Anomaly Detection Architecture Worksheet

Walk the five-layer stack against four numbers — latency, cameras, labels, compliance.

1. The four numbers — fill these in first

Latency budget

Sub-150 ms → edge.
200–500 ms → hybrid.
>500 ms → cloud.

your answer

Camera count

<20 → SaaS likely.
20–100 → hybrid.
100+ → custom likely.

your answer

Labelled examples

<100 → CLIP zero-shot.
100–10K → MIL.
>10K → fully supervised.

your answer

Compliance regime

EU AI Act → edge-first.
GDPR Art. 9 / Art. 35.
HIPAA / BIPA / CCPA.

your answer

2. The five-layer stack — check the layers your build needs

Layer	Where	Typical model	Latency	When to skip
1. Candidate filter	Edge	I3D / SlowFast INT8, MemAE	30–80 ms	Never
2. Class classifier	Edge or cloud	MIL: RTFM, MGFN, MTFL	30–60 ms	Pure object detection use cases
3. VLM re-scorer	Cloud	VadCLIP, TPWNG, VadCLIP++	100–300 ms	Single fixed anomaly class with many labels
4. Frontier explanation	Cloud	Gemini 2.5 Pro, Claude Opus 4.7	1–3 s	No regulator or audit requirement
5. Signal anomaly	Sidecar	Isolation Forest, LSTM-AE	Sec to min	Never — it catches the silent failures

3. Anomaly class — name it before picking a model

- ☐ Behavioural — running, climbing, loitering, collapse. 3D-CNN + MIL.
- ☐ Physical — unattended bag, spill, missing object. YOLO + tracking.
- ☐ Density / crowd — queue overflow, crowd crush. Heat-map regression.
- ☐ Schedule / temporal — motion at 03:00, door propped open. Rules + normality.
- ☐ Statistical — frozen feed, encoder drift. Isolation Forest on metrics.
- ☐ Combined — multiple of the above in the same product. Layered stack.

4. Topology decision

Edge only

Jetson Orin Nano Super or Hailo-8 per camera. $200–600/unit.
Sub-200 ms. Frames never leave site.
Strongest AI Act posture.
Fit: small site, regulated, poor connectivity.

Hybrid (default)

Edge runs layers 1–2 high-precision; cloud re-scores top 5–20%.
6× cloud-volume reduction, 95%+ true-anomaly retention.
Fit: 20+ cameras, real-time SLA, regulated footage.

Cloud only

Stream all cameras to L4 / L40S / A100 in managed region.
$1–2.50/GPU-hour. 100+ streams per GPU.
Fit: forensic-only, central retraining, no sub-300 ms SLA.

5. EU AI Act + audit trail — engineering artefacts, not legal paperwork

High-risk under Annex III (Articles 9–15) → mandatory: documented risk management, dataset governance, technical documentation, automatic logging, human oversight, accuracy and robustness requirements, conformity assessment. Fines up to 3% of global turnover or €15M. Engineering artefacts: model card, data sheet, training log, drift log, alert audit trail, reviewer queue with sign-off, immutable storage (S3 Object Lock). Preferred class: gait, posture, object, behaviour — avoid face-based where the use case allows. Article 50 disclosure for AI-generated or AI-augmented content.