

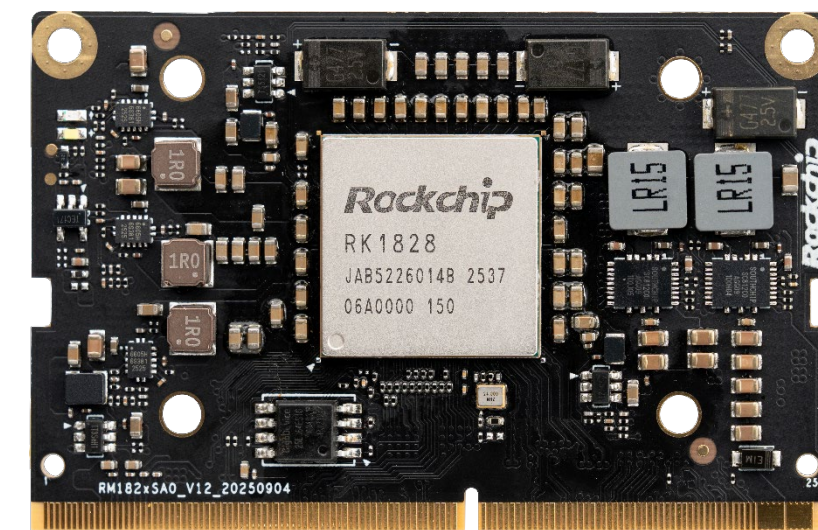


RK1828 Computing Card

- RM1828SA0 (SODIMM Module)
- RM1828MC0 (M.2 Module)

2026-10-29

FIREFLY TECHNOLOGY

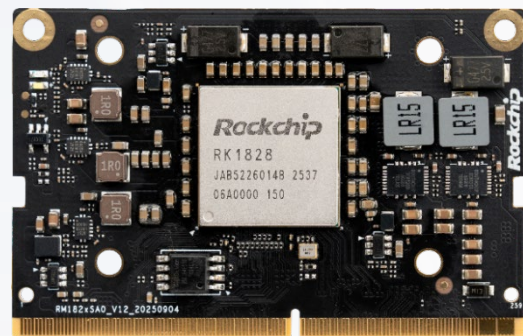


RM1828SA0



RM1828MC0

Product features



RM1828SA0



RM1828MC0



3D RAM Stack Architecture

RM1828SA0 / RM1828MC0 are AI inference modules for on-device LLMs. Built with 3D RAM stacking architecture and 5GB high-bandwidth DRAM to cut memory access overhead, they achieve over 100 tokens/s generation speed and end-to-end latency down to 0.1s for smooth low-latency local LLM interaction.



20 TOPS NPU Computing Power

NPU delivers up to 20 TOPS (INT8), supporting multi-precision computation of INT4/INT8/INT16. Compatible with CNN, RNN, LSTM and other classic networks, it suits inference of various vision and time-series algorithms.



Multi-Card Cascading Expansion

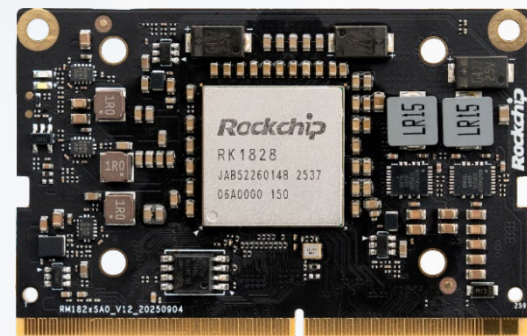
Supports flexible multi-card cascading expansion. A single card enables lightweight inference for 8B models with 61.11 TPS@8B LLM and 70 TPS@7B VLM. Dual cards support 12B models, quad cards for 27B models, and five cards for 35B MoE LLMs, delivering elastic computing power for LLM inference of different scales.



Comprehensive Model Support

Compatible with LLMs (Qwen, GLM Edge), vision models (SigLIP, DINO), CNN (YOLO) and VLMs (MiniCPM, InternVL3). It supports mainstream AI algorithms for diverse on-device inference development.

