

科技前瞻系列专题

科技公司的端侧AI布局

西南证券研究发展中心
海外研究团队
2024年6月

目录

◆ 1 苹果AI布局

◆ 2 高通AI布局

◆ 3 三星AI布局

◆ 4 相关标的

◆ 5 风险提示

1.1.1 苹果AI：模型侧：MM1

- 3月苹果正式公布自家的多模态大模型MM1，MM1是一个具有最高300亿参数的多模态LLM系列。**MM1由密集模型和混合专家 (MoE) 变体组成，具有300亿、70亿、30亿参数等多个版本。**MM1不仅在预训练指标中实现最优性能SOTA，在一系列已有多模态基准上监督微调后也能保持有竞争力的性能。
- 得益于大规模多模态预训练，MM1模型在上下文预测、多图像和思维链推理等方面表现不错。MM1模型在指令调优后展现出强大的少样本学习能力。在几乎所有基准测试中，苹果的MoE模型比密集模型取得更好的性能，显示MoE进一步扩展的巨大潜力。

(a) User:  { "smartphone": 1, "teddy bear": 1 } { "cat": 3 } { "book": 3, "vase": 1, "glass": 1 } { "dog": 2, "frisbee": 1 }

MM1-30B (Ours):

(b) User:  Red circle: "no parking anytime" Red circle: "Raffaello" Red circle: "Rue Saint-Paul" Red circle: "Hyde Park"

MM1-30B (Ours):

"Hyde Park"

(c) User:  furniture: bed frame, weight: 50 and 150 pounds (23 to 68 kg) furniture: sofa, weight: 100 to 200 pounds (45 to 91 kg) furniture: stove, weight: 150 to 300 pounds (68 to 136 kg) furniture: refrigerator, weight: 200 to 300 pounds (91 to 136 kg)

MM1-30B (Ours):

(d) User:  total: 1 + 3 = 4 total: 6 + 4 = 10 total: 4 + 1 = 5

MM1-30B (Ours):

由于其大规模多模态预训练，MM1可进行上下文预测。这使得MM1能够：a) 计数对象并遵循自定义格式；b) 引用图像的部分并执行OCR；c) 展示关于日常物品的常识和词汇知识；d) 执行基本的数学函数。

User: Photo:  Menu:  How much should I pay for all the beer on the table according to the price on the menu?

MM1-30B-Chat (Ours): 12

Emu-Chat-37B: According to the menu, the price for all the beer on the table is **15.99**.

LLaVA-NeXT-34B: The image shows a menu with beer prices, but it's not entirely clear which specific beers are on the table. However, I can provide you with an estimate based on the visible prices. From the menu, it appears that the prices for beer are as follows: Magna: 5, Medalla: 5, Heineken: 6 Presidente: 6 Assuming that the beers on the table are Magna, Medalla, Heineken, and Presidente, and there are two of each, you would pay: 2 x Magna: 2 x 5 = 10, 2 x Medalla: 2 x 5 = 10, 2 x Heineken: 2 x 6 = 12, 2 x Presidente: 2 x 6 = 12 Total: 10 + 10 + 12 + 12 = **44** Please note that this is an estimate based on the visible prices and the assumption that there are two of each type of beer. The actual cost may vary if the beers on the table are not the ones listed on the menu or if there are different quantities.

User: Explain why.

MM1-30B-Chat (Ours): There are 2 beers on the table, and each beer costs 6 according to the price on the menu. So, 2*6=12.

Emu-Chat-37B: The price for all the beer on the table is 15.99 because each beer has a price of 10.99, and there are a total of five beers on the table. When you multiply the number of beers (5) by the price per beer (10.99), you get a total cost of 54.95 (rounded up to 15.99).

LLaVA-NeXT-34B: The image shows a menu with beer prices, but it's not entirely clear which specific beers are on the table. However, I can provide you with an estimate based on the visible prices. From the menu, it appears that the prices for beer are as follows: (...)

MM1模型可以遵循指令进行跨图像推理。示例和图像来自VILA；当使用思维链提示时，VILA能够正确回答。

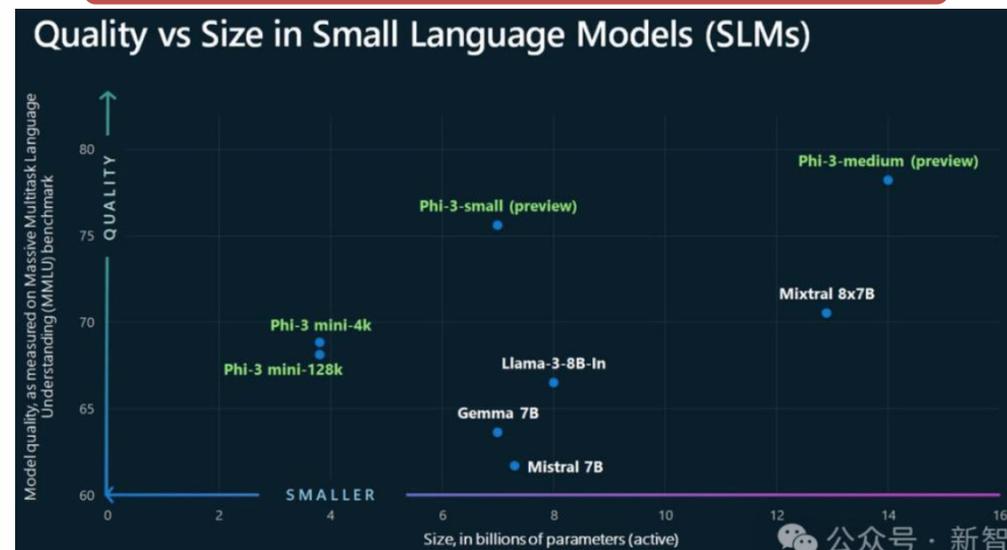
1.1.2 苹果AI：模型侧：OpenELM

- 4月26日，苹果宣布了更大的端侧AI推进，推出全新的开源大语言模型OpenELM。OpenELM包含2.7亿、4.5亿、11亿和30亿个参数的四种版本，定位于超小规模模型，运行成本更低，可在手机和笔记本电脑等设备上运行文本生成任务。同时，公司开源了OpenELM模型权重和推理代码、数据集、训练日志、神经网络库CoreNet。
- OpenELM使用了“分层缩放”策略，来有效分配Transformer模型每一层参数，从而提升准确率。在约10亿参数规模下，OpenELM与OLMo相比，准确率提高了2.36%，同时需要的预训练token数量减少了50%。
- OpenELM的基准测试（数据类型BFloat16）在工作站（配备英特尔i9-13900KF CPU、RTX 4090 GPU，24GB DDR5内存），以及M2 Max MacBook Pro（64GB内存）运行。

OpenELM与公开LLM对比

Model	Public dataset	Open		Model size	Pre-training tokens	Average acc. (in %)
		Code	Weights			
OPT [55]	✗	✓	✓	1.3 B	0.2 T	41.49
PyThia [5]	✓	✓	✓	1.4 B	0.3 T	41.83
MobiLlama [44]	✓	✓	✓	1.3 B	1.3 T	43.55
OLMo [17]	✓	✓	✓	1.2 B	3.0 T	43.57
OpenELM (Ours)	✓	✓	✓	1.1 B	1.5 T	45.93

常见的小模型性能对比



1.1.3 苹果AI : 模型侧 : Ferret-UI

- 4月8日，苹果展示了多模态模型Ferret-UI。Ferret-UI系统可以理解手机屏幕上的应用程序内容，专为增强对移动端UI屏幕的理解而定制，其配备了引用（referring）、定位（grounding）和推理（reasoning）功能。**Ferret-UI具备了解决现有大部分通用多模态大模型所缺乏的理解用户界面（UI）屏幕并与其有效交互的能力。**Ferret-UI与OpenAI近期发布的GPT-4o、谷歌近期发布的Gemini 1.5 Pro在功能上异曲同工。
- Ferret-UI可执行从基本到复杂任务等11种任务。它能在移动UI屏幕上使用灵活的输入格式（点、框、涂鸦）执行指代任务（例如控件分类、图标识别、OCR）和定位任务（例如查找控件、查找图标、查找文本、控件列表）。这些基本任务为模型提供了丰富的视觉和空间知识，使其能在粗略和精细级别（例如各种图标或文本元素之间）区分UI类型。这些基础知识对于执行更高级的任务至关重要。具体来说，Ferret-UI不仅能在详细描述和感知对话中讨论视觉元素，还可在交互对话中提出面向目标的动作，并通过功能推理推断出屏幕的整体功能。

Elementary Tasks
Referring
widget classification, icon recognition, OCR

- Classify the UI type of the widget [point0].
Button
- Identify the icon type of the widget [box1].
launch
- What text is displayed in the element [scribble0]?
Reminders
Don't forget. Use reminders

Grounding
find widget, find icon, find text, widget listing

- Where is the Open button?
Open button [box0]
- Where is the launch icon?
launch icon [box1]
- Where is the text "Reminders Don't forget. Use reminders"?
Reminders, Don't forget. Use reminders [box4]
- Please list all the UI elements on this screen.
UI widgets present in this screen include text displaying "Reminders, Don't forget. Use reminders" [box4], Open Button [box0], Launch icon [box1], ...

Advanced Tasks
Conversation (perception + interaction)

- How do I open the Reminders app?
You can open the Reminders app by tapping on the Open button [box0].
- I want to share the Reminders app with my friend, what should I do?
You can share the app by tapping on the launch icon [box1].
- How do I navigate to Games?
You can navigate to the Games tab by tapping on the Games tab [box2].
- Can my 15-year-old daughter use this app?
Yes, the app is suitable for all ages as indicated by the "4+" age rating [box3].

Detailed Description

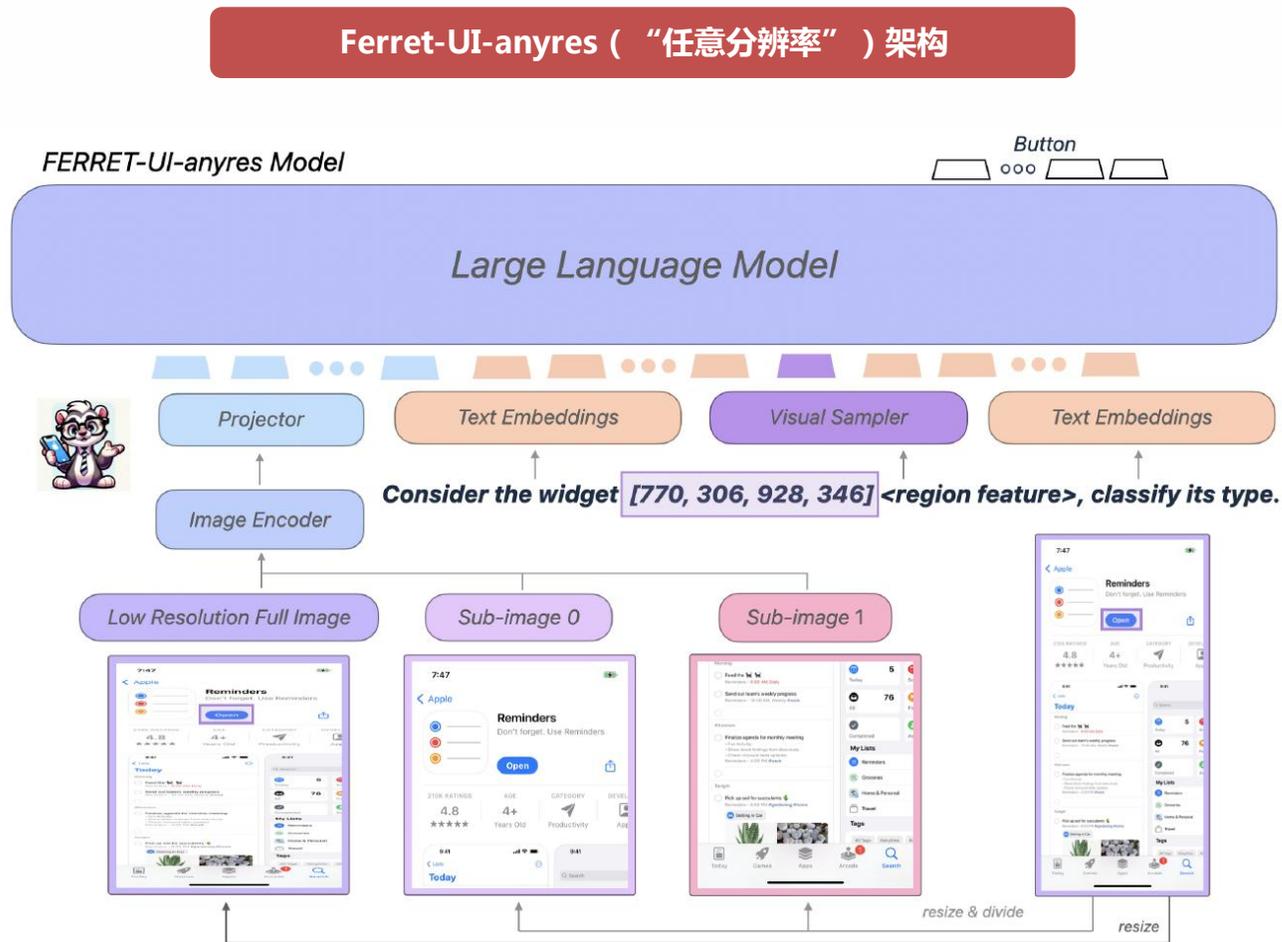
- Please describe this screen in detail.
The screen is for a reminders application. At the top left, there's a button labeled "Apple". The title of the screen, "Reminders", is located in the upper middle part of the screen. There's a text below the title that reads "Don't forget. Use reminders".
The screen is divided into several sections. The first section, labeled "210k+", is located just below the title. The second section, labeled "AGE", contains a text displaying "4+ Years Old" ...

Function Inference

- What's the high-level function of the screen?
The screen is the download page for a reminders application where you can set up reminders for various tasks such as sending out team's weekly progress, grocery shopping, traveling, and picking up kids.

1.1.3 苹果AI : 模型侧 : Ferret-UI

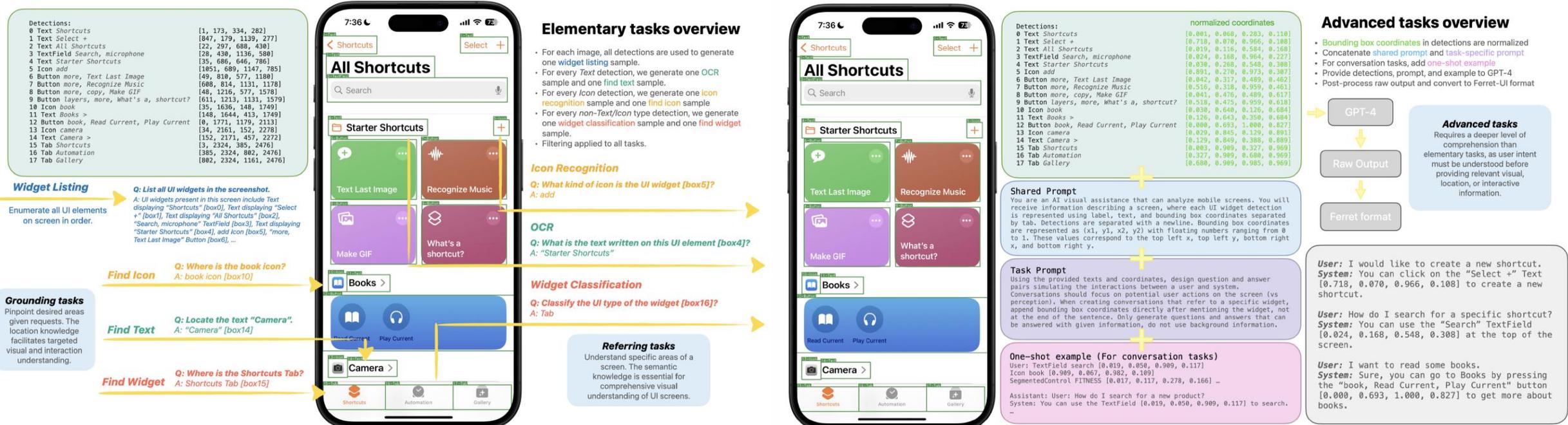
- Ferret-UI建立在Ferret的基础上。Ferret是一个MLLM，在不同形状和细节水平的自然图像中实现空间参考和基础。无论是点、框还是任何自由形式的形状，它都可以解释相应区域或对象并与之交互。Ferret包含一个预训练的视觉编码器（例如CLIP-ViT-L/14）和一个Decoder Only的语言模型（例如Vicuna）。此外，Ferret还采用了一种独特的混合表示技术，该技术将指定区域转换为适合LLM处理的格式。本质上，空间感知视觉采样器旨在熟练地管理不同稀疏度级别的区域形状的连续特征。
- 虽然Ferret-UI-base是基于Ferret的架构，但Ferret-UI-anyres包含额外的细粒度图像功能。特别是，预训练的图像编码器和投影层为整个屏幕生成图像特征。对于基于原始图像长宽比获得的每个子图像，都会生成额外的图像特征。对于具有区域参考的文本，视觉采样器会生成相应的区域连续要素。LLM使用全图像表示、子图像表示、区域特征和文本嵌入来生成响应。



