Text-To-Image Generation Using Stable Diffusion For Some Global Languages

Devika.D
IIT2021013@iiita.ac.in
Indian Institute Of Information Technology
India

Udayasree.B
IIT2021038@iiita.ac.in
Indian Institute Of Information Technology
India

Bhavana.M
IIT2021077@iiita.ac.in
Indian Institute Of Information Technology
India

Janvi.T
IIT2021222@iiita.ac.in
Indian Institute Of Information Technology
India

Abstract

The widespread adoption of diffusion models has demonstrated the potential of artificial intelligence (AI) systems to generate realistic images. This work undertakes a critical analysis of the process of creating and detecting realistic human faces using Stable Diffusion.Our project focuses on Stable Diffusion, a text-to-image diffusion model designed to create high-quality images from textual descriptions. This model, founded on an optimized iteration of traditional diffusion models, proves versatile in generating diverse images, ranging from realistic portraits to captivating landscapes and abstract art. Introducing a novel method for text-to-image synthesis, our approach seamlessly incorporates stable diffusion with a built-in dataset directly into the image creation process. By enhancing model stability through adaptive techniques, our method consistently produces diverse and realistic images across a range of textual inputs. Through extensive testing on standard benchmarks, our approach, combining stable diffusion and a built-in dataset, outperforms existing methods in terms of image quality, diversity, and faithfulness to the input text. Leveraging the benefits of a built-in dataset ensures that our model is trained on a comprehensive and diverse set of textual descriptions, contributing to its ability to accurately translate varied input into visually coherent representations.

keywords-Image quality, Diversity, Stable diffusion.

ACM Reference Format:
Devika.D, Udayasree.B, Bhavana.M, and Janvi.T. 2023. Text-To-Image Generation Using Stable Diffusion For Some Global Languages. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

I Introduction

In the fast-evolving fields of artificial intelligence and computer vision, the challenge of generating images from natural language descriptions has captivated many researchers and practitioners. The human brain has the capacity of processing the text with its imagination to see what it would look like. Similarly, this is a step taken to in-built the imagination of human vision into computer vision that sets a path to the thinking of text-to-image conversion using the description in the natural language. This endeavor finds applications in various domains, from artistic expression to practical tasks like computer-aided design. Moreover, it has spurred substantial progress in multi-modal learning, where combining visual and textual data represents one of the most dynamic areas of contemporary research[1].

Objective: The primary goal of this work is to leverage the capabilities of Stable Diffusion to generate high-quality images from textual descriptions, surpassing the performance of existing models[11]. By employing Stable Diffusion's novel diffusion process, we aim to achieve a significant reduction in both training time and loss compared to previous generative adversarial network (GAN) approaches[2].

GAN: Generative Adversarial Networks

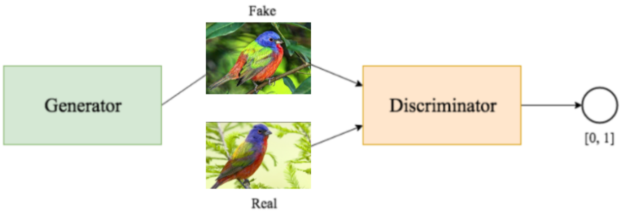


Figure 1. Basic idea of GAN.

Despite the transformative potential, access to these technologies has predominantly been confined to experts in the field due to the intricate knowledge required for generating realistic deepfakes. Contemporary image generation methods, notably those based on Generative Adversarial Networks (GANs), necessitate meticulous tuning and adaptation for specific scenarios to produce convincing false samples.

Working of GANS:

Generative Adversarial Network is a type of generative neural network model for unsupervised machine learning

Conference'17, July 2017, Washington, DC, USA
2023. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

which has drawn much attention in the arts and design due to its capacity to learn and generate creative products that are usually attributed to human endeavor. The basic architecture of GANs comprises two competing networks (Figure 1): a generator and a discriminator. During the training, the generator learns to map a high dimensional latent space z to generate new sample data. Inversely, the discriminator learns to distinguish the fake samples from the real samples in the dataset and then gives feedback to the generator. With this adversarial mechanism, their capacity to distinguish and generate fake samples improves along with the training. In this way, by exposing GANs to a particular set of arts or design examples in the training process, the networks can automatically learn the underlying pattern, or probabilistic distribution, of the given data and generate novel samples that share similar properties to the original images[14].