Product features



RM1828SA0



RM1828MC0



Supports Agent and Tool Calling

Beyond basic Q&A capabilities, the local LLM can serve as an Agent inference core. It supports integration with file readers, Shell and compilation tools, local code repositories, MCP services, document parsers and knowledge bases, and can call internal enterprise HTTP APIs. Combined with the OpenCode Code Agent, it accomplishes code development tasks.



Low-latency Edge Computing with Millisecond-level Response

Local AI processing eliminates cloud upload and network dependency, avoiding bandwidth fluctuation impacts to deliver millisecond response and robust data privacy. All business data is stored and computed locally to comply with data security and industry privacy regulations.



SODIMM / M.2 Standard Package

The module offers SODIMM and M.2 B+M-KEY interfaces. The SODIMM version is pin-compatible with NVIDIA Orin Nano, with strong interoperability.



Wide Range of Application Scenarios

Widely applied in robots, drones, AI edge computing, industrial automation, intelligent security, smart driving, traffic management, smart cities, home storage, retail automation and other products and fields.

Specifications



		RM1828SA0 (SODIMM Module)	RM1828MC0 (M.2 Module)
Basic Specifications	SOC	Rockchip RK1828	
	RGA	Built-in 2D graphics engine; supports BT.601/BT.709 pixel format conversion and dithering operation. Max resolution: source image 8192×8192, target image 4096×4096	
	NPU	NPU computing power: 20 TOPS (INT8); supports mixed-precision computation of INT4/INT8/INT16/FP8/FP16/BF16	
	Codec	JPEG Encoding: Supports baseline (sequential DCT), JFIF 1.02. Image size range: 16×16 ~ 65520×65520; supports YUV400/YUV420/YUV422/YUV444 JPEG Decoding: Supports baseline (sequential DCT), JFIF 1.02. Image size range: 48×48 ~ 65520×65520; supports YUV400/YUV420/YUV422/YUV440/YUV411/YUV444/RG888/RGB565	
	RAM	On-chip 5 GB DRAM	
	Power Supply	DC 5V	DC 12V
	OS	Linux	
	Software Support	Supports YOLO object-detection algorithms. Supports private deployment of large-scale Transformer-based models, such as Gemma, ChatGLM, Qwen, Phi and other LLM series. Supports classic network architectures including CNN, RNN and LSTM.	
	Interface Type	Gold-finger (SODIMM 260-Pin standard interface, 0.5 mm pitch)	M.2 B-KEY / M.2 M-KEY
	Dimensions	69.6mm × 45.0mm × 4.85mm (without heat sink) 69.6mm × 45.0mm × 20.67mm (with heat sink)	22mm × 80mm × 6mm (without heat sink) 25mm × 81mm × 23.05mm (with heat sink)
	Weight	21g (without heat sink), 75g (with heat sink)	11g (without heat sink), 53g (with heat sink)
Interface Specifications	Environment	Operating temperature: -20 °C ~ 60 °C, Storage temperature: -20 °C ~ 70 °C, Storage humidity: 10% ~ 90%RH (non-condensing)	
	Fan Interface	N/A	1 × FAN (4Pin-1.25mm)



Model Performance

RK182X LLM Running Performance

Model	Chip	TTFT(ms)	TPOT(ms)	Decode TPS
Qwen2.5-3B	RK182X	83.44	9.80	102.01
Qwen3-1.7B	RK1828	52.16	7.20	138.88
Qwen3-8B	RK1828	177.87	16.36	61.11

* Data based on RKNN3 V1.0 SDK, all models quantized with W4A16 G32.

RK182X VLM Running Performance

Model	Chip	Vision Resolution	Vision(ms)	LLM TTFT(ms)	LLM Decode TPS
FastVLM_1.5B_stage3	RK182X	512 * 512	144.13	47.99	148.47
InternVL3-2B	RK182X	448 * 448	190.80	47.93	148.26
Qwen2.5-VL-7B	RK1828	392 * 392	279.34	159.42	70.02
Qwen3-VL-4B	RK1828	384 * 384	158.89	108.29	89.69
MiniCPM-V_4	RK1828	448 * 448	237.55	94.94	106.62

* Data based on RKNN3 V1.0 SDK, all models quantized with W4A16 G32.

