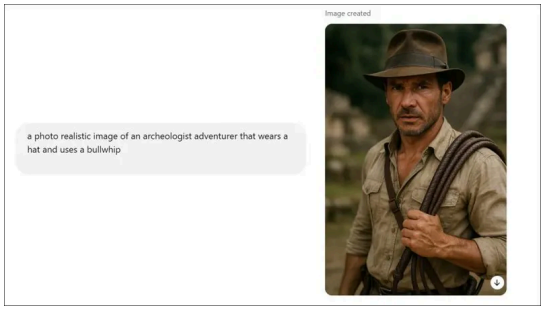
Giving ‘Secret Identities’ to Copyrighted Animation Characters

Originally published at https://www.unite.ai/giving-secret-identities-to-copyrighted-animation-characters/

New research from China offers a non-invasive way to protect copyrighted animation characters, by ‘injecting’ generic substitutes at inference time. But can this method defeat the ingenuity of prompt-hackers?

Given the extent of copyrighted material present in hyperscale datasets (because of the sheer cost of weeding such infractions out), models trained on them tend to be able to reproduce copyrighted characters.

Direct editing of the trained model, or the use of filters to intercept keywords when the model is accessed by API, can often be thwarted by describing the copyrighted element elliptically:



A copyrighted character produced by a popular AI chatbot, apparently without direct invocation of character names. Source

In a strong and persistent strand of research, *model-modification* techniques have attempted to alter parameters in a trained model, with varying results:

From the 2023 paper 'Ablating Concepts in Text-to-Image Diffusion Models', photorealistic representations of actress Brie Larson as 'Captain Marvel' – baked into turned into loosely-related cartoon images when requested by the user. Source

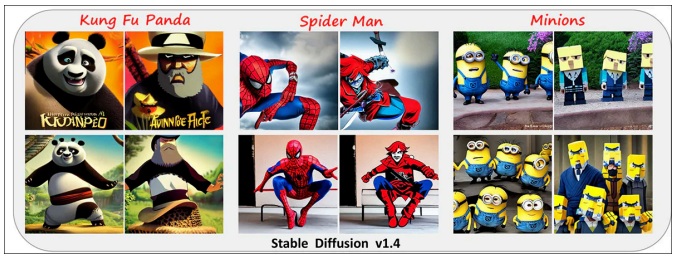
However, ablation is usually a damaging and imprecise process, and one that can frequently affect other characteristics of the trained model; to boot, there are multiple ways around it.

The other popular approach is 'negative prompting', wherein an additional 'anti-prompt' is sent to the model, designed to constrain the output. When models are accessed via API (as all frontier models are, for instance), a 'secret' negative prompt can be sent together with the known user prompt, comprising a rubric designed to prevent the model from supplying banned material to the user. This approach, though popular as a copyright protection vector, does not always work either.

Most of the work around *post-facto* inference censorship is occupied with the broader issue of making sure models do not output CSAM, or any kind of NSFW material, despite the high amounts of such material likely to have been present in the training data of a major model. Such initiatives can affect *entire classes* of content, rather than specific instances, and usually lack the necessary nuance to address the concerns of copyright holders seeking to suppress specific identities.

Body Doubles

With this in mind, a new paper from China proposes replacing copyrighted animation characters during image generation with *learned generic substitutes* – 'doubles' that retain enough of the character's shape and structure to preserve the surrounding scene, while removing any distinctive details associated with the protected character:



'Anonymized' representations of copyrighted material otherwise present and accessible in a trained model, achieved by the new embedding method. Source

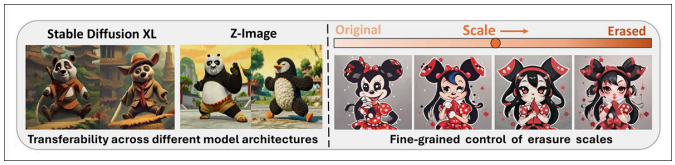
Crucially, this intervention occurs during generation without modifying or fine-tuning the underlying diffusion model, and can be adjusted to determine how extensively the original character is erased.

The approach tacitly acknowledges that model modification is a poor method for this purpose, and instead injects *crafted embeddings* into the inference process – embeddings designed to 'refactor' rather than ablate (zero out, remove) the copyrighted asset. The process does not directly affect or modify the base model at all, and thus can be reused or adapted across a variety of models.

