PRODUCT
INTRODUCTION

产品概述

DS-35R(J)

DS-35R(J) AI 推理服务器 大模型一体机

DS-35R(J) 采用瑞芯微 RK3588/RK3588J 旗舰芯片，搭载八核 64 位处理器（4×A76+4×A55），主频高达 2.4GHz，内置 6 TOPS@INT8 独立 NPU，支持 8K 超高清视频编解码。标准 1U 机架式设计，支持 Ubuntu、Debian、麒麟等多类操作系统，满足边缘计算、智能安防、工业物联网等场景的多样化部署需求。

配备 2 路M.2 PCIe高速接口，支持灵活扩20TOPS/60TOPS/160TOPS@INT8 系列 M.2 算力模组，最大可扩展至支持 4 路算力模组并行，整机最高 **646 TOPS@INT8** 异构算力，单个算力模组可稳定承载7B~35B 大模型运行。

产品特性

强悍算力内核，端侧 AI 高效处理

- ✓ **AI 算力强劲：**集成 6 TOPS@INT8 算力 NPU 神经网络处理器，支持 INT4/INT8/FP16 多精度运算，轻松应对图像识别、实时视频分析、智能检测等端侧 AI 密集任务。
- ✓ **多核算力支撑：**搭载八核 CPU（4×A76+4×A55），为系统运行、业务逻辑与多任务并发处理提供稳定、流畅的算力保障。

灵活算力扩展，按需弹性升级

- ✓ **专业扩展架构：**独立散热与供电系统，保障高性能持续稳定输出；标配 2 路 M.2 PCIe 高速接口，最高可扩展至4路，带宽充足、扩展灵活。
- ✓ **多档算力可选：**支持 20TOPS/60TOPS/160TOPS@INT8 算力模组灵活选配，可按需升级，**异构算力最高可达 646TOPS@INT8**，满足从小规模到超高性能场景需求。

专业级视频处理，超高清流畅体验

- ✓ **超高清视频能力：**支持8K@60fps H.265、8K@30fps H.264视频解码，8K@30fps H.265/H.264视频编码，支持多路高清摄像头接入，实现超高清可视化呈现。
- ✓ **强悍图像性能：**搭载 32MP 专业 ISP，支持 HDR、3D 降噪，解码能力强劲，低照度、复杂光影环境下画面依旧清晰。

工业级可靠设计，严苛环境稳定运行

- ✓ **灵活部署：**机架式安装（Φ482mm×181.4mm×45mm），能无缝融入现有机房机柜，也能单独部署在小型弱电间，兼顾“分布式边缘”和“集中化管理”的需求。
- ✓ **丰富工业接口：**双千兆网口、RS-485/232、DI/DO 等接口齐全，可直连传感器、PLC、工业相机，快速完成系统集成部署。

全场景扩展能力，适配多元部署需求

- ✓ 预留4路M.2 (M.2 NGFF*1、M.2 4G/5G*1、M.2 PCIe*2)扩展接口，支持 4G/5G、Wi-Fi 、算力模组、SATA SSD及高速 NVMe SSD 存储扩展，兼顾无线联网、算力扩展、高速读写与大容量存储，满足智慧工厂、安防监控、智能交通、智慧城市等多场景落地。

产品规格

规格参数		DS-35R	DS-35RJ
核心配置	芯片平台	RK3588	RK3588J
	主控处理器	八核64位大小核架构，4*Cortex-A76 + 4*Cortex-A55@2.4GHz	
	主控算力	6 TOPS@INT8	
	扩展算力(选配)	20TOPS / 60TOPS / 160TOPS @INT8（单张，最大支持 4 张）	
	GPU	ARM Mali-G610 MC4 GPU, 专用2D图形加速模块	
	ISP	32MP ISP，支持HDR 和3DNR	
	内存	默认8G LPDDR5，可选 16G	
	存储	默认128G eMMC，可选 64G	
	扩展存储	M.2 SSD：支持NGFF SSD 2242,2260,2280/NVMe SSD 2280（机箱内部） TF卡槽：支持插入TF存储卡	
	视频编解码	解码：支持 H.265/H.264/AV1/VP9/AVS2 ，最高 8K@60fps（7680×4320 分辨率） 编码：支持 H.264/H.265 ，最高 8K@30fps（7680×4320 分辨率）	
基本参数	电源	72W交直流电源	
	散热方式	主动散热	
	工作温度	0℃~60℃	-20°C ~ 70°C
	工作湿度	50%～90%RH，无凝结	
	安装方式	机架式安装	
	结构尺寸	约482mm x 181.4mm x 45mm	
	重量	约6kg（不含选配件）	
接口参数	网络接口	RJ45 ×2：支持接入10/100/1000Mbps网络 WiFi：支持板载扩展 WiFi/蓝牙模组 4G/5G：支持通过M.2 B-KEY接口扩展4G/5G模组	
	视频输出	HDMI IN ×1（支持4K@60fps视频输入） HDMI OUT ×1（支持8K@60fps视频输出）	
	音频接口	Line_IN ×1：音频输入，Line_OUT ×1：音频输出（标准3.5mm音频接口）	
	其他接口	USB3.0 ×4 / SIM卡槽 ×1 / RS-485 ×2 / RS-232 ×1 / DI ×1 / DO ×2 / DEBUG ×1 / 对外供电（5V）×1	
	指示灯	PWR（电源）×1 / SYS（系统）×1 / HDD（硬盘状态）×1 / WWAN（4G/5G模组状态） ×1	
	可扩展接口	M.2（4G/5G）×1 / M.2（NGFF SSD）×1 / M.2（NVMe SSD/算力模组）×2 / WIFI+BT×1 / GPS&北斗双模定位	
软件配置	系统版本	Debian12 / Ubuntu 22.04 / 麒麟 可选	
	软件支持	支持TensorFlow / ONNX / Caffe / PyTorch / MxNet / DarkNet等多种深度学习框架；	

