RK1820/RK1828 RKNN3 SDK Release Note

文件标识：RK-JC-YF-429

发布版本：V1.0.0

日期：2026-01-23

文件密级：□绝密 □秘密 □内部资料 ■公开

瑞芯微电子股份有限公司

Rockchip Electronics Co., Ltd.

地址： 福建省福州市铜盘路软件园A区18号

网址： www.rock-chips.com

客户服务电话： +86-4007-700-590

客户服务传真： +86-591-83951833

客户服务邮箱： fae@rock-chips.com

读者对象

本文档（本指南）主要适用于以下工程师：

技术支持工程师

软件开发工程师

修订记录

版本	修改人	修改日期	修改说明	核定人
V0.2.0	HPC	2025-08-16	初始版本	熊伟
V0.3.0b0	HPC	2025-09-12	1. 更新支持的模型列表 2. 更新精度、性能数据	熊伟
V0.4.0b0	HPC	2025-11-03	1. 更新支持的模型列表 2. 更新概述、精度、性能数据	熊伟
V1.0.0	HPC	2026-01-23	1. 更新支持的模型列表 2. 更新精度、性能数据	熊伟

目录

1 概述

RKNN3 SDK提供了AI模型部署到RK1820/RK1828协处理器所需要的软件栈，包括PC端开发套件（RKNN3 Toolkit）、板端运行API（RKNN3 Runtime）以及模型转换部署示例（RKNN3 Model Zoo）等。本次发布的SDK支持RK1820/RK1828运行模式为协处理器模式，即主控 SoC 通过 PCIe/USB 高速接口连接 RK1820/RK1828 协处理器。

- **主控 SoC**: 作为系统核心，负责任务调度、资源分配和整体控制。
- **RK1820/RK1828协处理器**: 作为计算加速单元，专注高性能专用计算任务。
- **PCIe/USB 高速接口**: 实现低延迟、高带宽的数据交互。
- **支持的硬件平台**
  - RK3588/RK3576 + RK1820/RK1828协处理器
- **支持的系统**
  - Android/Linux

2 主要特性

- 大幅提升LLM/ViT性能，LLM Decode性能整体提升超过15%
- 扩展模型支持范围，新增适配 Qwen3-VL / Qwen2.5-Omni(Thinker) / GLM Edge / SmolVLM 等模型
- 支持连板精度分析
- 支持数据传输与推理并行
- 支持 mRoPE
- 支持 Function Call 功能
- 支持 YUV 格式输入
- rkllm3 server 新增支持 embedding 模型，并兼容音频输入
- 支持多核多模型同时推理
- 支持用户在协处理器上自定义模型后处理
- 优化exSDPA、exMatMul、Resize、Transpose等算子实现
- 提供RKNN3 Toolkit Lite工具包，支持在开发板上进行Python API调用

3 支持的模型

目前支持的模型列表如下：

模型名称	模型来源
Qwen2.5-0.5B	https://huggingface.co/Qwen/Qwen2.5-0.5B
Qwen2.5-3B	https://huggingface.co/Qwen/Qwen2.5-3B-Instruct
Qwen2.5-7B	https://huggingface.co/Qwen/Qwen2.5-7B-Instruct
Qwen3-0.6B	https://huggingface.co/Qwen/Qwen3-0.6B
Qwen3-1.7B	https://huggingface.co/Qwen/Qwen3-1.7B
Qwen3-4B	https://huggingface.co/Qwen/Qwen3-4B
Qwen3-8B	https://huggingface.co/Qwen/Qwen3-8B
HY-MT1.5-1.8B	https://huggingface.co/tencent/HY-MT1.5-1.8B
Youtu-LLM-2B	https://huggingface.co/tencent/Youtu-LLM-2B
FastVLM	https://github.com/apple/ml-fastvlm
Qwen2.5-VL-3B	https://huggingface.co/Qwen/Qwen2.5-VL-3B-Instruct
Qwen2.5-VL-7B	https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct
Qwen2.5-Omni-3B (Thinker)	https://huggingface.co/Qwen/Qwen2.5-Omni-3B
Qwen3-VL-2B	https://huggingface.co/Qwen/Qwen3-VL-2B-Instruct
Qwen3-VL-4B	https://huggingface.co/Qwen/Qwen3-VL-4B-Instruct
InternVL3-2B	https://huggingface.co/OpenGVLab/InternVL3-2B
InternVL3_5-4B	https://huggingface.co/OpenGVLab/InternVL3_5-4B-Instruct
MiMo-VL-7B-RL	https://huggingface.co/XiaomiMiMo/MiMo-VL-7B-RL
GLM-Edge-1.5B-Chat	https://modelscope.cn/models/ZhipuAI/glm-edge-1.5b-chat
SmolVLM-500M-Instruct	https://huggingface.co/HuggingFaceTB/SmolVLM-500M-Instruct
UI-TARS-2B-SFT	https://huggingface.co/ByteDance-Seed/UI-TARS-2B-SFT
gme-Qwen2-VL-2B-Instruct	https://huggingface.co/Alibaba-NLP/gme-Qwen2-VL-2B-Instruct
Siglip2-so400m	https://huggingface.co/google/siglip2-so400m-patch14-384
Dinov3	https://huggingface.co/facebook/dinov3-vits16-pretrain-lvd1689m
MobilenetV1	https://ftrg.zbox.filez.com/v2/delivery/data/95f00b0fc900458ba134f8b180b3f7a1/examples/mobilenet_v1/mobilenet_v1_1.0_224.tflite

