GPUpartner

VALIDATION REPORT · IMAGE GENERATION

Krea 2 Turbo

on M4 24 GB · 5-bit (GGUF Q5_K_M)

STATUS · NOV 22, 2026

Not validated

1 of 3 checks pass

Krea 2 Turbo at 5-bit runs on the M4 24 GB, about 8.9 min per image.

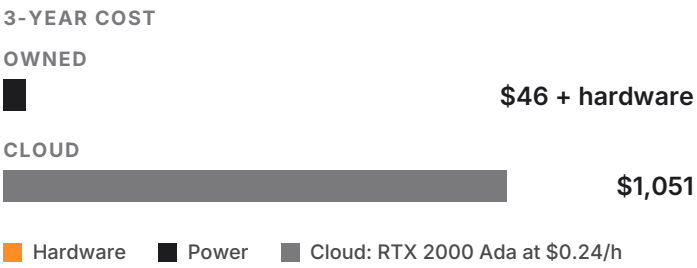
VALIDATION

CHECK	TARGET	RESULT	STATUS
Fits in GPU memory	16.0 GB usable	15.8 GB needed	PASS
Headroom	5% free	219 MB free (1%)	CAUTION
Time per image	under 30 s	8.9 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	$15 · 59 W at the wall
CLOUD INSTEAD, PER YEAR	$350
3-YEAR TOTAL, OWNED	Hardware + $46 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

Krea 2 Turbo

on M4 24 GB · 5-bit (GGUF Q5_K_M)

SETUP

PARAMETERS	12.8B
PRECISION	5-bit (GGUF Q5_K_M)
OUTPUT	1024 × 1024
STEPS	8
HARDWARE	M4 24 GB
GPU CAN USE	16.0 GB of 24.0 GB

MEMORY

WEIGHTS	12.6 GB
WORKING	2.3 GB
RUNTIME	819 MB
SYSTEM	8.0 GB
FREE	219 MB

SUGGESTED CHANGES

4-BIT WEIGHTS	Fits, 8.9 min an image
768² OUTPUT	Fits, 5.2 min an image
RTX PRO 2000 BLACKWELL	Runs it, 31 s an image
M5 MAX · 36 GB	Runs it, 39 s an image

SAME MODEL, OTHER MACHINES

RTX 4090	8-bit · 12 s
RTX 5090	16-bit · 9.4 s
RTX PRO 6000 BLACKWELL	16-bit · 4.5 s
M5 MAX 128 GB	16-bit · 34 s
M3 ULTRA 256 GB	16-bit · 66 s
DGX SPARK 128 GB	16-bit · 29 s
RYZEN AI MAX+ 395 128 GB	16-bit · 52 s

WHERE TO GET IT

FP8	huggingface.co/Comfy-Org/Krea-2
GGUF	huggingface.co/vantagewithai/Krea-2-Turbo-GGUF
OFFICIAL	huggingface.co/krea/Krea-2-Turbo

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-064SUWN, generated November 22, 2026.