Study of Generative Artificial Intelligence's Biases on the Example of Images Produced by the Stable Diffusion Model

Oksana Herasymenko and Maksym Borysenko

Taras Shevchenko National University of Kyiv, 60 Volodymyrska Street, Kyiv, 01033, Ukraine

Abstract

Generative artificial intelligence (AI) has revolutionized various fields by allowing algorithms to create content that closely resembles the output of human labor. However, this progress also raises concerns about generative models' behavior, predictability, and transparency. With AI's ever-increasing pervasive impact on society, it is crucial to delve into this topic, given its relevance. This work aims to empirically research the subtleties of behavior in the context of discriminativeness of generative AI using the stable diffusion model as an example. By better understanding these phenomena, we will be able to meet the challenges and ensure the responsible and ethical use of generative AI in our rapidly changing world.

Keywords [1]

Generative model, images generation, content, bias, AI predictability, limitations of generative AI, variational autoencoder

1. Introduction

Technologies are constantly developing, and artificial intelligence is increasingly penetrating people's everyday lives. From virtual assistants to recommendation algorithms, the impact of AI is quite tangible. At the same time, there is a problem of predictability and understanding of the behavior of generative AI models. Their results must be scrutinized to ensure compliance with societal norms, ethics, and legality, given the unpredictable nature of generative AI and the potential impact of its generated content on society. In addition, the discriminativeness and selectivity of these models also cause particular concerns since generative models blur the boundaries between real and synthetic data. Neural networks can detect and understand complex patterns and correlations in data [1]. Unlike traditional rule-based systems, neural networks can learn these patterns automatically from the data they are trained on, without the need for explicit programming or human understanding of the underlying mechanisms. This characteristic of neural networks is often called their «black box» nature [2].

Analysts can also process vast amounts of data, but human cognitive limitations do not provide an opportunity to detect hidden patterns and establish multidimensional relationships. Neural networks, on the other hand, are excellent at capturing complex relationships and revealing subtle details that are beyond human comprehension or may be difficult to detect. They can identify nonlinear relationships, detect hierarchical structures, and perceive complex interactions between different input data characteristics.

Another advantage of neural networks is their ability to extract abstract data representations. They excel at processing high-dimensional data, which allows them to process complex information and identify complex patterns. Neural networks consist of interconnected layers of nodes that learn to convert input data into a hierarchical arrangement of features [3], and each layer captures increasingly complex representations. These hierarchical representations allow them to detect and understand patterns that may not be obvious to humans.

The nature of neural networks as «black box» systems creates some challenges. Because of their

Information Technology and Implementation (IT&I-2023), November 20-21, 2023, Kyiv, Ukraine
EMAIL: oksgerasymenko@knu.ua; bormaxv@knu.ua
ORCID: 0000-0001-6804-2125; 0009-0008-5653-4964

complex and nonlinear nature, it is challenging to understand why a network makes a particular decision or makes a particular prediction. The lack of interpretability concerns sensitive areas such as healthcare or finance, where explainability is critical [4]. Therefore, it is possible to understand the nature of their work better by conducting empirical studies of neural networks[5], including generative ones.

2. Literature review

Today, «artificial intelligence» refers to a wide range of models. They play an essential role in realizing various capabilities such as machine learning, computer vision, natural language understanding, natural language generation, natural language processing, and robotics.

Generally, AI can be divided into two types: Narrow AI and General AI [6]. General AI describes systems with human-level intelligence that can understand, learn, and use knowledge in various domains. General AI is currently a hypothetical concept and still unattainable.

Narrow AI, in turn, is designed to perform specific tasks within a narrow range of capabilities. Examples of narrow AI include voice assistants, image recognition systems, recommendation algorithms, etc. Let us dwell in more detail on the concept of generative artificial intelligence.

2.1. Generative AI and its varieties

Generative artificial intelligence is a field of AI designed to produce content or data that resembles human-made material. These systems use machine learning capabilities, particularly generative models, to create new data that exhibit patterns similar to the training data [7]. In the field of generative AI, various types of models are simultaneously being developed and used, including:

- Generative Adversarial Networks (GAN): GANs consist of two inseparable components, namely a generator and a discriminator, which are trained simultaneously. A generator attempts to generate realistic data samples covering images or text, while a discriminator distinguishes actual samples from artificially generated ones. These two components enter into a competitive interaction, and as the generator learns, it gradually masters the ability to generate increasingly plausible content;
- Transformer models: Their example is the GPT model. They rely on the mechanism of self-attention, which enables them to distinguish complex and subtle connections inherent in inputdata. These models are successfully used in numerous generative tasks, particularly for creating texts, synthesizing images, and musical compositions;
- Variational autoencoders (VAEs): These generative models can handle data encoding and decoding. Typically, a VAE consists of an encoder that transforms the input data to representthe latent space and a decoder responsible for accurately reconstructing the received data fromthe latent space. VAEs allow the creation of new samples by sampling points from the latentspace and decoding them;
- Autoregressive models: Autoregressive models are probabilistic models that generate new data through sequential production, where each element is conditionally dependent on its predecessors. For example, language models GPT-3 and GPT-4 embody autoregressive models capable of producing complex and contextually relevant textual content;
- Deep Reinforcement Learning: Deep reinforcement learning combines deep learning techniques with reinforcement learning algorithms, offering generative AI capabilities. Thesemodels master the art of interacting with the environment and are rewarded or penalized depending on their actions. By taking risks and optimizing actions that yield greater rewards,they can create new patterns of behavior or strategies.

The models mentioned above are only a small part of the wide range of generative AI models. Researchers and developers are constantly developing different permutations and combinations of methods to create models capable of generating more diverse and unique content in many fields.

