Project 6: Text-to-image models

Enrico Benedetti, Filippos Papandreou and Simona Scala
{ enrico.benedetti5, simona.scala6 }@studio.unibo.it, fpapan01@ucy.ac.cy

Contents

1 Abstract

This paper examines Text-to-Image (T2I) models, focusing on two popular approaches: Generative Adversarial Network (GAN)-based methods and diffusion-based methods for image generation. The first section provides a theoretical foundation by explaining these methods and comparing their limitations. The challenges faced by T2I models are also highlighted. In the second section, practical applications of T2I technologies are explored, including coding resources, local deployment options, and user-accessible services. The third section delves into the ethical, legal, and safety concerns associated with T2I models, discussing issues such as social, racial, and gender biases, the potential harm caused by fake images, propaganda dissemination, offensive content, and intellectual property violations. By examining these aspects, this paper aims to contribute to a comprehensive understanding of T2I models and their implications.

2 Introduction

Text-to-Image (T2I) models have emerged as a powerful tool for generating realistic images from textual descriptions. These models employ advanced techniques such as Generative Adversarial Networks (GANs) and diffusion-based methods to bridge the gap between text and visual content. Understanding and harnessing the capabilities of T2I models have significant implications across various domains, including creative arts, design, virtual reality, and content generation.

In recent years, two primary approaches have gained prominence in T2I research: GAN-based methods and diffusion-based methods. GANs, with their adversarial training framework, have shown remarkable success in generating high-quality images. On the other hand, diffusion-based methods leverage diffusion processes to progressively refine the generated images. While both approaches have their merits, they also exhibit limitations that need to be addressed for further advancement and wider adoption.

This paper is structured into three sections to comprehensively address different aspects of T2I models. The first section provides a theoretical overview of GAN-based methods and diffusion-based methods, discussing their underlying principles and highlighting the challenges they face. By comparing and contrasting these approaches, we aim to shed light on their respective strengths and limitations, setting the stage for future improvements in T2I research.

Moving beyond theory, the second section of this paper explores the practical application of T2I technologies. We examine various resources available for coding T2I models, making it accessible to researchers and practitioners with different levels of expertise. Additionally, we delve into options for local deployment of T2I models without requiring extensive coding knowledge, as well as user-friendly services that enable individuals to leverage T2I capabilities.

However, as T2I models become increasingly powerful, ethical, legal, and safety concerns emerge. These concerns form the basis of the third section of this paper, where we discuss the potential pitfalls and implications associated with T2I technologies. Specific issues addressed include the perpetuation of social, racial, and gender biases in generated images, the potential harm caused by the proliferation of fake images, the spread of propaganda and hurtful content, and the violation of intellectual property rights.

By examining the theoretical foundations, practical applications, and ethical considerations surrounding T2I models, this paper aims to provide a comprehensive understanding of the current landscape and pave the way for future advancements that mitigate challenges and ensure responsible and beneficial use of this technology.

3 Comparative Analysis of GAN-Based and Diffusion-Based Methods for Text-to-Image Generation

3.1 GAN-based methods

3.1.1 Introduction to GANs

GANs, which stands for Generative Adversarial Networks, are a class of machine learning models that are widely used for generating synthetic data. GANs were introduced by Ian Goodfellow and his

colleagues in 2014 [GPAM+14] and have since gained significant attention and popularity in the AI field.

The fundamental concept behind GANs is the interplay between two neural networks: the generator and the discriminator. The generator network learns to generate new samples that resemble the training data, while the discriminator network learns to distinguish between real and fake samples. These two networks are trained simultaneously, with the generator trying to generate realistic samples to fool the discriminator, and the discriminator improving its ability to differentiate between real and fake samples. For a graphic intuition, see Figure 1 below.

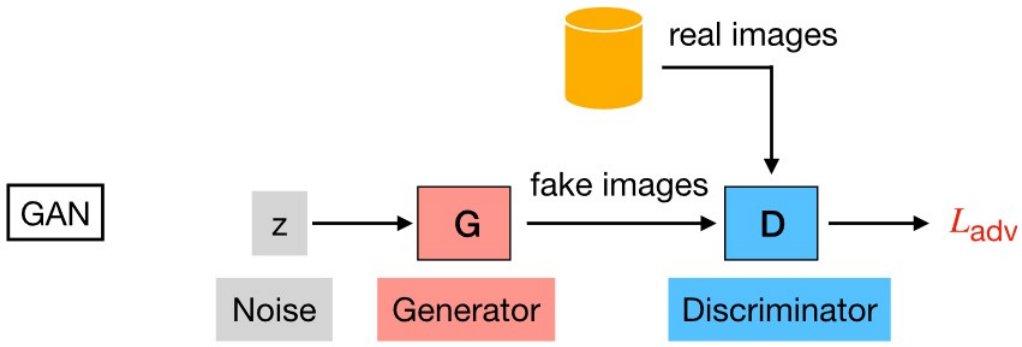

Figure 1: Simplified architecture of a GAN, from [GPAM+14]. Given noise input z randomly sampled from a normal distribution, the generator is trained to produce images to fool the discriminator. The discriminator is trained to distinguish between real and generated images.

The training process of GANs can be described as a min-max game. The generator's objective is to minimize the discriminator's ability to distinguish between real and fake samples, while the discriminator's objective is to maximize its accuracy in classifying real and fake samples. Through this adversarial process, both networks gradually improve their performance, resulting in a generator that can produce increasingly realistic synthetic data.

One of the key advantages of GANs is their ability to generate highly realistic and diverse samples across a variety of domains, including images, text, music, and even video. GANs have been successfully applied to tasks such as image synthesis, image-to-image translation, style transfer, text generation, and data augmentation.

However, training GANs can be challenging and often requires careful tuning of hyperparameters and architectures. GANs are known for their instability during training, including problems like mode collapse (where the generator only produces a limited set of samples) and vanishing gradients. Researchers have proposed various techniques to mitigate these issues, such as adding regularization terms, using different loss functions, and employing architectural modifications. [FHR+21]

Despite these challenges, GANs have made significant contributions to the field of machine learning and have opened up new possibilities for generating synthetic data. They have found applications in various domains, including art, entertainment, computer vision, and natural language processing. GANs continue to be an active area of research, with ongoing efforts to improve their stability, performance, and applicability to real-world problems.

3.1.2 Conditional approach for text-to-image synthesis using GANs

Text-to-image synthesis refers to the task of generating realistic images from textual descriptions. GANs have been successfully applied to this problem, enabling the generation of visually coherent images that align with given textual prompts. The combination of generative models and text embeddings allows GANs to bridge the gap between language and visual content.

The typical architecture for text-to-image synthesis using GANs consists of two main components: a text encoder and an image generator. The text encoder encodes the textual descriptions into a

latent space representation, often using techniques like word embeddings or recurrent neural networks (RNNs). The image generator takes this latent representation as input and produces synthesized images.

The discriminator in text-to-image GANs plays a crucial role in evaluating the generated images' quality and determining how well they match the textual descriptions. It learns to discriminate between real images and synthesized images. Both the generator and discriminator receive the textual description as a conditioning input.

This approach builds upon the work of [MO14], which introduced conditional GANs (cGANS).

To further improve the quality and coherence of generated images, researchers have explored additional techniques. For example, using attention mechanisms can help the generator focus on relevant parts of the image while synthesizing (AttnGAN by [XZH+17]). Another technique involves using multi-stage generation, where the generator first generates a rough image layout and then refines it with fine details.

Text Encoding The textual descriptions are encoded into a latent representation or embedding. Techniques such as word embeddings (e.g., Word2Vec or GloVe) or recurrent neural networks (RNNs) can be used to encode the text into a fixed-length vector. [RASL16]

Image Generation The generator network takes the text embedding as input and generates synthetic images. The goal is to generate images that are visually coherent and relevant to the provided textual description.

Discriminator The discriminator network receives both real images from the training dataset and generated images along with their corresponding text embeddings. It learns to discriminate between real and fake images and provides feedback to the generator to improve the generated image quality.

Training Process The generator and discriminator are trained in an adversarial manner. The generator tries to produce images that the discriminator cannot distinguish from real images, while the discriminator aims to correctly classify real and fake images. This iterative training process continues until the generator can generate realistic images that align with the provided textual descriptions.

