JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A Comprehensive Review on Generative AI- Text To Image Generator

Ritika Shaw
Department of CSE
Chandigarh University
Gharuan, (Mohali)
21BCS8073@cuchd.in

Gaurav Kashyap
Department of CSE
Chandigarh University
Gharuan, (Mohali)
21BCS8170@cuchd.in

Sahil
Department of CSE
Chandigarh University
Gharuan, (Mohali)
20BCS5133@cuchd.in

Bhavesh Dwivedi
Department of CSE
Chandigarh University
Gharuan, (Mohali)
20BCS5162@cuchd.in

Akhil Khandelwal
Department of CSE
Chandigarh University
Gharuan, (Mohali)
20BCS5058@cuchd.in

Purnima Sharma
Department of CSE
Chandigarh University
Gharuan, (Mohali)
purnima.r210@cumail.in

***Abstract*—This article provides an overview of Generative AI Text-to-Image Generation, a cutting-edge technology that leverages artificial intelligence to create visual content based on textual input. This innovation eliminates the need for traditional image creation methods. In 2010, the concept of Generative Adversarial Networks (GANs) was introduced. This innovative approach enables the generation of images, illustrations, and graphics by interpreting and translating text prompts into visual representations, eliminating the need for traditional image creation methods. In this article, we delve into the history of this field, summarize its core functionalities, explore potential solutions to enhance accuracy and quality, comprehend the underlying mechanisms, and discuss potential research gaps to further advance Generative AI Text-to-Image Generation.**

Keywords—Attentional Generative Network, EntityDrawBench, FID score, IS score, CLIP models, Latent Space, DrawBench Benchmark, R-precison(%), Guided Diffusion, semantic consistency, T5-XXL encoder, Diffusion Dequantization-Inpainting Model, COCO dataset, CUB dataset, STEM, GAN, Stable Diffusion.

I. INTRODUCTION

Generative AI Image-to-Text Generation is an innovative technology that harnesses the power of artificial intelligence to convert visual content into textual descriptions. This transformative approach eliminates the need for manual image captioning and leverages deep learning techniques to interpret images and generate accurate textual representations. As the demand for efficient image analysis and interpretation continues to grow, this paper provides an extensive overview and performance evaluation of Generative AI Image-to-Text Generation systems. It delves into the core principles, technical challenges, underlying mechanisms, and real-world applications of this technology.

The ever-increasing need for effective image understanding and content indexing has prompted a surge in research and development efforts to enhance visual data processing. Often referred to as AI-based image captioning, Generative AI Image-to-Text Generation utilizes neural networks to analyze images and create descriptive text, offering the potential for precise image interpretation and reduced human intervention. The inherent benefits of this technology, such as its ability to handle diverse visual content and streamline content management, have spurred significant interest from both academics and industry.

Nonetheless, the deployment of Generative AI Image-to-Text systems is not without its challenges. Issues such as image recognition accuracy, handling complex scenes, and ethical considerations, including potential biases, need to be addressed. Overcoming these challenges necessitates innovative solutions in neural network architectures, training data diversity, and ethical AI design. Consequently, gaining a comprehensive understanding of the principles and strategies underpinning Generative AI Image-to-Text Generation is crucial for optimizing its performance and realizing its full potential.

This review article aims to provide a holistic perspective on Generative AI Image-to-Text Generation systems. By examining the fundamental components of these systems, investigating the technical hurdles they face, exploring neural network architectures and training approaches for improved performance.

II. LITERATURE REVIEW

The 2016 article “Generative Adversarial Text to Image Synthesis” [1] by Scott Reed describes how a combination of GAN and encoding process enables the generation of images from prompts.

GAN architecture has 2 neural network components, called generator network and discriminator network, where the generator network is responsible for image generation and the discriminator network is responsible for distinguishing fake images. Both work in a loop, where the image output is improved by providing the discriminator with a more realistic dummy image, where it is expected that the discriminator will not recognize it.

The key contributions and findings of this work include:

Disentangling Style and Content: The authors show that the model can disentangle style and content in image generation. Noise vectors help maintain image contrast while not disturbing image content.

Interpolation Regularization: They introduced a variety of interpolation regularization functions that improve image quality by creating more visually appealing images.

Generalizability: The paper demonstrates the generalization ability of their approach on datasets such as the Oxford-102 Flowers [10] and the MS-COCO dataset [11], showing that the model can handle images with multiple objects and variable backgrounds.