1.1.3 苹果AI : 模型侧 : Ferret-UI

❑ Ferret-UI对基本任务的流程：UI检测器输出所有检测到的元素，以及每个元素的类型、文本和边界框。这些检测用于为基本任务创建训练样本。对于定位任务，使用所有元素检测来创建一个用于控件列表的样本，而其余任务一次专注于一个元素。将元素分为图标、文本和非图标/文本控件。对于每种类型，创建一个指代样本和一个定位样本。

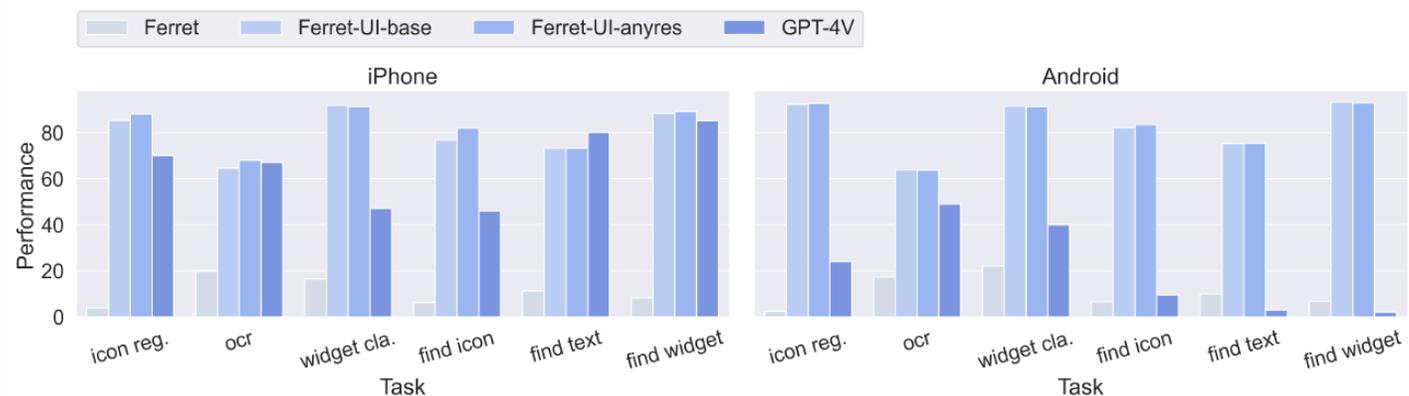
❑ Ferret-UI对复杂任务的流程：首先从检测输出中归一化边界框坐标，然后将检测、提示和可选的单个示例发送到 GPT-4。对于详细的描述和函数推理，将生成的响应与预先选择的提示配对，以训练Ferret-UI。对于对话任务，直接将 GPT-4输出转换为多回合对话。



1.1.3 苹果AI：模型侧：Ferret-UI

- Ferret-UI在简单任务处理上击败了GPT-4V。但在复杂任务处理上还是不如GPT-4V。
- 通过精心设计“任意分辨率”（anyres）以适应各种屏幕宽高比，以及策划包含广泛的基本和高级UI任务的训练样本，Ferret-UI在引用、定位和推理方面表现出显著的熟练程度。这些增强能力的引入预示着在众多下游UI应用中或将取得重大进步，从而扩大Ferret-UI在这一领域所能提供的潜在益处。

	Public Benchmark			Elementary Tasks				Advanced Tasks	
	S2W	WiC	TaP	Ref-i	Ref-A	Grd-i	Grd-A	iPhone	Android
Spotlight [30]	106.7	141.8	88.4	-	-	-	-	-	-
Ferret [53]	17.6	1.2	46.2	13.3	13.9	8.6	12.9	20.0	20.7
Ferret-UI-base	113.4	142.0	78.4	80.5	82.4	79.4	83.5	73.4	80.5
Ferret-UI-anyres	115.6	140.3	72.9	82.4	82.4	81.4	83.8	93.9	71.7
GPT-4V [1]	34.8	23.5	47.6	61.3	37.7	70.3	4.7	114.3	128.2



	iPhone					Android				
	DetDes	ConvP	ConvI	FuncIn	Avg	DetDes	ConvP	ConvI	FuncIn	Avg
Ferret [53]	2.5	34.7	23.7	19.1	20.0	2.0	33.9	24.9	21.9	20.7
Fuyu [6]	5.0	24.6	18.8	35.7	21.0	2.0	20.8	44.5	36.1	25.9
CogAgent [20]	53.1	59.7	74.8	71.9	64.9	28.0	58.5	90.1	90.5	66.8
Ferret-UI-base	64.5	75.0	77.5	76.5	73.4	90.8	72.8	79.3	79.2	80.5
Ferret-UI-anyres	97.4	92.1	91.1	95.2	93.9	86.4	70.3	50.2	77.3	70.1
GPT-4V [1]	66.8	105.6	198.5	86.3	114.3	126.6	109.4	188.6	88.3	128.2

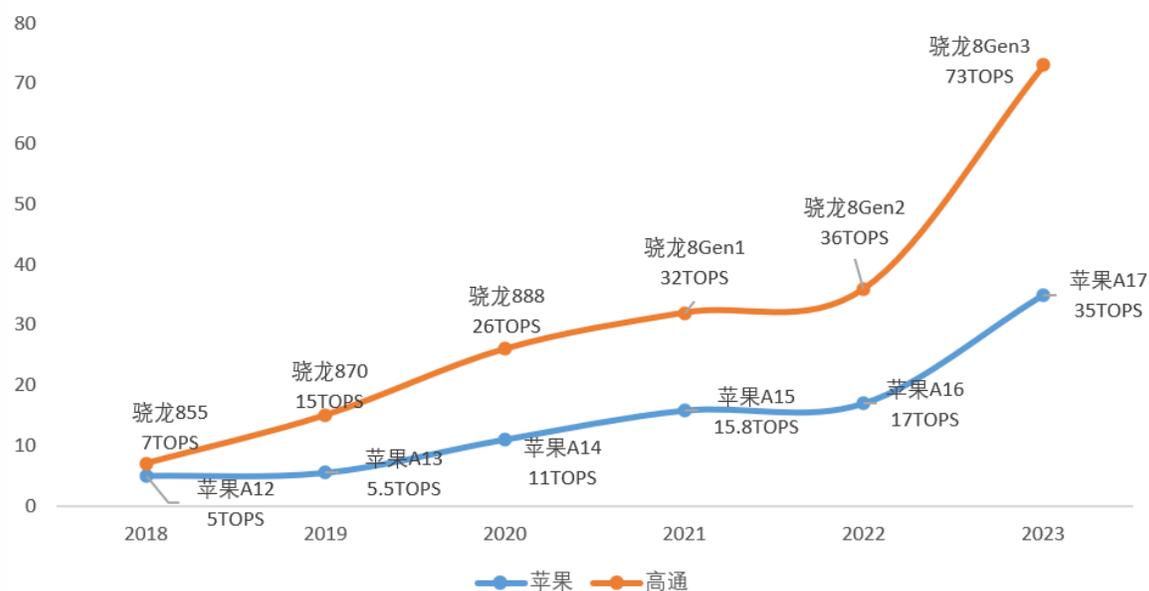
1.2.1 苹果AI：硬件侧：手机AP后劲不容小觑

- 苹果是较早介入AI计算单元的厂商，早在2017年就在SoC中引入了NPU计算单元，并搭载于iPhone 8、iPhone 8 Plus及iPhone X上。苹果A17芯片虽然算力为35TOPS，仅相当于高通骁龙8Gen2的AI性能（8Gen2可在手机侧运行10+亿参数的AI小模型），但手机上运行AI模型的效果还需看CPU/GPU/NPU的综合能力，软件生态上的适配优化也是提升用户体验的重要操作。
- 随着AI布局的深入，苹果将充分发挥在架构设计、iOS生态和软件优化方面的优势，其后劲不容小觑。

苹果从A11开始引入端侧NPU AI加速器

型号	CPU性能	GPU性能	NPU性能	RAM	工艺制程
A10	2×Hurricane大核, 2.34GHz 2×Zephyr小核, 1.09GHz	PowerVR GT7600 Plus 900MHz 345.6 GFLOPS	—	64-bit单通道 LPDDR4, 1600MHz 25.6 GB/s	TSMC 16nm
A11	2×Monsoon大核, 2.39GHz 4×Mistral小核, 1.19GHz	3核GPU自研 1066MHz 408 GFLPOS	2×神经网络引擎 600 BOPS	64-bit单通道 LPDDR4, 2133MHz 34.1 GB/s	TSMC 10nm
A12	2×Vortex大核, 2.49GHz 4×Tempest小核, 1.59GHz	4核GPU自研 1125MHz 576 GFLOPS	8×神经网络引擎 5 TOPS	64-bit单通道 LPDDR4, 2133MHz 34.1 GB/s	TSMC N7 7nm
A13	2×Lighting大核, 2.65GHz 4×Thunder小核, 1.80GHz	4核GPU自研 1575MHz 806 GFLOPS	8×神经网络引擎 5.5 TOPS	64-bit单通道 LPDDR4X, 2133MHz 34.1 GB/s	TSMC N7P 7nm
A14	2×Firestorm大核, 2.99GHz 4×Icestorm小核, 1.82GHz	4核GPU自研 1700MHz 870 GFLOPS	8×神经网络引擎 11 TOPS	64-bit单通道 LPDDR4X, 2133MHz 34.1 GB/s	TSMC N5 5nm
A15	2×Avalanche大核, 2.93-3.23GHz 4×Blizzard小核, 1.82GHz	4/5核GPU自研 1175 GFLOPS	16×神经网络引擎 15.8 TOPS	LPDDR4X, 容量6GB 3200MHz 42.7 GB/s	TSMC N5P 5nm
A16	2×Avalanche大核, 3.46GHz 4×Blizzard小核, 2.02GHz	5核GPU自研 1468 GFLOPS	16×神经网络引擎 17 TOPS	LPDDR5, 容量6GB	TSMC N4 4nm
A17	2×大核, 3.70GHz 4×小核, 2.02GHz	6核GPU自研	16×神经网络引擎 35 TOPS	LPDDR5, 容量6GB	TSMC 3nm

手机NPU算力升级趋势



1.2.2 硬件侧：M系列芯片：统一内存架构+能效比优势在端侧AI时代极具想象空间

- 苹果自2020年11月开始推出针对Mac和iPad的M系列芯片，M系列芯片至今迭代了4代。
- 苹果M系列芯片采用统一内存架构设计，允许CPU、GPU及其他协处理器共享和访问相同的内存池，使得并行计算更快速高效。M4芯片支持120GB/s的系统内存带宽，M3芯片最高支持128GB内存，M2 Ultra配置高达192GB的统一内存和800GB/s的系统内存带宽，相当于支持了近似容量的显存，甚至可能容纳下LLaMA-65B模型（650亿参数，显存需求是130GB左右）。在苹果统一内存架构下，CPU、GPU和内存直接通过硅中介层连在一起，数据传输带宽极高。M3/M4系列芯片采用ARM架构设计，台积电3nm工艺制造，在能效比方面优势显著。**苹果统一内存架构+ARM架构的能效比优势**在端侧AI时代极具想象空间。

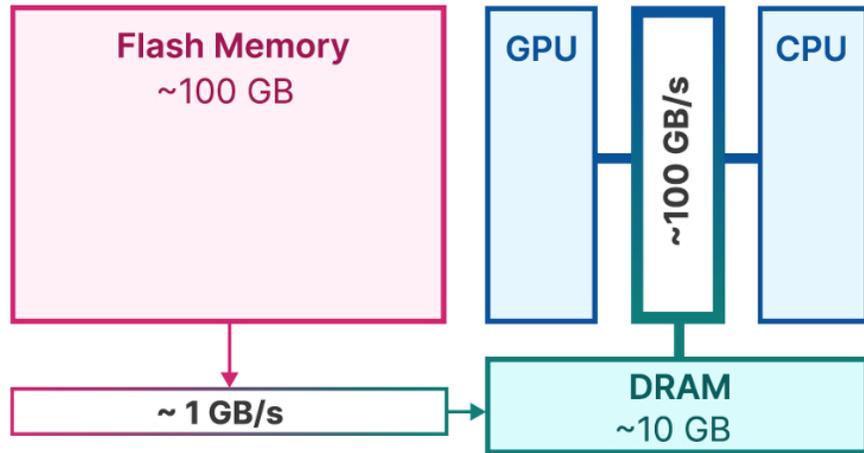
苹果M系列芯片规格

参数指标	M2 Ultra	M3	M3 Pro	M3 Max	M4
CPU	24 核中央处理器，16×性能核心+8×能效核心	8 核，4×性能核心+4×能效核心	12 核，6×性能核心+6×能效核心	16 核，12×性能核心+4×能效核心	10 核，4×性能核心+6×能效核心
GPU	60/76 核图形处理器，27.2 TFLOPS (FP32)	10 核，硬件加速光线追踪	18核，硬件加速光线追踪	40核，硬件加速光线追踪 17.04 TFLOPS (FP32)	10 核，硬件加速光线追踪
NPU	32 核神经网络引擎，31.6TOPS	16 核神经网络引擎，18TOPS	16 核神经网络引擎，18TOPS	16 核神经网络引擎，18TOPS	16 核神经网络引擎，38TOPS
支持内存	最高192GB	最高24GB	最高36GB	最高128GB	最高24GB
内存带宽	800GB/s	100GB/s	150GB/s	400GB/s	120GB/s
片间带宽	2.5TB/s	——	——	——	——
媒体处理引擎	支持 H.264、HEVC、ProRes 和 ProRes RAW 硬件加速 2个视频解码引擎，4个视频编码引擎，4个ProRes编解码引擎	支持 H.264、HEVC、ProRes 和 ProRes RAW 硬件加速 视频解码引擎，视频编码引擎，ProRes编解码引擎，AV1 解码		支持 H.264、HEVC、ProRes 和 ProRes RAW 硬件加速 视频解码引擎，2个视频编码引擎 2个ProRes编解码引擎，AV1 解码	支持 8K H.264、HEVC、ProRes 和 ProRes RAW 硬件加速 视频解码引擎，视频编码引擎，ProRes编解码引擎，AV1 解码
工艺制程	5nm	第一代3nm	第一代3nm	第一代3nm	第二代3nm
晶体管数量	1340亿	250亿	370亿	920亿	280亿

1.2.3 苹果AI：在内存有限的环境下如何增强LLM的推理效率

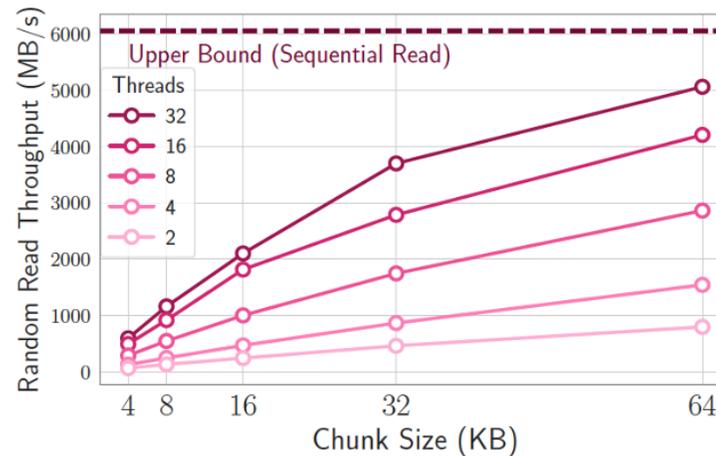
- ❑ 苹果在其发表的论文《LLM in a flash: Efficient Large Language Model Inference with Limited Memory》中详细阐述了如何在DRAM容量有限的设备中高效地运行LLMs。
- ❑ 苹果将模型参数存储在Flash中，按需将其带入DRAM，从而高效地运行超出可用DRAM容量的LLMs。为此，苹果构建了一个推理成本模型，其通过引入两种主要技术，即“窗口化”（通过重用之前激活的神经元战略性地减少数据传输）、以及“行列打包”（针对闪存的顺序数据访问优势，增加了从闪存读取的数据块大小），减少了从闪存传输的数据量、以及以更大更连续的块读取数据。这些方法使得能运行的模型大小达到可用DRAM容量的两倍，与CPU和GPU中的简单加载方法相比，推理速度分别提高了4-5倍和20-25倍。这些方法促成了数据负载的显著减少和内存使用效率的提高，对于端侧部署先进的AI模型尤为关键。

