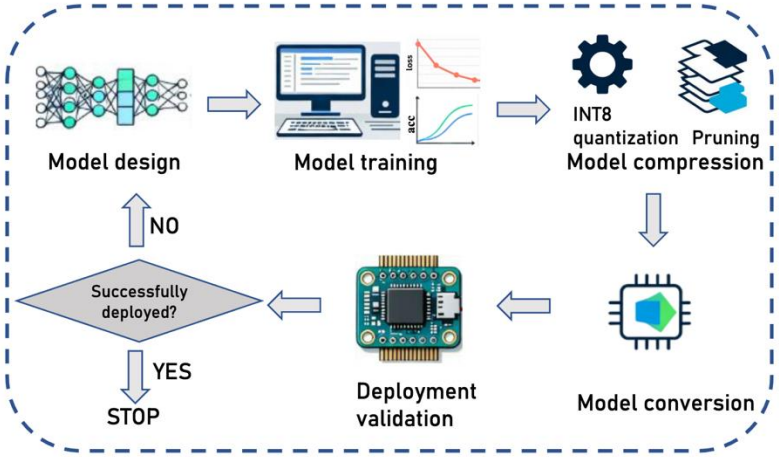
(a) Baseline　　(b) REPipe

研究内容二： LLM驱动的端侧模型定制

背景

MCU资源极度受限： 压缩/轻量化设计，HW-NAS搜索成本高，人工介入

研究内容



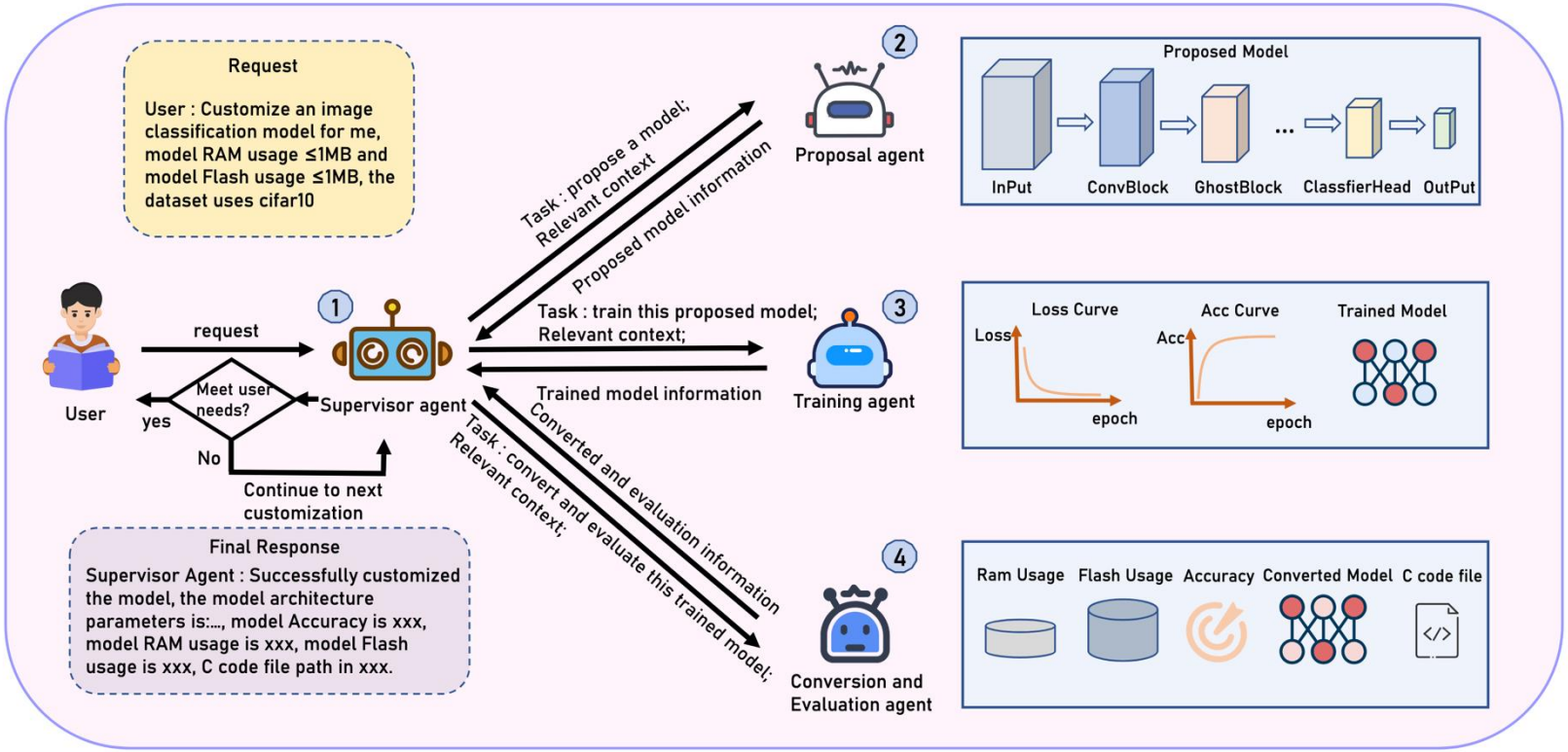
- **端到端框架**：实现从自然语言需求到可部署C代码的全流程自动化模型定制，比传统方法快100倍以上
- **硬件感知机制**：设计硬件在环过滤与历史性能库双重反馈机制，在训练前验证架构可行性并引导高效收敛
- **智能体调度**：提出状态隔离的多智能体调度机制，通过结构化摘要通信确保长周期优化过程的稳定性与可控性