Style Transfer: The study demonstrates a novel technique where the network is capable of conducting style transfer from query images onto textual descriptions. This innovative approach enables the transformation of a textual description into images that embody the pose and background of the corresponding query image, effectively bridging the visual and textual domains.

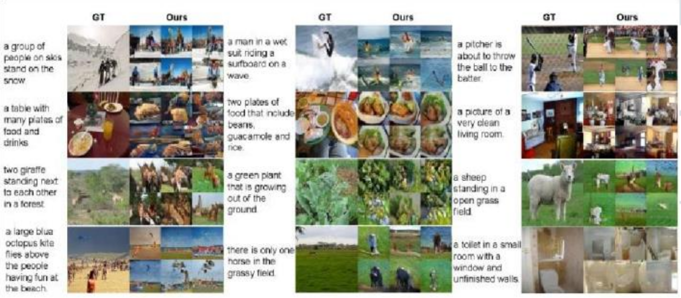

GT
Ours
a group of people on skis stand on the snow
a table with many plates of food and drinks
two giraffe standing next to each other in a forest
a large blue octopus kite flies above the people having fun at the beach
GT
Ours
a man in a wet suit riding a surfboard on a wave
two plates of food that include beans, guacamole and rice
a green plant that is growing out of the ground
there is only one horse in the grassy field
GT
Ours
a pitcher is about to throw the ball to the batter
a picture of a very clean living room
a sheep standing in a open grass field
a toilet in a small room with a window and unfinished walls

Figure 1. Image generation using GAN-CLS

The combination of GANs with text representations shows promise for various applications, including computer vision and creative content generation.

In the year 2017, Tao Xu authored an IEEE paper that introduced an innovative concept known as the Attentional Generative Adversarial Network (AttnGAN) [2]. This approach represents a significant advancement in the field of fine-grained text-to-image synthesis.

It distinguishes itself from earlier techniques by integrating an attention mechanism designed to meticulously capture intricate details at both the word and sub-region levels. It empowers the generation of superior quality images, demonstrating a more accurate alignment with the provided text descriptions. The core elements of the AttnGAN model encompass:

Attentional Generative Network: This model utilizes an attention mechanism to selectively highlight pertinent terms within the text description. It uses a global sentence vector to generate a low-resolution image initially and then refines it through multiple stages. At each stage of its operation, the network utilizes word vectors and attention mechanisms to systematically create more refined sub-regions of the image.

Deep Attentional Multimodal Similarity Model (DAMSM): The global sentence and fine-grained word level information helps DAMSAM to calculate the level of similarity between generated images and sentences. For training the generator, it offers more precise image-text matching loss. Examined on two datasets, CUB [12] and COCO [11], the AttnGAN performs noticeably better than earlier the most advanced models. Additionally, there was a noticeable increase of around 14% in the CUB Inception Score and roughly 170% on COCO.

The paper includes detailed analyses and visualizations of the attention layers of AttnGAN, which demonstrate its effectiveness in capturing fine-grained details.

The paper also highlights the significance of attention mechanisms in text-to-image creation, that has gained prominence in research discipline in recent years but hasn't fully explored the potential of attention models. It's worth noting that the paper presents both quantitative and qualitative results to support its claims and findings.

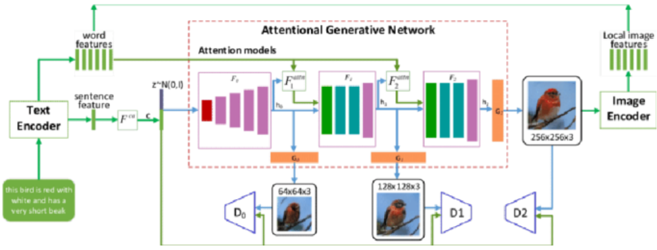


Figure 2. AttnGAN Architecture

Its key innovation lies in the attention mechanism that allows for more precise and detailed image synthesis, which can have applications in various domains.

In 2019, “MirrorGAN: Learning Text-to-image Generation by Redescription” authored by Tingting Qiao, introduces a cutting-edge framework known as MirrorGAN [3], which tackles the intricate challenge of maintaining semantic consistency while generating images.

Despite notable advancements in generative adversarial networks (GANs) that produce high-quality and visually realistic images, ensuring that these images align semantically with their accompanying textual descriptions remains a formidable hurdle.

