UGC Moderation Pipeline - Build Blueprint

Fora Soft Learn · AI for Video Engineering · Capstone 10.6

1 - The pipeline: two rules, five stages (a funnel)

Rule 1	Cheapest check first - each stage decides or escalates; expensive stages never see the bulk.
Rule 2	Every decision logged with a reason; illegal content reported + preserved, not deleted.
Stages	Ingest -> Stage0 hash -> Stage1 classifiers -> Stage2 VLM -> Stage3 human, over a governance floor.
Key move	Moderate sampled keyframes + audio transcript, NOT whole videos by the minute.

2 - Build vs buy (2026): rent the models, partner on hashes, build your policy

```
BUY / HOST  classifiers (Rekognition/Hive/Azure/NudeNet/CLIP) + VLM (Gemini/Claude/GPT/LLaVA/Qwen-VL)
PARTNER     known-bad hash systems (PhotoDNA, NCMEC/IWF, PDQ+TMK, Lantern) - never build/hold yourself
BUILD       routing + thresholds + decision store + audit + reporting + humane review tooling = product
ABSTRACT    classifier and VLM behind one interface each, so a vendor swap is a config change.
```

3 - The 2026 tool field (verify prices before you ship)

```
CLASSIFIERS  Azure ~$1.50/1k images (~$0.0015/img) | Rekognition ~$0.10/min video | Google | Hive | open
VLM          frontier Gemini/Claude/GPT for the hard cases | open LLaVA/Qwen-VL for routine context
HASH         PhotoDNA + NCMEC/IWF sets + Meta PDQ/TMK + Tech Coalition Lantern/VHIP (784k+ videos hashed)
NOTE         hash catches the KNOWN only; AI-generated CSAM has no prior hash -> classifiers + VLM matter.
```

4 - The cost model (per day at 100,000 videos)

```
NAIVE   200,000 video-min/day x $0.10/min = $20,000/day ~ $7.3M/year (and too slow).
FUNNEL  hash $50 + classifiers $1,800 + VLM(~7%) $140 + humans(~1%) $250 = ~$2,240/day ~ $0.8M/year.
RESULT  ~9x cheaper AND more accurate - the costly stages only see what truly needs them.
LEVER   tune Stage 1 thresholds (the share it clears), not the model, to win the bill.
```

5 - Build order: each milestone protects users

```
1 Hash matching + mandatory reporting (Stage 0 + legal floor) - cheapest and legally required, first.
2 Cheap classifier pass, conservative thresholds (Stage 1) - escalate generously, tighten with evidence.
3 Human review console + queue (Stage 3) BEFORE the VLM - with wellbeing features built in from day one.
4 VLM context stage (Stage 2) to shrink the human queue.   5 Scale + measurement + adversarial loop.
```

6 - Governance checklist (engineering context, not legal advice)

- [ ] Every chunk carries source + timecode from ingestion, so every decision can be logged with its reason.
- [ ] On a Stage 0 match: BLOCK + REPORT (NCMEC CyberTipline) + PRESERVE - never a quiet delete (18 U.S.C. 2258A).
- [ ] Thresholds set PER HARM CLASS - gravest harms deterministic (hash + human), not a tunable score.
- [ ] DSA statement-of-reasons + transparency; minors protected (Art. 28); EU AI Act Art. 50 marks AI media (Aug 2026).
- [ ] Reviewer wellbeing as a requirement: blurred previews, hard exposure caps per shift, rotation, living wage.
- [ ] Held-out known-answer eval set run continuously; novel confirmed harms fed back to the hash authorities.