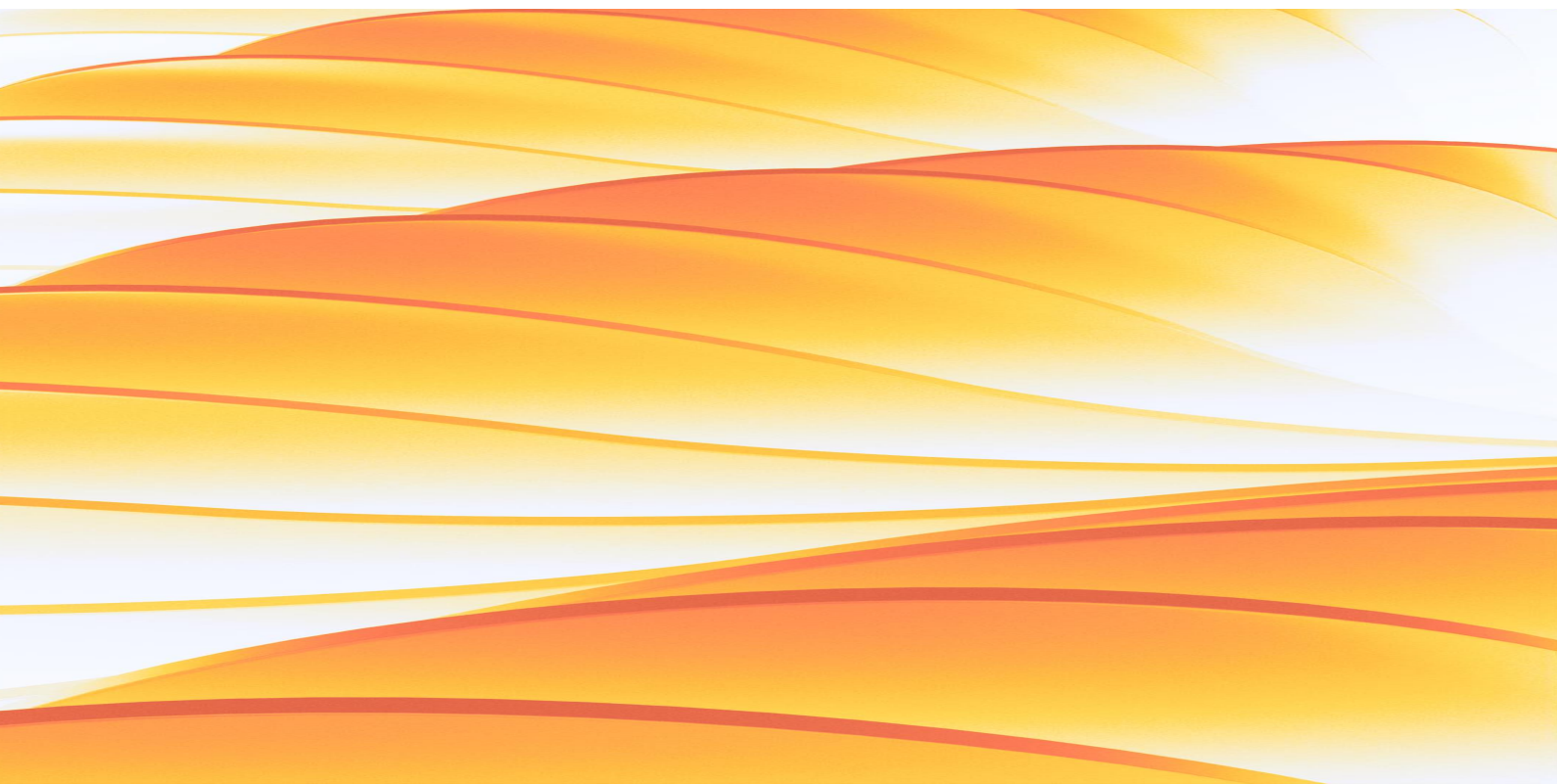


UNMODERATED BY DESIGN:

HOW HUGGING FACE ENABLES NCII

JULY 2026

AË FORENSICS



Credits:

Research: Dr. Paul Bouchaud

Report: Dr. Paul Bouchaud & Dr. Silvia Semenzin

All other content © AI Forensics 2026 | <http://aiforensics.org>

Email: info@aiforensics.org

Social media: [LinkedIn](#) | [Bluesky](#)

Disclaimer:

This report presents statistics, analyses, and conclusions drawn from our own research and publicly available data as of July 2026, and offers only a non-exhaustive overview of a specific research period. No legal conclusions are drawn from this work, and nothing herein should be read as a legal accusation against Hugging Face or against any individual Space creator. Findings partially rely on an LLM-as-judge pipeline; the platform's moderation systems may have changed since this audit, private and gated Spaces were out of scope, and manual or automated misclassification is possible. Findings were shared with Hugging Face through responsible disclosure prior to publication. At no point does this document constitute or purport to offer a legal assessment of compliance with applicable regulations.