Datasets Datasets for text-to-image synthesis often include pairs of textual descriptions and corresponding real images. These datasets are used to train the GAN models, where the generator aims to produce images that are visually similar to the real images given the textual input. [FHR+21]

3.1.3 Applications and challenges of GAN-based methods

Text-to-image synthesis using GANs has numerous applications, including generating scene descriptions from text, creating visual content from textual prompts, and assisting in creative tasks such as generating artwork based on textual concepts. It also has potential applications in areas like virtual reality, gaming, and content generation for storytelling and advertising.

However, challenges remain in text-to-image synthesis. Generating highly detailed and diverse images that precisely match the provided textual descriptions is a complex task. Ensuring semantic consistency, dealing with ambiguous or incomplete text descriptions, and maintaining diversity in generated samples are areas of ongoing research.

Overall, GANs have shown promising results in text-to-image synthesis, enabling the generation of visually compelling and contextually relevant images based on textual input. Continued advancements in this field hold potential for further applications and improvements in generating realistic images from textual descriptions. For example, the GFPGAN model [WLZS21] is capable of restoring faces in images, and is often used on images generated by another class of models (diffusion models) to improve perceptual quality.

3.2 Diffusion-based methods

In the following, a general formulation for the problem will be given, which will in part follow [SDWMG15]'s work. Then, after an overview of other formulations (noise score matching and differential equations), we will look at the use of diffusion models in the text-to-image task.

3.2.1 Introduction to diffusion models

Diffusion models are a family of probabilistic generative models that progressively destruct data by injecting noise, then learn to reverse this process for sample generation[YZS+23].

[SDWMG15] in their paper which introduced diffusion models in the machine learning field, describe the intuition as such:

> The main idea, inspired by non-equilibruim statistical physics, is to systematically and slowly destroy structure in a data distribution through an iterative forward diffusion process. We then learn a reverse diffusion process that restores structure in data, yielding a highly flexible and tractable generative model of the data. This approach allows us to rapidly learn, sample from, and evaluate probabilities in deep generative models with thousands of layers or time steps, as well as to compute conditional and posterior probabilities under the learned model.

In summary, diffusion models aim to learn a probability distribution over the data $p(x)$, which could be any data, images for example. Estimating it directly is usually intractable for complex distributions, because of the normalization constants. With diffusion models, instead of learning the distribution $p(x)$ directly, we learn a 'Force' $F(x) = -\nabla \log p(x)$ that, under certain conditions, has an equilibrium distribution which is exactly $p(x)$, i.e. instead of learning $p(x)$, we learn how to make some starting distribution closer to $p(x)$.

3.2.2 Formulation

The forward diffusion process transforms any data distribution into a tractable one ($\pi(y)$, 1) (for example a Normal distribution), by repeated applications of a suitable Markov diffusion kernel (T_π, 2) where β_t is the diffusion rate at 'time' t.

$$\pi(y) = \int dy' \; T_\pi(y|y';\beta) \; \pi(y') \tag{1}$$

$$q(x_t|x_{t-1}) = T_\pi(x_t|x_{t-1};\beta_t) \tag{2}$$

To represent the forward trajectory, which means to start from the data distribution (3) and to perform T diffusion steps iteratively, we use the chain rule of probability (4).

$$q(x_0) = p_{\text{data}}(x) \tag{3}$$

$$q(x_0, \ldots, x_T) = q(x_{0:T}) = q_0(x_0) \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{4}$$

The reverse diffusion process is the one used for generation. The same trajectory can be described but in reverse. It starts with the analytically tractable final distribution (5) and ends at the original data distribution (6, derived the same as 4). This is the part that has to be learned by a model through optimization.

$$p(x_T) = \pi(x_T) \tag{5}$$

$$p(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p(x_{t-1}|x_t) \tag{6}$$

The probability given by this model for the data is also

$$p(x_0) = \int dx_{1:T} p(x_{0:T}). \tag{7}$$

This distribution cannot be computed because we usually do not have the ability to compute all possible trajectories. Using annealed importance sampling [Nea98] and the Jarzynski equality [Jar97],

the relative probability of the forward and reverse trajectories, averaged over forward trajectories can be evaluated, if we assume a Markov representation, to (9).

$$p(x_0) = \int dx_{1:T}\ p(x_{0:T})\ \frac{q(x_{0:T})}{q(x_{0:T})} \tag{8}$$

$$= \int dx_{1:T}\ q(x_{1:T}|x_0) \left(p(x_T) \prod_{t=1}^{T} \frac{p(x_{t-1}|x_t)}{q(x_t|x_{t-1})} \right) \tag{9}$$

In (8) a resolution of unity $\frac{q(x_{0:T})}{q(x_{0:T})}$ is added to (7). Note that using Bayes' rule and considering it is a Markov process, it holds that

$$q(x_{0:T}) = p(x_0) \prod_{t=1}^{T} q(x_t|x_{t-1}) = q(x_{1:T}|x_0)p(x_0) \tag{10}$$

which can be used to deal with the intractability. Equation 9 can be evaluated by averaging over samples from the forward trajectory $q(x_{1:T}|x_0)$.

Then, training amounts to maximizing the model likelihood (finding parameters that maximize the probability of observing training data). Log likelihood is used because of its properties of converting products into sums, which avoids numerical overflow.

$$L = \int dx_0\ q(x_0) \log p(x_0) \tag{11}$$

$$= \int dx_0\ q(x_0) \log \left[\int dx_{1:T}\ q(x_{1:T}|x_0) \left(p(x_T) \prod_{t=1}^{T} \frac{p(x_{t-1}|x_t)}{q(x_t|x_{t-1})} \right) \right] \tag{12}$$

L is not directly maximized, but instead a lower bound is found using Jensen's inequality from which it follows that $\log \int q(z)f(z)dz \geq \int q(z) \log f(z)dz$, to move the log inside the integral. To get from (12) to (13) the equalities of the forward process (10) and a grouping of the integration variables are used.

$$L \geq \int dx_{0:T}\ q(x_{0:T}) \log \left[p(x_T) \prod_{t=1}^{T} \frac{p(x_{t-1}|x_t)}{q(x_t|x_{t-1})} \right] \tag{13}$$

$$= E_q \left[\log \left(p(x_T) \prod_{t=1}^{T} \frac{p(x_{t-1}|x_t)}{q(x_t|x_{t-1})} \right) \right] \tag{14}$$

The authors of [SDWMG15] also show (in their Appendix B) that the lower bound can be further rewritten in terms of Kullback-Leibler divergences and entropy terms that can be analytically computed. Note that the KL divergence is between the reverse transition and the forward process posterior conditioned on the starting sample, $q(x_{t-1}|x_t, x_0)$. So in a way we are computing the similarity between the posterior on the forward process and the learned reverse process.

$$L \geq K \tag{15}$$

$$K = -\sum_{t=2}^{T} \int dx_0\ dx_T\ q(x_0, x_t)\ D_{KL}(q(x_{t-1}|x_t, x_0)||p(x_{t-1}|x_t))$$
$$+H_q(X_T|X_0) - H_q(X_1|X_0) - H_p(X_T) \tag{16}$$

In general, training a diffusion probabilistic models amounts to finding the reverse Markov transition kernel that maximizes the lower bound on the log likelihood.

$$\hat{p}(x_{t-1}|x_t) = \underset{p(x_{t-1}|x_t)}{\text{argmax}} \; K \tag{17}$$

The authors then give an explicit formulation of the forward and backward process when using a Standard Normal distribution as the endpoint of the forward process $\pi(x_T) \sim \mathcal{N}(0,\, I)$. Depending on a continuous or discrete formulation, we get different diffusion models: DDPMs (Denoising Diffusion Probabilistic Models) and diffusion models based on SDEs (Stochastic Differential Equations).

3.2.3 Denoising Diffusion Probabilistic Models (DDPMs)

Forward Process The first two works that introduced DDPMs are [SDWMG15] and [HJA20]. They have made experiments on image datasets to study how the diffusion formulation performs. If Gaussian noise is added to corrupt data samples, there are a number of simplifications that can be made to the previous equations.

