From Gans to Diffusion Models: Text-To-Image Generation

Yian Xiao *

Wuhan Britain-China School, Wuhan, 430000, China

* Corresponding Author Email: aereasynx2008@gmail.com

Abstract. This paper traces the evolution of text-to-image generation (TIG) techniques from Generative Adversarial Networks (GANs) to Diffusion Models (DMs). It first introduces GAN variants, including DCGAN, WGAN, MGGAN, and StyleGAN. While these popular GANs pioneered image synthesis through adversarial training of the generator and discriminator, they suffered from training instability, mode collapse, and the lack of diversity. Therefore, the systematic introduction of representative DMs—including DDPM, Guided Diffusion, GLIDE, Stable Diffusion, and Imagen—shows how they address these issues through iteratively denoising, achieving unprecedented image fidelity, semantic alignment, and generation stability. Quantitative comparisons on datasets such as COCO and CUB show that DMs consistently outperform GANs in metrics like FID, IS, and CLIP score, though GANs retain shorter inference time. Nevertheless, critical challenges such as generation efficiency, understanding of complex prompts, and safety controls remain. This paper analyses possible reasons for those problems while pointing out key directions for future work.

Keywords: GANs; Diffusion Models; text-to-image generation.

1. Introduction

Text-to-image generation (TIG) has advanced rapidly, driven by the growing demand for high-quality visual content in art, design, education, and media. Yet, generating realistic and diverse images from textual descriptions remains a challenging task. Early models often struggle with unstable training, low image fidelity, and limited diversity. In real-world applications, these issues result in inconsistent visual results, poor alignment with text, and reduced usability. For domains such as advertising or scientific visualisation, these limitations can undermine trust and practical value.

In 2014, a fresh approach to image synthesis, Generative Adversarial Network (GAN), was introduced. The image synthesis problem was formulated as a two-player game between two competing neural networks: a generator and a discriminator. The generator network is responsible for generating realistic images, while the discriminator network is trained to distinguish between authentic and synthesised photos [1]. However, due to the adversarial learning process, they're unstable and would produce inconsistent results while training. Consequently, they would easily collapse. The instability of GANs leads to two significant problems: the generation of low-quality images and a lack of image diversity [2]. Although they're widely used, these current issues limit their application in real-world scenarios.

Diffusion Models (DMs) were introduced in 2015 to cope with some of these problems. The Diffusion Model (DM) generates images by learning to reverse a gradual noise process [3]. During training, noise is progressively added to the data; during generation, the model learns to iteratively denoise a random noise input into a structured image guided by the text prompt. This denoising approach stabilises the training process while producing highly detailed and semantically aligned images. The iterative refinement process in DM enables them to prevent model collapse while generating diverse and high-quality photos [4].

Since their introduction by Goodfellow, GANs have been widely adopted in numerous application domains. Early use cases included image generation and style transfer, which soon expanded to high-resolution image and video synthesis. More recent studies have highlighted the role of GANs in scientific domains, such as medical imaging tasks, including the synthesis of MRI and PET images [5].

Since the original formulation of DMs, the technique has undergone rapid progress. The introduction of DDPMs enabled stable training and high-fidelity image generation. Further

improvements enhanced sampling efficiency and visual quality. In 2022, Imagen demonstrated photo-realistic text-to-image generation by integrating large language models, while Stable Diffusion (SD) made high-resolution synthesis more scalable and accessible.

This paper aims to trace the evolution of TIG from GANs to DMs, elaborating how diffusion techniques have addressed key limitations of GANs such as model collapse and training instability. It then compares representative models on benchmarks using metrics such as Fréchet Inception Distance, Inception Score, and CLIP Score. Finally, it discusses ongoing challenges in efficiency, compositional understanding, and bias control, providing insights into future research directions. The goal is to clarify why DMs have become dominant in TIG and to outline paths for further improvement.

