

Open-Frontier VLM Selection Sheet

Fora Soft Learn · AI for Video Engineering

1 · The shared three-part recipe

Every open VLM is an image reader (vision encoder, e.g. CLIP) + a small connector (projector/adaptor) + a language model. LLaVA proved in 2023 that you only have to train the small connector. The four families differ in how good each part is and how they are trained, not in the shape.

2 · The four families (2026 snapshot)

Qwen-VL (Qwen3-VL) : Apache 2.0 | 2B-32B dense + 235B MoE | 256K->1M context | BEST ALL-ROUND
 InternVL (InternVL3): MIT-style | 1B-78B | ~72% MMMU | cleanest licence, top quality
 Pixtral 12B (Mistral): Apache 2.0 | 12B | 128K context | fits on ONE modest GPU
 LLaVA (OneVision-1.5): Apache 2.0 | fully open data | research baseline, not production

3 · The licence usually beats the benchmark

Apache 2.0 (Qwen, Pixtral) and MIT (InternVL) allow unrestricted commercial use at any scale. Community licences (some other models) allow commercial use but add conditions and a hyperscale cap, and are source-available, not formally open-source. A model two points higher with a licence your lawyers reject ships nothing. Read the licence before the leaderboard.

4 · Self-host vs closed API — the break-even

The arithmetic

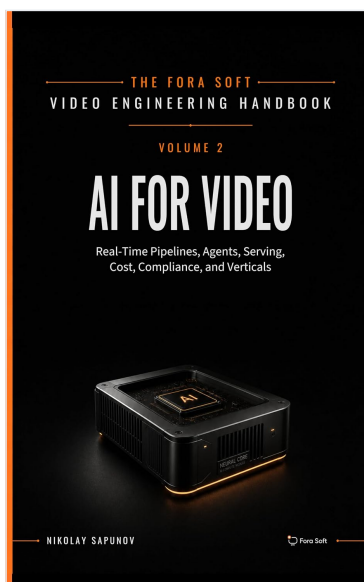
\$2/hr GPU x 24 x 30 = \$1,440 / month, fixed.
 \$1,440 / \$0.05 per video = ~28,800 videos.
 Above ~29k videos/month: own a GPU wins.

The hidden bills

- Idle GPU: paying full price for little work.
- Operations: deploy, patch, monitor, restart.
- vLLM eases serving, but free weights != free to run.

5 · Which open model for which job

- ☐ Best all-round open model, the safe default? Qwen3-VL - strongest scores, Apache 2.0, long video context.
- ☐ Need top quality under the cleanest possible (MIT-style) licence? InternVL3.
- ☐ Must fit on one modest GPU (edge, batch, near a camera)? Pixtral 12B - small and capable.
- ☐ Want a fully reproducible research baseline, not production? LLaVA-OneVision-1.5.
- ☐ Data can leave servers, low volume, need fast prototype? Rent a closed API instead.
- ☐ For a VIDEO product, weight video benchmarks (Video-MME), not still-image scores (MMMU).



GO DEEPER · VOLUME 7 OF 7

AI for Video: A Complete Guide to Intelligence in Video Products

From Real-Time Pipelines and Agents to Serving, Cost, Compliance, and Verticals

By Nikolay Sapunov, CEO at Fora Soft · 1,221 pages · Kindle & paperback on Amazon

[Read it on Amazon →](#)
amazon.com/dp/B0GX3DGBH4

This sheet compares Qwen-VL, InternVL, Pixtral and LLaVA, the licence question and the self-host break-even. Volume 7 covers running those models in production: serving with vLLM, Triton and TensorRT, cost per stream-hour, real-time pipelines, and compliance such as the EU AI Act.

WHAT YOU WILL FIND INSIDE

- Serving VLMs with vLLM, Triton and TensorRT
- Cost per stream-hour: self-host versus API
- Real-time video pipelines with VLM steps
- EU AI Act duties for video AI products

THE FULL HANDBOOK · 7 VOLUMES · 9,000+ PAGES · ALL ON AMAZON

Vol. 1 · Video Encoding — pixels and color to H.264, HEVC, AV1, quality metrics and AI

Vol. 2 · Video Streaming I — network transport, contribution and ingest, HLS, MPEG-DASH, CMAF

Vol. 3 · Video Streaming II — ABR, CDNs, players, WebRTC, DRM and the economics of delivery

Vol. 4 · Audio for Video I — PCM, codecs, loudness, accessibility, real-time voice processing

Vol. 5 · Audio for Video II — jitter buffers, lip sync, Atmos, quality metrics, neural codecs

Vol. 6 · AI for Video I — models and inference, computer vision, speech, search, generative video

Vol. 7 · AI for Video II — real-time pipelines, agents, serving, cost, compliance, verticals

Choosing an open VLM and where to run it?

We will benchmark candidate models on your videos, check licences, and model self-host versus API cost at your volume. 250+ video products shipped since 2005.

[Book a 30-min call →](#)
[WhatsApp →](#)
[Email us →](#)