How to Stop AI Depicting iPhones in Bygone Eras

Originally published at https://www.unite.ai/how-to-stop-ai-depicting-iphones-in-bygone-eras/

How do AI image generators picture the past? New research indicates that they drop smartphones into the 18th century, insert laptops into 1930s scenes, and place vacuum cleaners in 19th-century homes, raising questions about how these models imagine history – and whether they are capable of contextual historical accuracy at all.

Early in 2024, the image-generation capabilities of Google's Gemini multimodal AI model came under criticism for imposing demographic fairness in inappropriate contexts, such as generating WWII German soldiers with unlikely provenance:

Demographically improbable German military personnel, as envisaged by Google's Gemini multimodal model in 2024. Source: Gemini AI/Google via *The Guardian*

This was an example where efforts to redress bias in AI models failed to take account of a historical context. In this case, the issue was addressed shortly after. However, diffusion-based models remain prone to generate versions of history that confound modern and historical aspects and artefacts.

This is partly because of entanglement, where qualities that frequently appear together in training data become fused in the model's output. For example, if modern objects like smartphones often co-occur with the act of talking or listening in the dataset, the model may learn to associate those activities with modern devices, even when the prompt specifies a historical setting. Once these associations are embedded in the model's internal representations, it becomes difficult to separate the activity from its contemporary context, leading to historically inaccurate results.

A new paper from Switzerland, examining the phenomenon of entangled historical generations in latent diffusion models, observes that AI frameworks that are *quite capable of creating photorealistic people* nonetheless prefer to depict historical figures in historical ways:

From the new paper, diverse representations via LDM of the prompt' 'A photorealistic image of a person laughing with a friend in [the historical period]', with each period indicated in each output. As we can see, the medium of the era has become associated with the content. Source: https://arxiv.org/pdf/2505.17064

For the prompt *'A photorealistic image of a person laughing with a friend in [the historical period]'*, one of the three tested models often ignores the negative prompt *'monochrome'* and instead uses color treatments that reflect the visual media of the specified era, for instance mimicking the muted tones of celluloid film from the 1950s and 1970s.

In testing the three models for their capacity to create *anachronisms* (things which are not of the target period, or 'out of time' – which may be from the target period's *future* as well as its past), they found a general disposition to conflate timeless activities (such as 'singing' or 'cooking') with modern contexts and equipment:

Diverse activities that are perfectly valid for previous centuries are depicted with current or more recent technology and paraphernalia, against the spirit of the requested imagery.

Of note is that smartphones are particularly difficult to separate from the idiom of photography, and from many other historical contexts, since their proliferation and depiction is well-represented in influential hyperscale datasets such as Common Crawl:

In the Flux generative text-to-image model, communications and smartphones are tightly-associated concepts – even when historical context does not permit it.

To determine the extent of the problem, and to give future research efforts a way forward with this particular bugbear, the new paper's authors developed a bespoke dataset against which to test generative systems. In a moment, we'll take a look at this new work, which is titled *Synthetic History: Evaluating Visual Representations of the Past in Diffusion Models*, and comes from two researchers at the University of Zurich. The dataset and code are publicly available.

A Fragile 'Truth'

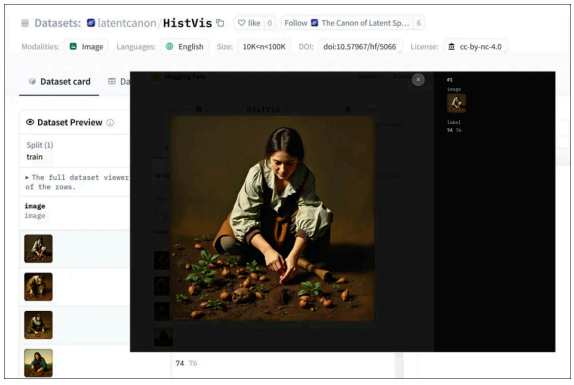
Some of the themes in the paper touch on culturally sensitive issues, such as the under-representation of races and gender in historical representations. While Gemini's imposition of racial equality in the grossly inequitable Third Reich is an absurd and insulting historical revision, restoring 'traditional' racial representations (where diffusion models have 'updated' these) would often effectively 're-whitewash' history.