An uncorrupted sample from the data distribution is $x_0 \sim p(x_0)$. The forward process then becomes

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t,\, \sqrt{1-\beta_t}x_{t-1}\, \beta_t I) \tag{18}$$

where β_t is the variance coefficient whose value is set by a variance schedule $\beta_1, ..., \beta_T$. The important property that needs to hold for the variance schedule is that $\beta_t \longrightarrow 0$ so that the final distribution approaches a Standard Gaussian (which has by definition zero variance). Also, if a schedule is used such that $\beta_t \ll 1$ then the reverse diffusion kernel has the same functional form as the forward one. The schedule can be learned or can be set as an hyperparameter, for example in [HJA20] they choose a linear schedule with 1000 timesteps. Other works experimented and found cosine schedules work well.

The formulation in Equation 18 allows the direct sampling of x_t (a sample from a certain timestep in the diffusion process for a starting x_0) as

$$q(x_t|x_0) = \mathcal{N}(x_t,\, \sqrt{\overline{\alpha}_t}x_0\, (1-\overline{\alpha}_t)I) \tag{19}$$

with $\alpha_t \coloneqq 1-\beta_t$ and $\overline{\alpha}_t \coloneqq \sum_{s=1}^{t} \alpha_s$. This can be proven by applying the recursive definition and using induction.

The sampling can be performed through a reparametrization trick (that uses the inverse standardization of the samples), so that it is enough to sample from a Standard Normal distribution (20) and linearly transform the original sample x_0 as follows:

$$z_t \sim \mathcal{N}(0,\, I) \tag{20}$$

$$x_t \sim q(x_t|x_0) \tag{21}$$

$$x_t = \sqrt{\overline{\alpha}_t}x_0 + \sqrt{(1-\overline{\alpha}_t)} \cdot z_t \tag{22}$$

Reverse process The reverse process is also defined as a Markov chain, like the forward process, but in this case the transition kernel has to be learned. The starting distribution for the reverse process is a Standard Normal Distribution (23), and the learnable parameters for the chain can be set as the mean and variance at each timestep (24).

$$\pi(y) = p(x_T) = \mathcal{N}(x_T,\, 0,\, I) \tag{23}$$

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1},\, \mu_\theta(x_t, t),\, \Sigma_\theta(x_t, t)) \tag{24}$$

Sampling using the reverse process can be done in the following way:

Algorithm 1: DDPM sampling method

Input:
T - the number of diffusion steps.
$\sigma_1, ..., \sigma_T$ - the standard deviations for the reverse transitions.

Output:
x_0 - the sampled image.

Steps:

$x_T \sim \mathcal{N}(0,\, I)$
for $t = T, ..., 1$ **do**
  if $t \geq 1$ **then**
    $z \sim \mathcal{N}(0,\, I)$
  else
    $z = 0$
  $\mu_\theta = \frac{1}{\sqrt{\alpha_t}} \cdot \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}} \cdot z_\theta(x_t, t) \right)$
  $x_{t-1} = \mu_\theta + \sigma_t \cdot z$

A sample from a Standard Normal distribution is taken and then the Markov chain for the reverse process is traversed until we get to the uncorrupted sample. Thanks to the reverse process formulation each x_t can be easily computed iteratively based on the previous data point in the reverse trajectory. There is research on different and more efficient sampling methods, for instance in [KP21] and [ZC23].

Training Training is performed by optimizing the usual variational bound on log likelihood (the same as Equation 16). In [HJA20] the authors further simplify the variational bound, by fixing the covariance to a constant value (25) and rewriting the mean as a function of noise (26):

$$\Sigma_\theta(x_t, t) = \sigma_t^2 \tag{25}$$

$$\mu_\theta = \frac{1}{\sqrt{\alpha_t}} \cdot \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}} \cdot z_\theta(x_t, t) \right) \tag{26}$$

The variational bound can be rewritten (remembering the expectation formulation (14) and using the above parametrization (26) as:

$$E_{x_0,\epsilon,t} \left[\frac{\beta_t^2}{2\sigma_t^2 \alpha_t (1-\overline{\alpha}_t)} \|\epsilon - \epsilon_\theta(\sqrt{\overline{\alpha}_t} x_0 + \sqrt{1-\overline{\alpha}_t}\epsilon, t)\|^2 \right] \tag{27}$$

Note that the weighting term that multiplies the 2-norm can be removed, and [HJA20] found this simplified version to bring better sample quality in their experiments, because more loss weight is given to the more difficult task of denoising at larger t compared to when the noise level is small (t closer to 0).

The basic algorithm for training can be summarized as:

Algorithm 2: DDPM training method

Steps:

repeat
  $x_0 \sim q(x_0)$
  $t \sim \mathcal{U}(\{1, ..., T\})$
  $\epsilon \sim \mathcal{N}(0,\, I)$
  Gradient descent step on $\nabla_\theta \|\epsilon - \epsilon_\theta(\sqrt{\overline{\alpha}_t} x_0 + \sqrt{1-\overline{\alpha}_t}\epsilon, t)\|^2$
until *convergence*

So, in the case of unconditional image generation, the model is trained to predict the noise from the image, and the mean is determined according to (26), while the covariance is fixed to a constant (25). We will see later that to condition with text as well (for the text-to-image task) this formulation is still usable with minor additions.

3.2.4 Noise Conditioned Score Networks (NCSNs)

At the core of score-based generative models is the concept of score or score function [Hyv05]. NCSNs generates the samples towards decreasing noise levels and trains the model by estimating the score functions for noisy data distribution. It is estimated as the gradient of log probability density for p.

Despite different motivations, NCSNs share a similar optimization objective with DDPMs during training, and in [HJA20] it is demonstrated that they are equivalent following a certain parametrization.

3.2.5 Stochastic Differential Equations (SDEs)

The DDPM is the discrete formulation of the diffusion process.

There is a connection with Stochastic Differential Equations (SDEs) in the case of a continous formulation, which have demonstrated useful properties in practical applications of these models.

Let's briefly consider the SDE formulation for the forward process

$$dx = f(x,t)dt + g(t)dW \tag{28}$$

We have an Ordinary Differential Equation (ODE) term where there is a function of time and input, plus a value from a stochastic source (e.g. Brownian motion in the thermodynamics and physics setting; usually in machine learning it's noise from a Gaussian distribution).

As for the reverse process, we have:

$$dx = \left[f(x,t) - g^2(t)\nabla_x \log p_t(x)\right] dt + g(t)dW \tag{29}$$

where dt can be taken as a small negative step in order to go backwards. The gradient of the log of $p_t(x)$ has a connection with the score function and Noise Conditioned Score networks.

Sampling consists of integrating (29), for example with the Euler–Maruyama method [BT95]. There are papers which developed more efficient methods for sampling, such as one in which the SDE for sampling is transformed into an ODE for which better off-the-shelf solving methods exist. For a large survey of efficient techniques for diffusion models one can go to [UAP22].

3.2.6 High-Resolution Image Synthesis with Latent Diffusion Models (& Stable Diffusion)

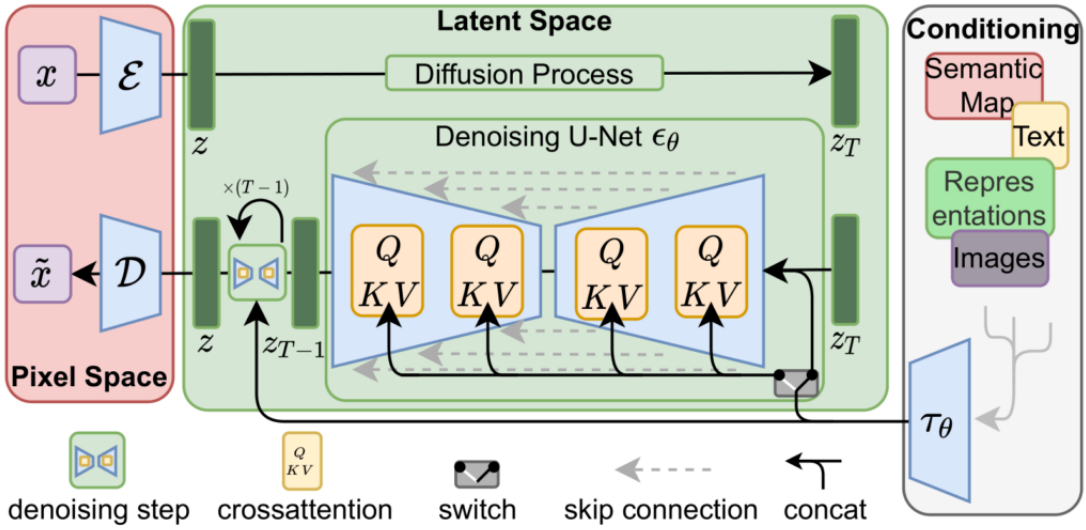

Figure 2: Stable diffusion architecture, from [RBL+22].