2. GANs in TIG

2.1. Principles of GAN

GANs were first introduced by Goodfellow, a class of generative models that learn to synthesise data by framing the generation process as a two-player minimax game of two competing neural networks: a generator G and a discriminator D. The generator network is responsible for generating realistic images. The discriminator network is trained to differentiate between authentic and synthesised photos [1]. The generator takes in random four-ounce $z \sim p_z(z)$ and attempts to produce realistic samples $G(z)$. At the same time, the discriminator tries to distinguish between real dafoura $x \sim p_{data}(x)$ and generated data $G(z)$.

The training objective can be defined as follows:

$$\min_G \max_D V(D,G) = E_{x\sim p_{dt}(x)}[\log D(x)] + E_{z\sim p_z(z)}\left[\log\left(1 - D\big(G(z)\big)\right)\right] \quad (1)$$

The discriminator attempts to maximise its accuracy of classification, while the generator tries to "deceive" the discriminator from differentiating authentic images and synthesised images correctly. Ideally, training converges to a Nash Equilibrium to achieve this.

Although the conception is ideal, the training of GANS faces numerous challenges in practice. The training process is susceptible to the balance between two networks, while the convergence is not guaranteed. Furthermore, since the generator receives gradients indirectly through the discriminator, instability, oscillation, or even divergence issues can easily arise during the learning process. Later improvements, such as WGAN and WGAN-GP, attempt to address this [6,7].

2.2. Main Variants and Algorithm Improvements

Since the invention of GANs, numerous variants have been suggested by researchers to solve the inherent limitations. For example, training instability, mode collapse and poor convergence. These improvements contribute to the growth of TIG in various aspects, such as network structure and loss function.

One of the earliest and most influential variants was DCGAN (Deep Convolutional GAN), proposed [8]. This model incorporated convolutional and transposed convolutional layers into the GAN architecture, enabling more stable training and improved image quality. For subsequent model developments, DCGAN has laid a solid foundation.

To address the instability problem, Wasserstein GAN (WGAN) replaced the Jensen-Shannon divergence with the Wasserstein distance [6]. This offers smoother gradients and improves the model's trainability. WGAN-GP made further improvements by introducing the Gradient Penalty to overcome the limitations of weight clipping and significantly enhance training stability [7].

To address mode collapse, namely the lack of diversity in the generator's outputs, MGGAN introduced the Manifold Guidance Network [2]. These approaches guide the generator to explore broader regions of the data manifold, thereby enhancing image diversity.

In terms of detail control, StyleGAN proposed a generator architecture based on image styles, which enables hierarchical control during generation [9]. Thus, images with higher resolution and stronger visual consistency can be synthesised.

These evolution traces indicate that the development of GAN has seen significant progress in training methods, stability, diversity, and image quality, laying a solid foundation for modern generative modelling.

2.3. Remaining Limitations for GANs

While GANs have made remarkable progress, several core limitations remain unresolved. These issues not only hinder further exploration but also pose challenges to real-world applications.

The training process of GAN is susceptible to the model architecture, hyperparameters and initialisation. A slight deviation would lead to subsequent problems. Even with methods like WGAN-GP, the outcome still depends on a delicate balance between the generator and discriminator [7].

Mode collapse is a common problem in both conditional and unconditional generation tasks. Although in MGGAN, Bang & Shim introduced a guidance network to induce the generator to learn the accurate distribution without missing modes, quality degradation (e.g., image blurs) still exists in their experiment [2].

Lack of likelihood estimation: GANs do not provide explicit likelihood or density functions, making quantitative evaluation through log‑likelihood impossible [10]. This limitation hinders their utility in statistical inference and assessment.

These limitations have motivated a growing interest among researchers to explore alternative models, such as DMs, which offer more stable optimisation and richer sample diversity.

3. DMs in TIG

DM generate images by learning to reverse a gradual noising process. During training, noise is progressively added to the data; during generation, the model learns to iteratively denoise a random noise input into a structured image guided by the text prompt. This approach is based on a Markov chain. The text prompt guides the denoising process to ensure the alignment between itself and the synthesised images. The operation of DMs involves a sequence of steps that improve stability, image quality, and semantic alignment [2].