大模型算力扩展-M.2 算力模组可选配置

M.2模组

DS-35R(J) 支持2路（可扩展至4路）M.2 PCIe 标准接口扩展，可选配三款差异化算力模组，按需升级大模型算力；采用“通用主控+专用AI算力”灵活组合，覆盖从轻量化推理到工业级复杂任务的全场景需求。

系列型号	DS-35R(J)-R182X	DS-35R(J)-HM50	DS-35R(J)-DL20
算力模组 算力指标	20 TOPS@INT8	160 TOPS@INT8 100 FLOPS @bFP16	60 TOPS@INT8 120 TOPS@INT4 30TFLOPS@BF16/FP16
算力模组 核心架构	3D 堆叠 DRAM+8 核 NPU （近存计算）	SRAM-CIM 存算一体 （第二代 “天璇” IPU）	异构计算架构 （Minsky™平台）
核心定位	高性价比大模型协处理器	高能效中端大模型主力引擎	工业级高性能多模态算力标杆
显存容量	5GB DRAM	12GB/24GB/48GB LPDDR5 （192-bit 位宽）	8/16GB LPDDR5 （128-bit 位宽）
显存带宽	1TB/s+	153.6GB/s	102.4GB/s
适配模型参数	3B-7B LLM/VLM （Qwen2.5-3B、 DeepSeek-R1-Distill-7B 等）	7B-35B LLM/VLM （Qwen3.5、ChatGLM、 Llama2 等）	7B-13B LLM/VLM （DeepSeek、InternVL 等）
核心优势	同源适配，驱动零开发；成本可控，延迟低至 0.1s	能效比 16 T/W；支持多芯扩容，多系统兼容	工业级宽温（-20℃~70℃）；32 路视频解码；7x24 小时稳定运行
算力模组 典型功耗	≤15W （同声传译场景仅 6W）	≤15W	≤25W

核心价值

本地部署 · 低延迟 · 零API 成本

- **覆盖全量级模型：**从 1.5B 轻量化模型到 35B 大参数量模型，可按需搭配算力模组，适配不同复杂度场景需求。
- **主流模型全面适配：**支持 Gemma-2B、LlaMa2-7B、Qwen2.5、Qwen3、 Qwen3.5 等主流大模型。
- **极致多模态性能：**算力模组搭载专用 NPU+RK3588 ISP 视觉处理单元，支持文本、图像、视频一体化多模态推理，端到端延迟低至 0.1s。

模型类型	支持代表模型	推荐算力模组
轻量化 LLM（3B-7B）	Gemma-2B Qwen1.5-1.8B ChatGLM3-6B	20TOPS
主流 LLM（7B-13B）	LlaMa2-7B Qwen2.5-8B DeepSeek-7B	160TOPS 60TOPS
百亿参数 LLM（35B+）	Qwen3-30B Qwen3.5-35B-A3B	160TOPS
多模态 VLM	Qwen2.5-VL-1.5B~14B Qwen3-VL-2B~14B InternVL3、YOLO	20 TOPS 160TOPS

大模型官方性能指标（截止2026年1月，持续优化）

➢ DS-35R-R182X——LLM

模型名称	Input Tokens	New Tokens	TTFT(ms)	TPOT(ms)	Decode TPS
Qwen2.5-0.5B	128	128	21.89	4.63	215.86
Qwen2.5-1.5B	128	128	47.47	6.78	147.56
Qwen2.5-3B	128	128	83.44	9.8	102.01
Qwen2.5-7B	128	128	158.06	14.23	70.26
Qwen3-0.6B	128	128	27.53	5.58	179.33
Qwen3-1.7B	128	128	52.16	7.2	138.88
Qwen3-4B	128	128	106.7	11.42	87.56
Qwen3-8B	128	128	177.87	16.36	61.11

➢ DS-35R-R182X——VLM

模型名称	Vision 分辨率	Vision(ms)	LLM TTFT(ms)	LLM Decode TPS
FastVLM_1.5B_stage3	512 * 512	144.13	47.99	148.47
MiniCPM-3o	448 * 448	234.43	62.74	116.7
InternVL3-2B	448 * 448	190.8	47.93	148.26
InternVL3_5-4B	448 * 448	183.96	107.12	87.86
Qwen2.5-VL-3B	392 * 392	275.85	94.46	51.3
Qwen2.5-VL-3B	392 * 392	274.8	84.69	102.58
Qwen2.5-VL-7B	392 * 392	279.34	159.42	70.02
Qwen3-VL-2B	384 * 384	155.33	53.39	142.37
Qwen3-VL-4B	384 * 384	158.89	108.29	89.69
MiMo-VL-7B-RL	392 * 392	280.53	169.11	65.17
MiniCPM_V_4	448 * 448	237.55	94.94	106.62