相关成果

- 已投稿 **1 篇论文**：投至 IEEE TMC (CCF A 类/中科院 1 区国际期刊)
- 下一步工作：研究MCU端模型的自动调试与故障诊断机制，并扩展至GPU/NPU/FPGA等异构设备

端到端自动化：自然语言需求→可部署C代码，LLM推理能力替代NAS搜索

研究内容二： 多智能体协同框架

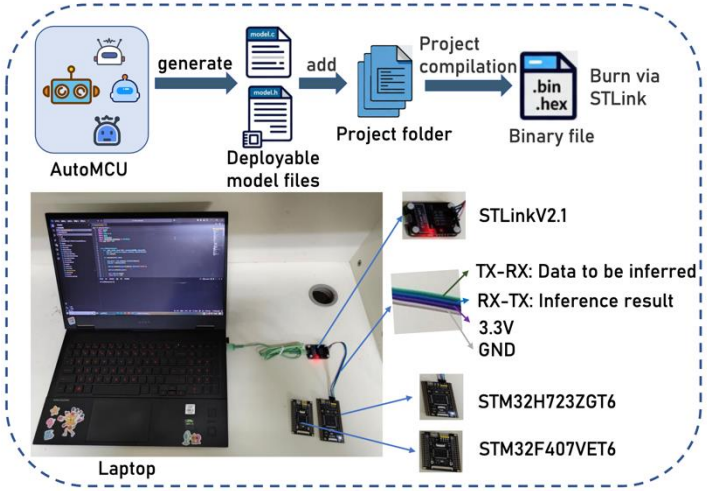


- ❑ Supervisor
  - ✓ 任务分解与调度
  - ✓ 集中控制，状态隔离
- ❑ Proposal Agent
  - ✓ 生成候选架构
  - ✓ 硬件可行性过滤 + 历史性能学习
- ❑ Training Agent
  - ✓ 训练与早停
  - ✓ 有限预算内评估性能
- ❑ Eval & Convert
  - ✓ 资源评估+C代码生成
  - ✓ STM32Cube.AI工具链验证

