GPUpartner

VALIDATION REPORT · VIDEO GENERATION

# MiniMax H3

on DGX Spark 128 GB · 16-bit (BF16)

STATUS · NOV 22, 2026

**Validated with conditions**

2 of 3 checks pass

**MiniMax H3 at 16-bit runs on the DGX Spark 128 GB, about 16 min per clip.**

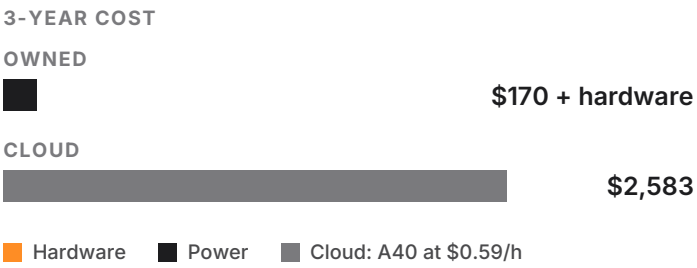
## VALIDATION

| CHECK | TARGET | RESULT | STATUS |
|---|---|---|---|
| Fits in GPU memory | 120 GB usable | 72.3 GB needed | PASS |
| Headroom | 5% free | 48.0 GB free (40%) | PASS |
| Time per clip | under 15 min | 16 min | CAUTION |

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 | $57 · 216 W at the wall |
| CLOUD INSTEAD, PER YEAR | $861 |
| 3-YEAR TOTAL, OWNED | Hardware + $170 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 DGX Spark 128 GB · 16-bit (BF16)

## SETUP

| | |
|---|---|
| PARAMETERS | 20.1B |
| PRECISION | 16-bit (BF16) |
| OUTPUT | 1360 × 768, 121 frames |
| STEPS | 20 |
| HARDWARE | DGX Spark 128 GB |
| GPU CAN USE | 120 GB of 128 GB |

## MEMORY

| | |
|---|---|
| WEIGHTS | 68.5 GB |
| WORKING | 3.0 GB |
| RUNTIME | 819 MB |
| SYSTEM | 7.7 GB |
| FREE | 48.0 GB |

## 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 |
| RYZEN AI MAX+ 395 128 GB | 16-bit · 28 min |
| DGX STATION GB300 | 16-bit · 43 s |

## 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-1JU02K2, generated November 22, 2026.