RK182X CNN Model Running Performance

Model	Resolution	Single-core Performance (FPS)	Multi-core Performance (FPS)
YOLOv5s	640*640	35.4	212.6
ResNet50V2	224*224	113.7	851.3
DINOv3	224*224	54.8	331.0
DepthAnythingV2	448*252	/	27.3

* Data based on RKNN3 V1.0 SDK, all models quantized with W8A8.

RK1828 Multimodal Model Running Performance

Model	Chip	Vision 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

* Data based on RKNN3 V1.0 SDK, all models quantized with W4A16 G32.

Model Performance



Multi-Card Model Performance

Test Conditions: RK3588 Host · RK1828 SO-DIMM · Model Quantization: W4A16

Model	Chip	Input	Output	Interface	TTFT (s)	Prefill TPS	Decode TPS
Qwen3.5-9B	2x RK1828	4K	128	PCIe	4.80	853.7	33.19
		4K	128	USB3.0	4.90	836.0	33.52
		8K	128	PCIe	9.73	841.7	32.33
		8K	128	USB3.0	9.94	824.0	32.65
Qwen3.8-27B	4x RK1828	4K	128	PCIe	6.92	592.1	13.02
		4K	128	USB3.0	7.15	573.2	13.02
		8K	128	PCIe	13.67	599.4	12.74
		8K	128	USB3.0	14.08	581.7	12.83

Note: Under current 4K/8K actual tests, the PCIe interface delivers an approximate 2%-3% advantage in TTFT and Prefill.

Host System Overhead

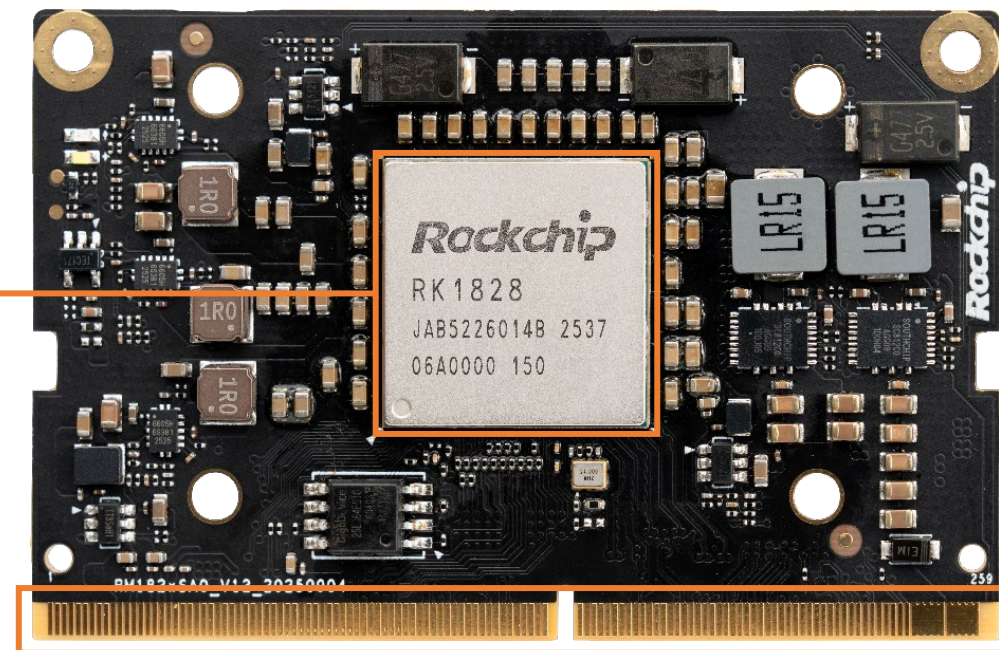
Test Conditions: RK3588 Host · RK1828 SO-DIMM · Based on 4-core A76 CPU running at 100% full load

Model	Chip	Interface	Avg. CPU	Avg. Memory	Peak Memory
Qwen3.5-9B	2x RK1828	PCIe	2.4%	179 MB	310 MB
		USB3.0	2.0%	181 MB	312 MB
Qwen3.8-27B	4x RK1828	PCIe	2.6%	191 MB	524 MB
		USB3.0	2.2%	187 MB	523 MB