闪存提供更高的容量，但带宽较小



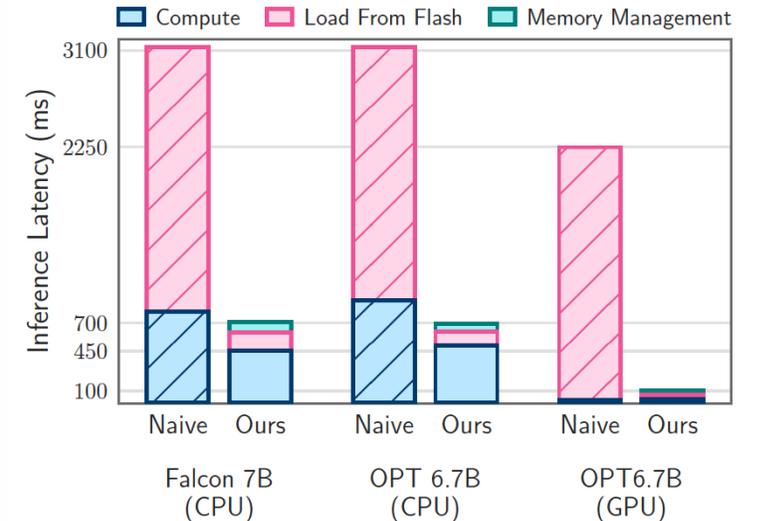
(a) Bandwidth in a unified memory architecture

闪存中随机读取的吞吐量随着顺序块大小和线程数量增加而增加



(b) Random read throughput of flash memory

苹果的方法使得单token的推理延迟大幅缩减



1.3.1 苹果AI：应用侧：改造Siri，紧握AI时代流量入口

- 据纽约时报报道，苹果预计将在6月举行的WWDC 2024上展示其人工智能领域的进展，预计将发布改进后的Siri语音助手，新Siri将更具对话性和多功能性。**Siri的底层技术将包括一个新的生成式AI系统，该系统将赋予Siri聊天功能，而不是一次回答一个问题。苹果专注于让Siri更好地处理其现有任务，包括设置定时器、创建会面日程和向杂货清单添加物品等，其功能更倾向于个人助手。**
- 根据苹果之前在语音领域发表的论文来看新版Siri可能的改进方向，主要有：1) 无需唤醒词，就能使用Siri助手，或者更准确地检测唤醒词；2) 改善用户与Siri助手之间的交流效果，比如让AI更好地理解模棱两可的问题、理解罕见词；3) 生成答案时不再冗长，用户更容易理解等。
- 结合5月OpenAI和谷歌的发布会来看，在AI语音方面，GPT-4o的拟人化风格，包括能识别人类感情，并根据用户感情做出相应的情绪反应，在语音交互中可随时打断，更贴近真实对话场景等，打破传统的类机器人体验；谷歌DeepMind展示了基于Gemini开发的智能体Astra的语音交互能力，包括强化的智能体发音、更广泛的语调、更好地理解上下文，并在对话中快速做出响应等。鉴于苹果与OpenAI和谷歌未来可能的合作，预计这些体验未来或将加入到Siri的功能当中。

GPT-4o拟人化语音交流

多模态GPT-4o可实时交互、检测人的情绪



it looks like you're feeling pretty happy and cheerful

看起来你很开心

科技增长

Google Astra演示用手机实时推理

Google Astra项目发布

有趣的AI视界

实时推理所处的环境
双设备丝滑运行

使用剪辑翻译，仅供参考



this appears to be the Kings Cross area of London
这似乎是伦敦国王十字区

1.3.2 苹果AI：应用侧：苹果原生APP与AI融合，或提升服务业务价值

- 据彭博社报道，预计苹果将在iOS 18中添加AI大模型能力，推出基于AI技术的新功能。
- 当前主流的AI应用或是苹果未来AI原生应用的发展方向。长期来看，苹果原生APP与AI的融合将为服务业务带来更大的价值提升。

苹果原生APP及对标AI应用

工具助手类



Siri

AI 助手



FaceTime

AI 通话/会议



Maps

AI 地图/定位



Translate

AI 翻译



办公类



Pages

智能写作



Microsoft Word



Numbers

智能制表

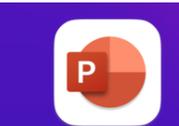


Microsoft Excel



Keynote

智能PPT创建

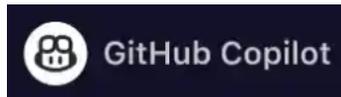


Microsoft PowerPoint



Xcode

AI 编程



生活娱乐类



Camera



Photos

AI 摄影/图像



Apple Music

AI 音乐



Apple TV+

AI 视频/影视



Apple Arcade

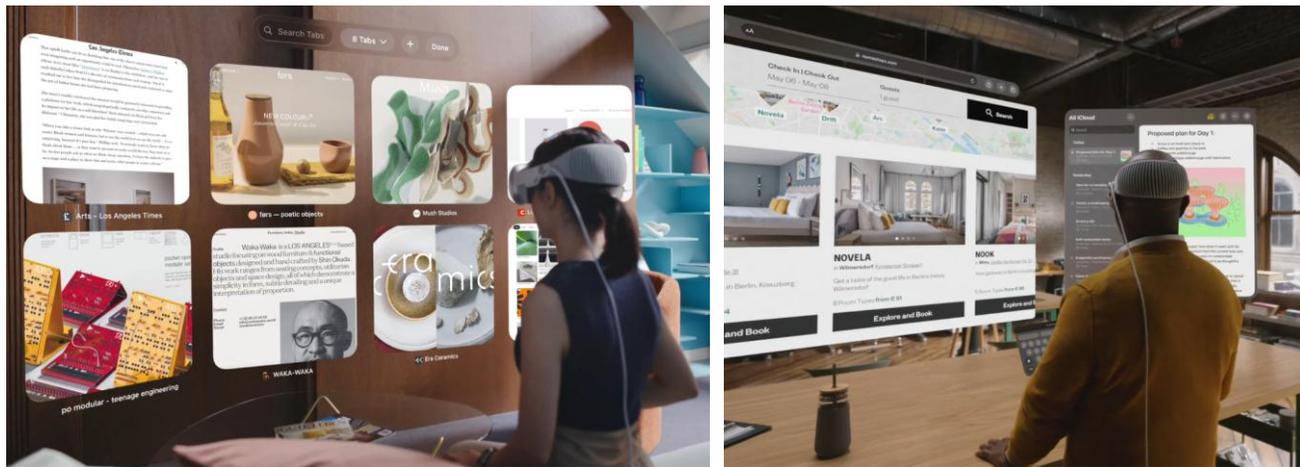
AI+游戏



1.4 苹果AI终端未来的杀手锏：AI Vision Pro打造全新生态

- ❑ XR设备可承载海量的信息流和深度的内容数据，其独特的优势使它能成为个人空间的延伸。计算机视觉、自然语言交互、深度学习等先进AI技术能使XR设备提供逼真的沉浸感、交互性和无缝的MR体验。内容方面，XR产业通过GPT等生成式AI的编码能力和虚拟世界中模型/环境/角色的自创能力，帮助企业持续提升生产力。AI与XR的融合或创造全新的生态，推动数字体验的发展，并将其扩展到制造业、虚拟生产、电子商务、教育等行业。
- ❑ 苹果的Vision Pro开启了空间计算时代，带来了全新的3D交互方式。生成式AI赋能空间计算将大幅推动产业的技术革新并提升用户体验。AI构建的底层框架，与空间计算的结合，有望创造出更宏伟的元宇宙上层建筑。

苹果Vision Pro交互界面



苹果Vision Pro重新定义XR设计



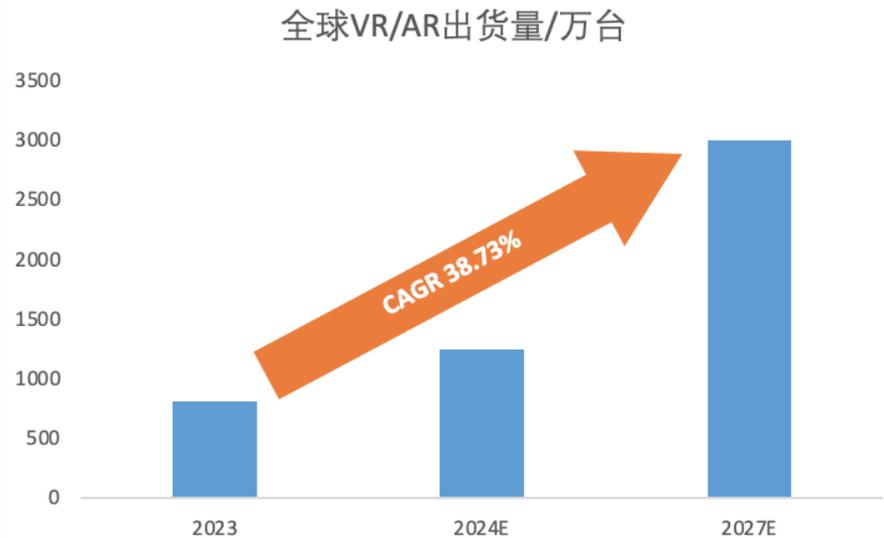
1.4 苹果AI终端未来的杀手锏：AI Vision Pro打造全新生态

- 苹果的Vision Pro开启了空间计算时代，带来了全新的3D交互方式。VisionOS搭载全新的3D界面，让数字内容看起来就像在用户的真实世界存在。Vision Pro带来全新的输入系统，用户通过眼睛、双手和语音来控制。VisionOS作为革命性的操作系统，利用用户周围的空间提供强大的空间体验，解锁办公和居家的全新机会。硬件方面，Vision Pro突破性的显示和先进的音频体验、以及高效的眼睛追踪系统，由M2和R1芯片来驱动。**生成式AI赋能空间计算将大幅推动技术革新和提升用户体验，基于苹果强大的芯片能力，Vision Pro将有力承接AI模型和生成式AI在端侧的部署，开创新的元宇宙时空。**
- IDC预计2024年VR/AR头显市场将增长至1250万台左右，主要来自苹果Vision Pro和Meta Quest 3的推动。到2027年，预计全球VR/AR销量将达到3000万台。

XR、智能眼镜的实时推理能力在不断进化



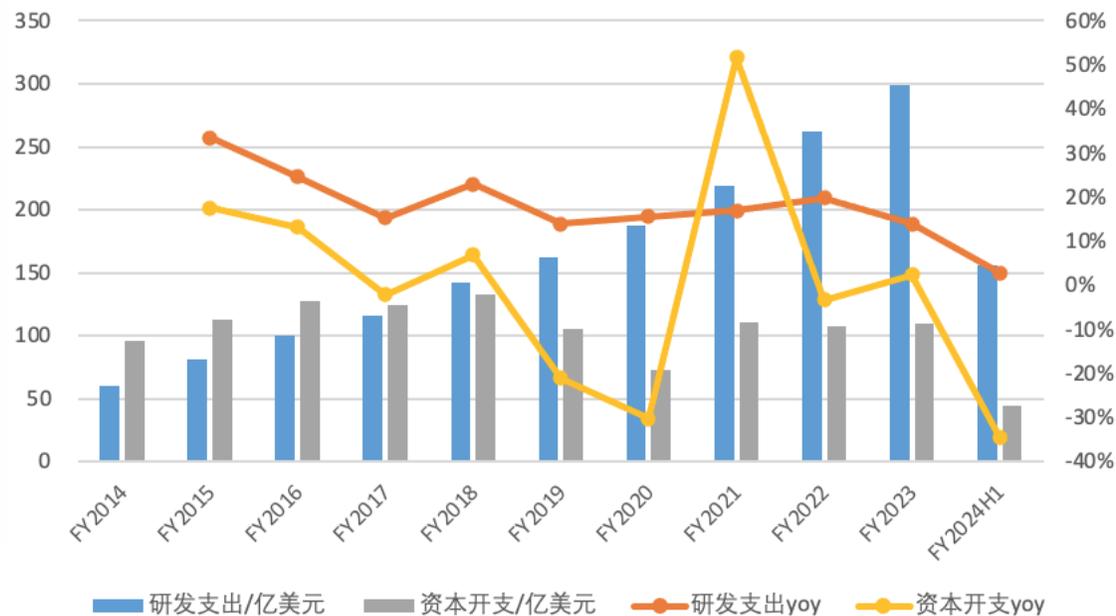
全球VR/AR出货量



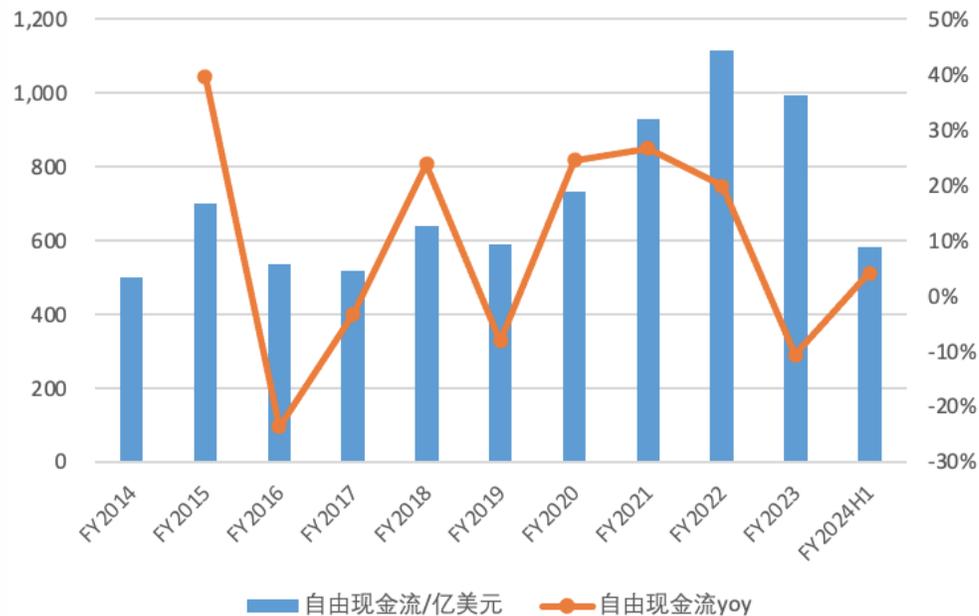
1.5 良好的现金流状况和对新技术的巨大投入是苹果深度布局AI的保障

- 苹果一直将AI和机器学习视为基础技术，并融入到其大部分产品中。苹果研发支出稳步提升，2023财年研发投入近300亿美元，2024财年上半年研发投入156亿美元，同比增长2.9%。苹果资本支出常年超过100亿美元，预计AI投入将提升苹果的资本支出规模。自由现金流2023财年达到995.8亿美元，2024财年上半年为582亿美元，同比增长4.2%。良好的现金流状况和对新技术的巨大投入是苹果深度布局AI的保障。

苹果研发投入和资本支出情况



苹果自由现金流情况



目录

◆ 1 苹果AI布局

◆ 2 高通AI布局

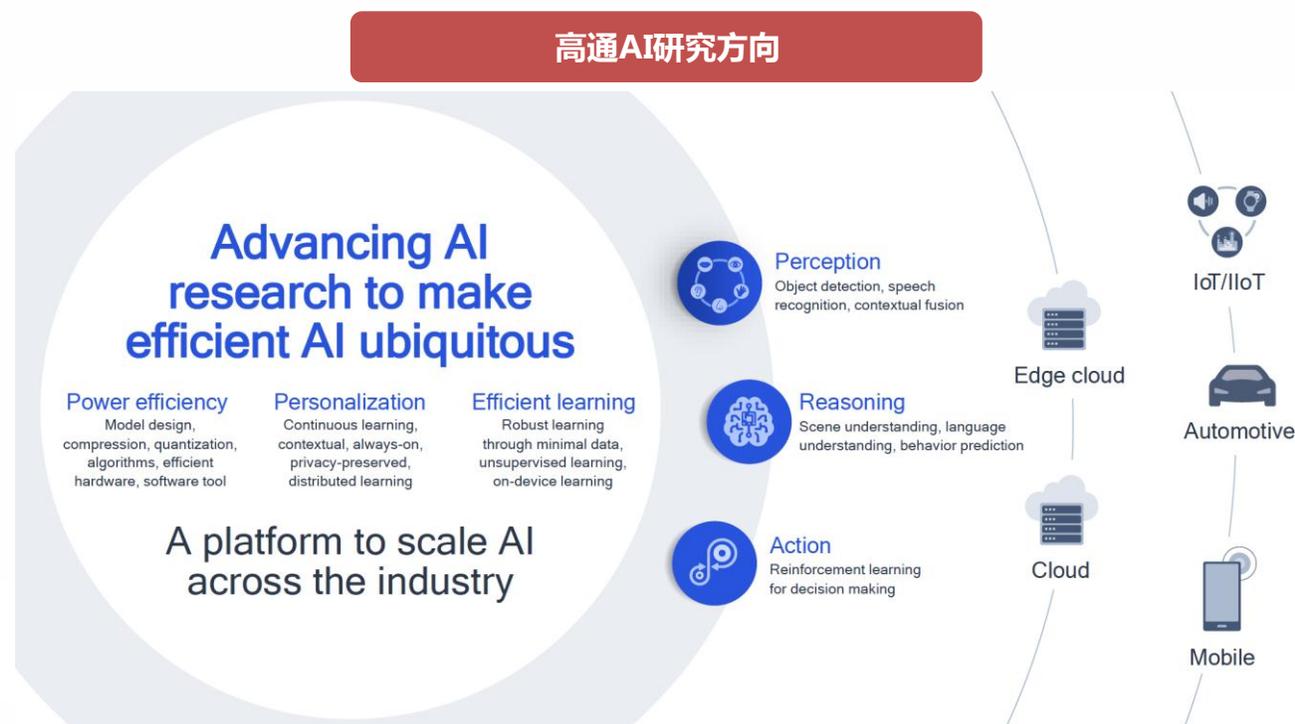
◆ 3 三星AI布局

◆ 4 相关标的

◆ 5 风险提示

2.1 高通旨在推动端侧AI更加高效

- 高通旨在推动 AI 更加高效，主要集中在**功耗效率**，**个性化**，**高效深度学习**三个方向。高通的AI平台可向多个行业扩展。
 - ✓ 针对功耗效率的优化，主要涵盖了**模型设计、压缩、量化、算法、高效硬件、软件工具**等。
 - ✓ 针对AI个性化，主要涵盖了**持续学习、情境感知、长期在线、隐私保护、分布式学习**等方面。
 - ✓ 针对高效深度学习，主要是**通过最少数据进行稳健学习、无监督学习和设备端学习**。
- 高通的AI框架主要涵盖三个层面。**感知层：物体检测、语音识别、情境融合。推理层：场景理解、语言理解、行为预测。执行层：用于决策的强化学习。**



