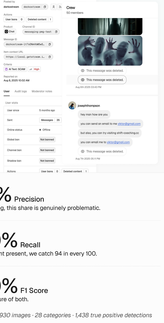
stream



AI Moderation Benchmark Report

Image & Live Video Moderation
Performance Across 28 Content Categories

01 The Numbers Upfront

91.36% Precision
Of everything we flag, this share is genuinely problematic.

94.30% Recall
Of all harmful content present, we catch 94 in every 100.

92.80% F1 Score
The balanced measure of both.

Benchmark sample: 930 images · 28 categories · 1,438 true positive detections

02 Two fears walk into every Trust & Safety review

The first is existential.
A piece of CSAM that slips into a live stream. A self-harm video that goes viral before a human reviewer can act. A hate symbol that sits in a public chat room for six hours. Any one of those can end a platform's relationship with regulators, advertisers, and the public at the same time.

The second fear is commercial.
Over-moderate innocent content, and you suspend legitimate creators, train your community to distrust the platform, and hand competitors an easy pitch. Both fears are real. They pull in opposite directions. And every moderation infrastructure decision you make is a bet on where to land between them.

Good moderation infrastructure measures itself against both fears explicitly, and publishes the results. That's what this report does.

We tested Stream's moderation engine on 930 images across 28 content categories. We're publishing precision, recall, and F1 scores for every single category, including the three where we fall short of 90%. T&S teams deserve evidence, not vendor slides.

03 What the headline numbers actually mean

91.36% Precision
The number your creators care about. Fewer than 9 in every 100 flagged items turned out to be safe content incorrectly held. Your moderation queue stays actionable. Your false-strike rate stays low.

94.30% Recall
The number your legal team cares about. For every 100 pieces of truly harmful content that enters your platform, Stream's engine catches 94 of them automatically before a human ever has to see them.

92.80% F1 Score
The number that tells you neither metric was achieved at the expense of the other.

136 false positives. 87 false negatives. Across 930 images.

04 How the benchmark was built

The sample was designed to reflect what a real platform actually encounters — not a lab environment optimized to make the numbers look good.

Safe content dominates, as it does on any live platform. Problematic categories appear in realistic proportions rather than artificially equal class splits. Edge cases like artistic nudity, educational content depicting harm, borderline violence were deliberately included to test whether the engine discriminates rather than simply over-flags. A moderation system that achieves 100% recall by flagging 100% of everything isn't a moderation system. It's a content blackout.

930 Images evaluated

28 Content categories

1,438 True positive detections

87 False negatives (missed content)

05 The categories where failure isn't an option

Not every category carries the same operational weight. A false positive on swimwear is a UX annoyance. A false negative on CSAM or a missed self-harm image is a regulatory event, a press cycle, and potentially a platform-ending moment. The categories below are the ones T&S teams most commonly define as non-negotiable.

Six categories achieved **100% precision** and **100% recall**. Zero false positives. Zero false negatives.

`SELF_HARM` `PHYSICAL_VIOLENCE` `OFFENSIVE_GESTURES` `ADULT_TOYS` `DRUGS_AND_CONTROLLED_SUBSTANCES` `TOBACCO_AND_VAPING`

Eight additional categories achieved 100% recall — prioritizing comprehensive catch rate on content where a miss is the more dangerous failure mode.

Category	Precision	Recall	F1
Child safety (MINORS)	93.75%	93.75%	93.75%
Self-harm	100%	100%	100%
Physical violence	100%	100%	100%
Extreme violence & gore	88.89%	100%	94.12%
Hate speech & symbols	92.31%	100%	96.00%
Graphic nudity	98.35%	83.22%	90.15%
Sexual Activity	88.64%	86.67%	87.64%
Drugs & controlled substances	100%	100%	100%
Weapons	88.89%	80.00%	84.21%

The MINORS category is worth a closer look.
At 93.75% on both precision and recall, the scores are intentionally symmetrical because both failure modes are unacceptable. Missing genuine child-safety content is catastrophic. But a system that over-flags adult users as minors creates its own harms and its own liability. Neither is acceptable. The engine treats them equally.

06 What this means for live video

Live video is where the stakes are highest and the response window is smallest. A harmful image in a static gallery has time for human review. A harmful moment in a live stream does not.

Stream's live video moderation works by **extracting key frames** from the stream at regular intervals and **scoring each one through the same image moderation model** benchmarked in this report.

Frame extraction frequency is configurable. For platforms with higher risk profiles, live gaming, creator streams, interactive video, sampling rates can be tuned to increase coverage. The model accuracy stays constant; coverage is a policy variable.

