GPUpartner

VALIDATION REPORT · VIDEO GENERATION

Wan 2.1 T2V-14B

on DGX Station GB300 · 16-bit (BF16)

STATUS · OCT 22, 2026

Validated

3 of 3 checks pass

Wan 2.1 T2V-14B at 16-bit runs on the DGX Station GB300, about 2.8 min per clip.

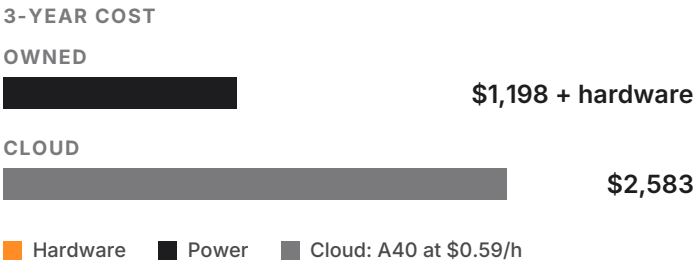
VALIDATION

CHECK	TARGET	RESULT	STATUS
Fits in GPU memory	252 GB usable	33.2 GB needed	PASS
Headroom	5% free	219 GB free (87%)	PASS
Time per clip	under 15 min	2.8 min	PASS

Targets: a clip in under 15 min, 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	$861
3-YEAR TOTAL, OWNED	Hardware + $1,198 power
3-YEAR TOTAL, CLOUD	$2,583
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 Oct 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 · VIDEO GENERATION

Wan 2.1 T2V-14B

on DGX Station GB300 · 16-bit (BF16)

SETUP

PARAMETERS	14.3B
PRECISION	16-bit (BF16)
OUTPUT	1280 × 720, 81 frames
STEPS	50
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	8-bit · 39 min
RTX 5090	8-bit · 31 min
RTX PRO 6000 BLACKWELL	16-bit · 13 min
M5 MAX 128 GB	16-bit · 2.0 h
M3 ULTRA 256 GB	16-bit · 3.9 h
DGX SPARK 128 GB	16-bit · 1.7 h
RYZEN AI MAX+ 395 128 GB	16-bit · 3.1 h

MEMORY

WEIGHTS	26.6 GB
WORKING	5.8 GB
RUNTIME	819 MB
SYSTEM	0.0 GB
FREE	219 GB

WHERE TO GET IT

FP8	huggingface.co/Comfy-Org/Wan_2.1_ComfyUI_repackaged
GGUF	huggingface.co/city96/Wan2.1-T2V-14B-gguf
OFFICIAL	huggingface.co/Wan-AI/Wan2.1-T2V-14B

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 Oct 2026; cloud prices are RunPod's on-demand list as of Oct 22, 2026. Report VR-0C4KT0Y, generated October 22, 2026.