Rockchip
瑞芯微电子

RK182X

让设备真正“听懂、看懂、思考”

—— AIoT 2.0 重塑智能硬件

智能设备的重塑

IoT

万物智联

连接、控制

典型芯片

RK3188, RK3288

数据处理

无

2014

首推搭载Android进入IoT市场

AIoT 1.0

深度学习CNN

目标识别、分类、检测

典型芯片

RK3568, RV1106, RK3588

数据处理

按规则识别 “人车非”

2018

首推内置NPU芯片

AIoT 2.0

大模型

听懂、看懂、思考

典型芯片

RK1820, RK1828

数据处理

泛化、理解、规划、决策

2025

首推端侧大模型协处理器芯片

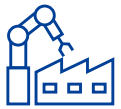
Rockchip
瑞芯微电子

大模型从云到端

边端侧模型能力及产品需求

Agent能力

融合感知理解、自主规划、工具调用、多步执行
本地知识库(RAG)、习惯、记忆

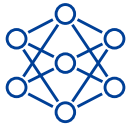


大模型泛化能力

自主分析异常，解决各类边界问题；
视频分析、工业检测、语音处理等应用

小参数大模型类型广泛

ASR、TTS、翻译、视觉编码、OCR、多模态、3D深度估计等各类模型层出不穷



小参数大模型能力迅速变强

Qwen 3.5-4B > Qwen3-VL-8B > Qwen2.5-VL-14B

边端侧部署的必要性

隐私安全 (Privacy)

敏感数据 (视频、音频) “不出端侧”，建立用户信任。

实时性 (Latency)

毫秒级响应，满足工业控制、自然交互等场景需求。

Token带宽成本 (Bandwidth Cost)

端云结合产品，仅传输端侧过滤后需深度思考数据；
全端侧产品，无持续云端Token需求

可靠性 (Reliability)

无网弱网可用，不依赖网络，稳定可靠

Rockchip
瑞芯微电子

大模型端侧部署的挑战

大模型性能需求挑战

- 大模型运行的超大带宽需求
- Transformer架构/Attention机制支持

端侧大模型运行效果和生态挑战

- 模型精度挑战 -> 量化方案支持
- 模型生态挑战 -> 主流模型支持和厂商合作

端侧设备的功耗挑战

- AI算力功耗 -> 高算力带来的功耗
- DRAM功耗 -> 超大访存带来功耗

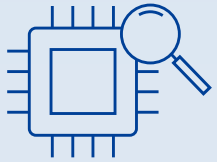
端侧设备可商业化挑战

- 商业化性价比 -- 晶圆面积限制 ->芯片性能
- 商业化性价比 -- 外挂DDR容量及成本

Transformer时代的“内存墙”危机

算力特性转变：从计算密集到内存密集

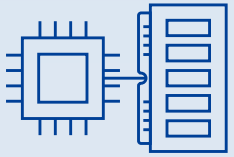
CNN (AIoT 1.0)



计算密集型 (Compute-Intensive)

算力要求高
带宽要求相对较低

Transformer (大模型)



内存密集型 (Memory-Intensive)

Decode 阶段逐个生成 Token，每次需加载全部权重 (算术强度低) 。

算术强度的困境与带宽需求

(以 7B 模型 INT4 量化为例)

速度	需求带宽
人类阅读 (约 20 Tokens/s)	~70 GB/s
自然对话 (约 50 Tokens/s)	~175 GB/s
极致响应 (约 100 Tokens/s)	~350 GB/s

传统SoC (如RK3588)
理论带宽峰值:
100 GB/S (44GB)

传统 SoC 的物理极限与能耗挑战

物理瓶颈与性能受限 (带宽受限)

推理速度慢 (~10 Tokens/s)，体验卡顿，“智能” 变 “智障” 。

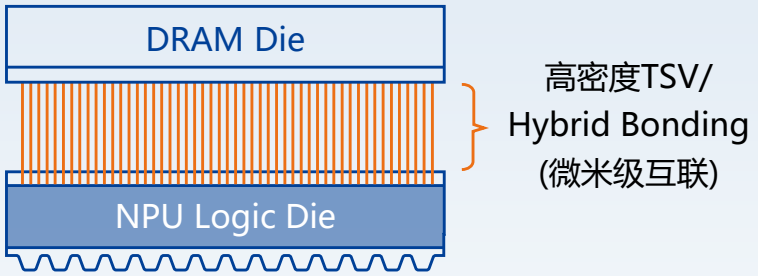
能耗噩梦 (数据搬运代价)

片外数据搬运能耗远高于计算，导致设备发热、降频。

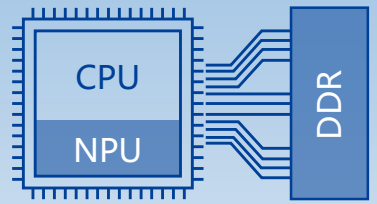
Rockchip 瑞芯微电子

业内首颗3D堆叠端侧AI芯片 — 突破DRAM带宽瓶颈

面对“内存墙” 和 “能耗墙”，RK182X 采用革命性 3D 堆叠技术，将 DRAM 晶圆直接堆叠在 NPU 逻辑晶圆之上，带来物理层面的降维打击。



传统封装



(PCB走线长、引脚有限)