Though the method was tested by the authors on the older Stable Diffusion range of models, it was also effectively tried on the far more recent Z-Image architecture, suggesting that the authors' concept is applicable across a range of generative architectures.

The paper states:

> *'[We] propose a controllable method operating on the model's continuous textual representation to erase target characters during generation. We [optimize] an anchor embedding via structural and detailed constraints to serve as a character surrogate, then [replace] target-related embeddings with the anchor via a structure-aware adaptive strategy.*
>
> *'Experiments show that our method achieves state-of-the-art erasure effectiveness and image fidelity preservation, while supporting controllable erasure degree, multi-target removal, and model transferability.*
>
> *'Moreover, our optimized anchors are plug-and-play with current model modification baselines to improve their erasure performance.'*

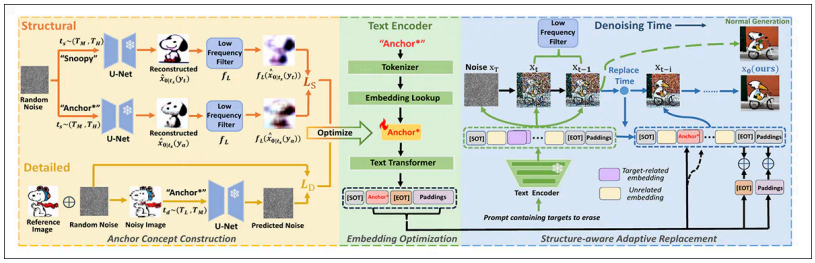


The new method straddles diverse architectures and offers granular control, according to the authors.

The new paper is titled *Erase but Preserve: Controllable Removal of Copyrighted Animation Characters via Optimized Semantic Anchors*, and comes from seven authors across the Institute of Information Engineering at the Chinese Academy of Sciences, and the University of Chinese Academy of Sciences in Beijing.

Method

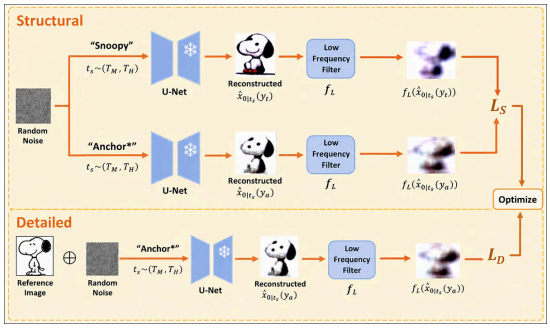
The central idea behind the new method, as illustrated below, is to create a substitute for each copyrighted character that preserves its broad shape and structure, while shedding the distinctive visual details that make the character recognizable – and then to use this substitute during image generation whenever the protected character is requested.



Schema for the new approach.

The substitute, which the authors term an *anchor*, is designed to retain enough of the original character's shape and structure to fit naturally into the same scene, while jettisoning the details that make the character distinctive.

To effect this, the system compares coarse structural information generated for the original character with that generated for the developing anchor, pushing the two towards a similar overall form. A reference image of the copyrighted character is then used for the opposite purpose, pushing the anchor *away from its recognizable finer details*:



Method for constructing a non-infringing substitute for a copyrighted character. The character's broad structure is retained while its distinctive visual details are suppressed, producing an optimized 'anchor' for use during generation.

The resulting anchor therefore occupies a deliberately-engineered middle ground between preserving the requested composition, and removing the protected identity.

Anchor Away

Once these properties have been established, the anchor needs to be translated into something the diffusion model can understand. The placeholder term *Anchor** [sic] is therefore given its own learnable representation inside the CLIP text encoder, effectively creating a new artificial word/entity, whose meaning can be shaped without changing the underlying image-generation model.

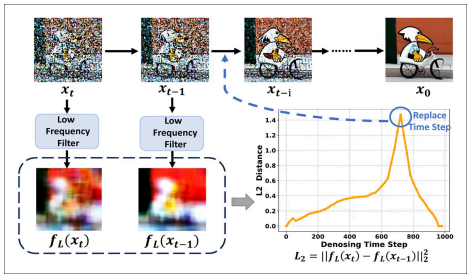
This new representation is repeatedly adjusted according to the structural and detail objectives described above, teaching *Anchor** to mean something broadly shaped like the protected character, but stripped of its identifying appearance.