2.2. Limitations of generative AI

Although AI has significantly advanced in recent years, it still has many limitations and drawbacks that must be considered for a deep dive into the subject. The main limitations include:

1. Dependence on the quantity and quality of data [8]. Generative models must be trained on something, and the volume and quality of this data directly affect the output of the generations.The generated product may be inaccurate, inconsistent, or biased without a sufficiently large and relevant dataset;
2. Difficulty with rare or new examples. As a consequence of the previous point, artificial intelligence may not cope appropriately with situations that differ from the training data. In such cases, it may produce unrealistic or false results;
3. Limitation in the creation of original content [9]. Generative AI learns from certain data sets and tries to reproduce respective patterns and regularities, combine existing data, and compose something new. However, this is the drawback. AI cannot produce genuinely new results because it relies on this data as a foundation, which causes originality to suffer;
4. Interpretability and explainability. Many generative AI models, such as deep neural networks, are considered «black boxes», meaning there is no way to understand how and why they get the results they do [9]. The lack of interpretability and explainability is critical in areas where accountability and transparency are crucial;
5. Lack of control. Generative models often lack detailed control over the output [10]. Although they can create fairly complex content, directing or influencing the generation process to achieve specific goals or adhere to certain constraints may be difficult;
6. Ethical problems and bias. Generative AI models can accidentally amplify bias or uneven quantitative distribution of training data [11, 12]. If the data contains biases, such as gender orracial biases, the generated content may reflect and consider these biases. Ensuring fairness andreducing bias in generative AI models remains a significant challenge;
7. Lack of context. Models often do not understand the context of the generated results, which can lead to inconsistent results [13]. They frequently generate reasonable content but fail to demonstrate a more profound understanding or logical consistency;
8. Computing and time resources. The processes of training and operation of AI models are very demanding on computing resources, in particular, powerful equipment and large amounts of memory [14]. Also, the load and processing time often grow exponentially with a linear increase in quality, which raises the question of feasibility. Computing units also consume a lotof electricity, creating many related costs.

3. Problem statement

The work aims to empirically study the behavior of generative neural networks using different training methods and combinations of image generation tags. The research aims to study the capabilities of these neural networks as «black box» that can detect hidden patterns that may be imperceptible to humans, even when analyzing large volumes of data. This research is important and relevant because, despite the impressive performance of neural networks in various tasks, there is still no clear explanation of why they make some choices or make certain decisions.

In particular, this experiment is designed to explore their behavior with different training methods and image generation tags. Thus, gaining additional insights into their decision-making processes is possible. These studies are essential because they allow us to more effectively use the potential of neural networks and improve their performance in various fields of application. Understanding the factors that influence neural network behavior and studying the impact of different training methods and image generation tags can lead to advances in fields such as computer vision, natural language processing, and creative content creation [15].

As mentioned above, today, researchers pay considerable attention to eliminating bias in various AI models [16]. One of the ways in this direction is eliminating a specific part of the characteristics of the object's description, which, according to scientists, will avoid bias. On the other hand, the loss of significant characteristics can lead to non-objectivity. The solution to this problem for generative models can be a clear and complete description of the characteristics of the object whose image is generated. The main task of this study is to compare two ways of generating images, without detailing the characteristics and with detailing, to determine empirically the drawbacks of each of them.

4. Experiment description

Images of bags were chosen for the experiment. The bag was the main object of the image, accordingly. The main idea of the experiment is to distinguish two main characteristics of image objects and to explore the generation of images with and without specifying the corresponding values of the characteristics. Such characteristics determine the color of the bag and its type, which determines its shape. Five colors and five types of bags were considered during the experiment. The implementation of the experiment was carried out in several stages, the main of which are:

- creating a set of images of bags of different colors and types, and in this set, the amount of each type of bags and the amount of each color of bags are the same;
- training of two models with and without indicating the color and type of the bag, respectively;
- generation of images with different prompt tags utilizing the models obtained at the previous stage;
- analysis of the results.

Each stage is described in more detail in this section's paragraphs, and the results' analysis is in the next section of this document.

4.1. Specifying a generative model for the experiment

The Stable Diffusion generative neural model was chosen for the study. Stable Diffusion is an open-source text-to-image model that can generate images based on textual descriptions. It is a variant of the latent diffusion model, a type of deep generative neural network that can produce realistic images by iteratively denoising random noise [17]. Stable Diffusion works by using a variational autoencoder (VAE) to encode an image into a latent vector that represents the high-level features of the image. It then uses U-Net, a convolutional neural network with feedforward connections, to decode the latent vector into an image. U-Net also accepts asinput a text encoder that encodes the text description into another vector that controls the image generation process. The text encoder is based on the CLIP model, which can learn the semantic similarity between images and text. Stable Diffusion generates an image by rejecting random noise and applying a sequence ofsmoothing steps, each of which makes the image more realistic and closer to the textual description. The model learns to recognize images by reversing the process of adding noise to the training images. Thus, the model can be trained on both noisy and clean images and generate diverse and high-quality images. Stable Diffusion can be used for various tasks, such as creating images from scratch, modifying existing images based on text, filling in missing details, or enhancing low-resolution images.

4.2. Dataset creation

In the beginning, choosing a suitable dataset has also to be done. After an unsuccessful attempt to find the required number of similar images on the Internet, it was decided to generate the corresponding photos using the same Stable Diffusion. An open-source web interface was locally deployed on GitHub for convenient and high-level interaction with generation models (the owner is the user AUTOMATIC1111). A general prompt was created, and 250 images of different bags were generated. As mentioned above, two characteristics were selected among the main ones: bag color and type. We settled on the types of bags such as Backpack, Tote Bag, Duffel Bag, Clutch Wallet, Drawstring Bag. The chosen colors were black, yellow, green, blue, and red. For each combination of these criteria, ten images were generated. Figure 1 shows an example of bags generated with the tag «Backpack».

