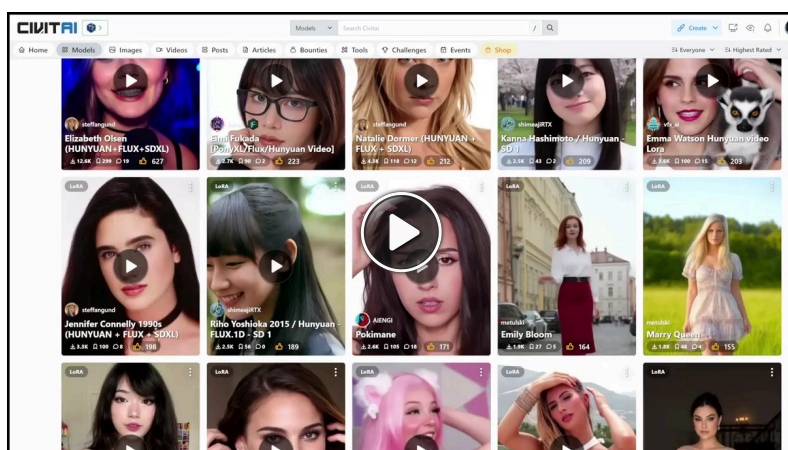


CivitAI Tightens Deepfake Rules Under Pressure From Mastercard and Visa

Originally published at <https://www.unite.ai/civitai-tightens-deepfake-rules-under-pressure-from-mastercard-and-visa/>



CivitAI, possibly the most popular AI model repository on the internet, has finally conceded to pressure from payment facilitators MasterCard and Visa to radically revise their policies on NSFW content – and particularly their TOS regarding celebrity LoRAs, one of the site's most popular user-submitted content streams, which enable people to use freely downloadable adjunct models such as [LoRAs](#) to create AI depictions (including [video depictions](#)) of well-known people.



CLICK TO PLAY VIDEO ON SITE

Click to play. Though Civit does not provide exact figures (and mislabeling/miscategorizing is not unusual, which would likely skew the figures), the number of celebrity AI models, primarily LoRAs, is clearly in the multiple hundreds or even thousands, and clearly dominated by female subjects. Source: civitai.com

Clearly operating under pressure in a [Twitch live-stream](#)** on behalf of the company, the company's Community Engagement Manager Alasdair Nicoll, himself a [creator](#) of (SFW) models at Civit, admitted that the changes have been forced upon the site by their payment processors' concerns about adult content, and depiction of real people. He also admits the likelihood that the primary forces behind these processors, Visa and MasterCard, are likely to demand even greater changes later:

'These are not changes that we wanted to make. This boils down to new and impending legislation. There's deepfake laws out there, there's AI porn laws...payment processors, and ultimately Visa and MasterCard are spooked; they don't want to be sued, and they're ultimately driving these changes.'

'Some of the other options that we were given were removing not-safe-for-work altogether; removing x and triple-x content off of CivitAI and on to a completely new platform, and then geo-blocking that platform, because over half the US states require some form of porn geo-blocking, [as well as] many Asian countries and the United Kingdom...'

'The third option was going full crypto, crypto payments only...so there were literally no good options for this.'

The Civit domain has been down periodically for revisions over the last few days, apparently to effect the changes. Though the site had already banned the use of NSFW themes in celebrity LoRA/model depictions, it is now impossible to browse the model section of Civit and see celebrity LoRA previews side-by-side with the very large number of generic NSFW models designed to produce mature content.

The [official announcement](#) states:

'Content tagged with real person names (like "Tom Cruise") or flagged as POI (real-person) resources will be hidden from feeds.'

In the Twitch session, Nicoll revealed further details of measures designed to protect famous figures and real people. Civit has always allowed real people to request that a Civit-hosted AI model depicting them be taken down, but now Nicoll alludes to a system that can prevent such images being re-uploaded after initial rejection, with the facility to identify a 'protected' personage even in images that the system has never seen before.

To this end, the site is now partnering with the Clavata AI moderation system – though the extent to which Clavata will be powering these new facilities is not yet clear.

Nicoll said:

'Tom Hanks has claimed his likeness from us, for example. Quite a lot of the adult actresses have; quite a lot of A-list actors and actresses have...'

'I think the first one that we ever had was Barbara Eden, her estate – she was one of the first to claim her likeness, which is sort of funny, because she's old.'*

Protected by Default?

