

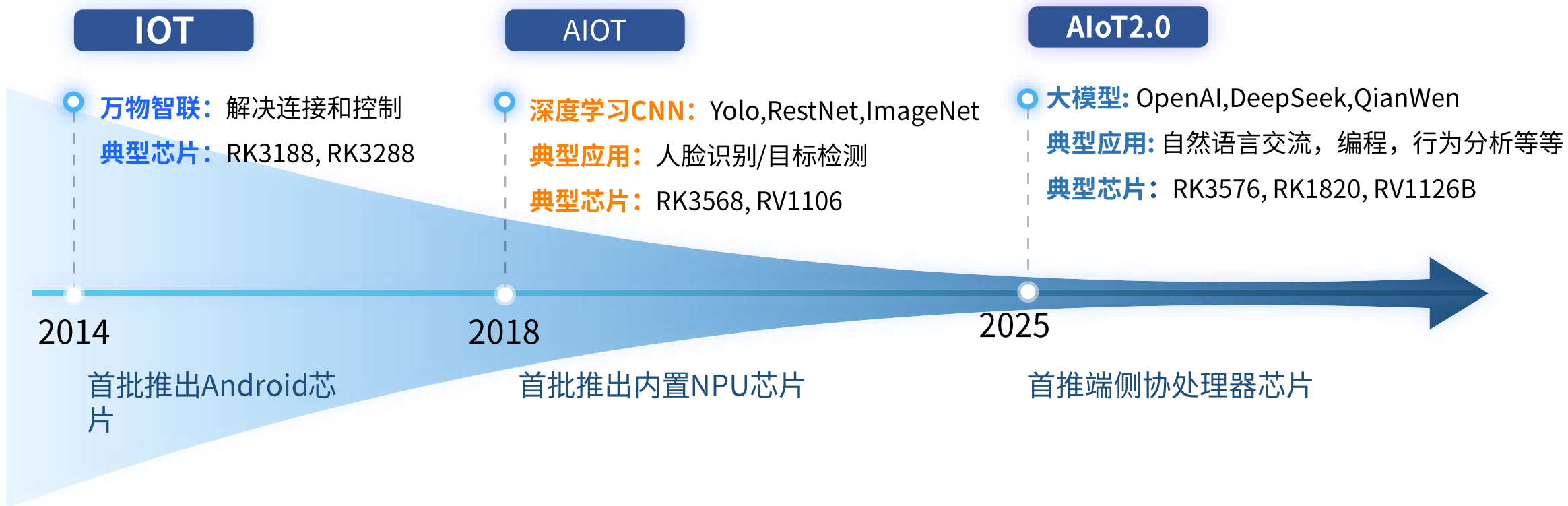


RK182X 大模型算力协处理器

——引领 AIoT 2.0 端侧落地

2026.01

>> 智能设备的重塑



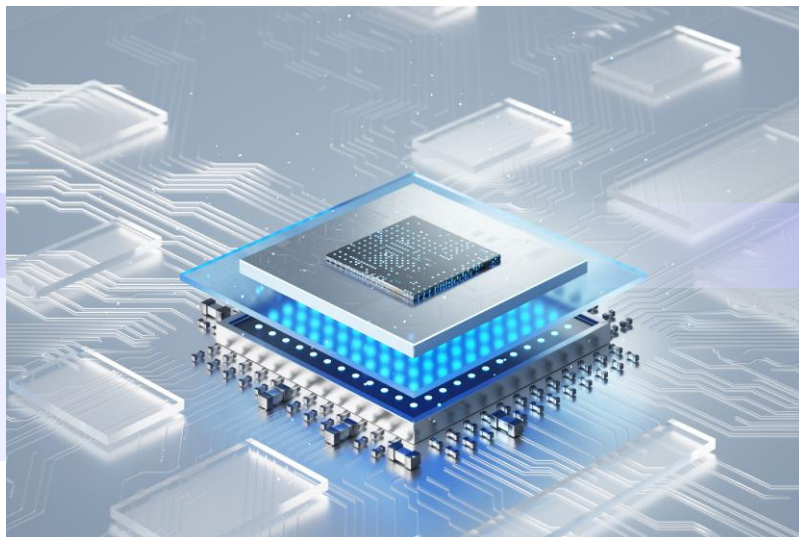
端侧大模型软硬件发展趋势

模型算法迅速发展

能力接近，SIZE变小

- ✓ Qwen2-7B (2024.07)
- ✓ Qwen2.5-3B (2024.12)
- ✓ Qwen3-1.7B (2025.05)

算法能力



硬件算力

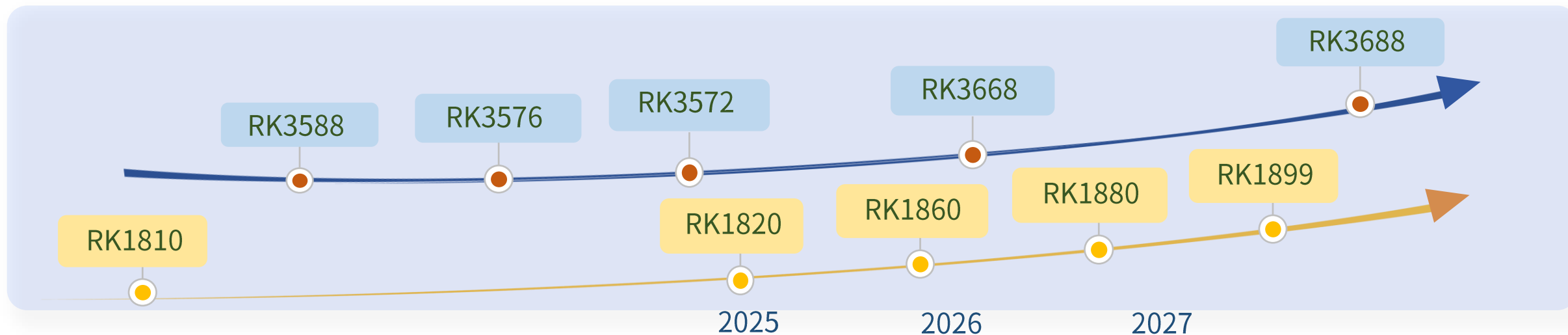
端侧硬件能力提升

端侧芯片迅速迭代

- ✓ 算力提升
- ✓ 带宽提升
- ✓ 算子和模型支持

瑞芯微 AIoT 2.0 新战略

主芯片+协处理器，成为瑞芯微并行研发、快速迭代的双轨重要资源线



双轨研发策略

- 旗舰主控姊妹片迅速迭代
- 研发一代成熟一代
- 协处理器紧跟旗舰主控工艺

主控+协处理器搭配模式

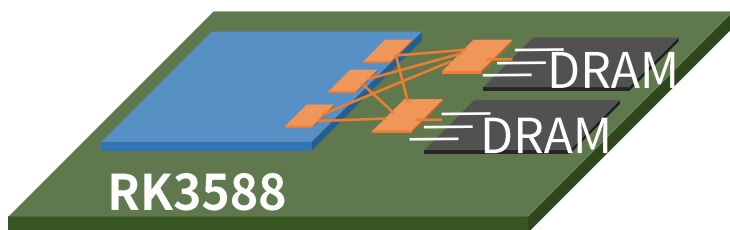
- (高中低端)主控聚焦通用任务
- 协处理器聚焦AI算力和带宽
- 灵活配置->最高性价比

》》 RK182X 3D堆叠技术 - 突破DRAM带宽瓶颈

大模型运算严重依赖SoC的DDR带宽

7B大模型30TPS的带宽需求：3.5GBx30=105GB/s

传统SoC外置DDR的方式，仅提供44GB-100GB的理论带宽，实际带宽7折

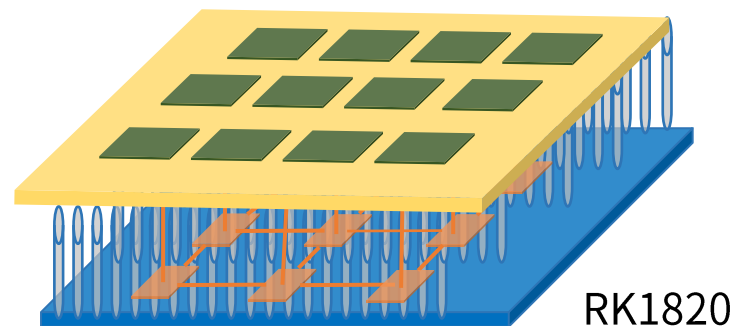


传统 2D 芯片，DDR外置
DRAM接口数量少 (32-128)
带宽受限
接口频率高，功耗大



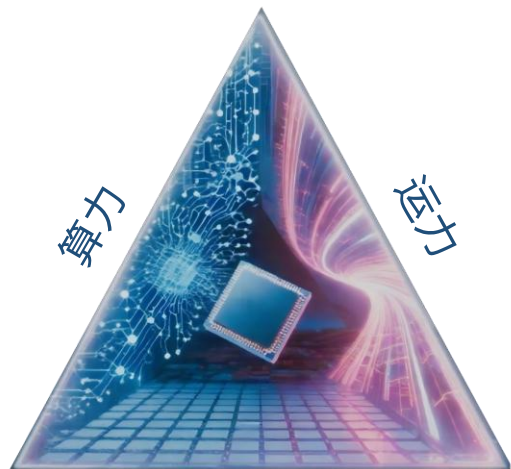
44GB/s >> 1024GB/s

2D芯片向3D芯片持续探索



3D堆叠芯片，DDR内置
DRAM IO接口数量 (>10000bit)
高带宽
接口频率低，功耗低

➤➤ RK182X - 目标 3B/7B大模型



存力

广义算力=算力&&存力&&运力

□运力(DRAM带宽)

- 1TB DRAM带宽

□算力(NPU能力)

- 满足 3B/ 7B 算力要求

□存力(DRAM容量)