In response to this challenge, MirrorGAN takes a distinctive approach, emphasizing the process of learning text-to-image synthesis through a mechanism of redescription. The framework is composed of three pivotal modules: The Semantic Text Embedding Module is responsible for generating embeddings at both the word and sentence levels. This makes it easier to understand the textual input in a more thorough way. Next, the Global-Local Collaborative Attentive Module, which is responsible for maintaining semantic uniformity across domains while creating images. Finally, the Semantic Text Regeneration and Alignment Module gives the framework an additional degree of adaptability and completeness by enabling the alignment and retrieval of textual data depending on the visual output.

MirrorGAN lays a robust foundation for enhancing text-to-image generation and its vice-versa, making use of the concept of learning through redescription. This implies that the system employs a specific model that combines both a wide (global) and a detailed (local) perspective to assist guarantee that the generated visuals are coherent and compatible with the textual descriptions. When converting text into graphics, this method not only preserves the general sense and context of the text, but it also refines the creation process, making it more efficient and effective.

By capturing nuanced semantic distinctions in text descriptions, MirrorGAN exhibits substantial potential for applications requiring cross-media content modelling and alignment. This research not only pushes the boundaries of text-to-image generation but also paves the way for more accurate and semantically consistent AI systems, with far reaching implications across various domains.

It scores better in Human perceptual test than AttnGAN [2]. It is better at mimicking real world visuals and thus score high in authenticity tests. It also excels in semantic consistency test where it was able to create more detailed image with respect to the given text descriptions.

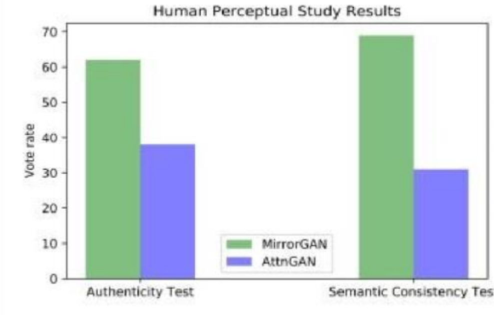


Figure 3. MirrorGAN and AttnGAN comparison on the basis of Authenticity and Semantic Consistency Tests

In 2021, a research paper called "SAM-GAN: SelfAttention supporting Multi-stage Generative Adversarial Networks for text-to-image synthesis" authored by Dunlu Peng, did Qualitative evaluation of SAM-GAN [4] model. When compared to AttnGAN, SAM-GAN-generated images exhibit more precise features and clear backgrounds. It is attributed to the self-attention mechanism, which fosters long-range dependencies between image regions.

The SAM-GAN method produces photos with a more comprehensive and diversified selection of backdrops. The clever treatment of noise vectors by the mode-seeking regularization module helps to this variety, guaranteeing that the produced pictures do not rely on a restricted set of backgrounds, but instead display a wide range of contextual changes. Notably, AttnGAN-generated images manifest defects like overly plump bodies, poor head-to-body ratios, unnatural contours, and incomplete subjects. SAM-GAN rectifies these issues, emphasizing its enhanced understanding of text semantics and ability to generate images with clearer structures and natural colors.

Method	CUB	COCO
	$R-Precision(\%)$	$R-Precision(\%)$
MirrorGAN	60.42	82.44
AttnGAN	67.82	85.64
SAM-GAN	**70.28**	**86.47**

Figure 4. comparitive R-precison(%) of SAM-GAN with AttnGAN and MirrorGAN

SAM-GAN's proficiency is further evident in its capability to create initial images with reasonable layouts and refine Stage-I images into more photo-realistic renditions. This sequential generation process allows SAM-GAN to provide gradually improved ultra clear images based on the input text. For instance, the initial images from SAM-GAN are relatively complete, allowing subsequent stages to refine and enhance the image's details.

By introducing different noise vectors, it can produce a range of images that are both distinct from one another and conceptually linked, demonstrating the versatility and coherence of the generated content. These images show variations, particularly in background content, underscoring SAM-GAN's capacity to produce diverse images.

In 2022, The groundbreaking research described in the paper titled "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding" discusses Imagen T2I model. The ideas and conclusions presented in the Imagen paper is as follows:

Preface: The Imagen model [5] is introduced as a text-to-image synthesis approach that merges diffusion models with the capabilities of extensive transformer language models, such as T5. Additionally, the significance of multimodal learning is explored in general.

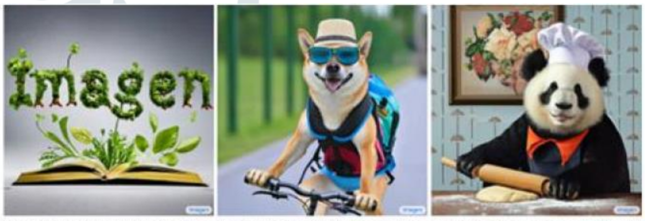
Making Use of Language Models: The research emphasizes that the choice of the large language model and its size significantly impacts the overall performance of text-to-image formation, underscoring the importance of the text encoder's scale in achieving higher-quality results.

