

Latency & Accuracy Tier Cheat Sheet

Match each surveillance analytic to the tier that meets its performance budget: how fast a detection arrives, and how often it is right. Representative 2026 figures.

A. The detection-latency budget (four legs)

- Capture & encode: ~15-33 ms, the same at every tier (it happens inside the camera).
- Frame to the compute: ~0 on the camera, 1-5 ms over the LAN, 40-200 ms across the internet.
- Inference: 8-30 ms small model on a camera NPU; 10-25 ms larger model on a GPU.
- Result delivery: 2-10 ms to a local relay, or 40-200 ms back across the internet.

B. End-to-end latency, and the placement rule

- On-camera ~25-100 ms; on-prem server ~30-120 ms; cloud ~300-800 ms (cellular >1,200 ms).
- The network round-trip - not the AI math - is what makes the cloud slow.
- Under 200 ms must run at the edge; 200 ms-1 s = hybrid; multi-second = cloud is fine.
- A person walks ~1.1 m in 800 ms (a cloud round-trip) but ~8 cm in 60 ms (on-camera).

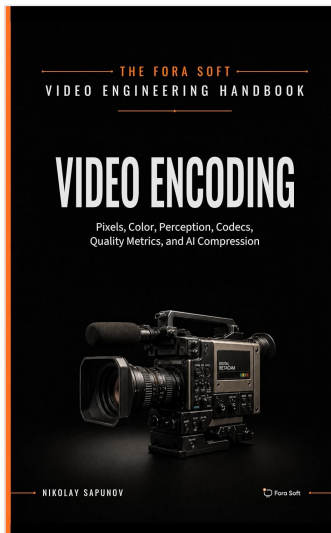
C. The accuracy ceiling by tier

- Model size sets the ceiling: nano ~40% mAP (camera) to xlarge ~55% mAP (cloud) on COCO.
- Quantizing to fit a camera costs 5-8% mAP; under 3% with calibration on real footage.
- The cloud's real edge is reasoning across cameras (re-ID, VLM), not single-frame detection.
- Accuracy is a precision/recall range: LPR 90-98%, face recognition condition-dependent. Never 100%.

D. When milliseconds matter (and when they do not)

- Edge (sub-200 ms): perimeter, intrusion, point-of-sale loss prevention, line-stop, fall detection.
- On-prem server: real-time heavy detection across many cameras on one local network.
- Cloud (seconds are fine): people-counting, dwell time, heatmaps, trend analytics.
- Cloud reasoning: cross-camera re-identification, VLM scene description, forensic search.

Place detection near the camera where reflexes matter and reserve the cloud for cross-camera reasoning where judgment matters and seconds are fine. State accuracy as a precision/recall range measured on your own footage, never a datasheet number. Engineering guidance, not legal advice - face and plate recognition are biometric and legally gated (EU AI Act high-risk; Illinois BIPA consent) regardless of tier. Figures are representative for 2026 hardware (a camera-class NPU, an on-prem GPU server, a cloud GPU region) and move with model size, resolution, frame rate, and network path. Sources: NIST FRVT; ONVIF Profile M; IETF RFC 3550; Ultralytics/COCO YOLO benchmarks; NVIDIA DeepStream; current edge-vs-cloud latency analyses.



GO DEEPER · VOLUME 1 OF 7

Video Encoding: A Complete Guide to Video Compression

From Pixels and Color to H.264, HEVC, AV1, Quality Metrics, and AI

By Nikolay Sapunov, CEO at Fora Soft · 1,522 pages · Kindle & paperback on Amazon

Read it on Amazon →

amazon.com/dp/B0HF52B65R

This sheet splits detection latency into four legs and maps each analytic to camera, server or cloud. Volume 1 covers the encoding side of that budget in depth: how capture and encode settings set latency, how H.264, HEVC and AV1 trade bitrate for quality, and how compression affects what a model can still see.

WHAT YOU WILL FIND INSIDE

- Capture and encode latency: GOP, lookahead, B-frames
- H.264, HEVC and AV1 for camera streams
- How compression artifacts affect analytics
- Quality metrics for checking camera streams

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

Deciding which analytics run on the camera, the server or the cloud?

We will measure latency and accuracy per tier on your cameras and footage and give you a placement plan per analytic. 250+ video products shipped since 2005.

Book a 30-min call →

WhatsApp →

Email us →