xCoAx 2024
12th Conference on Computation, Communication, Aesthetics & X
Fabrica, Treviso, Italy
2024.xCoAx.org

Language Is Leaving Me: An AI Cinematic Biometric Performance

Ellen Pearlman
ellenluminescense@gmail.com
RISEBA University, Riga, Latvia
NYU Tandon School of Engineering, New York, U.S.A.

In 2021 Open AI introduced CLIP and VQGAN, neural network architectures for still and moving text-to-image models, redefining components of contemporary visual culture. How these images are rendered and generated including their data sources, algorithmic parameters, and textual inputs are a complex and obscure process for those not well versed in computer science. This paper presents the developmental process and outcome for *Language Is Leaving Me – An AI Cinematic Opera Of The Skin*, a mixed reality performance installation focusing on AI, computer vision, sound, biometrics and epigenetic, or inherited traumatic memories of cultures of diaspora. It examines multi-lingual presentations of simple text prompts, commonly referred to as 'prompt engineering' and how, combined with VQGAN, CLIP and Stable Diffusion's image-to-image comparisons, reveal biases and flaws in their structures pointing towards difficulties for emerging visual taxonomies.

1. Introduction

Machine learning models use massive amounts of image data scrapped from contemporary visual sources through embedded tagging systems commonly referred to as Large Language Models (LLM). These systems emerged around 2017 built on transformer models (Vaswani et al. 2017). Transformer models are a type of neural network that learns context or pattern recognition to understand predictive models of sequential data, similar to how language functions, as humans predict sentence structure as they form thoughts, words, and sentences. AI models pulled back on tagging systems for predictive systems, but this is not feasible to do with AI image systems, as once an image bank is created that information is not easily deleted. If the algorithm and data are modified, that changes the training model and its sub relations. Once information links are broken, there are no guarantees the model won't make similar mistakes using billions of other connected linked information. Data that is anonymized for privacy reasons can delete taxonomy focal points or add unnecessary noise to the results and large data models can even be manipulated into reproducing parts of the training data that were previously taken out (Carlini et al. 2021, Greengard 2022). According to OpenAI, it is difficult if not impossible to prevent an opensource model from being used to cause harm, and many input classifiers can amplify bias. This problem of unsafe content output is one of the thorniest problems in visual AI. Epigenetics is understood to mean chromosome modifications that are not part of DNA structures, and more specifically its rDNA (Felsenfeld 2014).

Keywords AI, Stable Diffusion CLIP, Visual Culture, Linguistics, Algorithmic Bias, Prompt Engineering, Algorithmic Cinema, Taxonomy.
DOI 10.34626/2024_xcoax_001

It is controversial to state that human psychological trauma can definitively change chromosomal rDNA, much less human behavior. Yet, psychologists are beginning to understand that trauma can be passed through generations and manifest in various ways (Hübl, Jordan, and Vii 2020). As an artist and researcher my practice focuses on investigating newly emerging artificial intelligence cinema and synthetic media driven by Large Language Models (LLMs). Over three years I created, *Language Is Leaving Me – An AI Cinematic Opera Of the Skin*. (LILM), a biometric performed AI installation partially powered by EMG human biometric measurements, meaning the facial muscles of a volunteer audience member. EMG patches are placed on the volunteer's face to generate sections of the live time sonic environment. The volunteer's smile (positive) or frown (negative) while watching the AI cinema, altered the sounds. LILM used different linguistic prompts or texts translated from English into Yiddish, Chinese, Tamil, and Xhosa with Google Translate. It reveals hidden and devastating aspects of the multilingual algorithmic processes, and how LLMs alter and erase memory and identity of cultures of diaspora, distorting cultural and personal signified visual and verbal memories into an AI induced cognitive aphasia.