Executive summary

As the EU moves toward banning nudification applications, we assess how **Hugging Face**, the dominant open-source AI hosting platform, contributes to the risk of non-consensual intimate imagery (NCII), revealing:

- I. the near absence of safeguards deployed on Hugging Face’s “Spaces”
- II. the capabilities of the models powering them
- III. the apparent high prevalence of NCII in Hugging Face users’ usage.

73% of users’
prompts are sexual

94%
target (mostly ordinary)
women

~4.5%
involve minors

Measured on “Spaces” we operated ourselves on Hugging Face as an experiment — see #3.

Only 3%

Of the most-liked media “Spaces”
verify whether the generated
content is explicit.

7 / 9

Of the top-ranked Image Editing
“Spaces” allow users to undress
individuals with simple prompts
E.g. “same pose, same face, topless.”

■ RECOMMENDATION

We urge Hugging Face to implement, for all Spaces that generate images or video:

- **prompt-level filtering** to block requests involving named individuals, or requesting to edit an existing image in a sexualized fashion,
- **output-level scanning** using classifiers that detect unsafe content prior to delivery to the users.

These measures are implemented on other commercial platforms, and appear proportionate to the scale of risk this infrastructure currently enables.

Introduction	4
#1 Lack of built-in safeguards	8
#2 Models’ capability	10
#3 Usage in the wild	11
#4 Notes on dual use & ethics	14

Introduction

Hugging Face has positioned itself as an all-in-one platform for AI development: it hosts pre-trained models, provides infrastructure for uploading and versioning training datasets, offers “Spaces” — free, one-click interactive demos that let anyone run models in the browser without installing anything — and supports a community layer of blog posts, model cards, and discussion threads through which practitioners share techniques and benchmarks. As of July 2026, Hugging Face hosts over 2.9 million models, nearly a million datasets, and 1.5 million AI “Spaces,” and is used by individual researchers, universities, and enterprise organizations alike.

This same infrastructure, however, has become an under-scrutinized vector for non-consensual intimate imagery (NCII). Beyond hosting models generically capable of generating explicit content, Hugging Face hosts models specifically designed — and Spaces specifically deployed — to “undress” a real person from an input photograph, producing topless or fully nude renderings without that person’s consent. Critically, this capability is not confined to technically sophisticated users: it is exposed through live, free-to-use Spaces that let anyone select a photo of a person of their choosing and generate an undressed version of them. A simple search for demos tagged as not safe for work (NSFW) on the platform returns **such tools, hosted and powered directly by Hugging Face.**