Outstanding Outcomes: Imagen accomplished 7.27 FID (FID (Fréchet Inception Distance) score with no training on COCO dataset. Furthermore, the generated output visuals scored high in human perceptual tests as well.

DrawBench Benchmark: DrawBench is a benchmark used to assess the capabilities of image generation models, such as their ability to handle complex prompts, uncommon words, and odd interactions. The model was evaluated using FID and CLIP scores, but human perceptual assessments were given top priority. Imagen consistently receives higher ratings from human raters than other models. Imagen uses T5-XXL [13] text encoders to evaluate its performance.

Effective Architecture: Imagen's architecture is explained, emphasizing changes for convergence, inference speed, and memory efficiency.

Analysis of Imagen: The study presents in-depth conclusions from Imagen, such as the importance of text encoder size scaling, the fact that people prefer T5-XXL [13] over CLIP [14] and the effect of noise conditioning augmentation. It also exposes the shortcomings of the model, especially when it comes to producing pictures of people and the presence of social biases in generated content.



Sprouts in the shape of text 'Imagen' coming out of a fairytale book.

A photo of a Shiba Inu dog with a backpack riding a bike. It is wearing sunglasses and a beach hat.

A high contrast portrait of a very happy fuzzy panda dressed as a chef in a high end kitchen making dough. There is a painting of flowers on the wall behind him.

Figure 5. Imagen generated images

In 2022, the paper "Retrieval-Augmented Text-to-Image Generator (Re-Imagen) [6]", a novel generative model is presented to overcome the shortcomings in text-to-image generation, especially when it comes to producing images of unusual entities.

Problem Statement: Modern image generation models are able to generate crisp and clear images for common entities, but they have trouble with rare or invisible entities. This restriction may cause inaccurate images to be generated, which is problematic for real-world applications.

Re-Imagen as a Solution: Through this approach, the model acquires relevant and meaningful image-text-retrieval triples, which are then used as a valuable resource for the image generation process. The Imagen dataset serves as the foundation for this dataset. Furthermore, it uses WikiImages and COCO for evaluation. It then makes use of this knowledge to increase the faithfulness and accuracy of generated images for less common entities.

Evaluation: EntityDrawBench, a benchmark for assessing the capability of a model to produce real-world objects and subjects. Furthermore, it excels at creating visuals with exceptional precision that reflect unique or less often encountered entities.

Comparison to Existing Models: Re-Imagen performs significantly better in terms of FID score than KNN-Diffusion and Memory-Driven T2I models when compared to other text-to-image generation models. Even with fewer parameters, Re-Imagen-small performs comparably to normal-sized Imagen.

Model Architecture: Re-Imagen is an Imagen-like cascaded diffusion model that uses a variety of diffusion models to produce high-resolution images.

Sampling Strategy: During image generation, the paper presents an interleaved classifier-free guidance sampling strategy that balances the alignment of text and retrieval conditions.

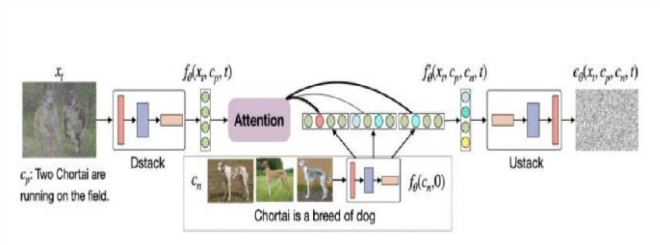

x_t
$f_\theta(x_t, c_p, t)$
Attention
$f'_\theta(x_t, c_p, c_n, t)$
$\epsilon_\theta(x_t, c_p, c_n, t)$
Dstack
Ustack
c_p: Two Chortai are running on the field.
c_n
$f_\theta(c_n, 0)$
Chortai is a breed of dog

Figure 6. Model Architecture of Re-Imagen

In 2022, "GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models," [7] the authors provide a thorough analysis of their new model for text-guided image generation and editing. The paper discusses using diffusion models for inpainting and how they explicitly fine-tune the model for this task, achieving better results compared to standard diffusion model inpainting.

Introduction: The difficulty of creating photorealistic images from text descriptions is discussed in the paper. It presents GLIDE (Guided Language to Image Diffusion for Generation and Editing), a 3.5 billion parameter text-conditional diffusion model. Rich visual content creation and fine-grained image editing with natural language prompts are made possible by text-conditional image models.