3.1. DDPM

3.1.1 Principles and Algorithm

A Denoising Diffusion Probabilistic Model (DDPM) is a parameterised Markov chain that uses variational inference to train, producing data-matched samples in a limited amount of time. The transitions of this chain are trained to invert a diffusion process. This Markov chain incrementally corrupts the data by adding noise in the reverse direction of sampling until the original signal is entirely degraded. In cases where the diffusion involves small-scale Gaussian noise, the sampling chain transitions can likewise be modelled as conditional Gaussians. This approach enables an exceptionally straightforward parameterisation of neural networks [11].

DDPM consists of two core components: a forward diffusion process that gradually adds Gaussian noise to corrupt the data, and a reverse denoising process that recovers the data from the noise.

Forward process: Given a data sample $x_0 \sim q(x_0)$, the forward process generates a sequence $x_1, x_2, \ldots, x_T$ by progressively adding Gaussian noise:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t I\right) \tag{2}$$

A closed-form expression for $q(x_t \mid x_0)$ is derived as:

$$q(x_t \mid x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}\, x_0, (1-\bar{\alpha}_t) I\right) \tag{3}$$

With $\bar{\alpha}t = \prod s = 1^{t}(1 - \beta_s)$.

Reverse process: The reverse process is modeled as another Markov chain:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)\right) \tag{4}$$

Instead of learning both mean and variance, the authors train a neural network to predict the noise ϵ added at each step:

$$\epsilon_\theta(x_t, t) \approx \epsilon \quad \text{(true noise)} \tag{5}$$

Training minimises the simplified loss:

$$L_{\text{simple}} = Ex_0, \epsilon, t\left[|\epsilon - \epsilon\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\,\epsilon, t\right)|^2\right] \tag{6}$$

This objective corresponds to denoising score matching under a reweighted variational lower bound, enabling stable training.

3.1.2 Innovation and Improvement

The DDPM, proposed by Ho, addresses several limitations presented in GAN models, based on a probabilistic model, while employing a non-adversarial training approach [11].

Unlike GANs, which rely on a delicate adversarial balance between generator and discriminator, DDPM replaces this with a simple mean squared error objective to predict the added noise during the forward process. This loss formulation is practical to implement and stable to optimise, hence effectively avoiding the oscillation or collapse that is commonly seen in GANs [11].

GANs have one significant weakness - they can't clearly describe how data is distributed. On the other hand, DDPMs learn by optimising a variational lower bound (ELBO), which lets them measure probabilities directly. The second section of the paper, written by Ho, explains how this training works and provides additional mathematical details in Appendix B [11]. Because every step in DDPMs uses simple Gaussian distributions, these models are easier to understand and track statistically than GANs.

GANs often suffer from mode collapse, where their outputs become too similar and exhibit only a limited number of variations. DDPMs work differently. They start from random Gaussian noise and slowly remove noise to create images. This helps them capture all types of variations in the data.

Without relying on an adversarial training process, DDPM generates images of higher quality and provides better stability, as well as a theoretical basis. It lays a foundation for further development of DMs.

3.2. Guided diffusion

Following the success of DDPM, Dhariwal and Nichol proposed Guided Diffusion (GD), a significant advancement aimed at improving both the fidelity and controllability of diffusion-based TIG [12]. Unlike DDPM, GD introduced classifier guidance, which enables the accurate generation of samples with semantic class labels.

The core concept is using gradients from a pre-trained classifier to guide the reverse noise process. Specifically, the generated sample will be directed toward a more semantically matched direction based on the classifier's calculation of the current image. This mechanism dramatically strengthens the coherence of the outputs and conditions without retraining the DM itself.

According to the experiment, the model achieves state-of-the-art FID scores on datasets such as ImageNet, outperforming GANs, including BigGAN. GANs regularly encounter instability and model collapse, especially in high-resolution conditional generation tasks. By contrast, GD exhibits a stable training process and consistently produces diverse, semantically aligned images.

By preserving DDPM's likelihood estimation, GD effectively addresses the long-term challenges of controllability and generation quality faced by GANs. It also marks the beginning of a new phase in the development of DMs.