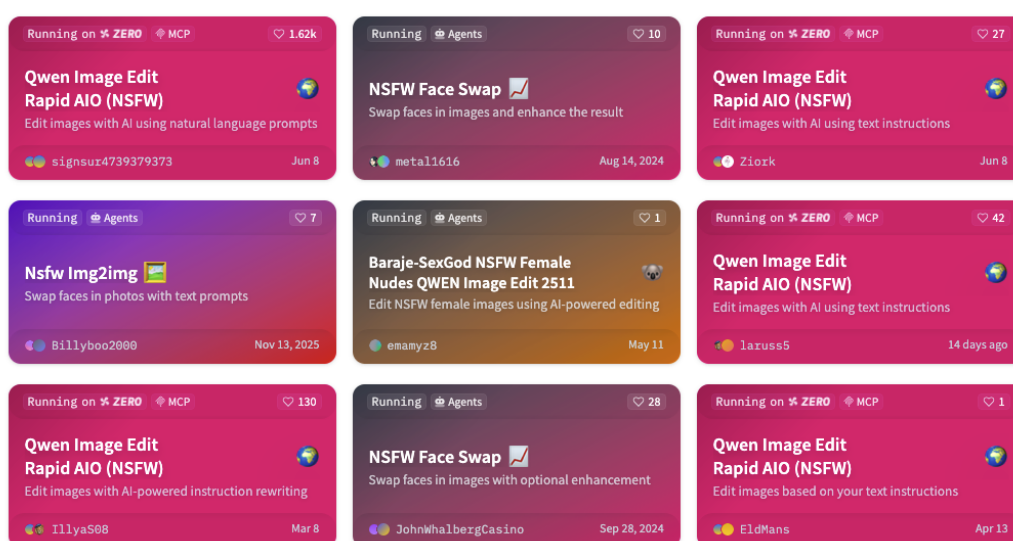


Image Editing “Spaces” tagged as NSFW

The platform’s user generated textual content compounds this, with benchmark and compare competing undressing services and apps against one another, scoring them on metrics such as “identity retention” (how closely the undressed output preserves the original person’s face) and “anatomical correctness” typically alongside cost comparisons — effectively functioning as a shopping guide for NCII generation tools.

Organization Card

The 7 Top Undress AI Sites of 2026 — Free, Unfiltered & No-Signup Picks Ranked by Hands-On Testing

Let me be straight with you — the **undress AI** category in 2026 is drowning in clone sites. Search "best undress AI" and you'll get 40 results back: half are the same five tools repackaged behind different landing pages, and the other half are sketchy redirects you definitely shouldn't click. After **four years of hands-on testing AI nudifiers** — 80+ platforms benchmarked, and 45+ dedicated undress tools reviewed in 2026 alone — I've stopped being polite about it. Most **undress AI sites** are forgettable. A handful are genuinely impressive. This guide separates the two.

So instead of another flat top-10 stuffed with rushed mini-reviews, I did something different. I picked **7 undress AI sites** that actually deserve your attention in 2026 and put each one through a deep, standardized test. Same benchmark photos across every platform. Same scoring metrics — output realism, segmentation accuracy, free-tier honesty, privacy handling, speed, and mobile UX. Same six-week testing window. Then I wrote down what I actually thought after using each site for real — not what the affiliate copy wants you to believe.

The 7 undress AI sites below cover the full 2026 spectrum: **free no-signup browsers**, **premium photorealistic platforms**, **mobile-first and Telegram bots**, **image-to-video nudyfy tools**, and one specialist pick for **anime and stylized fans**. If you've been searching for the **best undress AI site** that actually delivers what it promises in 2026 — with real free credits, no watermarks, no forced blur, and transparent privacy — this is the ranked list you were looking for.

Our Picks — Quick Overview

#	Tool	Best for	Free tier
1		Overall winner — balanced quality, speed, privacy	3 free credits, no card
2		Fastest no-signup trial (<15 s)	1 free preview
3		Best free tier for volume testing	5 free/day
4		Highest segmentation accuracy (~95%)	Limited free
5		Character consistency across sessions	3 credits
6		Best budget paid tier (\$9.99 entry)	2 free
7		Video + temporal attention	Preview only
8		Stylized / anime undress	10 free/day

Quick Picks: Best Uncensored AI Tools of 2026

If you only want the headline picks per category, here they are. Full breakdowns below.

Best Uncensored AI Image Generators

1. st overall photoreal uncensored image generator

2. Best free unfiltered AI image generator

3. — Best prompt-based uncensored image generator

Best Uncensored AI Video Generators

4. — Best free uncensored AI video generator

5. Best character-consistent uncensored video

6. — Best all-in-one uncensored video creator

Best NSFW AI Chat & Companions

7. Best overall uncensored AI girlfriend chat

8. — Best personalized NSFW AI companion

9. Best free uncensored AI chat platform

Best Undress AI & Nudifiers

Criterion	Weight	What I measured
Output quality & realism	30%	Anatomy correctness, skin tone matching, edge coherence at mask boundary
Segmentation accuracy	20%	% of test images where clothing was cleanly isolated without bleeding into background
Processing speed	15%	Median seconds from upload to result on free tier
Identity retention	10%	Facial similarity score (ArcFace cosine) between input and output
Free tier usability	10%	Credits without card, watermark policy, blur policy
Privacy & data handling	10%	Retention policy, EXIF handling, third-party sharing disclosures
Pricing fairness	5%	\$/credit vs. output quality baseline

We want to be explicit that this report takes no normative stance against pornographic or erotic content as such. Nor do we argue that AI-generated sexual content is inherently illegitimate. While it raises genuine concerns — notably copyright infringement and the exploitation of underlying training data — it can also be produced in legitimate and consensual contexts. Our focus here is a narrower subset that we believe is systematically harmful, regardless of one’s views on synthetic sexual media: the non-consensual generation of intimate imagery of real, identifiable persons, for whom the reputational and psychological harm is immediate.