研究内容二：实验结果

实验配置

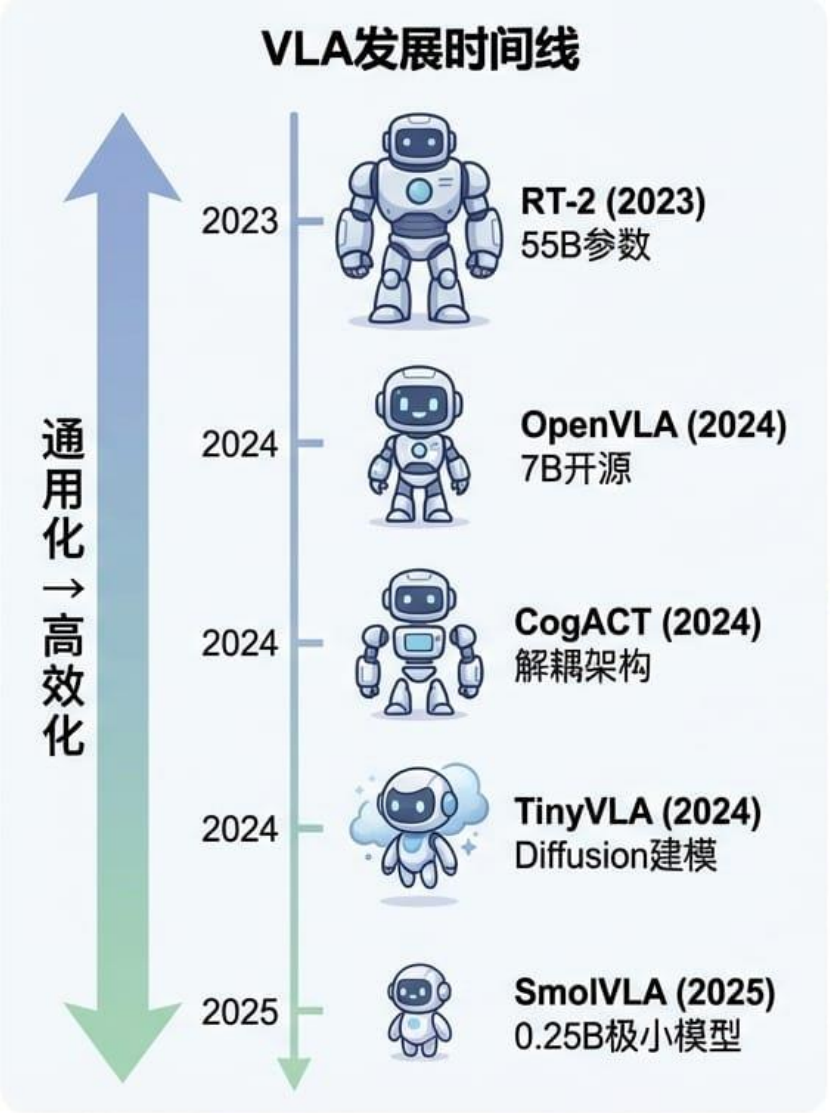
- 数据集：CIFAR-10 / CIFAR-100
- 硬件约束：RAM ≤ 256KB，Flash ≤ 512KB
- 真实设备：STM32F4/F7/H7系列MCU