带宽数量级提升

- 建立**数万**个**微米级垂直数据通道。**
- RK182X 等效带宽可达数百**GB**/s (相比传统 LPDDR4/5 的几十GB/s)
- 彻底消除 3B + 大模型带宽瓶颈。

数百GB vs 数十GB

极致的推理性能

实现远高于普通 SoC 的推理速度
跑3B模型推理输出超100TPS

内置DRAM

- 无需外挂DDR
- 布板简单面积小

突破性的能效比

- 传输距离极短，访存功耗低
  单位比特能耗 (pJ/bit) **量级降低**
- 同样电池下运行更久

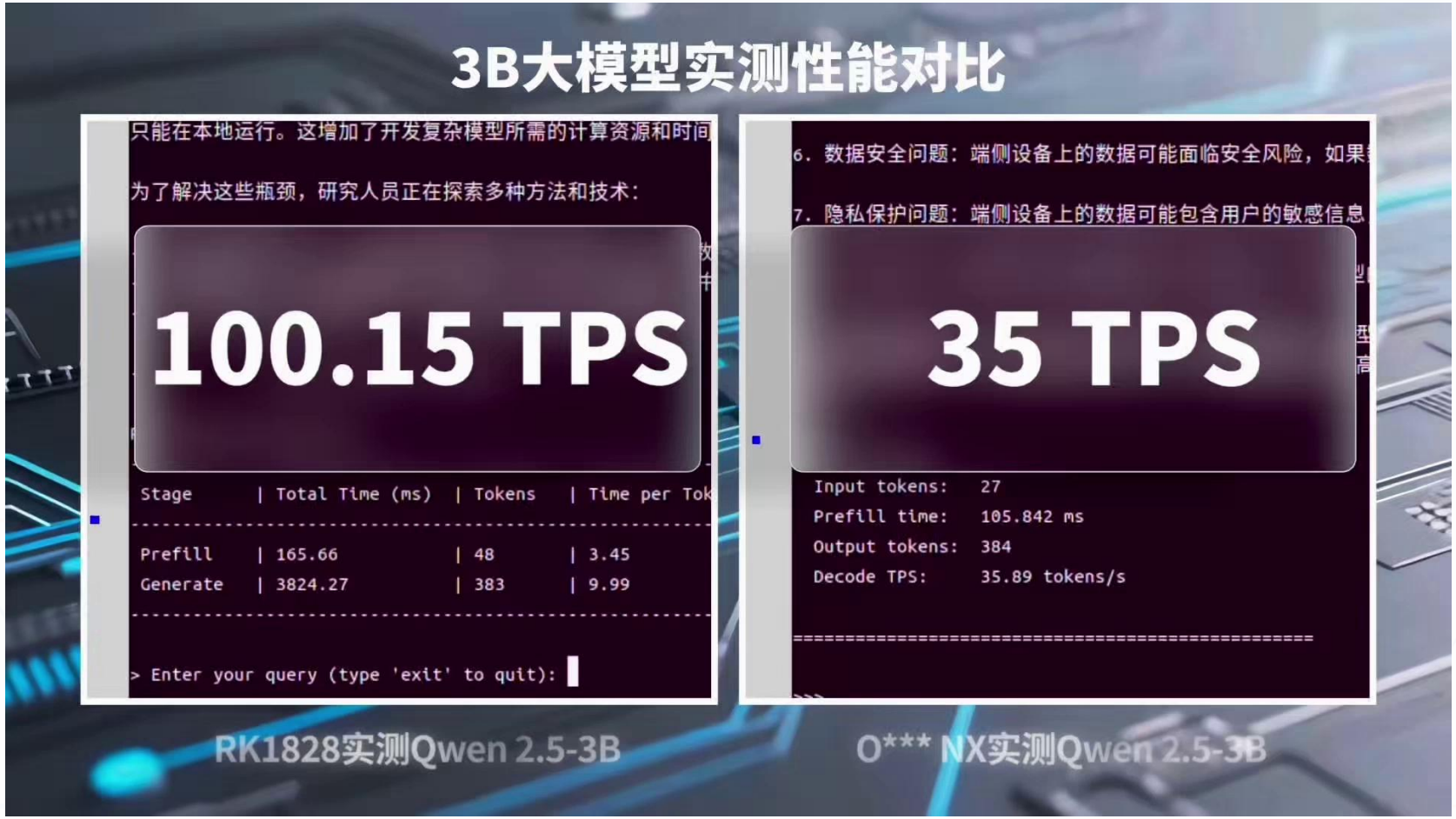
Rockchip 瑞芯微电子

RK182X：专为端侧大模型设计的AI推理芯片

	优势项	详细内容	收益
3D DRAM	低功耗	访存功耗比SoC低一个数量级	平均能耗比远高于竞品
	DRAM内置	内置2.5GB、5GB 3D DRAM	无需外挂DDR
	超高带宽	实测达到几百GB (端侧唯一)	高性能， 3B模型推理输出100+TPS
NPU设计	架构/算子加速	硬件支持Transformer及Attention机制	
	量化支持低比特	支持W4A16量化，内存占用少	5GB版支持8B大模型
	模型精度提升	支持FP16激活参数，Group量化	模型效果好
平台方案	主+协 算力解耦	真多核，大小模型同时运行 保障实时任务(ViT/LLM任务时间长)	不影响主控已有业务 独立选配升级
AI生态	模型适配全面	支持CNN/ViT/LLM/VLM/Omni/语音/ 嵌入/扩散等各类开源、商用模型	与模型厂商直接合作： 千问、智谱、面壁、阶跃、讯飞...

