GPUpartner

VALIDATION REPORT · VIDEO GENERATION

Wan 2.2 A14B

on RTX 5070 · 16-bit (BF16)

STATUS · NOV 22, 2026

Not validated

0 of 2 checks pass

Wan 2.2 A14B at 16-bit needs 33.2 GB; the RTX 5070 can use 12.0 GB. The smallest machine we checked that runs it is the A100 40 GB PCIe.

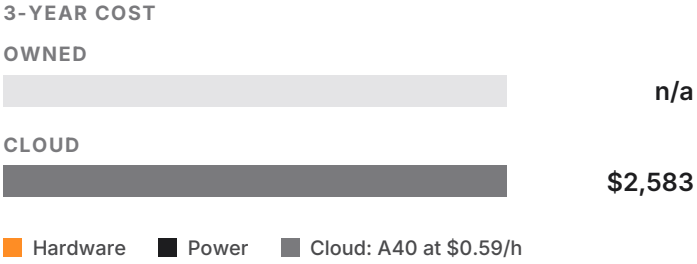
VALIDATION

CHECK	TARGET	RESULT	STATUS
Fits in GPU memory	12.0 GB usable	33.2 GB needed	FAIL
Headroom	5% free	none	FAIL
Time per clip	under 15 min	–	NOT TESTED

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	Doesn't run on this machine
CLOUD INSTEAD, PER YEAR	$861
3-YEAR TOTAL, OWNED	n/a
3-YEAR TOTAL, CLOUD	$2,583
BREAK-EVEN	n/a



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 · VIDEO GENERATION

Wan 2.2 A14B

on RTX 5070 · 16-bit (BF16)

SETUP

PARAMETERS	14.3B
PRECISION	16-bit (BF16)
HARDWARE	RTX 5070
GPU CAN USE	12.0 GB of 12.0 GB

MEMORY

WEIGHTS	26.6 GB
WORKING	5.8 GB
RUNTIME	819 MB
SYSTEM	0.0 GB
OVER BY	21.2 GB

SUGGESTED CHANGES

SPILL TO RAM	Runs, slower, 83 min a clip
A100 40 GB PCIE	Runs it, 17 min a clip
M5 MAX · 64 GB	Runs it, 1.6 h a clip
RYZEN AI MAX+ 395 · 64 GB	Runs it, 2.4 h a clip

SAME MODEL, OTHER MACHINES

RTX 4090	8-bit · 31 min
RTX 5090	8-bit · 24 min
RTX PRO 6000 BLACKWELL	16-bit · 10 min
M5 MAX 128 GB	16-bit · 1.6 h
M3 ULTRA 256 GB	16-bit · 3.1 h
DGX SPARK 128 GB	16-bit · 81 min
RYZEN AI MAX+ 395 128 GB	16-bit · 2.4 h

WHERE TO GET IT

FP8	huggingface.co/Comfy-Org/Wan_2.2_ComfyUI_Repackaged
GGUF	huggingface.co/QuantStack/Wan2.2-T2V-A14B-GGUF
OFFICIAL	huggingface.co/Wan-AI/Wan2.2-T2V-A14B

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-1EIBCQ2, generated November 22, 2026.