- 2.5GB, 满足3B模型需求
- 5GB, 满足7B模型需求

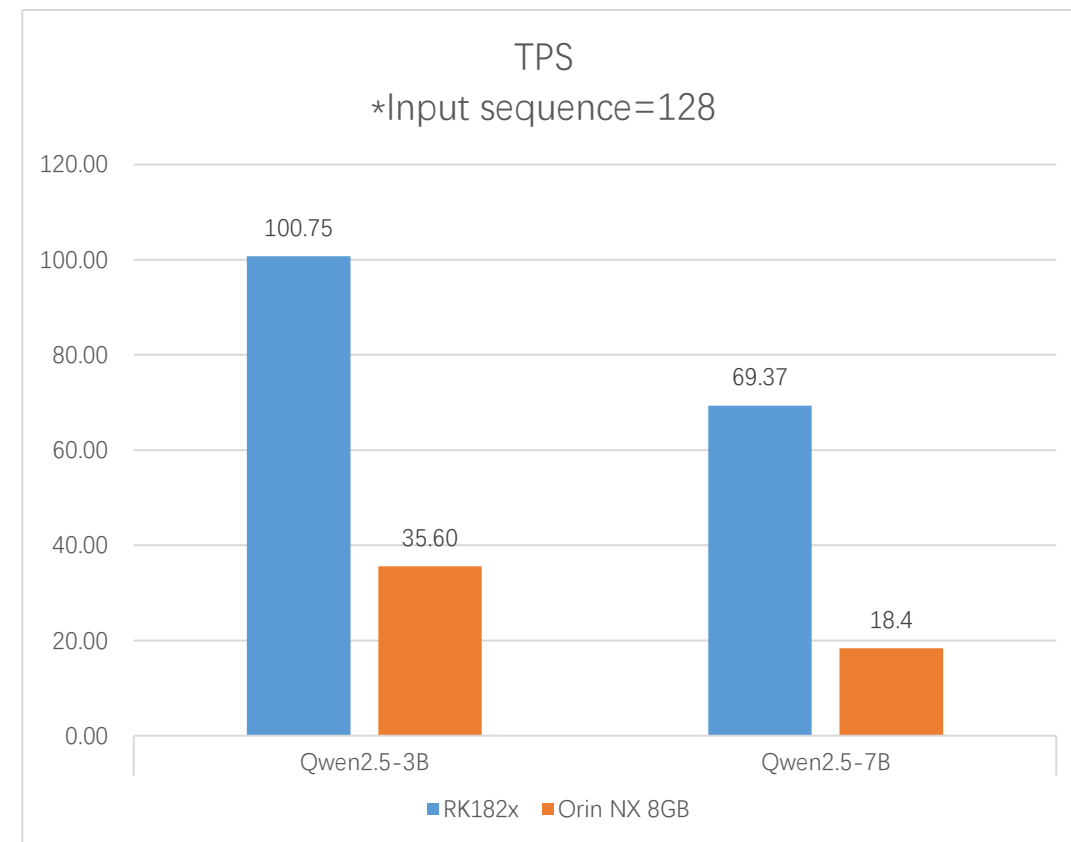
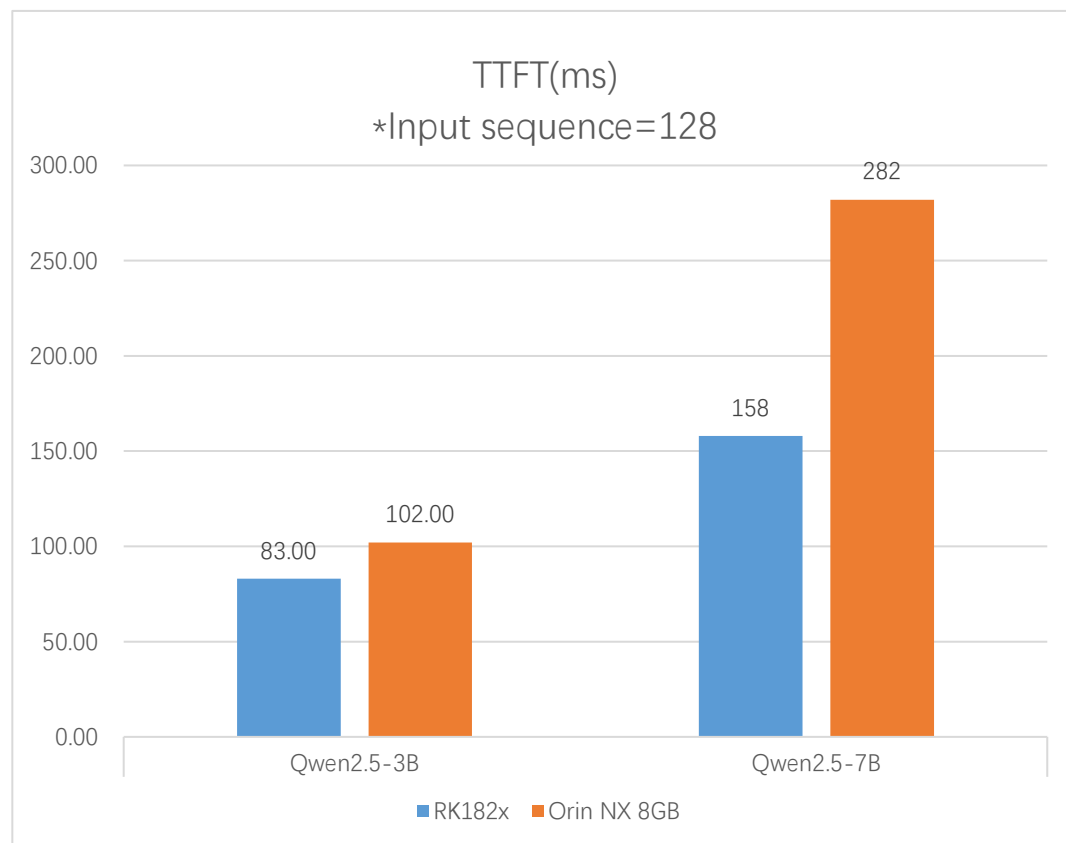
□性能提升

