Self-Hosting Open-Weights Video - Decision Sheet

Fora Soft Learn · AI for Video Engineering

1 - The five open-weights families (read license + VRAM first)

Model family	Best version	License	Min VRAM	Standout
Wan (Alibaba)	2.2 (open)	**Apache 2.0**	~8 GB (5B)	top open VBench 84.7%
HunyuanVideo (Tencent)	1.5	**Community (restr.)**	~14 GB (FP8)	top quality T2V+I2V
CogVideoX (Zhipu)	1.5 / 5B	**Apache (2B) / custom (5B)**	~8 GB (5B)	easiest to run
Mochi 1 (Genmo)	Mochi 1	**Apache 2.0**	~20-60 GB	permissive; strong motion
LTX-Video / LTX-2	LTX-2	**Apache 2.0**	consumer-DC	real-time; 4K + audio

2 - The two gates (run in this order, before any benchmark)

```
GATE 1  LICENSE -> Does it let me SELL the output, at my user scale, in my market?
GATE 2  VRAM    -> Does it FIT the GPU I can afford? (quantization can shrink it to fit)
Quality is the TIE-BREAKER among models that clear both gates - never the filter.
```

3 - License traps to check before you build

- [ ] Apache 2.0 = clean commercial: sell, fine-tune, ship, no user cap. CogVideoX-2B, Mochi 1, Wan 2.1/2.2, LTX-2 qualify.
- [ ] HunyuanVideo ships under the Tencent Community License - commercial-restricted; read the thresholds before any paid product.
- [ ] CogVideoX-5B uses its own license (NOT Apache like the 2B). Wan 2.5+ moved API-first - verify the license on the EXACT version you ship.

4 - Cost the bill: build vs rent

- [ ] Cost per self-hosted clip = (render minutes / 60) x GPU hourly rate. H100 ~ $2-2.70/hr; A100 sub-$1/hr spot.
- [ ] Fully loaded cost = GPU rental + idle hours + amortised 20-40h setup + a 3-5x operations multiplier. NOT 'free model x electricity'.
- [ ] Break-even sits north of ~5,000 clips/month (GPUs kept busy). Below it, a paid API is cheaper AND far less work.

5 - The three non-price reasons to self-host

- [ ] Data residency: sensitive footage (telemedicine, surveillance, private uploads) must stay on machines you control.
- [ ] Volume: above the break-even with high GPU utilisation, the owned cost per clip drops below the API charge.
- [ ] License + fine-tuning: only a permissive open model lets you retrain on your data and ship the result inside a product you sell.