When we choose GANS then NLP goes hand in hand. Natural Language Processing (NLP) involves using computers to extract information from human language, advancing language modeling through high-dimensional vector spaces and neural networks[15]. NLP models are now integral in language translation, speech recognition, and personal assistants. In architectural computational techniques, there is a shift towards qualitative aspects of human experiences, where NLP plays a vital role in analyzing the semantic dimension of human-machine design processes[14].

The language of design is pivotal in conceptualization, development, and evaluation, serving as an operator in negotiating outcomes and documenting the design process [16]. NLP techniques are increasingly applied to analyze linguistic behavior in collaborative design, recognizing the ongoing importance of verbal expression in design and communication [17]. Integrating NLP with computational generative design creates a unique interface between machines and humans, combining Generative Adversarial Networks (GANs) with NLP to introduce a qualitative dimension to the design process.

II Understanding Stable Diffusion :

The theoretical foundation of stable diffusion in artificial intelligence and machine learning involves sophisticated mathematical and statistical principles. It serves as a pivotal framework for controlled data transformations, aiming to generate novel and coherent samples from a learned distribution. This framework integrates probability distributions, stochastic processes, and optimization algorithms. The controlled transformation process balances exploration and exploitation, navigating data space with stability and coherence. deep generative models play key roles in the theoretical framework. Stable diffusion addresses challenges in high-dimensional data spaces, offering effective exploration and revealing latent structures[10]. It ensures adaptability by generating data samples that adhere to learned distributions with meaningful variations. The theoretical sophistication guides algorithmic mechanisms for data generation. As datasets grow in complexity, stable diffusion provides a means to explore high-dimensional spaces. In the evolving landscape of artificial intelligence, these theoretical foundations remain instrumental for pushing the boundaries of data generation.

Working Principle Of Stable Diffusion

Stable Diffusion, a innovative model in generative systems, seamlessly blends the strengths of diffusion models and latent variable models to propel text-to-image generation into new realms of quality and efficiency. At its core, Stable Diffusion excels in transforming textual prompts into high-fidelity images through a meticulously orchestrated process.

The journey begins with the encoding of a text prompt, where a text encoder translates the linguistic input into a latent vector. This vector encapsulates the underlying semantic essence of the text, acting as a guiding force throughout the subsequent image generation process[10].

The model then introduces a random noise tensor, representing the initial 'noisy' version of an image. This tensor embarks on a diffusion process, systematically shedding noise at each step. Crucially, a diffusion model comes into play, predicting and regulating the amount of noise present. The latent vector contributes significantly here, ensuring that the generated image aligns coherently with the semantics of the input text[11].

Following the diffusion process, the de-noised noise tensor undergoes a transformative journey through a U-Net decoder. This specialized neural network excels in the nuanced task of image reconstruction, piecing together the de-noised tensor into a refined and visually coherent image.

For added sophistication, Stable Diffusion accommodates depth guidance if pertinent information is available in the text prompt. The extracted depth details guide the U-Net decoder, ensuring the generated image maintains consistent depth relationships and object placements[10].

Output: The final output is a high-quality image that corresponds to the given text prompt. The integration of stable diffusion ensures the generated images are not only of superior quality but also exhibit increased stability and efficiency compared to traditional methods like Generative Adversarial Networks (GANs).

This simplified working of stable diffusion provides a glimpse into its transformative potential in various applications, ranging from creating artwork and enhancing images to generating concepts for design and aiding in educational or scientific visualizations[10].

In summary, the theory of stable diffusion represents a significant advancement in the field of generative models. With its ability to transform noisy data back into coherent and high-quality outputs, stable diffusion models offer a wide

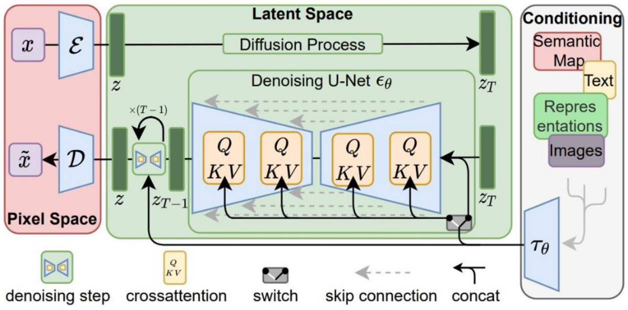


Figure 2. Process of stable diffusion

range of applications that are only beginning to be fully realized and appreciated. As research and development in this area continue to grow, we can expect to see even more remarkable applications and innovations arising from this fascinating domain[10].

Applications of Stable Diffusion:

Stable diffusion has a wide range of applications that span numerous businesses and creative disciplines, notably in the context of AI and machine learning. The following are a few of the major uses: 1.