- 端侧模型突破百 token 输出
- 端到端响应延时低至0.1S

□全方位模型算法支持

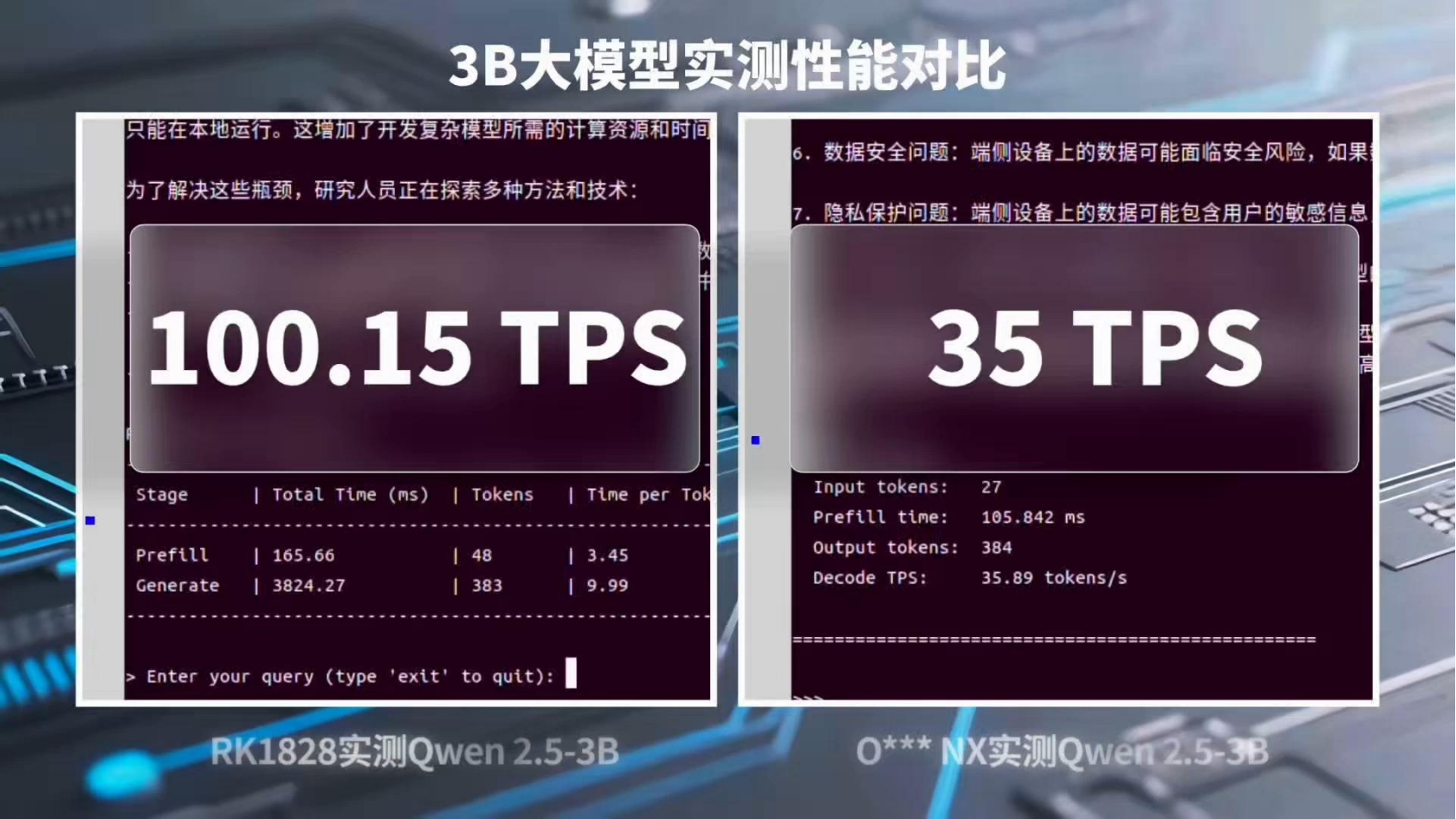
- 文本, 语音大模型
- 图片, 视频大模型
- 传统 CNN 模型

>> LLM 性能对比 – RK182X vs Orin NX 8GB



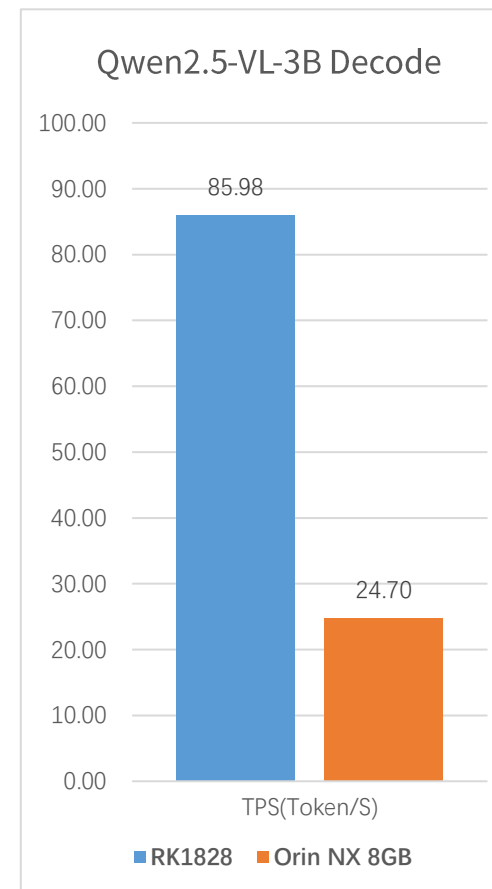
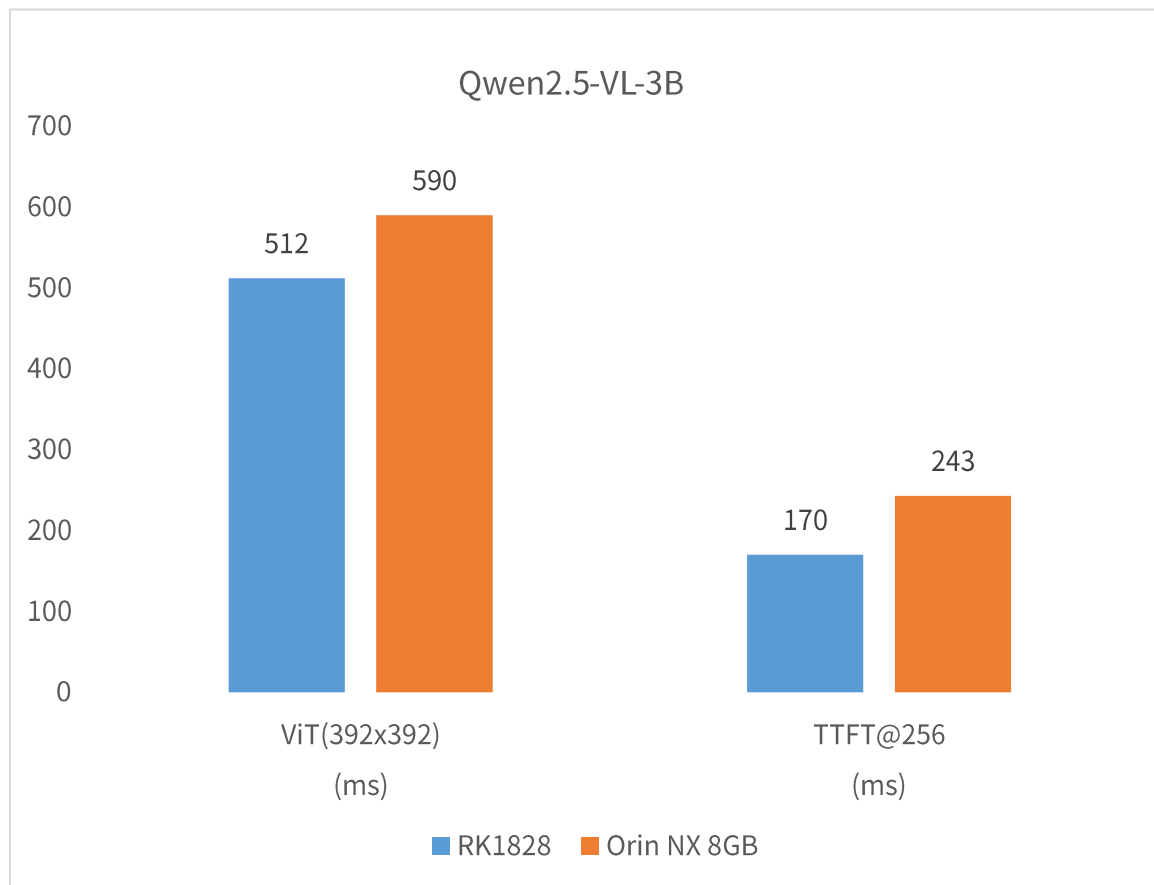
数据来源: Orin NX使用mlc_llm(0.19.0)框架测试

>> LLM 实测性能：体验升级，3B大模型输出破百！



注：数据基于RKNN3 SDK V0.5.0
所有模型使用w4a16量化，VL的LLM部分input为128

>> VLM 性能对比 – RK182X vs Orin NX 8GB



数据来源: Orin NX 数据基于 Llama.cpp 测试

>> 能耗对比 – RK182X vs Orin NX 8GB

模型名称	平台	输入数量 Tokens	输出数量 Tokens	首字时延 TTFT(ms)	输出速度 (tokens/s)	平均功耗	能量效率 (TPS/W)
Qwen 2.5 3B	RK182X	128	128	84.49	102.8	Avg 11.4W	9.02
	Jetson Orin NX 8G			102.00	35.60	Avg 19.1W	1.86

对比竞品：

- 三倍性能
- 四倍能耗比

>> RK182X 典型 LLM 性能

模型名称	输入长度	输出长度	首字时延 TTFT (ms)	输出速度 TPS
Qwen 2.5-0.5B	128	128	22	210.54
Qwen 2.5-3B	128	128	83.09	100.75
Qwen 2.5-7B	128	128	158.28	69.37
Qwen 3-1.7B	128	128	52	134.59
Qwen 3-4B	128	128	106.58	86.87

注：数据基于RKNN3 SDK V0.5.0
所有模型使用w4a16量化，VL的LLM部分input为128

>> RK182X 典型 VLM 性能

模型	图像分辨率	ViT(ms)	LLM TTFT (ms)	输出速度 TPS
FastVLM_1.5B	512*512	146	48.19	144.58
InternVL3-2B	448*448	233.98	48.08	145.38
InternVL3_5-4B	448*448	240.53	106.84	86.63
Qwen2.5-VL-3B	392*392	293	90.39	99.05
Qwen2.5-VL-7B	392*392	284	159	69.25
Qwen3-VL-2B	384*384	166.39	52.73	99.62
Qwen3-VL-4B	384*384	190.30	106.94	66.46

注：数据基于RKNN3 SDK V0.5.0
所有模型使用w4a16量化，VL的LLM部分input为128

端侧大模型应用方向



智能座舱



智慧家庭存算中心



教室监控分析



会议纪要设备



智能机器人



安防视频分析
结构化处理

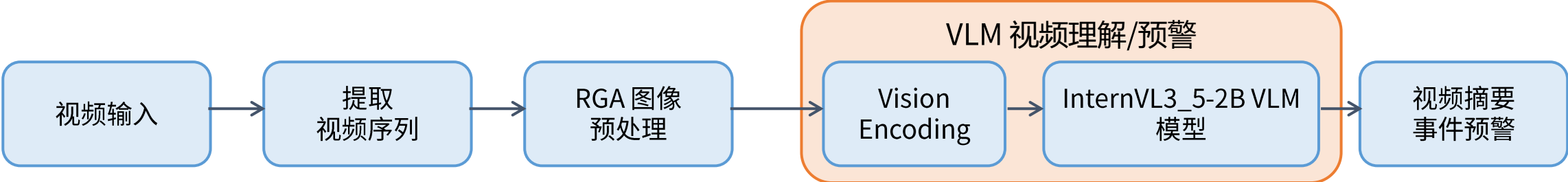


边缘AI计算



工业缺陷检测
运动控制

>> 典型应用 – 视频摘要 (RK3588+RK1820)



- 技术方案：
 - RK3588 检测目标后，提取视频序列（YOLO等）
 - RK1820 运行 InternVL3_5-2B
 - RK1820 执行视频摘要和事件预警推理
- 主要特性：
 - 监控视频压缩为精华片段，输出视频摘要
 - 支持任意定制的万能异常事件预警

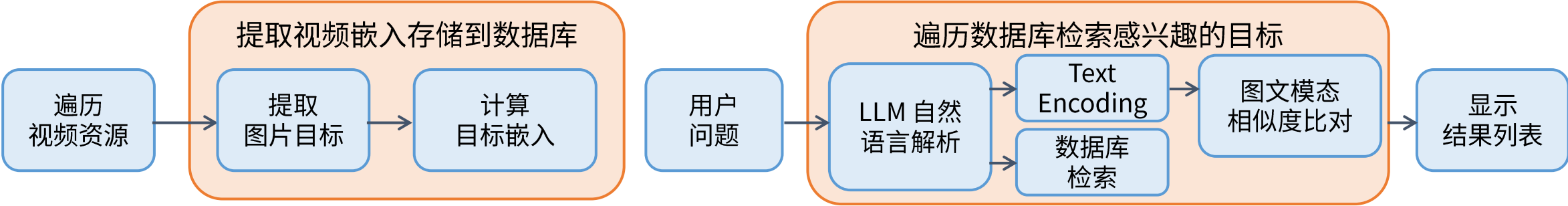
• 性能：

模块	Image Encoding（单张）	总结摘要（5s）	预警触发
耗时	<350ms	0.9s - 1.2s	0.9s-1.6s

典型应用 – 视频摘要 (RK3588+RK1820)



>> 典型应用 – 文搜图搜 (RK3588+RK1820)

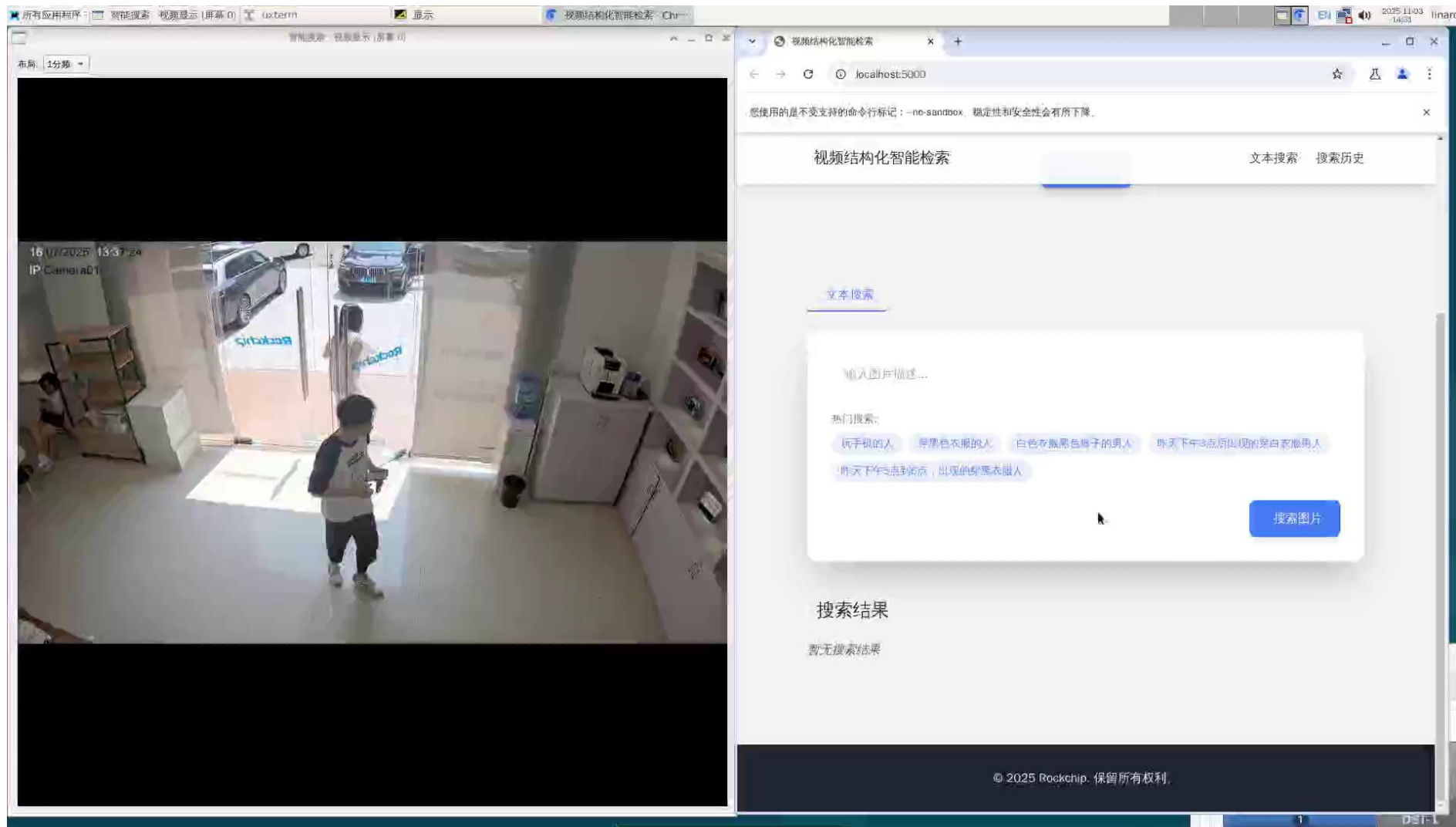


- 技术方案：
 - RK3588 进行视频解码，YOLO 检测
 - RK1820 运行 CLIP 和结构化模型提取目标特征
 - RK3588 执行目标检索
- 主要特性：
 - 使用图片和文本的 CLIP 特征
 - 使用 RK1820 加速提取和检索效率