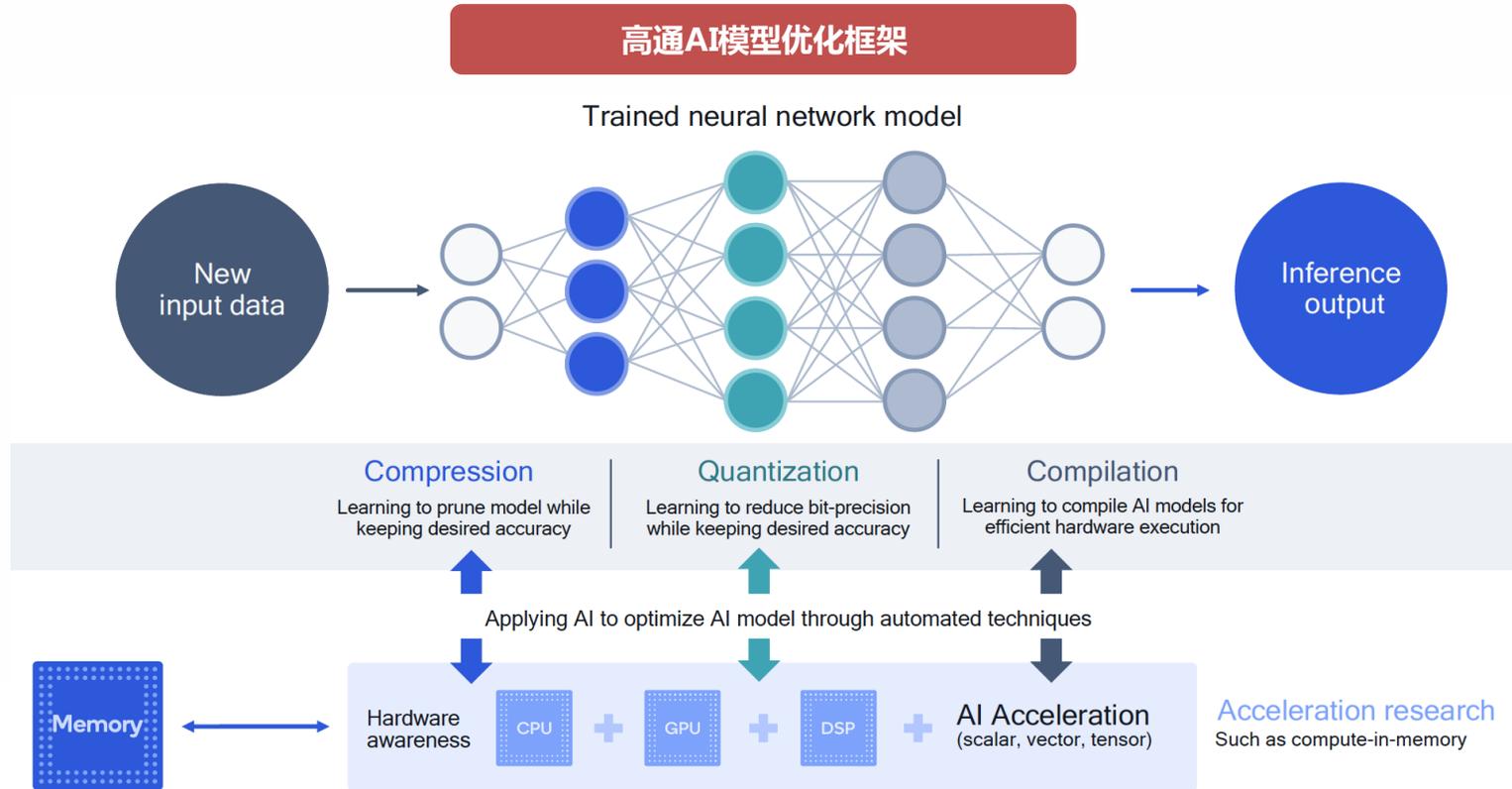
2.2 高通对端侧AI的挑战的看法

- AI工作负载的挑战在于：计算密集度高，复杂的并发，实时运行，长期在线等。
- 而终端环境受以下限制：对于终端轻薄设计，必须考虑**热效率**；长时间使用需要**较长的电池寿命**；**存储/内存带宽的限制**。



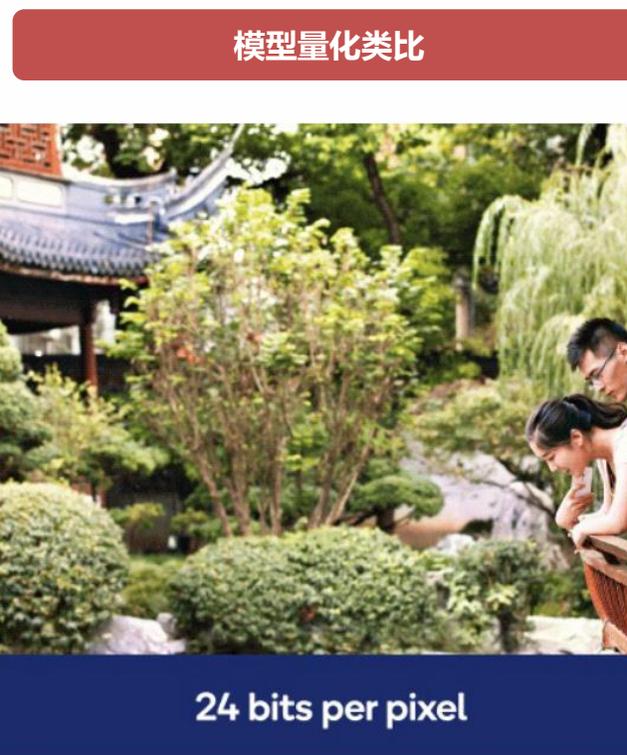
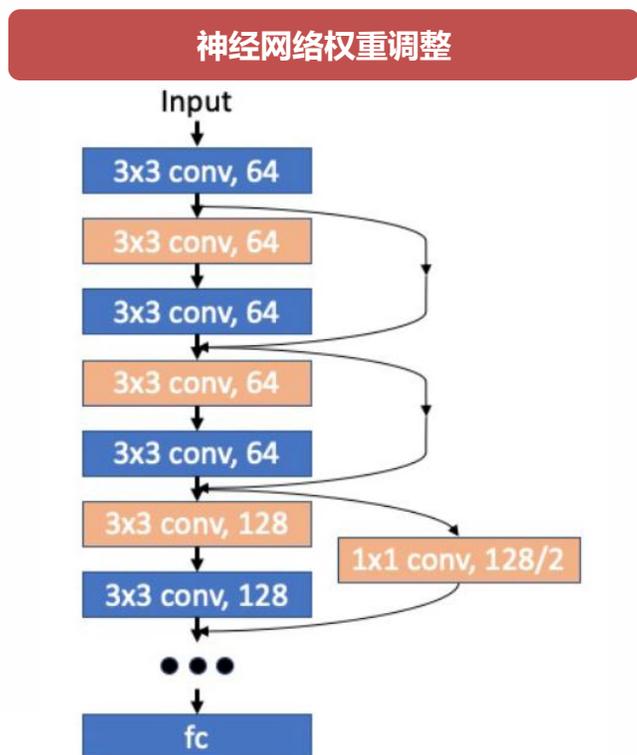
2.3.1 模型侧：高通对AI模型的优化方向

- 高通通过模型压缩、量化、编译等操作对AI模型进行优化。
 - ✓ **模型压缩**：学习对模型进行裁剪，同时保持所需的准确性；
 - ✓ **模型量化**：在保持模型所需精度的同时降低比特精度；
 - ✓ **编译**：学习编译 AI 模型以实现高效的硬件执行。



2.3.2 模型侧：高通AI模型量化策略

- 对于任何给定的经过训练的神经网络：将权重存储在低位（INT8），以低位进行计算
- 量化类比：使用较少的位表示图像中的每个像素，在保持模型所需精度的同时降低比特精度。



2.3.3 模型侧：高通AI模型量化策略带来的好处

减少内存占用



01010101

功耗
显著降低计算和内存访问的能耗

Add energy (pJ)		Mem access energy (pJ)	
INT8	FP32	Cache (64-bit)	
0.03	0.9	8KB	10
30X energy reduction		32KB	20
		1MB	100
Mult energy (pJ)		DRAM	1300-2600
INT8	FP32	Up to 4X energy reduction	
0.2	3.7		
18.5X energy reduction			

延迟
通过更少的内存访问和更简单的计算，可以减少延迟

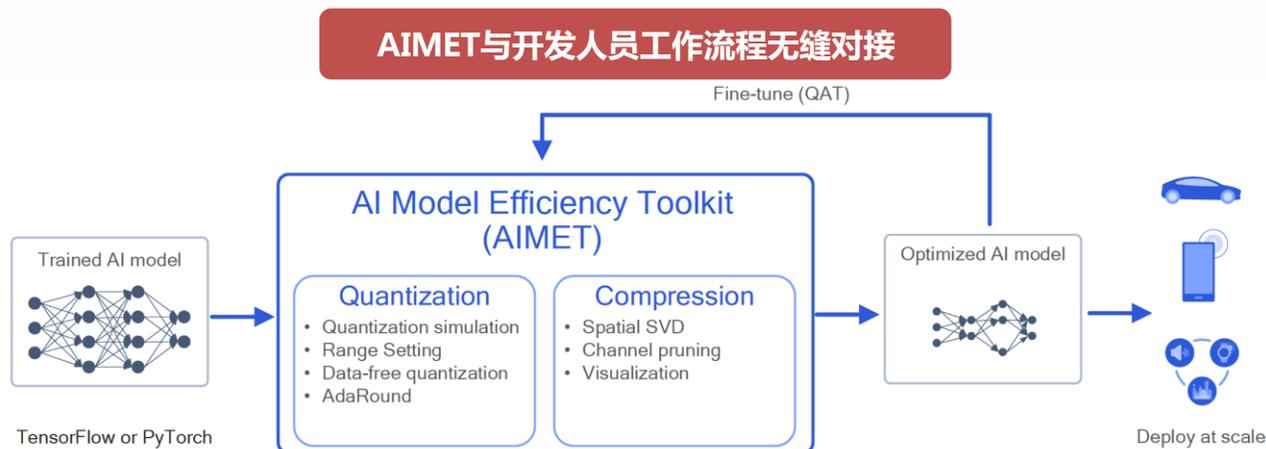


半导体
与浮点数运算和更多位数相比，整数运算或较少的位数需要较少的芯片面积

Add area (μm^2)	
INT8	FP32
36	4184
116X area reduction	
Mult area (μm^2)	
INT8	FP32
282	7700
27X area reduction	

2.3.4 模型侧：高通AIMET: 革新的AI模型优化工具

- 高通推出了针对端侧AI的优化工具库AIMET (AI Model Efficiency Toolkit)。AIMET是开源项目，旨在提供先进的深度学习模型优化技术。AIMET工具集提供了多种工具，以提高模型性能，降低计算资源需求，从而加速AI模型的部署。
 - ✓ **模型压缩与量化**：AIMET提供了一套完整的量化解方案，支持对模型进行INT8或更低精度量化，同时保持高精度。它采用了动态量化策略，可以在训练过程中调整权重的表示范围，保证模型在量化后的性能。
 - ✓ **知识蒸馏**：知识蒸馏是将大型复杂模型（教师模型）的知识转移到小型模型（学生模型）的过程。AIMET实现了高效的蒸馏算法，使得小模型能在保持高性能的同时减少内存占用和推理时间。
 - ✓ **剪枝与结构优化**：通过网络结构分析，AIMET可以识别并删除对模型性能影响较小的连接，实现网络的稀疏化和结构优化，进一步减小模型大小。
 - ✓ **元训练与迁移学习**：该项目提供了元训练算法，允许模型快速适应新任务，减少了对大量新数据的需求。这对于边缘设备上的应用尤其重要，因为它们通常受限于存储和计算资源。
- AIMET支持多硬件平台和深度学习框架，如TensorFlow、PyTorch等。清晰的API设计使它能无缝地融入现有开发流程，更易于集成。用户可以根据自己的需求选择不同的优化策略或实现自定义优化器，实现高度定制化。随着社区的发展和研究的进步，AIMET持续更新，不断引入新的技术和特性。
- AIMET能帮助开发者在终端侧创建更小巧、高效的模型，也能减少云端的计算资源消耗，降低成本。



2.3.4 模型侧：高通AIMET: 革新的AI模型优化工具

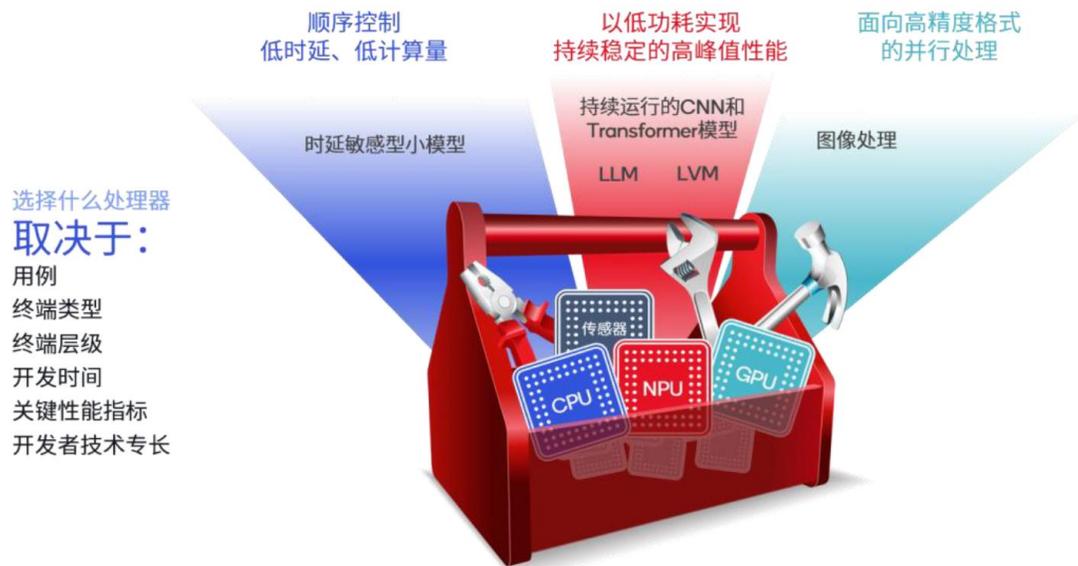
- 高通AIMET包括主流的量化人工智能模型。
- INT8的精度得以保持——精度损失小于1%。

AIMET量化模型FP32与INT8精度的比较



2.4.1 硬件侧：生成式AI对异构计算的要求

- 端侧生成式AI模型日益复杂，参数规模从十亿级别到百亿级别，且多模态趋势日益增强，这意味着模型能接受多种输入形式，并生成多种输出结果。此外，许多用例需同时运行多个模型。生成式AI工作负载的复杂性、并发性和多样性要利用SoC中所有处理器的能力。最佳的解决方案：1) 跨处理器和处理器内核扩展生成式AI处理；2) 将生成式AI模型和用例映射至一个或多个处理器及内核
- 大多数生成式AI用例可分为按需型、持续型或泛在型用例。按需型应用的关键性能指标是时延，这些应用使用小模型时，通常选择CPU。当模型变大(比如数十亿参数)时，GPU和NPU更合适。电池续航和能效对于持续和泛在型用例至关重要，NPU是最佳选择。
- 另一个关键区别在于AI模型为内存限制型(即性能表现受限于内存带宽)，还是计算限制型(性能表现受限于处理器性能)。当前的LLM在生成文本时受内存限制，因此需要关注CPU、GPU或NPU的内存效率。对于可能受计算或内存限制的大视觉模型，可使用GPU或NPU，但NPU可提供更好的能效。
- 提供自然语音用户界面(UI)以提高生产力并增强用户体验的个人助手预计将成为一类流行的生成式AI应用。语音识别、大语言模型和语音模型必将以某种并行方式运行，理想的情况是在NPU、GPU、CPU和传感处理器之间分布处理模型。