Diffusion Models: To create images, the study employs Gaussian diffusion models with progressive noise injection. By using a method for learning Σθ, it can improve sample quality while requiring fewer diffusion steps. Diffusion models produced innovative results on imagine production benchmarks.

Guided Diffusion: To condition diffusion models on text prompts, classifier guidance and classifier-free guidance are investigated. In the case of classifier-free guiding, the model navigates between predictions given with and without explicit labels, providing nuanced interpolation throughout the spectrum of labelled and unlabelled examples. Classifier guidance, on the other hand, comprises using the labels supplied by a classifier, bringing a more organized and supervised component to the conditioning process.

CLIP Guidance: Contrastive Language-Image Pretraining, or CLIP, is presented as a method for teaching text and image together to form joint representations. It can be used to guide generative models since it gives a score that indicates the proximity of an image to a given text caption. Several applications have made use of CLIP to direct the creation of images based on user-defined text prompts.

Results: Classifier-free guidance outperforms CLIP guidance and is liked by humans for its naturalism and textual alignment. Samples generated by GLIDE using classifier-free guidance are preferred over those from DALLE in 87% of cases for photorealism and 69% for caption similarity. GLIDE demonstrates the ability to produce lifelike images with realistic edits like shadows and reflections. The model excels at text driven image inpainting, allowing for realistic edits to existing images.

CLIP Model Application in Diffusion: The replacement of classifiers in diffusion models with CLIP models for guidance is discussed in the paper. They successfully control the diffusion process by perturbing the reverse-process mean. Diffusion models are guided by explicitly trained noised CLIP models on noisy images.

Comparing CLIP Guidance with Noise: Generic CLIP models lack specialized training in handling noisy images. Therefore, the paper emphasizes the significance of using noised CLIP guidance models in order to improve performance during the diffusion sampling process,

Model Education: The authors train a text-conditional diffusion model with 3.5 billion parameters at 64x64 resolution, and a text-conditional up-sampling diffusion model with 1.5 billion parameters at 256x256 resolution. For CLIP guidance, a noised 64x64 ViT-L CLIP model is trained. The models, whose architectures correspond to the corresponding requirements, are trained using the same dataset as DALL-E.

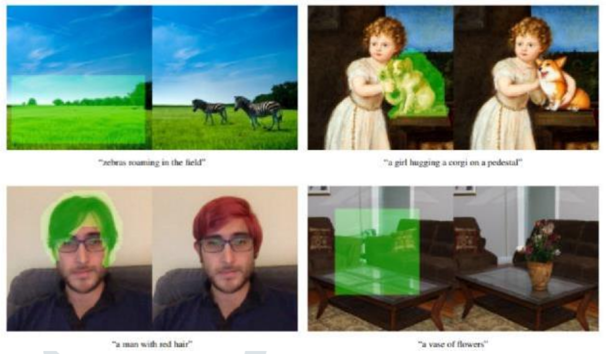

"zebras roaming in the field"
"a girl hugging a corgi on a pedestal"
"a man with red hair"
"a vase of flowers"

Figure 7. Image Editing possible with GLIDE

In 2022, The paper "Hierarchical Text-Conditional Image Generation with CLIP Latents" by Aditya Ramesh, presents a two-stage model for text-conditional image generation using CLIP embeddings. Creating UnCLIP[8] was approach for creating DALLE-2 model by OpenAI.

Introduction: The paper introduces the concept of UnCLIP, a model for text-guided image generation. It builds upon the CLIP model, which embeds images and text into a shared latent space. CLIP embeddings are known for their robustness, zero-shot capabilities, and success in vision and language tasks.

Approach: The authors propose a three-stage model: a CLIP model for encoding text and images, a diffusion prior model, and a generative stack (decoder). The generative stack employs DDIM (Diffusion-Dequantization-Inpainting Model) to produce images. UnCLIP uses spherical interpolation to create intermediate representations during image generation.

Key Contributions:

1.Image Manipulations: The paper demonstrates various image manipulations enabled by UnCLIP, including generating variations and interpolations of images. This allows for creating diverse images while preserving the essential content.

2.Text Diffs: UnCLIP enables language-guided image manipulations by interpolating between image CLIP embeddings and text differences. This allows for modifying images to match new text descriptions.

3.Probing CLIP Latent Space: The authors use UnCLIP to explore the CLIP latent space and investigate cases where CLIP makes incorrect predictions. They find that unCLIP can still generate relevant images in cases of typographic attacks.

