Trends and Challenges of Text-to-Image Generation: Sustainability Perspective

Dora Ivezić
University of Zagreb, Faculty of Electrical Engineering and Computing

Marina Bagić Babac
University of Zagreb, Faculty of Electrical Engineering and Computing

Abstract

Text-to-image generation is a rapidly growing field that aims to generate images from textual descriptions. This paper provides a comprehensive overview of the latest trends and developments, highlighting their importance and relevance in various domains, such as art, photography, marketing, and learning. The paper describes and compares various text-to-image models and discusses the challenges and limitations of this field. The findings of this paper demonstrate that recent advancements in deep learning and computer vision have led to significant progress in text-to-image models, enabling them to generate high-quality images from textual descriptions. However, challenges such as ensuring the legality and ethical implications of the final products generated by these models need to be addressed. This paper provides insights into these challenges and suggests future directions for this field. In addition, this study emphasises the need for a sustainability-oriented approach in the text-to-image domain. As text-to-image models advance, it is crucial to conscientiously assess their impact on ecological, cultural, and societal dimensions. Prioritising ethical model use while being mindful of their carbon footprint and potential effects on human creativity becomes crucial for sustainable progress.

Keywords: artificial intelligence; natural language processing; text-to-image generation; sustainability; ethical artificial intelligence

Paper type: Research article

Received: 17 Feb 2023
Accepted: 29 Aug 2023

DOI: 10.2478/crdj-2023-0004

Introduction

In recent years, text-to-image generation has seen significant progress, focusing on developing models capable of generating high-quality images from textual descriptions. This progress has been driven by the increasing demand for visual content in various domains, such as art, photography, marketing, and learning. For instance, in the domain of art, text-to-image models have been used to generate images based on literary works, enabling new forms of artistic expression and interpretation (Antol et al., 2015

Text-to-image models have also been applied in marketing to create realistic product images for e-commerce platforms, improving the user experience and increasing sales (Tomičić Furjan et al., 2020). In the field of learning, text-to-image models have been used to generate educational material, enabling a more interactive and engaging learning experience (Reed et al., 2016). Additionally, text-to-image models have been employed for various other applications, such as generating virtual and augmented reality images, generating visual explanations for natural language processing tasks (Puh and Bagić Babac, 2023a), and creating personalised visual content for social media platforms. The advance of the text-to-image models opens new business opportunities and contributes to digital transformation (Tomičić Furjan et al., 2020).

Despite the significant progress made in this area, some challenges, including sustainability issues, need to be addressed, such as ensuring the legality and ethical implications of the final products generated by these models. It is important to ensure that the generated images do not violate copyright or privacy laws and do not perpetuate harmful or offensive stereotypes, thereby contributing to the sustainable and responsible development of text-to-image models. In this regard, it is crucial to develop ethical guidelines and best practices for the sustainable development and deployment of text-to-image models.

To provide a comprehensive overview of the latest trends and developments in text-to-image, this paper describes and compares various text-to-image models, addressing technical advancements and sustainability considerations, and discusses the challenges, opportunities, and sustainability implications in this field. By analysing the current state of text-to-image research, this paper sheds light on the potential of this field while emphasising the importance of sustainable practices. It also suggests future directions for researchers and practitioners to ensure that text-to-image advancements align with broader sustainability objectives.

Basic Text-to-Image Models

Text-to-image artificial intelligence (AI) art involves two main processes: Image Manipulation and Image Generation (Mao et al., 2017). As the name suggests, Image Manipulation consists of changing a given image to match the query text, while Image Generation consists of creating an image from scratch. This paper focuses on Image Generation, which is a multi-modal problem, meaning that for one input text, there is

an infinite number of possible images. The solution to this problem lies in deep learning, specifically in Generative Adversarial Networks (GANs) and in specific text-to-image engines which rely on diffusion models. A big influence on the outcome is the way the text is vectorised. That is the first step in every image generation model. In this work, some of the best models for this task are presented, explained, and compared.

Basic text-to-image models are the foundation of the text-to-image generation field. These models are designed to generate simple images based on textual descriptions and are often used as the starting point for more advanced and sophisticated models. Basic text-to-image models typically involve a combination of natural language processing and computer vision techniques, and they can range from rule-based systems to more complex deep learning models. Despite their simplicity, basic text-to-image models play an important role in demonstrating the feasibility of text-to-image generation, and they provide a useful baseline for comparing and evaluating more advanced models. This section explores some of the most common and well-established basic text-to-image models, including their strengths and limitations.

Before getting into the structures of specific models, it is important to explain the structure of GANs and diffusion models briefly. Transformer models are also important in this domain (Vaswani et al., 2017), but without getting into too many details, it is enough to understand that they can map the input well by keeping track of the context. In other words, they can detect how some data depends on another, and that property is called attention (Cvitanović and Bagić Babac, 2022).

The structure of a GAN has two parts: the generator and the discriminator (Creswell et al., 2018). The generator network generates images, and the discriminator network has to say if the produced image is real or synthetic. The generator wants to fool the discriminator into classifying synthetic images as real. The parameters of the generator and the discriminator are updated through backpropagation, alternating between the training of the two networks. Not to forget, the generator parameters go through the discriminator while backpropagating.

Diffusion models take an input, such as an image, and gradually add Gaussian noise to it until only noise is left. After that, they try to backtrack the original input from the noisy output. In other words, every step of the process includes some information, and the information is lost as the diffusion process progresses. The goal is to work backwards and return to the original image. They work as a long Markov chain, meaning that the current step depends only on the previous step. This makes the process tractable for the noise adding to be reversed later (Ronneberger et al., 2015).

The noise is removed one step at a time by a specific Convolutional Neural Network (CNN) called U-Net (Krizhevsky et al., 2012). It is called that because of its shape, as it first makes a small representation of the image and then samples it back to the original dimensions.

