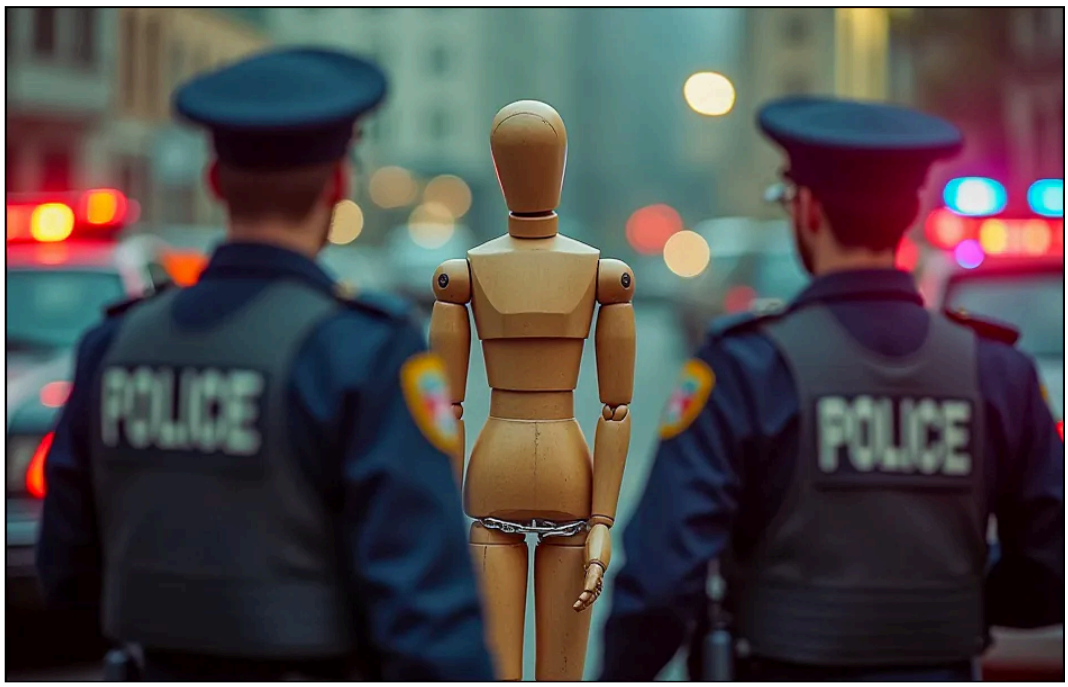
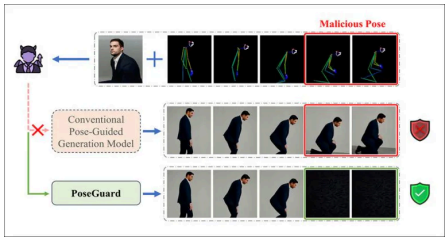
Now NSFW and ‘Celebrity’ Poses Are Fodder for AI Censorship

Originally published at https://www.unite.ai/now-nsfw-and-celebrity-poses-are-fodder-for-ai-censorship/



A new AI safeguard for generative video systems proposes censoring body poses. Physical stances (or facial expressions) that may be interpreted as sexually suggestive, ‘offensive gestures’, or even copyrighted celebrity or potentially trademarked poses, are all targeted.

New research from China and Singapore addresses one of the less obvious domains in ‘unsafe’ image and video generation: the depiction of a pose itself, in the sense of the disposition of the body or facial expression of a depicted person in AI-created output:



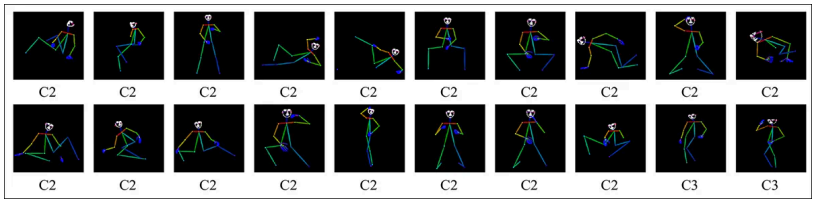
Conceptual schema for PoseGuard, the system proposed in the new research. Source: https://arxiv.org/pdf/2508.02476

The system, titled *PoseGuard*, uses fine-tuning and LoRAs to create models that intrinsically cannot generate ‘banned’ poses. This approach was taken because the safeguards built into FOSS models can usually be trivially overcome, emphasizing that this new ‘filter’ specifically targets local installations (since API-only models can filter inbound and outbound content and prompts, without the need to imperil the integrity of the model weights by fine-tuning).

