GPUpartner

VALIDATION REPORT · IMAGE GENERATION

# FLUX.2 [klein] 9B

on RTX 4070 · 8-bit (FP8 · Q8_0)

STATUS · NOV 22, 2026

**Validated**

3 of 3 checks pass

**FLUX.2 [klein] 9B at 8-bit runs on the RTX 4070, about 8.0 s per image.**

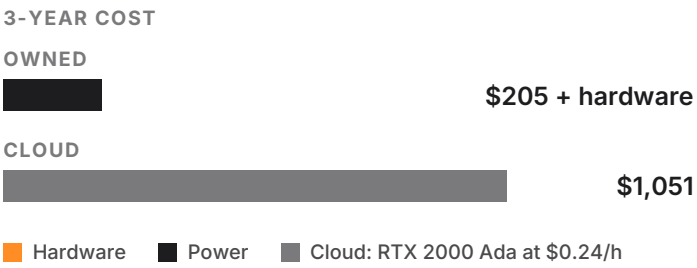
## VALIDATION

| CHECK | TARGET | RESULT | STATUS |
|---|---|---|---|
| Fits in GPU memory | 12.0 GB usable | 9.9 GB needed | PASS |
| Headroom | 5% free | 2.1 GB free (18%) | PASS |
| Time per image | under 30 s | 8.0 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 | $68 · 260 W at the wall |
| CLOUD INSTEAD, PER YEAR | $350 |
| 3-YEAR TOTAL, OWNED | Hardware + $205 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 RTX 4070 · 8-bit (FP8 · Q8_0)

## SETUP

| | |
|---|---|
| PARAMETERS | 9.1B |
| PRECISION | 8-bit (FP8 · Q8_0) |
| OUTPUT | 1024 × 1024 |
| STEPS | 4 |
| HARDWARE | RTX 4070 |
| GPU CAN USE | 12.0 GB of 12.0 GB |

## MEMORY

| | |
|---|---|
| WEIGHTS | 8.8 GB |
| WORKING | 288 MB |
| RUNTIME | 819 MB |
| SYSTEM | 0.0 GB |
| FREE | 2.1 GB |

## 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-1O16K9T, generated November 22, 2026.