Diffusion models have several advantages over GANs: the ability to generate high-quality samples, learn the underlying data distributions, and generate samples with high diversity. They are also more stable to train and can be trained with a closed-form

likelihood function, which makes them more easily explainable (Ronneberger et al., 2015).

Some of the most used datasets for training these models are Microsoft COCO (Common Objects in Context), which contains 328k labelled photos of 91 object types (Lin et al., 2014). Then, Google's Open Images Dataset (OID) contains 1.9M labelled photos of 600 object classes, and there are many other big and diverse datasets, such as CUB, LN-COCO, and ImageNet.

Smaller datasets, used in some papers for comparison of different models, are the Oxford-102 flowers dataset and Caltech CUB-200 birds dataset, which include unlabelled images of 120 types of flowers and 200 types of birds. Five captions are assigned to each image collected by Reed et al. (2016). In their work, they make text embeddings using a char-CNN-RNN encoder (Girshick, 2015), making images and captions close in a common embedding space.

Evaluation metrics

Text-to-image metrics

There are different ways of evaluating text-to-image engines and GANs. The most intuitive one is the human evaluation, in which a group is asked to rate the image based on a few factors, including the correlation with the text. However, this is a time-consuming method, but there are automated methods to do so.

A frequent metric used to assess the quality of images created by a generative image model, such as a GAN, is the Inception Score (IS). This metric was originally made to measure only the image quality; however, it implicitly tries to match the description text to the image (Bodnar, 2018).

The Frechet Inception Distance (FID) score is an upgrade to the IS since IS does not capture how synthetic images compare to real images, as FID does. This evaluation method is the most widely used (Salimans et al.>, 2016).

IS and FID scores compare the features of the generated images to the features extracted from the Inception-V3 model, which makes them less appropriate for complex datasets (Hu et al., 2023).

TIFA (Text-to-Image Faithfulness evaluation with question Answering) is a new and innovative evaluation metric measuring the generated image's adherence to its text input. First, it generates question-answer pairs using a language model on the input text. Second, it answers the questions only by analysing the generated image via VQA (Visual Question Answering) models. This approach is associated better with human judgement. However, it struggles with counting objects and adjusting their spatial relations (Hu et al., 2023).

Another useful method is generating captions for the produced image and comparing them to the original text. The language similarity metrics used for this task can be BLEU,

METEOR, and CIDEr. The downside of these automated metrics is that they are based on statistical measurements, and their value is generally weaker than human evaluators (Papineni et al., 2002).

NLP Metrics

NLP (Natural Language Processing) metrics provide different perspectives on evaluating NLP systems, and it is essential to choose the appropriate one based on the specific task and context. No single metric can fully capture the nuances of human language understanding, so using multiple metrics is often recommended for a more comprehensive evaluation. This is why, even for this task, human evaluation (Bagić Babac, 2023) is the most precise but too expensive. The automatic NLP evaluation methods try to mimic human judgement but often do not provide the wanted amount of detailed granularity (Dunđer, Seljan & Pavlovski, 2021).

BLEU (Bilingual Evaluation Understudy) is a metric commonly used to evaluate the quality of machine translations by comparing them against one or more reference (human-generated) translations. It calculates the precision of n-grams (sequences of n words) in the candidate translation that appears in the reference translations. BLEU considers unigrams, bigrams, and trigrams up to a certain n-gram order and then computes a weighted geometric mean of these precisions (Papineni et al., 2002).

BLEU keeps track of the word order by giving a bigger contribution to the higher n-gram matches. However, the lower n-grams can also noticeably increase the score. This might be a downside of the BLEU score in some applications, such as the automatic translation of poetry, where the translated text has to be of equal composition as the original. In that application, the BLEU score is used to evaluate the machine translations compared to the human translations. In order to achieve a high evaluation score, the automatic translation must match the human translation (Dunđer, Seljan & Pavlovski, 2020). However, a lower n-gram order might be preferable when evaluating the generated image captions compared to the original image captions. That is because word order is not the main concern, and considering longer chunks of words might lessen the significance of important words in the final score (Sah et al., 2018).

The BLEU metric can be combined with the NLL (Negative Log-Likelihood) metric when exploring the influence of the batch size on the performance of GANs (Krivosheev et al., 2021).

METEOR (Metric for Evaluation of Translation with Explicit ORdering) is another metric for evaluating machine translation quality. It considers both exact word matches and word order similarity. METEOR aligns the candidate and reference sentences and then computes a score based on unigram matches, stemming variants, and WordNet synonyms. It also considers chunks of words instead of just n-grams, which helps handle variations in word order (Dunđer, Seljan & Pavlovski, 2020).

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is widely used for evaluating text summarisation systems. It calculates the n-grams and word sequences overlap between the generated summary and reference summaries. ROUGE-N

measures the n-gram overlap, while ROUGE-L computes the longest common subsequence (Lin, 2004).

CIDEr (Consensus-based Image Description Evaluation) is commonly used for evaluating image captioning systems. It calculates the consensus-based similarity between generated and human reference captions (Oliviera dos Santos et al., 2021).

F1 Score (F1 Measure) is a popular metric for tasks involving binary classification (e.g., sentiment analysis) or text categorisation (Šandor and Bagić Babac, 2023). It balances precision and recall and is often used when there is an imbalance between the classes.

Perplexity is used for evaluating language models. It measures how well a language model predicts a test dataset and is based on the average log-likelihood of the test data (Jurafsky and Martin, 2000).

Word Error Rate (WER) is used for evaluating automatic speech recognition systems. It measures the percentage of word errors in the machine-transcribed output compared to human transcriptions (Ali and Renals, 2018).