大模型官方性能指标——HM50协处理器（截止2026年4月）

➢ DS-35R-HM50——VLM性能测试汇总

Model	Size	PrefillLen(k)	Ctx(k)	Batch	NChip	Input(k)	Output(k)	Prefill(tps)	Decode(tps)	Vision(fps)
Qwen2.5-VL	7B	0.25	8	1	1	0.25	0.05	673.25	23.94	5.19
Qwen2.5-VL	7B	0.25	8	1	1	0.5	0.05	673.73	23.91	5.22
Qwen2.5-VL	7B	0.25	8	1	1	1	0.05	667.65	23.31	5.18
Qwen2.5-VL	7B	0.25	8	1	1	2	0.05	667.88	22.79	5.26
Qwen2.5-VL	7B	0.25	8	1	1	4	0.05	654.72	21.67	5.18
Qwen2.5-VL	7B	0.25	8	1	1	7.95	0.05	627.87	20.19	5.22
DeepSeek-r1-Qwen3	8B	0.25	32	1	1	16	0.05	1241.86	16.43	0
Qwen3-VL	30b_a3b	0.25	8	1	1	0.25	0.05	1153.54	22.81	8.32
Qwen3-VL	30b_a3b	0.25	8	1	1	0.5	0.05	1157.04	21.08	8.34
Qwen3-VL	30b_a3b	0.25	8	1	1	1	0.05	1157.23	19.44	8.39
Qwen3-VL	30b_a3b	0.25	8	1	1	2	0.05	1072.27	17.09	8.37
Qwen3-VL	30b_a3b	0.25	8	1	1	4	0.05	948.24	13.88	8.35
Qwen3-VL	30b_a3b	0.25	8	1	1	7.95	0.05	746.73	10.44	8.34
Qwen3-VL	4B	0.25	32	1	1	0.25	0.05	2964.03	26.91	17.29
Qwen3-VL	4B	0.25	32	1	1	0.5	0.05	2933.48	26.84	17.21
Qwen3-VL	4B	0.25	32	1	1	1	0.05	2876.75	25.44	17.38
Qwen3-VL	4B	0.25	32	1	1	2	0.05	2657.84	24.75	17.3
Qwen3-VL	4B	0.25	32	1	1	4	0.05	2325.67	23.71	17.29
Qwen3-VL	4B	0.25	32	1	1	8	0.05	1865.69	21.71	17.31
Qwen3-VL	4B	0.25	32	1	1	16	0.05	1339.17	18.75	17.38
Qwen3-VL	4B	0.25	32	1	1	31.95	0.05	854.52	15.16	17.29
Qwen3-VL	8B	0.25	32	1	1	0.25	0.05	2167.69	19.1	11.08
Qwen3-VL	8B	0.25	32	1	1	0.5	0.05	2155.21	18.87	11.05
Qwen3-VL	8B	0.25	32	1	1	1	0.05	2121.52	18.06	11.02
Qwen3-VL	8B	0.25	32	1	1	2	0.05	1998.93	17.71	11.04
Qwen3-VL	8B	0.25	32	1	1	4	0.05	1803.98	17.17	11.04
Qwen3-VL	8B	0.25	32	1	1	8	0.05	1513.25	16.12	11.05
Qwen3-VL	8B	0.25	32	1	1	16	0.05	1147.42	14.33	10.99
Qwen3-VL	8B	0.25	32	1	1	31.95	0.05	772.38	12.19	11.05
Qwen3-VL	8B	0.25	8	1	1	0.25	0.05	2178.46	18.88	11.02
Qwen3-VL	8B	0.25	8	1	1	0.5	0.05	2163.91	18.76	11.06
Qwen3-VL	8B	0.25	8	1	1	1	0.05	2139.6	18.45	11.08
Qwen3-VL	8B	0.25	8	1	1	2	0.05	2013.25	17.78	11
Qwen3-VL	8B	0.25	8	1	1	4	0.05	1811.63	17.19	11.05
Qwen3-VL	8B	0.25	8	1	1	7.95	0.05	1511.12	16.82	11.02

