Real-Time Content Moderation in an SFU - Cheat Sheet

1 - Real-time moderation has no slack

```
After-the-fact: check an upload before it publishes - time to spare.
Real-time: content is already live - act before the harmful frame reaches viewers.
```

2 - The SFU is the one place that sees every stream

```
Tap video, audio, and chat once at the media router you already run.
Moderate three media: frames, the spoken track, and text on the RFC 8831 data channel.
```

3 - The encryption fork (settle this FIRST)

```
True E2EE (SFrame, RFC 9605) -> SFU sees sealed frames -> server moderation impossible.
Transport-only -> SFU reads media -> moderate at the server (open social / dating / market).
You cannot have both on one stream. Else: moderate on-device, or hybrid downgrade.
```

4 - Sample, don't scan; cheap+certain before expensive+fuzzy

```
30 fps x 60 s x $0.001 = $1.80/min/stream. Sample 2 fps -> $0.12/min (15x cheaper).
Add keyframes + scene changes; VAD-gate audio so checks run only on speech.
Order: 1 hash match (CSAM) -> 2 visual classifier -> 3 audio toxicity -> 4 text chat.
```

5 - Confidence -> action; CSAM is a legal path

```
>0.95 act automatically (blur/mute/suspend); 0.70-0.95 reversible + human review; else allow+log.
False positive punishes the innocent; false negative ships harm. Set the balance per category.
CSAM hash hit: not a threshold. Stop, SEAL evidence, report to NCMEC (18 U.S.C. 2258A). Never delete.
```

6 - Build-vs-buy checklist

- [ ] Decide the encryption fork before anything else (server-readable vs E2EE vs hybrid).
- [ ] Media layer: mediasoup / Janus / LiveKit; tap frames via Encoded Transform / Insertable Streams.
- [ ] Classifier layer: Hive / Rekognition / Azure / Google / OpenAI; PhotoDNA via Microsoft / NCMEC.
- [ ] Buy a small broadcast delay (1-2 s) so you can act before content reaches viewers.
- [ ] Staff a human review queue for the uncertain middle band; keep an audit log (EU DSA).
- [ ] Build the CSAM evidence + report path with counsel on day one, not after the first incident.