VLM vs. Custom CV — Decision Checklist

Run the six questions, then the cost and latency math, before you build a single feature.

1. The six-question decision tree (stop at the first clear answer)

- ☐ Categories known and fixed? NO -> VLM. YES -> next question.
- ☐ Real-time live video? YES -> custom detector (cloud VLM can't meet ms latency). NO -> next.
- ☐ Over ~100K frames / month? YES -> custom detector (better economics). NO -> next.
- ☐ Audited automated decision on the output? YES -> custom detector (determinism). NO -> next.
- ☐ ML team + time to train and maintain? NO -> VLM (ships in days). YES -> next.
- ☐ Need context beyond object location? YES -> VLM. NO -> custom detector.

2. Cost math — find your break-even

- ☐ VLM cost per SD frame ~ $0.0006 (Gemini 2.5 Flash: 960x540 = 6 tiles x 258 tokens @ $0.30/M in).
- ☐ Per camera at 5 FPS ~ 13M frames/month -> ~$7,500/month if every frame hits the VLM.
- ☐ Custom detector: high one-time build (~$8K-$20K), then near-zero per frame forever.
- ☐ Crossover rule of thumb ~ 100,000 frames/month: above it the detector wins, and the gap grows.

3. Latency and consistency guardrails

- ☐ Real-time stream? Custom detector (~1.5 ms server GPU, ~25 ms edge). Cloud VLM = seconds.
- ☐ Never ask a VLM for exact counts or precise positions -- it hallucinates (GroundCount, OmniSpatial 2026).
- ☐ Automated audited decision? Need deterministic output -> detector, not a probabilistic VLM.
- ☐ If you must combine, feed the detector's boxes into the VLM prompt so it reasons over real coords.

4. Ship the hybrid funnel (cheap gates expensive)

- ☐ Stage 1 - Motion / tiny detector on-device drops ~99% of frames for near-free.
- ☐ Stage 2 - Custom detector + geometry locates the object and checks the zone reliably.
- ☐ Stage 3 - VLM with a structured yes/no prompt, only on the rare frames that earn the cost.
- ☐ Constrain VLM output to fixed choices; require agreement across 2 frames for high-stakes calls.