➢ DS-35R-HM50——LLM性能测试汇总

Model	Size	PrefillLen(k)	Ctx(k)	Batch	NChip	Input(k)	Output(k)	Prefill(tps)	Decode(tps)	Vision(fps)
DeepSeek-r1-Qwen3	8B	0.25	32	1	1	0.25	0.05	2531.47	22.17	0
DeepSeek-r1-Qwen3	8B	0.25	32	1	1	0.5	0.05	2500.32	21.88	0
DeepSeek-r1-Qwen3	8B	0.25	32	1	1	1	0.05	2469.89	21.5	0
DeepSeek-r1-Qwen3	8B	0.25	32	1	1	2	0.05	2298.26	20.63	0
DeepSeek-r1-Qwen3	8B	0.25	32	1	1	4	0.05	2047.11	20.2	0
DeepSeek-r1-Qwen3	8B	0.25	32	1	1	8	0.05	1680.74	18.57	0
DeepSeek-r1-Qwen3	8B	0.25	32	1	1	16	0.05	1241.86	16.43	0
DeepSeek-r1-Qwen3	8B	0.25	32	1	1	31.95	0.05	814.07	13.38	0
DeepSeek-r1-Qwen3	8B	0.25	4	1	1	0.25	0.05	2533.73	22.42	0
DeepSeek-r1-Qwen3	8B	0.25	4	1	1	0.5	0.05	2517.83	22.15	0
DeepSeek-r1-Qwen3	8B	0.25	4	1	1	1	0.05	2475.4	21.12	0
DeepSeek-r1-Qwen3	8B	0.25	4	1	1	2	0.05	2313.39	20.95	0
DeepSeek-r1-Qwen3	8B	0.25	4	1	1	3.95	0.05	2026.01	20.89	0
DeepSeek-r1-Qwen3	8B	0.25	4	2	1	0.25	0.05	2540.26	37.75	0
DeepSeek-r1-Qwen3	8B	0.25	4	2	1	0.5	0.05	2516.1	37.01	0
DeepSeek-r1-Qwen3	8B	0.25	4	2	1	1	0.05	2475.2	34.92	0
DeepSeek-r1-Qwen3	8B	0.25	4	2	1	2	0.05	2313.12	33.95	0
DeepSeek-r1-Qwen3	8B	0.25	4	2	1	3.95	0.05	2027.99	34.35	0
Qwen2.5	7B	0.25	8	1	1	0.25	0.05	664.15	25.07	0
Qwen2.5	7B	0.25	8	1	1	0.5	0.05	666.3	25	0
Qwen2.5	7B	0.25	8	1	1	1	0.05	665.98	24.64	0
Qwen2.5	7B	0.25	8	1	1	2	0.05	659.23	24.04	0
Qwen2.5	7B	0.25	8	1	1	4	0.05	646.36	22.3	0
Qwen2.5	7B	0.25	8	1	1	7.95	0.05	620.84	21.24	0
Qwen3.5	2B	0.25	32	1	1	0.25	0.05	3897.47	65.02	0
Qwen3.5	2B	0.25	32	1	1	0.5	0.05	3857.63	64.78	0
Qwen3.5	2B	0.25	32	1	1	1	0.05	3853.94	62.87	0
Qwen3.5	2B	0.25	32	1	1	2	0.05	3834.09	61.86	0
Qwen3.5	2B	0.25	32	1	1	4	0.05	3776.79	59.62	0
Qwen3.5	2B	0.25	32	1	1	8	0.05	3711.53	56.82	0
Qwen3.5	2B	0.25	32	1	1	16	0.05	3574.24	50.61	0
Qwen3.5	2B	0.25	32	1	1	31.95	0.05	3299.95	42.06	0

➢ DS-35R-HM50——LLM性能测试汇总

Model	Size	PrefillLen(k)	Ctx(k)	Batch	NChip	Input(k)	Output(k)	Prefill(tps)	Decode(tps)	Vision(fps)
Qwen3.5	4B	0.25	32	1	1	0.25	0.05	1876.22	31.36	0
Qwen3.5	4B	0.25	32	1	1	0.5	0.05	1884.67	31.51	0
Qwen3.5	4B	0.25	32	1	1	1	0.05	1875.4	30.7	0
Qwen3.5	4B	0.25	32	1	1	2	0.05	1864.21	30.2	0
Qwen3.5	4B	0.25	32	1	1	4	0.05	1839.07	29.16	0
Qwen3.5	4B	0.25	32	1	1	8	0.05	1794.37	27.36	0
Qwen3.5	4B	0.25	32	1	1	16	0.05	1706.28	24.3	0
Qwen3.5	4B	0.25	32	1	1	31.95	0.05	1557.63	20.15	0
Qwen3.5	9B	0.25	32	1	1	0.25	0.05	1760.22	21.49	9.02
Qwen3.5	9B	0.25	32	1	1	0.5	0.05	1752.15	21.21	9.02
Qwen3.5	9B	0.25	32	1	1	1	0.05	1746.68	20.95	9.01
Qwen3.5	9B	0.25	32	1	1	2	0.05	1738.89	20.65	9.02
Qwen3.5	9B	0.25	32	1	1	4	0.05	1719.23	20.27	9.02
Qwen3.5	9B	0.25	32	1	1	8	0.05	1681.58	19.47	9.04
Qwen3.5	9B	0.25	32	1	1	16	0.05	1608.24	17.88	9.04
Qwen3.5	9B	0.25	32	1	1	31.95	0.05	1473.82	15.51	9.02
Qwen3	14B	0.25	16	1	1	0.25	0.05	1314.07	11.53	0
Qwen3	14B	0.25	16	1	1	0.5	0.05	1304.87	11.49	0
Qwen3	14B	0.25	16	1	1	1	0.05	1288.16	11.25	0
Qwen3	14B	0.25	16	1	1	2	0.05	1209.7	11.12	0
Qwen3	14B	0.25	16	1	1	4	0.05	1079.45	10.98	0
Qwen3	14B	0.25	16	1	1	8	0.05	891.91	10.42	0
Qwen3	14B	0.25	16	1	1	15.95	0.05	660.59	9.77	0
Qwen3	14B	0.25	32	1	2	0.25	0.05	1732.91	18.39	0
Qwen3	14B	0.25	32	1	2	0.5	0.05	1722.26	18.44	0
Qwen3	14B	0.25	32	1	2	1	0.05	1704.91	17.73	0
Qwen3	14B	0.25	32	1	2	2	0.05	1633.29	17.42	0
Qwen3	14B	0.25	32	1	2	4	0.05	1511.01	16.98	0
Qwen3	14B	0.25	32	1	2	8	0.05	1315.64	16.14	0
Qwen3	14B	0.25	32	1	2	16	0.05	1049.92	14.82	0
Qwen3	14B	0.25	32	1	2	31.95	0.05	747.2	12.91	0
Qwen3	14B	0.25	8	1	1	0.25	0.05	1333.3	11.51	0
Qwen3	14B	0.25	8	1	1	0.5	0.05	1325.83	11.49	0
Qwen3	14B	0.25	8	1	1	1	0.05	1311.85	11.31	0
Qwen3	14B	0.25	8	1	1	2	0.05	1226.2	11.27	0
Qwen3	14B	0.25	8	1	1	4	0.05	1088.32	10.79	0
Qwen3	14B	0.25	8	1	1	7.95	0.05	892.36	10.62	0