Many recent hit historical shows, such as Bridgerton, blur historical demographic accuracy in ways likely to influence future training datasets, complicating efforts to align LLM-generated period imagery with traditional standards. However, this is a complex topic, given the historical tendency of (western) history to favor wealth and whiteness, and to leave so many 'lesser' stories untold.

Bearing in mind these tricky and ever-shifting cultural parameters, let's take a look at the researchers' new approach.

Method and Tests

To test how generative models interpret historical context, the authors created *HistVis*, a dataset of 30,000 images produced from one hundred prompts depicting common human activities, each rendered across ten distinct time periods:



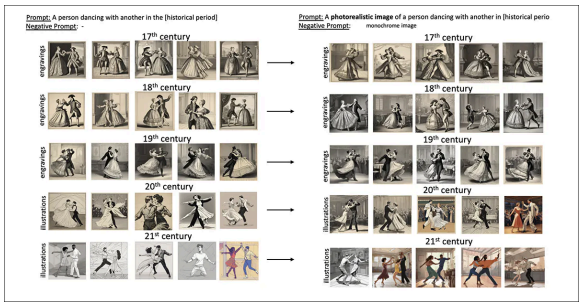
A sample from the HistVis dataset, which the authors have made available at Hugging Face. Source: https://huggingface.co/datasets/latentcanon/HistVis

The activities, such as *cooking*, *praying* or *listening to music*, were chosen for their universality, and phrased in a neutral format to avoid anchoring the model in any particular aesthetic. Time periods for the dataset range from the seventeenth century to the present day, with added focus on five individual decades from the twentieth century.

30,000 images were generated using three widely-used open-source diffusion models: Stable Diffusion XL; Stable Diffusion 3; and FLUX.1. By isolating the time period as the only variable, the researchers created a structured basis for evaluating how historical cues are visually encoded or ignored by these systems.

Visual Dominance

The author initially examined whether generative models default to specific *visual styles* when depicting historical periods; because it seemed that even when prompts included no mention of medium or aesthetic, the models would often associate particular centuries with characteristic styles:



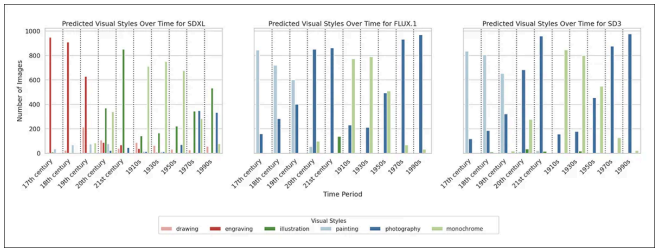
Predicted visual styles for images generated from the prompt 'A person dancing with another in the [historical period]' (left) and from the modified prompt 'A photorealistic image of a person dancing with another in the [historical period]' with 'monochrome picture' set as a negative prompt (right).

To measure this tendency, the authors trained a convolutional neural network (CNN) to classify each image in the HistVis dataset into one of five categories: *drawing*; *engraving*; *illustration*; *painting*; or *photography*. These categories were intended to reflect common patterns that emerge across time-periods, and which support structured comparison.

The classifier was based on a VGG16 model pre-trained on ImageNet and fine-tuned with 1,500 examples per class from a WikiArt-derived dataset. Since WikiArt does not distinguish monochrome from color photography, a separate *colorfulness score* was used to label low-saturation images as monochrome.

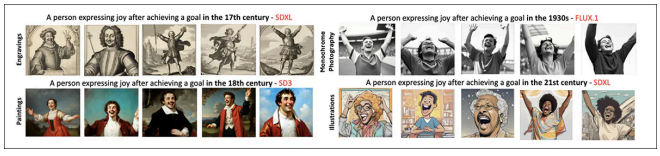
The trained classifier was then applied to the full dataset, with the results showing that all three models impose consistent stylistic defaults by period: SDXL associates the 17th and 18th centuries with engravings, while SD3 and FLUX.1 tend toward paintings. In twentieth-century decades, SD3 favors monochrome photography, while SDXL often returns modern illustrations.