1.1. Background

On July 8, 2021, Katherine Crowson, using the Python software programming language published experiments between VQGAN (Vector Quantized Generative Adversarial Network) and CLIP (Contrastive Language-Image Pre-training) transformer and generative technologies to produce text-to-image conversions (Crowson 2021a). Her work was based on Open AI's blog dated January 5, 2021, with the subsequent paper published on February 26, 2021 demonstrating the accelerated speed of developments over 41 days in the field of transformers and visual technologies. CLIP arose from investigations into zero-shot transfer, natural language supervision, and multimodal learning. VQGAN arose from previous text-to-image transfer models like AttnGAN (Attention GAN), (Radford et al. 2021, Open AI 2021a, 2021b, NerdyRodent 2021, Xu et al. 2018). CLIP decides which tagged captions VQGAN, aligns best with so it can be assigned to a specific image generation. The CLIP model includes its own separate transformer architecture as smart indexing but does not generate anything itself and can be used with other convolutional neural network image architectures. It deals with unseen datasets, referred to as zero-shot learning, meaning the model can observe samples from information classes that were not included in the original training model. (Larochelle, Erhan, and Bengio 2008). CLIP is open source, whereas other robust visual classification models like DALL-E, a diffusion model that uses a different set of transformers and also developed by OpenAI were not (Ramesh et al. 2021). Diffusion models, such as Google Brain's photorealistic Imagen are considered too controversial for public release and are also not opensource (Saharia et al. 2022). Katherine Crowson refers to CLIP as the "Perceptor" (or selector) and VQGAN as the "Generator" meaning the generator of images. This means there are not smart discriminators for filtering information outside of common contemporary data harvesting techniques.

2. Method and Results – Phase One

The Google Colab notebook "VQGAN+CLIP (with pooling).ipynb" was used for initial tests, even though these models have attracted widespread criticism (Whitaker 2021, Wiggers 2022, Ruiz 2019, Crawford, Kate, and Paglen 2019). The Google Colab notebook accessed for these first experiments allowed a download of either partial or entire databanks (Crowson 2021b). The first text prompt used to render text-to-image was the English language word "queer", chosen because it has different meanings depending on its historical context. Initially the British English vernacular meant odd or strange, such as the phrase "that's a queer fellow". In recent decades the word queer has shifted to a gendered and sexualized meaning.

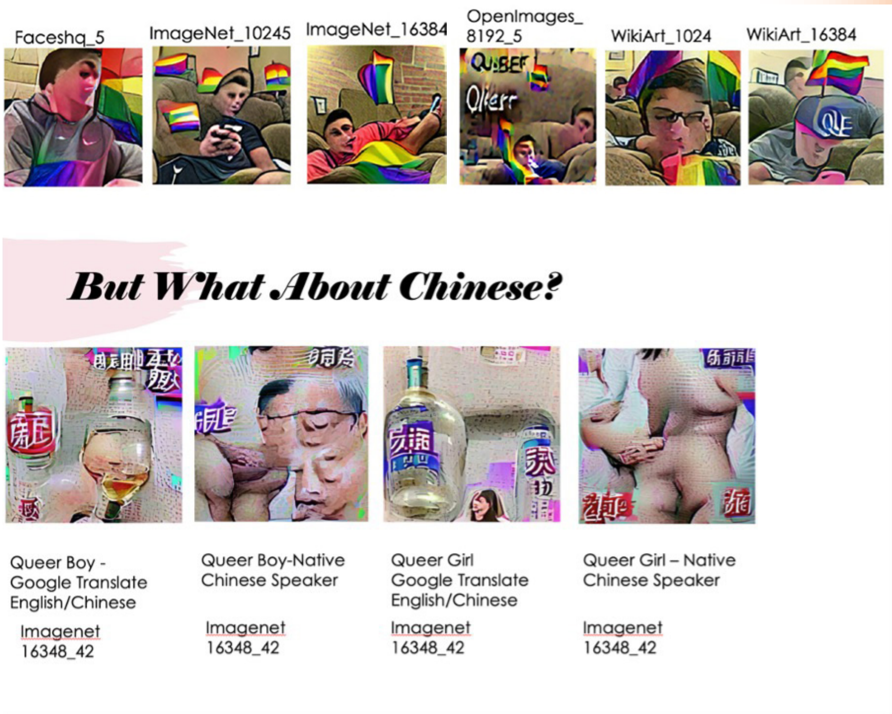

Fig. 1. Six different renderings of the word "queer" Chinese "queer boy" and "queer girl".

