Screening User-Supplied Video for Sexual Content Before Processing: A Two-Classifier, Human-Reviewed Gate and 50 Days of Production Data

Author: Vedansh Chauhan (Zoupyu) · ORCID 0009-0003-8885-6566
Type: Technical note · **Version:** 1.0 · **Language:** English
DOI: 10.5281/zenodo.23880340
Data window: 2026-09-13 07:18 UTC to 2026-11-02 04:13 UTC (50 days)
Licence: CC BY 4.0 (text and data files)

Abstract

Services that take a user's video and transform it (clipping, captioning, reframing, dubbing) are expected by payment processors and by law to keep sexually explicit material out. Published evaluations of NSFW detectors mostly use curated still images. Few report what happens when such a detector runs on the frames of ordinary user video in production.

We describe a pre-processing content gate in a short-form video clipping service. It screens every job at up to three points before any output exists. It scores sampled frames with an open-weights image classifier, counts a flag only when a second, independently trained classifier agrees, and sends anything short of a clear block to a human. It fails closed when frames cannot be read, and it stores scores and timestamps, never images.

Over 50 days the gate ran 3,369 screens on 1,599 jobs and scored 88,822 frames. Setting aside our own explicit test video, 177 jobs drew a flag under at least one rule. Human review confirmed 2 as violations and cleared 128 as false positives, so about 1.5% of reviewed flags were correct. The second classifier overturned the primary's flag in 126 screens, with a clear margin, and none of the jobs it allowed has been found to be a violation. Recall on unflagged traffic was not measured.

On this traffic the dominant problem is false positives on ordinary video, including children's animation, sport and music. A single high-scoring frame is noise. A platform's own signals (views, age restriction) are useful evidence but are not a substitute for review. We publish the design, the aggregate data, our mistakes, and recommendations for teams building similar systems.

Keywords: content moderation; NSFW detection; video screening; user-generated content; image classification; human-in-the-loop; false positives; trust and safety; ONNX; short-form video

---

1. Why this matters

Any product that accepts video from the public and does work on it carries the content it is given. Three pressures make screening necessary:

- **Payment processors.** Merchant policies commonly prohibit explicit and suggestive sexual content. In our case, the processor recommended taking no payments until automated moderation was live *(engineering log, 2026-09-11)*.

- **Law.** In the United States, the TAKE IT DOWN Act requires covered platforms to remove non-consensual intimate imagery, and known identical copies, within 48 hours of a valid request, from 19 May 2026 [FTC]. India's IT Amendment Rules 2026, in force from 20 February 2026, cut the window for removing non-consensual nudity or morphed imagery to 2 hours [Khaitan]. Rule 3(1)(g) of the IT Rules requires records of removed content to be kept for 180 days [IT Rules].
- **Distribution platforms.** Creators publish clips to short-video platforms, each with its own sexual-content rules. A tool that turns one long video into many short ones also multiplies one policy question.

The engineering question is not whether to screen. It is how to screen *video*, which is not the same as screening images, at a cost a small team can carry, without blocking legitimate users.

2. Setting

The service takes a long video (an upload, a Google Drive file, or a YouTube or Twitch link), finds moments worth clipping, lets the user choose and trim them, and renders vertical clips with burned-in captions. A second mode burns captions onto a whole video. Output is private to the account.

Source of screened video	Screens	Share
YouTube link	3,013	89.4%
Upload	203	6.0%
Twitch VOD link	118	3.5%
Google Drive file	35	1.0%
Total	**3,369**	

3,360 screens were for clip jobs and 9 for caption jobs. The traffic is mostly public online video: streamer content, challenge videos, gaming, sport, music, film and TV excerpts, and children's programming.

3. Design

3.1 Principle: stop it before anything is made

The design goal was that an explicit video must never be processed. Screening runs before the costly and visible stages (transcription, AI moment selection, rendering), not after output exists. A held job produces no clip, and the user's credit is refunded.