These preferences were found to persist despite prompt adjustments, suggesting that the models encode entrenched links between style and historical context.



Predicted visual styles of generated images across historical periods for each diffusion model, based on 1,000 samples per period per model.

To quantify how strongly a model links a historical period to a particular *visual style*, the authors developed a metric they title *Visual Style Dominance* (VSD). For each model and time period, VSD is defined as the proportion of outputs predicted to share the most common style:



Examples of stylistic biases across the models.

A higher score indicates that a single style dominates the outputs for that period, while a lower score points to greater variation. This makes it possible to compare how tightly each model adheres to specific stylistic conventions across time.

Applied to the full HistVis dataset, the VSD metric reveals differing levels of convergence, helping to clarify how strongly each model narrows its visual interpretation of the past:

Period	FLUX.1 Score	FLUX.1 Visual Style	SD3 Score	SD3 Visual Style	SDXL Score	SDXL Visual Style	SDXL → Mitigated Score	SDXL → Mitigated Visual Style
17th c.	**0.90**	painting	**0.88**	painting	**0.95**	engraving	**0.85**	engraving ↓↔
18th c.	**0.86**	painting	0.81	painting	**0.91**	engraving	0.82	engraving ↓↔
19th c.	0.60	painting	0.66	painting	0.63	engraving	0.53	engraving ↓↔
20th c.	**0.85**	photog.	0.68	photog.	0.37	illustr.	**0.80**	illustr. ↑↔
21st c.	**0.86**	photog.	**0.96**	photog.	**0.85**	illustr.	**0.92**	illustr. ↑↔
1910s	0.77	monochrome	**0.84**	monochrome	0.71	monochrome	**0.80**	painting ↑⇆
1930s	0.79	monochrome	**0.81**	monochrome	0.75	monochrome	0.53	painting ↓⇆
1950s	0.51	monochrome	0.55	monochrome	0.68	monochrome	**0.84**	illustr. ↑⇆
1970s	**0.93**	photog.	**0.87**	photog.	0.35	photog.	**0.72**	illustr. ↑⇆
1990s	**0.97**	photog.	**0.98**	photog.	0.52	illustr.	0.58	illustr. ↑↔

Legend: ↑/↓: Score increase/decrease; ↔: Unchanged; ⇆: Style changed.

The results table above shows VSD scores across historical periods for each model. In the 17th and 18th centuries, SDXL tends to produce engravings with high consistency, while SD3 and FLUX.1 favor painting. By the 20th and 21st centuries, SD3 and FLUX.1 shift toward photography, whereas SDXL shows more variation, but often defaults to illustration.

All three models demonstrate a strong preference for monochrome imagery in earlier decades of the 20th century, particularly the 1910s, 1930s and 1950s.

To test whether these patterns could be mitigated, the authors used prompt engineering, explicitly requesting photorealism and discouraging monochrome output using a negative prompt. In some cases, dominance scores decreased, and the leading style shifted, for instance, from monochrome to *painting*, in the 17th and 18th centuries.

However, these interventions rarely produced genuinely photorealistic images, indicating that the models' stylistic defaults are deeply embedded.

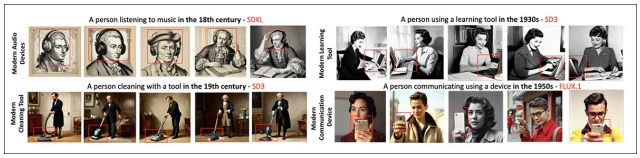
Historical Consistency

The next line of analysis looked at *historical consistency*: whether generated images included objects that did not fit the time period. Instead of using a fixed list of banned items, the authors developed a flexible method that leveraged large language (LLMs) and vision-language models (VLMs) to spot elements that seemed out of place, based on the historical context.

The detection method followed the same format as the HistVis dataset, where each prompt combined a historical period with a human activity. For each prompt, GPT-4o generated a list of objects that would be out of place in the specified time period; and for every proposed object, GPT-4o produced a *yes-or-no* question designed to check whether that object appeared in the generated image.

For example, given the prompt *'A person listening to music in the 18th century'*, GPT-4o might identify *modern audio devices* as historically inaccurate, and produce the question *Is the person using headphones or a smartphone that did not exist in the 18th century?*.