Interface description



Rockchip RK1828
(Up to 850MHz)



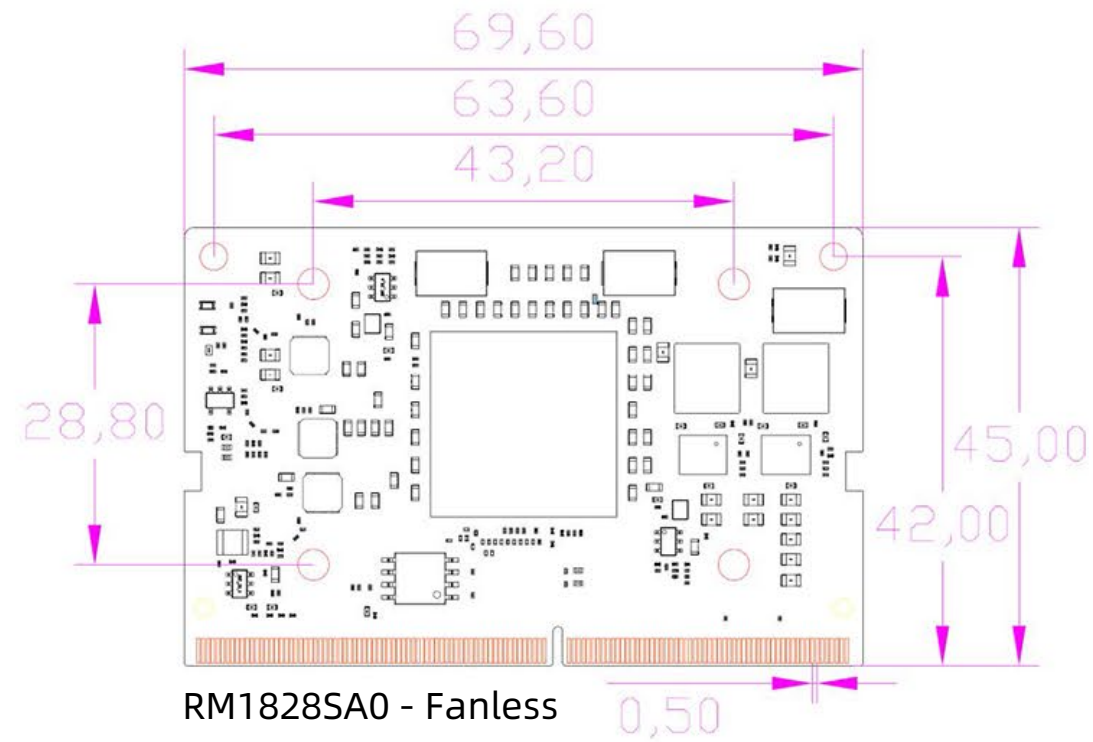