4.Aesthetic Quality Comparison: An aesthetic quality evaluation is conducted, comparing UnCLIP to GLIDE. Presence of guidance has significantly enhanced the aesthetic appeal for both models, with UnCLIP maintaining a balance between aesthetics and recall.

5.State-of-the-Art Performance: UnCLIP achieves a high FID score on the MS-COCO validation set, demonstrating its competitive performance in generating realistic images from text prompts.

6.Human Evaluations: More diverse images can be produced by UnCLIP in comparison to GLIDE. It offers competitive photorealism and caption similarity. The hierarchical nature of UnCLIP and its ability to balance diversity and fidelity in image generation sets it apart from other approaches.

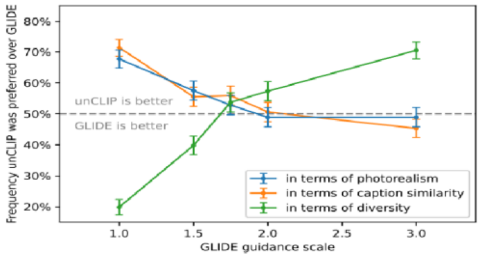


Figure 7. Human evaluators prefer unclip, as it yields greater diversity, photorealism and caption similarity.

In 2022, a research paper titled "On Distillation of Guided Diffusion Models" [9] which is part of Google Research, Brain Team. Here, authors offer a novel method for distilling classifier-free guided diffusion models in this review paper. These models have been criticized for being too computationally intensive, despite being very effective at producing high-resolution images. To solve this problem, a two-stage distillation process is introduced in the suggested method.

Approach:

The first step of the distillation process is to achieve alignment between the collective outputs generated by the two diffusion models within the teacher model. By combining the conditional and unconditional components of the original model, this helps produce a more efficient model. In the second step, the main aim is to make the first model simpler. A diffusion model is created that doesn't need as many steps to work well. So, the focus is to gradually simplify the first model to make it more efficient while still keeping it effective. This helps the overall diffusion process work better with fewer complications.

Key Contributions:

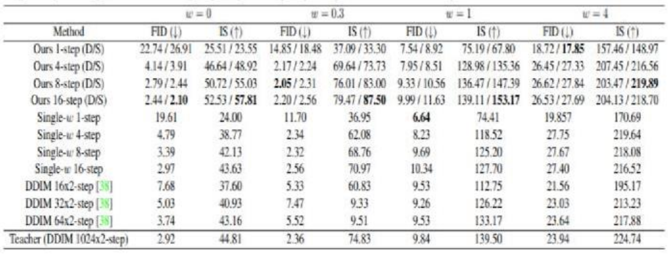
The distilled model can generate high-definition images using as few as four sampling steps for pixel-space models, offering a speedup of up to 256 times while keeping competitive FID/IS scores.

The paper also introduces a stochastic sampling process to enhance the distillation process, providing a solution to the computational efficiency problem associated with classifier free guided diffusion models.

For latent-space text-to-image models, the approach offers an efficient solution, highlighting its versatility across different types of diffusion models.

The fine-grained control over guidance strength allows users to tailor the trade-off between sample quality and diversity, enhancing the applicability of the method. Experimental results demonstrate the practical applicability of the distilled models in various image generation and manipulation tasks.

Figure 8. Improving the inference efficiency in Stable Diffusion in 8 steps.



	$w = 0$		$w = 0.3$		$w = 1$		$w = 4$	
Method	FID (↓)	IS (↑)	FID (↓)	IS (↑)	FID (↓)	IS (↑)	FID (↓)	IS (↑)
Ours 1-step (D/S)	22.74 / 26.91	25.51 / 23.55	14.85 / 18.48	37.09 / 33.30	7.54 / 8.92	75.19 / 67.80	18.72 / **17.85**	157.46 / 148.97
Ours 4-step (D/S)	4.14 / 3.91	46.64 / 48.92	2.17 / 2.24	69.64 / 73.73	7.95 / 8.51	128.98 / 135.36	26.45 / 27.33	207.45 / 216.56
Ours 8-step (D/S)	2.79 / 2.44	50.72 / 55.03	**2.05** / 2.31	76.01 / 83.00	9.33 / 10.56	136.47 / 147.39	26.62 / 27.84	203.47 / **219.89**
Ours 16-step (D/S)	2.44 / **2.10**	52.53 / **57.81**	2.20 / 2.56	79.47 / **87.50**	9.99 / 11.63	139.11 / **153.17**	26.53 / 27.69	204.13 / 218.70
Single-w 1-step	19.61	24.00	11.70	36.95	**6.64**	74.41	19.857	170.69
Single-w 4-step	4.79	38.77	2.34	62.08	8.23	118.52	27.75	219.64
Single-w 8-step	3.39	42.13	2.32	68.76	9.69	125.20	27.67	218.08
Single-w 16-step	2.97	43.63	2.56	70.97	10.34	127.70	27.40	216.52
DDIM 16x2-step [38]	7.68	37.60	5.33	60.83	9.53	112.75	21.56	195.17
DDIM 32x2-step [38]	5.03	40.93	7.47	9.33	9.26	126.22	23.03	213.23
DDIM 64x2-step [38]	3.74	43.16	5.52	9.51	9.53	133.17	23.64	217.88
Teacher (DDIM 1024x2-step)	2.92	44.81	2.36	74.83	9.84	139.50	23.94	224.74