2.4.2 硬件侧：高通AI引擎

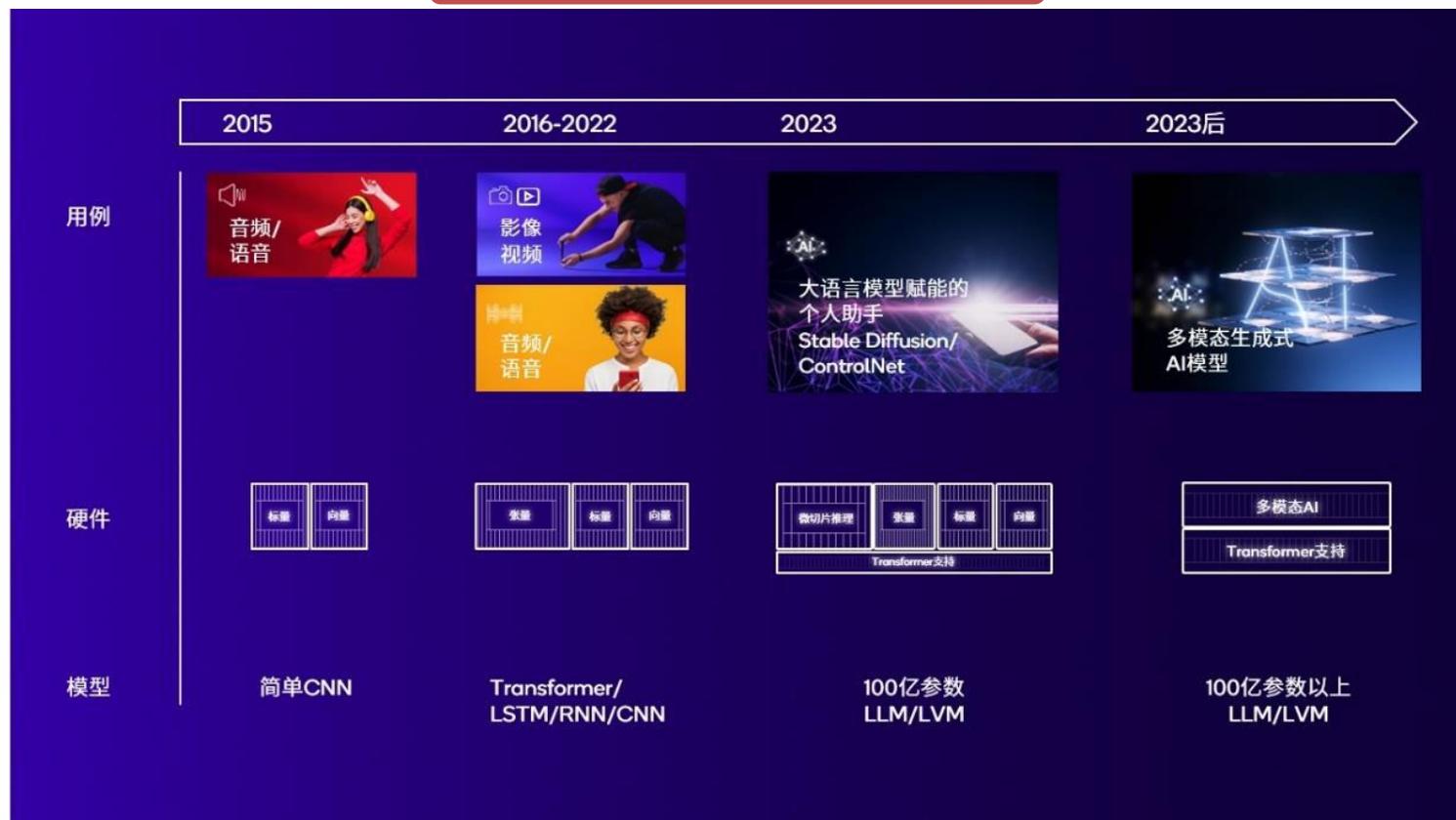
- ❑ 高通AI引擎包含多个硬件和软件组件，包括Hexagon NPU、Adreno GPU、Kryo/Oryon CPU、传感器中枢和内存子系统。
- ❑ 高通AI引擎的所有处理器相辅相成，实现效率提升：骁龙8Gen3相比上代NPU性能提升98%，能效提升40%；GPU能效提升25%；CPU性能达竞品2倍、功耗1/3；传感器中枢性能提升3.5倍，内存增加30%。具备专用电源传输轨道，在SoC其余部分关闭时运行，大幅节省电量。
- ❑ 相比于很多芯片组厂商选择授权多个三方处理器封装在一起的方式，高通采用差异化的定制设计方式，通过评估多样化处理器、系统架构和软件三个层级之间的关键约束条件、需求和依赖关系来设计SoC，包括设计共享内存子系统、决定不同处理器应支持的数据类型，为每类处理器插入全新的AI指令或硬件加速器等。
- ❑ 目前搭载高通AI引擎的产品出货超过20亿，涵盖广泛的终端品类，包括智能手机、XR、平板电脑、汽车、安防、机器人等。



2.4.3 硬件侧：高通NPU发展方向

- 高通NPU随着不断变化的AI模型和用例持续演进，实现高性能低功耗。
- NPU需考虑计算、内存和系统设计，通过减少内存数据传输提高性能和能效。未来或出现对更多规模模型和多模态模型的需求。随着AI持续快速演进，需要在性能、功耗、效率、可编程性和芯片面积之间权衡取舍，定制化NPU与AI行业方向保持高度一致。

高通NPU演进方向



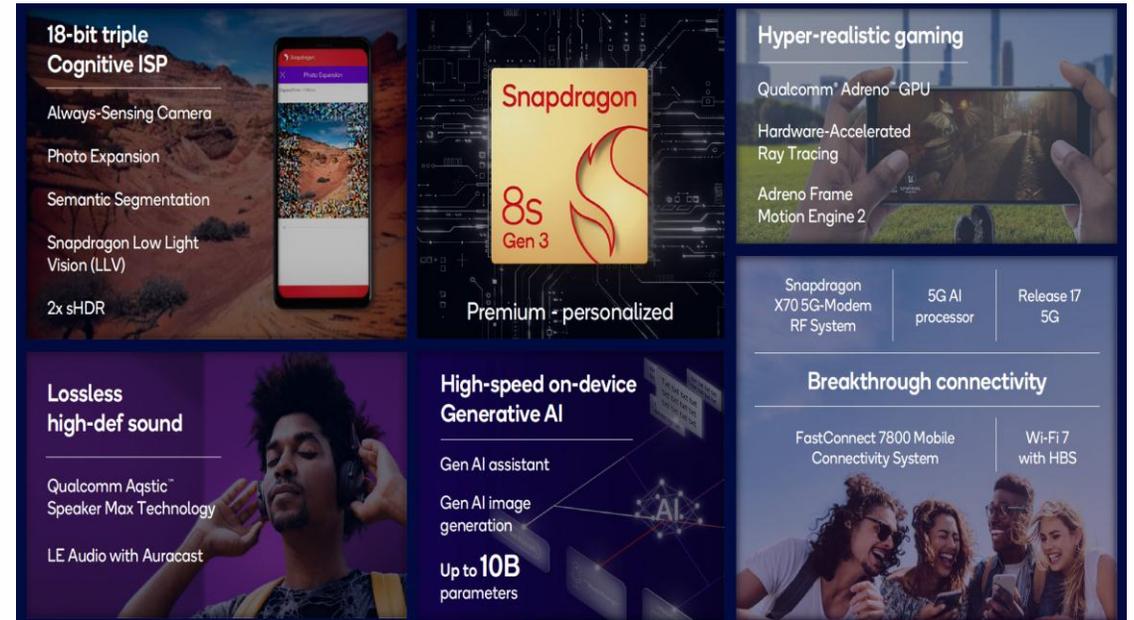
2.4.4 高通骁龙8Gen3 & 8s Gen3平台

- 高通骁龙8Gen3平台将高性能AI注入整个系统，将赋能用户创作独特内容、帮助生产力提升，并实现其他突破性的用例，进一步推动端侧AI的规模化扩展。骁龙8Gen3采用4nm制程。相比上一代骁龙8平台，骁龙8Gen3在NPU性能上有98%的提升，CPU性能有30%的提升，GPU性能有25%的提升，整体功耗降低10%。
- 高通骁龙8s Gen3平台支持端侧生成式AI功能、始终感知的ISP、超现实的移动游戏、突破性的连接和无损高清声音等。该平台支持广泛的AI模型，包括主流的大型语言模型，比如Llama 2、Gemini Nano、百川（70亿参数）和ChatGLM等。
- 骁龙8Gen3和8s Gen3成为支持终端上运行100亿参数模型的平台，并针对70亿参数LLM每秒生成20个tokens，各类复杂的AI应用可在搭载8Gen3和8s Gen3的手机上稳定运行。

高通骁龙8Gen3主要性能指标



高通骁龙8s Gen3主要性能指标



2.4.5 高通骁龙X Elite & X Plus平台

- ❑ 骁龙X Elite是高通第一款针对PC市场推出的SoC产品，骁龙X Elite采用高通Oryon CPU，CPU性能为竞品两倍。算力为75 TOPS，支持在终端运行超130亿参数的生成式AI模型，针对70亿参数LLM每秒可生成30个tokens，AI处理速度为竞品4.5倍。
- ❑ 近期高通推出AI PC品牌骁龙X Plus平台。X Plus采用10核CPU设计，GPU算力3.8 TFLOPS，NPU算力为45 TOPS。在Snapdragon X Plus的发布中，高通展示了在45 TOPS NPU上运行的新的AI优化应用程序和功能，包括：Codegen的Visual Studio代码生成，通过端侧生成式AI即时生成新代码来帮助程序员。Audacity中的音乐生成，使用设备AI上的Riffusion，根据提示或预先存在的音乐生成新音乐。OBS Studio中的实时字幕，在直播过程中使用设备上的Whisper实时将100种口语自动翻译成100种语言的实时字幕。

高通骁龙X Elite主要性能指标



高通骁龙X Plus主要性能指标



www.swsc.com.cn

数据来源：高通官网，西南证券整理

2.5.1 应用侧：高效视频/图像处理

- 随着对高效视频处理的需求不断增加，在各种设备和使用场景中，如何从视频流中提取有价值的信息至关重要。
- 高通提供的能力包括：通过对象检测和语义分割增强感知能力；高级相机功能，如视频增强和超高分辨率；高级视频理解，如搜索和监控；提高视频压缩以应对丰富媒体的需求。
- 由于视频帧之间存在冗余，高通开发了AI工具以减少不必要的计算，通过识别视频帧之间的冗余，以避免重复计算相同内容。通过该项计算，ResNet18的特征图在时间上基本保持不变。

从各种使用场景中提取视频流价值信息



Enhanced perception through object detection and semantic segmentation



Advanced camera features, like video enhancement and super-resolution

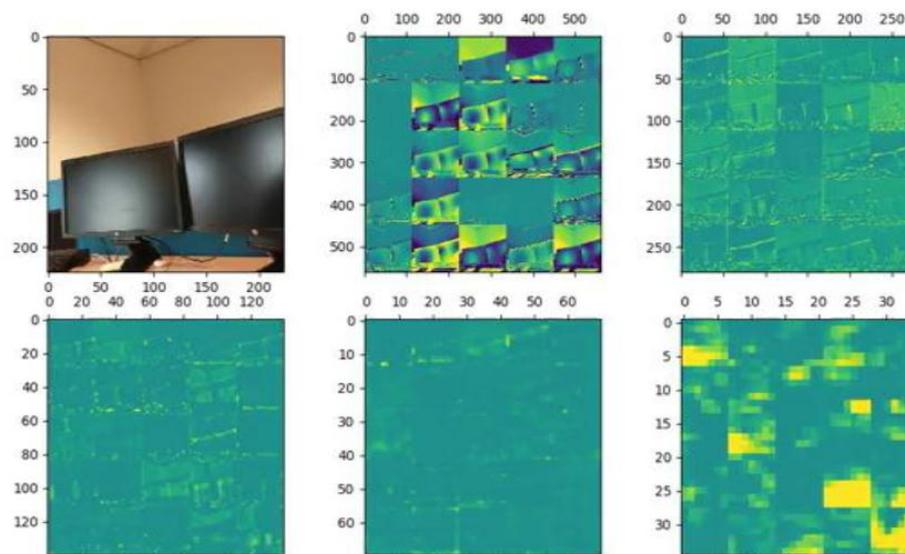


Advanced video understanding, like search and surveillance



Increased video compression to address the demand for rich media

识别视频帧之间的冗余，避免重复计算

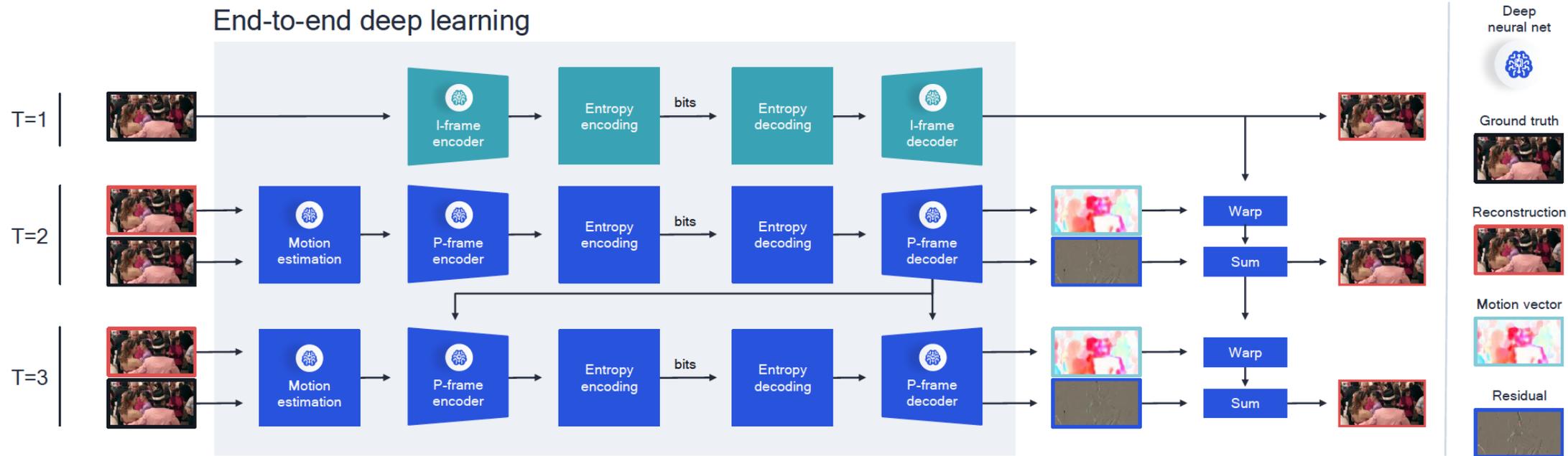


Feature maps over time for ResNet18 remain mostly constant

2.5.1 应用侧：高效视频/图像处理

- 针对视频编解码，基于新型神经网络学习进行帧和P帧压缩。
- 与其他学习型视频压缩解决方案相比，高通的方案实现了更先进的码率失真性能。
- 基于高效视频流/图像处理的能力，未来可以拓展延伸至AR/VR/MR、机器视觉、视频监控、智能驾驶等多个领域。

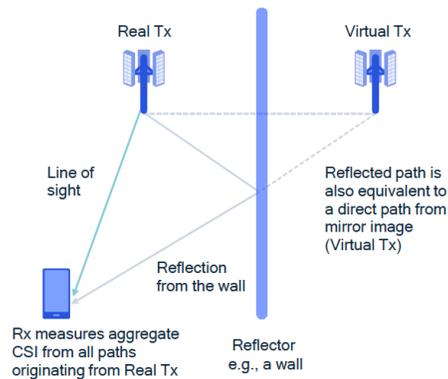
基于新型机器学习的视频编解码研究



2.5.2 应用侧：射频无线定位

- 高通利用深厚的无线领域知识最优地利用AI能力，通过学习建模、训练策略和计算上可实现的解决方案、学习补偿非线性因素等操作，以解决无线领域难以建模、最优解决方案的计算非可行性、高效的调制解调器参数优化等问题。
- 高通将AI应用于射频实现厘米级精确定位。**精确的室内定位在各行业和位置感知服务中都很价值，比如机器人、智能眼镜、智慧工厂等。**由于室内射频传播建模很复杂，用于室内定位的 AI 通过学习复杂的传播物理学和神经网络增强来实现高度准确的定位，包括解决射频信号的反射、衍射（射频路径在穿过物体时弯曲）、散射等问题。AI模型在无监督的情况下学习定位，通过一些标记的测量，解决了地图模糊性问题，以实现厘米级室内定位。

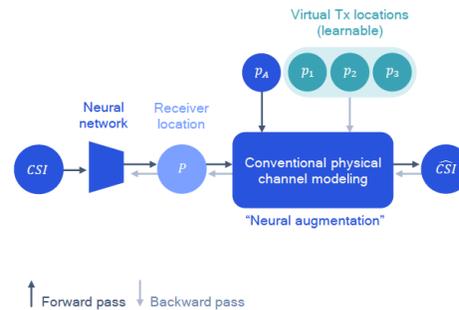
利用射频进行无监督学习以实现定位



Physics of reflections

The receiver (Rx) collects unlabeled channel state info (CSI) observations throughout the environment.

