GPUpartner

VALIDATION REPORT · IMAGE GENERATION

MOSS-Image

on M4 Pro 24 GB · 8-bit (FP8 · Q8_0)

STATUS · NOV 22, 2026

Not validated

1 of 3 checks pass

MOSS-Image at 8-bit runs on the M4 Pro 24 GB, about 21 min per image.

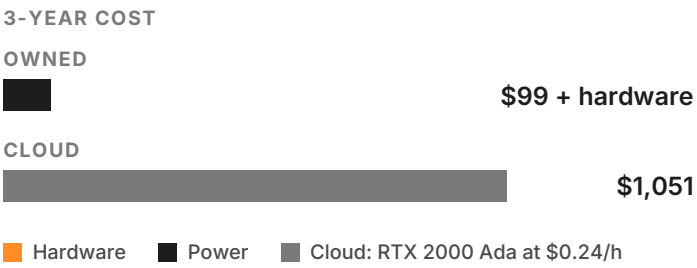
VALIDATION

CHECK	TARGET	RESULT	STATUS
Fits in GPU memory	16.0 GB usable	15.9 GB needed	PASS
Headroom	5% free	135 MB free (1%)	CAUTION
Time per image	under 30 s	21 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	$350
3-YEAR TOTAL, OWNED	Hardware + $99 power
3-YEAR TOTAL, CLOUD	$1,051
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

MOSS-Image

on M4 Pro 24 GB · 8-bit (FP8 · Q8_0)

SETUP

PARAMETERS	5.1B
PRECISION	8-bit (FP8 · Q8_0)
OUTPUT	1024 × 1024
STEPS	50
HARDWARE	M4 Pro 24 GB
GPU CAN USE	16.0 GB of 24.0 GB

MEMORY

WEIGHTS	12.7 GB
WORKING	2.3 GB
RUNTIME	819 MB
SYSTEM	8.0 GB
FREE	135 MB

SUGGESTED CHANGES

5-BIT WEIGHTS	Fits, 24 min an image
768² OUTPUT	Fits, 12 min an image
RTX PRO 2000 BLACKWELL	Runs it, 2.4 min an image
M5 MAX · 36 GB	Runs it, 3.0 min an image

SAME MODEL, OTHER MACHINES

RTX 4090	16-bit · 59 s
RTX 5090	16-bit · 47 s
RTX PRO 6000 BLACKWELL	16-bit · 20 s
M5 MAX 128 GB	16-bit · 3.0 min
M3 ULTRA 256 GB	16-bit · 5.9 min
DGX SPARK 128 GB	16-bit · 2.6 min
RYZEN AI MAX+ 395 128 GB	16-bit · 4.6 min

WHERE TO GET IT

GGUF	huggingface.co/realrebelai/MOSS-Image_GGUFs
OFFICIAL	huggingface.co/OpenMOSS-Team/MOSS-Image-SFT

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-1Q6S4JW, generated November 22, 2026.