GPUpartner

VALIDATION REPORT · VIDEO GENERATION

MiniMax H3

on M4 Pro 64 GB · 4-bit (GGUF Q4_K_M · NVFP4)

STATUS · NOV 22, 2026

Not validated

1 of 3 checks pass

MiniMax H3 at 4-bit runs on the M4 Pro 64 GB, about 2.4 h per clip.

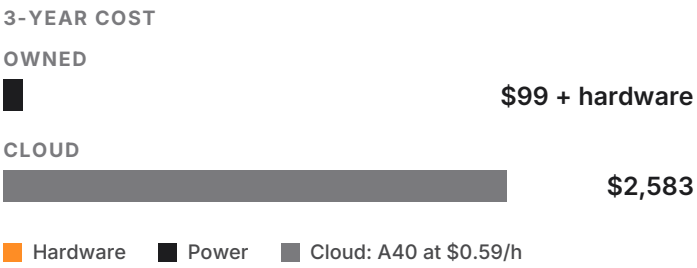
VALIDATION

CHECK	TARGET	RESULT	STATUS
Fits in GPU memory	48.0 GB usable	46.1 GB needed	PASS
Headroom	5% free	1.9 GB free (4%)	CAUTION
Time per clip	under 15 min	2.4 h	FAIL

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	$33 · 126 W at the wall
CLOUD INSTEAD, PER YEAR	$861
3-YEAR TOTAL, OWNED	Hardware + $99 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 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

MiniMax H3

on M4 Pro 64 GB · 4-bit (GGUF Q4_K_M · NVFP4)

SETUP

PARAMETERS	20.1B
PRECISION	4-bit (GGUF Q4_K_M · NVFP4)
OUTPUT	1360 × 768, 121 frames
STEPS	20
HARDWARE	M4 Pro 64 GB
GPU CAN USE	48.0 GB of 64.0 GB

MEMORY

WEIGHTS	42.3 GB
WORKING	3.0 GB
RUNTIME	819 MB
SYSTEM	16.0 GB
FREE	1.9 GB

SUGGESTED CHANGES

720P OUTPUT	Fits, 2.0 h a clip
RTX 5090	Runs it, 5.6 min a clip
M5 ULTRA · 96 GB	Runs it, 13 min a clip
RYZEN AI MAX+ 395 · 96 GB	Runs it, 32 min a clip

SAME MODEL, OTHER MACHINES

RTX 4090	Too big
RTX 5090	8-bit · 4.9 min
RTX PRO 6000 BLACKWELL	16-bit · 2.2 min
M5 MAX 128 GB	16-bit · 18 min
M3 ULTRA 256 GB	16-bit · 35 min
DGX SPARK 128 GB	16-bit · 16 min
RYZEN AI MAX+ 395 128 GB	16-bit · 28 min

WHERE TO GET IT

FP8	huggingface.co/Comfy-Org/MiniMax-H3
NVFP4	huggingface.co/lilcheaty/MiniMax-H3-NVFP4
GGUF	huggingface.co/Abiray/MiniMax-H3-GGUF
OFFICIAL	huggingface.co/MiniMaxAI/MiniMax-H3

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-0P1XJFC, generated November 22, 2026.