Similarly, an image of bags with tags «Tote Bag», «Duffel Bag», «Clutch Wallet» and «Drawstring Bag» was generated.

4.3. Models’ training

After creating the dataset, it was necessary to decide on a basic model for further training since there are already many variations of them. It was decided to choose the last original model from Stability AI, trained to generate photos with an extension of 512 by 512 pixels – stable-diffusion-2-1-base [18].

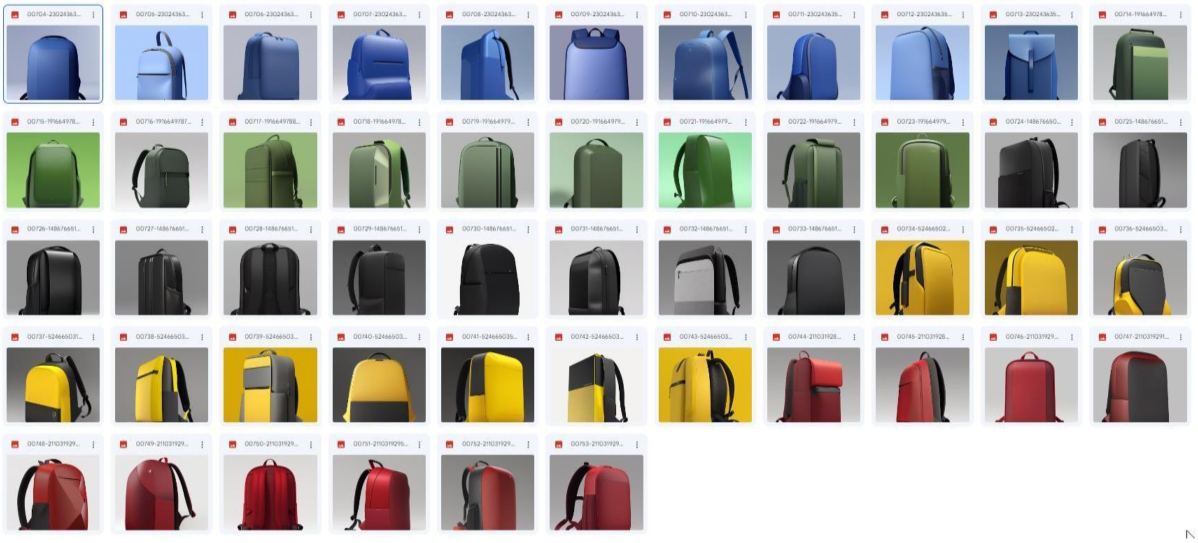


Figure 1: Images of bags generated with the tag «Backpack»

A large amount of video memory is required to model's training, so finding an external mean of calculations was necessary. After reviewing the available options, it was decided to use the Google Colab service. Among the available GitHub repositories, the project TheLastBen/fast-stable-diffusion [19] was found, which was already configured for convenient training such models. It was used as a base for the experiment. First, the selected project was cloned to a separate account. Next, the cells of the project were executed in a predefined sequence, including Google Drive with the data attaching to the project, loading all dependencies into the runtime, the neural network version selection in the Model_Version item (V2.1-512px). The last step was creation of a training session name. This step is essential, so let us discuss it in more detail.

The neural network has its database of known words (tokens) and associations with them. Therefore, using a token for learning in the form of a commonly used word is not advisable; otherwise, learning this concept will not be clean. In this case, the network will already know this concept, and it will not be easy to understand in what proportions the model used old and new information during generation. So, to teach a model of a specific concept, it is advisable to use a fictional word-token.

In this case, it was necessary to teach the network a collective image of the concept of «bag». The token that is understandable to a human but new to the network was chosen: bbbaaaggg. Next, according to the instructions from the project, all photos for training were renamed in the format bbbaaaggg(1).png, bbbaaaggg(2).png, bbbaaaggg(3).png, etc. The next step was to load all the images into the runtime. After that it was necessary to set parametersfor training, an example of which can be seen in Figure 2. We stopped at 15000 steps for training the model and 1000 steps for training the text encoder. Other parameters were left unchanged. At the end of the training, which lasted about 4 hours, the finished trained model was saved to the connected Google drive.