Figure 9. FID & Inception Score are quantitative measures of how realistic and diverse the pictures produced are., whereas the Inception Score alone assesses the created image distribution. Stable Diffusion's FID and IS score increases at each denoising step.

The evolution of Text-To-Image Generation models from 2016 to 2022 highlights the continuous improvements and adaptations introduced to enhance their practicality and applicability in real-world scenarios.

Papers referred	Key Findings	Techniques Used	Conclusion
“Generative Adversarial Text to Image Synthesis" [1]	-Disentangling Style and Content - Interpolation Regularization -Generalizability - Style Transfer	Generative adversarial network	Model can generate images based on given prompts successfully
AttnGAN: Fine-Grained Text to Image Generation With Attentional Generative Adversarial Networks” [2]	Uses Attentional Generative Network at each stage to generate more detailed local regions of the output visuals.	Attentional Generative Adversarial Network (AttnGAN)	The attention mechanism produce better image outputs but sometimes it can generate incomplete subjects.
“MirrorGAN: Learning TextTo-Image Generation by Re-description” [3]	-Semantically consistent model -It scores better in Human perceptual test than AttnGAN.	MirrorGAN	MirrorGAN uses redescription for enhancing text to image conversion and vice-versa.
“SAM-GAN: Self-Attention supporting Multi-stage Generative Adversarial Networks for text-to-image synthesis” [4]	-improved clear and crisp images based on the input text -Generated diverse images with clear backgrounds in comparison to AttGAN	SAM-GAN	Enhanced understanding of text semantics and ability to generate images with clearer structures and natural colors.
“Photo-realistic Textto-Image Diffusion Models with Deep Language Understanding” [5]	-Speed and efficiency -Better FID score -Higher DrawBench Benchmark scores	IMAGEN	Capable of producing high-fidelity image that combines diffusion models with the power of large language models, like T5 but can have social biases in generated content.
“Re-IMAGEN: Retrieval-Augmented Text-To-Image Generator” [6]	-Can produce images of unusual entities. -Uses EntityDrawBench for evaluation -Trained by dataset consisting of image-text-retrieval triples	Re-IMAGEN	Model can generate accurate images of rear or unusual entities.
"GLIDE: Towards Photorealistic Image Generation and Editing with TextGuided Diffusion Models," [7]	-creating photorealistic images from text descriptions -Editing images according to text description is possible -Need fewer diffusion steps	GLIDE	The model excels at textdriven image inpainting, allowing for realistic edits to existing image and has the ability to produce photorealistic images with shadows, reflections.
“Hierarchical TextConditional Image Generation with CLIP Latents”[8]	-Maintains a balance between aesthetics and recall. - UnCLIP was approach for creating DALLE-2 model by OpenAI. - creating diverse images while preserving the essential content. -Risk of unwanted content generation and biases.	UnCLIP	UnCLIP can be a powerful tool for generating images based on textual prompts with enhanced diversity and quality compared to GLIDE.
“On Distillation of Guided Diffusion Models” [9]	The distilled model can generate lifelike images using as few as four sampling steps for pixel-space models, offering a speedup of up to 256 times while keeping competitive FID/IS scores.	Stable Diffusion	Experimental results demonstrate the practical applicability of the distilled models in various image generation and manipulation tasks.

TABLE I. COMPARISON AND ANALYSIS OF PAPERS REVIEWED

IV. RESEARCH GAPS

The current landscape of text-to-image (T2I) generation models faces a notable research gap that warrants exploration and investigation. One of the major gaps is the model's ability to generalize to various target cultures.