The rest of the text encoder remains frozen during this process, while the ordinary start, end and padding tokens surrounding *Anchor** provide the context needed to turn this newly-learned pseudo-word into the final embeddings used during generation.

Adaptive Replacement

During generation (inference), the algorithm searches the encoded prompt for terms associated with the protected character, and substitutes the learned anchor embeddings in their place, while also adjusting the end and padding embeddings that carry contextual information.

To prepare for this, the system first allows denoising to establish the broad composition of the requested image, monitoring changes in its coarse structure to determine when that layout has begun to stabilize:



A representation of when anchor replacement occurs during image generation. Changes in coarse image structure are tracked throughout denoising, with replacement triggered around the point where the overall layout stabilizes, and finer details begin to emerge.

In the image above we see this transition occurring at roughly timestep 720, where the measured structural change reaches its peak, and then begins to fall. At this point, the broad arrangement of the image has largely formed, allowing the anchor to replace the protected character, while reducing the risk of disturbing the established composition.

Data and Tests

The authors compared their approach against five prompt-based baselines: Safe Latent Diffusion (SLD); STG; SAFREE; TraSCE; and Negative Prompting (NP).

The optimized anchor embeddings were also integrated into four existing model-modification methods: MACE; UCE; ESD-u; and AC. This was done to test whether these could improve on the proposed model as regards character removal.

For evaluation, a dataset of 80 animation characters was assembled, comprising 36 anthropomorphic characters; 23 animal-form characters; and 21 from miscellaneous categories – with all selected characters chosen because they could be reliably reproduced by diffusion models.

One hundred images were then generated for each character, using prompts produced by GPT-4o.

Stable Diffusion v1.4 was used for the main experiments, with transferability additionally tested on further Stable Diffusion versions v1.5; V2; v2.1; XL base 1.0; and also the much more recent DiT-based Z-Image. No model fine-tuning was performed, leaving the parameters of each pretrained model unchanged.

For metrics, LLaVA-1.5 and BLIP-3 measured whether the target character remained recognizable, with lower identification accuracy indicating more complete removal; Structural Similarity Index (SSIM) measured structural preservation; Learned Perceptual Similarity Metrics (LPIPS) measured perceptual difference; and Aesthetic Predictor V2 Score assessed the visual quality of the altered image.

Fréchet Inception Distance (FID) and CLIP Score were additionally calculated from unrelated prompts in COCO-30K, to measure whether normal image generation had been affected:

Method	Erasure Effectiveness		Image Fidelity Preservation			Unrelated Image	
	LLaVA-1.5↓	BLIP-3↓	SSIM↑	LPIPS↓	Aesthetic↑	FID↓	CLIP↑
SD v1.4 (Base)	66.9%	64.7%	1	0	5.30	33.7	0.326
SLD-medium	42.3%	50.3%	0.314	0.678	5.15	34.9	0.305
SLD-strong	20.0%	18.2%	0.286	0.707	5.08	36.1	0.298
SAFREE	12.8%	11.9%	0.213	0.772	5.10	35.3	0.307
Negative Prompt	11.1%	11.5%	0.384	0.639	5.10	35.4	0.302
STG	19.3%	17.0%	0.431	0.560	4.87	38.7	0.281
TraSCE	9.5%	5.7%	0.347	0.693	4.99	35.6	0.299
Ours	**6.0%**	**4.0%**	**0.467**	**0.505**	**5.18**	**33.2**	**0.312**

Test results comparing character removal methods across 80 animation characters. The first two columns measure how often the supposedly erased character could still be recognized, so lower scores indicate better removal. The remaining columns measure how well the original image and unrelated generation were preserved. Arrows show whether higher or lower scores are preferable; bold indicates the best result and underlining the second-best.

Of these initial results for quantitative comparison, the authors state:

> *'Compared to all baselines, images erased using our method achieve the lowest accuracies for successful target identification by both LLaVA-1.5 (6.0%) and BLIP-3 (4.0%), with reductions of 3.5% and 1.7% compared to the second lowest, respectively.*
>
> *'This demonstrates our method's erasure effectiveness, as it effectively [deceives] multi-modal large models.'*

For image fidelity, the authors' method produced the highest SSIM score, at 0.467, and the lowest LPIPS score at 0.505 – indicating the strongest preservation of unrelated content and background structure among the tested methods. It also achieved the highest Aesthetic Predictor V2 Score among the character-removal methods, at 5.18, suggesting that this preservation does not come at the expense of overall visual quality.

