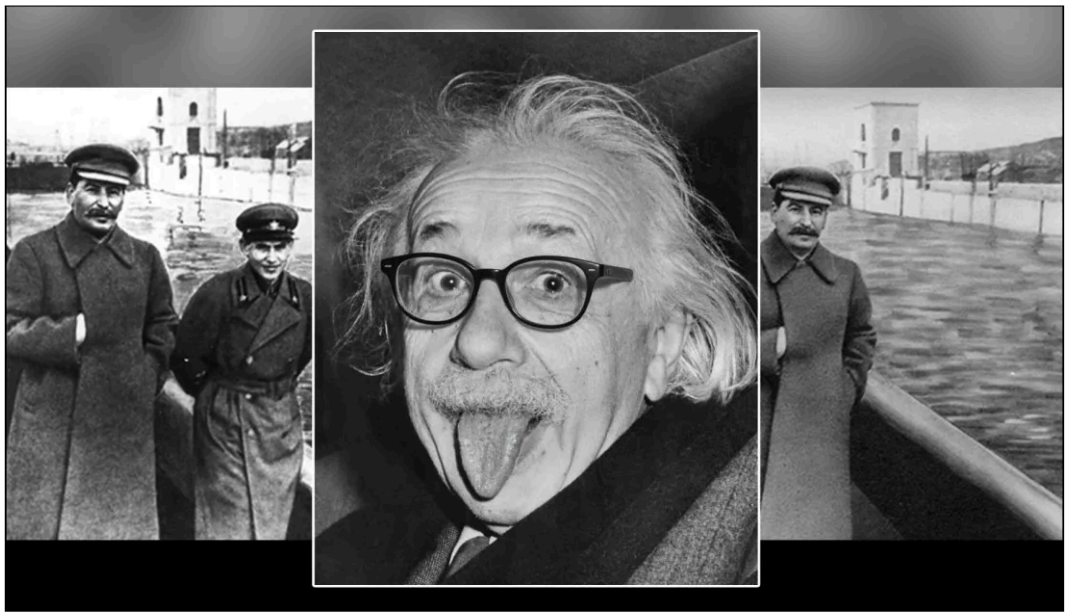
Smaller Deepfakes May Be the Bigger Threat

Originally published at https://www.unite.ai/smaller-deepfakes-may-be-the-bigger-threat/



Conversational AI tools such as ChatGPT and Google Gemini are now being used to create deepfakes that do not swap faces, but in more subtle ways can rewrite the whole story inside an image. By changing gestures, props and backgrounds, these edits fool both AI detectors and humans, raising the stakes for spotting what is real online.

In the current climate, particularly in the wake of significant legislation such as the TAKE IT DOWN act, many of us associate deepfakes and AI-driven identity synthesis with non-consensual AI porn and political manipulation – in general, *gross* distortions of the truth.

This acclimatizes us to expect AI-manipulated images to always be going for high-stakes content, where the quality of the rendering and the manipulation of context may succeed in achieving a credibility coup, at least in the short term.

Historically, however, far subtler alterations have often had a more sinister and enduring effect – such as the state-of-the-art photographic trickery that allowed Stalin to remove those who had fallen out of favor from the photographic record, as satirized in the George Orwell novel *Nineteen Eighty-Four*, where protagonist Winston Smith spends his days rewriting history and having photos created, destroyed and 'amended'.