We will give an overview of a text to image diffusion model (Stable Diffusion) [RBL+22] to understand how it works and how the theory that we used previously is put into practice to get impressive results.

Introduction Previous work that introduced diffusion models to deal with images trained them in pixel space, while the idea of [RBL+22] is that working on image (aka pixel) space is still somewhat inefficient because computing functions in pixel space is costly. The authors propose to circumvent this by explicitly separating compression and generation phases. They use an autoencoding model which aims to learn a space that is perceptually equivalent to the image space, but which offers significantly reduced computational complexity. They empirically found that a reduction factor of 4 or 8 times provided good results. The latent space of this autoencoder model is then used to learn multiple diffusion models (e.g. DDPMs), each capable of an image-related task such as inpainting, text-to-image generation and superresolution.

Perceptual Image Compression As the red part in Figure 2 shows, the image gets compressed from pixel space to a smaller latent space through the encoder $\mathcal{E}$, and gets decoded by $\mathcal{D}$. The autoencoder network is separately trained (in an adversarial manner) to minimize the reconstruction loss between x and $\tilde{x}$.

Latent Diffusion Model and Conditioning Mechanisms The training objective is the same as the simplified from [HJA20] (equation 27 without the scaling constants).

The base architecture for the model is the U-Net [RFB15] [DN21]. To perform conditional generation (e.g on text), the additional input y is fed to a domain specific encoder τ_θ and then into the denoising U-Net through cross-attention layers. In the case of text conditioning, for instance, the text is tokenized and encoded by a transformer [VSP+17].

The final training objective then is:

$$E_{x_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(\sqrt{\overline{\alpha}_t}x_0 + \sqrt{1-\overline{\alpha}_t}\epsilon, t, \tau_\theta(y))\|_2^2\right]. \tag{30}$$

Stable Diffusion The text-to-image model (the actual Stable Diffusion model) has been made open source in a collaboration with Stability AI and it is now very popular, and many finedtuned versions have been created. The training data used for Stable Diffusion v1.0 for example is from a subset of LAION 5B [SBV+22], which is even bigger than the previous dataset used for text-to-image in the paper.

3.3 Comparisons

3.3.1 Evaluation Metrics

In this section we summarize the main methods of evaluating text-to-image methods, following the works of [FHR+21], [DNH22] and [ZZZK23]. Most of the existing metrics that assess the quality of a model are based on two aspects: image quality and text-image alignment. To assess the image quality of the model, the most common metrics are Inception Score (IS) and Fréchet Inception Distance (FID). These metrics originally come from traditional GAN tasks used to evaluate image quality. For evaluating text-image alignment, the most commonly used metric is R-precision (RP).

Inception Score IS [SGZ+16] leverages a pre-trained Inception-v3 network [DDS+09] to classify the generated images and determine their conditional label distribution. This metric aims to determine whether a decent generator can generate samples with low label distribution entropy, meaning they are easily recognizable and belong to specific categories. It also considers the diversity of generated images, looking for high entropy in the overall label distribution. A higher IS value indicates better image quality and diversity.

Fréchet Inception Distance FID [HRU+18] calculates the Fréchet distance between two sets of images: model-generated images and real images. To calculate the FID, features from each set are extracted by a pre-trained Inception-v3 network [DDS+09]. Then, these two feature sets are modeled as two multivariate Gaussian distributions. Finally, the Fréchet distance is calculated between the two distributions. A lower FID value indicates better image quality and diversity.

R-precision The concept behind RP [XZH+17] involves utilizing a synthesized image query against an input caption. Specifically, it generates an image based on a ground truth caption and employs this image to query the input description among a set of 100 candidate captions. To assess the success of the retrieval, the matching score between the generated image and the ground truth caption is evaluated. The matching score is determined by the cosine similarity between the image encoding vector and the caption encoding vector. RP represents the ratio of successful retrievals, and a higher score indicates better quality.

Human evaluation Apart from the automatic methods for evaluation, researchers have also compiled benchmarks for human evaluation (see Section 5 from [ZZZK23]). To provide a concrete example, DrawBench was introduced along with Google's Imagen [SCS+22], and it consists of a list of prompts designed to test image quality and fidelity to the caption (e.g., "Two cats and one dog sitting on the grass"). Then, human raters may be asked to give a quality score or select which set of generated images looks best (e.g., when comparing different models).

3.3.2 Models' Performances

After providing an overview of the commonly used evaluation metrics and strategies, we will now explore the challenges and limitations that are inherent in these approaches. Our analysis will specifically concentrate on two prominent works in the field of text-to-image tasks over time, namely GAN-based methods and diffusion-based methods.

GAN-based methods The survey by Frolov et al. [FHR+21] highlights that GAN-based methods have achieved the maximum performance in terms of IS and RP when evaluated against real images from the COCO dataset [LMB+14]. However, despite these high scores, the generated images remain noticeably unrealistic, as can be seen in Figure 3, raising concerns about the reliability of these metrics.

In particular, the saturation of IS can occur when the generator becomes overfitted to specific patterns or characteristics in the training set. This means that the generator focuses too much on replicating a limited set of images, resulting in a lack of diversity and realism beyond that particular context. Consequently, the inflated IS score does not accurately reflect the overall quality and realism of the generated images. Increasing the batch size during training has been suggested as a potential solution to improve the IS score. In fact, a larger batch size provides the generator with more diverse examples, potentially leading to better image quality and increased diversity in the generated samples.

Moreover, the study observed that RP scores for some models surpass those of real images, suggesting potential overfitting of this metric during training. It was hypothesized that this overfitting may occur because many models use the same text-encoder for both training and final RP evaluation. To address this issue, the proposed solution is to evaluate RP using a different model that is pre-trained on a dataset distinct from COCO.

On the other hand, the reported Fréchet Inception Distance (FID) scores of GAN-based methods are worse than those computed on real COCO images, which aligns with the challenges that these methods still face in synthesizing detailed and distinct objects. Additionally, GANs are known for being challenging to train as they often collapse unless specific hyperparameters and regularizers are carefully selected.

Diffusion-based methods Diffusion-based methods have emerged as formidable contenders in the field of image generation, surpassing the performance of GANs. Figure 4 illustrates some remarkable samples generated by Stable Diffusion model, already mentioned in Section 3.2.6. This significant advancement was demonstrated in the research conducted by Dhariwal and Nichol [DN21], where they established the superiority of diffusion-based models by showcasing their substantial improvements in terms of FID. However, compared to GANs, they tend to be slower in terms of sampling due to the inclusion of multiple denoising steps, which require additional forward passes.

Researchers, such as [LL21], are exploring methods to simplify the sampling process of diffusion models by condensing it into a single-step model. Although the samples generated by the single-step model are not currently on par with GANs, they show significant improvements compared to previous single-step likelihood-based models. This line of investigation holds promise for narrowing the sampling speed gap between diffusion models and GANs without sacrificing image quality.

Figure 3: Examples of images generated by GAN-based models trained on the COCO dataset.

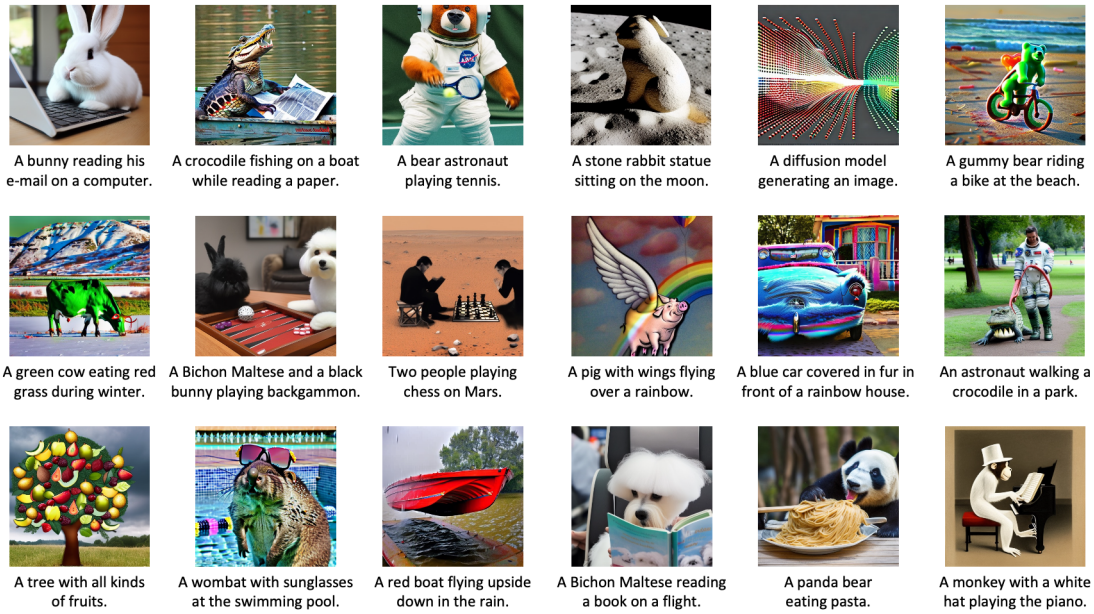

Figure 4: Images generated by Stable Diffusion based on various text prompts.