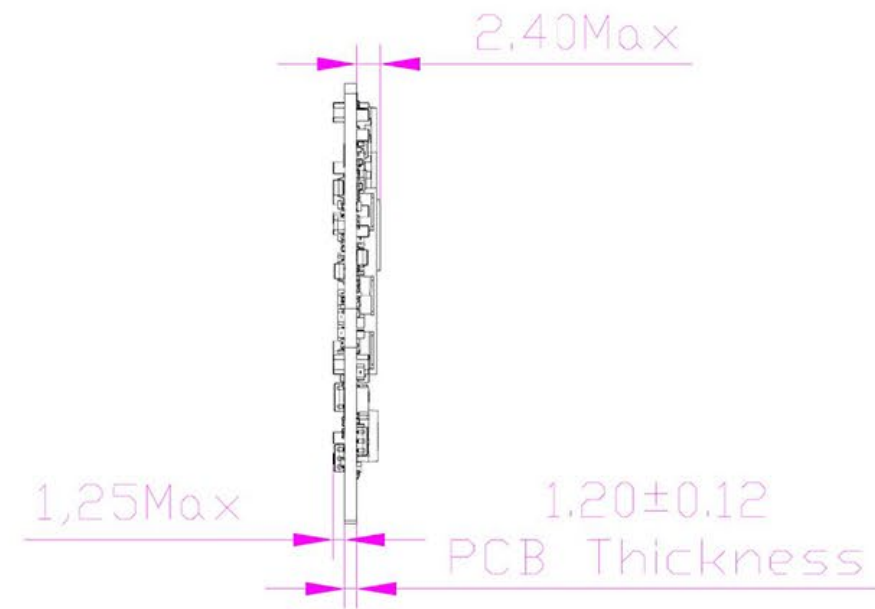
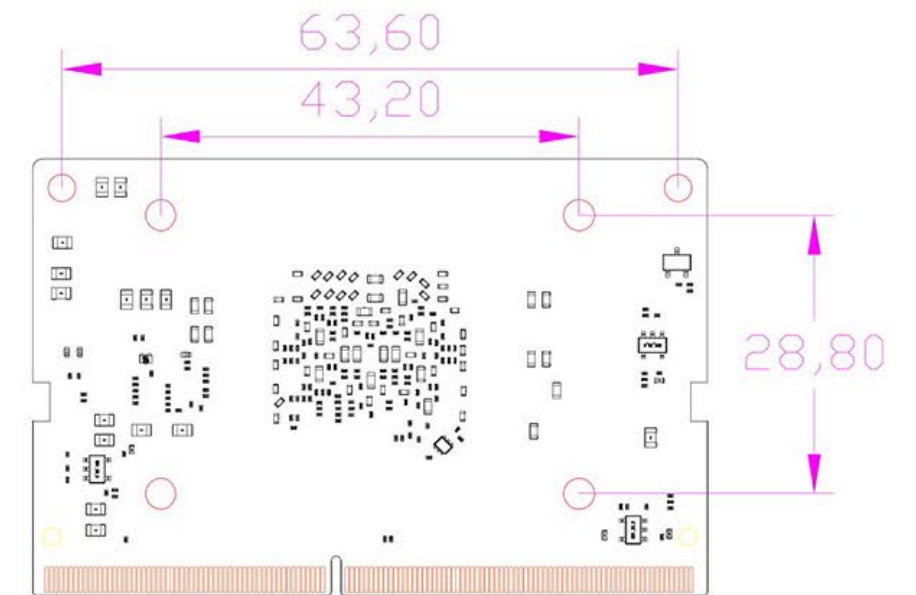
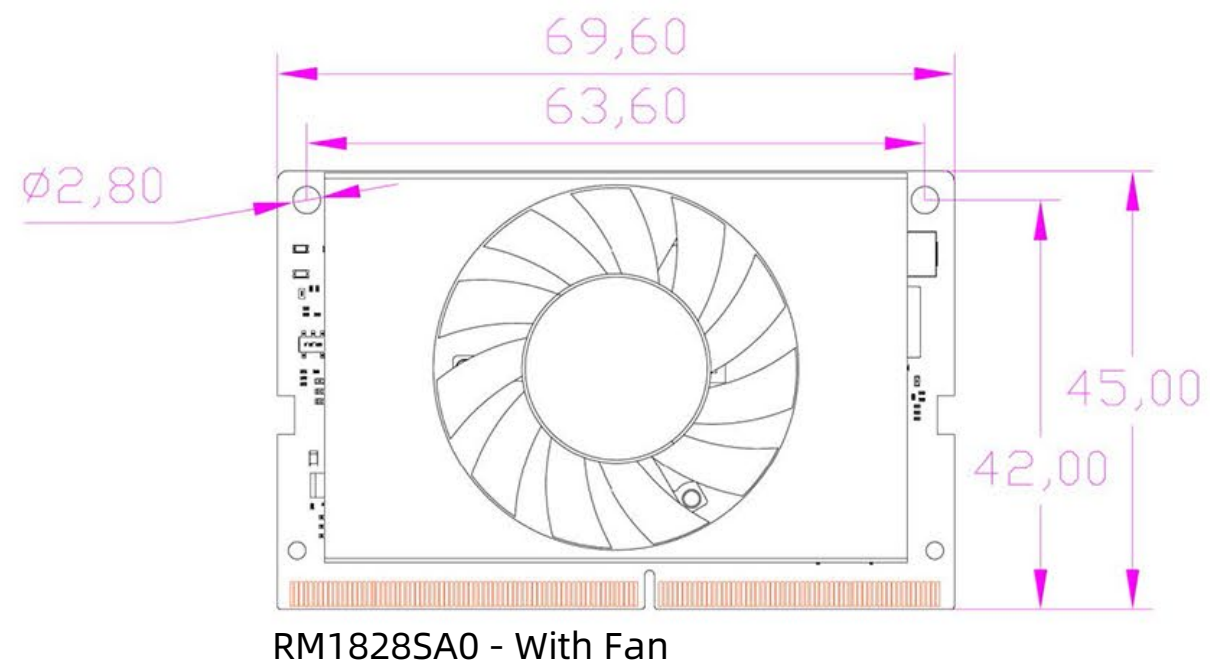