➢ DS-35R-HM50——LLM性能测试汇总

Model	Size	PrefillLen(k)	Ctx(k)	Batch	NChip	Input(k)	Output(k)	Prefill(tps)	Decode(tps)	Vision(fps)
Qwen3.5	35B-A3B	0.25	64	1	1	0.25	0.05	833.39	33.94	0
Qwen3.5	35B-A3B	0.25	64	1	1	0.5	0.05	861.22	34.08	0
Qwen3.5	35B-A3B	0.25	64	1	1	1	0.05	838.39	33.29	0
Qwen3.5	35B-A3B	0.25	64	1	1	2	0.05	813.24	32.79	0
Qwen3.5	35B-A3B	0.25	64	1	1	4	0.05	787.76	31.95	0
Qwen3.5	35B-A3B	0.25	64	1	1	8	0.05	800.57	30.78	0
Qwen3.5	35B-A3B	0.25	64	1	1	16	0.05	765.31	28.26	0
Qwen3.5	35B-A3B	0.25	64	1	1	32	0.05	738.46	24.06	0
Qwen3.5	35B-A3B	0.25	64	1	1	63.95	0.05	641.1	18.64	0
Qwen3	30b_a3b	0.25	128	1	1	0.25	0.05	1042.66	27.53	0
Qwen3	30b_a3b	0.25	128	1	1	0.5	0.05	1076.11	27.37	0
Qwen3	30b_a3b	0.25	128	1	1	1	0.05	1081.35	25.86	0
Qwen3	30b_a3b	0.25	128	1	1	2	0.05	1036.87	25.02	0
Qwen3	30b_a3b	0.25	128	1	1	4	0.05	981.76	24.03	0
Qwen3	30b_a3b	0.25	128	1	1	8	0.05	871.16	22.14	0
Qwen3	30b_a3b	0.25	128	1	1	16	0.05	698.94	19.14	0
Qwen3	30b_a3b	0.25	128	1	1	32	0.05	511.07	15.22	0
Qwen3	30b_a3b	0.25	128	1	1	64	0.05	332.04	10.85	0
Qwen3	30b_a3b	0.25	128	1	1	127.95	0.05	195.13	7.03	0
Qwen3	30b_a3b	0.25	256	1	2	0.25	0.05	1597.97	32.5	0
Qwen3	30b_a3b	0.25	256	1	2	0.5	0.05	1628.3	32.31	0
Qwen3	30b_a3b	0.25	256	1	2	1	0.05	1623.88	30.22	0
Qwen3	30b_a3b	0.25	256	1	2	2	0.05	1594.14	29.57	0
Qwen3	30b_a3b	0.25	256	1	2	4	0.05	1507.96	28.33	0
Qwen3	30b_a3b	0.25	256	1	2	8	0.05	1373.52	26.36	0
Qwen3	30b_a3b	0.25	256	1	2	16	0.05	1159.61	23.24	0
Qwen3	30b_a3b	0.25	256	1	2	32	0.05	883.22	18.77	0
Qwen3	30b_a3b	0.25	256	1	2	64	0.05	604.92	13.71	0
Qwen3	30b_a3b	0.25	256	1	2	128	0.05	369.56	8.87	0
Qwen3	30b_a3b	0.25	256	1	2	255.95	0.05	207.51	5.27	0
Qwen3	30b_a3b	0.25	32	1	1	0.25	0.05	1074.17	28.08	0
Qwen3	30b_a3b	0.25	32	1	1	0.5	0.05	1077.79	27.88	0
Qwen3	30b_a3b	0.25	32	1	1	1	0.05	1108.02	26.3	0
Qwen3	30b_a3b	0.25	32	1	1	2	0.05	1071.21	25.77	0
Qwen3	30b_a3b	0.25	32	1	1	4	0.05	990.17	24.35	0
Qwen3	30b_a3b	0.25	32	1	1	8	0.05	879	22.8	0
Qwen3	30b_a3b	0.25	32	1	1	16	0.05	713.99	19.88	0
Qwen3	30b_a3b	0.25	32	1	1	31.95	0.05	513.25	16.54	0