Rockchip
瑞芯微电子

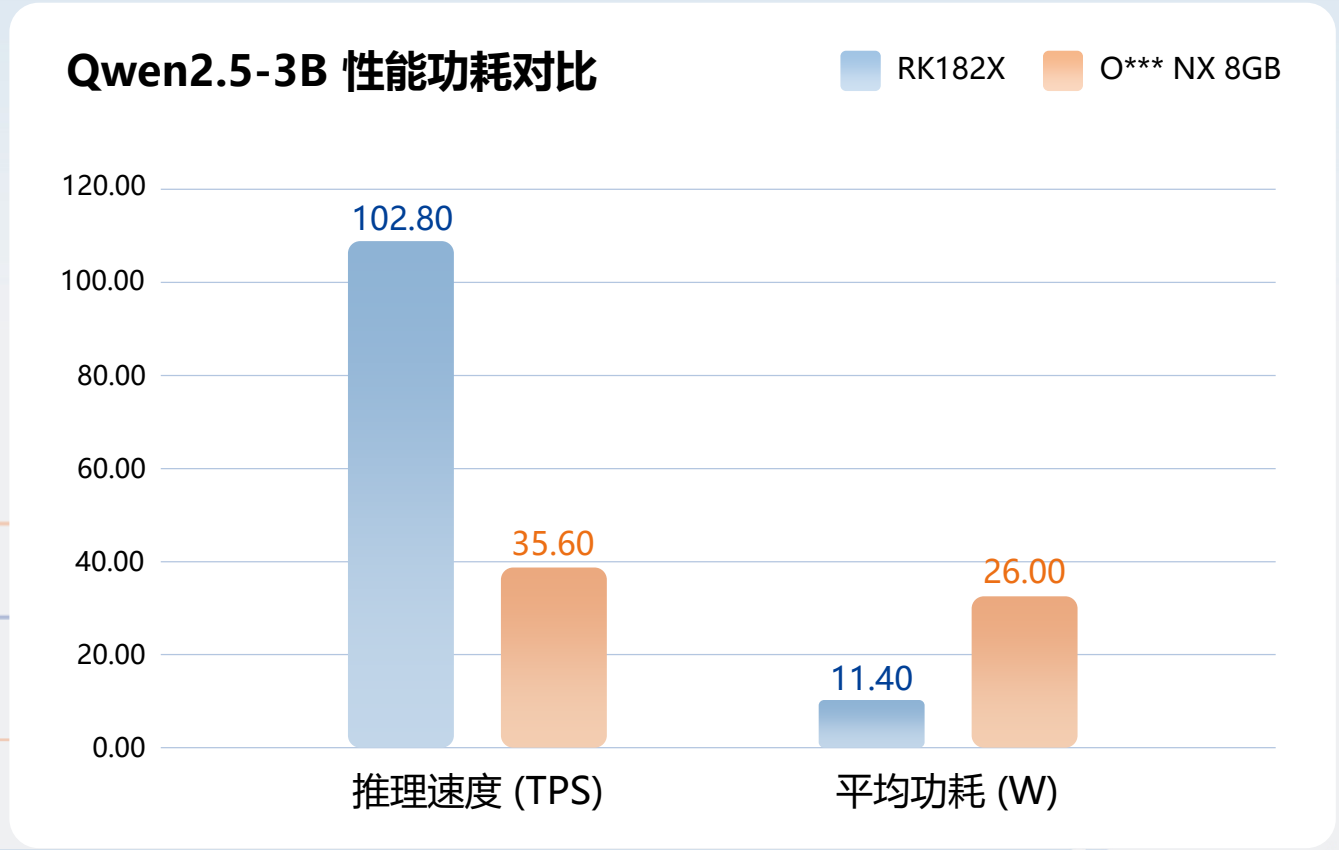
LLM 性能实测对比 – 体验升级，3B大模型输出破百！

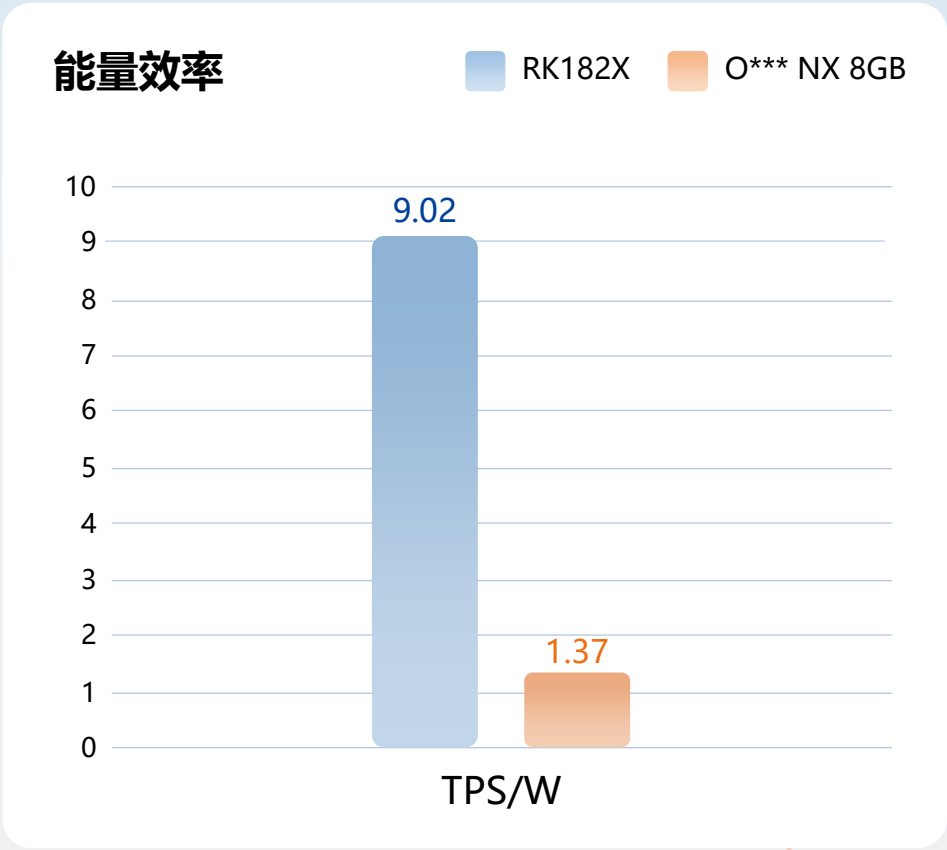


Rockchip 瑞芯微电子

能耗对比 — RK182X vs O**n NX 8GB

三倍性能，六倍能耗比





Rockchip 瑞芯微电子

瑞芯微 AIoT 2.0 产品生态

音频拾音算法

回声消除、噪声抑制、自动增益

音频处理算法

ASR、TTS、翻译、声纹、多音轨

视觉算法(CNN)

目标检测、分类、
图像分割、关键点检测

算法伙伴

科大讯飞 iFLYTEK　思必驰 AISPEECH　Ultralytics yolo　千问 Qwen　智谱 GLM Edge　面壁 MiniCPM　阶跃星辰 STEP Edge

从 “听清” 到 “听懂”

AIoT人机交互(机器感强)