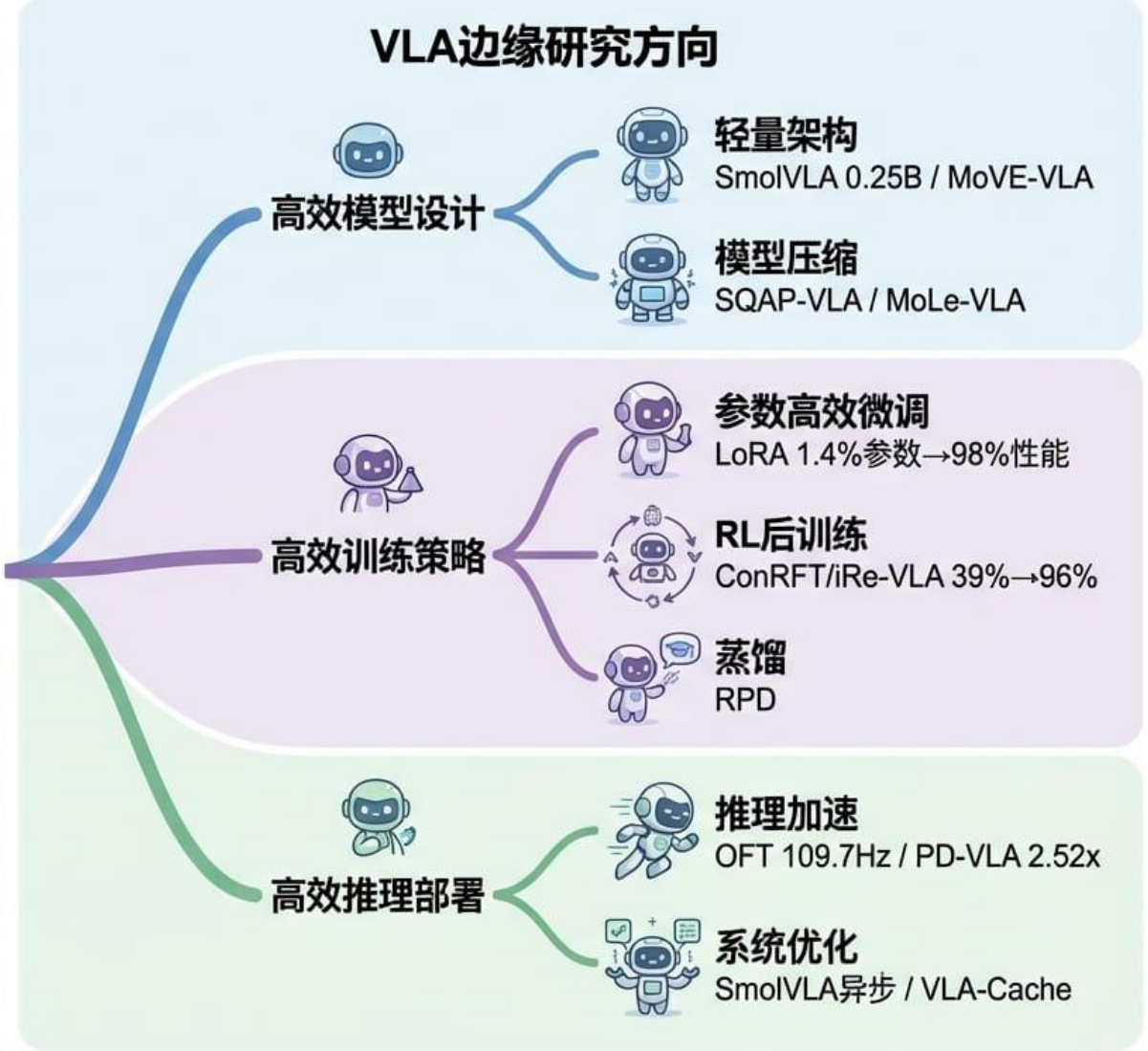


Model	CIFAR10				
	Accuracy (%) ↑	RAM (KB) ↓	Flash (KB) ↓	Time (GPU Hours) ↓	Cost ($) ↓
MobiletNetV2 [21]	84.50	110.44	8670.41	N/A	N/A
ShuffleNetV2 [22]	79.34	**24.00**	4905.33	N/A	N/A
SqueezeNet [51]	82.91	72.00	2892.79	N/A	N/A
MCUNet-in2 [16]	83.69	42.63	2197.57	N/A	N/A
MnasNet [17]	83.61	54.75	12094.38	N/A	N/A
FairNAS-C [18]	84.25	56.44	12159.85	N/A	N/A
μNAS_1024 [20]	**89.75**	1159.99	2902.15	257.48	53.35
AutoMCU_1024 (DeepSeek-V3.2)	89.31±0.16	218.81±70.71 [135.00 , 364.31]	848.97±99.16 [724.85 , 988.67]	1.58±1.07	0.45±0.31
AutoMCU_1024 (MiMo-V2-Flash)	89.30±0.34	210.24±65.74 [128.00 , 384.00]	876.00±126.11 [629.28 , 1015.82]	**1.00±0.48**	**0.23±0.11**
AutoMCU_1024 (qwen-plus)	88.77±0.46	185.66±59.76 [83.38 , 257.12]	888.31±98.52 [707.18 , 1006.38]	1.54±1.04	0.47±0.30
μNAS_256 [20]	87.88	580.92	872.34	173.87	36.03
AutoMCU_256 (DeepSeek-V3.2)	87.62±0.52	124.84±23.80 [83.38 , 160.00]	**466.55±34.61 [402.77 , 504.52]**	1.56±1.25	0.45±0.37