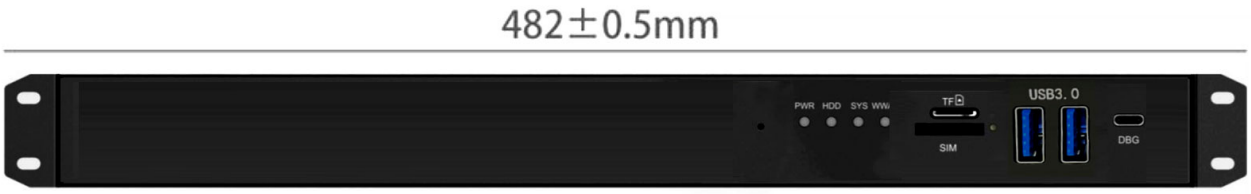
➢ DS-35R-HM50——LLM性能测试汇总

Model	Size	PrefillLen(k)	Ctx(k)	Batch	NChip	Input(k)	Output(k)	Prefill(tps)	Decode(tps)	Vision(fps)
Qwen3	32B	0.25	32	1	2	0.25	0.05	924.16	9.45	0
Qwen3	32B	0.25	32	1	2	0.5	0.05	918.35	9.33	0
Qwen3	32B	0.25	32	1	2	1	0.05	910.31	9.08	0
Qwen3	32B	0.25	32	1	2	2	0.05	871.87	8.95	0
Qwen3	32B	0.25	32	1	2	4	0.05	806.35	8.73	0
Qwen3	32B	0.25	32	1	2	8	0.05	705.06	8.47	0
Qwen3	32B	0.25	32	1	2	16	0.05	564.07	7.73	0
Qwen3	32B	0.25	32	1	2	31.95	0.05	402.83	6.88	0
Qwen3	4B	0.25	32	1	1	0.25	0.05	3258.35	20.96	0
Qwen3	4B	0.25	32	1	1	0.5	0.05	3227.25	20.79	0
Qwen3	4B	0.25	32	1	1	1	0.05	3158.23	20.04	0
Qwen3	4B	0.25	32	1	1	2	0.05	2897.2	19.57	0
Qwen3	4B	0.25	32	1	1	4	0.05	2505.56	18.99	0
Qwen3	4B	0.25	32	1	1	8	0.05	1977.62	17.68	0
Qwen3	4B	0.25	32	1	1	16	0.05	1392.88	15.54	0
Qwen3	4B	0.25	32	1	1	31.95	0.05	875.3	12.97	0
Qwen3	8B	0.25	16	4	1	0.25	0.05	2378.92	52.48	0
Qwen3	8B	0.25	16	4	1	0.5	0.05	2361.71	51.65	0
Qwen3	8B	0.25	16	4	1	1	0.05	2328.13	48.05	0
Qwen3	8B	0.25	16	4	1	2	0.05	2181.29	45.88	0
Qwen3	8B	0.25	16	4	1	4	0.05	1949.57	42.23	0
Qwen3	8B	0.25	16	4	1	8	0.05	1613.78	36.13	0
Qwen3	8B	0.25	16	4	1	15.95	0.05	1199.36	30.69	0
Qwen3	8B	0.25	32	1	1	0.25	0.05	2367.05	19.75	0
Qwen3	8B	0.25	32	1	1	0.5	0.05	2347.65	19.51	0
Qwen3	8B	0.25	32	1	1	1	0.05	2316.51	19.06	0
Qwen3	8B	0.25	32	1	1	2	0.05	2167.09	18.57	0
Qwen3	8B	0.25	32	1	1	4	0.05	1934.19	17.78	0
Qwen3	8B	0.25	32	1	1	8	0.05	1603.88	16.66	0
Qwen3	8B	0.25	32	1	1	16	0.05	1196.89	14.83	0
Qwen3	8B	0.25	32	1	1	31.95	0.05	792.84	12.59	0
Qwen3	8B	0.25	8	1	1	0.25	0.05	2379.2	19.67	0
Qwen3	8B	0.25	8	1	1	0.5	0.05	2361.17	19.48	0
Qwen3	8B	0.25	8	1	1	1	0.05	2323.55	18.84	0
Qwen3	8B	0.25	8	1	1	2	0.05	2178.92	18.42	0
Qwen3	8B	0.25	8	1	1	4	0.05	1949.5	18.01	0
Qwen3	8B	0.25	8	1	1	7.95	0.05	1605.18	17.56	0