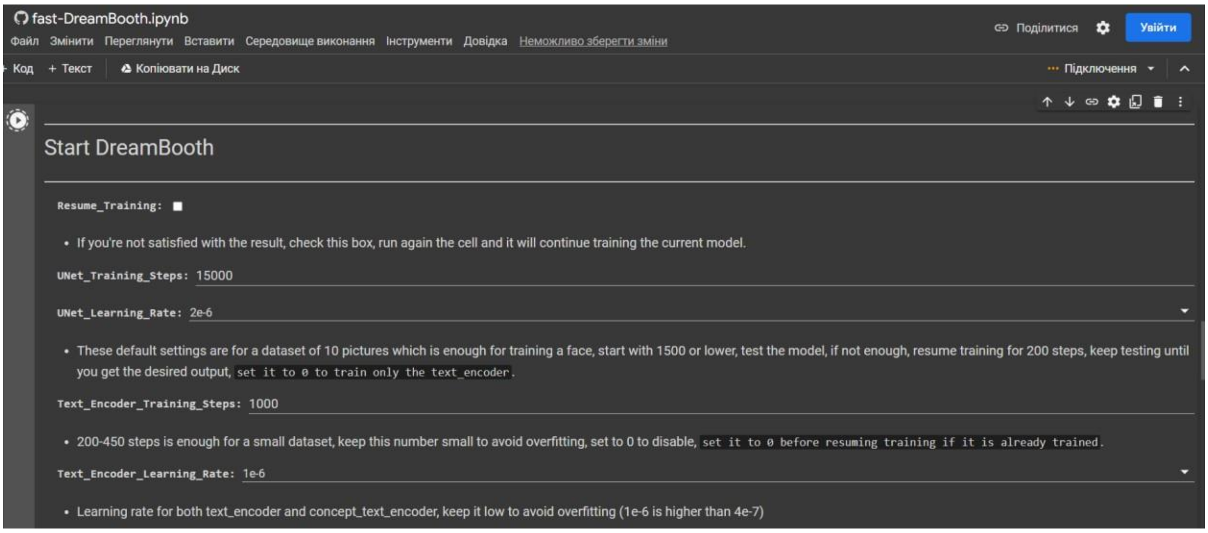
After training the regular model, it was decided to make its counterpart but with an additional description of each photo for better results. So, another training of the model was carried out, which consisted of repeating all the steps described above, but this time every image was accompanied by a corresponding text file with the same name as the image itself (bbbaaaggg(1).txt, bbbaaaggg(2).txt, bbbaaaggg(3).txt, etc.). The file contained tags that described the bag according to its type and color parameters (example can be found at Figure 3). A different session name (bbbaaagggdes) was also chosen for this training, and an option to use external descriptions was activated during the setup phase for the training. As a result, another model was created.

After training several sets of images with different tags were generated to explore the differences of the trained models, and the results were analyzed and presented in the next section of this paper.

3. Results and discussion

After creating the models, the images generated were analyzed for compliance with the specified parameters. Verification started with the model trained without description (hereafter, the first model).

Image sets were generated with different prompts to obtain the images. Figure 4 shows an example of 100 generated images for the prompt «bbbaaaggg» without additional parameters.


fast-DreamBooth.ipynb
Файл Змінити Переглянути Вставити Середовище виконання Інструменти Довідка Неможливо зберегти зміни
Поділитися
Увійти
Код
+ Текст
Копіювати на Диск
Підключення

Start DreamBooth

Resume_Training: □

- If you're not satisfied with the result, check this box, run again the cell and it will continue training the current model.

UNet_Training_Steps: 15000

UNet_Learning_Rate: 2e-6

- These default settings are for a dataset of 10 pictures which is enough for training a face, start with 1500 or lower, test the model, if not enough, resume training for 200 steps, keep testing until you get the desired output, `set it to 0 to train only the text_encoder`.

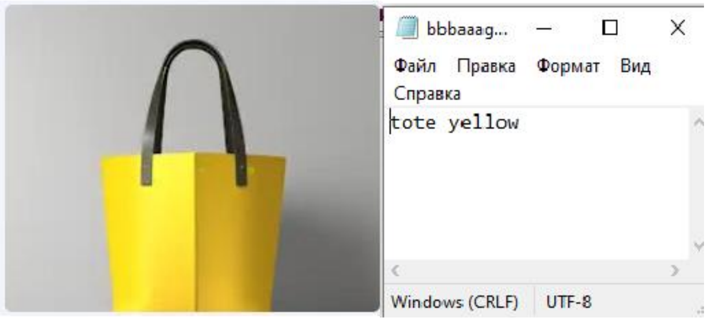
Text_Encoder_Training_Steps: 1000

- 200-450 steps is enough for a small dataset, keep this number small to avoid overfitting, set to 0 to disable, `set it to 0 before resuming training if it is already trained`.

Text_Encoder_Learning_Rate: 1e-6

- Learning rate for both text_encoder and concept_text_encoder, keep it low to avoid overfitting (1e-6 is higher than 4e-7)

Figure 2: Setting of the model training parameters



tote yellow

Figure 3: An example of a yellow tote bag and the content of a file with its description

As it turned out, the model learned to generate images of bags similar to the input data quite well. The style, angle, composition, shapes, and color scheme are similar to the original dataset. All types of bags and colors that were in the references are present in the pictures. Images with artifacts, such as missing the expected object (Figure 5), may sometimes appear, but such cases are quite rare.

We can often see the mixing of different parameters and features. For example, colors appear in images between pre-made colors (Figure 6 a), such as orange, a combination of red and yellow. In the generated sets, we can also notice completely new colors compared to the original images, such as white (Figure 6 b), or even a combination of elements inherent in their various types in one bag (Figure 6 c), which is devoid of logic in practical terms.

Upon analyzing the generation results, it was observed that green and black are the dominant colors for the depicted bags. Tote Bags and Drawstring Bags are the most popular types of bags in generated sets. Moreover, Drawstring bags have a significant advantage in terms of quantity. At the same time, it is worth noting that the training sample contained the same number of each type of bag and the same number of bags of each color. It can be assumed that the model during generation prefers a specific type of bag or color based on some dependencies, unforeseeable for a human, that it has detected. It can be an example of a bias even for such simple characteristics as the color and shape of the object. These conclusions are drawn from analyzing a relatively small number of image sets. In order to formulate generalizing judgments in this context, it is necessary to conduct additional studies designed to analyze the proposed hypothesis with statistical evaluations. Several generations with various more specific prompts were conducted to check the controllabilityof the results. For example, Figure 7 shows 100 generated pictures for the prompt «bbbaaaggg, blue». In general, the trained model is able to cope with such a task. Although there were deviations from theexpected result, most of the resulting images followed the prompt. Nevertheless, in this case, we can see that mainly Tote Bags and Drawstring Bags were generated, confirming the previously put forwardhypothesis.