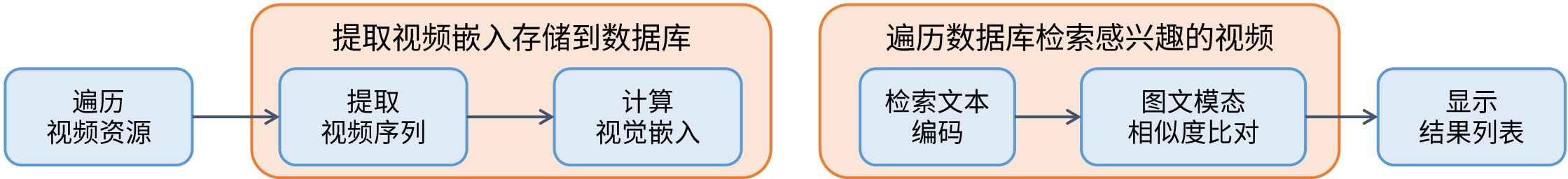
• 性能：

模块	目标检测	目标特征提取	LLM 自然语言理解
耗时	< 20ms	< 350ms	3.5s - 4s

典型应用 – 文搜图搜 (RK3588+RK1820)



>> 典型应用 – 视频检索 (RK3588+RK1820)



• 技术方案:

- RK3588 提取视频序列
- RK1820 运行 Vison-Embedding-2B 模型
- RK1820 执行视频检索

• 主要特性:

- 使用时序序列的视觉嵌入，而非 CLIP 特征
- 使用 RK1820 加速提取和检索效率

• 性能:

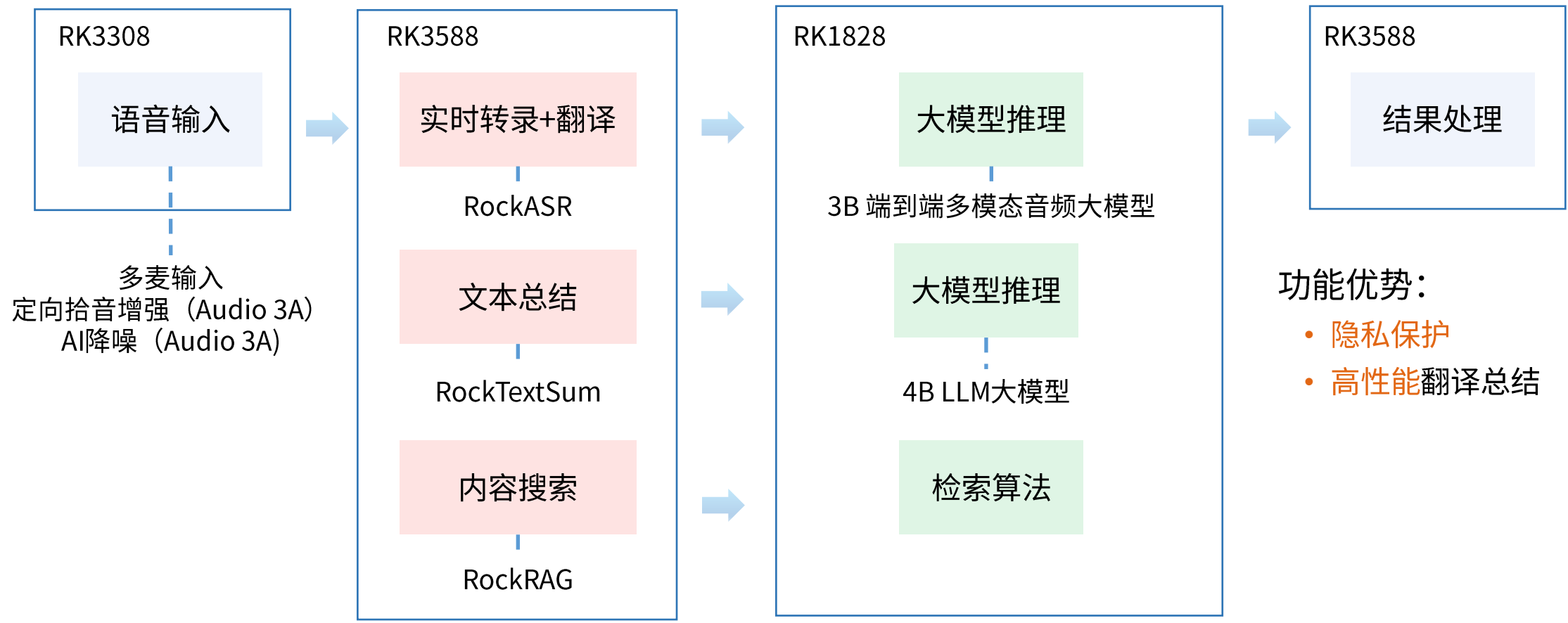
模块	Image Encoding (单张)	Text Encoding
耗时	<350ms	60ms - 80ms

典型应用 – 视频检索 (RK3588+RK1820)



>> 典型应用 – 会议音频

- 端侧多模态音频大模型，实现语音识别转文字、语音翻译、会议摘要总结等功能



- 功能优势:
- 隐私保护
 - 高性能翻译总结

>> 典型应用 – 会议音频

- 产品应用方向：
 - 涉密会议记录与摘要生成
 - 涉密会议、无网/ 弱网环境语音转写与总结
 - 个性化语音交互终端
- 应用模型：
 - ASR, TTS: 3B 端到端音频多模态大模型
 - 文本总结: 4B LLM 大模型

- 方案优势：
 - 高安全性: 数据本地化, 适配涉密环境
 - 低延迟: 语音识别与翻译响应迅速
 - 可定制化: 一次投入, 长期使用; 无持续 API 调用费; 可做深度定制
- 性能:

模块	TTS	ASR
首包耗时 (ms)	210	3500
RTF	0.08-0.10	0.089-0.196

模块	文本总结
处理速度 (s/ 千音频 token 输入)	10.13

典型应用 – 会议音频

- 会议转写和实时翻译

会议记录

会议纪要

Speaker A: Good morning, everyone. Thank you for joining today's conference call. Let's start with the sales performance review. David, could you provide us with the latest figures?

Speaker B: Certainly. Our sales have increased by 10% this quarter compared to the previous one. The new marketing strategy has significantly boosted our online sales.

Speaker A: That's excellent news. Emily, how are our international markets performing?

Speaker C: We've seen a 15% growth in Europe, but sales in Asia have been stable, with no significant increase. We're planning to adjust our strategy in Asia to better target local preferences.

Speaker A: Good idea, John. Any updates on the customer feedback?

Speaker D: Yes, customer satisfaction has improved by 8% this quarter. However, there are still some concerns about delivery times, which we are addressing with our logistics team.

Speaker A: Thank you for the updates. Are there any issues we need to discuss further?

Speaker B: We might need additional resources to support the increased demand. Can we discuss this in detail in our next meeting?

Speaker A: Absolutely, let's schedule a follow-up call for next Monday. Thank you everyone for your contributions.

发言人A: 大家早上好。感谢各位参加今天的电话会议。我们先从销售业绩回顾开始。大卫，能给我们提供最新数据吗？

发言人B: 当然可以。本季度我们的销售额较上一季度增长了10%。新的营销策略极大地推动了线上销售。

发言人A: 这真是个好消息。艾米丽，我们的国际市场表现如何？

发言人C: 欧洲市场增长了15%，但亚洲市场的销售额一直很稳定，没有明显增长。我们计划调整亚洲市场的策略，以更好地迎合当地偏好。

发言人A: 好主意。约翰，客户反馈方面有什么更新吗？

发言人D: 是的，本季度客户满意度提升了8%。不过，关于交货时间仍有一些问题，我们正在和物流团队一起解决。

发言人A: 感谢各位的更新。还有什么问题需要我们进一步讨论的吗？

发言人B: 我们可能需要额外的资源来支持增长的需求。我们能在下次会议中详细讨论这个问题吗？

发言人A: 当然可以，我们把跟进电话会议安排在下周一吧。感谢大家的付出。

- 会议纪要总结

2025_07_09_08-19_NOTES.txt

会议内容：

发言人A: 大家早上好。感谢各位参加今天的电话会议。我们先从销售业绩回顾开始。大卫，能给我们提供最新数据吗？

发言人B: 当然可以。本季度我们的销售额较上一季度增长了10%。新的营销策略极大地推动了线上销售。

发言人A: 这真是个好消息。艾米丽，我们的国际市场表现如何？

发言人C: 欧洲市场增长了15%，但亚洲市场的销售额一直很稳定，没有明显增长。我们计划调整亚洲市场的策略，以更好地迎合当地偏好。

发言人A: 好主意。约翰，客户反馈方面有什么更新吗？

发言人D: 是的，本季度客户满意度提升了8%。不过，关于交货时间仍有一些问题，我们正在和物流团队一起解决。

发言人A: 感谢各位的更新。还有什么问题需要我们进一步讨论的吗？

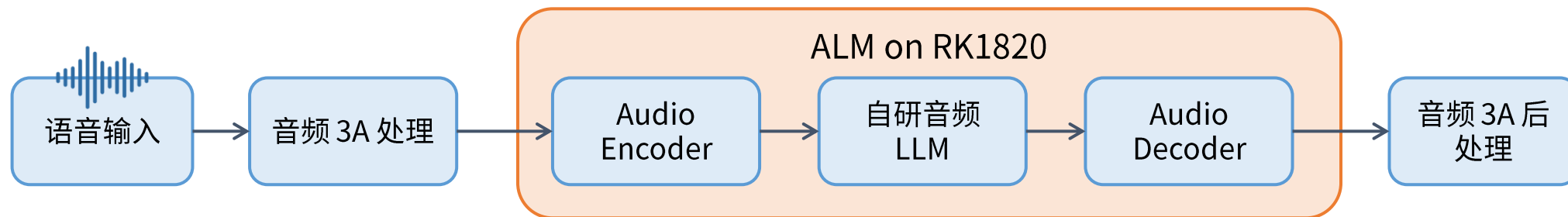
发言人B: 我们可能需要额外的资源来支持增长的需求。我们能在下次会议中详细讨论这个问题吗？

总结内容：

本次电话会议主要讨论了销售业绩、国际市场表现、客户反馈及后续跟进安排。销售额较上一季度增长10%，这得益于新营销策略的推动，线上销售占比提升。欧洲市场表现出色，同比增长15%，而亚洲市场的销售额保持稳定，但计划调整亚洲市场策略以进一步优化表现。客户满意度提升了8%，不过交货时间仍存在一些问題，需要重点关注和改进。会议还指出当前增长需求需要额外资源支持，并安排下周一进行跟进，确保各项措施得到有效落实。