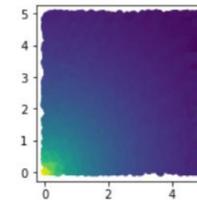
The goal is to learn the Virtual Tx locations and how to triangulate using CSI.



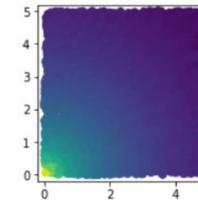
Neural augmentation

The neural network uses a generative auto-encoder plus conventional channel modeling (based on physics of propagation) to train on the observations and learn the environment.

Real map (CSI vs position)



Learned map Unsupervised



Incredible results

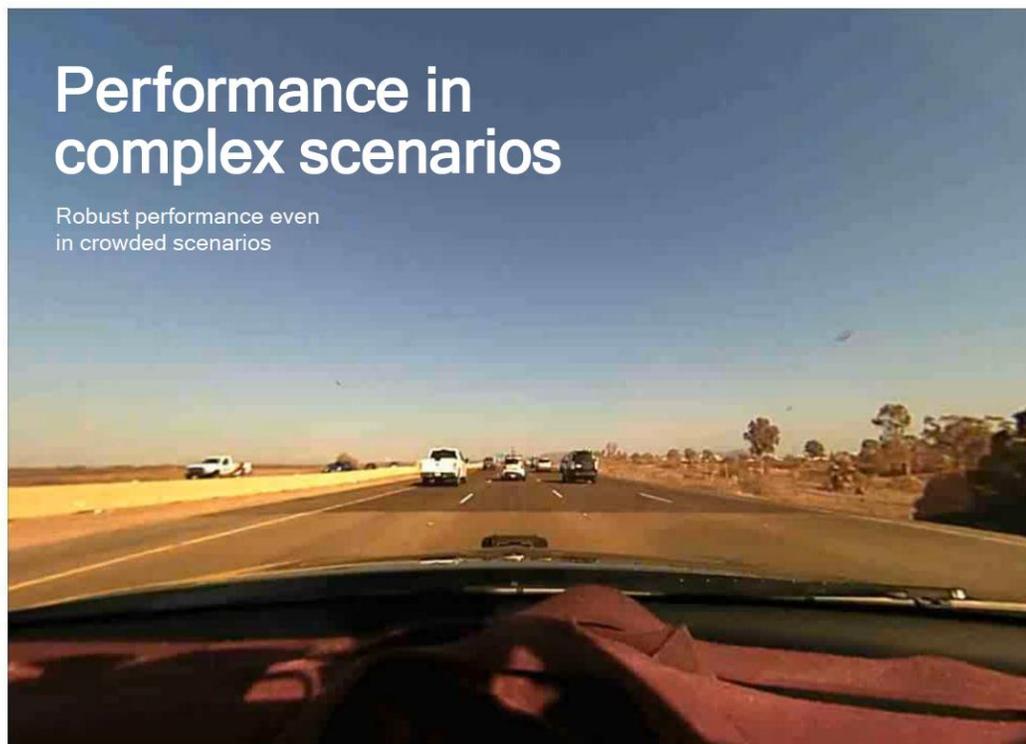
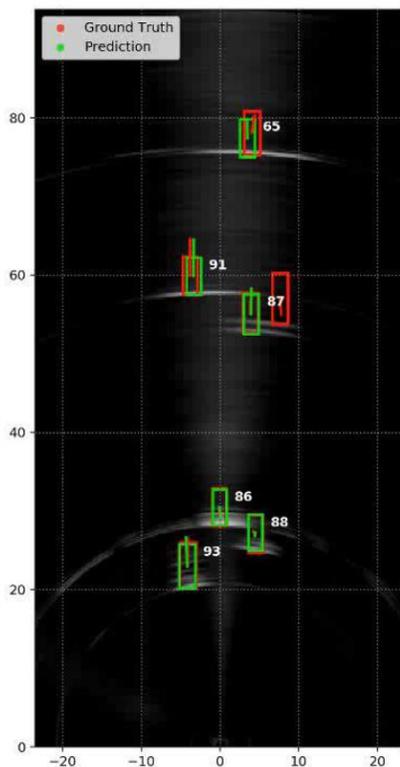
The neural network learns the virtual transmitter locations up to isometries completely unsupervised.

With a few labeled measurements, map ambiguity is resolved to achieve cm-level positioning.

2.5.3 应用侧：AI雷达

- 传统雷达价格实惠且响应迅速，具有长距离、直接测量速度、不受光照或天气条件影响等特点。高通将深度学习直接应用于雷达信号，并通过传感器融合，获得最佳视野，改进几乎所有现有的雷达功能。
- 高通将AI应用于雷达产品，在位置和尺寸估计、速度估计、目标分类以及不确定性估计方面取得显著改进，在复杂场景中也能表现出色。未来还可以拓展至雷达压缩、可行驶空间、稀疏雷达感知、行人感知、范围扩展和自适应采样研究等领域。

AI雷达在复杂场景中检测被遮挡的车辆



目录

◆ 1 苹果AI布局

◆ 2 高通AI布局

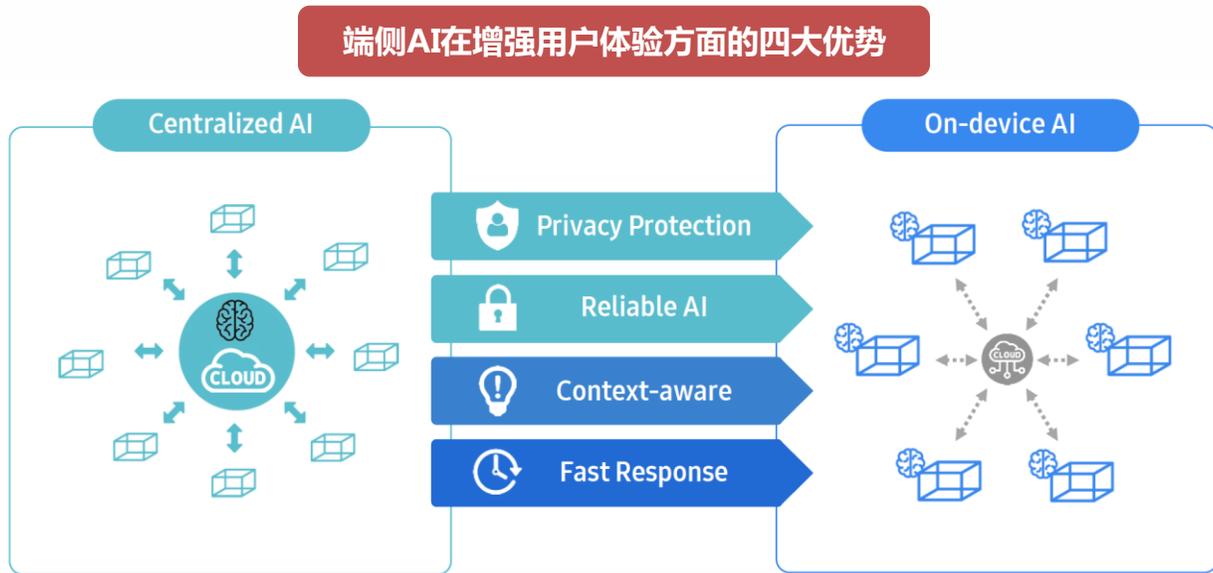
◆ 3 三星AI布局

◆ 4 相关标的

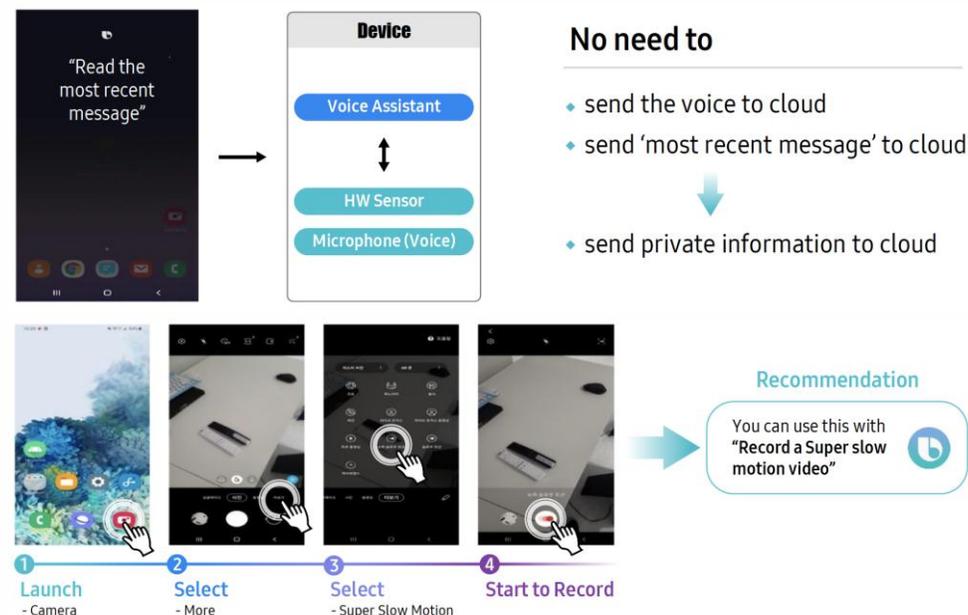
◆ 5 风险提示

3.1 三星对端侧AI优势的看法

- 三星研究院对设备端人工智能的探讨，涵盖了从**深度神经网络模型压缩**到**特定领域加速器**等领域。三星认为，端侧 AI 可在用户体验方面有效增强云端人工智能的效果。
- 三星认为端侧 AI 在增强用户体验方面具备四大优势：**隐私保护**，**随时可用的可靠性**，**更好的上下文感知**，**快速相应能力**。
 - ✓ **隐私保护**：重要信息通过端侧 AI 完成本地处理，无需上传云端，避免造成隐私泄露。
 - ✓ **随时可用的可靠性**：无论有无网络连接，或服务器出现问题，本地 AI 模型亦可实时工作。
 - ✓ **更好的上下文感知**：设备对用户的了解相当深入，端侧 AI 可以向用户提供最佳选项建议，倾向于个人定制化的AI助理。
 - ✓ **快速响应**：由于没有网络和服务器的延迟，执行速度很快。

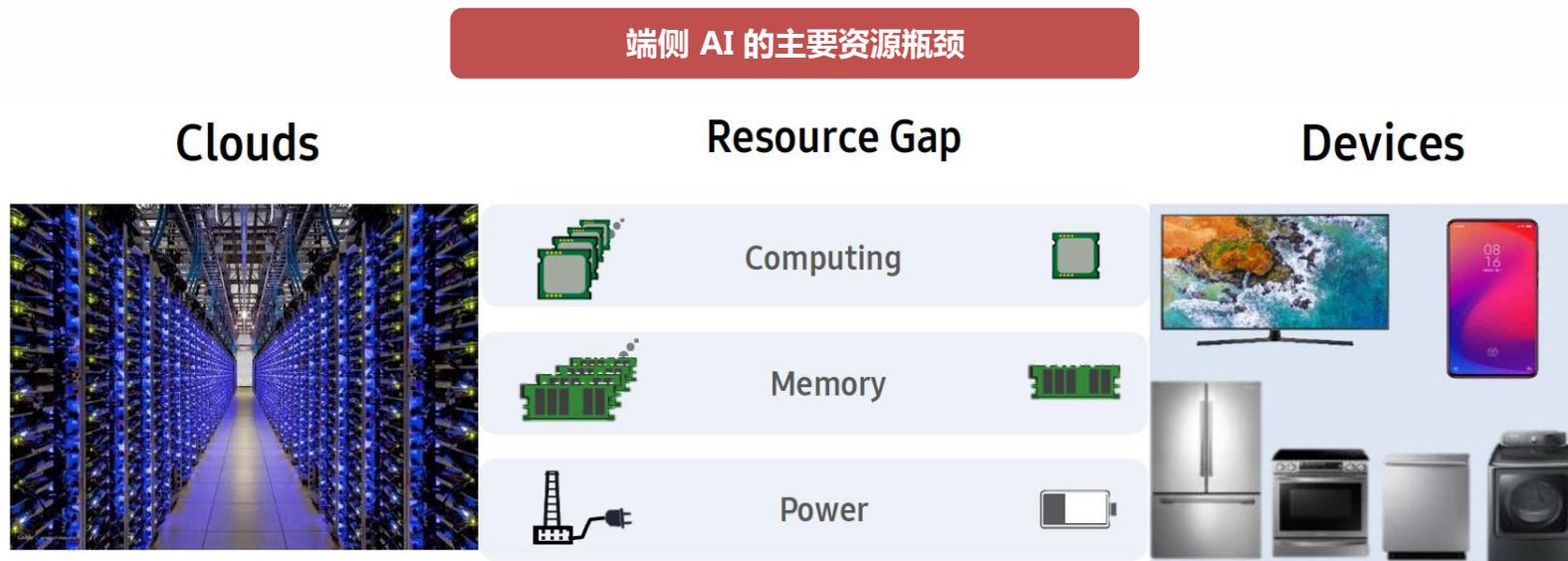


端侧AI在隐私保护、个性化助理方面优势



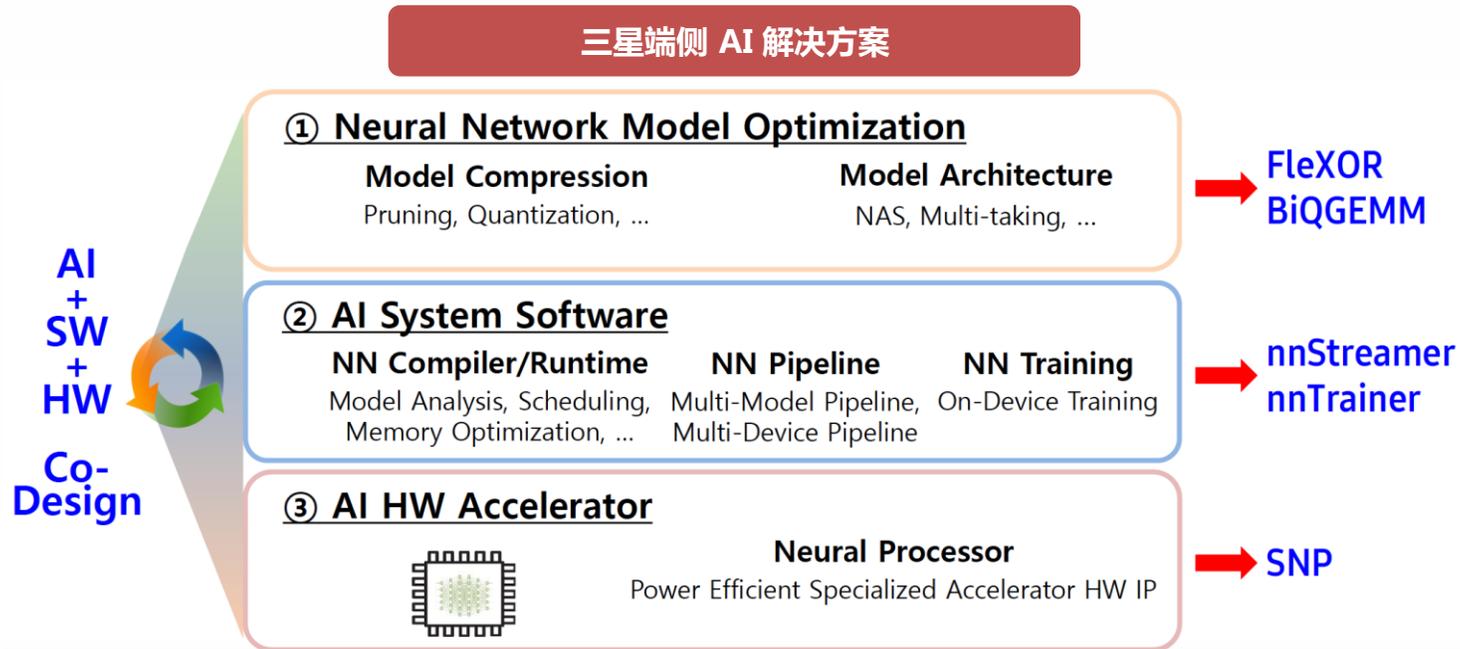
3.2 三星对端侧AI的挑战的看法

- 相比云端 AI 丰富的资源，端侧 AI 需要在有限的资源上运行，这对模型提出新的要求。
- 三星认为，端侧 AI 的主要资源瓶颈在于：**算力，存储，电力3大领域**。
 - ✓ 本地算力芯片在性能上与云端GPU有较大差异，要综合考虑性能、功耗、成本、便携性等因素。
 - ✓ 本地运行AI模型对缓存和硬盘提出更高的要求，主要集中在容量、规格等性能指标。
 - ✓ Scaling Law驱动 AI 模型向更强、更快发展，推理需求也在逐渐增多，云端和终端设备的供电或成为未来发展的制约因素。



