GPUpartner

VALIDATION REPORT · IMAGE GENERATION

# Qwen-Image 2.1

on RTX 4070 · 8-bit (FP8 · Q8_0)

STATUS · NOV 22, 2026

**Not validated**

2 of 3 checks pass

**Qwen-Image 2.1 at 8-bit runs on the RTX 4070, about 4.3 min per image.**

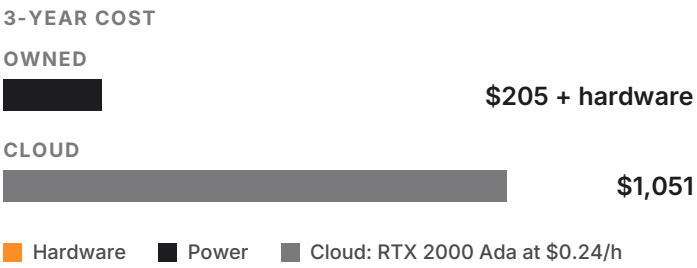
## VALIDATION

| CHECK | TARGET | RESULT | STATUS |
|---|---|---|---|
| Fits in GPU memory | 12.0 GB usable | 8.7 GB needed | PASS |
| Headroom | 5% free | 3.3 GB free (27%) | PASS |
| Time per image | under 30 s | 4.3 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 | $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

# Qwen-Image 2.1

on RTX 4070 · 8-bit (FP8 · Q8_0)

## SETUP

| | |
|---|---|
| PARAMETERS | 7.1B |
| PRECISION | 8-bit (FP8 · Q8_0) |
| OUTPUT | 2048 × 2048 |
| STEPS | 40 |
| HARDWARE | RTX 4070 |
| GPU CAN USE | 12.0 GB of 12.0 GB |

## MEMORY

| | |
|---|---|
| WEIGHTS | 6.9 GB |
| WORKING | 1.0 GB |
| RUNTIME | 819 MB |
| SYSTEM | 0.0 GB |
| FREE | 3.3 GB |

## SAME MODEL, OTHER MACHINES

| | |
|---|---|
| RTX 4090 | 16-bit · 1.5 min |
| RTX 5090 | 16-bit · 72 s |
| RTX PRO 6000 BLACKWELL | 16-bit · 31 s |
| M5 MAX 128 GB | 16-bit · 4.7 min |
| M3 ULTRA 256 GB | 16-bit · 9.1 min |
| DGX SPARK 128 GB | 16-bit · 4.0 min |
| RYZEN AI MAX+ 395 128 GB | 16-bit · 7.1 min |

## WHERE TO GET IT

| | |
|---|---|
| FP8 | huggingface.co/unsloth/Qwen-Image-2.1-FP8 |
| GGUF | huggingface.co/unsloth/Qwen-Image-2.1-GGUF |
| OFFICIAL | huggingface.co/Qwen/Qwen-Image-2.1 |

## 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-197VF08, generated November 22, 2026.