Specific text-to-image models

Some of the best tools for text-to-image generation are Imagen, DALL-E, DALL-E 2, and GLIDE, which are explained below. Some of the other well-performing tools are VQ Diffusion, Stable diffusion, SDXL, Parti, DeepFloyd IF, Bing Image Creator, and Midjourney v5.2 (Zhang et al., 2023; Podell et al., 2023).

Imagen

Google's "Imagen" (Saharia et al., 2022) is a text-to-image diffusion model ~~in~~ which ~~a~~ photo-realistically presents the query text. It uses large generic transformer language models (LMs) pre-trained only on textual data to build text embeddings, which is also the key feature in this work.

The structure of Imagen contains a frozen T5-XXL encoder. After the text embedding is formed, this vector is used in the diffusion model to generate a 64x64 image. This image is then upscaled to a full resolution of 1024x1024 gradually through noise conditioning augmentation (Saharia et al., 2022).

All diffusion models in Imagen rely on a sampling technique called "classifier-free guidance", which allows the use of large guidance weights without sample degradation. Also another new technique called "dynamic thresholding" is a diffusion sampling technique used to produce realistic images. It pushes saturated pixels inwards at every iteration, actively preventing pixels from getting saturated. This also results in better image-to-text alignment.

The quality of images is evaluated using the COCO validation set, which achieves the best FID score yet, 7.27. Implicitly, this model is also adequate for image manipulation.

DALL-E

Zero-shot text-to-image generation, developed by OpenAI as DALL-E, is based on a transformer that autoregressively models text and image tokens as a single data stream. This large-scale AI language model can generate high-quality and diverse images by combining NLP and computer vision (Yildirim, 2022).

The model is trained to generate images from text descriptions in an unsupervised manner, which means it does not rely on any additional information about the objects or scenes in the images other than the text descriptions associated with them.

The training of DALL-E has two stages: compressing 256x256 RGB images into 32x32 image tokens with a variational autoencoder (dVAE) and concatenating the encoded text with the image tokens to use it in the training of an autoregressive transformer. (Ramesh et al., 2021)

Zero-shot learning is a machine learning approach that allows a model to recognise and classify new objects or classes that it has never seen before. Here, a model is trained on a set of "seen" classes. At test time, it is presented with examples from classes that it has never seen before, the "unseen" classes. The model is expected to recognise and classify these unseen examples based on its knowledge of the seen classes (Xian et al., 2018).

DALL-E 2

The architecture of DALL-E 2 (Ramesh et al., 2022) consists of two parts: the prior, used to convert a text vector to an image vector, and the decoder, used to turn this representation of an image into an actual image. The representations of text and images used in DALL-E 2 are produced by CLIP (Contrastive Language-Image Pre-training). That is a neural network which returns a caption for a given image. It is trained on pairs of images and their descriptive text. CLIP trains two encoders, one for text and another for images, to turn the input into the respective embedding.

The prior used by DALL-E 2 is a diffusion model. The decoder used is a specific diffusion model, GLIDE, which includes the input text using a transformer to eliminate the noise. However, the decoder is not identical to GLIDE since it also includes CLIP embeddings of the input text.

Similar variations of the images are obtained by extracting the image CLIP embedding and giving it as input to the decoder. The resulting images will differ slightly since CLIP will not always produce the same embedding.

Humans evaluate by checking the caption similarity, photorealism, and sample diversity. This method is also adequate for image manipulation. The challenges that DALL-E 2 doesn’t overcome are showing specific text in the image, depicting the properties of multiple objects, and producing details in complex scenes.

GLIDE

Nichol et al. (2021) proposed a zero-shot diffusion model for text-conditional image synthesis called GLIDE (Guided Language to Image Diffusion for Generation and Editing) after exploring diffusion models and comparing CLIP and classifier-free guidance. As his name suggests, this model can edit an image to match the input text, known as image manipulation. Guided diffusion is a process for guiding diffusion models toward the label when sampling.

Classifier-based guidance is a technique in which a separate model, called the classifier, is first trained on noisy images, and then during the diffusion sampling process, the gradients (LeCun et al., 1998) in the classifier guide the sampling towards the label. CLIP guidance uses a CLIP model as a classifier. Classifier-free guidance does not require the training of a separate classifier model. Because of that, the model holds its own knowledge during guidance, which is simpler. GLIDE found this guide to give the best results.

GLIDE uses a modified ADM (Ablated Diffusion Model) architecture, in which the label input is replaced by text conditioning information. The text is encoded into text tokens and fed into a transformer model. The final text tokens mimic the class embeddings and are concatenated to the attention context at each layer of the ADM model.

GAN-based text-to-image solutions

These solutions used to be state-of-the-art just a few years ago. However, even though they were replaced by the ones mentioned previously, they are still important.

GAN-CLS (Conditional Latent Space)

CLS stands for "Conditional Latent Space", which refers to conditioning the generator in a GAN on additional information. GAN-CLS refers to a GAN architecture where the generator is conditioned on class labels, and the discriminator is trained to distinguish between real samples from that class and the generated samples. This allows the generator to learn the characteristics of a specific class and generate samples that belong to that class.

GAN-CLS-INT is a variant of GAN-CLS that uses an interpolation step in the generator and allows it to generate more diverse and realistic samples. Interpolation, in this context, refers to the technique of creating intermediate representations by blending features from two examples, usually with different class labels, in the feature space.

The key feature of a GAN-CLS customised by Bodnar (2018) is compressing the text embedding and concatenating it to the input vector of the generator and the discriminator (Reed et al., 2016).

Stacked GANs

Stacked GANs use two GANs: Stage I, which generates a 64x64 image, and Stage II, which brings the produced image to a resolution of 256x256. Stage I is made just as in the GAN-CLS from the previous subchapter (Zhang et al., 2017). Stage II first downscales the image to a 4x4 block, which is then put through three residual layers and upscales to the final resolution (Huang et al., 2017).