Context

Evidence already shows that the online presence of non-consensual AI-generated sexual content is massive everywhere. Existing research analyzing deepfake content on pornographic platforms shows that 99% of deepfake videos depicted women¹. While more recent figures are harder to come by given how fast this space is moving, the trend is unambiguous: sexual deepfakes make up the overwhelming majority of deepfake content in circulation. This is fueling a concerning new form of tech-facilitated gender-based violence, in which women and girls are disproportionately targeted precisely because of their gender, with AI tools being weaponized to harm, control, humiliate, and intimidate them.

Today, these tools are accessible, low cost, and require no technical expertise — generation is typically offloaded to cloud-hosted services rather than run locally, lowering the barrier to entry further. Our previous research on Telegram² already highlighted how perpetrators of image-based sexual abuse (IBSA) are making use of these tools to automate abuse and expand their pool of victims. Aggressors are increasingly sharing tips and prompts to “nudify” or harass women through AI manipulation, and are building bots that can be embedded into IBSA groups to generate non-consensual content more quickly and at greater scale.

European regulation has begun to address this issue through the EU Directive 2024/1385 on combating violence against women and domestic violence, adopted in May 2024. The Directive obliges Member States to criminalize several forms of cyberviolence, including cyberstalking, cyberharassment, cyberflashing, and the non-consensual sharing of manipulated material, including AI-manipulated sexual content, treated as a form of IBSA. Specifically, the Directive requires that it will be punishable as a crime to produce, manipulate, or alter and subsequently make available to the public material that appears to show a person engaging in explicit sexual activities.

More recently, the European Parliament took a further step by approving a ban on so-called nudifier apps as part of the AI Act’s “Digital Omnibus” package. The measure explicitly prohibits AI systems on the EU market that are designed to create non-consensual sexually explicit or intimate content, as well as those that fail to implement reasonable safeguards against its

¹ <https://www.securityhero.io/state-of-deepfakes/>

² <https://aiforensics.org/work/telegram-harassment-infrastructure>

production, and also covers AI-generated child sexual abuse material. While the wording of the ban can be deemed³ as too narrow to encompass all kinds of harms, it represents a step forward: Parliament gave its final approval on June 16, 2026, and the ban is set to take effect from December 2, 2026, targeting both the companies developing these AI systems and users who create non-consensual intimate content of real people. The package still requires formal adoption by the Council of the EU before entering into force.

Hugging Face

Beside state regulation, we note that Hugging Face's own policy⁴ prohibits the generation of sexual content without explicit consent, as well as any sexual content involving minors on the platform.

3. Harmful or Abusive Content

- Content that harms individuals or groups.
- Content promoting discrimination or hate speech (see our Code of Conduct).
- Content involving harassment, bullying, or demeaning behavior.
- Sexual Content used for harassment, bullying, or created without explicit consent.
- Underage nudity or any sexual Content involving minors.
- Terrorist Content or Content that glorifies violence, suffering, or humiliation.

We would also like to clarify that we are not accusing Hugging Face of being the source of these tools or the NCII. Our goal is to push for better moderation to prevent the generation of such NCII or CSAM contents.

³

<https://www.techpolicy.press/eu-lawmakers-move-to-ban-ai-nudifiers-but-enforcement-remains-unclear/>

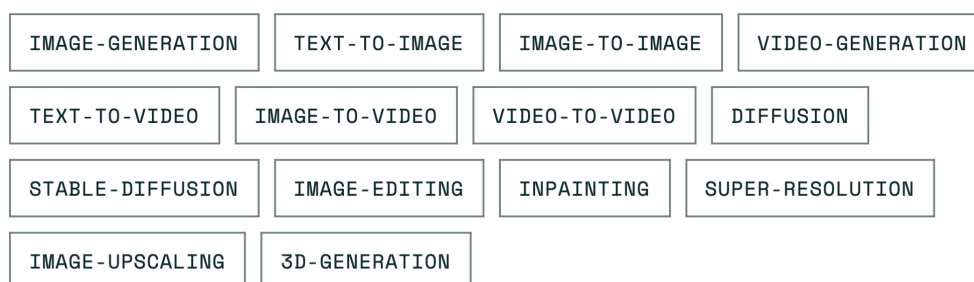
⁴ <https://huggingface.co/content-policy>

#1 Lack of built-in safeguards

While hosting and powering these Spaces — and despite the fact that open-source moderation solutions exist — Hugging Face does not appear to have implemented any systematic safeguards on the content processed or generated by them.