1. Image Generation and Enhancement: One of the primary applications of stable diffusion models is in generating high-quality images. This includes creating entirely new images or enhancing existing ones, such as improving image resolution or repairing damaged photographs. This technology is extensively used in fields like digital art and graphic design.

2. Audio Generation and Editing: Similar to image generation, stable diffusion can be used to generate and edit audio. This includes creating realistic sound effects, generating music, or enhancing the clarity of recorded audio. This includes creating realistic sound effects, generating music, or enhancing the clarity of recorded audio.Its use can be noteworthy in the construction of realistic soundscapes for virtual reality experiences, as well as in the sound design of movies and video games.

3. Text Generation: In the realm of natural language processing, stable diffusion models can be employed to generate coherent and contextually relevant text. This has applications in content creation, automated journalism, chatbots, and language translation services[2].

4. Data Augmentation: Stable diffusion can be used to augment datasets, especially when the available data is limited. By generating new data points that are coherent with the existing dataset, these models can significantly enhance machine learning model training, particularly in domains where data collection is challenging or expensive.

5. Medical Imaging: In healthcare, stable diffusion models can assist in generating high-resolution medical images from lower-resolution inputs. This can aid in better diagnosis and analysis, particularly in areas like radiology where image clarity is crucial.

6. Scientific Visualization: By creating visuals of complex scientific data, these models can facilitate simpler interpretation and comprehension of the data. In disciplines like environmental science, molecular biology, and astrophysics, this can be especially helpful.

7. Fashion and Interior Design: In the fashion and design industries, stable diffusion models can help in creating and visualizing new designs, providing a tool for designers to experiment with different styles and patterns more efficiently.

8. Video Games and Virtual Reality: In video game development and virtual reality, these models can generate realistic textures, environments, and characters, enhancing the visual quality and immersive experience of these digital worlds.

9. Interactive Storytelling and Film Production: Stable diffusion models can be used to create visual elements for interactive storytelling and film production, aiding in the creation of realistic and engaging visual narratives.

10. Education and Training: These models can also be applied in educational and training scenarios, such as creating simulations or visual aids that enhance learning and understanding of complex concepts.

The applications of stable diffusion are continually evolving, as researchers and practitioners find new and innovative ways to apply these models across different fields. The versatility and effectiveness of these models in generating high-quality, coherent outputs make them a valuable tool in numerous domains.

Choosing Stability: The Advantages of Stable Diffusion over GANs in Facial Transformation-
Choosing stable diffusion models over Generative Adversarial Networks (GANs) for text-to-image conversion tasks can be advantageous for several reasons:

- **Higher Quality and Detail in Outputs:**Stable diffusion models tend to generate images with higher fidelity and more detail compared to GANs. They are particularly adept at handling complex scenes and intricate textures, which is crucial for creating images from descriptive text.
- **Better Handling of Ambiguity:** Text descriptions can often be ambiguous or contain elements that are not explicitly detailed. Stable diffusion models are generally better at interpreting and filling in these gaps to create coherent images, whereas GANs might struggle with such ambiguities.
- **Greater Diversity in Generated Images:**Stable diffusion models can offer a broader diversity in the images

they generate from the same text input. This is because they explore the data distribution more thoroughly, whereas GANs might latch onto more common patterns, leading to less variation in outputs.

- **Improved Stability and Training Efficiency:** Training GANs can be challenging due to issues like mode collapse, where the generator starts producing a limited variety of outputs. Stable diffusion models, by design, tend to have more stable training dynamics and are less prone to such issues.
- **Flexibility in Content Generation:** Stable diffusion models can be more flexible in terms of content generation, capable of creating a wider range of styles and abstractions. This makes them particularly useful for creative and artistic applications where diversity and uniqueness are valued.
- **Better Control Over the Generation Process:** With stable diffusion models, there is often better control over the image generation process, as they provide more intuitive ways to steer the generation towards desired characteristics. This can be particularly useful when trying to match specific aspects of the text description.
- **Efficient Scaling and Adaptation:** Stable diffusion models are generally more efficient to scale and adapt to different datasets and domains. This scalability is important when dealing with diverse text descriptions and the need for varied image outputs.
- **Energy Efficiency:** Some studies suggest that stable diffusion models can be more energy-efficient compared to GANs, especially at scale, which is an important consideration given the growing concerns around the environmental impact of AI systems.
  In summary, while GANs have been a popular choice for text-to-image conversion tasks, the emergence of stable diffusion models offers several advantages in terms of image quality, diversity, stability, and flexibility, making them a compelling alternative in this domain.