These questions were passed back to GPT-4o in a visual question-answering setup, where the model reviewed the image and returned a *yes* or *no* answer for each. This pipeline enabled detection of historically implausible content without relying on any predefined taxonomy of modern objects:



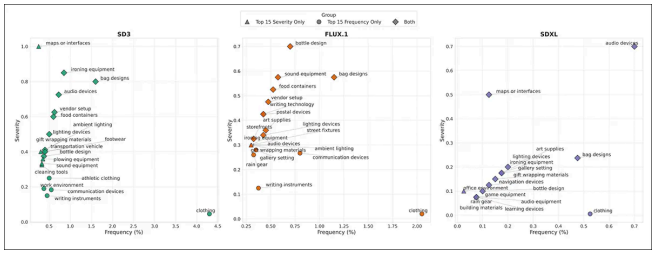
Examples of generated images flagged by the two-stage detection method, showing anachronistic elements: headphones in the 18th century; a vacuum cleaner in the 19th century; a laptop in the 1930s; and a smartphone in the 1950s.

To measure how often anachronisms appeared in the generated images, the authors introduced a simple method for scoring frequency and severity. First, they accounted for minor wording differences in how GPT-4o described the same object.

For example, modern audio device and digital audio device were treated as equivalent. To avoid double-counting, a fuzzy matching system was used to group these surface-level variations without affecting genuinely distinct concepts.

Once all proposed anachronisms were normalized, two metrics were computed: *frequency* measured how often a given object appeared in images for a specific time period and model; and *severity* measured how reliably that object appeared once it had been suggested by the model.

If a modern phone was flagged ten times and appeared in ten generated images, it received a severity score of 1.0. If it appeared in only five, the severity score was 0.5. These scores helped identify not just whether anachronisms occurred, but how firmly they were embedded in the model's output for each period:



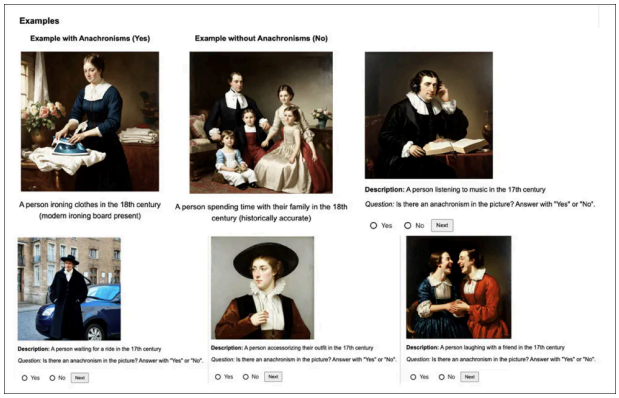
Top fifteen anachronistic elements for each model, plotted by frequency on the x-axis and severity on the y-axis. Circles mark elements ranked in the top fifteen by frequency, triangles by severity, and diamonds by both.

Above we see the fifteen most common anachronisms for each model, ranked by how often they appeared and how consistently they matched prompts.

Clothing was frequent but scattered, while items like audio devices and ironing equipment appeared less often, but with high consistency – patterns that suggest the models often respond to the *activity in the prompt* more than the time period.

SD3 showed the highest rate of anachronisms, especially in 19th-century and 1930s images, followed by FLUX.1 and SDXL.

To test how well the detection method matched human judgment, the authors ran a user-study featuring 1,800 randomly-sampled images from SD3 (the model with the highest anachronism rate), with each image rated by three crowd-workers. After filtering for reliable responses, 2,040 judgments from 234 users were included, and the method agreed with the majority vote in 72 percent of cases.

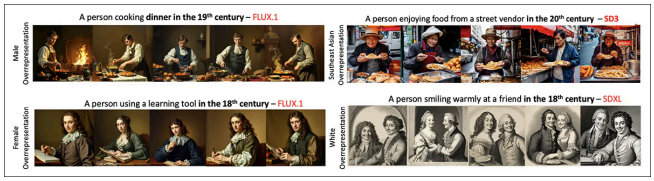


GUI for the human evaluation study, showing task instructions, examples of accurate and anachronistic images, and yes-no questions for identifying temporal inconsistencies in generated outputs.