Model	CIFAR100				
	Accuracy (%) ↑	RAM (KB) ↓	Flash (KB) ↓	Time (GPU Hours) ↓	Cost ($)
MobiletNetV2 [21]	50.70	110.44	9120.77	N/A	N/A
ShuffleNetV2 [22]	44.73	**24.00**	5265.68	N/A	N/A
SqueezeNet [51]	42.82	72.00	3073.14	N/A	N/A
MCUNet-in2 [16]	51.06	42.63	2254.17	N/A	N/A
MnasNet [17]	52.34	54.75	12544.73	N/A	N/A
FairNAS-C [18]	45.75	56.44	12610.20	N/A	N/A
μNAS_1024 [20]	60.26	1024.10	2522.94	188.60	39.08
AutoMCU_1024 (DeepSeek-V3.2)	61.63±0.50	157.35±49.24 [112.50 , 257.12]	861.72±81.10 [674.17 , 958.78]	**0.72±1.00**	0.25±0.31
AutoMCU_1024 (MiMo-V2-Flash)	**61.63±0.42**	183.13±65.27 [93.80 , 289.00]	872.20±82.83 [728.46 , 1013.46]	1.28±1.35	0.32±0.34
AutoMCU_1024 (qwen-plus)	61.55±0.81	201.52±43.28 [125.06 , 256.00]	866.86 ± 116.07 [680.83 ,1008.65]	0.83±0.56	**0.24±0.15**
μNAS_256 [20]	58.82	349.00	1448.15	160.80	33.32
AutoMCU_256 (DeepSeek-V3.2)	58.70±0.71	138.21±57.10 [72.95 , 256.00]	**481.49±28.43 [430.25 , 509.80]**	2.43±1.15	0.76±0.37