Then we check images, obtained by specifying only the type of bag. So, Figure 8 depicts 100 generated pictures for the prompt «bbbaaaggg, clutch». The model generates bags of this type quite accurately. Although there were deviations from the expected result, most of the resulting images followed the prompt. The model generated mainly bags in cold tones, which also may indicate model's bias. After testing the model trained without image description, the model trained on images with description was tested (hereafter, the second model). Some images were generated utilizing it to accomplish that. The first and most interesting feature is that the model did not learn to generate the general appearance of the bag but instead created a somewhat abstract object that resembled the input references in form and style.

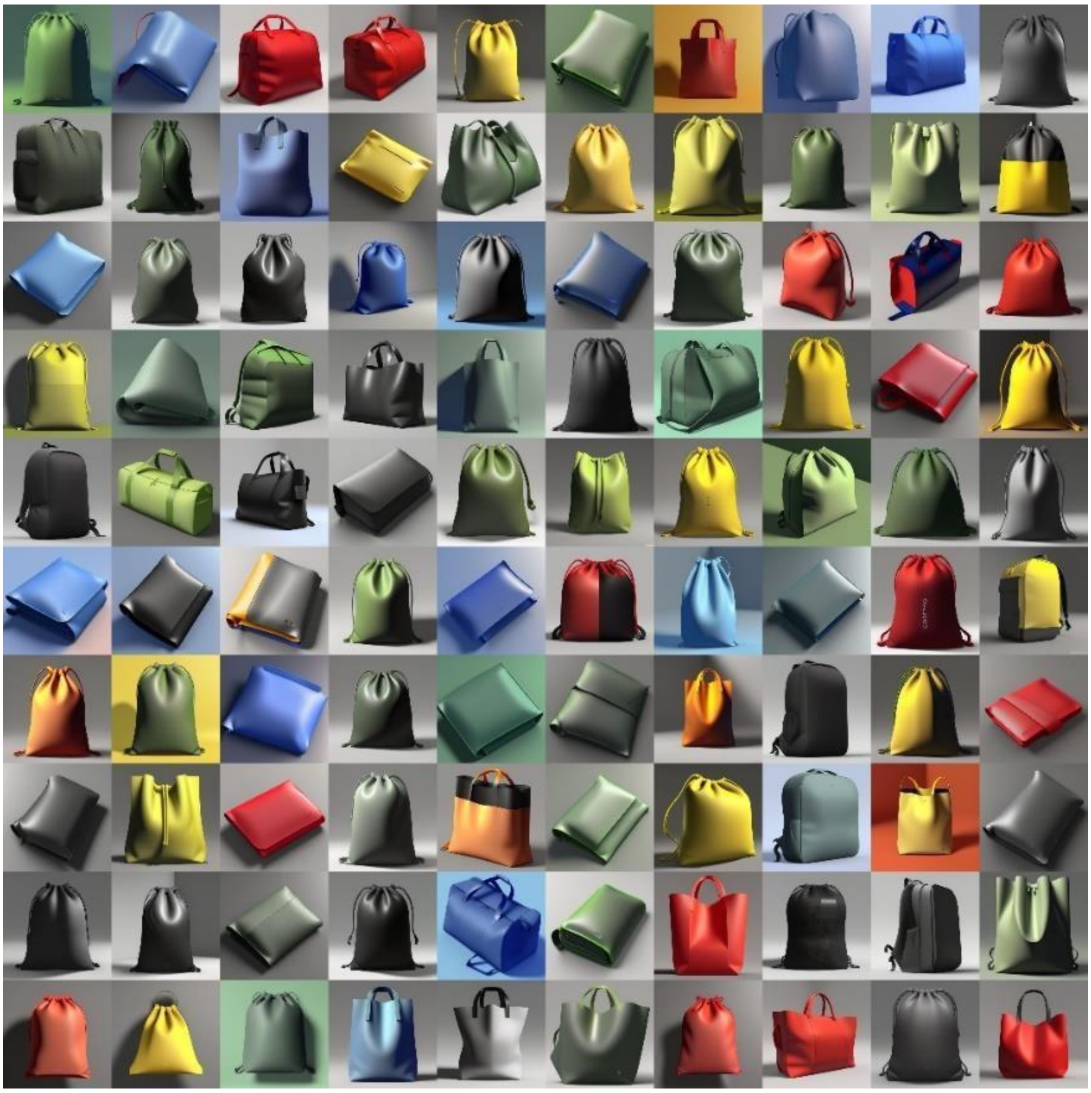


Figure 4: An example of images, generated by the first model with the prompt «bbbaaaggg»