3.3. GLIDE

GLIDE, a diffusion-based model designed explicitly for TIG, was introduced by OpenAI [13]. As one of the earliest attempts to incorporate natural language directly into DMs, GLIDE marked a key shift from class-conditional image generation to multi-modal generation based on text prompts.

Rather than relying on class labels or classifiers, GLIDE utilised learned text embeddings generated by a text encoder (e.g., Transformer or CLIP) as conditional inputs to the DMs. Furthermore, it also employs a classifier-free guidance, which interpolates between conditional and unconditional predictions during sampling. Therefore, there's no requirement for an external classifier to control generated content. This approach not only simplifies the training process but also enhances it. And improves the ability to express complicated semantics.

In numerous TIG tasks, GLIDE demonstrates outstanding synthesis quality, outperforming previous models in image fidelity and TI alignment. However, it still fails to capture prompts sometimes when it comes to highly unusual and complex objects or scenarios. GLIDE is considered less favourable for use in real-time applications compared to previous models due to its slow sampling time.

Although GLIDE is not as widely used as later models, it established significant concepts and laid a foundation for combining language understanding and diffusion-based synthesis.

3.4. Stable Diffusion

3.4.1 Principles and algorithm

Stable Diffusion (SD) is based on the Latent Diffusion Model (LDM) framework introduced by Rombach [14]. Instead of adding noise to pixel-level images, the model compresses image data into a lower-dimensional latent space through a pretrained Variational Autoencoder (VAE). The denoising process is operated by a U-Net denoiser conditioned on text embeddings via a cross-attention mechanism (typically from a CLIP text encoder).

LDM first use pretrained VAE encoder to map an image x to a lower-dimensional latent vector $\mathrm{z} = \mathcal{E}(\mathrm{x})$, and reconstructs it using a decoder $\mathcal{D}(\cdot)$. It leverages the image-specific inductive biases inherent in model architecture. This is achieved by constructing the core U-Net primarily with 2D convolutional layers, while using the reweighted bound to prioritise the perceptually most significant elements in the objective function.

$$\mathcal{L}\mathrm{LDM} \coloneqq \mathrm{E}\mathcal{E}(\mathrm{x}),\ \epsilon \backslash \mathrm{sim}\ \mathcal{N}(0,\ 1),\ \mathrm{t}\left[\ |\ \epsilon\ -\ \epsilon_{\theta}(\mathrm{z_t},\ \mathrm{t})\ |_2^2\ \right] \tag{7}$$

$\mathrm{z_t}$ can be obtained from encoder $\mathcal{E}$ during training due to the fixed forward process. Samples from $\mathrm{p}(\mathrm{z})$ can be decoded to image space through $\mathcal{D}$.

3.4.2 Advantages and Improvements

The proposed method achieves competitive performance across multiple image generation tasks, including unconditional synthesis, inpainting, and stochastic super-resolution, while substantially reducing computational requirements. Compared to pixel-based diffusion approaches, this framework demonstrates significantly lower inference costs without compromising output quality.

LDM- {4-16} achieves a good balance between efficiency and perceptually faithful results. LDM-8's 38-point FID advantage shows this over pixel-based diffusion after training. LDM- {4-8} outperform models with poor perceptual-conceptual compression ratios. Compared to pixel-based LDM-1, they achieve much lower FID scores while substantially increasing sample throughput. For complex datasets like ImageNet, reduced compression rates are needed to maintain quality. Overall, LDM-4 and -8 provide the best conditions for high-quality synthesis.

Likelihood-based models offer key advantages over other approaches. First, they do not suffer from the mode-collapse and training instabilities commonly seen in GANs. Second, thanks to heavy reliance on parameter sharing, they can model the intricate distributions of natural images very efficiently. This stands in contrast to massive AR models, which often require billions of parameters to achieve similar complexity.

Introducing cross-attention-based conditioning into LDMs allows for the use of various conditioning modalities that were previously unexplored in DMs. This combination of domain-specific experts in learning language representation and visual synthesis creates a strong model that generalises well to complex, user-defined text prompts. Applying classifier-free diffusion guidance significantly improves sample quality while also reducing the number of required parameters.