Six image datasets were used for the initial Google colab inquiry ; Faceshq_5; ImageNet_10245; ImageNet_16394; OpenImages_8192_5; WikiArt_1024; and WikiArt_16384 and various seeds were used to render the images. The larger the seed number, the longer it takes to generate an image and the more GPU it uses (Olah et al. 2017). All datasets returned strikingly similar images. One other modifier or text prompt was added to the query to see if it changed the type of image by using the Chinese language - queer boy and queer girl. Only one dataset, Imagenet 16384 was used for this inquiry with a consistent cycle seed value of 42. Google Translate was used for a literal translation and the slang or colloquial vernacular of the term of native speakers was also used. The Google Translate Colab initially included a translate option, but it was subsequently removed without notice. The standard image returned alcoholic drinks, ambiguous and nonsensical Chinese characters and a rainbow flag off to the side. The slang translation returned a male face with the eyebrow morphing into possible eyeglasses or an eyebrow with pinkish skin tone, a nude amorphous body, and indecipherable Chinese characters. The same variables were run for Chinese queer girl. They returned bottles of alcohol, pink colors, and a fleshy colored background with a figure with long hair. The slang variation returned a possible rainbow flag, a sexualized rear torso view of luscious pink flesh, and incomprehensible Chinese characters. The Hindi and Tamil fluent programmer used the Coco image bank with a seed of 42. Hindi queer returned pseudo-Hindi script, a wheat-colored and multi-colored object, and a splotchy green background. Queer in the Tamil language resembled a landscape with a pink river against a pale blue background without any representation of a human being. Queer boy in Hindi and Tamil used the Coco repository with a seed of 42. The Hindi queer boy looked like a topographical map with pseudo-Hindi characters and the Tamil queer boy had pink sections, and a dark brown undistinguishable object. Queer girl in Hindi resembled an orange, white, and green flag, and landscape with indecipherable script. Queer girl in Tamil resembled an old sandstone temple with a blue sky and grotesque reddish-brown faces.

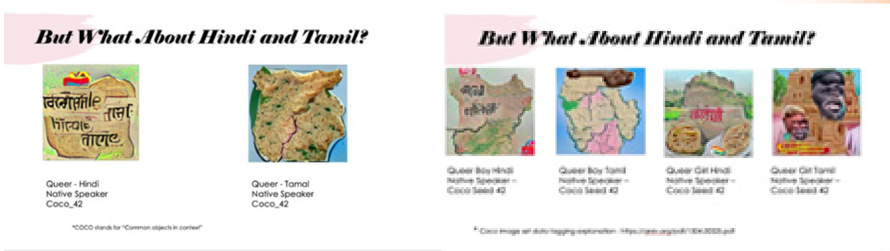

Fig. 2. Hindi and Tamil interpretation of the word "queer", Hindi, and Tamil native speaker "queer boy" and "queer girl".