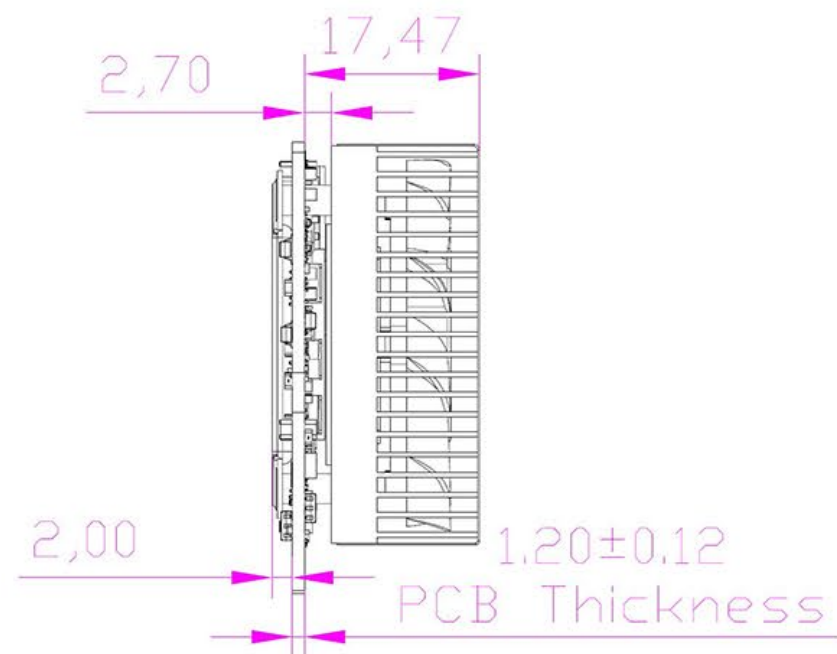
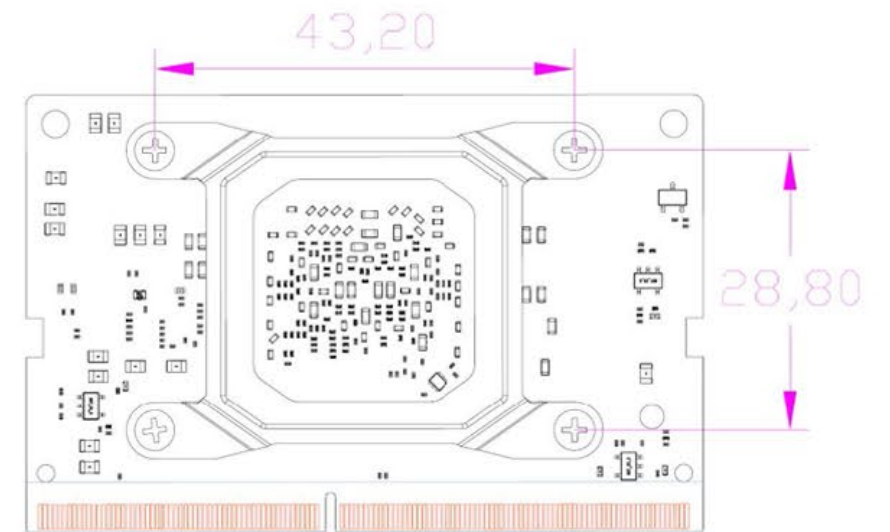
Gold Finger
(260-Pin, SODIMM, 0.5mm Pitch)