In conclusion, the development and application of Text-to-Image (T2I) models face several significant

challenges that must be addressed in order to promote fairness, effectiveness, and accurate evaluation. One crucial challenge is the limitation of data and computation, which hampers the success of these models. They heavily rely on labeled data and require substantial computational resources for training. Additionally, the large size of these models presents deployment difficulties, particularly in environments that prioritize efficiency, such as edge devices.

Another challenge lies in evaluating T2I models. Existing automatic evaluation metrics have their limitations, and human evaluation is subjective and constrained by prompt design efficiency and biases. It is essential to develop improved automatic evaluation criteria and more robust human evaluation methodologies to ensure comprehensive and reliable assessments.

Furthermore, dataset bias is a prominent issue as these models are trained on text-image pairs collected from real-world data, which can introduce biases related to race, gender, and other factors. Mitigating dataset bias requires the development of diverse and balanced datasets, coupled with new methods to reduce the impact of bias. These challenges give rise to ethical implications, which will be discussed in the final section of this paper.

4 Practical Applications and Deployment Options for Text-to-Image Technologies

In this section we propose an overview and some starting points for using text-to-image technologies in practice. First we briefly point to some resources for coding, then we talk about a way to use the models locally without directly coding, and finish with the more user accessible services.

4.1 Coding notebooks

In order to understand how to work with those kinds of models a hands-on approach is recommended.

As the most popular libraries and implementation for deep learning are commonly in Python, a good starting point could be one of Hugging Face's Jupyter Notebooks which showcase an API for diffusion models.[1]

Additionally, for a more in-depth understanding, one can look at code which goes deeper into the model's components, such as Binxu Wang's.[2]

4.2 Using text-to-image models locally

Thanks to the open source status of models like Stable Diffusion, it did not take long before the text-to-image community grew and built many applications that facilitate the use of text-to-image models locally. While it is possible to generate images online, using others' computing power, this sometimes subjected to limitations (such as having to pay for credits to use in generation).

There are also implementations that use only the CPU, for instance, but they are of course much slower than running on GPUs. There is no doubt that with research the quality-resources trade-off will improve.

It is fairly straightforward to set up a web user interface for using text-to-image generative models. One of them is the Stable diffusion webui, which is an open-source project that allows to set up a local server that a browser can access to start generating images from text prompts (and more features such as inpainting, img2img, etc.), providing an interface between the user and the many machine learning models.

One can follow an installation guide to install the application, like GitHub user AUTOMATIC1111's.[3] Things to consider when trying to make a model work locally are also the operative system, the machine learning libraries versions and the machine's graphics card type (e.g. AMD or NVIDIA).

If one wants to explore the different models' capabilities, e.g. anime style fine-tuned models, there exist websites where those models can be shared. Some popular ones are Hugging Face[4] and Civitai.[5]

For some example images of the interface refer to Appendix B.

[1] link: Diffusers API notebook.
[2] link: DiffusionFromScratch repository.
[3] link: stable-diffusion-webui repository.
[4] link: Hugging Face website.
[5] link: Civitai website.

4.3 Services for the broader public

Recently, many websites offering services regarding text-to-image were set up. They do not require any technical knowledge, so some of them have seen tremendous success and started providing those services commercially. One of the most notable ones is Midjourney,[6] which is used through the Discord message program. It requires a subscription fee to create images for commercial purposes. Generating images can also be done for free, but they cannot be used commercially and are publically displayed.[7]

Other relevant websites are those which collect prompts and generated images, along with information about the model which produced them, like Lexica.[8]

5 Ethical Implications of Text-to-Image Technology

The recent improvements in text-to-image generation and image generation in general have been met with excitement about the 'democratization of art' by the tech community, and many commercial services are already selling the technology, e.g. midjourney and many others [cite]. This very rapid development should not make us ignore the many ethical, legal and safety problems that such technology, which could become even more powerful down the line, comes with.

Academics papers almost always contain the ethics/societal implication section of their works, and the highlights are social/racial/gender bias, harm from fake images, propaganda, hurtful content, intellectual property violation. In the following we will discuss some of these issues and examples.

5.1 Bias

'Bias' in this context is the tendency of the models to generate images that reflect or reinforce existing stereotypes or prejudices based on social attributes, such as gender, race, culture, etc.

Some research on the fairness and ethics in AI applications focuses on methods, guidelines and techniques to detect and mitigate bias. It can have many forms, but it is almost inevitable and it is introduced as early as when starting to collect data for vision datasets. Bias present in collected data can be amplified by learning algorithms and can then be propagated to the tasks which build upon those models.

For a survey on bias in vision datasets and a comprehensive description of possible bias types, one can look at [FPNK22].

The work of [JOS+21] warns about the effects of using models that use synthetic data:

> The generated data learns a distribution shifted from that of the real world, one which exacerbates these biases and disproportionately underrepresents those already in the minority, both in number and quality. This poses serious ethical implications on any downstream tasks trained on a synthetically-augmented dataset.

As an example, we can take the case of PULSE [MDH+20], an upscaling model that was trained using images generated from StyleGAN [KLA19] (itself trained on CelebA-HQ [HLH+18], a dataset of high-resolution human faces). The authors of PULSE, after investigating why the algorithm worked less well for non-white people (see figure 6), conclude that it could be mainly because 'StyleGAN yields white faces much more frequently than faces of people of color, indicating more of the prior density may be dedicated to white faces'.

In the specific area of text-to-image, where as we know different prompts produce different images, a paper by [BYMC22] studied the effects of 'ethical interventions' in generation prompts, following the work from [ZKK+21] which explored this approach for textual generative models.

Such interventions consist simply in a phrase that is added to the generation prompt such as 'a photo of a doctor if all individuals can be one irrespective of gender/skin color' (see figure 7). The study found evidence that a large change in image generation can be caused by certain keyphrases in the context of the gender and cultural bias, even though this is just one of the possible research directions.

[6] link: Midjourney start guide.
[7] link: Midjourney showcase.
[8] link: Lexica website.

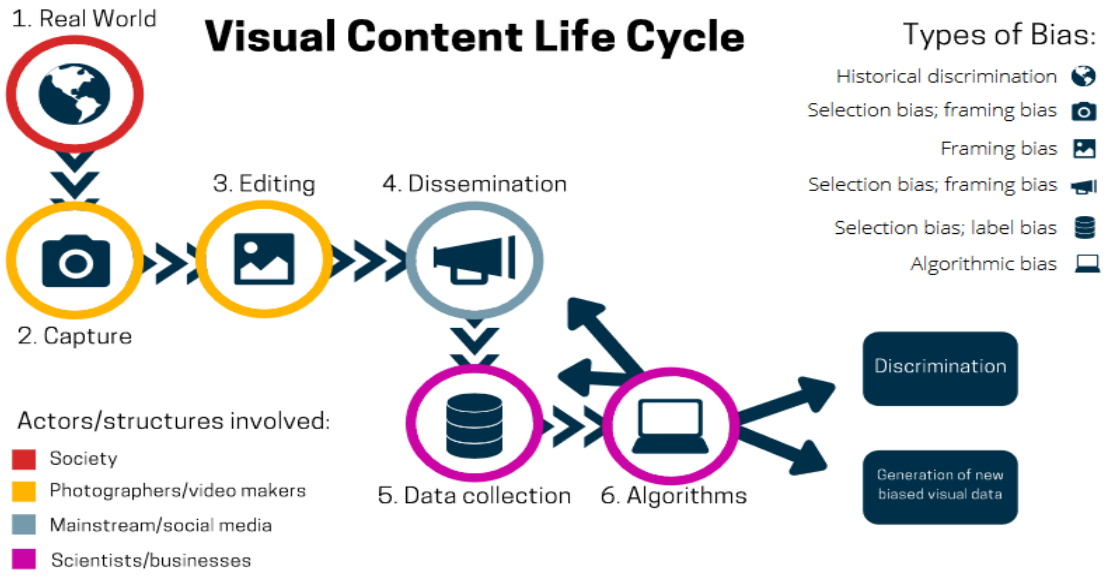

Figure 5: Simplified illustration of visual content life cycle and associated sources of bias, taken from [FPNK22].