This exploration would involve examining how cross-cultural image generation can be harnessed to improve user experience and foster user acceptance across applications such as advertising, content creation, accessibility tools, and language learning. Furthermore, this review could shed light on the ethical considerations surrounding T2I models, especially in terms of addressing cultural representation and avoiding stereotypes.

Second is the Sample Efficiency Gap It states that despite the promise of Stable Diffusion models in reducing data requirements, the extent to which they can enhance sample efficiency while maintaining image quality remains unexplored. Identifying the boundaries of sample efficiency is essential for practical deployment. Furthermore, existing evaluation metrics have limitations, and identifying more appropriate evaluation criteria is essential.

Addressing these research gaps is essential for refining the generative model capabilities and practical applications.

V. CONCLUSION

In conclusion, the evolution of text-to-image generation models from 2016 to 2022 has marked a remarkable journey filled with transformative developments and pivotal breakthroughs, reshaping the landscape of artificial intelligence and computer vision. The inception of Generative Adversarial Networks (GANs) in 2016 laid the foundational groundwork for translating textual descriptions into direct image pixels. However, the pursuit of crafting images with impeccable body proportions, diverse backgrounds, and complete object representations found its resolution in SAMGAN, while semantic consistency was achieved through the innovative approach of redescription.

The year 2022 witnessed a significant milestone with the introduction of IMAGEN, a model capable of creating photorealistic high-resolution images, addressing a prior limitation of its predecessors. Subsequently, the notion of leveraging the IMAGEN dataset led to the development of Re-Imagen, which excelled in generating images of fantastical entities.

The advent of new models such as GLIDE and UnCLIP enabled users to edit images according to textual descriptions, upholding aesthetic and photorealistic qualities. However, these enhancements came with computational efficiency challenges, which were effectively mitigated through the introduction of the Stable Diffusion model. This innovation not only gained substantial popularity but also rendered text-to-image generation models applicable in real-world scenarios.

Yet, the journey is far from complete. The need for culturally diverse text-to-image generation models that can generate non-biased images remains. Moreover, fostering interdisciplinary collaborations, where experts from fields such as linguistics, psychology, art, and ethics come together, holds the promise of offering unique insights that enrich the development of these models. Such partnerships can pave the way for holistic solutions that align more closely with human cognition and values. As we move forward, these endeavours will undoubtedly continue to shape the future of generative text-to-image creation models.

REFERENCES

[1] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, Honglak Lee Proceedings of The 33rd International Conference on Machine Learning, PMLR 48:1060-1069, 2016

[2] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Finegrained text to image generation with attentional generative adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1316– 1324, 2018

[3] Tingting Qiao, Jing Zhang, Duanqing Xu, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1505-1514

[4] Dunlu Peng,Wuchen Yang,Cong Liu,Shuairui Lü SAM-GAN: Self-Attention supporting Multi-stage Generative Adversarial Networks for textto-image synthesis, Neural Networks Volume 138, June 2021, Pages 57-67

[5] https://proceedings.neurips.cc/paper_files/ paper/2022/hash/ec795aeadae0b7d230fa35cbaf04c04 1-Abstract-Conference.html

[6] RE-IMAGEN: RETRIEVAL-AUGMENTED TEXT-TO-IMAGE GENERATOR Wenhu Chen, Hexiang Hu, Chitwan Saharia, William W. Cohen Google Research

{wenhuchen,sahariac,hexiang,wcohen}@google.com

https://arxiv.org/pdf/2209.14491.pdf

[7] GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models Alex Nichol * Prafulla Dhariwal * Aditya Ramesh * Pranav Shyam Pamela Mishkin Bob McGrew Ilya Sutskever Mark Chen https://arxiv.org/pdf/2112.10741.pdf

[8] "Hierarchical Text-Conditional Image Generation with CLIP Latents" Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, Mark Chen https://3dvar.com/Ramesh2022Hierarchical.pdf [9] https://openaccess.thecvf.com/content/CV PR2023/papers/Meng_On_Distillation_of_Guided_Di ffusion_Models_CVPR_2023_paper.pdf

[10] Z. Nilsback, Maria-Elena and Andrew, "Automated flower classification over a large number of classes," in 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing. IEEE, 2008, pp. 722–729.

[11] Microsoft COCO: Common Objects in Contexthttps://link.springer.com/chapter/10.1007/978-3-31910602-1_48

[12] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset.

[13] Technical Report CNS-TR2011-001,

California Institute of Technology, 2011.

[14] *Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. JMLR, 21(140), 2020*

[15] *Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In ICML,2021.*