GPUpartner

VALIDATION REPORT · VIDEO GENERATION

# HunyuanVideo 1.5

on M3 Ultra 256 GB · 16-bit (BF16)

STATUS · NOV 22, 2026

**Not validated**

2 of 3 checks pass

**HunyuanVideo 1.5 at 16-bit runs on the M3 Ultra 256 GB, about 3.8 h per clip.**

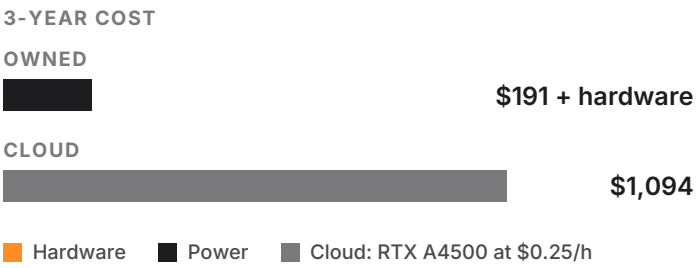
## VALIDATION

| CHECK | TARGET | RESULT | STATUS |
|---|---|---|---|
| Fits in GPU memory | 192 GB usable | 26.5 GB needed | PASS |
| Headroom | 5% free | 165 GB free (86%) | PASS |
| Time per clip | under 15 min | 3.8 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 | $64 · 243 W at the wall |
| CLOUD INSTEAD, PER YEAR | $365 |
| 3-YEAR TOTAL, OWNED | Hardware + $191 power |
| 3-YEAR TOTAL, CLOUD | $1,094 |
| 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

# HunyuanVideo 1.5

on M3 Ultra 256 GB · 16-bit (BF16)

## SETUP

| | |
|---|---|
| PARAMETERS | 8.3B |
| PRECISION | 16-bit (BF16) |
| OUTPUT | 1280 × 720, 121 frames |
| STEPS | 50 |
| HARDWARE | M3 Ultra 256 GB |
| GPU CAN USE | 192 GB of 256 GB |

## MEMORY

| | |
|---|---|
| WEIGHTS | 22.3 GB |
| WORKING | 3.4 GB |
| RUNTIME | 819 MB |
| SYSTEM | 64.0 GB |
| FREE | 165 GB |

## SAME MODEL, OTHER MACHINES

| | |
|---|---|
| RTX 4090 | 16-bit · 38 min |
| RTX 5090 | 16-bit · 30 min |
| RTX PRO 6000 BLACKWELL | 16-bit · 13 min |
| M5 MAX 128 GB | 16-bit · 2.0 h |
| DGX SPARK 128 GB | 16-bit · 1.7 h |
| RYZEN AI MAX+ 395 128 GB | 16-bit · 3.0 h |
| DGX STATION GB300 | 16-bit · 2.8 min |

## WHERE TO GET IT

| | |
|---|---|
| FP8 | huggingface.co/Comfy-Org/HunyuanVideo_1.5_repackaged |
| GGUF | huggingface.co/jayn7/HunyuanVideo-1.5_T2V_720p-GGUF |
| OFFICIAL | huggingface.co/tencent/HunyuanVideo-1.5 |

## 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-1PXX166, generated November 22, 2026.