In the following example, the problem with the *second* picture is that we 'don't know what we don't know' – that the former head of Stalin's secret police, Nikolai Yezhov, used to occupy the space where now there is only a safety barrier:

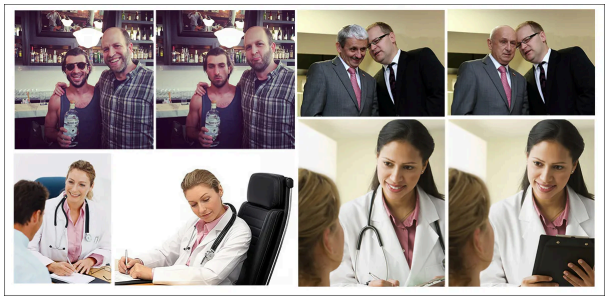
Now you see him, now he's…vapor. Stalin-era photographic manipulation removes a disgraced party member from history. Source: Public domain, via https://www.rferl.org/a/soviet-airbrushing-the-censors-who-scratched-out-history/29361426.html

Currents of this kind, oft-repeated, persist in many ways; not only culturally, but in computer vision itself, which derives trends from statistically dominant themes and motifs in training datasets. To give one example, the fact that smartphones have lowered the barrier to entry, and *massively* lowered the cost of photography, means that their iconography has become ineluctably associated with many abstract concepts, even when this is not appropriate.

If conventional deepfaking can be perceived as an act of 'assault', pernicious and persistent minor alterations in audio-visual media are more akin to 'gaslighting'. Additionally, the capacity for this kind of deepfaking to go unnoticed makes it hard to identify via state-of-the-art deepfake detections systems (which are looking for gross changes). This approach is more akin to water wearing away rock over a sustained period, than a rock aimed at a head.

MultiFakeVerse

Researchers from Australia have made a bid to address the lack of attention to 'subtle' deepfaking in the literature, by curating a substantial new dataset of person-centric image manipulations that alter context, emotion, and narrative without changing the subject's core identity:



Sampled from the new collection, real/fake pairs, with some alterations more subtle than others. Note, for instance, the loss of authority for the Asian woman, lower-right, as her doctor's stethoscope is removed by AI. At the same time, the substitution of the doctor's pad for the clipboard has no obvious semantic angle. Source: https://huggingface.co/datasets/parulgupta/MultiFakeVerse_preview

Titled *MultiFakeVerse*, the collection consists of 845,826 images generated via vision language models (VLMs), which can be accessed online and downloaded, with permission.

The authors state:

> *'This VLM-driven approach enables semantic, context-aware alterations such as modifying actions, scenes, and human-object interactions rather than synthetic or low-level identity swaps and region-specific edits that are common in existing datasets.*
>
> *'Our experiments reveal that current state-of-the-art deepfake detection models and human observers struggle to detect these subtle yet meaningful manipulations.'*

The researchers tested both humans and leading deepfake detection systems on their new dataset to see how well these subtle manipulations could be identified. Human participants struggled, correctly classifying images as real or fake only about 62% of the time, and had even greater difficulty pinpointing which parts of the image had been altered.

Existing deepfake detectors, trained mostly on more obvious face-swapping or inpainting datasets, performed poorly as well, often failing to register that any manipulation had occurred. Even after fine-tuning on MultiFakeVerse, detection rates stayed low, exposing how poorly current systems handle these subtle, narrative-driven edits.

The new paper is titled *Multiverse Through Deepfakes: The MultiFakeVerse Dataset of Person-Centric Visual and Conceptual Manipulations*, and comes from five researchers across Monash University at Melbourne, and Curtin University at Perth. Code and related data has been released at GitHub, in addition to the Hugging Face hosting mentioned earlier.

Method

The MultiFakeVerse dataset was built from four real-world image sets featuring people in diverse situations: EMOTIC; PISC, PIPA, and PIC 2.0. Starting with 86,952 original images, the researchers produced 758,041 manipulated versions.

The Gemini-2.0-Flash and ChatGPT-4o frameworks were used to propose six minimal edits for each image – edits designed to subtly alter how the most prominent person in the image would be perceived by a viewer.

The models were instructed to generate modifications that would make the subject appear *naive*, *proud*, *remorseful*, *inexperienced*, or *nonchalant*, or to adjust some factual element within the scene. Along with each edit, the models also produced a *referring expression* to clearly identify the target of the modification, ensuring the subsequent editing process could apply changes to the correct person or object within each image.