3.5. Imagen

3.5.1 Principles and algorithm

Imagen is a TIG DM that combines large language models with cascaded diffusion to achieve unprecedented photorealism and deep language learning. It proposed a cascaded LDM architecture, where DMs upscale from low to high resolution in multiple stages: starting from a 64×64 resolution, it progressively upscales outputs through intermediate 256×256 an d final 1024×1024 stages. Separate DMs and conditions are applied to each stage, operating independently, and are used in the production of the previous stage [15].

Imagen contains a large pretrained language model, frozen T5-XXL, as the text encoder to map input text into a sequence of embeddings. This language model can extract rich semantic information to provide a more precise text prompt for TIG. Furthermore, Imagen introduced classifier-free guidance during sampling to enhance the semantic alignment between the image and the text description [15].

Classifier-free guidance is an alternative technique of classifier guidance. It trains a single DM on conditional and unconditional objectives through randomly dropping c. The guidance sampling process can be expressed as:

$$\epsilon\theta^{\text{guided}} = (1 + w) \cdot \epsilon\theta(xt, c) - w \cdot \epsilon\theta(x_t) \quad (8)$$

To increase semantic alignment.

3.5.2 Advantages and Improvements

Imagen emphasises the significance of frozen large pretrained language models as text encoders—the advantages of freezing, such as offline computation of embedding results, result in negligible computation during the training process. In addition, human evaluators explore the advanced performance of frozen T5-XXL in both TI alignment and image fidelity [15].

The utility of two text-conditional super-resolution DMs and a multi-step diffusion pipeline allows Imagen to generate highly detailed and high-resolution images. These cascaded DMs with noise conditioning augmentation have proven very useful in generating high-fidelity images in a progressive manner. Each stage helps to preserve image structure and fine textures [15].

The integration of classifier-free guidance in all stages of the pipeline enables more consistent and expressive TI alignment, while avoiding the bias and complexity introduced by relying on auxiliary classifiers [15].

Overall, Imagen makes significant progress in diffusion models by proposing three key strengths: powerful language models, multi-stage resolution control, and efficient sampling. It creates a flexible system that deeply understands text while producing high-quality images.

4. Data Analysis

4.1. Datasets Introduction

To evaluate the performance of the selected TIG models, people focus on three representative datasets that contain various scenes, objects and relationships: COCO, CUB and LAION. By combining these datasets, the experiment enables a balanced evaluation across both GAN-based and diffusion-based models.

4.2. Algorithm Performance Comparison

Table 1. Quantitative and efficiency comparison of GAN-based and diffusion-based models on COCO, CUB-200, and LAION datasets. Metrics include Fréchet Inception Distance (FID ↓), Inception Score (IS ↑), CLIP Score (↑), Resolution, and Inference Time.

Model	Dataset	FID ↓	IS ↑	CLIP ↑	Resolution	Inference Time*
DCGAN	COCO	68.1	6.2	–	64×64	~0.05s
WGAN	COCO	55.2	7.0	–	64×64	~0.05s
MGGAN	CUB	18.2	4.1	–	256×256	~0.2s
StyleGAN	COCO	18.3	9.1	–	1024×1024	~0.15s
DDPM	CUB	7.9	15.0	0.255	256×256	~20s
Guided Diffusion	COCO	7.3	28.0	0.270	256×256	~25s
GLIDE	COCO	12.2	–	0.290	256×256	~4.0s
SD	LAION-5B	7.27	–	0.297	512×512	~0.5s
Imagen	COCO	7.27	38.0	0.330	1024×1024	~5.0s

*Inference Time is approximate and based on reported values for a single NVIDIA V100 GPU.

In Table 1, DMs consistently outperform GANs on critical image quality (FID, IS) and TI alignment (CLIP Score), while GANs remain superior in inference speed, generating samples faster than most DMs. However, recent DMs, such as SD, GLIDE, and Imagen, demonstrate noticeable improvements in inference time compared to previous DMs, including DDPM and guided diffusion, thereby narrowing the efficiency gap between GANs and DMs while maintaining high quality. SD also shows a balance of inference time (~0.5s) and performance (low FID, high CLIP) at 512×512 resolution.