➢ DS-35R-HM50——LLM性能测试汇总

Model	Size	PrefillLen(k)	Ctx(k)	Batch	NChip	Input(k)	Output(k)	Prefill(tps)	Decode(tps)	Vision(fps)
GPT-OSS	20B	0.25	256	1	1	0.25	0.05	1712.05	32.75	0
GPT-OSS	20B	0.25	256	1	1	0.5	0.05	1758.98	32.63	0
GPT-OSS	20B	0.25	256	1	1	1	0.05	1670.71	31.36	0
GPT-OSS	20B	0.25	256	1	1	2	0.05	1655.88	30.31	0
GPT-OSS	20B	0.25	256	1	1	4	0.05	1564.14	28.18	0
GPT-OSS	20B	0.25	256	1	1	8	0.05	1395.24	24.7	0
GPT-OSS	20B	0.25	256	1	1	16	0.05	1164.7	19.96	0
GPT-OSS	20B	0.25	256	1	1	32	0.05	855.46	14.42	0
GPT-OSS	20B	0.25	256	1	1	64	0.05	560.97	9.29	0
GPT-OSS	20B	0.25	256	1	1	128	0.05	332.09	5.42	0
GPT-OSS	20B	0.25	256	1	1	255.95	0.05	182.39	2.96	0
GPT-OSS	20B	0.25	256	1	2	0.25	0.05	2409.48	42.4	0
GPT-OSS	20B	0.25	256	1	2	0.5	0.05	2449.14	41.55	0
GPT-OSS	20B	0.25	256	1	2	1	0.05	2349	40.82	0
GPT-OSS	20B	0.25	256	1	2	2	0.05	2326.75	39.73	0
GPT-OSS	20B	0.25	256	1	2	4	0.05	2224.48	37.33	0
GPT-OSS	20B	0.25	256	1	2	8	0.05	2062.65	33.48	0
GPT-OSS	20B	0.25	256	1	2	16	0.05	1788.12	28.86	0
GPT-OSS	20B	0.25	256	1	2	32	0.05	1421.65	21.87	0
GPT-OSS	20B	0.25	256	1	2	64	0.05	995.33	15.03	0
GPT-OSS	20B	0.25	256	1	2	128	0.05	625.1	9.1	0
GPT-OSS	20B	0.25	256	1	2	255.95	0.05	357.4	5.14	0
GPT-OSS	20B	0.25	32	1	1	0.25	0.05	1761.73	32.88	0
GPT-OSS	20B	0.25	32	1	1	0.5	0.05	1610.45	32.76	0
GPT-OSS	20B	0.25	32	1	1	1	0.05	1488.58	31.43	0
GPT-OSS	20B	0.25	32	1	1	2	0.05	1423.69	30.25	0
GPT-OSS	20B	0.25	32	1	1	4	0.05	1353.17	28.21	0
GPT-OSS	20B	0.25	32	1	1	8	0.05	1218.64	24.76	0
GPT-OSS	20B	0.25	32	1	1	16	0.05	1020.48	20.02	0
GPT-OSS	20B	0.25	32	1	1	31.95	0.05	776.11	14.63	0
GPT-OSS	20B	0.25	32	1	2	0.25	0.05	2370.82	42.25	0
GPT-OSS	20B	0.25	32	1	2	0.5	0.05	2458.59	41.32	0
GPT-OSS	20B	0.25	32	1	2	1	0.05	2353.25	40.89	0
GPT-OSS	20B	0.25	32	1	2	2	0.05	2308.37	38.71	0
GPT-OSS	20B	0.25	32	1	2	4	0.05	2229.15	37.72	0
GPT-OSS	20B	0.25	32	1	2	8	0.05	2064.95	33.5	0
GPT-OSS	20B	0.25	32	1	2	16	0.05	1785.79	28.73	0
GPT-OSS	20B	0.25	32	1	2	31.95	0.05	1413.97	22.4	0
GPT-OSS	20B	0.25	64	1	1	0.25	0.05	1735.27	32.88	0
GPT-OSS	20B	0.25	64	1	1	0.5	0.05	1775.8	32.78	0
GPT-OSS	20B	0.25	64	1	1	1	0.05	1597.51	31.55	0
GPT-OSS	20B	0.25	64	1	1	2	0.05	1557.96	30.27	0
GPT-OSS	20B	0.25	64	1	1	4	0.05	1474.75	28.21	0
GPT-OSS	20B	0.25	64	1	1	8	0.05	1319.13	24.73	0
GPT-OSS	20B	0.25	64	1	1	16	0.05	1105.39	19.98	0
GPT-OSS	20B	0.25	64	1	1	32	0.05	823.46	14.42	0
GPT-OSS	20B	0.25	64	1	1	63.95	0.05	546.16	9.35	0
GPT-OSS	20B	0.25	64	1	2	0.25	0.05	2402.88	41.49	0
GPT-OSS	20B	0.25	64	1	2	0.5	0.05	2450.4	41.36	0
GPT-OSS	20B	0.25	64	1	2	1	0.05	2340.66	41.04	0
GPT-OSS	20B	0.25	64	1	2	2	0.05	2320.79	39.17	0
GPT-OSS	20B	0.25	64	1	2	4	0.05	2223.39	37.28	0
GPT-OSS	20B	0.25	64	1	2	8	0.05	2063.98	34.53	0
GPT-OSS	20B	0.25	64	1	2	16	0.05	1789.58	28.67	0
GPT-OSS	20B	0.25	64	1	2	32	0.05	1414.39	21.85	0
GPT-OSS	20B	0.25	64	1	2	63.95	0.05	995.48	15.08	0