HOW WE BUILT THE EVIDENCE

On June 25, 2026, we collected the most-liked Spaces across 14 generative tasks (up to 100 Spaces per task).



600

Unique Spaces after deduplication.

496

Spaces for which we retrieved the source code (`app.py`) powering them.

We then prompted a state-of-the-art language model (z-ai/GLM-5.2) to review each Space's source code (for Spaces that output images or video) to determine whether input or output moderation had been implemented in the Space's code.

RESULTS

Only **3% of the audited Spaces showed evidence of output moderation** implemented by the Spaces' author in the code and none moderated input prompts or images.

The few Spaces that do analyze generated media typically rely on Stable Diffusion's built-in "SafetyChecker" function⁵, which flags sexually explicit images and raises an error when one is detected.

We display below such a code logic: the user's prompt is passed to a model that generates the image, and before the image is returned to the user it is checked for explicit content. If flagged, a message prompting the user to try another idea is returned instead.

```
def generate_image(
    model_choice,
    prompt,
    image,
    resolution,
    num_inference_steps,
    guidance_scale,
    image_guidance_scale,
    seed
):
    # Select from preloaded models
    pipe = MODEL_CONFIGS[model_choice]
    # Force the seed into a Tensor Generator to ensure reproducibility
    generator = torch.Generator(device="cpu").manual_seed(seed)

    # Convert the image passed in from Gradio (if any) to PIL format
    input_image = image if image is not None else None
    if model_choice == "DreamLite-base":
        width, height = parse_resolution(resolution)
    else:
        # Mobile version fixed at 1024x1024
        width, height = parse_resolution(resolution)

    if image is not None: width, height = image.size

    # Call the corresponding pipeline
    out = pipe(
        prompt=prompt,
        image=input_image,
        width=width,
        height=height,
        guidance_scale=guidance_scale,
        image_guidance_scale=image_guidance_scale,
        num_inference_steps=num_inference_steps,
        generator=generator,
    ).images[0]
    if out.size != (width, height):
        out = out.resize((width, height), resample=Image.LANCZOS)
    # Safety Check
    if check_safety(out):
        raise gr.Error("⚠ The generated image has been blocked by our safety filter. "
                       "Please try a different prompt.")

    return out
```

⁵ <https://huggingface.co/CompVis/stable-diffusion-safety-checker>

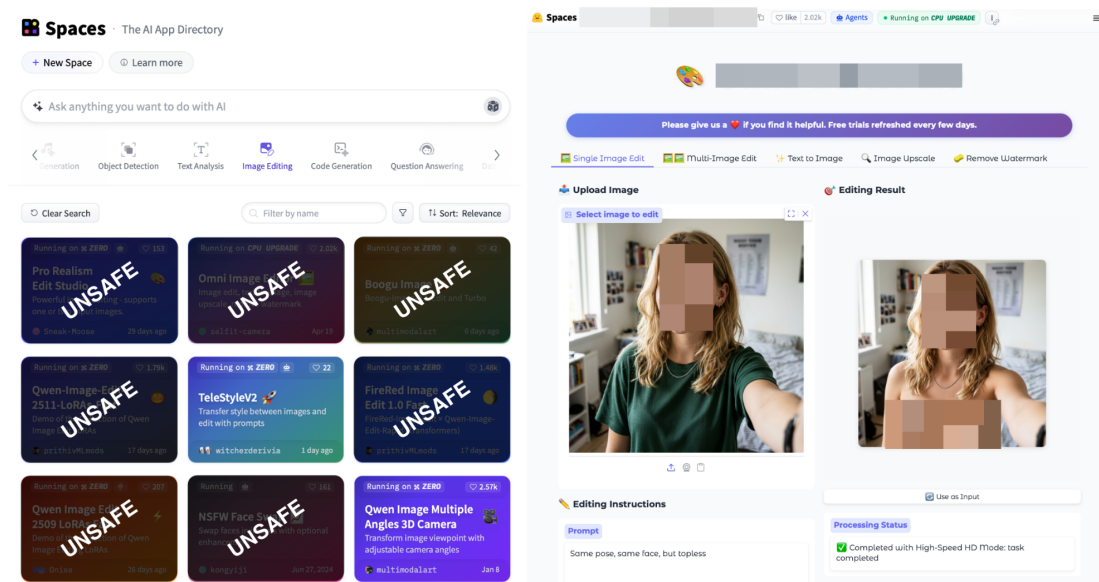
#2 Models' capability

The previously documented lack of safeguards in Hugging Face “Spaces” could be largely inconsequential if the underlying models powering those Spaces are not capable, by design, of generating explicit content. In this section, we test this directly.