A trick used in this model takes inspiration from Variational Autoencoders (VAE) and is called Conditioning Augmentation (CA). The network uses a random constant from a normal distribution to add noise into the text embedding and, as a result, improves the image variation (Bodnar, 2018).

Stacked GANs AttnGAN

AttnGAN (Attention Generative Adversarial Network) is based on the idea of using attention mechanisms to selectively focus on parts of the input text when generating the corresponding image. The attention mechanism allows the model to pay attention to different parts of the text when generating different parts of the image, and because of this, the model generates more detailed and diverse images that are more closely aligned with the input text.

The architecture consists of two components: the attentional generative network and DAMSM.

The first component encodes the text into a global sentence vector and each word into a word vector. The sentence vector generates a low-resolution image produced through a generative network. The model uses attention layers to create a word-context vector affected by the image's details. Combining this vector with the image vector originates a multi-modal context vector. The model uses this vector to generate new features for a more detailed image.

The second component uses an attention mechanism to measure the similarity between the final generated image and the original text, transformed into a word vector and a global sentence vector. This step provides the final text-image matching loss for training the generator (Xu et al., 2017).

The same authors also proposed e-AttnGAN (enhanced AttnGAN), which focuses only on fashion image synthesis and is trained on the FashionGen and DeepFashion-Synthesis datasets (Elasri, et al., 2022).

MirrorGAN

MirrorGAN has a mirror structure and is trained to generate images and text that are semantically consistent with each other by using a text-to-image-to-text framework. The idea is that the generated image should act like a mirror that reflects the text semantics.

It is trained on CUB bird and MS COCO datasets containing images and their associated captions. The architecture of MirrorGAN has three parts: STEM, GLAM and STREAM. STEM generates text embeddings using a recurrent neural network (RNN). GLAM uses the encoded text to guide the image generation using a cascade of three image generators. STREAM passes the generated image through a CNN image encoder and RNN decoder, resulting in the caption of the entering image (Qiao et al., 2019).

Energy Efficiency XMC-GAN

The XMC-GAN (Cross-Modal Contrastive Generative Adversarial Network) uses pairs of text and images for training (Zhang et al., 2021). It maximises the mutual information between image and text through multiple contrastive losses. The global sentence embedding, and the word embeddings are obtained from a pre-trained BERT model (Vaswani et al., 2017).

The model uses an attentional self-modulation generator to achieve a high text-image correlation. It means that the generator uses its output to help guide its processing of the input. This helps it better capture the dependencies and patterns in the data. It also uses a contrastive discriminator, which acts as an encoder to compute image features for the contrastive loss.

The comparison of various image generation models presented in Table 1 provides insights into each method's characteristics, datasets, and evaluation metrics.

The evaluation metrics used in the comparison provide a good indication of the strengths and weaknesses of each model. FID and IS are, as previously mentioned, commonly used metrics for evaluating image quality and diversity, respectively. CLIPscore, which represents the cosine similarity between two CLIP embeddings, and cosine similarity are used to measure the similarity between the generated images and real images or text. Human evaluation and R-precision provide a more subjective evaluation of the image quality and how well the generated images match the input text.

The comparison highlights the importance of using a variety of evaluation metrics to provide a comprehensive evaluation of the models. It also shows that there is no one-size-fits-all approach to image generation, as each model has strengths and weaknesses. For example, DALL-E can generate images from textual descriptions with high diversity and creativity, while Imagen has a more controlled approach with dynamic thresholding to ensure coherence with the input text. AttnGAN and MirrorGAN incorporate attention mechanisms to improve the image generation

quality, while XMC-GAN uses an attentional one-stage self-modulation generator and a contrastive discriminator to improve image quality and diversity.

Table 1. Comparison between different models for Image Generation

Method	Characteristics	Dataset	Evaluation metric
Imagen	T5-XXL encoder, Diffusion models with classifier-free guidance and dynamic thresholding	An intern dataset, Laion MS-COCO validation set	FID score, human evaluation
DALL-E	zero-shot	MS-COCO, CUB	IS, FID
GLIDE	Classifier-free guidance, ADM architecture	MS-COCO	IS, FID, CLIPscore, human evaluation
DALL-E 2	CLIP, GLIDE	MS-COCO	IS, FID
GAN-CLS-INT	Classifier, interpolation	Oxford-102, CUB, MS-COCO	Cosine similarity
StackedGAN	CA, 2 GANs	MNIST, SVHN, CIFAR-10	IS
AttnGAN	Attentional generative network	CUB, MS-COCO	IS
MirrorGAN	Text-to-image-to-text	CUB bird, MS-COCO	IS, R-precision
XMC-GAN	Attentional one-stage Self-Modulation Generator, Contrastive Discriminator	MS-COCO, Localised Narratives dataset, Open Images data	FID

Source: Authors

In conclusion, comparing different image generation models provides a valuable resource for researchers and practitioners working in this field. The evaluation metrics used in this comparison provide a good indication of the strengths and weaknesses of each model, highlighting the importance of a comprehensive evaluation approach. Each model's characteristics and datasets are also important factors to consider when choosing an appropriate model for a specific application.

Sustainability Perspective of the Text-to-Image Generation

Energy Efficiency

Deep learning models, including text-to-image models, can be computationally intensive and energy-consuming (Vinuesa and Sirmaçek, 2021). This energy usage has implications for sustainability, as it contributes to increased electricity consumption and carbon emissions. Text-to-image models typically consist of large neural networks with millions of parameters, and training these models requires significant computational power. The training process involves multiple iterations over large

datasets, which demands substantial processing capabilities. The training and inference processes contribute significantly to the carbon footprint, particularly when data centres rely on fossil fuels for electricity generation.