The authors clarify:

> *'Note that* referring expression *is a widely explored domain in the community, which means a phrase which can disambiguate the target in an image, e.g. for an image having two men sitting on a desk, one talking on the phone and the other looking through documents, a suitable referring expression of the later would be* the man on the left holding a piece of paper.*'*

Once the edits were defined, the actual image manipulation was carried out by prompting vision-language models to apply the specified changes while leaving the rest of the scene intact. The researchers tested three systems for this task: GPT-Image-1; Gemini-2.0-Flash-Image-Generation; and ICEdit.

After generating twenty-two thousand sample images, Gemini-2.0-Flash emerged as the most consistent method, producing edits that blended naturally into the scene without introducing visible artifacts; ICEdit often produced more obvious forgeries, with noticeable flaws in the altered regions; and GPT-Image-1 occasionally affected unintended parts of the image, partly due to its conformity to fixed output aspect ratios.

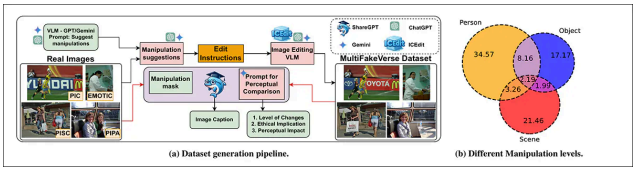
Image Analysis

Each manipulated image was compared to its original to determine how much of the image had been altered. The pixel-level differences between the two versions were calculated, with small random noise filtered out to focus on meaningful edits. In some images, only tiny areas were affected; in others, up to *eighty percent of the scene* was modified.

To evaluate how much the meaning of each image shifted in the light of these alterations, captions were generated for both the original and manipulated images using the ShareGPT-4V vision-language model.

These captions were then converted into embeddings using Long-CLIP, allowing a comparison of how far the content had diverged between versions. The strongest semantic changes were seen in cases where objects close to or directly involving the person had been altered, since these small adjustments could significantly change how the image was interpreted.

Gemini-2.0-Flash was then used to classify the *type* of manipulation applied to each image, based on where and how the edits were made. Manipulations were grouped into three categories: *person-level* edits involved changes to the subject's facial expression, pose, gaze, clothing, or other personal features; *object-level* edits affected items connected to the person, such as objects they were holding or interacting with in the foreground; and *scene-level* edits involved background elements or broader aspects of the setting that did not directly involve the person.



(a) Dataset generation pipeline. (b) Different Manipulation levels.

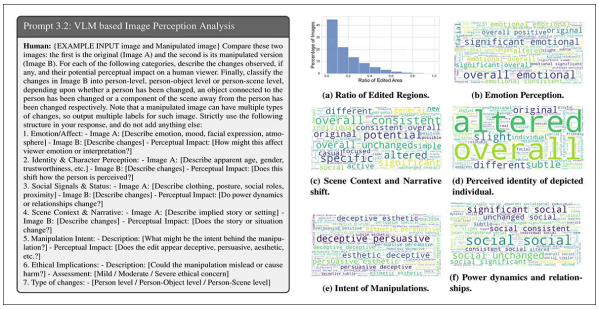
The MultiFakeVerse dataset generation pipeline begins with real images, where vision-language models propose narrative edits targeting people, objects, or scenes. These instructions are then applied by image editing models. The right panel shows the proportion of person-level, object-level, and scene-level manipulations across the dataset. Source: https://arxiv.org/pdf/2506.00868

Since individual images could contain multiple types of edits at once, the distribution of these categories was mapped across the dataset. Roughly one-third of the edits targeted only the person, about one-fifth affected only the scene, and around one-sixth were limited to objects.