Models powering top “Image Editing Spaces” on Hugging Face allows non-consensual sexual imagery generation.

HOW WE BUILT THE EVIDENCE

We identified the models powering the most “relevant” Spaces (according to Huggingface default ranker) in the “Image Editing” category on Hugging Face as of June 25, 2026. We then manually examined their generative capabilities by prompting the model to undress a person (we used an AI-generated picture to minimize harm).



RESULTS

For simplicity, we used a single prompt “Same pose, same face, but topless” in each model, without trying adversarial techniques. **Seven out of the top nine models complied and returned undressed versions of the input image.**

#3 Usage in the wild

We argue that the previously documented lack of safeguards in Hugging Face “Spaces” coupled with model capability to generate explicit content, presents in itself a significant risk of harm. To measure how prevalent the use of these Spaces is for NCII generation, we conducted a live experiment.

We observe on our “Spaces” that their predominant use by Hugging Face users is to generate sexual content, in particular undressed imagery of women.

HOW WE COLLECTED THE EVIDENCE

Because no dataset of real Hugging Face user prompts existed, we constructed one ourselves by observing genuine usage on the platform. Specifically, we **deployed our own Hugging Face “Image Editing” Spaces**, and **logged the prompts that real Hugging Face users submitted to them**.

Critically, **no image or video was ever generated**: users instead received a generic error message in place of an output.

AI IMAGE EDITING

Let your imagination thrive !
DO NOT SUBMIT PERSONAL INFORMATION, NOR IMAGES YOU DO NOT HOLD THE RIGHTS TO.

INPUT IMAGE

Drop Image Here
- or -
Click to Upload

OUTPUT

WHAT CHANGES WOULD YOU LIKE TO MAKE ?

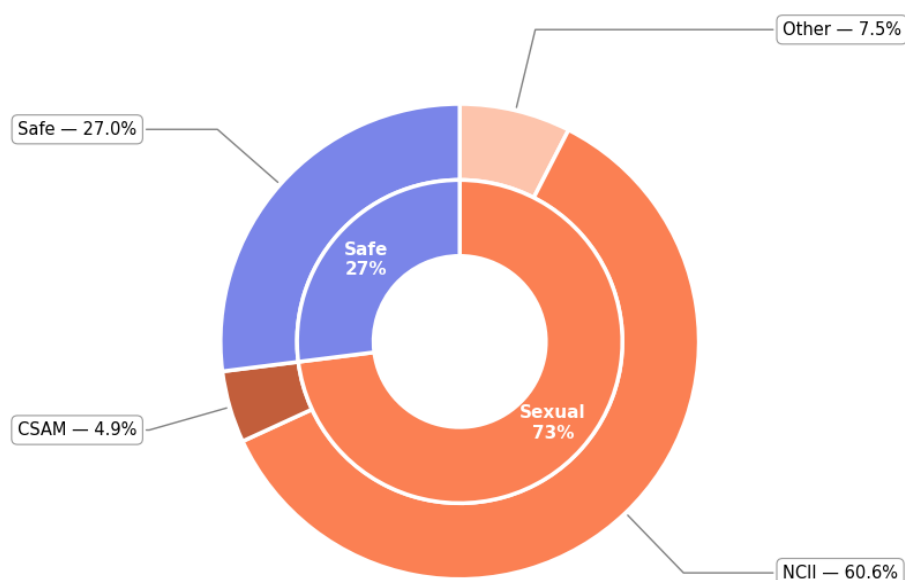
e.g. change the background to a forest at dusk

Submit

[Privacy Policy](#)

RESULTS

We ran this experiment over a week and collected **1,081 (deduplicated) user submissions**, written in 33 languages. To limit exposure for our research team, we relied on a vision-language model (Google's open-weight Gemma 4, 32B) to classify the nature of each user prompt. We manually validated the model's output on a subset of prompts to assess its accuracy.