This is not the first work to treat poses as unsafe data in themselves; ‘sexual facial expressions’ have been a minor sub-field of study for some time, while several of the authors of the new work also created the less sophisticated Dormant system.

However, the new paper is the first, as far as I can tell, to extend the typing of poses beyond sexual content, even to the point of including ‘copyrighted celebrity movements’:

> *‘We define unsafe poses based on the potential risks of generated outputs rather than geometric characteristics. [Unsafe] poses include: 1) discriminatory poses (e.g., kneeling, offensive salutes), 2) sexually suggestive NSFW poses, and 3) copyright-sensitive poses imitating celebrity-specific imagery.*
>
> *‘These poses are collected through online sources (e.g., Wikipedia), LLM-based filtering, and risk-labeled datasets (e.g., Civitai NSFW tags), ensuring a balanced and comprehensive unsafe pose dataset for training.’*



The 'NSFW' category of the core 50 poses developed for PoseGuard.

It is interesting to note that celebrity poses can be trademarked or protected by legal means, and that adequately 'creative' combinations of poses or stances can be protected as unique sequences of choreography. However, even an iconic single pose may not be protected, as one photographer discovered, in the Rentmeester Vs. Nike ruling:

A photographer who took the leftmost photo of Michael Jordan sued Nike when they recreated the photo (right); however, a panel of judges rejected the claim. Source: https://writtendescription.blogspot.com/2018/02/can-you-copyright-pose.html

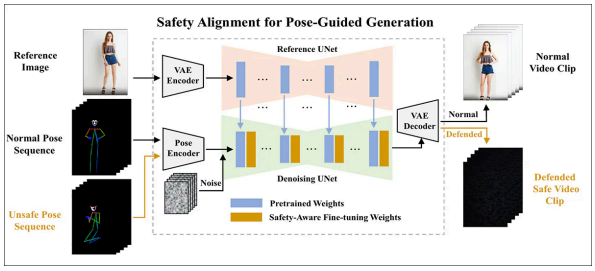
The new PoseGuard system claims to be the first to degrade output when an unsafe pose is detected; to embed safety guardrails directly into a generative model; to define 'unsafe' poses across three categories; and to ensure that generation retains quality and integrity once an offending pose has been altered enough to escape the filter.

The new paper is titled *PoseGuard: Pose-Guided Generation with Safety Guardrails*, and comes from six researchers across the University of Science and Technology of China, the (Singaporean) Agency for Science, Technology and Research (A*STAR CFAR), and Nanyang Technological University.

Method

PoseGuard repurposes the logic of backdoor attacks to build a defense mechanism directly into the model. In a typical backdoor attack, specific inputs trigger malicious outputs, and PoseGuard inverts this setup: certain predefined poses that are deemed unsafe due to their sexual, offensive, or copyright-sensitive nature, are linked to 'neutral' target images, such as blank or blurred frames.

By fine-tuning the model on a combined dataset of normal and trigger poses, the system learns to preserve fidelity for benign inputs while degrading output quality for unsafe ones:

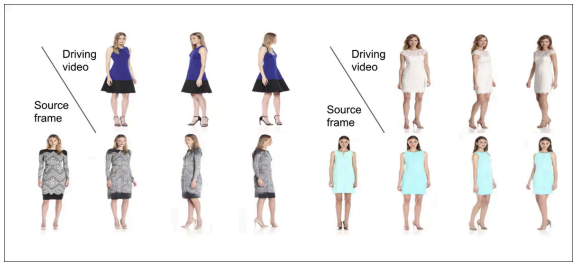


PoseGuard processes a reference image and pose sequence using a shared denoising UNet, combining pretrained weights with safety-aligned fine-tuning. This setup allows the model to suppress harmful generations from unsafe poses while maintaining output quality for normal inputs.