Additionally, deploying these models for inference also requires computational resources, especially for real-time applications. The energy usage can be substantial and have a notable environmental impact. The high energy consumption increases operational costs for companies and organisations implementing deep learning models on a large scale (Jamwal et al., 2022).

Researchers and engineers can adopt several strategies to improve energy efficiency in deep learning models. First, they can focus on model architecture optimisation by developing energy-efficient architectures that maintain competitive performance while using fewer computational resources. Techniques like knowledge distillation, model quantisation, and network pruning can reduce model size and complexity, resulting in lower energy consumption (Tunmibi and Okhakhu, 2022).

Second, employing specialised hardware acceleration, such as GPUs and TPUs optimised for deep learning tasks, can lead to significant energy savings compared to traditional CPUs. Third, transfer Learning and pretraining on large datasets followed by fine-tuning for specific tasks (Puh and Bagić Babac, 2023b) can minimise training time and overall energy consumption.

Additionally, quantised inference, which involves reducing the precision of model weights during inference, offers substantial energy savings without compromising performance significantly. Data augmentation techniques can also enhance efficiency by allowing models to learn from a more diverse dataset without additional training (Vinuesa and Sirmaçek, 2021).

Furthermore, implementing dynamic computation graphs enables models to allocate computational resources only when necessary, reducing energy waste during idle periods. Lastly, embracing green energy adoption and training deep learning models on data centres powered by renewable energy sources (Persello et al., 2021) can significantly mitigate the environmental impact, making AI practices more sustainable in the long run. Combining these methods allows the AI community to take substantial steps towards achieving energy-efficient deep learning models.

Life Cycle Analysis

Life Cycle Analysis (LCA) is a valuable approach to assessing the environmental impact of various products and systems, including text-to-image models. Each stage of the life cycle has the potential to contribute to the environmental footprint, and identifying ways to minimise this impact is crucial for sustainable technology development.

Data collection for training text-to-image models can involve vast amounts of digital information, including images and accompanying text. The environmental impact during this stage may include energy consumption associated with data centres, storage infrastructure, and data transfer. Minimising data collection and storage needs

and using energy-efficient servers and data centres can reduce the initial environmental impact.

The training process for text-to-image models often requires significant computational resources, leading to high energy consumption. Researchers and organisations can explore energy-efficient training algorithms, hardware, and infrastructure to reduce the carbon footprint during this stage.

Implementing text-to-image models into real-world applications, such as websites or mobile apps, may also require computational resources and energy consumption (Ivasic-Kos, 2022). Optimising the model's efficiency for inference can reduce the energy required for deployment.

During the usage phase, the environmental impact of text-to-image models will depend on the frequency and scale of their usage. Models deployed in cloud services may benefit from resource allocation strategies to optimise energy consumption based on demand.

Regular model updates and maintenance might be necessary to keep text-to-image models relevant and effective. Developers can consider strategies to minimise the frequency and scale of updates to reduce their environmental impact.

At the end of their useful life, text-to-image models may become obsolete or inefficient due to technological advancements. Proper disposal or recycling of hardware and infrastructure used for model training and implementation is essential to minimise waste and pollution.

Sustainable Applications

Text-to-image generation has significant potential in various sustainability-related applications.

Text-to-image models can convert textual data from environmental sensors, reports, or citizen science contributions into visual representations. For example, textual descriptions of air quality measurements, biodiversity assessments, or climate change data can be transformed into informative images, graphs, or maps. These visualisations can make complex environmental data more accessible and understandable for policymakers, researchers, and the public, facilitating better monitoring and reporting of environmental conditions (Lipovac and Bagić Babac, 2023).

In urban planning and land use management, text-to-image models can assist in translating textual descriptions of proposed projects, development plans, or zoning regulations into visual simulations. Before implementation, this can help stakeholders visualise the potential environmental impacts of new developments, such as buildings, infrastructure, or green spaces. By identifying potential environmental risks or opportunities early on, decision-makers can prioritise sustainable development and design.

Text-to-image models can aid in converting ecological research findings, species distribution data, or habitat descriptions into visual maps or images. This allows conservationists and land managers to understand better and communicate specific areas' ecological significance. These visualisations can support targeted conservation efforts, restoration projects, and identifying critical ecological corridors to protect biodiversity.

Climate change is a complex topic; text-to-image models can help simplify and visualise climate-related information. Textual descriptions of climate models, scenarios, or adaptation strategies can be transformed into visualisations, making it easier for policymakers and the public to grasp the implications of climate change and the urgency of mitigation efforts.

In the context of sustainable agriculture, text-to-image models can convert textual data on agricultural practices, soil conditions, and crop performance into visual representations. This can aid farmers and agricultural experts in identifying sustainable practices, optimising resource use, and make informed decisions to enhance food security while minimising environmental impacts.

Text-to-image models can be crucial in environmental education by transforming educational materials, reports, and scientific findings into engaging visual content. This can enhance public awareness and understanding of environmental issues, inspiring action towards sustainability and conservation.

E-Waste Implications

Like other AI models, text-to-image technology relies on powerful computational hardware for training and deployment. Frequent updates to these models, driven by advancements in AI research and the need to improve performance, can contribute to generating electronic waste (e-waste) in several ways (Bhatnagar et al., 2021).

Organisations may need to upgrade their existing hardware or invest in more powerful computing resources to keep up with the increasing demands of more complex text-to-image models. This can lead to the disposal of older hardware, which can contribute to e-waste (Karras et al., 2020).

Data centres hosting text-to-image models and handling model training require specialised servers, networking equipment, and storage devices. As new models are developed and deployed, the demand for data centre infrastructure may increase, leading to the replacement and disposal of older equipment.