3.3 三星的端侧 AI 解决方案

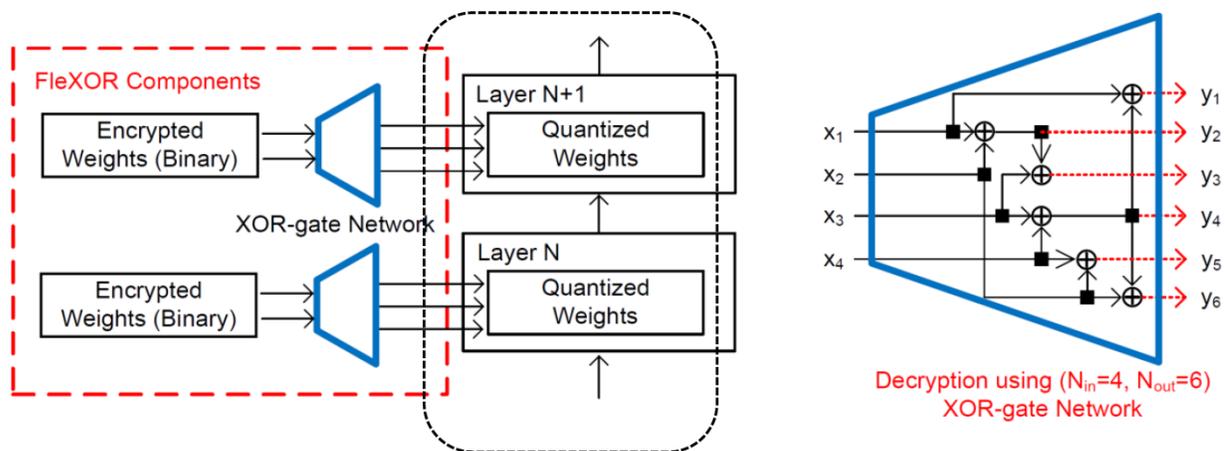
- 三星的端侧 AI 解决方案主要分为三部分：**模型优化**，**AI系统软件**，**AI硬件加速器**。
- 神经网络模型优化**：涵盖了模型压缩技术，包括模型裁剪、量化等；模型架构，包括神经架构搜索（NAS）、多任务学习等。主要由FleXOR和BiQGEMM实现。
- AI系统软件**：涵盖了神经网络编译器/Runtime，包括模型分析、调度、内存优化等；神经网络管道，包括多模型管道、多设备管道；设备端的模型训练。主要由nnStreamer和nnTrainer实现。
- AI硬件加速器**：增加NPU计算单元，嵌入节能型专用加速器硬件IP。



3.3.1 FleXOR : 可训练的分数量化

- ❑ FleXOR是一个灵活的加密算法/架构，它能使用小于1位的分数来表示每个权重。
- ❑ FleXOR基于异或的量化位加密增强了模型压缩比，异或感知训练算法学习加密权重，通过小于1位的量化实现较高的模型准确性。
- ❑ FleXOR允许减少内存占用和带宽，这对于节能的推理设计至关重要。

FleXOR架构



FleXOR减少了内存占用

Table 3: Weight compression comparison of ResNet-18 on ImageNet.

Methods	Bits/Weight	Top-1	Top-5	Storage Saving
Full Precision [10]	32	69.6%	89.2%	1×
BWN [20]	1	60.8%	83.0%	~ 32×
ABC-Net [18]	1	62.8%	84.4%	~ 32×
BinaryRelax [26]	1	63.2%	85.1%	~ 32×
DSQ [7]	1	63.7%	-	~ 32×
FleXOR ($N_{out} = 20$)	0.8	63.8%	84.8%	~ 40×
	0.63 (mixed) ²	63.3%	84.5%	~ 50.8×
	0.6	62.0%	83.7%	~ 53×

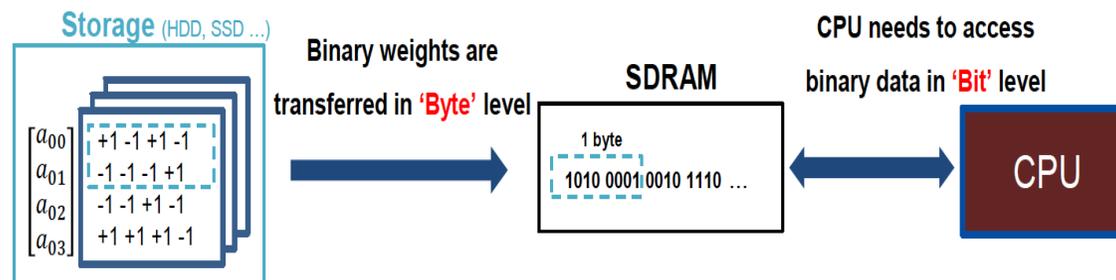
²To 4 groups of 3×3 conv layers in ResNet-18 (except the first conv layer connected to the inputs), we assign 0.9, 0.8, 0.7, and 0.6 bits/weight, respectively. To the remaining 1×1 conv layers (performing downsampling), we assign 0.95, 0.9, and 0.8 bits/weight, respectively.

3.3.2 BiQGEMM : 基于查找表的二进制编码量化DNN的矩阵乘法

之前的二进制编码量化实现：

- ✓ 用于带量化权重的矩阵乘法操作的特殊硬件
- ✓ 对于非均匀量化，CPU/GPU/NPU 在实际操作中需要在芯片上执行反量化
- ✓ 二进制编码仅用于减少内存需求（而不是延迟），不依赖特殊硬件

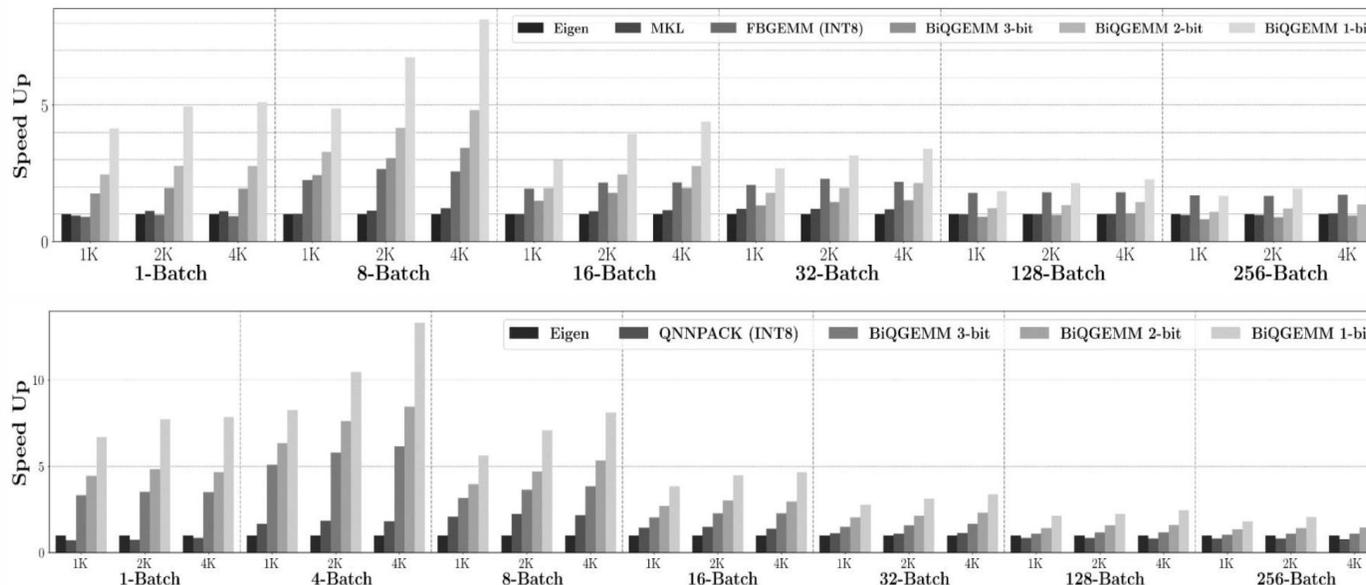
CPU/缓存/存储的运行流程



BiQGEMM使用1个线程的加速比 (上方PC, 下方移动终端)

BiQGEMM的优势，使得模型线程加速：

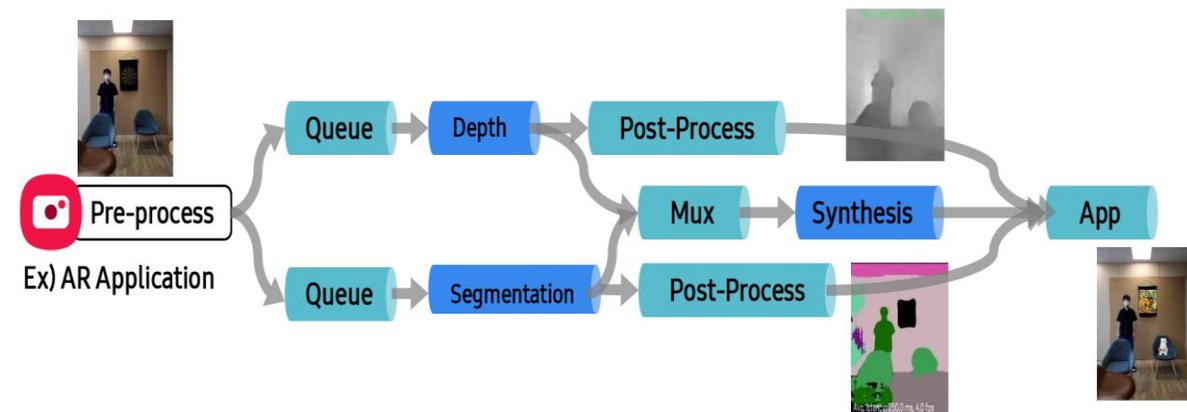
- ✓ 专用的基于查找表的二进制编码矩阵乘法单元内核设计
- ✓ CPU 和 GPU 可以利用量化矩阵来提高性能
- ✓ 采用无冗余计算，浮点乘法的数量大大减少



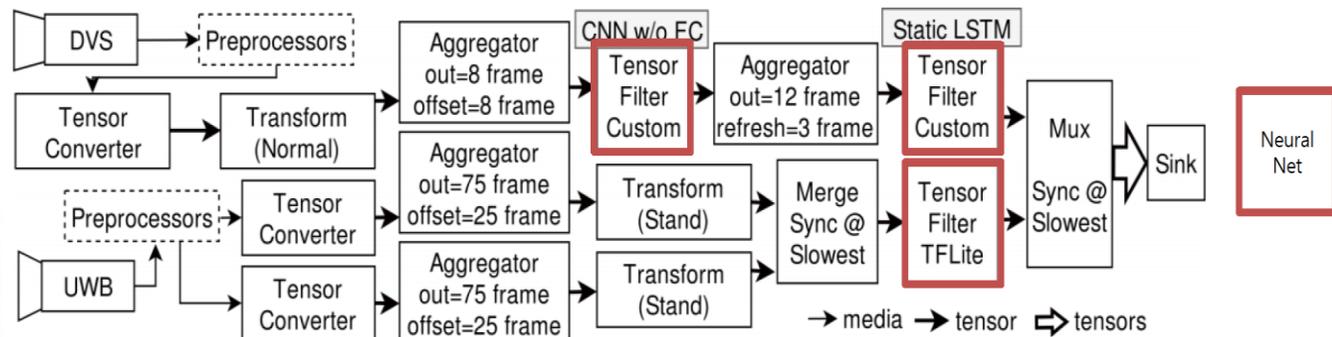
3.3.3 nnStreamer : 高效开发和执行机器学习管道

- nnStreamer是一个高效且灵活的神经网络管道：
 - ✓ 可将数千行代码转换为几十行的管道描述；
 - ✓ 将手动并行化升级为自动管道并行化；
 - ✓ 将直接媒体/硬件优化转变为可重复使用的媒体/硬件模块。
- nnStreamer实例：可将1000行代码压缩到16行；输入30FPS图像处理，CPU占用率从90.4%降至51.4%，内存占用空间从40MB减少至17MB。

nnStreamer神经网络管道



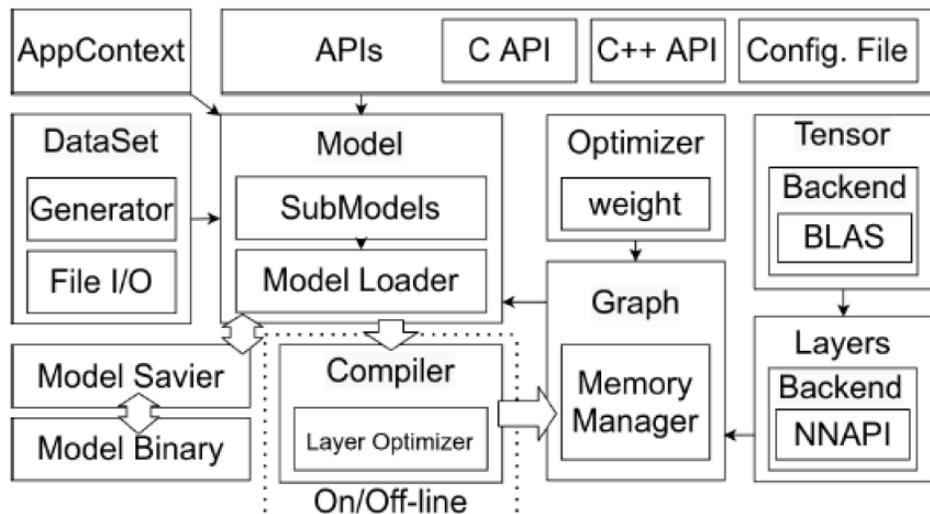
nnStreamer可有效减少CPU和内存占用



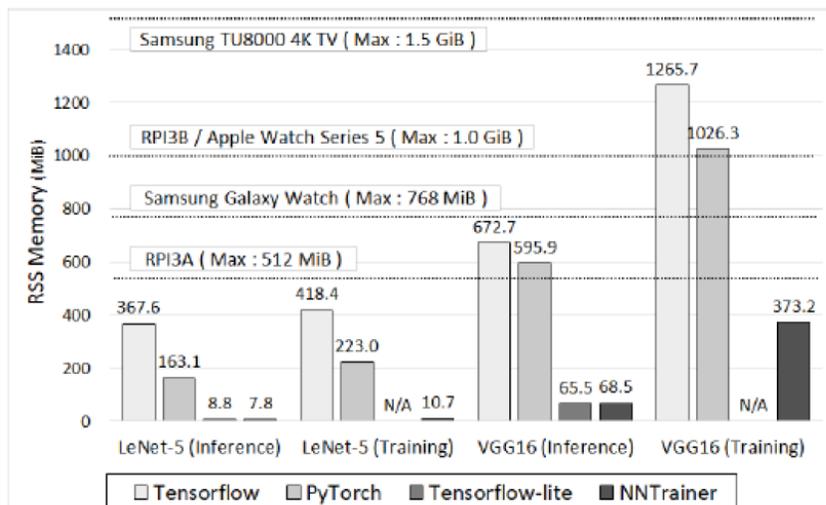
3.3.4 nnTrainer : 面向端侧的轻量化个性化深度学习框架

- nnTrainer是在嵌入式设备上训练神经网络的软件框架。
- 当用户使用AI应用时，nnTrainer将确保用户会获得更快（例如响应时间从100ms缩短到50ms）、更准确（比如精确度从88%提升到95%）、个性化的反馈体验。nnTrainer同时还提供隐私保护，将个人数据保留在用户设备上。对nnTrainer的挑战在于训练数据量小，算力和内存的资源有限。
- nnTrainer提供内存使用和训练时间的优化，迁移学习和元学习，TFLite/PyTorch模型级别的兼容性，易于实现自定义操作符。nnTrainer支持主流的Android、Tizen和Linux操作系统。相比于PyTorch和TensorFlow，nnTrainer最高可节省63.7%和69%的内存占用。

nnTrainer系统框架



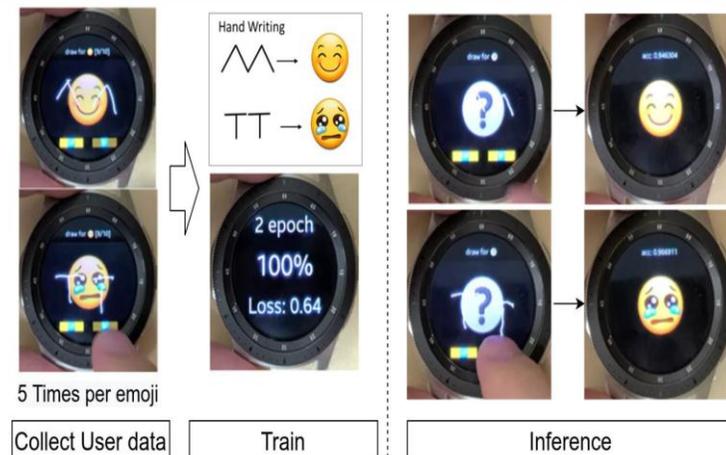
nnTrainer框架可有效节约内存消耗



Peak Memory Consumption

- PyTorch : 1.2 GiB
- TensorFlow : 1.02 GiB
- NNTrainer : 0.37 GiB

nnTrainer框架实例：定制化表情



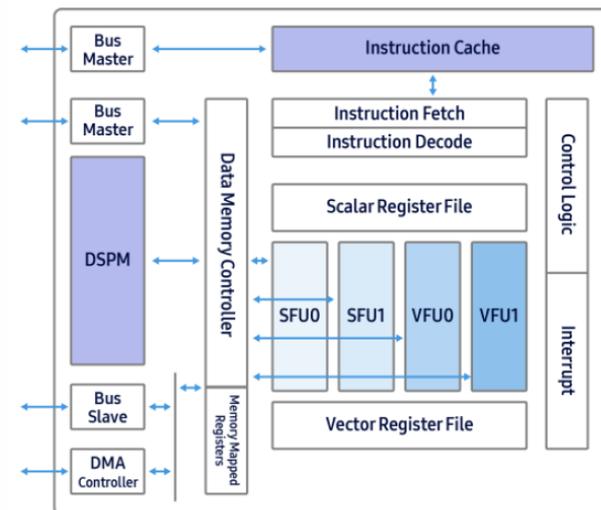
3.4 三星AI软硬件配合效果

- 体积缩小13倍的神经网络在语音识别方面可以表现出类似的性能。530MB时准确率为91.6%，升级后38MB时准确率为91.1%。
- 三星设计的用于语音识别的NPU，是一个应用于神经网络加速器的专用硬件，其功耗约为友商CPU产品的四分之一。