- 精准语音识别
  (不理解潜台词)
- 关键词唤醒
  (每次对话都“失忆”)
- 简单指令执行
  (固定操作)
- 基础问答
  (标准回复)

期待的人机交互

- 怎么说都懂，理解意图
- 连续对话，像朋友一样聊天
- 主动建议，自动完成复杂任务链
- 有温度，个性化的回应

Rockchip
瑞芯微电子

从“听清” 到“听懂”— AIoT 2.0 给语音交互装上大脑

连续情境对话

多轮对话，上下文记忆

深度语义理解

结合常识、消除歧义

复杂任务执行

自主规划、调用服务

多模态融合感知

情绪分析、
真 “察言观色”

个性化与人格化

长期记忆、学习
用户习惯

Rockchip
瑞芯微电子

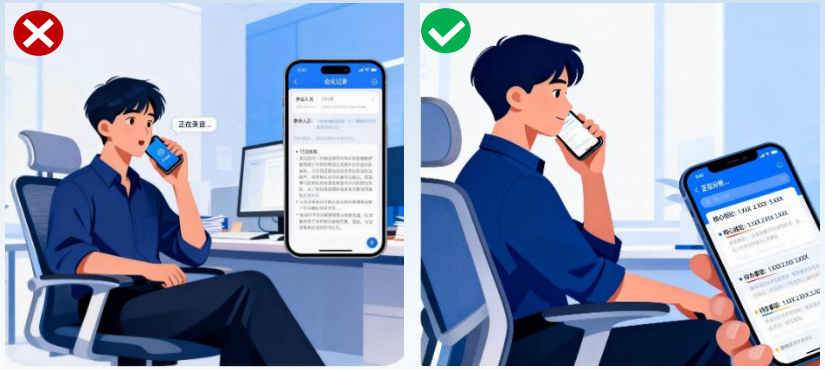
体验升级 — 智能助手

智能家居：从“遥控器”到“私人管家”

自动管理设备，了解家庭成员习惯，老幼报告生成

智能办公：从“录音笔”到“专属秘书”