III Literature survey:

Text-to-image synthesis, a captivating interdisciplinary field, is at the intersection of linguistics and computer vision. It presents an array of innovative strategies for translating textual input into visual output. Some of these strategies involve the orchestration of recurrent neural networks and generative adversarial networks, like in the case of AttnGAN and StackGAN.

Furthermore, methods like CLIP offer a two-way street, where text and images can be mapped to a common embedding space, allowing for powerful cross-modal understanding and synthesis. StyleGAN and StyleGAN2, originally designed for unconditional image generation, can also be adapted to generate images from text through conditional modifications of their input vectors[5].

We will examine the complexities and subtleties of these techniques, as well as how they have advanced the state-of-the-art in text-to-image synthesis, as we go through them. This chapter will serve as a comprehensive guide for researchers, developers, and enthusiasts seeking to harness the creative and practical potential of generating images from text[4].

some of the existing methods and techniques for text-to-image synthesis:

A. Conditional Generative Adversarial Networks (cGANs):

- Conditional GANs are widely used for text-to-image synthesis. They take textual descriptions as conditional input and generate corresponding images. Models like StackGAN and AttnGAN use this approach, allowing for more detailed and realistic image generation[2].
- **StackGAN:** The StackGAN paper was authored by Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, and Xiaogang Wang. It was published in 2017 https://arxiv.org/abs/1612.03242.
- ***Model Working***: StackGAN utilizes a two-stage process. The Stage-I GAN sketches the primitive shape and colors of the object based on the given text description, yielding Stage-I low-resolution images. The Stage-II GAN takes Stage-I results and text descriptions as inputs, and generates high-resolution images with photo-realistic details[5].

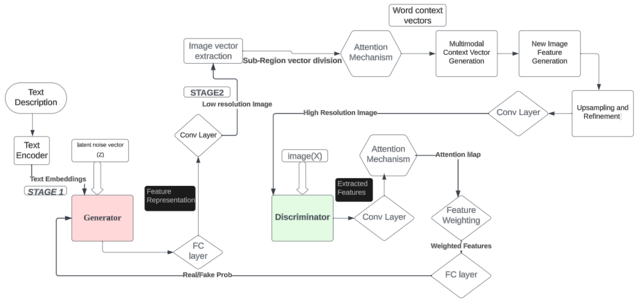


Figure 3. Basic idea of Stack GAN.

- **AttnGAN:** AttnGAN was developed by Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. The research paper was published in 2018. https://ieeexplore.ieee.org/document/8578241
- ***Model Working:*** Attention gan uses a two-step process in the first step it draws different subregions of the image by focusing on words that are more relevant to the subregion being drawn. Each word in the sentence

is encoded into a word vector and then the generative network utilizes the word vector to generate the low-resolution images.

- The generated image which is generated by the generator and other images are taken as input by the discriminator. The discriminator's ability to distinguish between real and fake images is what allows the GAN to learn to generate realistic images.

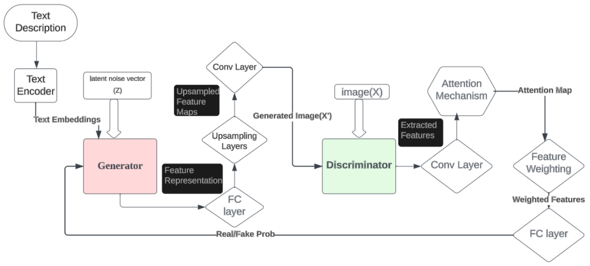


Figure 4. Basic idea of Attn GAN.

B. BigGAN:

- BigGAN is a powerful generative model capable of producing high-resolution images from text descriptions. It is renowned for its capacity to produce detailed, high-quality images from textual input.
- BigGAN was introduced by Andrew Brock, Jeff Donahue, and Karen Simonyan in their paper titled "Large Scale GAN Training for High Fidelity Natural Image Synthesis." The paper was published in 2019. https://arxiv.org/abs/1809.11096

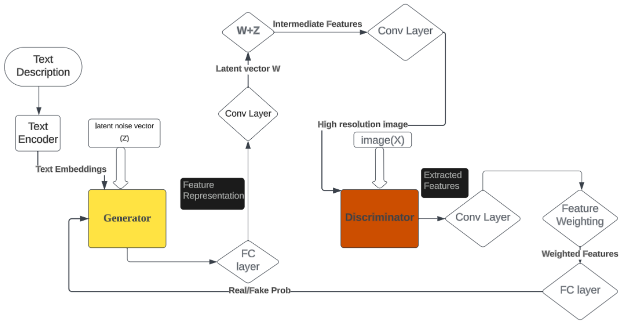


Figure 5. Basic idea of BIG GAN.

