GPUpartner

VALIDATION REPORT · VIDEO GENERATION

LTX-2.5

on RTX 4080 Super · 5-bit (GGUF Q5_K_M)

STATUS · NOV 22, 2026

Validated with conditions

2 of 3 checks pass

LTX-2.5 at 5-bit runs on the RTX 4080 Super, about 82 s per clip.

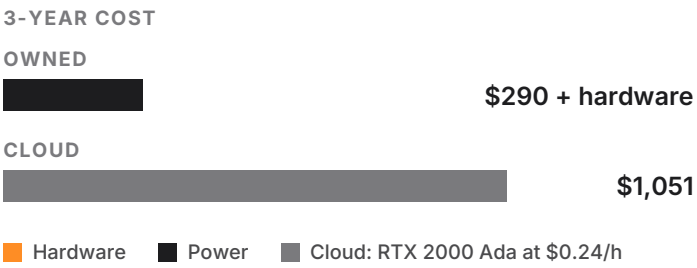
VALIDATION

CHECK	TARGET	RESULT	STATUS
Fits in GPU memory	16.0 GB usable	15.3 GB needed	PASS
Headroom	5% free	753 MB free (5%)	CAUTION
Time per clip	under 15 min	82 s	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	$97 · 368 W at the wall
CLOUD INSTEAD, PER YEAR	$350
3-YEAR TOTAL, OWNED	Hardware + $290 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 · VIDEO GENERATION

LTX-2.5

on RTX 4080 Super · 5-bit (GGUF Q5_K_M)

SETUP

PARAMETERS	21B
PRECISION	5-bit (GGUF Q5_K_M)
OUTPUT	960 × 544, 121 frames
STEPS	8
HARDWARE	RTX 4080 Super
GPU CAN USE	16.0 GB of 16.0 GB

MEMORY

WEIGHTS	13.9 GB
WORKING	542 MB
RUNTIME	819 MB
SYSTEM	0.0 GB
FREE	753 MB

SUGGESTED CHANGES

4-BIT WEIGHTS	Fits, 82 s a clip
480P OUTPUT	Fits, 69 s a clip
RTX 4000 SFF ADA	Runs it, 1.7 min a clip
M5 MAX · 48 GB	Runs it, 2.4 min a clip

SAME MODEL, OTHER MACHINES

RTX 4090	8-bit · 54 s
RTX 5090	8-bit · 47 s
RTX PRO 6000 BLACKWELL	16-bit · 31 s
M5 MAX 128 GB	16-bit · 2.1 min
M3 ULTRA 256 GB	16-bit · 3.8 min
DGX SPARK 128 GB	16-bit · 1.8 min
RYZEN AI MAX+ 395 128 GB	16-bit · 3.1 min

WHERE TO GET IT

FP8	huggingface.co/unsloth/LTX-2.5-FP8
GGUF	huggingface.co/unsloth/LTX-2.5-GGUF
OFFICIAL	huggingface.co/Lightricks/LTX-2.5

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-1FO5XSH, generated November 22, 2026.