自动完成：翻译、总结、待办提醒



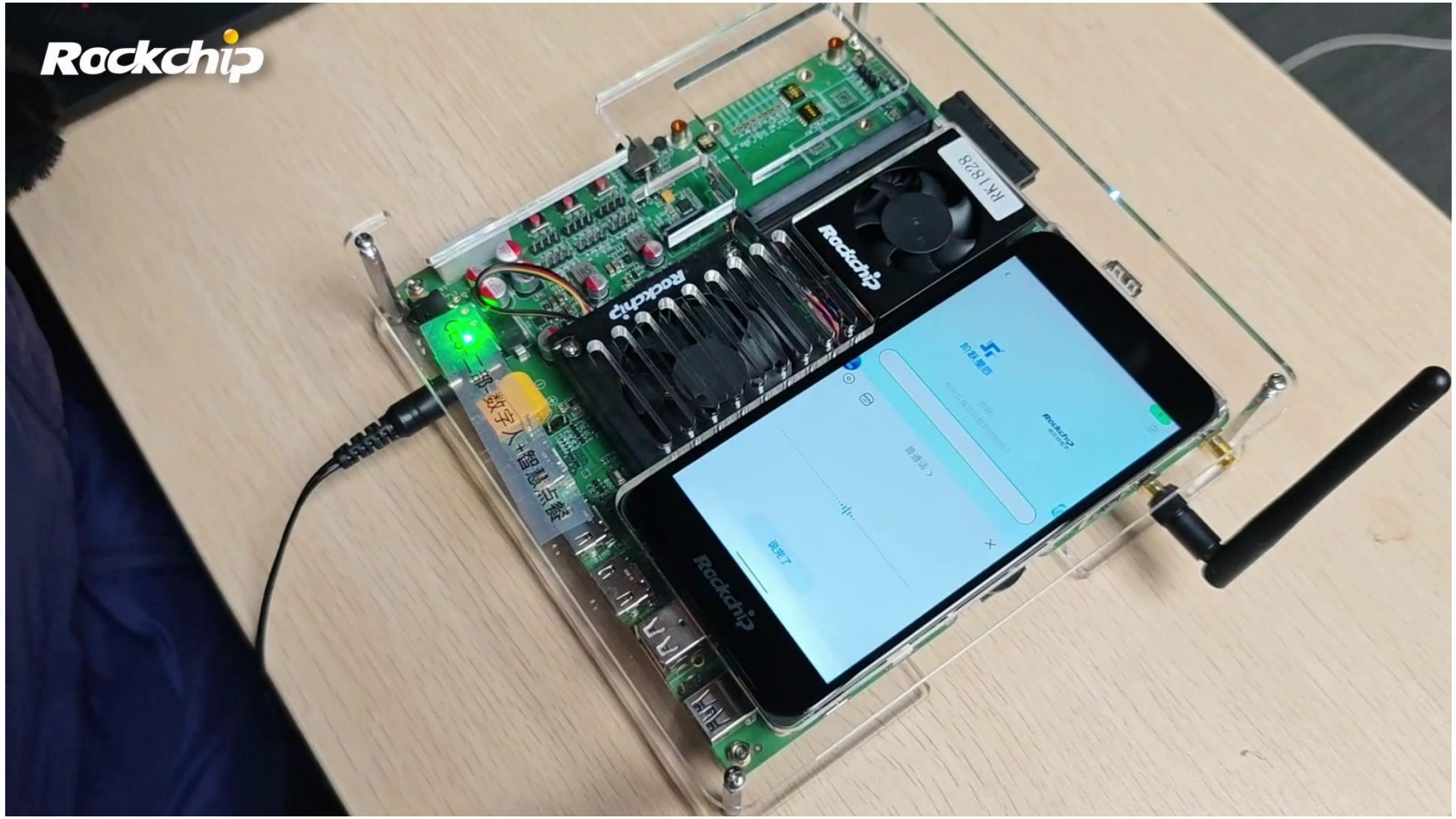
智能车载：从“导航仪”到“副驾伙伴”

成员识别，自然交互陪伴，基于时间地点自主规划任务

智能机器人：从“遥控”到“自主助理”