3.2 Screening points

Screen	When	What it samples	Recorded as
Source screen	Uploads and Drive: as soon as the file arrives. Stream links: after transcription, because seeking a remote stream many times is unreliable and an unreliable read must not hold a clean video.	Frames spread evenly across the stretch being processed. Caption jobs sample more densely, because the whole video is the deliverable.	Gate A
Board screen	After draft moments are found, before the user sees them.	Frames inside every draft moment and, for files on local disk, the whole section at a fixed interval.	Gate B
Clip screen	After the user's chosen clips are framed, before rendering.	Frames inside each stretch about to become a clip.	Gate B

Sample times are the midpoints of equal slices, so black intros and end cards are not over-sampled. The source screen is sparse; the later screens look again, more densely, at exactly the footage that will become output. Exact sampling intervals and frame caps are not published, because they would show where the gaps between samples fall.

Frames are decoded by ffmpeg and piped as raw RGB straight to the classifier, resized to the model's input size (224×224 for the primary, 384×384 for the second opinion). No intermediate JPEG is written.

3.3 Two classifiers that must agree

Primary: Falconsai/nsfw_image_detection [Falconsai]. A ViT (base `google/vit-base-patch16-224-in21k`) fine-tuned on about 80,000 images, Apache-2.0, eval accuracy 0.980 on its own data. We run the quantized ONNX export from `onnx-community/nsfw_image_detection-ONNX` on CPU with onnxruntime-node. Every sampled frame is scored.

Second opinion: Marqo/nsfw-image-detection-384 [Marqo]. A ViT-tiny fine-tune, Apache-2.0, trained on 220,000 images. We run the ONNX export published by the International Consortium of Investigative Journalists (ICIJ) [ICIJ-ONNX], a 22 MB file. Only frames the primary flags are sent to it.

A flagged frame counts only if the second model scores it at least 0.4, ICIJ's default threshold. On ICIJ's 670-file benchmark, at its best threshold of 0.39, Marqo had 5 false positives (1.04%) and missed 7 of 188 positives (3.72%). A flat black frame scores 0.3594 on it, which is why the line sits above that [ICIJ]. If the second model cannot score a frame (model down, frame unreadable), the primary's flag stands. A missing veto never lets through anything the primary alone would have stopped.

3.4 From frame scores to a decision

A decision is never made on one frame:

- **Blocked:** at least 2 counted frames score at least 0.90, **and** they are at least 15% of all frames scored.
- **Review:** any counted frame scores at least 0.80, short of a block.
- **Pass:** otherwise.

In enforce mode, both blocked and review stop the job. The difference is what happens next: a block also pauses new jobs from the account until a human reviews it. A job's status is the worst result across its screens.

Coverage rule. If fewer than 70% of the requested frames decode, the result is *incomplete*. The gate does not decide; the job waits and retries. If the classifier is unavailable, the job also waits. An outage of the screen must never become a way around it (fail closed).

3.5 Source-platform signals

For YouTube links, the gate records whether the platform itself reports the video as public, not age-restricted, widely viewed, and not brand new. When all of these hold, a flag is recorded but does not stop the job. Since 2026-10-26, a review-level flag on such a video is kept in the audit log only; a block-level flag is still emailed to a human. Uploads, Drive files and Twitch links are never trusted this way. The exact view and age cut-offs are not published, so they cannot be used as a target.

3.6 Human review, and remembering it

Every held job emails a reviewer and appears in an admin list. The reviewer sees timestamps and scores, not frames. They either **clear** the job, which then runs at no charge, or **confirm** the violation, which keeps it blocked and leads to account termination. A clear is stored against the source video's platform ID, so the same public video submitted later by anyone skips screening. A confirmed violation on that video overrides every clear. Uploads never match each other. Users are notified of every review outcome, naming the rule, never the scores or models.

3.7 What is stored

One audit row per screen, pass or not: gate, model versions, thresholds, frames requested and scored, top score, flagged timestamps with both models' scores, the decision, whether it was enforced, and the review outcome. Audit rows are stored apart from projects, so they survive deletion of the project or the account. No automatic expiry is set, which satisfies the 180-day minimum. **No frames or pixels are stored.** Holding possible CSAM carries its own legal duties, and scores are enough to audit a decision.