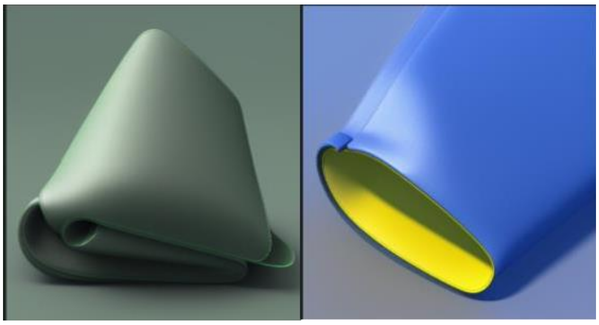


Figure 5: An example of artifacts among generated images

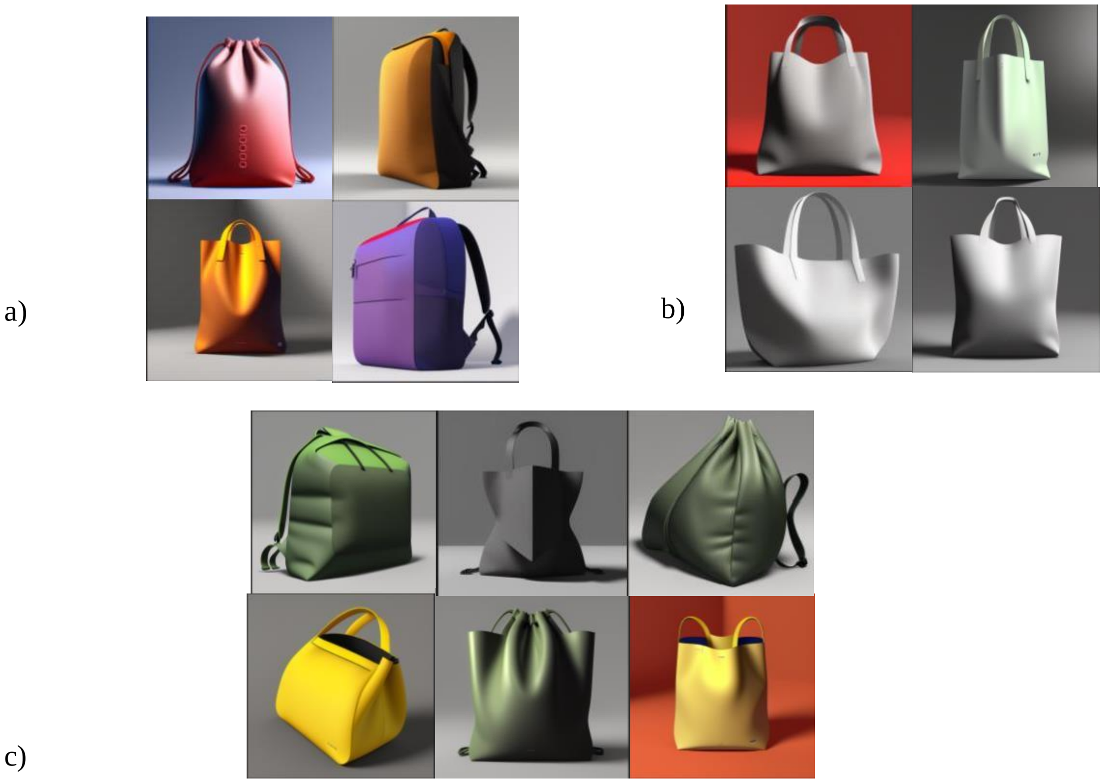


Figure 6: Examples of images with mixed or new characteristics (relative to the initial data set): a) color mixing; b) the new color - white; c) mixing elements of different types of bags

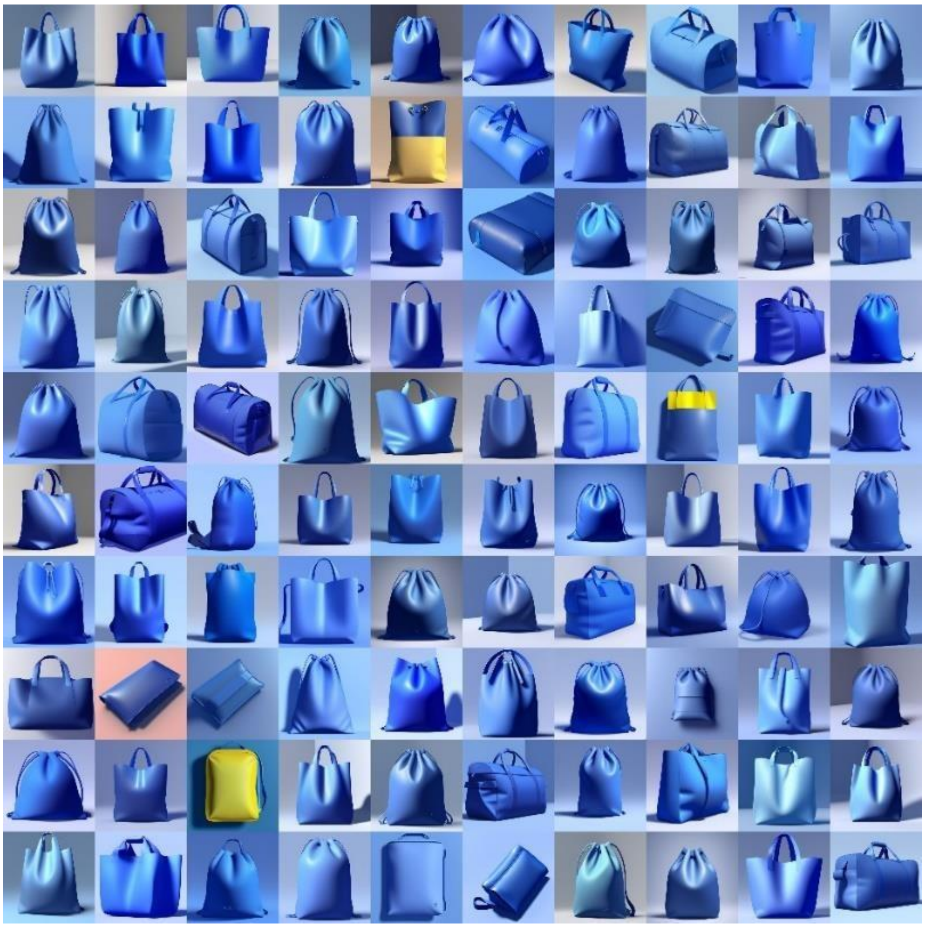


Figure 7: An example of images, generated by the first model with the prompt «bbbaaaggg, blue»