场景识别情绪分析，基于记忆个性化交互，自主完成任务

体验升级 — GUI Agent Demo



从“看清”到“看懂” — AIoT视觉

AIoT视觉优势

- **看得清：**高清画质，超低照度
- **认得准：**精准识别“人、车、物”等目标
- **反应快：**基于固定规则触发告警

AIoT视觉局限性

- **没有环境、行为、事件分析能力**
- **容易误报：**规则死板，环境干扰
- **检索效率低：**依赖关键词

Rockchip 瑞芯微电子

从“看清”到“看懂”— AIoT 2.0 给视觉装上大脑

场景理解
能理解环境

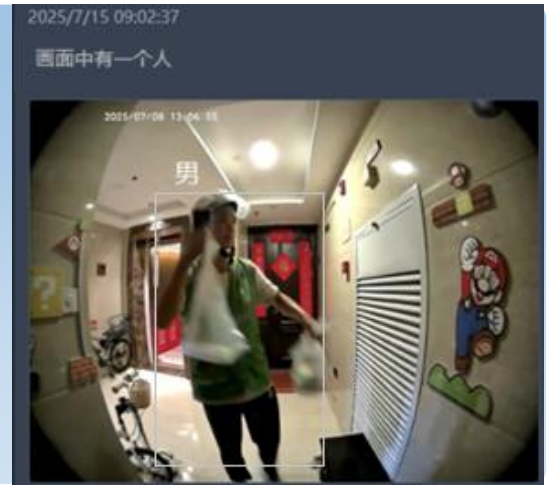
意图分析
能分析行为背后的意图

逻辑推理
能关联时间地点

自然交互
能用人类的语言问答总结

传统 AIoT

- 能认到有人出现



多模态大模型

- 场景和行为分析
- 异常行为预警
- 高性能
  2S内发出预警

Rockchip
瑞芯微电子

AIoT 2.0 视频分析 — 从被动记录到主动预警、分析总结

智慧家庭
老幼看护预警
日报周报生成
自动时光缩影

智慧厂区
规范操作预警
劳保用品穿戴

智慧交通
事故自动发现
拥堵原因分析

智慧零售
门店热力图
客流排队分析
员工动线分析

Rockchip 瑞芯微电子

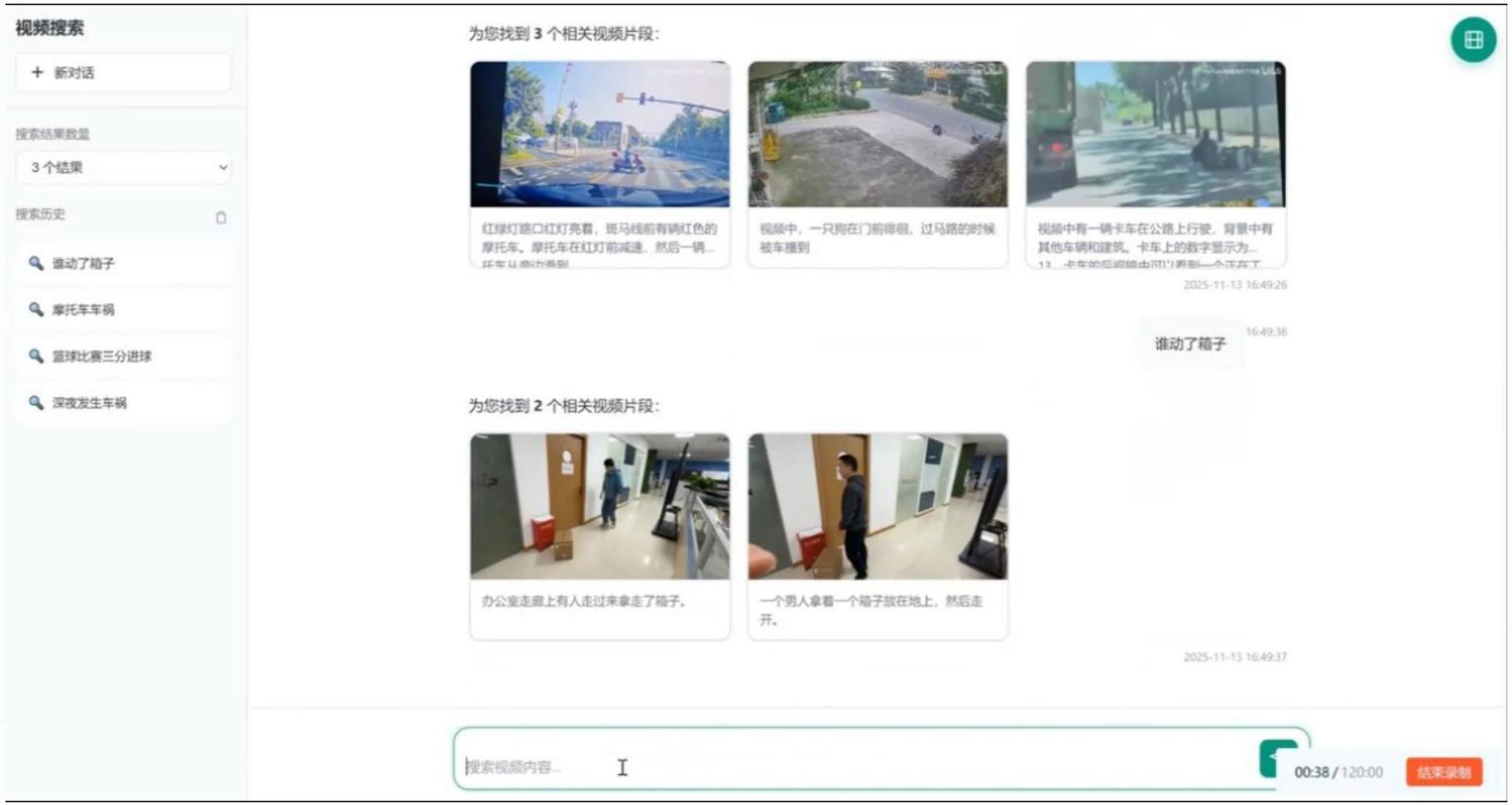
体验升级 — 事件检索“一键即达”

视频搜索

- 自然语言输入
- 秒级快速反应
- 精准匹配视频

找失物

“谁把箱子搬走了？”