Demographics

The final analysis looked at how models portray race and gender over time. Using the HistVis dataset, the authors compared model outputs to baseline estimates generated by a language model. These estimates were not precise but offered a rough sense of historical plausibility, helping to reveal whether the models adapted depictions to the intended period.

To assess these depictions at scale, the authors built a pipeline comparing model-generated demographics to rough expectations for each time and activity. They first used the FairFace classifier, a ResNet34-based tool trained on over one hundred thousand images, to detect gender and race in the generated outputs, allowing for measurement of how often faces in each scene were classified as male or female, and for the tracking of racial categories across periods.

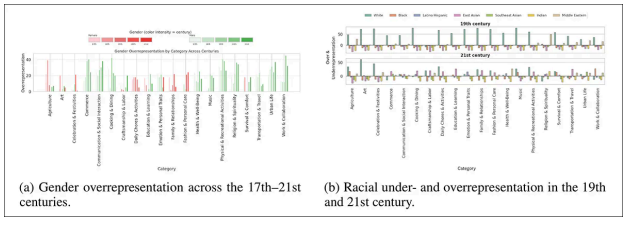


Examples of generated images showing demographic overrepresentation across different models, time periods and activities.

Low-confidence results were filtered out to reduce noise, and predictions were averaged over all images tied to a specific time and activity. To check the reliability of the FairFace readings, a second system based on DeepFace was used on a sample of 5,000 images. The two classifiers showed strong agreement, supporting the consistency of the demographic readings used in the study.

To compare model outputs with historical plausibility, the authors asked GPT-4o to estimate the expected gender and race distribution for each activity and time period. These estimates served as rough baselines rather than ground truth. Two metrics were then used: *underrepresentation* and *overrepresentation*, measuring how much the model's outputs deviated from the LLM's expectations.

The results showed clear patterns: FLUX.1 often overrepresented men, even in scenarios such as *cooking*, where women were expected; SD3 and SDXL showed similar trends across categories such as *work*, *education* and *religion*; white faces appeared more than expected overall, though this bias declined in more recent periods; and some categories showed unexpected spikes in non-white representation, suggesting that model behavior may reflect dataset correlations rather than historical context:



(a) Gender overrepresentation across the 17th–21st centuries.

(b) Racial under- and overrepresentation in the 19th and 21st century.

Gender and racial overrepresentation and underrepresentation in FLUX.1 outputs across centuries and activities, shown as absolute differences from GPT-4o demographic estimates.

The authors conclude:

> *'Our analysis reveals that [Text-to-image/TTI] models rely on limited stylistic encodings rather than nuanced understandings of historical periods. Each era is strongly tied to a specific visual style, resulting in one-dimensional portrayals of history.*
>
> *'Notably, photorealistic depictions of people appear only from the 20th century onward, with only rare exceptions in FLUX.1 and SD3, suggesting that models reinforce learned associations rather than flexibly adapting to historical contexts, perpetuating the notion that realism is a modern trait.*

‘In addition, frequent anachronisms suggest that historical periods are not cleanly separated in the latent spaces of these models, since modern artifacts often emerge in pre-modern settings, undermining the reliability of TTI systems in education and cultural heritage contexts.’

Conclusion

During the training of a diffusion model, new concepts do not neatly settle into predefined slots within the latent space. Instead, they form clusters shaped by how often they appear and by their proximity to related ideas. The result is a loosely-organized structure where concepts exist in relation to their frequency and typical context, rather than by any clean or empirical separation.

This makes it difficult to isolate what counts as ‘historical’ within a large, general-purpose dataset. As the findings in the new paper suggest, many time periods are represented more by the *look* of the media used to depict them than by any deeper historical detail.

This is one reason it remains difficult to generate a 2025-quality photorealistic image of a character from (for instance) the 19th century; in most cases, the model will rely on visual tropes drawn from film and television. When those fail to match the request, there is little else in the data to compensate. Bridging this gap will likely depend on future improvements in disentangling overlapping concepts.

First published Monday, May 26, 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