Over the last couple of years, the AI VFX company Metaphysic (full disclosure: I worked for Metaphysic.ai from early 2022 until late 2024) attempted to create [a proprietary system](#) that would allow anyone to register their own likeness, though primarily aimed at Hollywood names concerned about AI-based hijacking of their identities, with support from actors [such as](#) Anne Hathaway, Octavia Spencer and Tom Hanks (with whom the company worked on the Robert Zemeckis outing *Here* [2024]).

Logically, the utility of the system would always depend on eventual case law; based on the measures Civit is now being forced to take, the [subscription-based service](#)[†] proposed by Metaphysic could be redundant in the face of the rapid growth of deepfake laws, and potential (free) coverage under common law. It's not currently known whether the Metaphysic Pro offering will transfer to the Double Negative VFX company, which [acquired Metaphysic's assets](#) last year.

In any case, it increasingly seems that global law and general market pressures are more likely to provide protection and remedies, as opposed to commercial solutions of this kind.

Boiling the Frog

A 2023 report by 404 Media [brought attention](#) to the proclivity of celeb and porn AI models at Civit, though the site's founder Justin Maier [downplayed](#) the relationship between user-contributed celebrity likenesses and their use in generating pornographic material.

Though Civit makes money by facilitating the on-site use of LoRAs and other user-supplied models, Nicoll is clear that this is not the primary concern motivating Visa and MasterCard to stipulate changes to the site, in order that it can continue to be monetized:

'Some people are saying that the reason that we're in this mess is because we allow generation. That doesn't come into it. The hosting of these models, the hosting of this content, is enough to draw the eye of Sauron.'

Community comment threads have marveled in recent years that Civit has been allowed to host celebrity likenesses. Aware of the possibility, perhaps inevitability of a clampdown, a number of initiatives to preserve LoRAs either removed by Civit or by their uploaders, have been proposed or implemented, including the (until now) rather neglected subreddit r/CivitaiArchives.

Though many have suggested that a torrent-based initiative is the natural solution, no well-followed domain seems yet to have emerged – and in any case, this would seem certain to move activity banned at Civit and elsewhere to the outermost margins of the internet; to walled gardens; and, most likely, to the dark net, since most of the frameworks that could accommodate banned likeness LoRAs (such as Reddit and Discord) either already ban such content or seem certain to ban it imminently.

At the moment, celebrity LoRAs can still be seen with some restrictions at Civit, though most of the generated content has been de-listed and will be excluded from casual discovery. What seems likely, one commenter suggested to Nicoll in the Twitch session, is that the crackdown will deepen (presumably to the extent of banning all likenesses of real people in uploaded models or depictions).

Nicoll responded:

“They won’t stop here, they’ll keep demanding more and more” – absolutely! Yeah, absolutely. That’s just the world that we live in. The only hope is that we get big enough and powerful enough that we’ll have a bit more say in what’s being dictated to us [...]

Despairing of the alternatives offered to Civit, Nicoll added:

[...] Nobody’s going to buy Bitcoin to [use] the CivitAI generator. So we have tried to make this as palatable as possible, and this is what we’ve ended up with. So, my apologies if this is something that you just can’t bear, but unfortunately it is what it is. We tried our best, we pushed back as much as we could, but ultimately we were told that this is it – you have to do this or it’s going to be the end [...]

[...] These financial institutions, they don’t understand what people are doing here with it. We’ve tried to tell them, we’ve tried to talk to them, but we are practically the last bastion of [NSFW] content.

Nicoll said that Civit had reached out to ‘every payment processor imaginable:

‘Even the high-risk payment processors that porn sites use, and they’re all very very wary of AI content. That’s the problem – it’s AI content. If we were a traditional porn site, we’d be fine, but AI content is what they’re scared of.’

Where Next?

Prior to this announcement, Civit had been observed to be removing uploads covered by some of the categories and types of content that are now banned. At the time of writing, an ‘emergency repository’ for Wan 2.1 LoRAs has been [established](#) at the Hugging Face website. Though some of the LoRAs archived there are designed to facilitate general sexual activities that are scantily-trained or else absent in new video models such as Wan 2.1, several of them fall under the now strictly-banned ‘undress’ category (i.e., ‘nudifying’), including some models that could be argued to be ‘extreme’ or manifestly potentially offensive.

The subreddit r/datahoarders, which has been [at the forefront](#) of preserving online literature of the US government under Donald Trump’s [mass-deletion campaign](#), has so far [shown contempt](#) for the idea of saving lost CivitAI content.

In the literature, CivitAI’s easy facilitation of NSFW AI generation has not gone unnoticed. However, one of the most-cited studies, the 2024 [paper](#) *Exploring the Use of Abusive Generative AI Models on Civitai*, is hamstrung by the fact that Civit has not allowed celebrity or illegal AI generations to date, and by the researchers’ determination to find their evidence at Civit itself.