Rockchip 瑞芯微电子

RK182X 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

泛安防视频分析
结构化搜索

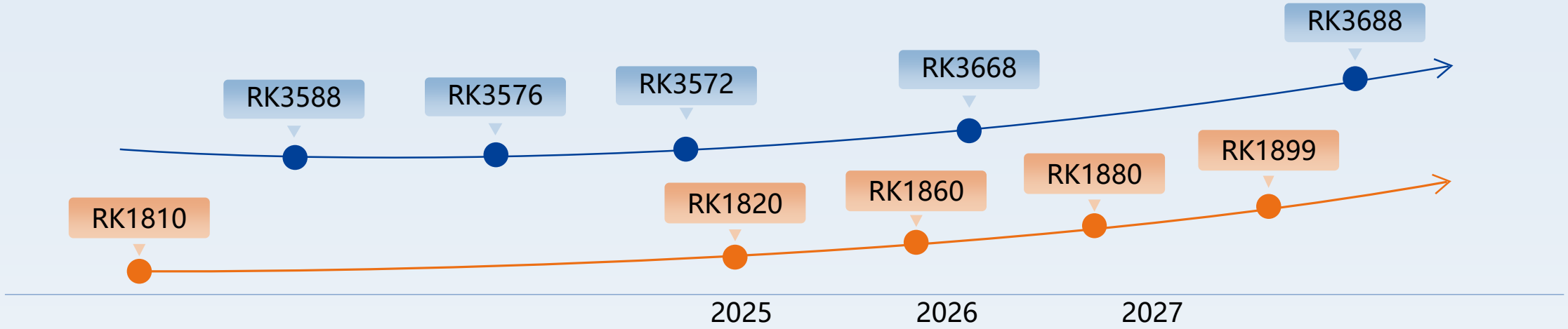
边缘AI计算

工业缺陷检测

Rockchip
瑞芯微电子

瑞芯微 AIoT 2.0 新战略 — 最佳端侧AI芯片方案

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

- (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

Rockchip 瑞芯微电子

主控+协处理器：解耦算力与控制

传统硬件迭代痛点：升级主控(AP)是 “痛苦” 的

研发成本高
重新设计 PCB、调试 BSP、重新认证

生命周期错配
设备周期长 (5 - 10 年) vs AI 迭代快 (几个月)

RK182X “协处理器” 商业模式：完美解决痛点

主控 (Host)

负责：OS、外设控制、网络连接
例如：RK3568 / MCU

USB/PCIe

协处理器 (CP)

负责：核心 “思考”、大模型推理
例如：RK182X (M.2模组)

战略优势

无缝升级

不改动原有主板，通过 USB/PCIe 外挂 M.2 模组，瞬间赋能 AIoT 2.0。

成本优势

专为端侧设计，内嵌 DRAM 内存；无需外挂 DRAM，BOM 成本低。

算力解耦

可选配不同算力的模组；
下一代AI模型出现时，仅升级协处理器模块，无需整机重设计。

Rockchip 瑞芯微电子

RK1828 典型 LLM/VLM 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 2.5-3B	128	128	83.09	100.75
Qwen 2.5-7B	128	128	158.28	69.37
Qwen 3-4B	128	128	106.58	86.87

模型	图像分辨率	ViT(ms)	LLM TTFT(ms)	输出速度 TPS
FastVLM_1.5B	512*512	144	48.19	148.47
InternVL3-2B	448*448	190.80	47.93	148.26
Qwen2.5-VL-7B	392*392	279.34	159.40	70.02
Qwen3-VL-4B	384*384	158.89	108.29	89.69

注：数据基于RKNN3 SDK V1.0.0
所有模型使用 w4a16 量化，VL 的 LLM 部分 input 为128

Rockchip 瑞芯微电子

RK182X 典型 ViT/CNN 性能

模型	图像分辨率	单核帧率	多核帧率
YOLOv5s	640x640	35.4	212.6
ResNet50V2	224x224	113.7	851.3
ViT_b_p16	224x224	48.9	/
DINOv3	224x224	54.8	331.0
DepthAnythingV2	448x252	/	27.3

模型精度保持

模型类别	模型名称	数据集	考察重点	原始精度	RKNN3精度	数据类型
LLM	Qwen3-4B	GSM8k	数学能力	90.6	89.84	W4A16
LLM	Qwen3-4B-Instruct	IFEval(strict)	指令遵循	88.73	**90.42**	
LLM	Qwen3-4B-Instruct	BFCL(avg)	函数调用	84.94	**85.34**	
VLM	InternVL3.5-4B	MMBench(cn)	图片理解	78.69	77.41	
CNN	YOLOv8s	Coco2017	目标检测	0.39(AP@.5:.95)	0.380	W8A8
CNN	ResNet50V2	Imagenet	图片分类	0.729(Top1)	0.721	

RK182X全面的模型支持

大语言模型 (LLM)

文本理解、生成、
机器翻译

- Qwen 2.5 0.5B / 1.5B / 3B / 7B
- Qwen 3 0.6B / 1.7B / 4B / 8B
- GLM Edge
- Hunyuan-MT1.5
- Youtu-LLM

语音模型

ASR、TTS

- SenseVoice
- Whisper

多模态/视觉语言模型 (VLM)

图像理解、视频分析
OCR、视觉问答

- Qwen2.5VL 3B / 7B
- Qwen3VL 2B / 4B
- GELab-Zero
- InternVL3 2B/4B
- MiniCPM-V-4
- FastVLM 1.6B
- MiMo-VL-7B-RL
- Janus-Pro
- Qwen2.5-Omni

视觉类ViT/CNN

图像分类、分割、
目标检测、深度估计

- SigLIP 1/2
- DINOv2/v3
- EVA02
- CLIP ViT-B/L

CNN类

- YoloV5/V6/V8/ World/26
- MobileNet
- ResNet

深度估计 (Depth Estimation)