3.8 Rollout

The gate shipped in *shadow* mode (score and log, never stop), then switched to *enforce*. Of the 3,369 screens in the window, 10 ran in shadow and 3,359 in enforce.

4. What happened

4.1 The first model failed in shadow mode

The first primary model was AdamCodd/vit-base-nsfw-detector, whose card states it was trained to be restrictive: "if the image shows cleavage or too much skin, it will be classified as NSFW" [AdamCodd]. In 8 shadow screens on production, it would have blocked 2 jobs and sent 2 more to review. One was an AI-generated meme video with revealing outfits and nothing explicit. Its source screen flagged three frames at 0.957, 0.946 and 0.880, which under the rule of the time (2 frames at or above 0.90) meant *blocked*. We replaced the model the same day.

Measured on our own footage with quantized exports *(engineering log, 2026-09-13; scores only, no frame viewed)*:

	Safe showcase footage, 114 frames: max score	Explicit test video, 40 frames: frames ≥ 0.90	Time per frame (laptop CPU)
AdamCodd	0.280	15	230–411 ms
Falconsai	**0.003**	**17**	**66–141 ms**

Padding frames to a square, instead of squashing them, was tried and rejected: it reduced detection of landscape explicit content for both models. Rescreened under Falconsai, the meme's highest frame scored 0.285 on the source screen and 0.527 on the clip screen, with no frame at or above 0.90 *(engineering log)*.

The explicit test video was a publicly available adult video of consenting adults, used only to produce scores; no frames were stored or viewed.

4.2 The first 21 hours of enforcement

From the first Falconsai screen in enforce mode (2026-09-13 08:44 UTC), 56 jobs were screened in 21 hours and 8 were stopped. Two were our own explicit test video, run twice: held for review at a 25% share threshold, then blocked once the share was lowered to 15%. The other six were ordinary videos:

- Four were stopped on **a single frame**: 1 of 60, 1 of 32 and 1 of 26 frames on the source screen, and 1 of 15 on the clip screen. They were a large-budget challenge video, a compilation of a streamer being startled by jump scares, a children's nursery rhyme, and a long-form talk *(engineering log)*.
- A K-pop fan-camera video was blocked with 6 of 9 frames scoring between 0.936 and 0.994.
- A user upload was blocked with 2 of its 8 frames at or above 0.90.

Reviewers cleared five. The fan-camera video's owner deleted the held project before review, so no outcome was recorded; the engineering log records it as a false positive. That owner also stayed paused overnight with nothing on screen to explain why. This led directly to emailing the reviewer on every flag and notifying users of every outcome. The second classifier and source signals were added the following day.

4.3 Fifty days in numbers

Per screen (3,369):

Final decision	Screens	Share
Pass	3,230	95.9%
Review	93	2.8%
Blocked	11	0.3%
Incomplete (too few frames readable)	35	1.0%

Per job. 1,599 jobs were screened. Two of them were our own explicit test video. The figures below cover the other 1,597 (real traffic):

	Jobs
Drew a flag under at least one rule	177 (11.1%)
Held in enforce mode	32
Confirmed as violations by a human	**2**
Cleared as false positives by a human	**128**
Never reviewed (§4.5–4.6)	47

Of the 130 flagged jobs a human reviewed, 2 were violations: **a precision of about 1.5%.** Both were **uploads**, from two different accounts, and both scored 8 of 8 frames at or above 0.90 on the source screen. No YouTube, Twitch or Drive job was confirmed as a violation.

Scores on screens that passed (top primary score per screen, n = 3,230): median 0.0019, 90th percentile 0.226, 95th 0.695, 99th 0.997. The primary model is usually very confident a video is safe. But 126 passed screens (3.9%) contained a frame it scored at or above 0.80, which the second opinion or the source signals then allowed.