研究内容三：边缘高效VLA





研究内容三：边缘高效VLA

架构优化

泛化能力增强

持续学习能力提升

π0 2024.10
PaliGemma 3B + Flow Matching

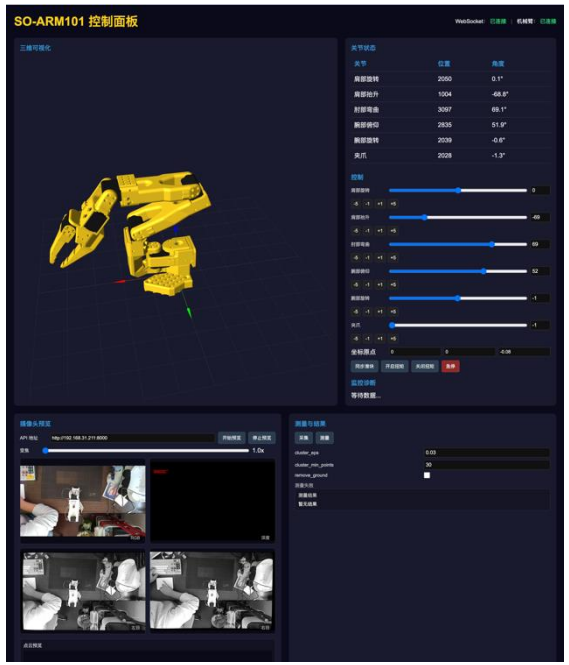
π0-FAST 2025.01
DCT动作压缩，训练5x加速

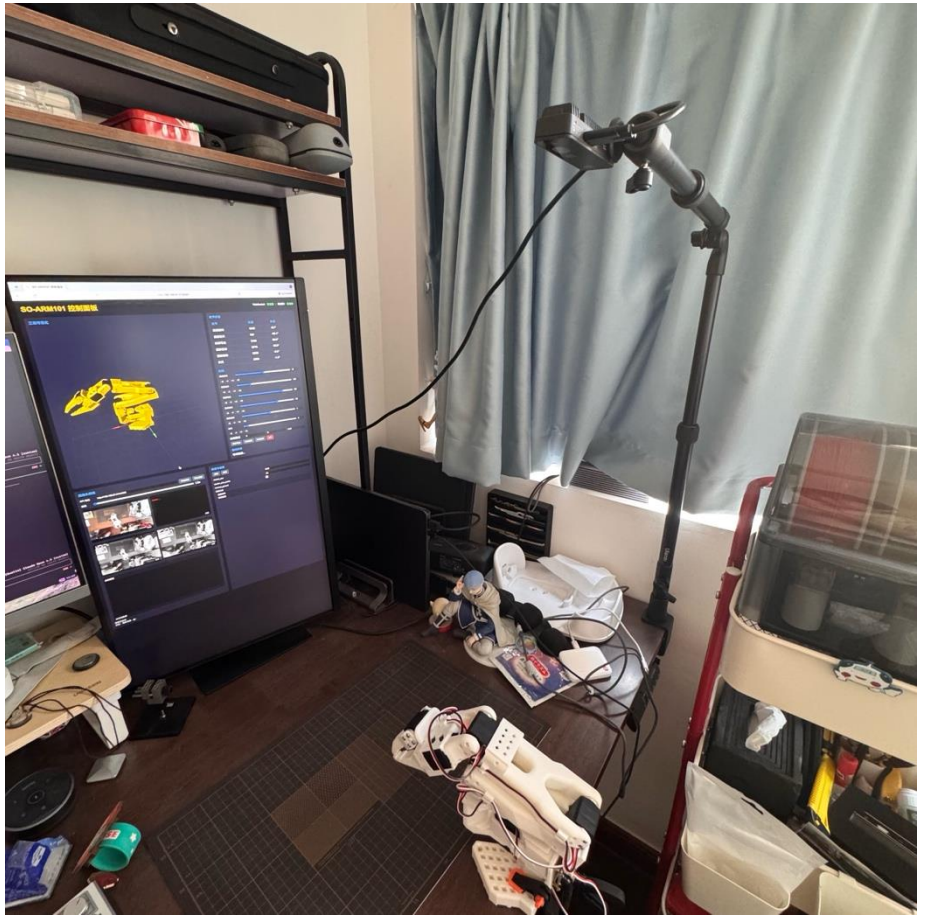
π0.5 2025.04
开放世界泛化，104个家庭

$\pi^*_{0.6}$ 2025.11
自经验学习，持续改进

为什么选择π0

- ✅ 开源生态最完整（OpenPI，权重+代码+微调）
- ✅ 双流架构（Flow Matching + FAST自回归）
- ✅ LeRobot集成，LoRA微调管线完备
- ✅ ~3.3B无边缘方案 → 我们的切入点





汇报提纲

1 个人简介

2 研究背景

3 研究内容

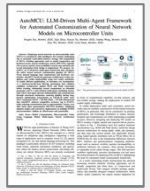
4 **后续计划**

后续计划：研究进展

论文发表/投稿情况

- **IEEE TMC**：2026年1月8日以通讯作者投稿中科院 1 区/CCF A类国际期刊论文 1 篇

- **IEEE TMC**：2026年1月21日以通讯作者投稿中科院 1 区/CCF A类国际期刊论文 1 篇



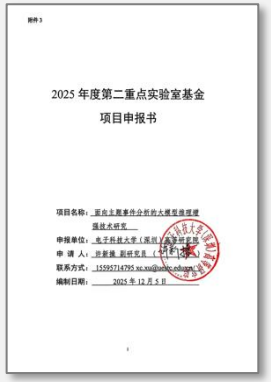
- **IEEE TWC**：2026年1月27日以通讯作者投稿中科院 1 区国际期刊论文 1 篇

- **IEEE/ACM TON**：2026年1月29日以通讯作者投稿中科院1 区/CCF A类国际期刊论文 1 篇

- **IEEE TNSE**：2025年2月2日以通讯作者返修中科院 2 区国际期刊论文 1 篇（Minor）

项目申报/参与情况

- 2025年12月5日申报中电10所第二重点实验室基金项目1项，经费20万，并参加答辩环节
- 2025年10月20日参与北理牵头平安中国重点专项子课题1项



奖项获得情况

- 2025年11月28日获得IEEE ISPCE-AS 国际会议最佳论文奖

后续计划：工作目标

短期目标

方向	目标与时间	产出
异构边缘协同推理	期刊论文扩展投稿（3/4月）	CCF A 会议论文投稿
端侧模型定制与部署	期刊论文扩展投稿（3/4月）	CCF A 会议论文投稿
边缘高效VLA	确定具体研究内容（1个月内）	研究Proposal

中期目标

方向	目标与时间	产出
科研论文	视觉推理+边缘部署联合工作	至少 1 篇论文
边缘高效VLA	完成1篇工作并投稿（1年内）	至少 1 篇论文
项目申报	参与团队项目申报	基金申请书

电子科技大学（深圳）高等研究院
Shenzhen Institute for Advanced Study, UESTC

数据智能研究中心
Data Intelligence Group

感谢聆听

THANKS FOR YOUR ATTENTION

许新操 | 数据智能研究中心

Powered by:

https://github.com/openclaw/openclaw

https://github.com/ResearAI/AutoFigure-Edit