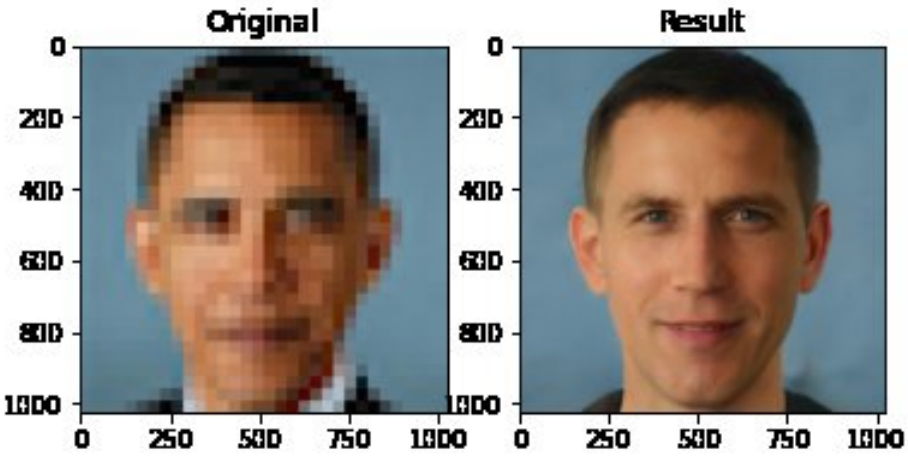

Figure 6: Instance of a non-white face upsampled with Caucasian features. PULSE upscaling result from a low-resolution picture of the american ex-president Barack Obama, taken from Twitter user @Chicken3gg (Image URL).

5.2 Harm

Harm is the potential of the models to generate images that cause physical, psychological, or emotional damage to individuals or groups, or instigate actions such as violence, abuse and hate speech.

Text to image models can also be used in a harmful way to produce fake images that could spread disinformation or support a given narrative for malicious actors such as authoritarian governments or power-hungry political or extremist movements, or cybercriminals.

Disinformation already existed long before any text to image model was developed. 'Cheapfakes' require very low technological skills and, thus, present the lowest barriers to entry. These are media products manipulated with a low level of technical sophistication, such as videos or images with a misleading or false caption. That said, the AI-powered 'deepfakes', in the right circumstances could become an increasingly more popular method of misinformation.

Currently, deepfakes are mostly created for pornography, causing great harm to the victims, who are mostly women. For a more in-depth discussion of those topics, refer to [BG19].

Figure 7: Models generations from Stable Diffusion conditioned on various prompts, taken from [BYMC22].

A report by [BSB22] identified the main conditions that could increase the threat of disinformation made using text-to-image technology.

- **Cost and time efficiency** If the choice of disinformation path will depend on which method is easiest and quickest to create fake media from, then the improvement of text-to-image models in this direction will probably make them the preferred choice for such operations. For example, if it becomes efficient to finetune a generative model on the appearance of a celebrity or political figure, then deepfakes about those specific targets could increase.

- **Access and democratization** If more models become freely usable by anyone, then malicious actors could use them to their benefit. The current trend seem to point in this direction, as models and their weights are developed following open-source or are leaked.

- **Filters and self regulation** Companies such as OpenAI have developed the first models and applied filters to reduce harmful content creation. It is in some cases possible to circumvent those filters (e.g. using the words 'red liquid' instead of 'blood' to generate violent or gory images). Also, access to source code can mean it is possible to remove filtering and content moderation entirely for the media creation process.

  Many social media platforms have already adopted policies to better prepare themselves to deal with manipulated media content and properly address the threat and misuse of synthetic media for disinformation purposes. This type of content should be removed and users posting it banned according to policy by most mainstream social media, which theoretically could help in containing the spread of disinformation.

- **Detection mechanisms and societal resilience** AI-generated content could become more difficult to distinguish from real images. The work of [GCM+21] on GAN-generated images conclude that we are still very far from having reliable tools for automatic fake image detection, especially considering the challenges of real-word scenarions (e.g. image compression, resizing, and new models).

  [BSB22] also worry that a very high stream of fake content could overwhelm detectors and undermine the credibility of journalists and reputable news agencies, or provide plausible deniability to malicious actors (e.g. a politician's scandalous conduct being defined as 'synthetic content'). If people approach social media with less skepticism, those spreading disinformation could influence

public opinion around information sources online, resulting in a reduced level of trust in news on social media. This is arguably true in some social communities already.

The authors therefore argue for the need of standardized regulations for providers of AI services and social media platforms, as well as increasing digital literacy.

Regulation Currently, there are no standard-regarded laws specifically regulating the major aspects of AI application, so there are no limitation in the ways models can be trained or on which data is possible to use. Still, AI system applications need to abide by existing laws such as those regarding data protection in the EU, and those that forbid certain activities such as fraud.

As an example, the existing legal framework (in this case the european GDPR [Eur16]) provided grounds for the italian privacy authority ordering a stop the services of OpenAI's ChatGPT [LHM+23] for italian users until the company complied, some weeks later.

The AI act [Com21], currently a draft of the European Commission, is a proposal for a legislative framework on AI technology. One of its aims is to ban or limit some applications of AI deemed too risky, such as social scoring or remote biometric identification (e.g. real-time facial recognition to match faces against a terrorists database), subliminal manipulation resulting in harm, or exploitation of children. It aims also to ensure more transparency on how AI systems were developed and work through legal requirements.

The work of [VB21] provides a more detailed analysis of the AI Act draft, also raising questions about the regulation of synthetic content.

As for the direction of research, the United States' National Science committee updated its guidelines for AI research in terms of safety, collaboration and standard benchmarks. [scotUS23]

5.3 Privacy, Creativity, and Copyright

Infringement is the violation of the intellectual property rights or personal privacy of the owners or subjects of the images, in the context of text to image models, it matters.

To understand more about the issue, we should start by looking more in depth at one of the most used open source datasets for pretraining text to image models: the LAION 5B [SBV+22] dataset.

For most of the models developed before it, the training data was not made public (only the resulting weight in the case of OpenAI's CLIP models). For example, the data used to train Midjourney has not been made public, but it most likely comes from internet scraping, like LAION's. Therefore, this was a good first step in transparency. Nevertheless, some issues with it have been raised also by the authors themselves, which hope it could be made safer in future research.

LAION-5B consists of over 5.8 billion examples (text-image pairs), which have been obtained from Common Crawl[9], a public web archive, and automatically flag as much pornographic, violent, etc. content as possible (which amounted to 3% of the total data). The dataset contains URLs to images and their caption.

In terms of privacy risks, the authors mention that the dataset does contain personal and/or sensitive information like face pictures and medical images, and hope future work will resolve those issues for example by anonimizing faces while maintaining visual quality with another AI model. After all, LAION-5B is a tool for research, like the authors state:

> LAION-5B is not a finished data product. [...] we do not only release our dataset, but also our software stack we built for assembling LAION-5B. We view our initial data release and this paper as a first step on the way towards a widely applicable pre-training dataset for multimodal models. As a result, we strongly recommend that LAION-5B should only be used for academic research purposes in its current form. We advise against any applications in deployed systems without carefully investigating behavior and possible biases of models trained on LAION-5B.

The problem then is that there are many commercial applications already operating that have most likely not resolved those issues.

[9] link: https://commoncrawl.org/.