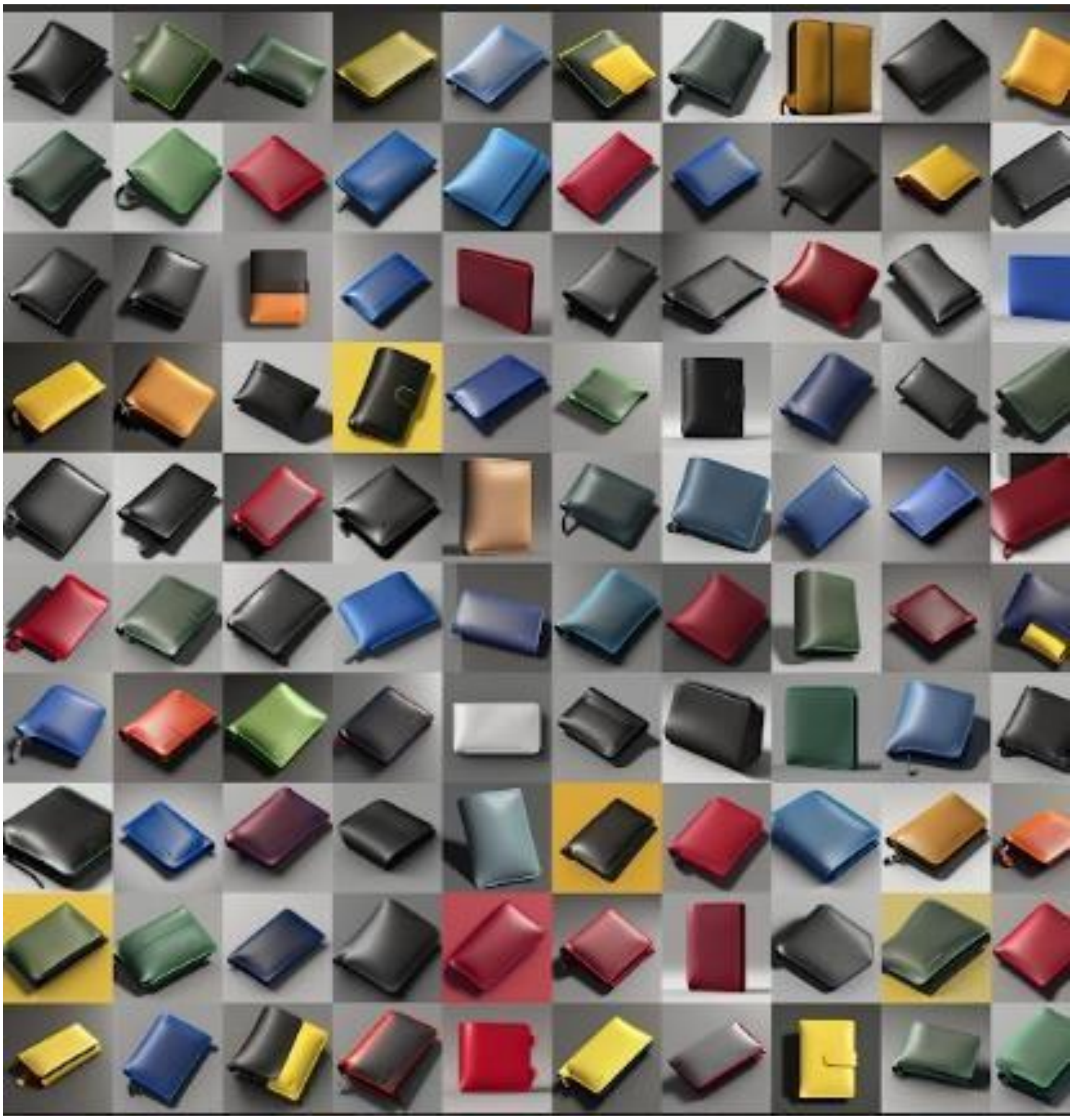


Figure 8: An example of images, generated by the first model with the prompt «bbbaaaggg, clutch»

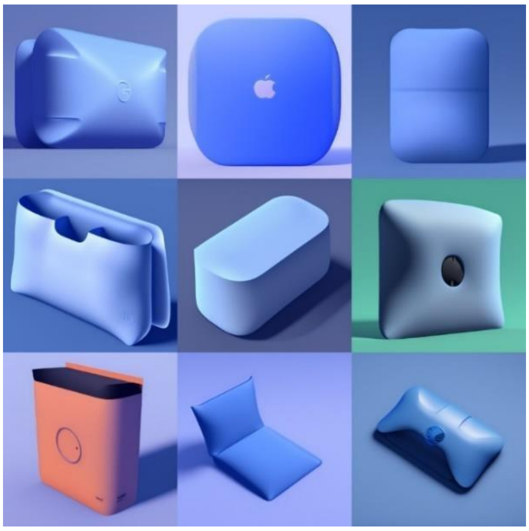


Figure 9: An example of images, generated by the second model with the prompt «bbbaaagggdes»

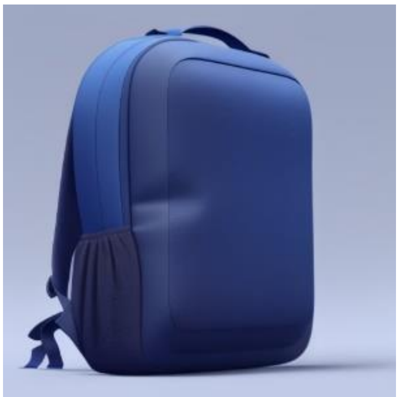


Figure 10: Image, generated by the second model with the prompt «bbbaaagggdes, backpack blue»

Figure 9 shows an example of a set with nine generated images for the prompt «bbbaaagggdes» created by the second model. This example can serve as a confirmation of the hypothesis put forward by other studies, which states that removing important characteristics from datasets is inappropriate to eliminate bias. It is better to specify characteristics more precisely when generating images or obtaining a forecast (for predictive models).