Assessing Perceptual Impact

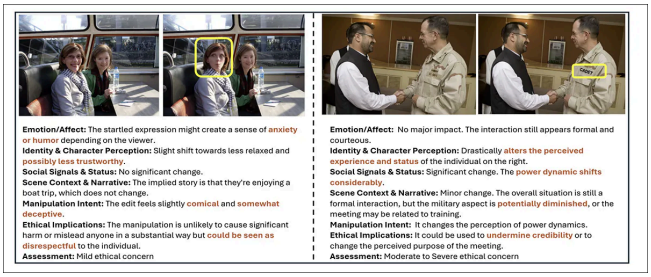
Gemini-2.0-Flash was used to assess how the manipulations might alter a viewer's perception across six areas: *emotion*, *personal identity*, *power dynamics*, *scene narrative*, *intent of manipulation*, and *ethical concerns*.

For *emotion*, the edits were often described with terms like *joyful*, *engaging*, or *approachable*, suggesting shifts in how subjects were emotionally framed. In narrative terms, words such as *professional* or *different* indicated changes to the implied story or setting:



Gemini-2.0-Flash was prompted to evaluate how each manipulation affected six aspects of viewer perception. Left: example prompt structure guiding the model's assessment. Right: word clouds summarizing shifts in emotion, identity, scene narrative, intent, power dynamics, and ethical concerns across the dataset.

Descriptions of identity shifts included terms like *younger*, *playful*, and *vulnerable*, showing how minor changes could influence how individuals were perceived. The intent behind many edits was labeled as *persuasive*, *deceptive*, or *aesthetic*. While most edits were judged to raise only mild ethical concerns, a small fraction were seen as carrying moderate or severe ethical implications.



Examples from MultiFakeVerse showing how small edits shift viewer perception. Yellow boxes highlight the altered regions, with accompanying analysis of changes in emotion, identity, narrative, and ethical concerns.

Metrics

The visual quality of the MultiFakeVerse collection was evaluated using three standard metrics: Peak Signal-to-Noise Ratio (PSNR); Structural Similarity Index (SSIM); and Fréchet Inception Distance (FID):

Dataset	PSNR(↑)	SSIM(↑)	FID(↓)
MultiFakeVerse	**66.30**	**0.5774**	**3.30**

Image quality scores for MultiFakeVerse measured by PSNR, SSIM, and FID.

The SSIM score of 0.5774 reflects a moderate degree of similarity, consistent with the goal of preserving most of the image while applying targeted edits; the FID score of 3.30 suggests that the generated images maintain high quality and diversity; and a PSNR value of 66.30 decibels indicates that the images retain good visual fidelity after manipulation.

User Study

A user study was run to see how well people could spot the subtle fakes in MultiFakeVerse. Eighteen participants were shown fifty images, evenly split between real and manipulated examples covering a range of edit types. Each person was asked to classify whether the image was real or fake, and, if fake, to identify what kind of manipulation had been applied.

The overall accuracy for deciding real versus fake was 61.67 percent, meaning participants misclassified images more than one-third of the time.

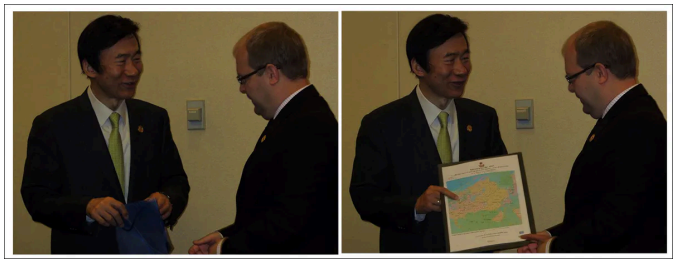
The authors state:

> *'Analyzing the human predictions of manipulation levels for the fake images, the average intersection over union between the predicted and actual manipulation levels was found to be 24.96%.*
>
> *'This shows that it is non-trivial for human observers to identify the regions of manipulations in our dataset.'*

Building the MultiFakeVerse dataset required extensive computational resources: for generating edit instructions, over 845,000 API calls were made to Gemini and GPT models, with these prompting tasks costing around $1000; producing the Gemini-based images cost approximately $2,867; and generating images using GPT-Image-1 cost roughly $200. ICEdit images were created locally on an NVIDIA A6000 GPU, completing the task in roughly twenty-four hours.