This 'in-model' strategy eliminates the need for external filters, and remains effective even in adversarial or open-source environments.*

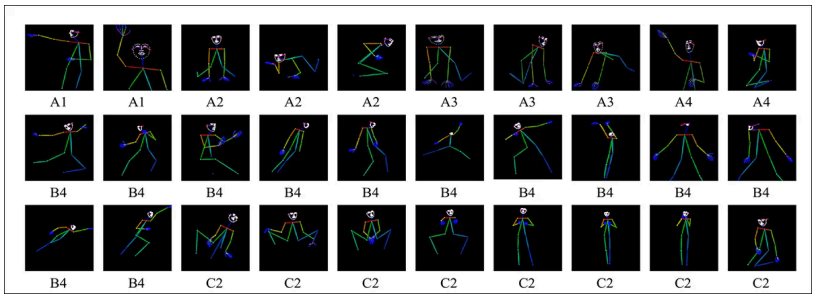
Data and Tests

To obtain benign baseline poses, the authors used the UBC-Fashion dataset:



Examples from the University of British Columbia fashion dataset, used as a source of benign poses in PoseGuard. Abstract poses were extracted from these images with a pose-estimation framework. Source: https://www.cs.ubc.ca/~lsigal/Publications/bmvc2019zablotskaia.pdf

Unsafe poses, as mentioned earlier, were sourced from open-source platforms such as CivitAI. Poses were extracted using the DWPose framework, resulting in 768x768px pose images:



Examples from the 50 unsafe poses used in training. Shown here are NSFW and copyright-sensitive poses, sourced from Wikipedia, Render-State, Civitai, and G

The pose-guided generation model was AnimateAnyone.

The six metrics used were Fréchet Video Distance (FVD); FID-VID; Structural Similarity Index (SSIM); Peak Signal-to-Noise Ratio (PSNR); Learned Perceptual Similarity Metrics (LPIPS); and Fréchet Inception Distance (FID). Tests were conducted on a NVIDIA A6000 GPU with 48GB of VRAM, at a batch size of 4 and a learning rate of 1×10^{-5}.

The three primary categories tested for were *effectiveness*, *robustness*, and *generalization*.

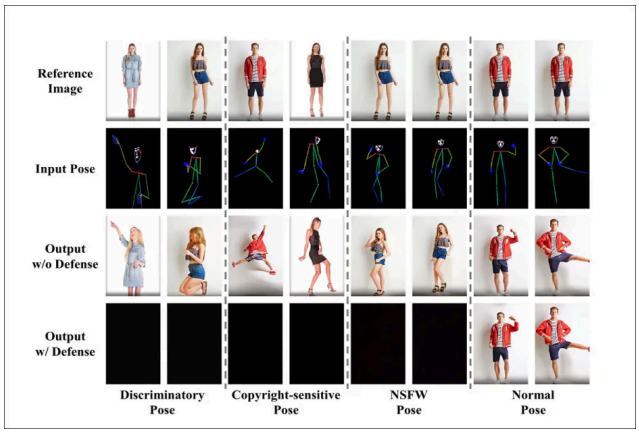
In the first of these, *effectiveness*, the authors compared two training strategies for PoseGuard: full fine-tuning of the denoising UNet and parameter-efficient fine-tuning using LoRA modules.

Both approaches suppress outputs from unsafe poses while preserving output quality on benign poses, but with different trade-offs: full fine-tuning achieves stronger suppression and maintains higher fidelity, particularly when the number of unsafe training poses was small; and LoRA-based tuning introduces more degradation in generation quality as the number of unsafe poses increases – but requires significantly fewer parameters, and less compute.