会议总结

典型应用 – 同声传译 (RK3576+RK1820)



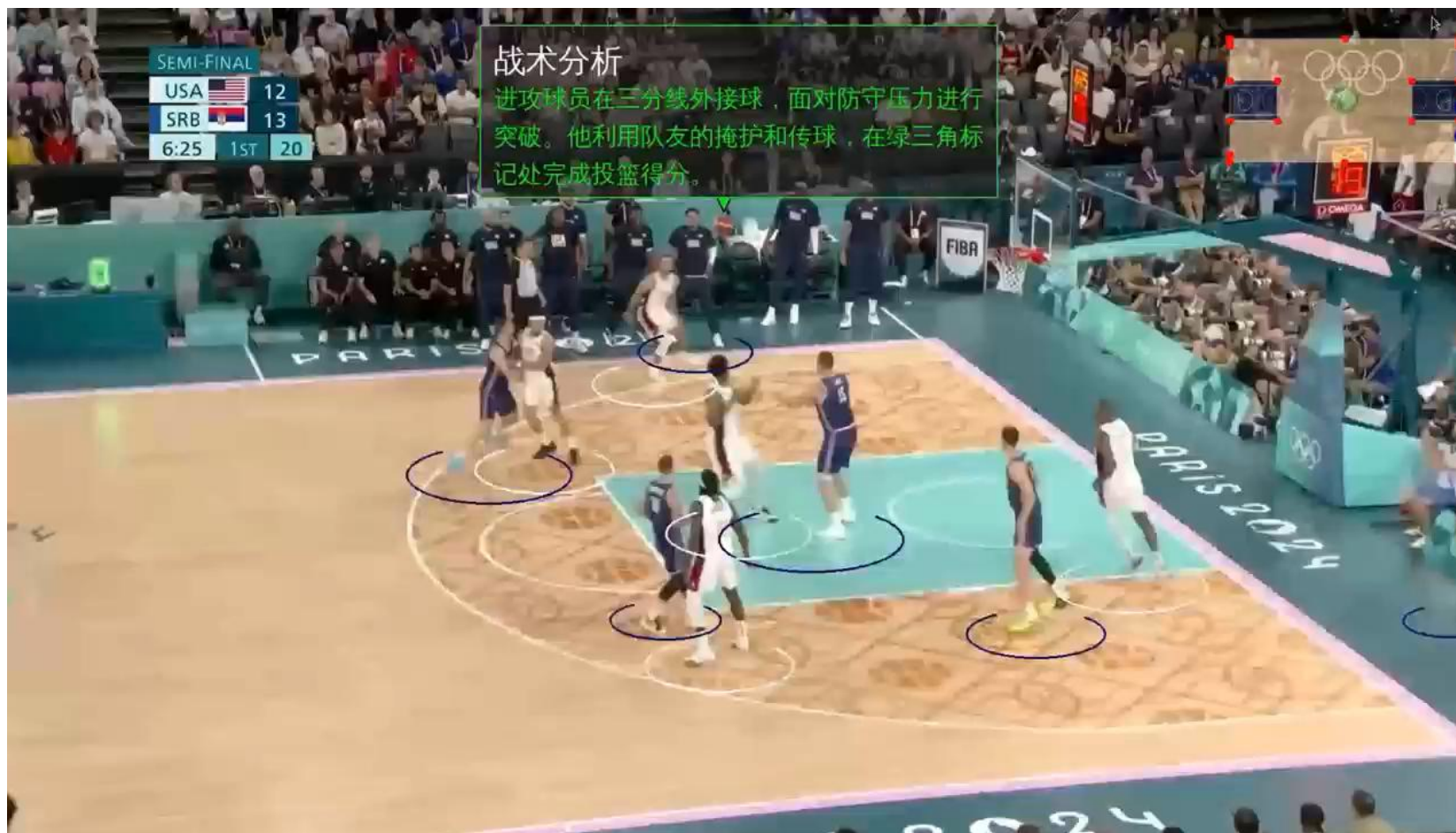
- 技术方案:
 - 音频 3A: RK3576 CPU处理
 - 端到端大模型: 自研模型, SIZE 约 3B
- 主要特性:
 - 支持音频预处理: AI 降噪、AI 回声消除、区分不同发言人声纹
 - 流式处理 翻译准确性高 支持中英文

- 性能:
 - 平均延迟 3s
 - RTF<0.8

典型应用 – 同声传译 (RK3576+RK1820)



应用案例 – AITV战术分析



应用案例 – AITV内容增强

使用VLM模型，打破传统“线性播放”模式 - 精彩剪辑

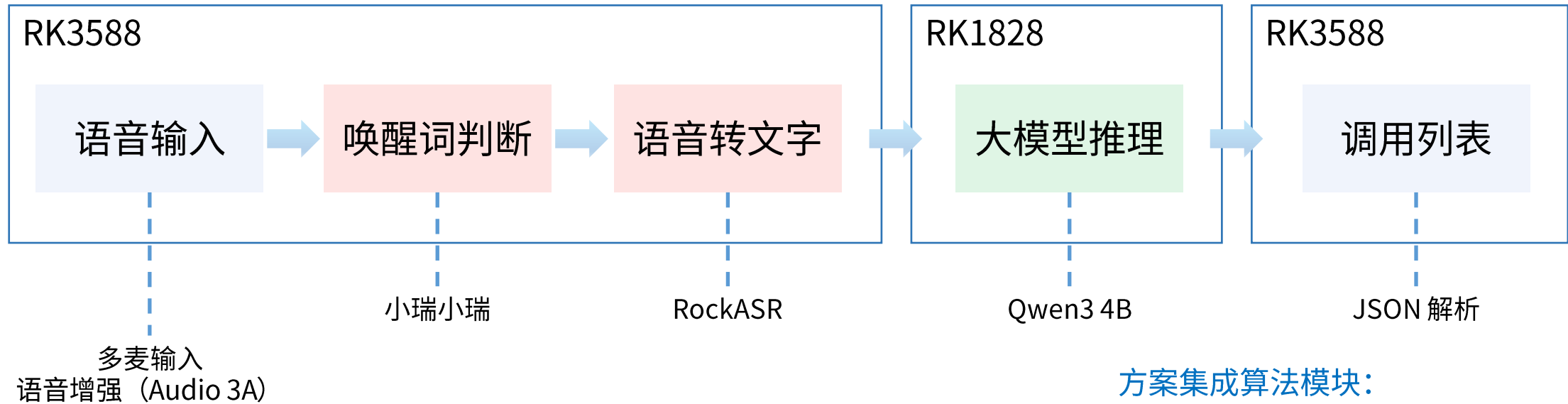
- 体育赛事 精彩时刻、PVR 集锦、实时战术分析
- 电视剧 精彩剧情提取，跳过广告和拖沓剧情(剧情完整挤干水分)



视频结构化 自定义高光时刻

>> 典型应用 – 边侧智能家居算力中枢

- RK182X 可作为本地语义理解与控制命令算力中枢，配合主 SoC 实现从语音唤醒到设备控制的全链路处理，在无网络环境下完成语音识别、语义理解和控制指令生成



构建安全、可离线的语音控制中心

- 方案集成算法模块：
- 语音降噪 (librkaudio_vqe)
 - 语音唤醒 (librkaudio_vqe)
 - 语音转文字 (RockASR)
 - 语义理解 (Qwen3-4B)
 - 设备控制 (JSON解析、MCP)

>> 典型应用 – 边侧智能家居算力中枢

- 产品应用方向：
 - 智能家居语音控制中心
 - 会议语音控制与自动化管理
- 方案优势：
 - 隐私安全：语音数据与指令均在本地处理
 - 离线运行：脱离网络环境仍可完整执行控制流程
 - 全链路处理：涵盖语音唤醒、ASR，语义理解
 - 可扩展性：支持多麦输入、语音增强与场景定制

- 性能：
 - 语音指令延时
测试条件：输入提示词长度1250 tokens，输出36 tokens
上下文支持：最大16K tokens

模块	prompt	decode	JSON Parse	总计
耗时（ms）	1891.454	743.918	10.85	2646.222

注：该测试中输入提示词长度为1250 tokens，输出长度为36 tokens

典型应用 – 边侧智能家居算力中枢



语音输入指令

打开卧室空调设置到23度风速调到2

卧室的空调设置温度23度

卧室的空调设置风速2

>> 典型应用 – 领域专家数字人

- 基本功能：

• 沉浸式真人形象数字分身

• ASR/ TTS 驱动语音对话

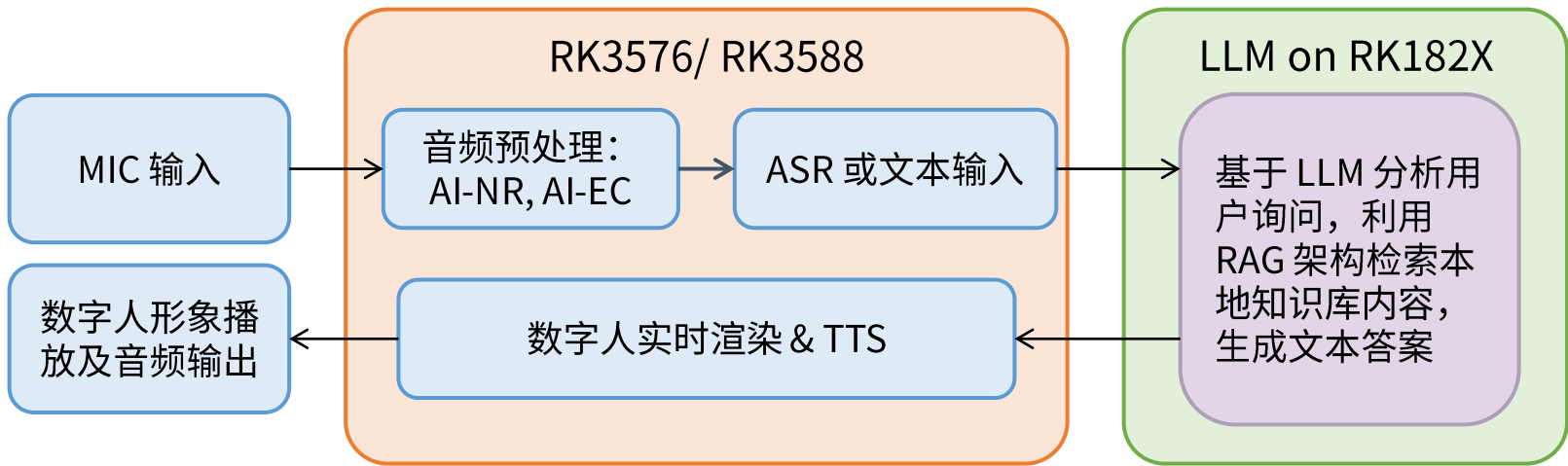
• 基于 RAG 的本地文档知识快速检索

• 基于 RAG 的本地知识库图谱精准检索
- 产品应用方向：

• 专业知识顾问（医院、政企等）：内部数据安全不出网、支持语音问答、实时讲解

• 教育培训（企业培训、课堂、展厅讲解）：数字讲师沉浸式互动

• 私有助理（个人 AI 助手、科研助理）：本地隐私、语音交互



- 算法模型：
- ASR , TTS , MiniCPM, Qwen3-4B
- 性能：

模块	TTS	ASR
首包耗时 (ms)	210	3500
RTF	0.08-0.10	0.089-0.196

典型应用 – 领域专家数字人



- 快速搜索 RAG:
可实时导入 .txt, .md 文档进行解析进行问答 (PDF 格式支持开发中)
- L1 快速搜索 RAG, 解析只需要 1s 即可进行问答
- L2 快速搜索 RAG, 增加向量化稠密检索, 需五分钟左右生成向量, 回答相比较 L1 更加精准
- 本地知识图谱检索:
离线构建的本地知识图谱, 比快速搜索 RAG 更精准