模型名称	模型来源
MobilenetV2	https://ftrg.zbox.filez.com/v2/delivery/data/95f00b0fc900458ba134f8b180b3f7a1/examples/mobilenet/mobilenetv2-12.onnx
Resnet50V2	https://ftrg.zbox.filez.com/v2/delivery/data/95f00b0fc900458ba134f8b180b3f7a1/examples/resnet/resnet50-v2-7.onnx
YOLOv5s	https://ftrg.zbox.filez.com/v2/delivery/data/95f00b0fc900458ba134f8b180b3f7a1/examples/yolov5/yolov5s_rknn3.onnx
YOLOv6s	https://ftrg.zbox.filez.com/v2/delivery/data/95f00b0fc900458ba134f8b180b3f7a1/examples/yolov6/yolov6s_rknn3.onnx
YOLOv8s	https://ftrg.zbox.filez.com/v2/delivery/data/95f00b0fc900458ba134f8b180b3f7a1/examples/yolov8/yolov8s_rknn3.onnx
SenseVoiceSmall	https://modelscope.cn/models/iic/SenseVoiceSmall
Depth-Anything-V2-small	https://huggingface.co/depth-anything/Depth-Anything-V2-Small

用户可以从如下网盘获取预先转好的rknn模型。RKNN3_SDK (https://console.box.lenovo.com/l/H1fig1, 提取码：rknn)。具体路径如下：RKNN3_SDK/rknn3_models/v1.0.0。

4 模型性能

- LLM模型性能

模型名称	加速芯片	Input Tokens	New Tokens	TTFT(ms)	TPOT(ms)	Decode TPS
Qwen2.5-0.5B	RK182X	128	128	21.89	4.63	215.86
Qwen2.5-1.5B	RK182X	128	128	47.47	6.78	147.56
Qwen2.5-3B	RK182X	128	128	83.44	9.80	102.01
Qwen2.5-7B	RK1828	128	128	158.06	14.23	70.26
Qwen3-0.6B	RK182X	128	128	27.53	5.58	179.33
Qwen3-1.7B	RK1828	128	128	52.16	7.20	138.88
Qwen3-4B	RK1828	128	128	106.70	11.42	87.56
Qwen3-8B	RK1828	128	128	177.87	16.36	61.11

- VLM模型性能

模型	加速芯片	Vision 分辨率	Vision(ms)	LLM TTFT (ms)	LLM Decode TPS
FastVLM_1.5B_stage3	RK182X	512 * 512	144.13	47.99	148.47
MiniCPM-3o	RK182X	448 * 448	234.43	62.74	116.70
InternVL3-2B	RK182X	448 * 448	190.80	47.93	148.26
InternVL3_5-4B	RK1828	448 * 448	183.96	107.12	87.86
Qwen2.5-VL-3B	RK182X	392 * 392	275.85	94.46	51.30
Qwen2.5-VL-3B	RK1828	392 * 392	274.80	84.69	102.58
Qwen2.5-VL-7B	RK1828	392 * 392	279.34	159.42	70.02
Qwen3-VL-2B	RK182X	384 * 384	155.33	53.39	142.37
Qwen3-VL-4B	RK1828	384 * 384	158.89	108.29	89.69
MiMo-VL-7B-RL	RK1828	392 * 392	280.53	169.11	65.17
MiniCPM_V_4	RK1828	448 * 448	237.55	94.94	106.62