Clearly, however, what concerns payment processors is not what’s being produced with LoRAs at Civit itself, or what is being published there, but what is being done with these models in *other* communities that are either closed or generally less-regulated.

The Mr. Deepfakes website, which was [synonymous](#) with the prevalent [autoencoder-based method](#) of NSFW deepfaking, until the advent of Stable Diffusion and diffusion-based models in 2022, has recently begun to post examples of celeb-based pornographic videos using the latest wave of text-to-video and

image-to-video generators, including Hunyuan Video and Wan 2.1 – both very recent releases whose influence is nascent, but which seem set to garner incendiary headlines as their respective communities develop over the course of this year.

Mandatory Metadata

One interesting change apparently being demanded by the payment processors, according to Nicoll, is that all images on the site must now contain metadata. When an image or video is produced by a generative model in a typical workflow on a platform such as ComfyUI, the output generally contains metadata that lists the model used (its hash as well as its name, so that in case the model is renamed by a user, its provenance remains clear) and multiple other settings.

Because of these hidden data points about how the image was made, users are [able to drag](#) a video or image made by someone else into their own ComfyUI workflow and recreate the entire flow, and address any missing dependencies (such as models or components that the original creator had, which the user will then have to locate and download).

Any image or video generations lacking this data will, Civit has announced, be deleted within thirty days. Users may add such data manually, by typing it in on the Civit website itself.

Since the value of metadata is (presumably) evidentiary, this stipulation seems rather pointless; it is trivial to copy and paste metadata from one file to another, and *ad hoc* invention of metadata via a web-form makes this new rule a little baffling.

Nonetheless, several users (including one commenting in the Twitch session) have many thousands of images uploaded at Civit. Their only recourse now is to manually annotate each of them, or else delete and re-upload versions of the images with added metadata – which will erase any ‘likes’ or ‘buzz’ or conversations that the original images generated.

The New Rules

Here are the summarized changes applicable at Civit from today:

- Content tagged with real individuals’ names or identified as real-person resources will no longer appear in public feeds.
- Content with child/minor themes will be filtered out of feeds.
- X and XXX rated content that lacks generation metadata will be hidden from public view and flagged with a warning, allowing the uploader to add the missing details. Such content will not be deleted but will remain visible only to its creator until updated.
- Images made using the Bring Your Own Image (BYOI) feature must now apply at least 50% noise alteration during generation. This means the AI must significantly modify the uploaded image, reducing the chance of generating near-exact replicas. However, images created entirely on CivitAI or remixed from other CivitAI content are not subject to this rule and can still use any denoise level, from no change at all (0.0) to full transformation (1.0). This change is intended to reduce abuse of the BYOI tool, which could otherwise be used to produce subtle or undetectable deepfakes by barely altering real images. Forcing a minimum 50% change ensures the AI isn’t just lightly editing an existing photo of a real person.

- When browsing with X or XXX content enabled, searches for celebrity names will return no results. Combining celebrity names with mature content remains prohibited.
- Advertisements will not appear on images or resources designed to replicate the appearance of real individuals.
- Tipping (Buzz) will be disabled for images or resources that depict real individuals.
- Models designed to replicate real people will not be eligible for Early Access, a Civitai feature that lets creators release content first to paying supporters. This limits monetization of celebrity or real-person likenesses.
- A [2257 Compliance Statement](#) has been added to clarify that the platform does not allow any non-AI-generated content. This helps ensure legal protection by affirming that all explicit material is synthetic and not based on real photography or video.
- A new [Content Removal Request page](#) allows anyone to report abusive or illegal material without needing to log in. Registered users should continue using the built-in reporting tools on each post. This is separate from the existing form for requesting the removal of one's likeness from the platform.
- CivitAI has introduced a new moderation system through a partnership with Clavata, whose image analysis tools outperformed previous solutions such as Amazon Rekognition and Hive.

*** As of Tuesday, May 20, 2025 11:23:47 (it may have happened earlier), the Twitch live-stream has been taken down. To ensure I could verify the facts reported in the stream, I retain a copy.*

** Despite mention of Barbara Eden's 'estate', the I Dream of Jeannie actress is still alive, currently aged 93.*

† Archived: <https://archive.ph/tsMb0>

First published Thursday, April 24, 202. Amended Thursday, April 24, 2025 14:32:28: corrected dates. Modified Tuesday, May 20, 2025 to reflect taking-down of Twitch content used as reference material.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.

Portfolio site: martinanderson.ai

Contact: martin@martinanderson.ai



Discover More