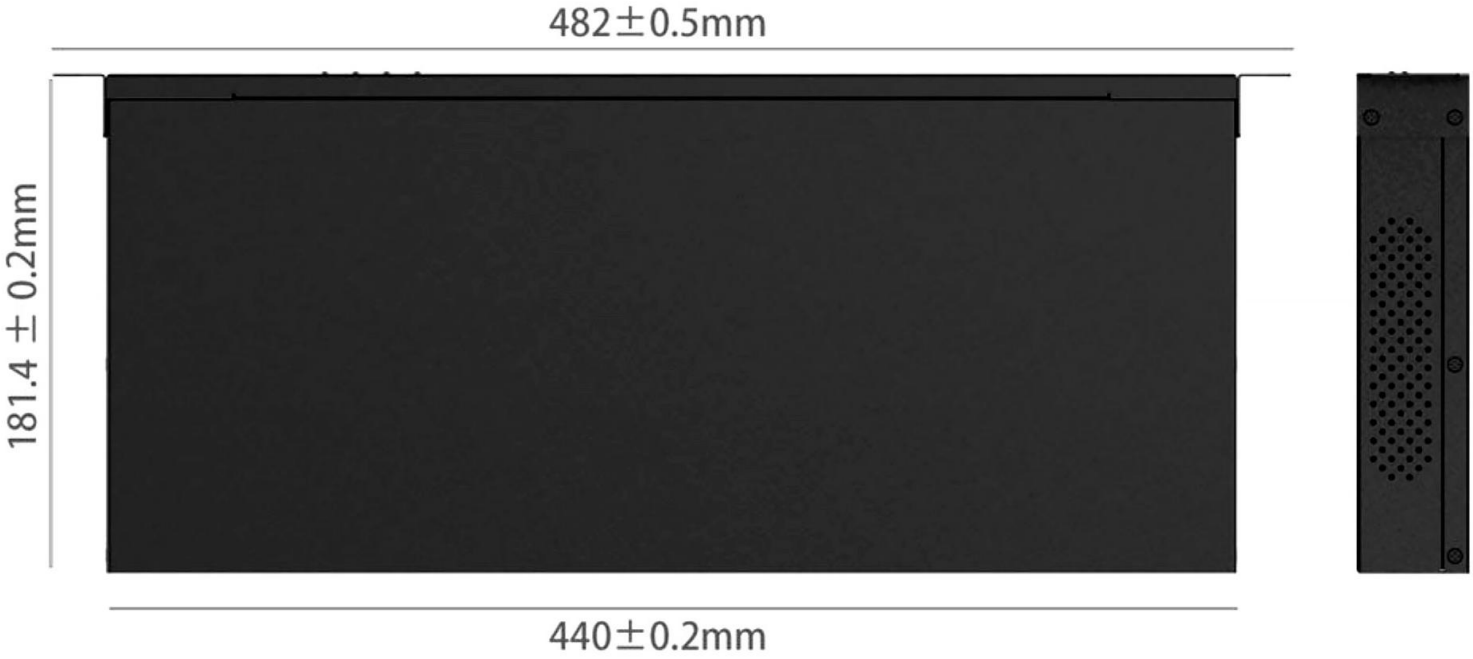
产品尺寸


482±0.5mm
PWR HDD SYS WWW
TF
SIM
USB3.0
DBG

前面板


LINE
IN OUT
USB3.0
HDMI
IN OUT
ETH0
ETH1
45±0.2mm

后面板


482±0.5mm
181.4±0.2mm
440±0.2mm

保修描述

保修范围

- 保修期限：主机保修1年（自购买日期起，具体以销售凭证为准）
- 免费服务：非人为故障，凭有效保修卡可免费维修
- 不保修范围：随机附件、人为损坏（如碰撞、电压异常、误操作等）及超出保修期的故障

不保修的情况

以下情形需支付维修费：

- 无法提供有效保修凭证，或凭证与产品不符
- 人为损坏（如带电插拔、运输磕碰、未按说明书操作等）
- 自行拆卸、改装或未经授权维修
- 自然灾害等不可抗力导致的故障

用户须知

- 配合义务：维修时需提供产品SN码及故障描述
- 数据安全：维修可能导致数据丢失，请提前备份，本公司不承担责任
- 更换部件：维修更换的故障部件归本公司所有

其他声明

- 本公司保留条款解释权及维修操作权限
- 超保或非保修故障可提供有偿服务，费用详询售后

注：购买后请核对保修信息，妥善保存凭证

官方公众号　官方抖音　官方淘宝

四川万物纵横科技股份有限公司

公司地址

成都市高新区天府五街花漾锦江B座7层

官方网站

http://www.iotdt.com

联系电话

191-1390-7060