07 Full results across all 28 categories

Categories in bold achieved 100% precision and 100% recall.

Category	Precision	Recall	F1	TP	FP	FN
SELF_HARM	100%	100%	100%	10	0	0
PHYSICAL_VIOLENCE	100%	100%	100%	10	0	0
DRUGS_AND_CONTROLLED_SUBSTANCES	100%	100%	100%	12	0	0
OFFENSIVE_GESTURES	100%	100%	100%	11	0	0
TOBACCO_AND_VAPING	100%	100%	100%	11	0	0
ADULT_TOYS	100%	100%	100%	16	0	0
HATE_SPEECH_AND_SYMBOLS	92.31%	100%	96.00%	12	1	0
NUDITY_IN_ART_OR_EDUCATION	92.31%	100%	96.00%	12	1	0
GAMBLING	92.31%	100%	96.00%	12	1	0
MALE_SWIMWEAR_OR_UNDERWEAR	93.33%	100%	96.55%	14	1	0
QR_CODE	93.33%	100%	96.55%	14	1	0
VISUALLY_DISTURBING	100%	90.91%	95.24%	10	0	1
SAFE_CONTENT	93.99%	96.03%	95.00%	266	17	11
PARTIAL_NUDITY	94.87%	93.67%	94.27%	74	4	6
VISUALLY_IDENTIFIABLE_PERSON	92.39%	95.94%	94.13%	425	35	18
EXTREME_VIOLENCE_AND_GORE	88.89%	100%	94.12%	16	2	0
MINORS	93.75%	93.75%	93.75%	15	1	1
FEMALE_SWIMWEAR_OR_UNDERWEAR	88.00%	97.78%	92.63%	44	6	1
SENSITIVE_PERSONAL_INFORMATION	96.15%	86.21%	90.91%	25	1	4
GRAPHIC_NUDITY	98.35%	83.22%	90.15%	119	2	24
REVEALING_ATTIRE	84.40%	95.83%	89.76%	92	17	4
ALCOHOL	88.89%	88.89%	88.89%	8	1	1
BODY_WASTE	91.67%	84.62%	88.00%	11	1	2
SEXUAL_ACTIVITY	88.64%	86.67%	87.64%	39	5	6
SUGGESTIVE_POSES_AND_EROTICA	78.79%	98.11%	87.39%	52	14	1
BARE_MALE_CHEST	81.82%	93.75%	87.38%	90	20	6
WEAPONS	88.89%	80.00%	84.21%	8	1	2
BAR_CODE	71.43%	100%	83.33%	10	4	0
OVERALL	91.36%	94.30%	92.80%	1,438	136	87

08 Where we fall short and why

Any moderation vendor who claims no weaknesses either hasn't tested carefully or isn't sharing the results. Three categories in this benchmark scored below 90% F1. Here's what's behind each one.

`BAR_CODE` **71.43% precision, 100% recall**
The engine catches every barcode present, but **occasionally misidentifies QR-like patterns as barcodes.** This is a precision problem, not a safety problem. For most platforms, barcodes aren't a content risk category in the first place — and the threshold is tunable if you want to suppress false positives entirely.

`SUGGESTIVE_POSES_AND_EROTICA` **78.79% precision, 98.11% recall**
This category lives on the boundary of subjective policy. Different platforms draw the line between suggestive and explicit in different places, and the engine reflects that genuine ambiguity. **High recall means Stream errs toward flagging rather than missing.** For platforms enforcing in this category, human moderator review on flagged content is recommended for final calls, and per-platform threshold configuration is available.

`BARE_MALE_CHEST` **81.82% precision, 93.75% recall**
An intentionally permissive category for most platforms. Fitness apps, sports communities, and outdoor platforms typically want this allowed, not flagged. For platforms that don't enforce in this category, the false positive rate is simply irrelevant. Policy-tunable per customer.

The pattern across all three: these are **policy ambiguity problems, not detection failures.** Stream exposes per-category threshold controls so each platform can tune enforcement to match its community standards rather than inheriting a one-size-fits-all policy.

09 What to take away

Stream's moderation engine catches **94 out of every 100 pieces** of harmful **content** across a realistic 28-category benchmark, with perfect scores on six of the most safety-critical categories and zero misses on eight more.

The remaining error is concentrated in categories shaped by policy subjectivity, not detection limits. All of it is tunable.

If you want to see how these numbers hold up on your content, your categories, and your edge cases, we'll run the benchmark.

Reach out at **getstream.io/contact**.