Many digital artists have protested how AI companies have used the result of their work for training AI models, without their consent (tracing image ownership for scraped content would be near impossible). Artists also feel threatened by the ability of text-to-image systems to replicate their work and style. For instance, they had a protest against the decision by some art hosting websites of accepting uploads of AI generated media.[10] In general, an increase in the accessibility of art generation, effectively removes the barrier of entry for many kinds of media creation - human artistic skills such as drawing, coloring, shading... and this could bring about great changes in the art field.

Can a picture generated by an AI be considered an art work and thus an original work? At the moment, a generative deep learning output is not protected by copyright if the creation lacks human authorship. For example, for a single image generated from a prompt there is no copyright, but if the authors arranges them into a book in an artistic fashion, then the whole could benefit form authorship rights. For a deeper analysis of the legal framework in the US and EU for generative art issues, see [FM22].

6 Conclusions

In conclusion, the advancements made in Text-to-Image (T2I) technology through state-of-the-art methodologies have showcased promising results. However, it is crucial to acknowledge that there is still significant room for improvement and refinement in this field.

One area that requires attention is the definition of comprehensive evaluation metrics. While existing automatic evaluation metrics provide some insights, they have limitations and fail to capture the full nuances of image generation quality. Therefore, the development of improved automatic evaluation criteria is necessary to ensure accurate and reliable assessments of T2I models.

Furthermore, human evaluation remains a valuable component in assessing the quality of generated images. However, it is essential to establish standard benchmarks and methodologies for human evaluation to enhance consistency and comparability across studies. This would enable researchers and practitioners to better understand the strengths and limitations of different T2I models and facilitate progress in the field.

Additionally, addressing the ethical, legal, and safety implications associated with T2I technology is of paramount importance. Measures need to be taken to mitigate biases, such as social, racial, and gender biases, in the generated images. Steps must also be taken to combat the potential harm caused by the dissemination of fake images, propaganda, offensive content, and violations of intellectual property rights.

Overall, while the current state-of-the-art methodologies for T2I have shown promising results, further advancements are necessary. Continued research and development efforts should focus on refining evaluation metrics, establishing standardized benchmarks for human evaluation, and addressing the ethical implications to ensure responsible and beneficial use of T2I technology. By doing so, we can unlock the full potential of T2I models and foster advancements that benefit various domains and society as a whole.

[10] link: Article on the ArtStation protest.

References

[BG19] Jacquelyn Burkell and Chandell Gosse, *Nothing new here: Emphasizing the social and cultural context of deepfakes*, First Monday **24** (2019), no. 12.

[BSB22] Lena-Maria Böswald, Beatriz Almeida Saab, and Jan Nicola Beyer, *What a pixel can tell: Text-to-image generation and its disinformation potential*, Democracy Reporting International (2022).

[BT95] Vlad Bally and Denis Talay, *The euler scheme for stochastic differential equations: error analysis with malliavin calculus*, Mathematics and Computers in Simulation **38** (1995), no. 1, 35–41.

[BYMC22] Hritik Bansal, Da Yin, Masoud Monajatipoor, and Kai-Wei Chang, *How well can text-to-image generative models understand ethical natural language interventions?*, 2022.

[Com21] European Commission, *Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts (document 52021pc0206)*, COM/2021/206 final (2021).

[DDS+09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, *Imagenet: A large-scale hierarchical image database*, 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.

[DN21] Prafulla Dhariwal and Alex Nichol, *Diffusion models beat gans on image synthesis*, 2021.

[DNH22] Tan M. Dinh, Rang Nguyen, and Binh-Son Hua, *Tise: Bag of metrics for text-to-image synthesis evaluation*, 2022.

[Eur16] European Commission, *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)*, 2016.

[FHR+21] Stanislav Frolov, Tobias Hinz, Federico Raue, Jörn Hees, and Andreas Dengel, *Adversarial text-to-image synthesis: A review*, Neural Networks **144** (2021), 187–209.

[FM22] Giorgio Franceschelli and Mirco Musolesi, *Copyright in generative deep learning*, Data amp; Policy **4** (2022), e17.

[FPNK22] Simone Fabbrizzi, Symeon Papadopoulos, Eirini Ntoutsi, and Ioannis Kompatsiaris, *A survey on bias in visual datasets*, 2022.

[GCM+21] Diego Gragnaniello, Davide Cozzolino, Francesco Marra, Giovanni Poggi, and Luisa Verdoliva, *Are gan generated images easy to detect? a critical analysis of the state-of-the-art*, 2021.

[GPAM+14] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, *Generative adversarial networks*, 2014.

[HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel, *Denoising diffusion probabilistic models*, 2020.

[HLH+18] Huaibo Huang, Zhihang Li, Ran He, Zhenan Sun, and Tieniu Tan, *Introvae: Introspective variational autoencoders for photographic image synthesis*, 2018.

[HRU+18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, *Gans trained by a two time-scale update rule converge to a local nash equilibrium*, 2018.

[Hyv05] Aapo Hyvärinen, *Estimation of non-normalized statistical models by score matching*, Journal of Machine Learning Research **6** (2005), no. 24, 695–709.

[Jar97] C. Jarzynski, *Nonequilibrium equality for free energy differences*, Physical Review Letters **78** (1997), no. 14, 2690–2693.

[JOS+21] Niharika Jain, Alberto Olmo, Sailik Sengupta, Lydia Manikonda, and Subbarao Kambhampati, *Imperfect imaganation: Implications of gans exacerbating biases on facial data augmentation and snapchat selfie lenses*, 2021.

[KLA19] Tero Karras, Samuli Laine, and Timo Aila, *A style-based generator architecture for generative adversarial networks*, 2019.

[KP21] Zhifeng Kong and Wei Ping, *On fast sampling of diffusion probabilistic models*, 2021.

[LHM+23] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, and Bao Ge, *Summary of chatgpt/gpt-4 research and perspective towards the future of large language models*, 2023.

[LL21] Eric Luhman and Troy Luhman, *Knowledge distillation in iterative generative models for improved sampling speed*, 2021.

[LMB+14] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick, *Microsoft COCO: common objects in context*, CoRR **abs/1405.0312** (2014).

[MDH+20] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin, *Pulse: Self-supervised photo upsampling via latent space exploration of generative models*, 2020.

[MO14] Mehdi Mirza and Simon Osindero, *Conditional generative adversarial nets*, 2014.

[Nea98] Radford M. Neal, *Annealed importance sampling*, 1998.

[RASL16] Scott Reed, Zeynep Akata, Bernt Schiele, and Honglak Lee, *Learning deep representations of fine-grained visual descriptions*, 2016.

[RBL+22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, *High-resolution image synthesis with latent diffusion models*, 2022.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, *U-net: Convolutional networks for biomedical image segmentation*, 2015.

[SBV+22] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev, *Laion-5b: An open large-scale dataset for training next generation image-text models*, 2022.

[scotUS23] National science council of the United States, *National ai research and development strategic plan*, 2023.

[SCS+22] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi, *Photorealistic text-to-image diffusion models with deep language understanding*, 2022.

[SDWMG15] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli, *Deep unsupervised learning using nonequilibrium thermodynamics.*

[SGZ+16] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen, *Improved techniques for training gans*, Advances in Neural Information Processing Systems (D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, eds.), vol. 29, Curran Associates, Inc., 2016.

[UAP22] Anwaar Ulhaq, Naveed Akhtar, and Ganna Pogrebna, *Efficient diffusion models for vision: A survey*, 2022.

[VB21] Michael Veale and Frederik Zuiderveen Borgesius, *Demystifying the draft EU artificial intelligence act — analysing the good, the bad, and the unclear elements of the proposed approach*, Computer Law Review International **22** (2021), no. 4, 97–112.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, *Attention is all you need*, 2017.

[WLZS21] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan, *Towards real-world blind face restoration with generative facial prior*, 2021.

[XZH+17] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He, *Attngan: Fine-grained text to image generation with attentional generative adversarial networks*, 2017.

[YZS+23] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang, *Diffusion models: A comprehensive survey of methods and applications*.

[ZC23] Qinsheng Zhang and Yongxin Chen, *Fast sampling of diffusion models with exponential integrator*, 2023.

[ZKK+21] Jieyu Zhao, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Kai-Wei Chang, *Ethical-advice taker: Do language models understand natural language interventions?*, 2021.

[ZZZK23] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon, *Text-to-image diffusion models in generative ai: A survey*, 2023.

Appendices

A Tables

Table A.3
Results on the COCO dataset, as reported in the corresponding reference. Rows marked with† indicate updated results in its open-source code.

