GPUpartner

VALIDATION REPORT · IMAGE GENERATION

Z-Image Turbo

on DGX Station GB300 · 16-bit (BF16)

STATUS · NOV 22, 2026

Validated

3 of 3 checks pass

Z-Image Turbo at 16-bit runs on the DGX Station GB300, about 1.4 s per image.

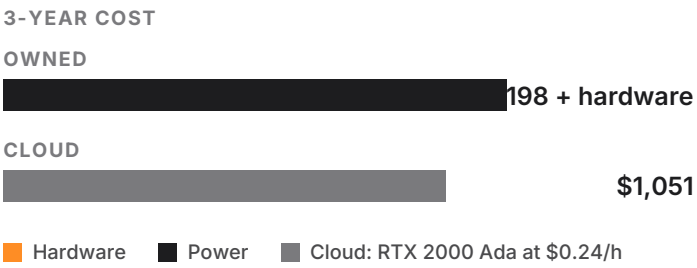
VALIDATION

CHECK	TARGET	RESULT	STATUS
Fits in GPU memory	252 GB usable	12.5 GB needed	PASS
Headroom	5% free	239 GB free (95%)	PASS
Time per image	under 30 s	1.4 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	$399 · 1520 W at the wall
CLOUD INSTEAD, PER YEAR	$350
3-YEAR TOTAL, OWNED	Hardware + $1,198 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

Z-Image Turbo

on DGX Station GB300 · 16-bit (BF16)

SETUP

PARAMETERS	6.2B
PRECISION	16-bit (BF16)
OUTPUT	1024 × 1024
STEPS	8
HARDWARE	DGX Station GB300
GPU CAN USE	252 GB of 252 GB
SPILL SPACE	480 GB LPDDR5X on the Grace CPU

SAME MODEL, OTHER MACHINES

RTX 4090	16-bit · 6.4 s
RTX 5090	16-bit · 5.2 s
RTX PRO 6000 BLACKWELL	16-bit · 2.8 s
M5 MAX 128 GB	16-bit · 18 s
M3 ULTRA 256 GB	16-bit · 33 s
DGX SPARK 128 GB	16-bit · 15 s
RYZEN AI MAX+ 395 128 GB	16-bit · 27 s

MEMORY

WEIGHTS	11.5 GB
WORKING	270 MB
RUNTIME	819 MB
SYSTEM	0.0 GB
FREE	239 GB

WHERE TO GET IT

FP8	huggingface.co/unsloth/Z-Image-Turbo-FP8
NVFP4	huggingface.co/Comfy-Org/z_image_turbo
GGUF	huggingface.co/unsloth/Z-Image-Turbo-GGUF
OFFICIAL	huggingface.co/Tongyi-MAI/Z-Image-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-0G2XU8J, generated November 22, 2026.