

Jetson Orin Edge-Vision Scoping Worksheet

Size a detector + tracker + segmenter deployment before the next hardware order.

1. Pick the board — match AI throughput to your camera count

Module (2026, Super)	Peak AI (INT8)	Good for
Orin Nano 8GB	67 TOPS	1 camera: detect + track, occasional segment
Orin NX 16GB	157 TOPS	Several cameras, more segmenting
AGX Orin 64GB	275 TOPS	Many cameras, larger models
AGX Thor	800 TOPS	Robotics, several large models at once

2. Budget the frame — 33 ms at 30 fps (representative Orin NX, 1 camera)

Stage	Cost	Note
Decode (NVDEC)	~2 ms	Hardware decoder, runs in parallel — near-free of AI budget
Detect (YOLO, INT8)	~15 ms	The big cost. Size the board around this stage
Track (ByteTrack)	< 1 ms	CPU-side geometry. Nearly free
Segment (NanoSAM)	~3 ms	Only when a rule needs it, amortised over frames
Rules + overlay	~3 ms	Decide + draw on the GPU
Headroom	~9 ms	Safety margin — absorbs busy frames + thermal throttle

3. The trade — total work is fixed; divide it between cameras and frame rate

- ☐ **High fps, few cameras** — licence-plate at a gate, fast events.
- ☐ **Batch frames** — run the detector once over all cameras' frames together.
- ☐ **Compile INT8 + TensorRT** — and validate the accuracy gap on your footage.
- ☐ **Low fps, many cameras** — 'was anyone here?' across a site.
- ☐ **~40 streams @ ~5 fps** — Orin NX 16GB, small YOLOv8 INT8 (NVIDIA figure).
- ☐ **Keep frames on the GPU** — move only results across the chip boundary.

4. The four traps that miss the frame budget

Quoting the benchmark frame rate

'65 FPS YOLO' is the detector alone. Your pipeline also decodes, tracks, segments, and draws.

Fix: budget the whole pipeline, on your hardware and footage.

Copying the frame around

Crossing the CPU/GPU boundary copies the frame — often costlier than the AI itself.

Fix: keep the frame on the GPU; move only boxes and masks.

Segmenting everything, every frame

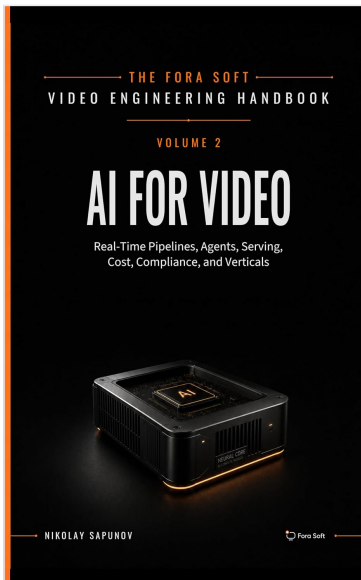
Full SAM 2 on every object blows the budget — it runs near 1 fps on small devices.

Fix: distilled SAM (NanoSAM), selected tracks only, reuse across frames.

Ignoring the enclosure heat

A throttled board in a hot housing silently drops frame rate below the bench number.

Fix: validate fps in the real enclosure, at real ambient temperature.



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 worksheet sizes the board, the 33 ms frame budget and the camera-versus-fps trade for a Jetson deployment. Volume 7 covers running vision under a real-time budget: real-time pipelines, serving and inference infrastructure such as TensorRT and Triton, cost per stream-hour, and vertical playbooks for video products.

WHAT YOU WILL FIND INSIDE

- Real-time pipelines and per-frame latency budgets
- Cost per stream-hour for edge and cloud
- Serving with TensorRT and Triton
- Vertical playbooks for vision 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

Sizing Jetson hardware for your camera count and frame rate?

We will profile your detector, tracker and segmenter on Orin, pick the board and deliver a pipeline that holds the frame budget in the real enclosure. 250+ video products shipped since 2005.

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