- 全模态模型

模型	加速芯片	Vision 分辨率	Vision(ms)	Audio(ms)	LLM TTFT (ms)	LLM Decode TPS
Qwen2.5-Omni-3B (Thinker)	RK1828	392 * 392	310.86	98.91	84.83	102.63

- CNN模型性能

模型名称	加速芯片	分辨率	单核性能（帧率）	多batch多核性能 （帧率）
MobilenetV1	RK182X	224 * 224	384.97	1505.06
MobilenetV2	RK182X	224 * 224	280.06	1319.91
Resnet50V2	RK182X	224 * 224	113.66	851.34
YOLOv5s	RK182X	640 * 640	35.41	212.65
YOLOv6s	RK182X	640 * 640	29.33	194.70
YOLOv8s	RK182X	640 * 640	32.07	210.73

注：

1. RK182X 表示加速芯片可为 RK1820 或 RK1828。
2. Qwen2.5-VL-3B：
    - 如果使用 RK1820 加速芯片，采用两段式运行方案（LMHead在RK3588上执行）；
    - 如果使用 RK1828 加速芯片，模型推理完全在协处理器端执行。
3. RK1820/RK1828协处理器NPU频率均为1GHZ。
4. 测试基于RK3588 + RK1820/RK1828，两者之间通过PCIe连接，RK3588使用performance模式。
5. TTFT：模型生成第一个 token所需的时间。
6. TPOT：生成每个输出 token所需的平均时间。
7. TPS：模型每秒能生成的 token数量。
8. VLM 的 Vision 和 LLM 耗时为独立测试，LLM部分的 Input Tokens 和 New Tokens 均设为128。

5 模型精度

- LLM模型精度

模型名称	加速芯片	数据集	原始模型 (float32)	RKNN3 模型 (W4A16 G32)
Qwen2.5-0.5B	RK182X	gsm8k	40.71	36.09
Qwen2.5-3B	RK182X	gsm8k	79.91	80.52
Qwen3-4B	RK1828	gsm8k	90.6	89.84

- VLM模型精度

模型名称	数据集	原始模型 (float32)	RKNN3 模型 (W4A16 G32)
FastVLM_1.6B	MMbench(cn)	58.42	60.48
Qwen2.5-VL-3B	MMbench(cn)	76.8	74.40
Qwen2.5-VL-7B	MMbench(cn)	79.98	81.44
InternVL3_2B	MMbench(cn)	77.23	72.77
InternVL3_5-4B	MMbench(cn)	78.69	72.42
mimo_vl_7b	MMbench(cn)	74.7	70.05
MiniCPM-3o	MMbench(cn)	68.99	69.67

- CNN模型精度

模型名称	数据集	原始模型 float32 （TOP1）	原始模型 float32 （TOP5）	RKNN3 模型 W8A8（TOP1）	RKNN3 模型 W8A8（TOP5）
MobilenetV1	imagenet	0.677	0.877	0.676	0.876
MobilenetV2	imagenet	0.694	0.888	0.680	0.881
Resnet50V2	imagenet	0.729	0.911	0.721	0.906

模型名称	数据集	原始模型 float32 AP@0.5:0.95	原始模型 float32 AP@0.5	RKNN3 模型 W8A8 AP@0.5:0.95	RKNN3 模型 W8A8 AP@0.5
Yolov5s	coco2017	0.326	0.481	0.314	0.474
Yolov6s	coco2017	0.403	0.551	0.386	0.533
Yolov8s	coco2017	0.39	0.525	0.383	0.517

- CNN模型性能测试方法

参考：**rknn3-runtime/examples/rknn3_model_test_demo/README_CN.md**

- LLM模型性能测试方法

参考：**rknn3-model-zoo/tools/rknn3_llm_test/README.md**

- CNN模型精度测试

目前rknn3-model-zoo集成了CNN模型精度相关的测试方法及代码，用户如果需要复测模型精度，可以参考其中的说明。

1. 分类模型参考: **rknn3-model-zoo/examples/mobilenet_v2/README.md**
2. 检测模型参考: **rknn3-model-zoo/examples/yolov8/README.md**

- LLM模型精度测试

rknn3-model-zoo也集成了LLM模型精度测试的方法及代码，支持CMMLU数据集。具体测试参考 **rknn3-model-zoo/tools/rknn3_llm_test/README.md**