Tests

Prior to tests, the dataset was divided into training, validation, and test sets by first selecting 70% of the real images for training; 10 percent for validation; and 20 percent for testing. The manipulated images generated from each real image were assigned to the same set as their corresponding original.



Further examples of real (left) and altered (right) content from the dataset.

Performance on detecting fakes was measured using image-level accuracy (whether the system correctly classifies the entire image as real or fake) and F1 scores. For locating manipulated regions, the evaluation used Area Under the Curve (AUC), F1 scores, and intersection over union (IoU).

The MultiFakeVerse dataset was used against leading deepfake detection systems on the full test set, with the rival frameworks being CnnSpot; AntifakePrompt; TruFor; and the vision-language-based SIDA. Each model was first evaluated in zero-shot mode, using its original pretrained weights without further adjustment.

Two models, CnnSpot and SIDA, were then fine-tuned on MultiFakeVerse training data to assess whether retraining improved performance.

Method	Year	Real		Fake		Overall	
		Acc (↑)	F1 (↑)	Acc (↑)	F1 (↑)	Acc (↑)	F1 (↑)
CnnSpot [37]	2021	100.00 (6.16↓)	19.00 (69.88↑)	0.00 (97.97↑)	0.00 (98.62↑)	50.02 (45.88↑)	9.54 (84.21↑)
Trufor [13]	2023	84.00	18.52	15.27	26.07	49.64	22.30
AntiFakePrompt [9]	2024	63.30	30.47	70.45	80.63	66.87	55.55
SIDA-13B [15]	2024	67.40(23.07↑)	21.30(64.09↑)	44.55(52.94↑)	60.08(38.09↑)	55.97(38.01↑)	40.69(51.09↑)

Deepfake detection results on MultiFakeVerse under zero-shot and fine-tuned conditions. Numbers in parentheses show changes after fine-tuning.

Of these results, the authors state:

> *'[The] models trained on earlier inpainting-based fakes struggle to identify our VLM-Editing based forgeries, particularly, CNNSpot tends to classify almost all the images as real. AntifakePrompt has the best zero-shot performance with 66.87% average class-wise accuracy and 55.55% F1 score.*
>
> *'After finetuning on our train set, we observe a performance improvement in both CNNSpot and SIDA-13B, with CNNSpot surpassing SIDA-13B in terms of both average class-wise accuracy (by 1.92%) as well as F1-Score (by 1.97%).'*

SIDA-13B was evaluated on MultiFakeVerse to measure how precisely it could locate the manipulated regions within each image. The model was tested both in zero-shot mode and after fine-tuning on the dataset.

In its original state, it reached an intersection-over-union score of 13.10, an F1 score of 19.92, and an AUC of 14.06, reflecting weak localization performance.

After fine-tuning, the scores improved to 24.74 for IoU, 39.40 for F1, and 37.53 for AUC. However, even with extra training, the model still had trouble finding exactly where the edits had been made, highlighting how difficult it can be to detect these kinds of small, targeted changes.

Conclusion

The new study exposes a blind spot both in human and machine perception: while much of the public debate around deepfakes has focused on headline-grabbing identity swaps, these quieter 'narrative edits' are harder to detect and potentially more corrosive in the long-term.

As systems such as ChatGPT and Gemini take a more active role in generating this kind of content, and as we ourselves increasingly participate in altering the reality of our own photo-streams, detection models that rely on spotting crude manipulations may offer inadequate defense.

What MultiFakeVerse demonstrates is not that detection has failed, but that at least part of the problem may be shifting into a more difficult, slower-moving form: one where small visual lies accumulate unnoticed.

First published Thursday, June 5, 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