典型应用 – AI Agent 智慧点餐助手

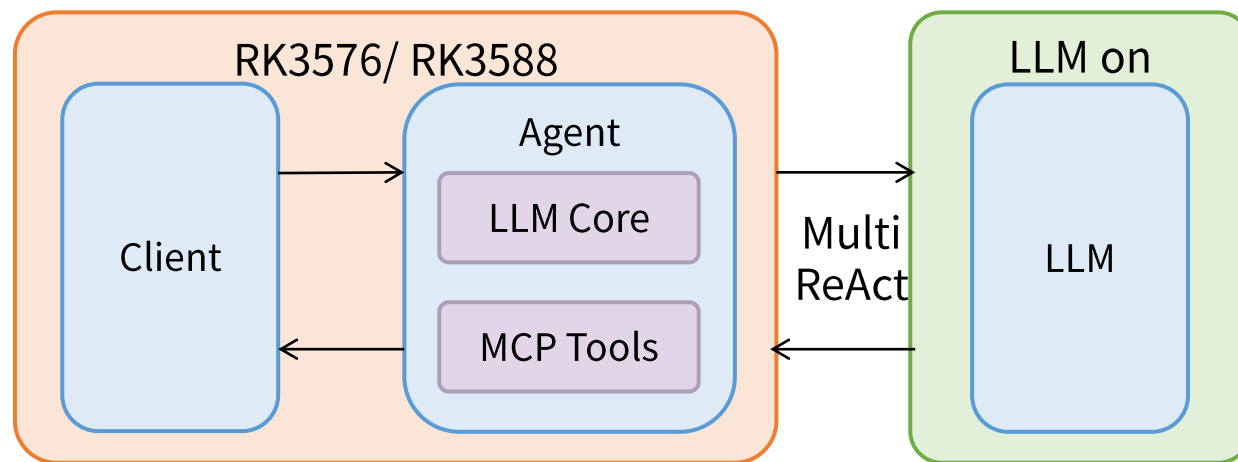


点餐助手

- 口语化交互
- 根据人数、口味偏好等信息,
- 自动规划菜单
- 执行点餐操作

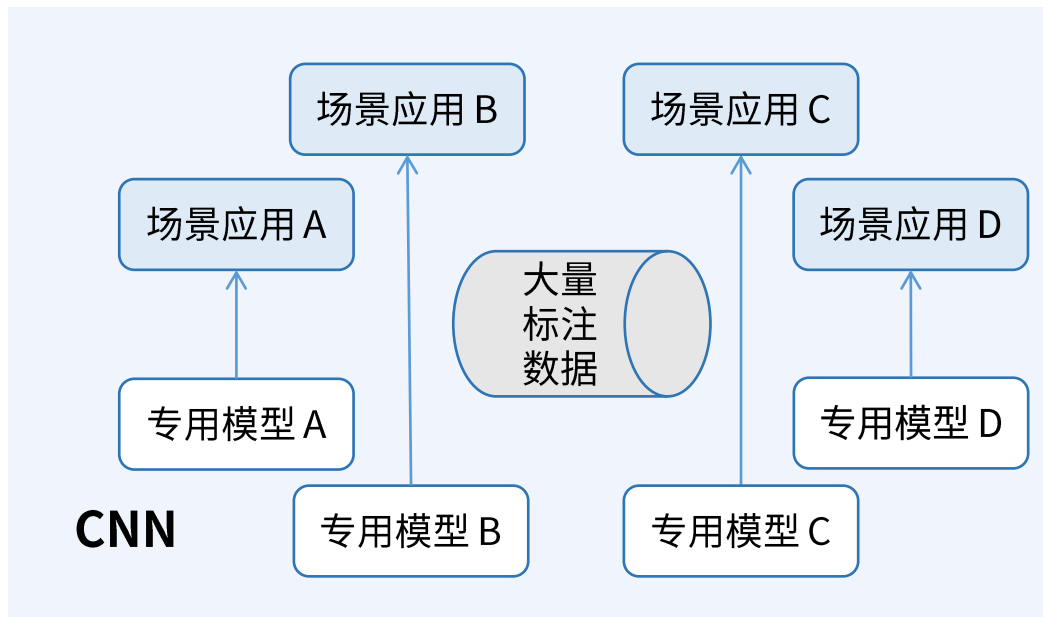
典型应用 – AI Agent 助理

- 基本功能：
 - 基于 LLM 分析理解用户需求，AI 自主决策，自主调用工具，实现产品的智能化升级
- 方案优势：
 - 基于 Android 平台，学习和开发成本低支持标准 MCP 协议
 - 支持标准 MCP 协议
 - 轻量级框架，快速部署，可端侧可云端
 - 支持单个/多个智能体协作框架



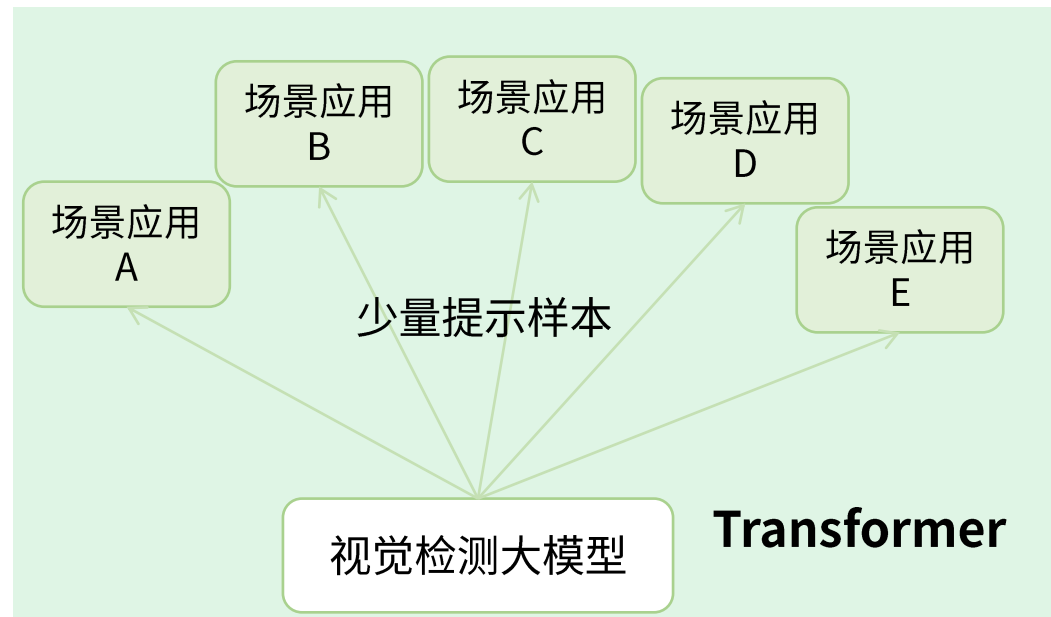
- 产品应用方向：
 - 服务业
 - 零售业
 - 数字政务助理
 - 家庭智能网关
 - 语音助手（车载）
 - 会议助手
 - ...

典型应用 – 工业视觉检测



行业痛点

1. **精度不足**: 对细微瑕疵、复杂纹理等高难场景漏检率高
2. **数据依赖**: 模型训练依赖大量标注数据，成本高、样本获取难
3. **迭代慢**: 产线、环境变更，模型重新训练与调试周期长

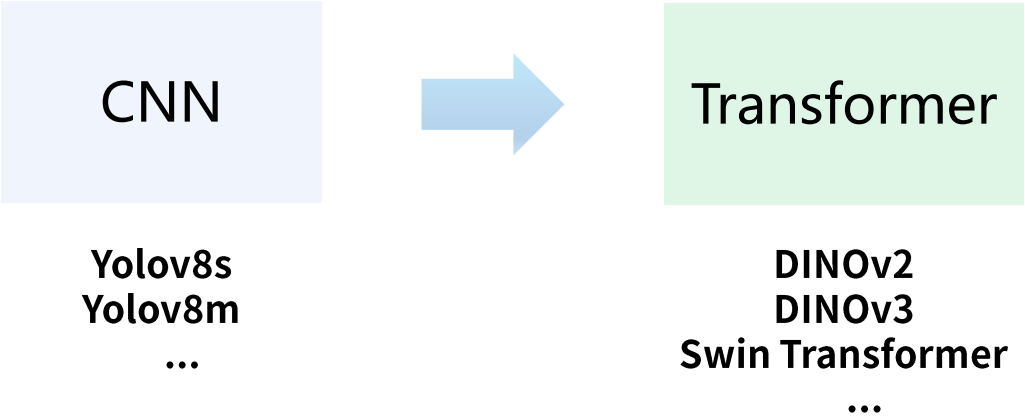


RK182x 视觉大模型方案

1. **样本学习**: 每类缺陷仅需数十张样本即可完成模型构建
2. **强泛化能力**: 单一模型可适配多种工业场景
3. **灵活部署**: 快速响应产线变更，提升灵活性。

>> 典型应用 – 工业视觉检测

- 工业产线检测对模型精度、泛化性与响应速度要求极高
- 基于 RK182X 高带宽、高算力平台，可高效部署 DINOv2、DINOv3，Swin Transformer 等模型，支持无监督与有监督学习，快速适配不同检测场景



- 产品应用方向：
 - 柔性材料表面检测（纺织、薄膜、锂电极片）
 - 高反光材质检测（金属、玻璃、汽车零部件）
 - 小批量多品类精密检测（3C电子、结构件等）

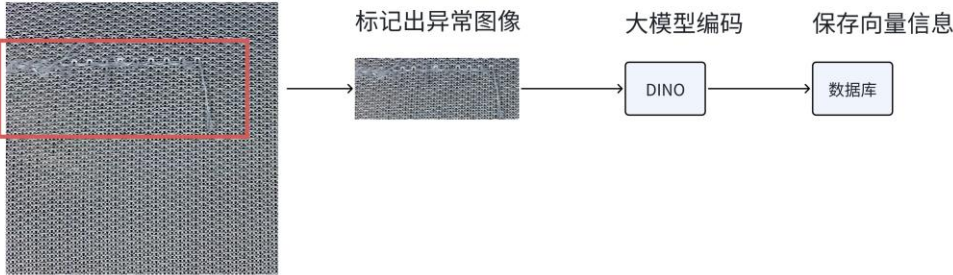
测试项目 DINOv2_vits16 (224×224)	RK3588	RK1820	RK1820 vs RK3588
推理吞吐量	31.36 FPS	180 FPS	+474%
单帧平均延迟	122 ms	44 ms	-64%