3. Methods and Result – Phase Two

Experimenting with VQGAN and CLIP I found the results unusable to make a movie, and instead made a five-minute narrated video, "Language Is Leaving Me" using original and copyright free archival footage. LILM explored my epigenetic trauma as an agnostic Jew whose ancestors fled to the United States from the Eastern European Pale of Settlement in 1906. In 2022 the opensource visual text-to-image model Stable Diffusion was released by Stability.ai (Runway 2023) drawing on data sets from LAION 5-B image bank (Schuhmann et al. 2022) based on a particular type of diffusion model called Latent Diffusion (Rombach et al. 2021). The dataset for Stable Diffusion was based on the 2-B English language label subset of LAION-5-B a general crawl of the internet created by the German charity LAION (Schuhmann et al. 2022). LAION-5-B uses an aesthetic predictor, a numeric value that indicates how much someone likes an image. Schuhuhmann states "4000 samples were annotated in a scale from 0 to 10 to be good looking or not" (Schuhmann et al. 2022). I found that simplistic approach problematic.

3.1. Implementation

Giovanni Lion, a software developer based in Hong Kong and I used a special VPN high end server to access his version of AUTOMATIC 1111 and build text-to-image and image-to-image scenarios in Stable Diffusion (AUTOMATIC1111 2023). I divided my original LILM movie into 57 scenes as individual .png files using a Split-To-Frames convertor written by software developer Anton Vykhovanets. These scenes rendered 50-200 .pngs per scene, the cinematic equivalent of individual frames per second. The .pngs were batch processed in an image-to-image comparison and the text prompts came from my Enlish language video translated into four different “librettos” in Chinese, Yiddish, Tamil and Xhosa through Google Translate.

English Original

The head of the residency, my friend and I all had coffee.
I told my host I was an artist from New York City.
That is all I told her.
She looked distressed and said she had something very important to give me.
We never met before.

Yiddish

רעד הויפט פון דער רעזידענץ, איך און מיין פריינד האבן אלע געהאט קאווע
 איך דערציילט מיין באַלעבאָס איך בין אַ קינסטלער פון ניו יארק סיטי.
דאָס איז אַלע איך דערציילט איר.
 זי האט אויסגעזען נויט און געזאגט אז זי האט עפעס זייער וויכטיק צו געבן מיר.
 מיר קיינמאָל באגעגנט פריער.

Mandarin

驻地负责人、我和我的朋友都喝了咖啡。
我告诉主人我是来自纽约市的艺术家。
这就是我告诉她的全部内容。
她看起来很苦恼，
说她有很重要的东西要给我。
我们以前从未见过面。

Tamil

ரசிெடன்சி தலைவர், என் நண்பர் மற்றும் நான் அனைவரும் காபி சாப்பிட்டோம்.
நான் நியூயார்க் நகரத்தைச் சேர்ந்த கலைஞர் என்று எனது தொகுப்பாளரிடம் கூறினேன்.
நான் அவளிடம் சொன்னது அவ்வளவுதான்.
அவள் மன உளைச்சலுக்கு ஆளாகியிருப்பதைப் பார்த்து, என்னிடம் மிக முக்கியமான ஒன்றைத் தருவதாகக் கூறினாள்.
நாங்கள் இதற்கு முன் சந்தித்ததில்லை.

Xhosa

Intloko yendawo yokuhlala, mna nomhlobo wam sasinekofu.
Ndaxelela undwendwe lwam ukuba ndiligcisa elisuka kwisiXeko saseNew York.
Yiloo nto kuphela endamxelela yona.
Wayekhangeleka enxunguphele, esithi unento ebalulekileyo afuna ukundinika yona.
Asizange sidibane ngaphambili.

1. Example of a FFMPG script written by Anton Vykhovanets.

The files rendered using the image-to-image comparisons and were made into a MP4 video through the use of an FFMPG script deployed in Terminal.

```
ffmpeg -r 7 -start_number `ls | grep -E '^[0-9]+\.jpg$'
| sed 's/\.jpg$//' | sort -n | head -n 1` -i %05d.jpg
resulting_video.mp4
```
[1]

The FFMPEG conversion frames per second in the .MP4 contained highly disruptive flickering. After many attempts I located a usable flicker rate of approximately 20 fps. The biometric implementation of the sound environment included a customized motherboard designed by Wiktor Krokosz. The board received EMG input from the facial muscles of an audience volunteer and registered a frown (negative) or a smile (positive) corresponding along with the general reactions of the audience.