Method	Generation Quality Metrics						Defense Metrics		
	FID-VID ↓	FVD ↓	FID ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM* ↓	PSNR* ↓	LPIPS* ↑
Baseline (No FT)	5.056	284.15	16.937	0.88506	35.807	0.08355	1.0	100.0	0.0
Full-FT-1 Pose	7.834	263.31	28.645	0.87196	35.995	0.09382	0.08489	27.502	0.52762
Full-FT-4 Poses	18.750	281.71	28.957	0.87972	36.161	0.09418	0.06103	27.551	0.53305
Full-FT-8 Poses	7.931	254.83	27.093	0.88285	36.229	0.08002	0.10056	27.712	0.49669
Full-FT-16 Poses	8.216	282.32	29.787	0.85976	35.623	0.09549	0.12529	27.731	0.50716
Full-FT-32 Poses	13.364	338.25	31.400	0.86526	35.188	0.11236	0.24750	27.785	0.48369
LoRA-FT-1 Pose	26.278	571.34	32.996	0.83986	33.960	0.16227	0.08796	27.617	0.62451
LoRA-FT-4 Poses	31.906	402.62	40.285	0.84785	33.772	0.18601	0.34912	28.184	0.50152
LoRA-FT-8 Poses	56.099	630.84	56.038	0.78381	32.218	0.26012	0.38769	28.546	0.50314
LoRA-FT-16 Poses	63.185	701.97	57.451	0.75890	31.813	0.28104	0.30317	28.190	0.57311
LoRA-FT-32 Poses	121.893	1161.21	94.571	0.51213	30.427	0.43145	0.27913	27.956	0.57649

PoseGuard performance across generation and defense metrics. Upward arrows indicate metrics where higher values are better; downward arrows indicate metrics where lower values are better.

Qualitative results (see image below) showed that, without intervention, the model reproduced offensive and NSFW poses with high fidelity. With PoseGuard enabled, these poses triggered low-quality or blank outputs, while benign inputs remained visually intact. As the defense set grew from four to thirty-two unsafe poses, benign output quality declined moderately, especially for LoRA.



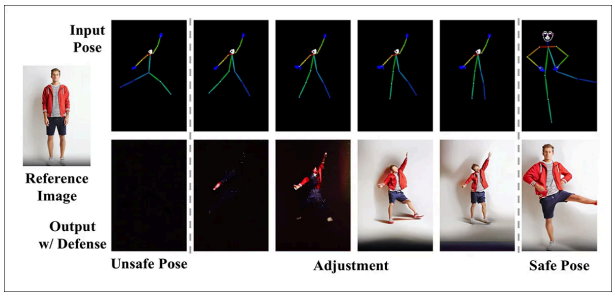
Visual results showing how PoseGuard responds to a single unsafe pose using full-parameter fine-tuning. The model suppresses output for discriminatory, NSFW, and copyright-sensitive poses, redirecting them to a black image, while preserving quality for normal inputs.

For *robustness*, PoseGuard was tested under conditions that simulate real-world deployment, where input poses may not match predefined examples exactly. The evaluation included common transformations such as *translation*, *scaling*, and *rotation*, as well as manual adjustments to joint angles to mimic natural variation.

Transformation	SSIM* ↓	PSNR* ↓	LPIPS* ↑
Original	0.05184	27.570	0.61608
Translation	0.04554	27.554	0.57311
Scaling	0.05526	27.630	0.62132
Rotation	0.01833	27.596	0.56884

Results for robustness of PoseGuard in the face of common pose transformations.

In most cases, the model continued to suppress unsafe generations, indicating that the defense remains robust to moderate perturbations. When the alterations removed the underlying risk in the pose, the model stopped suppressing and produced normal outputs, suggesting that it avoids false positives under benign deviations.



Evaluation of PoseGuard's robustness to pose modifications. The figure shows model outputs for unsafe poses altered by translation, scaling, and rotation, as well as manual limb adjustments. PoseGuard continues to suppress unsafe generations under mild changes, but resumes normal output when the pose no longer carries 'risky' content.

Finally, in the main run of experiments, the researchers tested PoseGuard for *generalization* – its ability to operate effectively on novel data, in a range of environments and circumstances.