4.4 What the second classifier did

- It overturned the primary's flag in **126 screens** (124 from review to pass, 2 from block to pass).
- 276 primary-flagged frames were vetoed. Their second-opinion scores had a median of 0.133, a 90th percentile of 0.303 and a maximum of 0.398.
- The 139 frames it agreed on scored at least 0.403, with a median of 0.757.
- **91 jobs** were allowed by the veto alone. Humans later reviewed 68 and cleared all 68; 23 were not reviewed. None has been found to be a violation.

Disagreements between the two models were well separated, not borderline: the useful property in a veto.

4.5 What the source signals did

39 jobs were allowed by the source signals alone. Humans cleared 23; 16 were not reviewed. The flagged trusted videos were overwhelmingly widely watched mainstream content; the most viewed

had more than 1.2 billion views.

A case worth describing. One unreviewed trusted job was a video whose title openly refers to adult content. It met every trust condition (public, not age-restricted, about 1.4 million views, more than a year old). Each of its two screens flagged one frame, and **both classifiers agreed** on both (primary 0.956 and 0.905; second opinion 0.764 and 0.621). One frame per screen is below the block rule, so the result was *review*, and trust let the job run without a human decision.

We reviewed it after the data window closed. It is a podcast in which a guest describes how scenes are made and edited in the adult film industry. It contains no nudity and no sexual activity, and it is within our policy. Two lessons:

- **The two models agreeing is not proof.** Both flagged frames of a fully clothed conversation. The veto removes most false positives (§4.4), but not all of them.
- **A title is a reason to look, not a verdict.** The title pointed one way and the footage the other. In this case the platform's signals were right, but only a human could tell.

This review happened after the window, so it is not counted in the figures above.

4.6 What false positives look like

Frames are not stored, so we cannot say what each model saw. We can say which kinds of video drew flags that a human then cleared. From the titles of the 128 cleared jobs:

- **Children's animation and nursery rhymes:** 6 distinct titles were cleared, including several from the largest preschool channels on YouTube. Two more were allowed without review.
- **Large-budget challenge and stunt videos, and IRL streamer content:** the largest group, matching the service's traffic.
- **Sport:** football, professional wrestling, and a women's combat-sport bout.
- **Music videos, a fashion show, film and TV excerpts, dubbed dramas.**
- **Gaming:** a game trailer, shooter gameplay, a streamer's Minecraft session.
- **Food:** chocolate-making and eating videos.

96 of the 128 cleared jobs (75%) were flagged on one or two frames, 23 on three to five, and 9 on six or more. The most-flagged cleared jobs were sustained skin-heavy but non-sexual footage: a comedy video (55 flagged frames across its screens), a TV comedy show (22), a Halloween party video (16) and a fashion show (15).

Our working explanation, which we have not tested, is that video frames sit outside these models' training distribution: motion blur, transitions, close-ups dominated by skin tones, saturated animation, and compression artefacts.

4.7 Unreadable frames

35 screens (1.0%) were incomplete. 30 were frame extractions from stored uploads that timed out at 30 seconds per seek; 5 were remote streams that refused to open. Every one waited and retried rather than passing.

4.8 Cost

All inference runs on CPU on the same 8-vCPU server as the rest of the pipeline; no frame leaves it. Mean wall-clock per screen, including seeking and decoding, over the 3,339 screens that recorded timing: **source screens 30.8 s (1.17 s per frame)** and **board and clip screens 13.5 s (0.50 s per frame; median 7.5 s)**. Seeking dominates. Model inference alone was 66–141 ms per frame in local measurement *(engineering log)*.

Human cost: 131 reviewed jobs in 50 days (including the test video), about 2.6 a day. Each was a decision from timestamps and scores, plus watching the public source where needed.

5. Comparison with published benchmarks

Published accuracy figures disagree, because each evaluation uses different data and a different definition of "unsafe."