Fig. 3. Frame from LILM -Left to right circular - Chinese “其中一个人就在他正在吃的坟墓边缘”, English (original) “One of the men right at the edge of the grave he was eating”, Yiddish” שטנעמ י ד ן ופ ר ענייא ןסעגעג ר ע ט אה , רבק ן ופ ד נאר ם ייב , ךיילג ן”, Xhosa “Enye indoda ibisitya kanye ekupheleni kwengcwaba”.

4. Discussion

The first tests rendered by Google Colab with VQ Gan and CLIP revealed outrageous gaps and misrepresentations with just the English language word queer. The six datasets should have rendered multiple perspectives but did not. In theory this could have been due to the consistent seed number of 42 but it seemed one standardized algorithmic transformer was interpretating a very nuanced human representation throughout six completely different datasets. For the second experiment with queer girl and queer boy not as many trials were run and not as many different image databanks were used, but the results were still flawed. The Chinese, Hindi, and Tamil scripts revealed cultural biases, incomplete or non-existent data and linguistic tagging and weighting of images that made no sense. For LILM I ran my short video through Stable Diffusion using Automatic 1111 web-based user interface. The text prompt was expanded to include newly introduced image-to-image transformers of the LAION 5-B dataset to see how AI could render my psychological, cultural, and spiritual trauma into another visual image bank’s interpretation. The result showed ghastly cultural deformations.

Ancient cultures expressed themselves thorough cave paintings using visual semiotic meanings understood by the individuals who created the images in the place-based context where they were created. Meaning in the age of algorithmic smoothing and machine learning creates semantic taxonomies containing seismic fault lines in terms of interpretating algorithms into images. This is referred to as the 'epistemics of training sets" or the fraught and complex relation between images and the concepts that tie those images to their linguistic meaning and that meaning to the semantic image tagging prompts deployed by algorithms (Cuzzolin 2021). This can also be thought of as "catastrophic forgetting". Currently image tagging consists mostly of nouns, with verbs and adjectives thrown in to enhance their meaning. Epistemic AI attempts to create new mathematical models for decision making. Its operating premise is that instead of inferring predictive models about data it has at its disposal, it will assume it has a paucity of data from which to make any sort of conclusion. Only since 2016 has the issue of epistemic uncertainty in machine language become an area of focus. It acknowledges the difficulty of comprehending in domain and out of domain sampling that causes model cognition to be insufficient. Another factor is the mundane human labor of images tagging by individuals whose first language is not English, and who lack the understanding and cultural nuance to comprehend what they are looking at. Paid by the word, it is advantageous for them to perform a cursory analysis of an image, tag it and move onto the next task at hand (Ruiz 2019; Crawford and Paglen 2019; Gebru et al. 2018; Joler and Crawford 2018; Zhou et al. 2021). As this is not a paper on the politics of the labor of image tagging, it is only briefly mentioned.

5. Conclusion

Text-to-image transformer technologies VQGAN, CLIP, and various diffusion models such as Dall-e and Stable Diffusion are powerful additions in building AI based visual technologies. In the coming years these technologies will be used throughout many areas of contemporary life, from medical diagnosis to immigration, employment, education, credit and banking, logistics, jurisprudence, advertising, the art world, NFTs, game development, the metaverse and many more. What these investigations demonstrate is how inaccurate and clumsy algorithmic process are when given open-ended multi-nuanced text prompts for visual information. When terms are linguistically misaligned with cultural norms or underrepresented linguistic and cultural image datasets, inaccuracies multiply. Deeply problematic is their final rendered output. Using different image banks Faceshq_5, Imagenet_10245, ImageNet_16384, OpenImages8192_5, WikiArt_1024, and wikiArt_16384 and Coco combined with English, Chinese and Tamil revealed serious issues of bias and relevance. This may have played a partial reason for the disappearance of the translate option from this specifically referenced Google Colab notebook. When I ran image-to-image comparisons with the Chinese, Yiddish, Tamil, and Xhosa language scripts the problems multiplied exponentially.