>> 典型应用 – 工业视觉检测

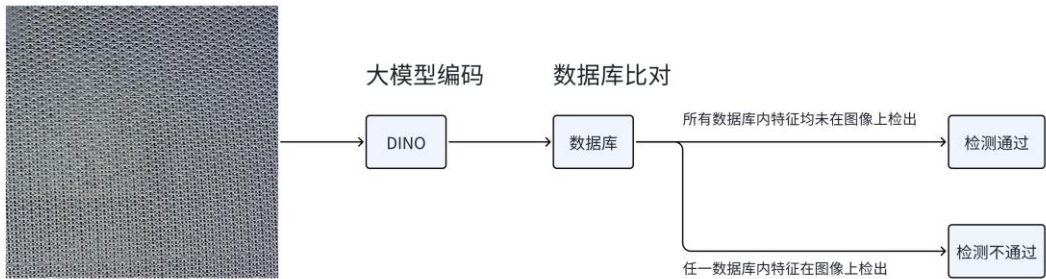
- 有监督学习
适用于人工标记异常，瑕疵样本多的场景

标注阶段

人工标注出异常部分



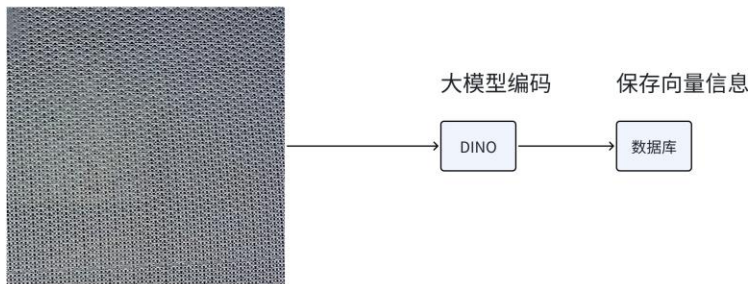
检测阶段



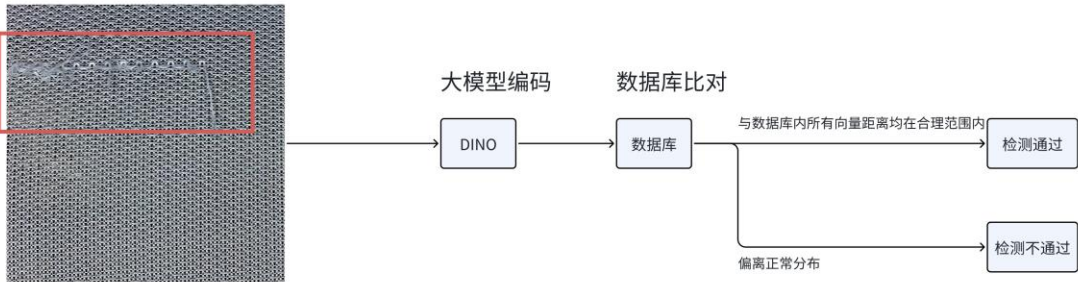
- 无监督学习
适用于只输入正常样本，瑕疵样品少的场景

标注阶段

使用一定数量的正常样本构建数据库

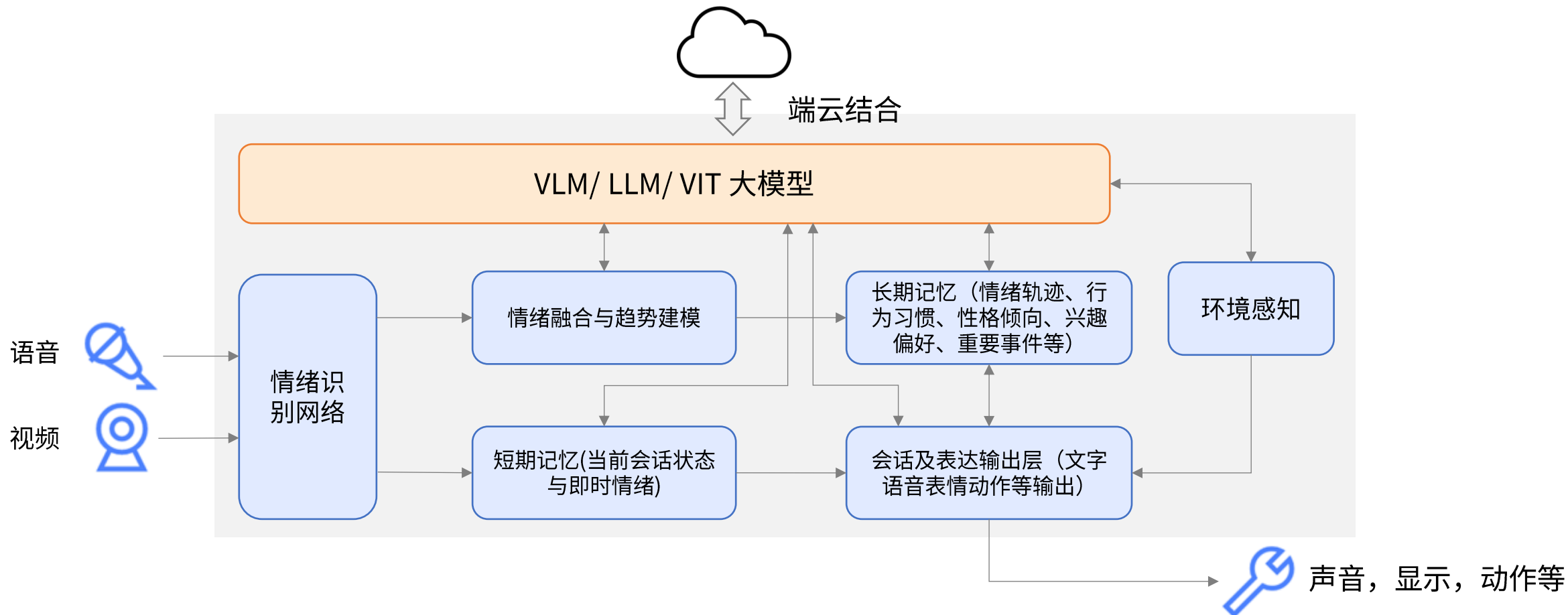


检测阶段



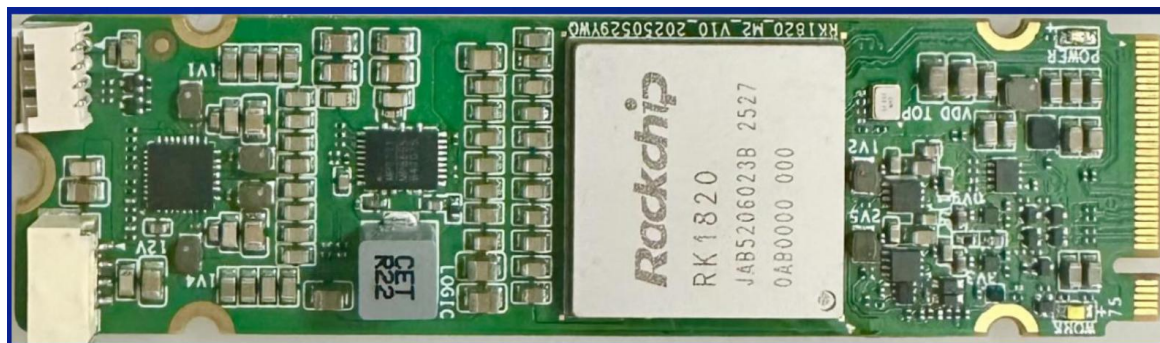
典型应用 – 陪伴机器人

- 对话（LLM）和场景分析（VLM）对于陪伴机器人是刚需，也是 RK182X 性能优势点
- 多种组合方式：VLM 运行在端侧隐私保护；LLM端云结合，兼顾效果和场景
- 情绪分析（表情，语义，语调）+长短期记忆+对话内容+环境感知，输出对应的声音/ 动作/ 表情