- Depth-Anything-V2-small

向量嵌入与重排序 (Embedding & Reranker)

RAG (检索增强生成) 核心组件

Embedding (嵌入)

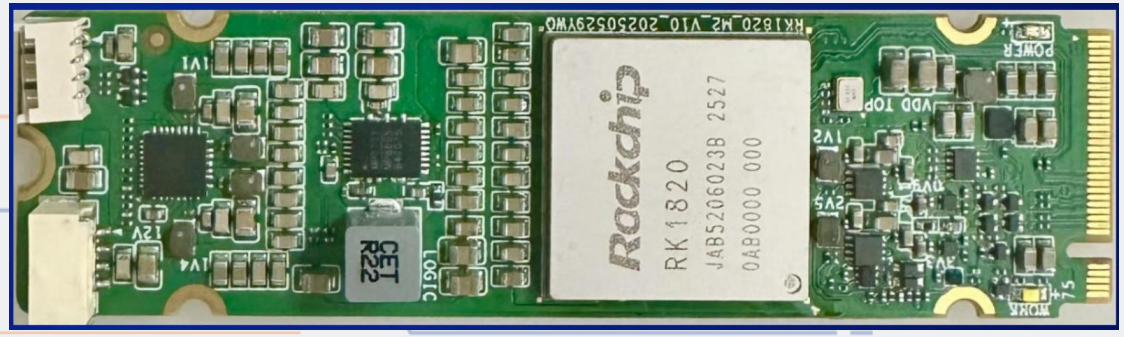
- M3E small
- Albert-base-v2

Reranker (重排序)

- Qwen3-0.6B-Reranker
- Qwen3-4B-Reranker

RK182X 模组

- NPU 加速卡，搭配主控使用，无需外挂DDR
- SO-DIMM 接口 (PCIe2x1，USB3.0)，MP
- M.2 2280 接口 (PCIe2x1)，ES



Rockchip 瑞芯微电子

RKNN3 SDK组成及开发

RKNN3 SDK开发路径

软硬件验证(跑ModelZoo模型) -> 模型转换 -> 精度仿真 -> 板端验证 -> 部署优化

RKNN3 Toolkit (PC Tool)

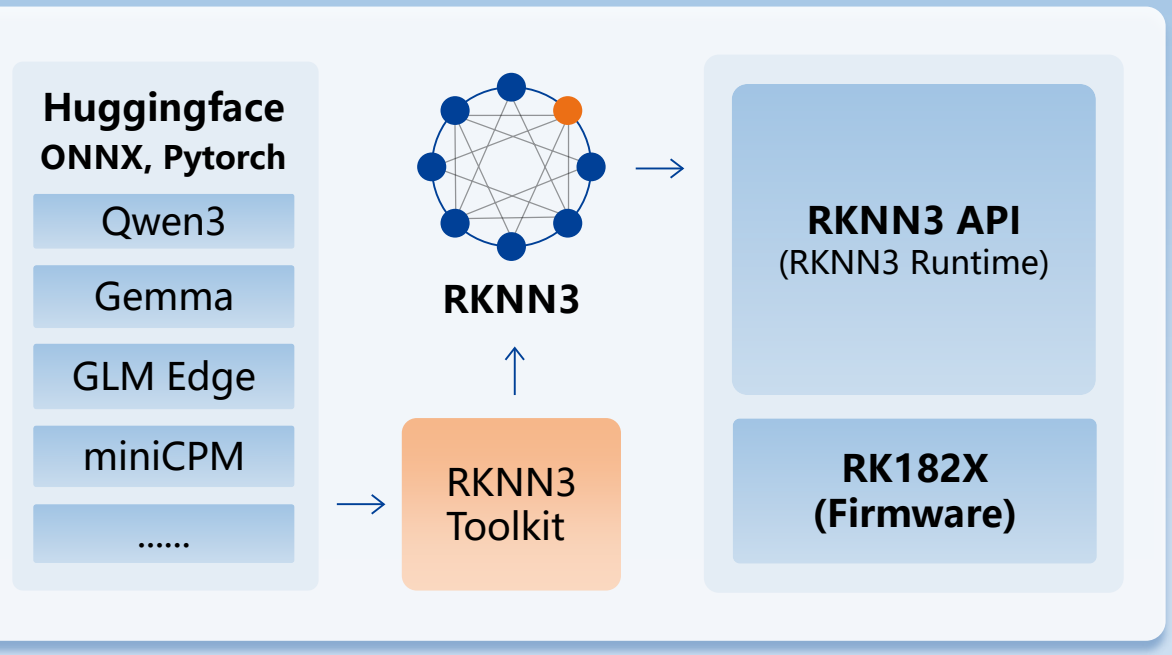
模型量化，模型转换，精度分析，模拟器

RKNN3 Runtime

用户层API库(Python, C/C++, OpenAI接口)

Model Zoo

模型仓库：CNN/ViT/LLM/VLM...



Rockchip 瑞芯微电子

感谢观看

地址　中国福建福州鼓楼区铜盘路软件园A区18号楼

邮箱　service@rock-chips.com

邮编　350003

电话　86-591-83991906

传真　86-591-83951833

微信公众号

微信视频号

微博号