However, if in the same model, a more specific description is given in the recording format it was trained to, the result will be correct. For example, Figure 10 depicts an example of generated images for the prompt «bbbaaagggdes, backpack blue». A similar result was obtained when adding a description of the name of a particular type of bag to «bbbaaagggdes» via a comma (as it was specified in the description files for training the corresponding model).

4. Conclusion

After analyzing the images generated by the two considered models, the option without additional description was more successful in a practical sense. Its results were more predictable without description and, with additional description, were no worse than the other model. In turn, the model trained with the descriptive tags could only generate a satisfactory result by specifying the details. Identifying the precise cause of this behavior without ambiguity can be challenging. However, we assume that such shortcomings were obtained due to an unsuccessful way of constructing the description or, in general, the impracticality of training the model on such a concept as a bag, with an additional description. On the other hand, although the first model was trained without a description of what exactly a blue backpack looks like, such concepts as «blue» and «backpack» are already embedded in the initial model and can be operated on as needed, directing the generation process in one direction or another. This experiment laid down a collective image of specific bags of certain types and colors.

So, after analyzing the generation results, it was found that when generating images without clearly specifying the parameters, black and green prevail in colors, and the Tote Bag and Drawstring Bag types prevail in the types of bags, although the training sample contained the same number of each type of bag and the same number of bags of every color. Grounded on that, we assume that the neural network during generation prefers some type of bag or color based on some unforeseeable dependencies it has detected and can be an example of a bias even for such simple characteristics as the color and shape of the object. These conclusions are based on analyzing a relatively small number of generated images. In order to formulate generalizing judgments in this context, it is necessary to conduct additional studies designed to analyze the proposed hypothesis with statistical evaluations.

Nevertheless, it is important to note that training the model for specific concepts without additional description and context may be more appropriate for a particular class of problems since the generative network does a pretty good job while combining different tags and peculiarities into a complete picture.

5. Reference

[1] M. H. A. Banna et al., Application of Artificial Intelligence in Predicting Earthquakes: State-of-the-Art and Future Challenges, in IEEE Access 8 (2020), 192880-192923. doi:

10.1109/ACCESS.2020.3029859.

[2] A. Rai, Explainable AI: from black box to glass box, Journal of the Academy of Marketing Science 48, (2020), 137–141. https://doi.org/10.1007/s11747-019-00710-5

[3] C. Shorten, Hierarchical Neural Architecture Search, 2019. URL: https://towardsdatascience.com/hierarchical-neural-architecture-search-aae6bbdc3624

[4] J.M. Durán, K.R. Jongsma, Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI, Journal of Medical Ethics 47, (2021), 329-335.

[5] M. Favaretto, E. De Clercq, B.s. Elger, Big Data and discrimination: perils, promises and solutions. A systematic review. Journal of Big Data 6, 12, (2019). https://doi.org/10.1186/s40537-019-0177-4

[6] T.L. Ang, M. Choolani, K.C. See, K.K. Poh, The rise of artificial intelligence: addressing the impact of large language models such as ChatGPT on scientific publications, Singapore Medical Journal 64(4), (2023), 219-221. doi: 10.4103/singaporemedj.SMJ-2023-055.

[7] Generative AI – What is it and How Does it Work?, 2023. URL: https://www.nvidia.com/en-us/glossary/data-science/generative-ai/

[8] Data Dependencies. Google for Developers, 2023. URL: https://developers.google.com/machine-learning/crash-course/data-dependencies/video-lecture

[9] N. Ameen, G.D. Sharma, Sh. Tarba, A. Rao, R. Chopra, Toward advancing theory on creativityin marketing and artificial intelligence, Psyhology & Marketing 39(9), (2022), 1802-1825.

[10] C. Gordon, What Are The Risks Of Google And Microsoft Advancing Their Generative AI Innovations?, 2023. URL: https://www.forbes.com/sites/cindygordon/2023/05/11/google- strikes-back-on-microsoft/?sh=255526db463d

[11] D. Varona, J.L. Suárez, Discrimination, Bias, Fairness, and Trustworthy AI, Applied Sciences 12(12), (2022). https://doi.org/10.3390/app12125826

[12] X. Ferrer, T. v. Nuenen, J. M. Such, M. Coté and N. Criado, Bias and Discrimination in AI: A Cross-Disciplinary Perspective, IEEE Technology and Society Magazine 40(2), (2021), 72-80.doi: 10.1109/MTS.2021.3056293.

[13] M. Kejriwal. Artificial Intelligence needs to be more context-aware, 2021. URL: https://mkejriwal1.medium.com/artificial-intelligence-needs-to-be-more-context-aware-ecc3097ee2ea

[14] AI is harming our planet: addressing AI's staggering energy cost, 2022. URL: https://www.numenta.com/blog/2022/05/24/ai-is-harming-our-planet/

[15] A. Bozkurt, Generative artificial intelligence (AI) powered conversational educational agents:The inevitable paradigm shift, Asian Journal of Distance Education 18(1), (2023).

[16] L. Moerel, Algorithms can reduce discrimination, but only with proper data, 2018. URL: https://iapp.org/news/a/algorithms-can-reduce-discrimination-but-only-with-proper-data/

[17] Stable Diffusion Public Release, 2022. URL: https://stability.ai/blog/stable-diffusion-public-release

[18] stabilityai/stable-diffusion-2-1-base. URL: https://huggingface.co/stabilityai/stable-diffusion-2-1-base

[19] GitHub – TheLastBen/fast-stable-diffusion. URL: https://github.com/TheLastBen/fast-stable-diffusion