GPUpartner

VALIDATION REPORT · IMAGE GENERATION

MOSS-Image

on RTX PRO 6000 Blackwell · 16-bit (BF16)

STATUS · OCT 23, 2026

Validated

3 of 3 checks pass

MOSS-Image at 16-bit runs on the RTX PRO 6000 Blackwell, about 20 s per image.

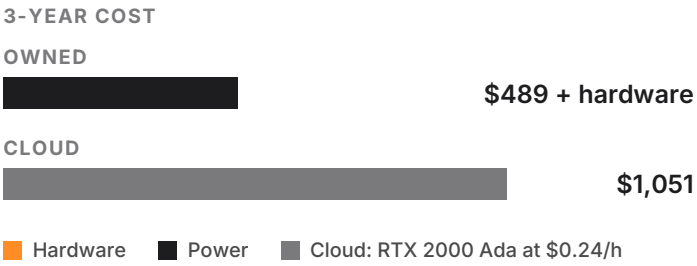
VALIDATION

CHECK	TARGET	RESULT	STATUS
Fits in GPU memory	96.0 GB usable	10.5 GB needed	PASS
Headroom	5% free	85.5 GB free (89%)	PASS
Time per image	under 30 s	20 s	PASS

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	$163 · 620 W at the wall
CLOUD INSTEAD, PER YEAR	$350
3-YEAR TOTAL, OWNED	Hardware + $489 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 Oct 23, 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 RTX PRO 6000 Blackwell · 16-bit (BF16)

SETUP

PARAMETERS	5.1B
PRECISION	16-bit (BF16)
OUTPUT	1024 × 1024
STEPS	50
HARDWARE	RTX PRO 6000 Blackwell
GPU CAN USE	96.0 GB of 96.0 GB

MEMORY

WEIGHTS	9.5 GB
WORKING	216 MB
RUNTIME	819 MB
SYSTEM	0.0 GB
FREE	85.5 GB

SAME MODEL, OTHER MACHINES

RTX 4090	16-bit · 59 s
RTX 5090	16-bit · 47 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
DGX STATION GB300	16-bit · 4.9 s

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 Oct 2026; cloud prices are RunPod's on-demand list as of Oct 23, 2026. Report VR-101PWJ2, generated October 23, 2026.