It's worth mentioning that Imagen is the top performer in terms of raw image quality (IS=38.0) and TI alignment (CLIP=0.330) on COCO. Furthermore, it maintains these features at high resolution (1024 × 1024).

5. Conclusion

Despite significant progress in TIG, critical limitations remain unresolved, even under current architectures. These gaps represent core challenges and future directions.

DMs (e.g., DDPM, Imagen) rely on iterative sampling, which results in prohibitive delays. Imagen requires ~5.0 seconds) and computational inefficiency. Current techniques, such as distillation, achieve acceleration by sacrificing image quality and are unable to obtain both real-time generation and image fidelity. Future directions could focus on non-Markovian sampling algorithms, as the sequential nature of Markov-chain denoising opposes parallelisation, as well as closed-form approximations of diffusion trajectories.

One of the other critical problems is that state-of-the-art models consistently fail to interpret prompts involving multiple objects, complicated attributes and spatial relationships. This results from the lack of explicit scene-graph parsing of text encoders and the omission of compositional logic. To enhance understanding of complex prompts and scenes, techniques such as Neuro-symbolic architectures that integrate scene graphs and generative feedback loops with large language models can be considered.

The amplification of biases and inadequate safety controls is also a problem with TIG. Studies demonstrate that these models often reinforce stereotypes. For instance, over-representing particular animal species when generating images of “pets”. Furthermore, current safety protection measures have a limited effect. Although the system will filter out content from apparent violations, some violating pictures can still slip through the detection. Future improvements can prioritise the development of bias-aware training objectives, enabling automatic identification and balance

between different types of content. Additionally, upgrading the detection system by conducting multiple rounds of security checks during the synthesis process, rather than after the synthesis, can improve the detection accuracy.

In conclusion, while DMs have surpassed GANs in synthesising high-fidelity and semantically aligned images, critical challenges remain. The iterative denoising process of DMs incurs high computational costs and lengthy inference times, which restrict their use in real-time applications. Furthermore, models still struggle to comprehend complex prompts and scenes, as well as the social biases inherent in the training data. Future work should focus on developing efficient sampling algorithms, integrating advanced reasoning models for better prompt comprehension, and refining safety mechanisms throughout the training and generation pipeline. Addressing these issues is crucial for developing trustworthy and user-aligned TIG systems.

References

[1] Iqbal T, Qureshi S. The survey: Text generation models in deep learning. Journal of King Saud University-Computer and Information Sciences, 2022, 34(6): 2515–2528.

[2] Bang D, Shim H. MGGAN: Solving mode collapse using manifold guided training. arXiv preprint arXiv:1804.04391, 2018.

[3] Sohl-Dickstein J, Weiss E, Maheswaranathan N, Ganguli S. Deep unsupervised learning using nonequilibrium thermodynamics. arXiv preprint arXiv:1503.03585, 2015.

[4] Zhou J. Text-to-image generation: A literature review focusing on the diffusion model. ITM Web of Conferences, 2025, 73: 02037.

[5] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. Advances in Neural Information Processing Systems, 2014, 27.

[6] Arjovsky M, Chintala S, Bottou L. Wasserstein GAN. arXiv preprint arXiv:1701.07875, 2017.

[7] Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of Wasserstein GANs. arXiv preprint arXiv:1704.00028, 2017.

[8] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434, 2015.

[9] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 4401–4410.

[10] Dieng A B, Ruiz F J R, Blei D M, Titsias M K. Prescribed GAN (PresGAN): Enabling explicit density estimation. arXiv preprint arXiv:1910.04302, 2019.

[11] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. arXiv preprint arXiv:2006.11239, 2020.

[12] Dhariwal P, Nichol A. Diffusion models beat GANs on image synthesis. arXiv preprint arXiv:2105.05233, 2021.

[13] Nichol A, Dhariwal P, Ramesh A, et al. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741, 2021.

[14] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022: 10684–10695.

[15] Saharia C, Chan W, Saxena S, et al. Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems, 2022.