There is an implicit but unstated assumption that whatever cannot be taxonomized and tagged will be rendered irrelevant or just cease to exist. This includes any culture, person, organization, or entity not aware of, capable of, or interested in digitizing their individual and collective visual memories and histories. Data not scrapped from contemporary easily accessible sources will be excluded, marginalized, misunderstood, and most certainly misrepresented if it manages to be represented at all. Since there are approximately 7000 languages in the world this is a problem multiplying exponentially when applied to linguistic models that are not notated and codified. Extrapolating to include styles of visual codification based on that linguistic tagging raises the likelihood for misunderstandings, misidentifications, exclusions, and obliteration for peoples of marginalized cultures. Facebook/Meta is attempting to tackle this problem with their "No Language Left Behind: project stating, "What does it take to double the language coverage of most existing translation models while ensuring high quality and safe translations?" (Costa-jussà et al. 2022). It is dubious, given the ethical and financial implications of Meta's business models that their approach will be enough to solve the issue as their model ultimately boils down to a fiscal model, not one for the benefit of the common good. What about indigenous cultures that have little or no representation? How will Meta deal with these issues? This problem or deficit is so serious it is something world organizations like UNESCO are paying close attention to. They state that AI contains, "potential negative implications for the diversity and pluralism of languages, media, cultural expressions, participation, and equality" (UNESCO n.d.). This means the responsibility falls squarely on the artists, coders, and others to reveal and reconfigure these glaring misrepresentations.

References

AUTOMATIC1111.
2023. "AUTOMATIC1111/Stable-Diffusion-Webui: Stable Diffusion Web UI." Github. January 24, 2023.
https://github.com/AUTOMATIC1111/stable-diffusion-webui.

Carlini, Nicholas, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts et al.
2021. "Extracting Training Data from Large Language Models." *Proceedings of the 30th USENIX Security Symposium*, 2633–50.

Costa-jussà, Marta R, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, et al.
2022. "No Language Left Behind: Scaling Human-Centered Machine Translation." Cambridge, Massachusetts: Harvard University.
https://github.com/facebookresearch/fairseq/tree/nllb.

Crawford, Kate, and Trevor Paglen.
2019. "Excavating AI." AI Now Institute, NYU. 2019.
https://excavating.ai/.

Crowson, Katherine.
2021a. "A Katherine Crowson VQGAN+CLIP Derived Google Colab Notebook." Python Awesome. 2021.
https://pythonawesome.com/a-katherine-crowson-vqgan-clip-derived-google-colab-notebook/.

———.
2021b. "A Katherine Crowson VQGAN+CLIP Derived Google Colab Notebook." Deep Learning Daily. 2021.
http://www.deeplearningdaily.com/a-katherine-crowson-vqgan-clip-derived-google-colab-notebook/.

Cuzzolin, Fabio.
2021. "Epistemic AI." Epistic AI. 2021.
https://www.epistemic-ai.eu/.

Felsenfeld, Gary.
2014. “A Brief History of Epigenetics.” *Cold Spring Harbor Perspectives in Biology* 6 (1). https://doi.org/10.1101/CSHPERSPECT.A018200.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé, and Kate Crawford.
2018. “Datasheets for Datasets.” *Arxiv.Org.* http://arxiv.org/abs/1803.09010.

Greengard, Samuel.
2022. “Can AI Learn to Forget?” *Communications of the ACM* 65 (4): 9–11. https://doi.org/10.1145/3516514.

Hübl, Thomas, Julie Jordan, and Avritt Vii.
2020. “Healing Collective Trauma a Process for Integrating Our Intergenerational & Cultural Wounds.” In *Healing Collective Trauma, a Process for Integrating Our Intergenerational & Cultural Wounds*, 1–11. Louisville, Colorado: Sounds True.