- **Freepik model card** [Freepik], accuracy by severity level:

Level	Freepik	Falconsai	AdamCodd
High	99.54%	97.92%	98.62%
Medium	97.02%	78.54%	91.65%
Low	98.31%	31.25%	89.66%
Neutral	99.87%	99.27%	98.37%

This matches what we saw. AdamCodd fires on mildly suggestive content; Falconsai mostly does not. - **ICIJ nuditag benchmark** [ICIJ], 670 files: AdamCodd fp32 had 7.05% false positives and a negative 99th-percentile score of 0.9475, leaving no clean threshold. Marqo had 1.04% false positives and 3.72% misses. - **KidsNanny** [Panchal et al.], UnsafeBench Sexual category (1,054 images, 683 unsafe): Falconsai had the highest precision (91.15%) but a recall of **40.70%**, against AdamCodd's 67.79% and Freepik's 76.13%. The authors note their results are not independently verified.

The last two points matter for anyone copying this design. **Our primary model is tuned toward precision.** Against a broad definition of sexual content that includes suggestive material, it may miss a large share. Our data cannot rule that out: we measured detection on one explicit test video, not on real traffic. We saw many false positives *despite* choosing a precise model, which says more about how hard real video is than about the model.

6. Recommendations for teams that process user video

1. **Screen before you process, and at more than one point.** A sparse pass over the source, then a denser pass over exactly what becomes output.
2. **Never decide on one frame.** Four of our first six false holds were a single frame. Require several high frames *and* a share of the video.
3. **Use two independently trained models, and let the second only veto.** Ours overturned 126 primary flags with a clean margin. Make a missing second score keep the flag, never drop it.

4. **Fail closed, with a coverage rule.** If you could not read enough of the video, you do not know what is in it.
5. **Ship in shadow mode first.** Our first model would have paused accounts over revealing outfits. We learned that from logs, not from users.
6. **Pick the model for your policy.** If your policy covers suggestive content, a precision-tuned model will under-flag it and a restrictive one will flood review. Decide which error you can afford, then measure on your own traffic.
7. **Store scores and timestamps, not pixels.** Enough to audit a decision, without holding material you must not hold.
8. **Remember human decisions per source.** Popular videos are submitted again and again; clearing one once saves review time and repeated false holds.
9. **Treat platform signals as evidence, not clearance.** "Not age-restricted" is the platform's policy, not yours. Review a sample of what trust lets through. A title that mentions adult content is a reason to look, not a verdict: in our one such case the footage was a clothed conversation.
10. **Tell the user.** A silent hold reads as a broken product. Name the rule, never the scores.
11. **Check licences.** NudeNet, a body-part detector, was rejected because of its AGPL-3.0 licence and YOLOv8 base [NudeNet], a poor fit for a closed commercial service.

For creators who clip other people's video. A screening tool may occasionally hold an ordinary video, including cartoons and sport. A hold is not an accusation, and a human clears false ones. Separately, each platform you publish to applies its own sexual-content rules to each short clip, and a clip cut from one suggestive moment carries that moment without the context around it.

7. Limitations

- **Recall on real traffic is unmeasured.** Unflagged jobs are not reviewed one by one, so we do not know how many violations passed without a flag.
- **Very few positives.** Two confirmed violations on real traffic support no statistical claim about detection.
- **No frames were kept,** so false positives are described by video type, not by what the model saw.
- **47 flagged real-traffic jobs were never reviewed** (vetoed, trusted, or deleted by their owner).
- **Rules changed during the window.** 8 screens used AdamCodd (review line 0.60), 3 used a 25% block share, and 97 ran before the second classifier existed.
- **One service, one traffic mix.** About 89% of screens were public YouTube videos, mostly entertainment. A service with more uploads would see a different balance.
- **Sexual content only.** The gate does not screen violence, self-harm or other categories.
- **Conflict of interest.** The author runs the service studied (see Declarations).