语音神经网络管道拥有更显著的性能优势

Bits	Hyper LRA	Korean			English		
		WER	xRT	Size	WER	xRT	Size
32	no	9.37	4.89	530.56	9.03	4.32	530.50
32	yes	9.85	0.99	140.18	8.91	1.15	153.98
32	+MWER	9.60	1.26	140.18	8.64	1.48	153.98
8	no	9.64	1.18	132.88	9.07	0.94	132.87
8	yes	10.21	0.33	35.34	9.24	0.38	38.77
8	+MWER	9.80	0.35	35.34	8.88	0.44	38.77

三星ASR(自动语音识别)加速器架构



	CPU	ASR Acceleration (NPU-based)
Power Consumption	982mW	276mW

* Measured under xRT(real-time factor) <1

3.5 三星端侧 AI 部署

- 三星高斯大模型包括**高斯语言、高斯代码和高斯图像**三部分。其中，语言组件作为生成式语言模型，能够在撰写、翻译、分析摘要等场景下提供AI辅助，帮助提升语音、文字处理的效率；代码组建主要用于支持软件开发；而图像组件则专注于生成和编辑图像。
- 三星Galaxy S24系列集成了Galaxy AI功能，海外版采用谷歌的Gemini，国内版采用百度的文心一言。三星Galaxy AI背后主要由Gemini提供支持，包括Galaxy S24搭载用于设备端的Gemini Nano，多项复杂功能则通过云端使用Gemini Pro处理计算任务。**本地AI模型深度集成在三星One UI 6.1中，不以具体的APP形态出现。**大部分涉及撰写、翻译、分析摘要、图片处理的原生应用，都能获得生成式AI技术的赋能，从而让手机AI不再被框定在聊天机器人的对话框当中。即圈即搜（Circle to Search）、实时翻译、笔记助手、图像处理等是Galaxy AI功能的应用亮点。

三星高斯生成式AI模型

Samsung Gauss

Samsung Research's Pioneering Generative AI Model



三星Galaxy S24的主打AI功能



目录

◆ 1 苹果AI布局

◆ 2 高通AI布局

◆ 3 三星AI布局

◆ 4 相关标的

◆ 5 风险提示

4.1 苹果 (AAPL.O)

- AI模型与手机/PC/穿戴设备/MR等整合，将给苹果产品线带来新一轮创新周期。基于庞大的用户基础、强大的生态系统、以及选择性提价的能力，苹果或成为未来端侧AI的最大受益者之一。
- iPhone 16或搭载本地AI模型，新Siri作为苹果的核心流量入口将得到AI加持，Mac/Vision等其他硬件产品有望未来陆续得到AI赋能。Vision Pro开启空间计算时代，多模态AI与Vision Pro的结合将创造全新生态，未来可期。这有利于推动苹果的硬件销售，以及提升软件和服务的价值空间。
- 预计24-25财年净利润为1020/1095亿美元。对应28xPE、26xPE，处于历史中枢水平。建议积极关注。

指标/年度↕	FY2022A↙	FY2023A↙	FY2024E↙	FY2025E↙	FY2026E↙
营业收入(百万美元)↕	394328.00	383285.00	389112.51	417906.14	438616.00
增长率↕	7.79%↙	-2.80%↙	1.52%↙	7.40%↙	4.96%↙
净利润(百万美元)↕	99803.00	96995.00	102046.91	109546.54	118615.53
增长率↕	5.41%↙	-2.81%↙	5.21%↙	7.35%↙	8.28%↙
每股收益 EPS↕	6.51	6.33	6.65	7.14	7.74
PE↕	28.12	28.94	27.51	25.62	23.66
PB↕	55.39	45.17	17.10	10.25	7.15

苹果在AI领域的布局

- 苹果今年加大对生成式AI的投入。预计将在iOS 18等OS中添加AI大模型能力，推出基于AI技术的新功能。更智能的Siri助手预计6月问世。苹果在本地和云端组合部署生成式AI功能。苹果将于6月举行WWDC 2024大会，AI布局将是本次活动的最大看点。
- 苹果正与OpenAI、谷歌和百度就AI模型上机进行合作洽谈。重点关注苹果自研的Ajax大模型的进展、以及相关资本支出对供应链的拉动。
- 苹果强于移动端的计算架构体系设计能力，且拥有从芯片到OS到应用的全栈能力，以及Siri等天然的应用场景，苹果理论上更有能力推出AI Agent 杀手级应用。



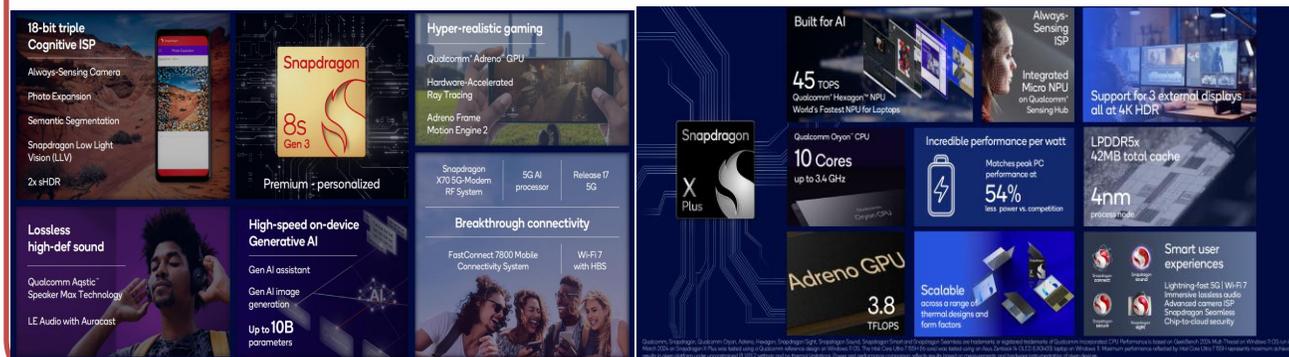
4.2 高通 (QCOM.O)

- **周期拐点将至，手机业绩即将复苏。**全球手机等消费电子需求初步稳定，消费电子周期拐点将至。随着智能手机市场复苏，公司作为安卓高端/旗舰机型的首选移动技术平台将直接受益，预计未来三年手机业务收入复合增速8.5%。
- **积极推动混合AI，AI硬件前景可期。**通过与生态伙伴合作，高通让百亿参数AI模型能在骁龙平台上运行，预计24H2开始AI手机和AI PC可看到高通的AI成果。公司积极推动混合AI战略，或成为终端生成式AI和移动计算的领军者。
- **预计24财年或迎来戴维斯双击。24年22xPE (GAAP)、20xPE (Non-GAAP)。**建议积极关注。

指标/年度	FY2022A	FY2023A	FY2024E	FY2025E	FY2026E
营业收入(百万美元)	44200.00	35820.00	39352.69	43518.57	46388.11
增长率	31.68%	-18.96%	9.86%	10.59%	6.59%
GAAP 净利润(百万美元)	12936.00	7232.00	10007.65	12227.52	13492.66
增长率	43.05%	-44.09%	38.38%	22.18%	10.35%
每股收益 EPS	11.62	6.50	8.97	10.96	12.09
Non-GAAP 净利润(百万美元)	14254.00	9486.00	11381.00	13244.00	14815.00
增长率	45.29%	-33.45%	19.98%	16.37%	11.87%
净资产收益率	71.81%	33.51%	31.68%	27.91%	23.54%
PE	10.28	18.39	20.03	16.40	14.86

高通在AI终端领域的布局

- 高通目前有4款主打终端本地运行的AI芯片，用于移动平台的骁龙8Gen3和8s Gen3、用于PC端的骁龙X Elite和骁龙X Plus。
 - ✓ 骁龙8Gen3采用4nm制程。相比上一代，8Gen3的NPU性能提升98%，CPU性能提升30%，GPU性能提升25%，整体功耗降低10%。8Gen3和8s Gen3成为支持终端上运行100亿参数模型的平台，并针对70亿参数LLM每秒生成20个tokens，各类复杂的AI应用可在搭载8Gen3和8s Gen3的手机上稳定运行。
 - ✓ 骁龙X Elite采用高通Oryon CPU，CPU性能为竞品两倍。算力为75 TOPS，支持在终端运行超130亿参数的生成式AI模型，针对70亿参数LLM每秒可生成30个tokens，AI处理速度为竞品4.5倍。近期高通推出AI PC品牌骁龙X Plus平台。X Plus采用10核CPU设计，GPU算力3.8 TFLOPS，NPU算力为45 TOPS。



目录

◆ 1 苹果AI布局

◆ 2 高通AI布局

◆ 3 三星AI布局

◆ 4 相关标的

◆ 5 风险提示

5 风险提示

- AI产业发展不及预期的风险；
- AI终端销售不及预期的风险；
- AI应用发展不及预期的风险。



分析师：王湘杰
执业证号：S1250521120002
电话：0755-26671517
邮箱：wxj@swsc.com.cn

分析师：杨镇宇
执业证号：S1250517090003
电话：023-67563924
邮箱：zyyu@swsc.com.cn

西南证券投资评级说明

报告中投资建议所涉及的评级分为公司评级和行业评级（另有说明的除外）。评级标准为报告发布日后6个月内的相对市场表现，即：以报告发布日后6个月内公司股价（或行业指数）相对同期相关证券市场代表性指数的涨跌幅作为基准。其中：A股市场以沪深300指数为基准，新三板市场以三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）为基准；香港市场以恒生指数为基准；美国市场以纳斯达克综合指数或标普500指数为基准。

公司评级	买入：未来6个月内，个股相对同期相关证券市场代表性指数涨幅在20%以上 持有：未来6个月内，个股相对同期相关证券市场代表性指数涨幅介于10%与20%之间 中性：未来6个月内，个股相对同期相关证券市场代表性指数涨幅介于-10%与10%之间 回避：未来6个月内，个股相对同期相关证券市场代表性指数涨幅介于-20%与-10%之间 卖出：未来6个月内，个股相对同期相关证券市场代表性指数涨幅在-20%以下
行业评级	强于大市：未来6个月内，行业整体回报高于同期相关证券市场代表性指数5%以上 跟随大市：未来6个月内，行业整体回报介于同期相关证券市场代表性指数-5%与5%之间 弱于大市：未来6个月内，行业整体回报低于同期相关证券市场代表性指数-5%以下

分析师承诺

报告署名分析师具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，报告所采用的数据均来自合法合规渠道，分析逻辑基于分析师的职业理解，通过合理判断得出结论，独立、客观地出具本报告。分析师承诺不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接获取任何形式的补偿。

重要声明

西南证券股份有限公司（以下简称“本公司”）具有中国证券监督管理委员会核准的证券投资咨询业务资格。

本公司与作者在自身所知情范围内，与本报告中所评价或推荐的证券不存在法律法规要求披露或采取限制、静默措施的利益冲突。

《证券期货投资者适当性管理办法》于2017年7月1日起正式实施，本报告仅供本公司签约客户使用，若您并非本公司签约客户，为控制投资风险，请取消接收、订阅或使用本报告中的任何信息。本公司也不会因接收人收到、阅读或关注自媒体推送本报告中的内容而视其为客户。本公司或关联机构可能会持有报告中提到的公司所发行的证券并进行交易，还可能为这些公司提供或争取提供投资银行或财务顾问服务。

本报告中的信息均来源于公开资料，本公司对这些信息的准确性、完整性或可靠性不作任何保证。本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，本报告所指的证券或投资标的的价格、价值及投资收入可升可跌，过往表现不应作为日后的表现依据。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告，本公司不保证本报告所含信息保持在最新状态。同时，本公司对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

本报告仅供参考之用，不构成出售或购买证券或其他投资标的的要约或邀请。在任何情况下，本报告中的信息和意见均不构成对任何个人的投资建议。投资者应结合自己的投资目标和财务状况自行判断是否采用本报告所载内容和信息并自行承担风险，本公司及雇员对投资者使用本报告及其内容而造成的一切后果不承担任何法律责任。

本报告及附录版权为西南证券所有，未经书面许可，任何机构和个人不得以任何形式翻版、复制和发布。如引用须注明出处为“西南证券”，且不得对本报告及附录进行有悖原意的引用、删节和修改。未经授权刊载或者转发本报告及附录的，本公司将保留向其追究法律责任的权利。



西南证券研究发展中心

西南证券研究发展中心

上海

地址：上海市浦东新区陆家嘴21世纪大厦10楼

邮编：200120

北京

地址：北京市西城区金融大街35号国际企业大厦A座8楼

邮编：100033

深圳

地址：深圳市福田区益田路6001号太平金融大厦22楼

邮编：518038

重庆

地址：重庆市江北区金沙门路32号西南证券总部大楼21楼

邮编：400025

西南证券机构销售团队

区域	姓名	职务	手机	邮箱	姓名	职务	手机	邮箱
上海	蒋诗烽	总经理助理、销售总监	18621310081	jsf@swsc.com.cn	张玉梅	销售经理	18957157330	zmyf@swsc.com.cn
	崔露文	销售副总监	15642960315	clw@swsc.com.cn	阚钰	销售经理	17275202601	kyu@swsc.com.cn
	谭世泽	高级销售经理	13122900886	tsz@swsc.com.cn	魏晓阳	销售经理	15026480118	wxyang@swsc.com.cn
	李煜	高级销售经理	18801732511	yfliyu@swsc.com.cn	欧若诗	销售经理	18223769969	ors@swsc.com.cn
	卞黎旻	高级销售经理	13262983309	bly@swsc.com.cn	李嘉隆	销售经理	15800507223	ljlong@swsc.com.cn
	田婧雯	高级销售经理	18817337408	tjw@swsc.com.cn	龚怡芸	销售经理	13524211935	gongyy@swsc.com.cn
北京	李杨	销售总监	18601139362	yfly@swsc.com.cn	张鑫	高级销售经理	15981953220	zhxin@swsc.com.cn
	张岚	销售副总监	18601241803	zhanglan@swsc.com.cn	王一菲	销售经理	18040060359	wyf@swsc.com.cn
	杨薇	资深销售经理	15652285702	yangwei@swsc.com.cn	王宇飞	销售经理	18500981866	wangyuf@swsc.com
	姚航	高级销售经理	15652026677	yhang@swsc.com.cn	路漫天	销售经理	18610741553	lmtyf@swsc.com.cn
	胡青璇	高级销售经理	18800123955	hqx@swsc.com.cn	马冰竹	销售经理	13126590325	mbz@swsc.com.cn
广深	郑龔	广深销售负责人	18825189744	zhengyan@swsc.com.cn	丁凡	销售经理	15559989681	dingfyf@swsc.com.cn
	杨新意	广深销售联席负责人	17628609919	yxy@swsc.com.cn	陈紫琳	销售经理	13266723634	chzlyf@swsc.com.cn
	张文锋	高级销售经理	13642639789	zwf@swsc.com.cn	陈韵然	销售经理	18208801355	cyryf@swsc.com.cn
	龚之涵	销售经理	15808001926	gongzh@swsc.com.cn				