Joler, Vladan, and Kate Crawford.
2018. “Anatomy of an AI System.” New York: SHARE Labs, SHARE Foundation, AINow Institute.

Larochelle, Hugo, Dumitru Erhan, and Yoshua Bengio.
2008. “Zero-Data Learning of New Tasks.” In *Proceedings of the National Conference on Artificial Intelligence*, 2:646–51. Chicago, Illinois: AAAI Press, Menlo Park.

NerdyRodent.
2021. “Just Playing with Getting VQGAN+CLIP Running Locally, Rather than Having to Use Colab.” Github.Com. 2021. https://github.com/nerdyrodent/VQGAN-CLIP.

Olah, Chris, Alexander Mordvintsev, and Ludwig Schubert.
2017. “Feature Visualization.” Distill Publishing. 2017. https://doi.org/10.23915/distill.00007.

Open AI.
2021a. “GitHub - Openai/CLIP: Contrastive Language-Image Pretraining.” Open AI. 2021. https://github.com/openai/CLIP.

Open, AI.
2021b. “Microscope OpenAI.” Open AI. 2021. https://microscope.openai.com/models.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal et al.
2021. “Learning Transferable Visual Models From Natural Language Supervision.” *ArXiv.Org.*

Ramesh, Aditya, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever.
2021. “Zero-Shot Text-to-Image Generation.” *Arxiv.Org.*

Rombach, Robin, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer.
2021. “High-Resolution Image Synthesis with Latent Diffusion Models.” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2022-June (December): 10674–85. https://doi.org/10.1109/CVPR52688.2022.01042.

Ruiz, Cristina.
2019. “Leading Online Database to Remove 600,000 Images after Art Project Reveals Its Racist Bias.” The Art Newspaper. 2019. https://www.theartnewspaper.com/2019/09/23/leading-online-database-to-remove-600000-images-after-art-project-reveals-its-racist-bias.

Runway, CompVis, and Stability AI.
2023. “Stability-AI/ Generative-Models.” Github. 2023. https://github.com/Stability-AI/generative-models.

Saharia, Chitwan, William Chan, Saxena Saurabh et al.
2022. “Imagen: Text-to-Image Diffusion Models.” Imagen Research Google. 2022. https://imagen.research.google/.

Schuhmann, Christoph, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes et al.
2022. “LAION-5B: An Open Large-Scale Dataset for Training next Generation Image-Text Models.” *ArXiv.Org*, October. http://arxiv.org/abs/2210.08402.

UNESCO.
n.d. “Recommendation on the Ethics of Artificial Intelligence.” UNESO. Accessed April 9, 2023. https://en.unesco.org/about-us/legal-affairs/recommendation-ethics-artificial-intelligence.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin.
2017. "Attention Is All You Need." *Arxiv.Org*, June. http://arxiv.org/abs/1706.03762.

Whitaker, Jimmy.
2021. "The Fall of ImageNet." Towards Data Science. 2021. https://towardsdatascience.com/the-fall-of-imagenet-5792061e5b8a.

Wiggers, Kyle.
2022. "3 Big Problems with Datasets in AI and Machine Learning." Venture Beat. 2022. https://venturebeat.com/2021/12/17/3-big-problems-with-datasets-in-ai-and-machine-learning/.

Xu, Tao, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He.
2018. "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks." *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1316–24. https://doi.org/10.1109/CVPR.2018.00143.

Zhou, Xinlei, Han Liu, Farhad Pourpanah, Tieyong Zeng, and Xizhao Wang.
2021. "A Survey on Epistemic (Model) Uncertainty in Supervised Learning: Recent Advances and Applications." *Arxiv.Org*, November. https://doi.org/10.1016/j.neucom.2021.10.119.