RK1820/RK1828 RKNN3 SDK Release Note

Document ID: RK-JC-YF-429

Version: V1.0.0

Date: 2026-01-23

Confidentiality: □Top Secret □Secret □Internal ■Public

DISCLAIMER

THIS DOCUMENT IS PROVIDED “AS IS”. ROCKCHIP ELECTRONICS CO., LTD.(“ROCKCHIP”)DOES NOT PROVIDE ANY WARRANTY OF ANY KIND, EXPRESSED, IMPLIED OR OTHERWISE, WITH RESPECT TO THE ACCURACY, RELIABILITY, COMPLETENESS,MERCHANTABILITY, FITNESS FOR ANY PARTICULAR PURPOSE OR NON-INFRINGEMENT OF ANY REPRESENTATION, INFORMATION AND CONTENT IN THIS DOCUMENT. THIS DOCUMENT IS FOR REFERENCE ONLY. THIS DOCUMENT MAY BE UPDATED OR CHANGED WITHOUT ANY NOTICE AT ANY TIME DUE TO THE UPGRADES OF THE PRODUCT OR ANY OTHER REASONS.

Trademark Statement

"Rockchip", "瑞芯微", "瑞芯" shall be Rockchip’s registered trademarks and owned by Rockchip. All the other trademarks or registered trademarks mentioned in this document shall be owned by their respective owners.

Rockchip Electronics Co., Ltd.

No.18 Building, A District, No.89, software Boulevard Fuzhou, Fujian,PRC

Website: www.rock-chips.com

Customer service Tel: +86-4007-700-590

Customer service Fax: +86-591-83951833

Customer service e-Mail: fae@rock-chips.com

Target Audience

This document (this guide) is primarily intended for the following engineers:

- Technical Support Engineers
- Software Development Engineers

Revision History

Version	Author	Date	Description	Approved By
V0.2.0	HPC	2025-08-16	Initial version	Vincent
V0.3.0b0	HPC	2025-09-12	1. Updated the list of supported models 2. Updated accuracy and performance data	Vincent
v0.4.0b0	HPC	2025-11-03	1. Update the list of supported models 2. Update the overview; model accuracy, and performance data	Vincent
V1.0.0	HPC	2026-01-23	1. Updated the list of supported models 2. Updated accuracy and performance data	Vincent

Table of Contents

1 Overview

The RKNN3 SDK provides the software stack required to deploy AI models to the RK1820/RK1828 coprocessor, including a PC development kit (RKNN3 Toolkit), on-device runtime API (RKNN3 Runtime), and model conversion and deployment examples (RKNN3 Model Zoo). This SDK release supports the RK1820/RK1828 in coprocessor mode, where a host SoC is connected to the RK1820/RK1828 coprocessor via a high-speed PCIe/USB interface.

- **Host SoC**: Acts as the system's core, responsible for task scheduling, resource allocation, and overall control.
- **RK1820/RK1828 Coprocessor**: Serves as a computation acceleration unit, focusing on high-performance, specialized computing tasks.
- **PCIe/USB High-Speed Interface**: Enables low-latency, high-bandwidth data interaction.
- **Supported Hardware Platforms**
  - RK3588/RK3576 + RK1820/RK1828 Coprocessor
- **Supported Systems**
  - Android/Linux

2 Key Features and Enhancements

- Significantly improved LLM/ViT performance; overall LLM decode performance improved by more than 15%.
- Expanded model support range, adding models such as Qwen3-VL / Qwen2.5-Omni(Thinker) / GLM Edge / SmolVLM.
- Added support for cross-board accuracy analysis.
- Added support for overlapping data transfer and inference.
- Added support for mRoPE.
- Added support for Function Call.
- Added support for YUV-format input.
- `rkllm3-server` now supports embedding models and audio input.
- Added support for concurrent multi-core, multi-model inference.
- Added support for custom model post-processing on the coprocessor.
- Optimized implementation of exSDPA, exMatMul, Resize, Transpose operators.
- Provides RKNN3 Toolkit Lite package to support Python API calls on development boards.