》》 RK182X 样品：模组

- NPU 加速卡，需在主控运行 runtime service
 - 无视频解码，需主控做前处理和后处理
- SO-DIMM 接口
- M.2 接口



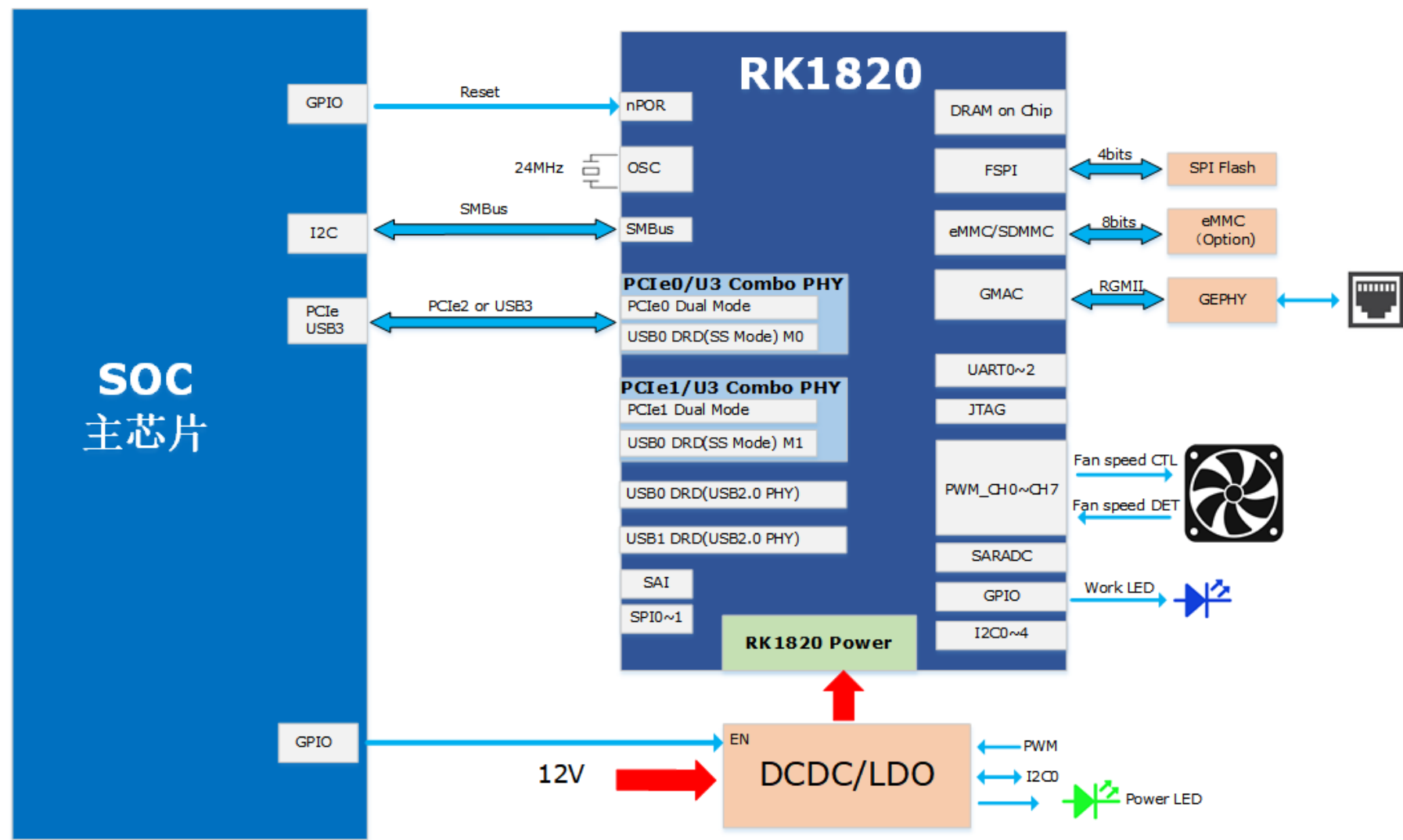
>> RK182X 模组性能及功耗数据

Model Name	Input Tokens	RK1828 NPU@1GHz SO-DIMM模组默认模式		RK1828 NPU@850MHz M.2模组默认模式	
		TTFT (ms)	TPS(tokens/s)	TTFT (ms)	TPS (tokens/s)
Qwen2.5 3B	128	84	102.8	91.7	97.9
Power Consumption		Avg 11.4W Peak 33.5W		Avg 10.1W Peak 26.7W	
Qwen2.5 7B	128	158.2	70	175.1	67.2
Power Consumption		Avg 15.3W Peak 36.4W		Avg 13.6W Peak 29.4W	

M.2模组 vs SO-DIMM模组: 80%功耗, 90% Prifill性能, 95% Decode性能

注1: 数据基于RKNN3 SDK V0.5.0
注2: 模组带风扇, 环境温度25°下测试的RK182x芯片功耗

>> RK182X 硬件应用框图



>> RK182X 开发套件

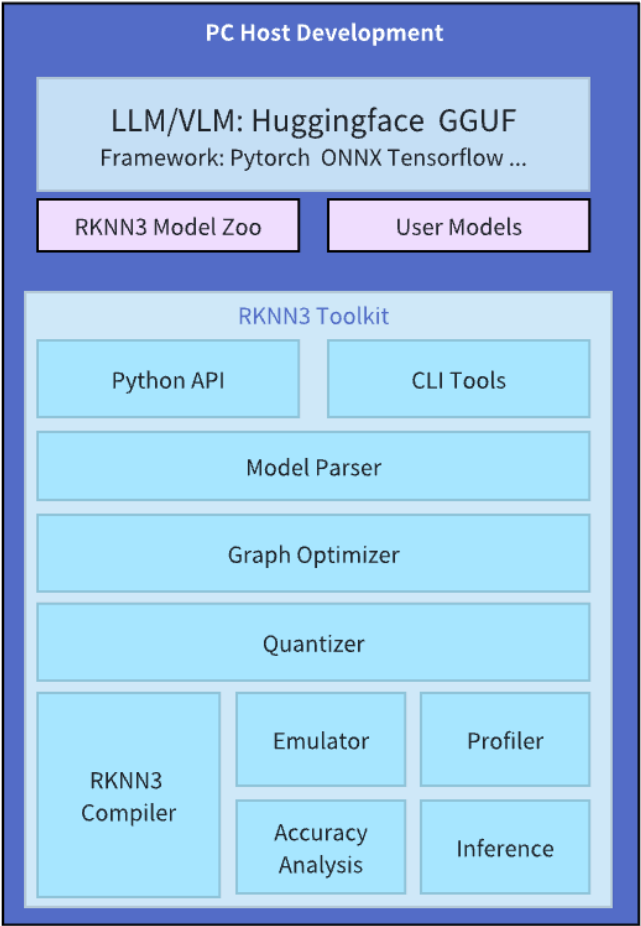
- 支持接多个SO-DIMM模组
 - RK1820 + RK3588 EVB10
 - RK1828 + RK3588 EVB10



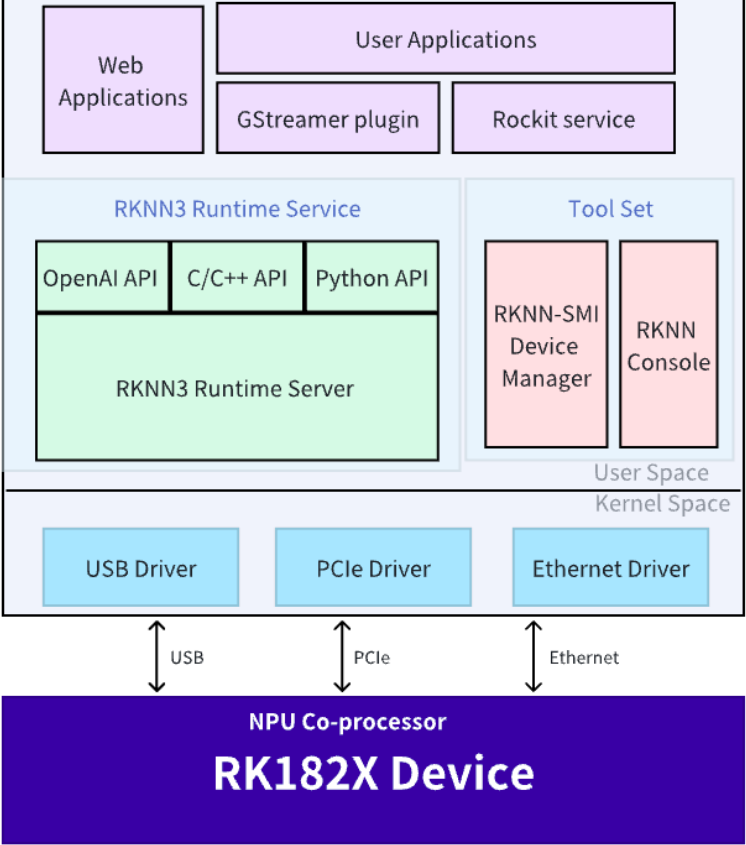
>> RK182X 系统应用框图

PC端

模型转换
模型仿真
模型调试



Application Processor(RK3588, RK3576 ...)



主控端

设备管理
Runtime服务

- Tokenizer
- 词表(约几百MB)
- 应用接口

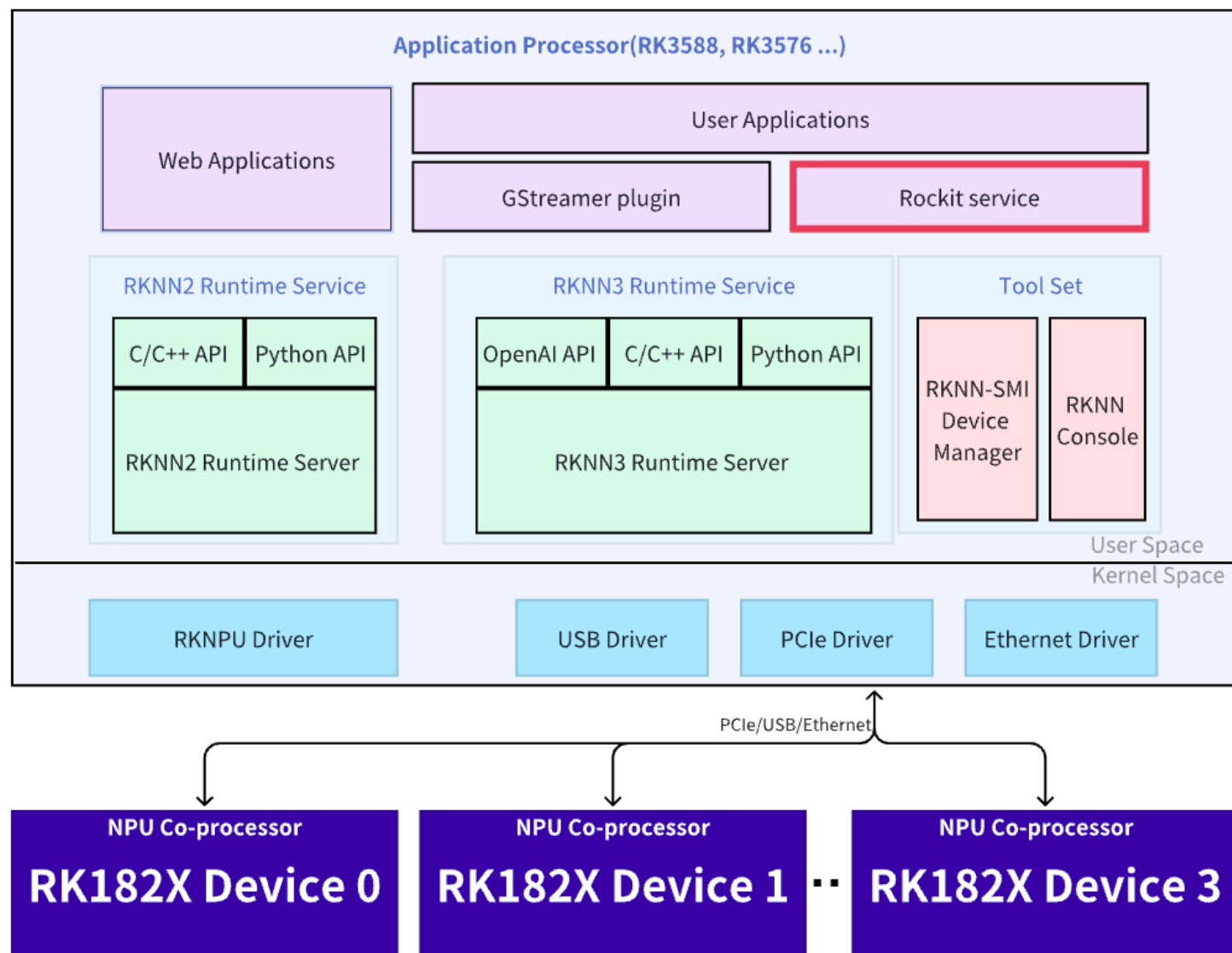
设备端

模型推理

不影响主控业务，主控CPU/GPU/NPU/DRAM带宽等资源均不受影响

基于主控+协处理器的应用框图

- Rockit框架算力融合
 - 可调度RKNN2和RKNN3
 - 应用自主拆分任务
- RK182X协处理器
 - 可挂载一个或多个
 - 每个NPU运行独立模型
- 主流模型框架和格式支持
 - Huggingface, Pytorch...
 - GGUF, ONNX...
- 通用API接口
 - OpenAI API
 - C/C++ API
 - Python API



Thank You!

Contact Us

Building No.18, A District,
Fuzhou Software Park,
89 Soft Avenue, Tongpan Road,
Gulou District, Fuzhou, Fujian, China
P.C: 350003

TEL: 86-591-83991906 FAX: 86-591-83951833

Rockchip

瑞芯微电子

SH603893