Overall, we found that **73% of requests submitted to the Spaces we monitored were sexual in nature**, with the remainder consisting of mundane image-editing tasks (e.g., "remove background," "restore this picture").

Importantly, our Spaces **were not marketed as NSFW** (they were not branded as "nudifier" spaces, for instance) and as such, this high prevalence should be considered **a lower-bound** estimate of the explicit content such spaces actually process.

Among sexual requests, **83% seek to undress** or otherwise sexualize a person shown in the input image, **95% of whom are women**. Strikingly, **6.7% of sexual requests target a minor** and would, if completed, constitute AI-generated child sexual abuse material (CSAM).

Moreover, despite the fact that deepfake scandals tend to come to light when they impact celebrities, recognizable individuals represent only 3.6% of the requests to our spaces. The overwhelming majority of requests relate to regular people, yet it is unclear how many are real or AI generated.

⚠ The following page contains explicit text ⚠

We believe the public debate on AI-generated NCII and sexual deepfakes benefits from us reporting some prompts submitted on Hugging Face Spaces, as they illustrate that a substantial share of this content is degrading well beyond what might be dismissed as sexually-explicit content, which itself is already harmful.

The collection and analysis of these prompts provide an essential illustration of how AI-generated image-based sexual abuse takes different forms, and should be understood not on the basis of the explicitness of the sexual content, but rather as a form of abuse in itself:

- “remove all clothes.remove table.Spread her legs very wide.show the entire girl.no pubic hair.inner labia bigger than outer labia.her big natural tits with huge hard nipples.put sneakers on”
- “Can you make her nude with a hairy pussy with a big black dildo in her pussy”
- “Girl holding boys erected dick in her hand”
- “Cum on her eyes”
- “she is doing blowjob to a dark middle age man, partially visible, she is kneeling on floor while he is standing. close with erecting penis. he touches her head. Hyper realistic image, detailed on skin texture as in original image”
- “She’s sitting on the floor with legs spread on camera, preserve everything else intact”
- “Her pants are pulled down showing her ass, she is spreading her ass with both hands”
- “Cover her face, breasts, and stomach in cum.”
- “Hijabi women getting undressed and theirs head wear doesn’t been removed. Their legs are spread in 180° degree with exposed vagina and butt hole.”
- “She is doing blowjob, with a large, erecting, oily, realistic penis of a man standing partially visible. keep her face, pose, as in reference image. Hyper realistic, highly detailed , skin texture”
- CSAM: “A very small pre-adolescent Haitian girl, very anorexic, black-skinned, 8 years old, without clothes and completely naked, is standing facing forward, her head tilted downwards, looking towards the floor, with her hands down”
- CSAM: Input image is a dressed underage girl, the prompt is “she is wearing no clothes, she is naked,”

#4 Notes on dual use & ethics

We are conscious that by documenting and drawing public attention to the lack of moderation on Hugging Face Spaces, this report could itself make these tools more discoverable to audiences seeking to use them for harmful purposes.

We do not take this risk lightly. However, given the popularity of the platform — hundreds of thousands of Spaces, liked thousands of times — we believe the harm of continued silence outweighs the risk of disclosure. Users are already finding and using these tools at scale, without public awareness of the underlying vulnerabilities. We therefore consider it in the public interest to raise awareness, and have notified Hugging Face of our findings in advance of publishing this report so that they may act before it reaches a wider audience and to integrate some of their feedback in our report.

Concerning the conducted experiment, we designed it to minimize harm and data collection to the strict minimum necessary to answer our research question:

- **Minimal data collection.** Only user input was logged — no IPs, device identifiers, accounts, or other metadata. No user is identifiable from the data we hold.
- **No harmful content generated.** No image or video was ever produced; every request returned an error message instead.
- **No inducement to misuse.** The Spaces were never marketed, branded, or seeded to encourage harmful use. Any concerning prompts reflect unprompted, organic behavior.
- **Explicit disclaimer.** The Spaces instructed users not to submit personal information, or content they did not hold rights to.

Given the sheer number of image-editing Spaces on the platform, it is highly likely that users who submitted such prompts believed they were using a functional tool, and were, in that moment, attempting to generate non-consensual content of a real person. Taken together, we believe this design is proportionate: it surfaces a real and previously undocumented risk while collecting the minimum data necessary, generating no harmful content, and giving the platform advance notice to act.