GPUpartner

VALIDATION REPORT · IMAGE GENERATION

FLUX.2 [klein] 9B

on M4 Pro 24 GB · 4-bit (GGUF Q4_K_M · NVFP4)

STATUS · NOV 22, 2026

Validated with conditions

1 of 3 checks pass

FLUX.2 [klein] 9B at 4-bit runs on the M4 Pro 24 GB, about 63 s per image.

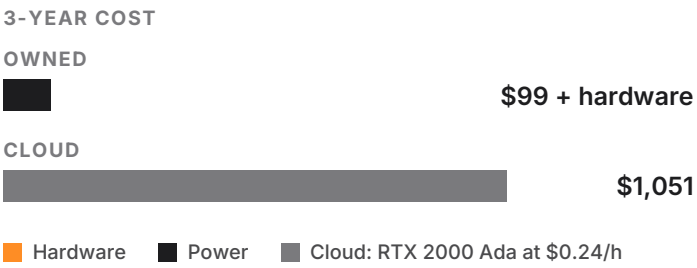
VALIDATION

CHECK	TARGET	RESULT	STATUS
Fits in GPU memory	16.0 GB usable	15.8 GB needed	PASS
Headroom	5% free	159 MB free (1%)	CAUTION
Time per image	under 30 s	63 s	CAUTION

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

FLUX.2 [klein] 9B

on M4 Pro 24 GB · 4-bit (GGUF Q4_K_M · NVFP4)

SETUP

PARAMETERS	9.1B
PRECISION	4-bit (GGUF Q4_K_M · NVFP4)
OUTPUT	1024 × 1024
STEPS	4
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	159 MB

SUGGESTED CHANGES

768² OUTPUT	Fits, 37 s an image
RTX PRO 2000 BLACKWELL	Runs it, 7.9 s an image
M5 MAX · 36 GB	Runs it, 9.9 s an image
RYZEN AI MAX+ 395 · 32 GB	Runs it, 15 s an image

SAME MODEL, OTHER MACHINES

RTX 4090	16-bit · 3.5 s
RTX 5090	16-bit · 3.0 s
RTX PRO 6000 BLACKWELL	16-bit · 1.8 s
M5 MAX 128 GB	16-bit · 8.8 s
M3 ULTRA 256 GB	16-bit · 16 s
DGX SPARK 128 GB	16-bit · 7.6 s
RYZEN AI MAX+ 395 128 GB	16-bit · 13 s

WHERE TO GET IT

FP8	huggingface.co/black-forest-labs/FLUX.2-klein-9b-fp8
NVFP4	huggingface.co/black-forest-labs/FLUX.2-klein-9b-nvfp4
GGUF	huggingface.co/unsloth/FLUX.2-klein-9B-GGUF
OFFICIAL	huggingface.co/black-forest-labs/FLUX.2-klein-9B

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-1BNV3X4, generated November 22, 2026.