Model	IS ↑	FID ↓	R-Prec. ↑
Real Images (Hinz et al., 2020)	34.88	6.09	68.58
GAN-INT-CLS (Reed, Akata, Yan et al., 2016)	7.88	60.62	–
StackGAN (Zhang et al., 2016)	8.45	74.05	–
StackGAN (Zhang et al., 2016)†	10.62		–
StackGAN++ (Zhang et al., 2017)	8.30	81.59	–
ChatPainter (Sharma et al., 2018b)	9.74	–	–
HDGAN (Zhang et al., 2018b)	11.86	–	–
HfGAN (Huang et al., 2019a)	27.53	–	–
Text2Scene (Tan et al., 2018)	24.77	–	–
AttnGAN (Xu et al., 2017)	25.89	–	85.47
MirrorGAN (Qiao et al., 2019)	26.47	–	74.52
Huang et al. (2019b)	26.92	34.52	89.69
AttnGAN+OP (Hinz et al., 2019)	24.76	33.35	82.44
OP-GAN (Hinz et al., 2020)	27.88	24.70	89.01
SEGAN (Tan et al., 2019)	27.86	32.28	–
ControlGAN (Li, Qi et al., 2019)	24.06	–	82.43
DM-GAN (Zhu et al., 2019)	30.49	32.64	88.56
DM-GAN (Zhu et al., 2019)†	32.43	24.24	92.23
Hong et al. (2018)	11.46	–	–
Obj-GAN (Li et al., 2019)	27.37	25.64	91.05
Obj-GAN (Li et al., 2019)†	27.32	24.70	91.91
SD-GAN (Yin et al., 2019)	35.69	–	–
textStyleGAN (Stap et al., 2020)	33.00	–	87.02
AGAN-CL (Wang, Lang, Liang, Lyu et al., 2020)	29.87	–	79.57
TVBi-GAN (Wang, Quan et al., 2020)	31.01	31.97	–
RiFeGAN (Cheng et al., 2020)	31.70	–	–
Wang, Lang, Liang, Feng et al. (2020)	29.03	16.28	82.70
Yuan and Peng (2019)	16.40	–	–
Rombach et al. (2020)	34.7	30.63	–
CPGAN (Liang, Pei, & Lu, 2020)	**52.73**	–	**93.59**
Pavllo et al. (2020)	–	19.65	–
XMC-GAN (Zhang et al., 2021)	30.45	**9.33**	–

(a)

TABLE 1
FID of representative methods on MS-COCO dataset

model	FID
CogView [12]	27.10
LAFITE [60]	26.94
DALLE [11]	17.89
GLIDE [15]	12.24
Imagen [16]	7.27
Stable Diffusion [17]	12.63
VQ-Diffussion [43]	13.86
DALL-E 2 [18]	10.39
Upainting [63]	8.34
ERNIE-ViLG 2.0 [64]	6.75
eDiff-I [65]	6.95

(b)

Figure 8: Comparison of results: (a) results of GAN-based models on the COCO dataset, as summarized in [FHR+21], (b) results of diffusion-based models on the COCO dataset, as summarized in [ZZZK23].

B Text-to-image webui images

In the following we present a couple of example screenshots of the webui mentioned in Section 4.2. It makes generating images, tuning the conditioning parameters and saving the outputs in an organized way very easy compared to working on the code directly.

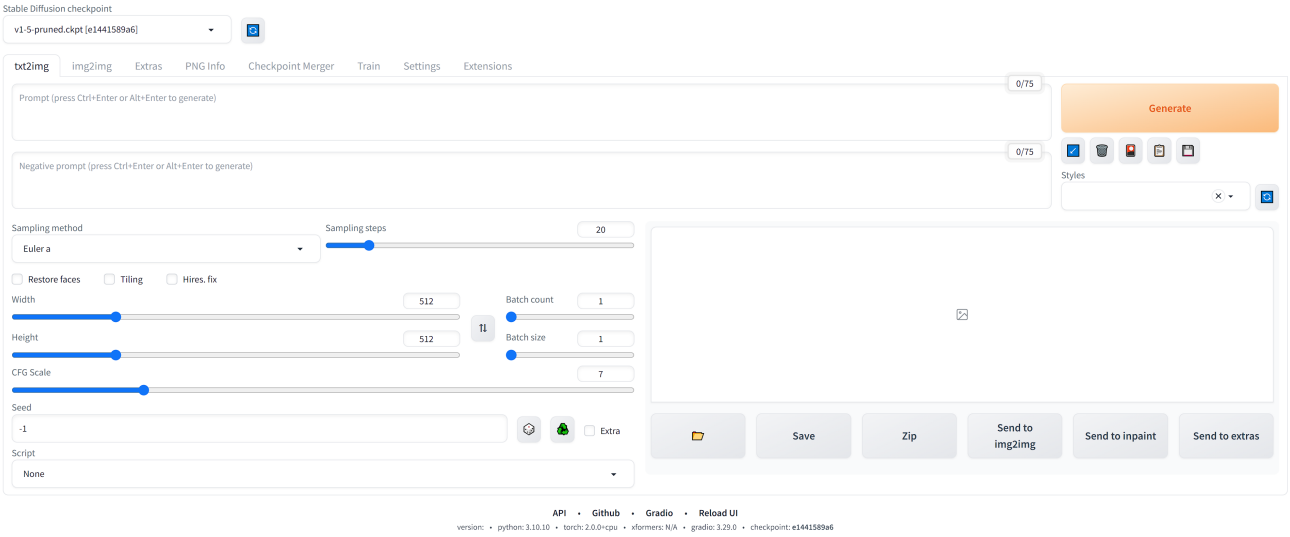

Figure 9: The main page of the webui interface.

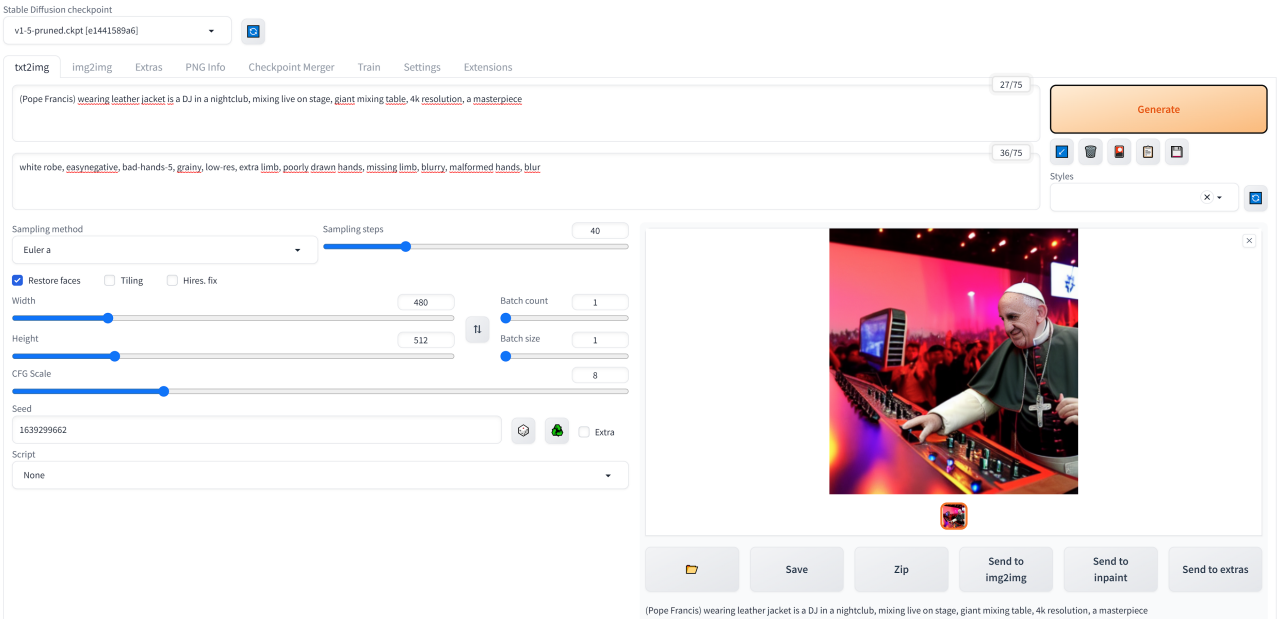

Figure 10: The final result of generating an image with a textual conditioning.