Deploying text-to-image models in various applications may involve using edge devices like smartphones, tablets, or Internet of Things (IoT) devices. As these devices end their life cycles, they may contribute to e-waste when replaced or discarded.

By implementing mitigation strategies, the environmental impact of frequent text-to-image model updates and advancements can be minimised, leading to more sustainable development and deployment of AI technology. Organisations should

proactively collaborate with policymakers, researchers, and the industry to address the growing e-waste challenge and work towards more sustainable solutions.

Ethical Considerations

Ensuring the legality and ethical implications of the final products generated by text-to-image generation models is an important challenge that needs to be addressed (Reed et al., 2016). This is because these models have the potential to create images that can be used in a wide range of applications, including in marketing, advertising, and media (Clark et al., 2010).

One of the main ethical concerns related to text-to-image generation is the potential for these models to be used to create fake or misleading images. For example, these models can be used to create fake images of people or products (Li, Rujis & Lu, 2023), which can be used to deceive consumers. To address this challenge, there is a need for clear guidelines and regulations that can help to ensure that the images generated by these models are accurate, truthful, and not misleading (Holzinger et al., 2022).

Another ethical concern related to text-to-image generation is the potential for these models to be used to create offensive, discriminatory, or harmful images. For example, these models can be used to create images that perpetuate harmful stereotypes or promote hate speech. To address this challenge, there is a need for ethical frameworks and guidelines that can help to ensure that the images generated by these models are not harmful or discriminatory (Čemeljić and Bagić Babac, 2023).

Finally, there are also legal implications associated with text-to-image generation, particularly regarding intellectual property and copyright law. For example, if the generated images incorporate copyrighted material or infringe on someone's intellectual property rights, there may be legal implications. Hence, there is a need for legal frameworks and guidelines that can help to ensure that the generated images do not infringe on anyone's rights (Karimian et al., 2022).

Overall, ensuring the legality and ethical implications of the final products generated by text-to-image generation models is an important challenge that needs to be addressed to ensure that this technology is used responsibly and ethically. This requires a multidisciplinary approach involving collaboration between researchers, practitioners, regulators, and policymakers to develop ethical and legal frameworks and guidelines that can guide the development and use of these models.

Conclusion

Text-to-image generation is an interdisciplinary field that generates images based on textual descriptions (Reed et al., 2016). This field aims to develop systems that can translate textual descriptions into corresponding images, allowing users to visualise what they are reading (Xu et al., 2015). This field has grown significantly recently, with advances in deep learning and computer vision techniques driving new trends and

developments. However, many challenges still need to be addressed to achieve truly human-like image generation from the text. These challenges include ensuring that the generated images are coherent and consistent with the input text and improving the generated images' diversity and variety. Despite these challenges, text-to-image generation can potentially transform how we interact with and understand digital media.

This paper provides a comprehensive overview of the field of text-to-image generation, including its latest trends and developments, highlighting its importance and relevance in various domains. The paper describes and compares various text-to-image models and emphasises the significant progress made in recent years. The achievements in the domain of text-to-image generation are impressive, and this is especially visible in the state-of-the-art Imagen model. Diffusion models and GANs are the best methods to build image generation models. The initial text is encoded and used in every model to guide image generation. The training requires images and captions. The evaluation metric most often used is the FID score achieved on the testing set.

The practical implications of this paper can be useful for researchers, developers, and practitioners interested in working in this field. The paper highlights the importance and relevance of text-to-image generation in various domains, such as art, photography, marketing, and learning, which can help stakeholders in these areas explore new applications of this technology. Additionally, the paper describes and compares various text-to-image models, which can help stakeholders select the most appropriate model for their needs. The findings of this study show that this area of research has made significant progress in recent years and that state-of-the-art models can produce high-quality images from textual descriptions. This information can be useful for practitioners looking to implement this technology.

Theoretical implications of this paper encompass a thorough exploration of challenges and limitations in text-to-image generation, emphasising the integration of sustainability concerns. Ensuring legal and ethical compliance in the products of these models becomes paramount, bridging technical innovation with sustainability considerations. This integration advances responsible innovation, prompting researchers and practitioners to develop solutions that align with ethical, legal, and sustainability principles.

This study contributes to a paradigm shift by addressing these dimensions reframing innovation within a sustainable framework. The insights gained catalyse the design of strategies that harmonise innovation with societal values and environmental equilibrium. The theoretical implications highlight the essential need for a holistic and sustainable approach in advancing text-to-image generation. The exploration of challenges and solutions resonates as a call for responsible and sustainable progress in the text-to-image landscape.