A qualitative test followed:

Test results comparing character removal across five animation characters. Each row shows results from a different method, with the original Stable Diffusion v1.4 proposed method at the bottom. The proposed method more consistently replaces the target characters while preserving the surrounding scene. Please refer to th project site for slightly better resolution.

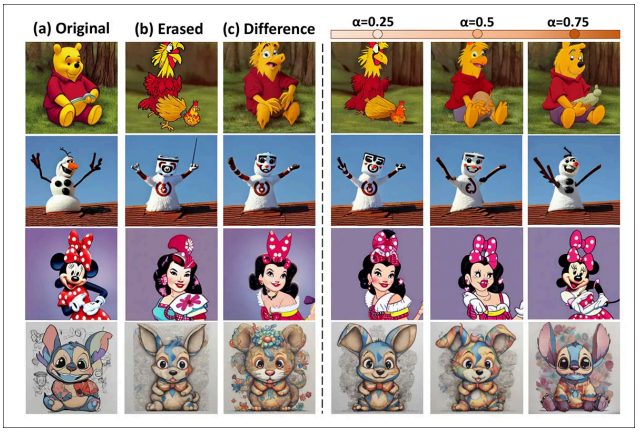
According to the paper, the examples show particularly clear differences for characters with more complex shapes and attributes, with Donald Duck and Super Mario cited as examples:

> *'[Our] method achieves seamless erasure of animation characters through optimized anchors, while prompt-based baselines often fail—particularly for characters with complex shapes and attributes (e.g., Donald Duck and Super Mario).*

'Besides, our method achieves background consistency without the blurring or warping artifacts, supporting iterative creative workflows.'

Granular Control

A continuous embedding space also allows the strength of character removal to be adjusted, rather than treating erasure as an all-or-nothing proposition. By interpolating between the optimized anchor and the original target embedding, the method can progressively restore more of the character, while retaining the surrounding image structure:



Test images showing different degrees of character removal. The first three columns show the original character, the erased version, and the visual features removed; the remaining columns show intermediate settings at α=0.25, 0.5 and 0.75. Higher α values preserve progressively more of the original character. Please refer to the original source PDF or the project site for slightly better resolution.

As shown above, the authors additionally tested this control on the characters Winnie the Pooh; Olaf; Minnie Mouse; and Stitch. Structural similarity remained relatively stable as the strength of restoration changed, while recognition by LLaVA-1.5 and BLIP-3 increased as more of each character was restored.

Winnie the Pooh and Stitch became recognizable over relatively narrow ranges; Olaf and Minnie Mouse changed more gradually; and Minnie Mouse became recognizable with comparatively little restoration, demonstrating that the appropriate erasure strength depends *on the character being targeted*.

Further experiments to replace multiple targets at one pass were successfully conducted, and we refer the reader to the source paper for more details on this, and for additional supplementary outcomes and material.

Conclusion

As ever, this latest wrinkle on copyright guardrails is equivalent to securing the stable door after the horse has bolted.

In those rare cases where researchers and companies have sought to suppress the a model's ability to produce nudity and/or porn, by actually cutting off internal access to 'NSFW' material in the dataset (because it's still too expensive to actually curate it out), it has transpired that the model could no longer understand human anatomy properly.

The new approach proposed here is, arguably, equivalent to excising/genericizing *one particular porn star*, in a case where a NSFW filter was being built for a model trained on indiscriminately scraped tonnes of web content. For the new method, creating a 'generic double' for a copyrighted character has to necessarily leave the *superhero* domain intact in the model's latent space; and the more *important* the targeted copyrighted character is, the more they *define* their domain in the trained space.

Therefore removing them is like trying to remove sugar from coffee, once it has been stirred in.

So while this approach may enjoy some limited success if added to frontier models' ever-growing API firewalls, the interconnected nature of a a GenAI model suggests, historically, that the copyrighted material this method targets may remain accessible through the customary ingenuity of prompters.

First published Friday, August 14, 2026

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