Rockchip RK1828
(Up to 850MHz)



M.2 B-KEY / M.2 M-KEY

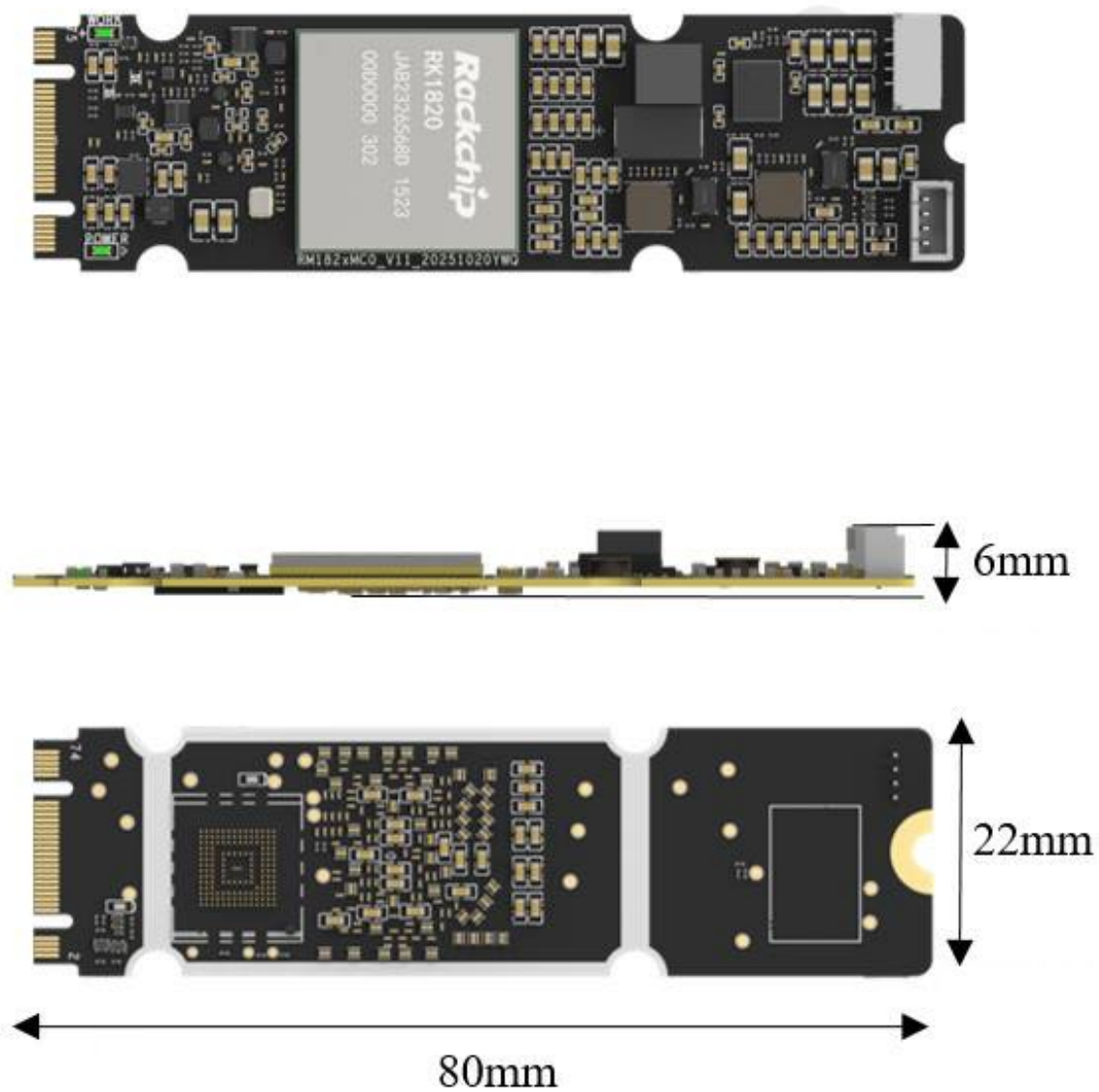
Dimension



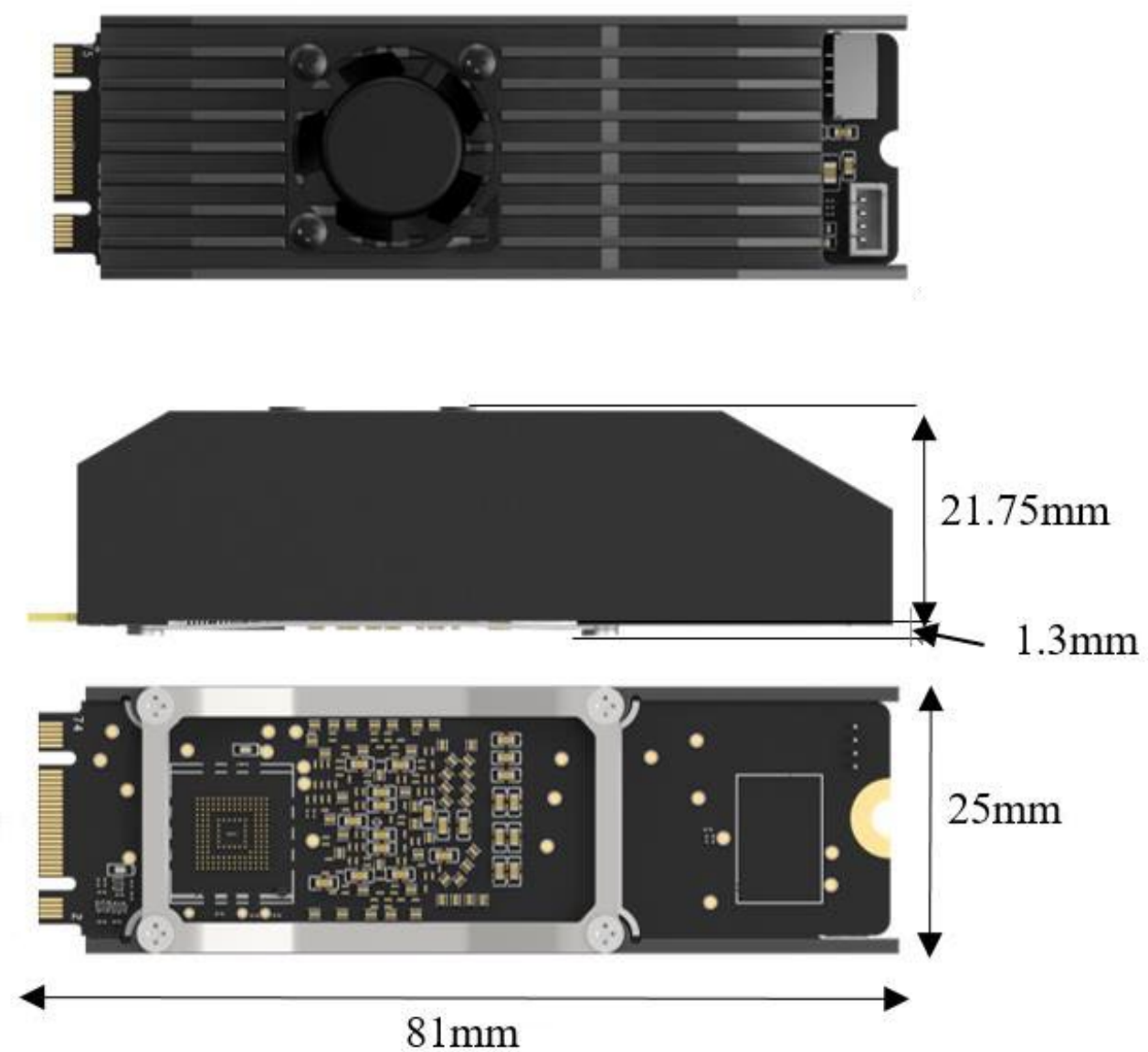
Dimension



RM1828MC0 - Fanless



RM1828MC0 - With Fan





FIREFLY TECHNOLOGY



Contact Us
(+86)18688117175



E-mail
global@t-firefly.com



Website
<https://en.t-firefly.com/>



Address
Room 2101, Hongyu Building, #57 Zhongshan 4Rd, East District,
Zhongshan, Guangdong, China.