3 Supported Models

The currently supported models are listed below:

Model Name	Model Source
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

Model Name	Model Source
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

Users can download pre-converted RKNN models from the following cloud drive: RKNN3_SDK (https://console.box.lenovo.com/l/H1fig1, Access Code: rknn). The specific path is as follows: `RKNN3_SDK/rknn3_models/v1.0.0`

4 Model Performance

This section shows the performance of typical LLM, VLM, full-modal and CNN models on the RK1820/RK1828 coprocessor.

- **LLM Model Performance**

Model Name	Accelerator	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 Model Performance**

Model	Accelerator	Vision Resolution	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

- **Full-Modal Model Performance**

Model Name	Accelerator	Resolution	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 Model Performance**

Model Name	Accelerator	Resolution	Single-Core Performance (fps)	Multi-Batch Multi-Core Performance (fps)
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

Note:

1. "RK182X" in the tables indicates that the accelerator chip can be either RK1820 or RK1828.
2. For `Qwen2.5-VL-3B`:

- When using RK1820 as the accelerator, a two-stage scheme is adopted (the LMHead runs on RK3588).
- When using RK1828 as the accelerator, the model runs entirely on the coprocessor.

3. The RK1820/RK1828 coprocessor NPU frequency is 1GHz.
4. Tests are based on an RK3588 + RK1820/RK1828 setup connected via PCIe, with the RK3588 set to performance mode.
5. TTFT: Time To First Token.
6. TPOT: Time Per Output Token.
7. TPS: Tokens Per Second.
8. The Vision and LLM timeouts of VLM were tested independently, and the Input Tokens and New Tokens of the LLM part were both set to 128.

5 Model Accuracy

- **LLM Model Accuracy**

Model Name	Accelerator	Dataset	Orig Model Acc (float32)	RKNN3 Acc (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 Model Accuracy**

Model Name	Dataset	Orig Model Acc (float32)	RKNN3 Acc (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 Model Accuracy**

Model Name	Dataset	Orig Model Float32 (TOP1)	Orig Model Float32 (TOP5)	RKNN3 W8A8 (TOP1)	RKNN3 W8A8 (TOP5)
MobilenetV1	imagenet	0.677	0.877	0.676	0.876
MobilenetV2	imagenet	0.694	0.888	0.680	0.882
Resnet50V2	imagenet	0.729	0.911	0.721	0.906

Model Name	Dataset	Orig Model Float32 AP@0.5:0.95	Orig Model Float32 AP@0.5	RKNN3 W8A8 AP@0.5:0.95	RKNN3 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

Note:

1. W4A16 G32 means that weights use 4-bit asymmetric quantization and activations use 16-bit floating point representation. Quantization parameters are assigned for every group of 32 weights along the input channel dimension.
2. W8A8 means that both weights and activations use 8-bit asymmetric quantization.

6 Recommended Server Configuration

The recommended server configurations and conversion time estimates are as follows:

Model Name	Recommended Server Configurations	Estimated Conversion Time
Qwen 2.5 0.5B	32-core CPU / 8 GB RAM / 1 TB SSD/HDD	~11 minutes
Qwen 2.5 1.5B	32-core CPU / 16 GB RAM / 1 TB SSD/HDD	~26 minutes
Qwen 2.5 3B	32-core CPU / 32 GB RAM / 1 TB SSD/HDD	~52 minutes
Qwen 2.5 7B	32-core CPU / 64 GB RAM / 1 TB SSD/HDD	~105 minutes

- If you encounter insufficient memory, you can try enabling a Swap partition to expand memory. See reference: https://wiki.debian.org/Swap.

7 References

7.1 Performance Testing Methods

- **CNN Model Performance Testing**

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

- **LLM Model Performance Testing**

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

7.2 Accuracy Testing Methods

- **CNN Model Accuracy Testing**

  The rknn3-model-zoo integrates testing methods and code for CNN model accuracy. Users who need to re-verify model accuracy can refer to the instructions shown below.

  1. Classification models: **rknn3-model-zoo/examples/mobilenet_v2/README.md**
  2. Detection models: **rknn3-model-zoo/examples/yolov8/README.md**

- **LLM Model Accuracy Testing**

  The rknn3-model-zoo also integrates methods and code for LLM model accuracy testing, with support for the CMMLU dataset. For specific testing procedures, refer to **rknn3-model-zoo/tools/rknn3_llm_test/README.md**