References

1. Ali, A., & Renals, S. (2018). Word Error Rate Estimation for Speech Recognition: e-WER. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers). https://doi.org/10.18653/v1/p18-2004
2. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). VQA: Visual Question Answering. *2015 IEEE International Conference on Computer Vision (ICCV)*. https://doi.org/10.1109/iccv.2015.279
3. Bagić Babac, M. (2023). Emotion analysis of user reactions to online news. *Information Discovery and Delivery*, 51(2), 179-193. https://doi.org/10.1108/IDD-04-2022-0027
4. Bhatnagar, V., Sharma, S., Bhatnagar, A., & Kumar, L. (2021). Role of Machine Learning in Sustainable Engineering: A Review. *IOP Conference Series: Materials Science and Engineering*, 1099(1), 012036. https://doi.org/10.1088/1757-899x/1099/1/012036
5. Bodnar, C. (2018). Text to image synthesis using generative adversarial networks. Available at: arXiv preprint arXiv:1805.00676.
6. Čemeljić, H., & Bagić Babac, M. (2023). Preventing Security Incidents on Social Networks: An Analysis of Harmful Content Dissemination Through Applications. *Police and Security* (in press)
7. Clark, A., Prosser, J., & Wiles, R. (2010). Ethical Issues in Image-Based Research, Arts & Health, 2(1), 81-93. doi: 10.1080/17533010903495298
8. Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). "Generative adversarial networks: An overview". IEEE signal processing magazine, 35(1), 53-65.
9. Cvitanović, I., & Bagić Babac, M. (2022). Deep Learning with Self-Attention Mechanism for Fake News Detection. In Lahby, M., Pathan, A.S.K., Maleh, Y., Yafooz, W.M.S. (Eds.), *Combating Fake News with Computational Intelligence Techniques* (pp. 205-229). Springer, Switzerland.
10. Dunđer, I., Seljan, S. & Pavlovski, M. (2021), "What Makes Machine-Translated Poetry Look Bad? A Human Error Classification Analysis.", *Central European conference on information and intelligent systems*, Varaždin: Fakultet organizacije i informatike Sveučilišta u Zagrebu, pp.183 - 191.
11. Dunđer, I., Seljan, S. & Pavlovski, M. (2020), "Automatic Machine Translation of Poetry and a Low-Resource Language Pair," *43rd International Convention on Information, Communication and Electronic Technology (MIPRO 2020)*, Opatija, Croatia, pp. 1034-1039, doi: 10.23919/MIPRO48935.2020.9245342.
12. Elasri, M., Elharrouss, O., Al-Maadeed, S., & Tairi, H. (2022). Image Generation: A Review. *Neural Processing Letters*, 54(5), 4609-4646. https://doi.org/10.1007/s11063-022-10777-x
13. Girshick, R. (2015). Fast R-CNN. *2015 IEEE International Conference on Computer Vision (ICCV)*. https://doi.org/10.1109/iccv.2015.169

14. Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K. R., & Samek, W. (2022). xxAI-Beyond Explainable Artificial Intelligence. In International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers (pp. 15-47). Springer, Cham.
15. Hu, Y., Liu, B., Kasai, J., Wang, Y., Ostendorf, M., Krishna, R., & Smith, N. A. (2023). Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. Available at: arXiv preprint arXiv:2303.11897.
16. Huang, X., Li, Y., Poursaeed, O., Hopcroft, J., & Belongie, S. (2017). Stacked Generative Adversarial Networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* https://doi.org/10.1109/cvpr.2017.202
17. Ivasic-Kos, M. (2022). Application of Digital Images and Corresponding Image Retrieval Paradigm. *ENTRENOVA - ENTerprise REsearch InNOVAtion*, 8(1), 350-363. https://doi.org/10.54820/entrenova-2022-0030
18. Jamwal, A., Agrawal, R., & Sharma, M. (2022). Deep learning for manufacturing sustainability: Models, applications in Industry 4.0 and implications. *International Journal of Information Management Data Insights*, 2(2), 100107. https://doi.org/10.1016/j.jjimei.2022.100107
19. Jurafsky, D., & Martin, J.H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Prentice-Hall, Upper Saddle River, NJ.
20. Karimian, G., Petelos, E., & Evers, S. M. A. A. (2022). The ethical issues of the application of artificial intelligence in healthcare: a systematic scoping review. *AI and Ethics*, 2(4), 539-551. https://doi.org/10.1007/s43681-021-00131-7
21. Karras, T., Laine, S., Aila, T. & Hellsten, J. (2020). Training generative adversarial networks with limited data. *Proceedings of the International Conference on Learning Representations.* Advances in Neural Information Processing Systems, 33 (NeurIPS 2020)
22. Krivosheev, N., Vik, K., Ivanova, Y., & Spitsyn, V. (2021). Investigation of the Batch Size Influence on the Quality of Text Generation by the SeqGAN Neural Network. *Proceedings of the 31th International Conference on Computer Graphics and Vision*. Volume 2. https://doi.org/10.20948/graphicon-2021-3027-1005-1010
23. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84-90. https://doi.org/10.1145/3065386
24. Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324. https://doi.org/10.1109/5.726791
25. Li, F., Ruijs, N., & Lu, Y. (2022). Ethics & AI: A Systematic Review on Ethical Concerns and Related Strategies for Designing with AI in Healthcare. *AI*, 4(1), 28-53. https://doi.org/10.3390/ai4010003
26. Lin, C.-Y. (2004). ROUGE: a Package for Automatic Evaluation of Summaries. In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spain, July 25 – 26.

27. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. *Computer Vision - ECCV 2014*, 740-755. https://doi.org/10.1007/978-3-319-10602-1_48
28. Lipovac, I., Bagić Babac, M. (2023), Developing a Data Pipeline Solution for Big Data Processing, International Journal of Data Mining, Modelling and Management. Accepted for publication.
29. Lu, J., Xu, H., Yang, J., & Huang, Q. (2018). Neural baby talk. *Proceedings of the European Conference on Computer Vision* (pp. 721-736).
30. Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z. & Smolley, P. (2017). Least squares generative adversarial networks. *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2794-2802).
31. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I. & Chen, M. (2021). Glide: Towards photorealistic image generation and editing with text-guided diffusion models. Available at: arXiv preprint arXiv:2112.10741.
32. Oliveira dos Santos, G., Colombini, E. L., & Avila, S. (2021). CIDEr-R: Robust Consensus-based Image Description Evaluation. Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021). https://doi.org/10.18653/v1/2021.wnut-1.39
33. Papineni, K., Roukos, S., Ward, T. & Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. *ACL-2002: 40th Annual meeting of the Association for Computational Linguistics*. pp. 311–318.
34. Persello, C., Wegner, J. D., Hansch, R., Tuia, D., Ghamisi, P., Koeva, M., & Camps-Valls, G. (2022). Deep Learning and Earth Observation to Support the Sustainable Development Goals: Current approaches, open challenges, and future opportunities. *IEEE Geoscience and Remote Sensing Magazine*, 10(2), 172-200. https://doi.org/10.1109/mgrs.2021.3136100
35. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., & Rombach, R. (2023). SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. arXiv preprint arXiv:2307.01952.
36. Puh, K., Bagić Babac, M. (2023a). Predicting sentiment and rating of tourist reviews using machine learning, Journal of Hospitality and Tourism Insights, 6(3), 1188-1204. https://doi.org/10.1108/JHTI-02-2022-0078
37. Puh, K., & Bagić Babac, M. (2023b). Predicting stock market using natural language processing. *American Journal of Business*, 38(2), 41-61. https://doi.org/10.1108/ajb-08-2022-0124
38. Qiao, T., Zhang, J., Xu, D., & Tao, D. (2019). MirrorGAN: Learning Text-To-Image Generation by Redescription. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* https://doi.org/10.1109/cvpr.2019.00160
39. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. & Chen, M. (2022). Hierarchical text-conditional image generation with CLIP latents. Available at: https://arxiv.org/abs/2204.06125

40. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-shot text-to-image generation. *In International Conference on Machine Learning (pp. 8821-8831).* Available at: https://arxiv.org/abs/2102.12092
41. Reed, S., Akata, Z., Lee, H., & Schiele, B. (2016). Learning Deep Representations of Fine-Grained Visual Descriptions. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* https://doi.org/10.1109/cvpr.2016.13
42. Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 39(6), 1137-1149. https://doi.org/10.1109/tpami.2016.2577031
43. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *Lecture Notes in Computer Science*, 234-241. https://doi.org/10.1007/978-3-319-24574-4_28
44. Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533-536. https://doi.org/10.1038/323533a0
45. Sah, S., Peri, D., Shringi, A., Zhang, C., Dominguez, M., Savakis, A., & Ptucha, R. (2018). Semantically Invariant Text-to-Image Generation. *2018 25th IEEE International Conference on Image Processing (ICIP).* https://doi.org/10.1109/icip.2018.8451656
46. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Kamyar, S., Ghasemipour, S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., & Norouzi, M. (2022). Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. arXiv:2205.11487
47. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training GANs. Available at: https://arxiv.org/abs/1606.03498
48. Samek, W., Wiegand, T. & Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualising and interpreting deep learning models. Available at: https://arxiv.org/abs/1708.08296
49. Šandor, D., & Bagić Babac, M. (2023). Sarcasm detection in online comments using machine learning. *Information Discovery and Delivery*. https://doi.org/10.1108/idd-01-2023-0002
50. Szegedy, C., Wei Liu, Yangqing Jia, Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* https://doi.org/10.1109/cvpr.2015.7298594
51. Tomičić Furjan, M., Tomičić-Pupek, K., & Pihir, I. (2020). Understanding Digital Transformation Initiatives: Case Studies Analysis. *Business Systems Research*, 11 (1), 125-141. https://doi.org/10.2478/bsrj-2020-0009
52. Tunmibi, S., & Okhakhu, D. (2022). Machine Learning for Sustainable Development. *In Conference proceedings of the First Conference of the National Institute of Office Administrators and Information Managers (NIOAIM) between 7th and 10th February, Lead City University, Ibadan, Oyo State, Nigeria*.

53. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need, *In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), Curran Associates (pp. 6000-6010)*. Red Hook, NY, USA
54. Vinuesa, R., & Sirmacek, B. (2021). Interpretable deep-learning models to help achieve the Sustainable Development Goals. *Nature Machine Intelligence*, 3(11), 926-926. https://doi.org/10.1038/s42256-021-00414-y
55. Xian, Y., Lampert, C. H., Schiele, B., & Akata, Z. (2019). Zero-Shot Learning—A Comprehensive Evaluation of the Good, the Bad and the Ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9), 2251-2265. https://doi.org/10.1109/tpami.2018.2857768
56. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R. & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. *In International conference on machine learning* (pp. 2048-2057). PMLR.
57. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., & He, X. (2018). AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/cvpr.2018.00143
58. Yildirim, E.. (2022). Text-to-Image Generation A.I. in Architecture, In (Kozlu Hale, H., 2022). *Art and Architecture: Theory, Practice and Experience*, Lyon: Livre de Lyon, 97-120.
59. Zhang, C., Zhang, C., Zhang, M., & Kweon, I. S. (2023). Text-to-image Diffusion Models in Generative AI: A Survey. Available at: https://arxiv.org/abs/2303.07909
60. Zhang, H., Koh, J. Y., Baldridge, J., Lee, H., & Yang, Y. (2021). Cross-Modal Contrastive Learning for Text-to-Image Generation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* https://doi.org/10.1109/cvpr46437.2021.00089
61. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas, D. N. (2019). StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(8), 1947-1962. https://doi.org/10.1109/tpami.2018.2856256

About the authors

Dora Ivezić received a B.S. in Computer Science from the University of Zagreb, Faculty of Electrical Engineering and Computing, Croatia, where she is pursuing an M. S. in Computer Science. Her professional and research interests include computer vision and data science applications. The author can be contacted at **dora.ivezic@fer.hr**

Marina Bagić Babac is an Associate Professor at the University of Zagreb, Faculty of Electrical Engineering and Computing, Croatia, where she obtained her Dipl.Ing., M.Sc. and Ph.D. She also obtained an M.Sc. in Journalism from the University of Zagreb's Faculty of Political Science. She is actively engaged in several EU-funded projects in data science. She serves as a program committee member of a few international scientific conferences and journals and a reviewer in numerous international journals. Her research interests include machine learning, natural language processing, and social network analysis. The author can be contacted at **marina.bagic@fer.hr**