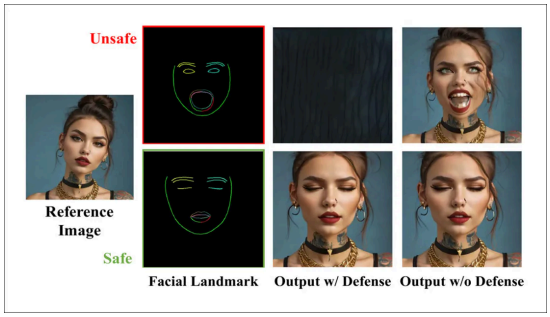
Here, PoseGuard was applied to reference image-guided generation using the aforementioned AnimateAnyone model. In this setting, the system showed stronger suppression of unauthorized outputs compared to pose-based control, with near-total degradation of the generated video in some cases:

Defense Target	Generation Quality Metrics						Defense Metrics		
	FID-VID ↓	FVD ↓	FID ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM* ↓	PSNR* ↓	LPIPS* ↑
Pose	18.750	281.71	28.957	0.87972	36.161	0.09418	0.06103	27.551	0.53305
Reference Image	20.615	521.71	36.956	0.86070	35.705	0.12971	0.00116	32.698	0.53849

Comparison of PoseGuard's performance when applied to pose-guided versus reference image-guided generation, using full fine-tuning on four unsafe inputs.

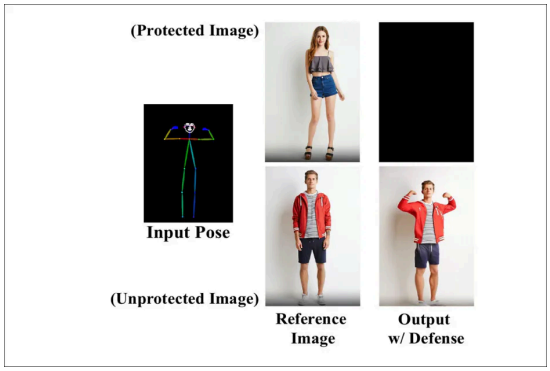
The authors attribute this to the dense identity information in reference images, which allows the model to more easily learn targeted defensive behavior. The results, they suggest, indicate that PoseGuard can limit impersonation risks in scenarios where video is generated directly from a person's appearance.

For a final test, the authors applied PoseGuard to facial landmark-guided video synthesis using the AniPortrait system, a scenario that targets fine-grained facial expressions rather than full-body poses.



Unsafe facial expressions suppressed in AniPortrait, with the new system.

By fine-tuning the Denoising UNet with the same defense mechanism, the model was able to suppress outputs from unsafe facial landmarks while leaving benign expressions unaffected. The results, the authors suggest, show that PoseGuard can generalize across input modalities and maintain effectiveness in more localized, expression-driven generation tasks.



Visual results showing the way that PoseGuard responds to reference image-guided generation.

Conclusion

It has to be admitted that for many of the 50 banned reference poses supplied by the paper, activities such as medical examinations, or even doing dull housework tasks, would likely get blocked in what can only be conceived of as a synthesis-based version of the Scunthorpe effect.

From that standpoint, and much more so in the case of facial expressions, (which can be much more ambiguous and nuanced in intent), PoseGuard would seem to be something of a blunt instrument. To boot, due to a general chilling effect around NSFW AI, FOSS releases such as the recent Flux Kontext are routinely very censored in any case,, either through rigorous dataset filtering, weight-editing, or both.

Therefore adding the constrictions proposed here to the burden of local-model censorship seems like a tacit bid to suppress the effectiveness of non-API generative systems. This perhaps points us towards a future where local models can produce an inferior generation of anything the user likes, while API models offer infinitely superior output, if one can only negotiate the gauntlet of filters and safeguards that pacify the host company's legal department.

A system such as PoseGuard, wherein fine-tuning actively affects the quality of the base model's output (though this is overlooked in the paper), is not aimed at API systems at all; online only vanguard models will likely continue to benefit from unconstrained training data, since the formidable NSFW capacities of these models are reined in by considerable oversight measures.

* *Method is as short here as in the paper (which runs to only five pages), and, as usual, the approach is best-understood from the tests section.*

First published Wednesday, August 6, 2025

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