GPUpartner

VALIDATION REPORT · IMAGE GENERATION

FLUX.2 [dev]

on Ryzen AI Max+ 395 64 GB · 5-bit (GGUF Q5_K_M)

STATUS · NOV 22, 2026

Not validated

1 of 3 checks pass

FLUX.2 [dev] at 5-bit runs on the Ryzen AI Max+ 395 64 GB, about 8.4 min per image.

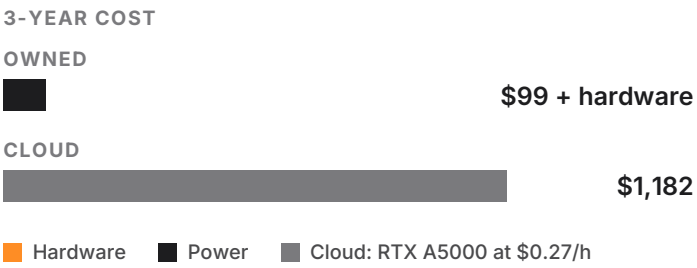
VALIDATION

CHECK	TARGET	RESULT	STATUS
Fits in GPU memory	48.0 GB usable	46.9 GB needed	PASS
Headroom	5% free	1.1 GB free (2%)	CAUTION
Time per image	under 30 s	8.4 min	FAIL

Targets: an image in under 30 s, and 5% of memory left free. Calculated from published specs, not benchmarked.

BUDGET

HARDWARE, ONE-TIME	MSRP not entered
POWER, PER YEAR	$33 · 126 W at the wall
CLOUD INSTEAD, PER YEAR	$394
3-YEAR TOTAL, OWNED	Hardware + $99 power
3-YEAR TOTAL, CLOUD	$1,182
BREAK-EVEN	Needs the hardware MSRP



Budget basis: 4 h a day, power at $0.18 per kWh, cloud at RunPod on-demand prices as of Nov 22, 2026. Hardware at the MSRP entered in the calculator, not a GPU Partner quote. A 3-year view; resale value, staff time and taxes aren't counted.

GPUpartner

TECHNICAL DETAIL · IMAGE GENERATION

FLUX.2 [dev]

on Ryzen AI Max+ 395 64 GB · 5-bit (GGUF Q5_K_M)

SETUP

PARAMETERS	32.2B
PRECISION	5-bit (GGUF Q5_K_M)
OUTPUT	1024 × 1024
STEPS	28
HARDWARE	Ryzen AI Max+ 395 64 GB
GPU CAN USE	48.0 GB of 64.0 GB

MEMORY

WEIGHTS	43.7 GB
WORKING	2.3 GB
RUNTIME	819 MB
SYSTEM	16.0 GB
FREE	1.1 GB

SUGGESTED CHANGES

4-BIT WEIGHTS	Fits, 8.4 min an image
768² OUTPUT	Fits, 5.0 min an image
RTX 4090	Runs it, 1.8 min an image
M5 ULTRA · 96 GB	Runs it, 3.4 min an image

SAME MODEL, OTHER MACHINES

RTX 4090	5-bit · 1.8 min
RTX 5090	6-bit · 85 s
RTX PRO 6000 BLACKWELL	16-bit · 31 s
M5 MAX 128 GB	16-bit · 4.8 min
M3 ULTRA 256 GB	16-bit · 9.3 min
DGX SPARK 128 GB	16-bit · 4.1 min
DGX STATION GB300	16-bit · 7.1 s

WHERE TO GET IT

FP8	huggingface.co/Comfy-Org/flux2-dev
NVFP4	huggingface.co/black-forest-labs/FLUX.2-dev-NVFP4
GGUF	huggingface.co/city96/FLUX.2-dev-gguf
OFFICIAL	huggingface.co/black-forest-labs/FLUX.2-dev

METHOD AND ASSUMPTIONS

Memory is the denoiser's weights at the chosen precision plus the bigger of one transformer block's working set and the VAE decode, plus the runtime; the text encoders run first and then wait in system RAM. Time is the denoising FLOPs per step times the steps, against the chip's dense 16-bit TFLOPS, scaled to published GeForce runs of this model. Running cost is power at the wall at your rate and hours; the cloud figure is the cheapest RunPod on-demand GPU that runs the same job. Accuracy: memory lands within about a gigabyte; speeds and times are rough, about ±25%. Model data and hardware specs checked Nov 2026; cloud prices are RunPod's on-demand list as of Nov 22, 2026. Report VR-0GI2TQI, generated November 22, 2026.