8. Data and reproducibility

Data files (CC BY 4.0, published with this note; no user IDs, emails, video titles or timestamps finer than one day):

File	Contents
`screens_daily.csv`	Screens and frames scored per day, by gate, source type, model, mode, primary and final decision, why allowed, and whether enforced. Sums to 3,369 screens and 88,822 frames.
`flagged_jobs.csv`	One row per flagged job (179): first-flag date, source type, model, worst primary and final decision, why allowed, held, review outcome, flagged frames, frames scored, internal-test flag.
`max_score_histogram.csv`	Each screen's top primary score, binned, by final decision.
`second_opinion_scores.csv`	Second-opinion scores for primary-flagged frames, vetoed vs counted, in 0.05 bins.

Models (pinned by repository revision and file SHA-256):

- Falconsai/nsfw_image_detection, ONNX export `onnx-community/nsfw_image_detection-ONNX` at revision `1ceb3c7fe1e9f3f2507e6df577437f23a9149fd5`, file `onnx/model_quantized.onnx`, SHA-256 `d9afb1e057104e6cc8616d174f0f6a8b8b0389c839eea7d5efa4bdf1a77efd27`
- Marqo/nsfw-image-detection-384, ONNX export `ICIJ/nsfw-image-detection-384-onnx` at revision `4f63fac119fb24a9d98393e57549b1427530ca52`, file `marqo-nsfw-384.onnx`, SHA-256 `50256dde930a4ceeb8953246ba7345158e02cc0a201a173402ab3e6e3a76f40a`

Software: Node.js, onnxruntime-node (package range ^1.19.0), ffmpeg. The service's source code is not public.

9. Ethics

No explicit material was viewed to tune thresholds; calibration used scores only. Screening runs on our own hardware, and frames are never sent to a third party. By policy, suspected CSAM is never sent to an external analysis service and is reported to the NCMEC CyberTipline and to Indian authorities. Users are told the outcome of every review. This note and its data contain no user-identifying information.

Declarations

- **Competing interests:** The author is the founder of Zoupyu (https://zoupyu.com), the commercial service studied. No product is offered or promoted here.
- **Funding:** None.
- **Data availability:** Aggregate data is published with this note (§8). Per-user records are not released, for privacy.

How to cite

Chauhan, V. (2026). *Screening User-Supplied Video for Sexual Content Before Processing: A Two-Classifier, Human-Reviewed Gate and 50 Days of Production Data* (Version 1.0) [Technical note]. Zenodo. https://doi.org/10.5281/zenodo.23880340

References

- [AdamCodd] AdamCodd. *vit-base-nsfw-detector*, model card. https://huggingface.co/AdamCodd/vit-base-nsfw-detector
- [Falconsai] Falconsai. *nsfw_image_detection*, model card. https://huggingface.co/Falconsai/nsfw_image_detection
- [Freepik] Freepik. *nsfw_image_detector*, model card. https://huggingface.co/Freepik/nsfw_image_detector
- [FTC] US Federal Trade Commission. *Complying with the TAKE IT DOWN Act*. https://www.ftc.gov/business-guidance/resources/complying-take-it-down-act
- [ICIJ] International Consortium of Investigative Journalists. *nuditag*, BENCHMARK.md. https://github.com/ICIJ/nuditag
- [ICIJ-ONNX] ICIJ. *nsfw-image-detection-384-onnx*. https://huggingface.co/ICIJ/nsfw-image-detection-384-onnx
- [IT Rules] Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, Rule 3(1)(g). https://indiankanoon.org/doc/45988656/
- [Khaitan] Khaitan & Co. *MeitY notifies the IT Amendment Rules, 2026*. https://www.khaitanco.com/thought-leadership/MeitY-notifies-the-IT-Amendment-Rules-2026
- [Marqo] Marqo. *nsfw-image-detection-384*, model card. https://huggingface.co/Marqo/nsfw-image-detection-384
- [NudeNet] notAI-tech. *NudeNet*. https://github.com/notai-tech/nudenet
- [Panchal et al.] Panchal, V., Talsaniya, T., Patel, P., Patel, M. (2026). *KidsNanny: A Two-Stage Multimodal Content Moderation Pipeline Integrating Visual Classification, Object Detection, OCR, and Contextual Reasoning for Child Safety*. arXiv:2603.16181. https://arxiv.org/abs/2603.16181