

Video AI Cost Optimization — 25-Lever Checklist

Fora Soft Learn · AI for Video Engineering

The one idea — levers multiply, they don't add

Three independent 3× wins compound to 27×, not 9×. Pull EVERY lever that applies to your workload. Two levers multiply only when they act on different parts of the cost (token count vs price per token).

The 25 levers, in five families

1-5 PROCESS LESS VIDEO (the biggest win for video)

- **1** Sample frames — don't process all 30 every second (1 fps = 30× fewer)
- **2** Downscale resolution before the model sees the frame
- **3** Crop to the region of interest with a cheap detector first
- **4** Gate the big model behind a cheap trigger (run on events, not a schedule)
- **5** Trim the system prompt; cap the output length (output tokens cost the most)

6-10 RUN A SMALLER, CHEAPER MODEL

- **6** Route the easy 90% to a cheaper tier (frontier is 8-30× the price)
- **7** Cascade: try the cheap model first, escalate only on low confidence
- **8** Swap a frontier API for a small open model where quality allows
- **9** Distil a smaller student model (e.g. Distil-Whisper, ~6× faster)
- **10** Quantize to INT8 / INT4 (~4× less memory, usually <1% accuracy)

11-15 SERVE THE MODEL MORE EFFICIENTLY (self-host)

- **11** Continuous batching — keep the GPU full (up to ~23× throughput)
- **12** Reuse the KV cache (PagedAttention)
- **13** Speculative decoding — 2-3× faster, mathematically identical output
- **14** A purpose-built runtime (vLLM / TensorRT-LLM / SGLang)
- **15** Right-size the GPU to the model's memory footprint

16-20 PAY LESS PER UNIT OF COMPUTE (mostly config, not code)

- **16** Batch API for non-real-time jobs — flat 50% off
- **17** Prompt caching — up to 90% off the repeated input prefix
- **18** Semantic answer caching + dedupe re-uploaded clips
- **19** Spot / preemptible GPUs — 60-90% off (interruptible work only)
- **20** Reserved / committed capacity — 20-70% off the steady baseline

21-25 ARCHITECT & GOVERN FOR COST

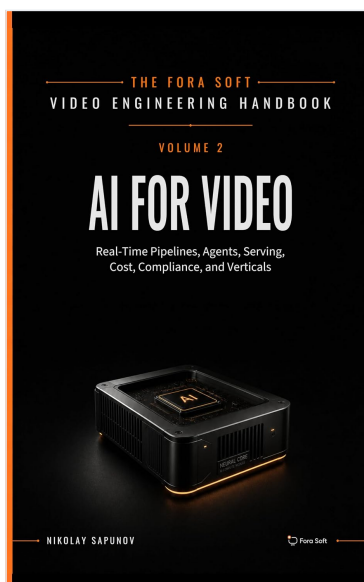
- **21** Serverless scale-to-zero — pay per second, idle costs nothing
- **22** Right provider (neocloud vs hyperscaler); kill egress/storage waste
- **23** Hybrid edge + cloud — constant work on-device, hard work in the cloud
- **24** Precompute derived artifacts once (transcript, embeddings, index)
- **25** Instrument per-feature cost budgets; alert when one drifts

Worked example — a 50,000-hour/month moderation pipeline

- Naive: 1.08M tok/hr × \$2.50/1M = \$2.70/hr × 50,000 = \$135,000 / month
- + process less video (low-res + sampling, ~1/3 tokens) → \$45,000 (/3)
- + route the easy 90% to a Flash tier (blended \$0.52/1M) → \$9,360 (/4.8)
- + run the offline job on the batch API (-50%) → \$4,680 (/2)
- Three independent levers → 29× cheaper, with no change the user sees.

Before you optimize, ask:

- ☐ Have we instrumented cost per feature BEFORE optimizing — do we know where the money actually goes?
- ☐ Are we processing less video first (frame sampling, low resolution, ROI crop, event-gating)?
- ☐ Is routine work on a cheaper model tier, with the flagship kept for the hard minority?
- ☐ Is every non-real-time job on the batch API, with shared prompts cached (up to 90% off)?
- ☐ Does the pricing mode match the load (spot = interruptible, reserved = baseline, serverless = spiky)?
- ☐ For self-hosted work, are we on an efficient runtime with continuous batching and a right-sized GPU?
- ☐ Are the stacked levers INDEPENDENT, and did we verify the total against a real measurement?



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 checklist lists 25 levers that cut the cost of video AI and shows how they multiply. Volume 7 covers running these systems in production: serving on vLLM, Triton and TensorRT, cost per stream-hour, real-time pipelines that process only what matters, and the infrastructure choices behind each line of the bill.

WHAT YOU WILL FIND INSIDE

- Serving with vLLM, Triton and TensorRT
- Real-time pipeline budgets
- Cost per stream-hour
- Infrastructure choices behind the bill

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

Is your video AI bill growing faster than usage?

We will instrument cost per feature, apply sampling, model routing, batching and caching where they fit, and verify the savings on real traffic. 250+ video products shipped since 2005.

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