- ***Model Working:*** BigGAN is designed for high-resolution image synthesis. It uses a modified GAN architecture and large-scale training to generate high-quality images based on textual descriptions. The model can produce detailed and realistic images, even at very high resolutions[2].

C. CLIP:

- CLIP (Contrastive Language-Image Pre-training) is a versatile model that can both understand and generate images based on text descriptions. It offers flexibility and can be fine-tuned for various text-to-image synthesis tasks.
- CLIP was developed by researchers at OpenAI. The key contributors to CLIP include Alec Radford, Ilya Sutskever, and Sam Altman, among others.
- ***Model Working:*** CLIP is a versatile model that can understand and generate images based on text descriptions. It aligns text and image embeddings in a shared space, allowing for powerful cross-modal understanding and synthesis. The model can be fine-tuned for various text-to-image tasks.

D. StyleGAN and StyleGAN2:

- Originally designed for image generation from random noise, StyleGAN and StyleGAN2 can be adapted for text-to-image generation by conditioning the generation process on textual prompts. These models produce images with specific characteristics based on text input[5].
- StyleGAN and StyleGAN2 were developed by Tero Karras, Samuli Laine, and Timo Aila from NVIDIA Research. The research papers for these models were published in 2019 and 2020, respectively.

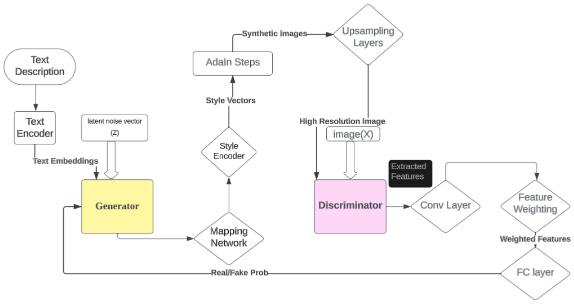


Figure 6. Basic idea of STYLE GAN2.

Model Working Of StyleGAN:

- StyleGAN is trained on a dataset of images and consists of a generator and discriminator then the generator starts with random noise and progressively creates an image.
- After creating the image For text-to-image synthesis, we will add a text description as input alongside the noise.
- The newly added text description is encoded into a latent vector and combined with the noise.Finally, the generator produces an image that matches the text description by adjusting its internal styles, like color and texture, based on the text input[5].

F. Hybrid Models:

- Some approaches combine generative models (like GANs) with text embeddings to generate images from text. Hybrid models can provide high-quality images while offering precise control over the output[9].

Furthermore, the selection among these methods should consider factors such as the dataset characteristics, computational resources available, and the desired level of control over the generation process. The diversity in available options empowers researchers and practitioners to tailor text-to-image synthesis approaches to the unique demands of their projects.

IV Proposed Methodology:

(1) TEXT-TO-IMAGE GENERATION:

1. Package Installations and Imports

The code begins with package installations and imports:

!pip install –upgrade diffusers transformers -q: Installs or upgrades the diffusers and transformers packages. These are essential for leveraging diffusion models and Transformer-based architectures.(-q stands for quiet mode without output)

Library Imports:

- from pathlib import Path: Allows working with filesystem paths in a more object-oriented manner.
- import tqdm: Provides a progress bar for iterations, useful for tracking loops.
- import torch: PyTorch library for machine learning, enabling GPU acceleration and tensor computation.
- import pandas as pd, import numpy as np: Libraries for data manipulation and numerical computation respectively.
- from diffusers import StableDiffusionPipeline: Imports the Stable Diffusion model from the diffusers package, which is a library for diffusion models.
- from transformers import pipeline,set-seed: Imports functionalities from the transformers library, which offers pre-trained models and tools for NLP tasks.
- import matplotlib.pyplot as plt: Matplotlib for data visualization.
- import cv2: OpenCV library, used for image processing tasks.

2. Configuration Class (CFG) The code defines a class CFG to hold configuration parameters:

- Device: ("cuda"): Specifies the device to be used,here TP GPU, for computational tasks.
- Seed: (42): A predefined seed value for ensuring reproducibility in random number generation.
- Generator: Sets up a torch random number generator with the specified seed. **Image Generation Parameters:**
- Steps: (35): Number of steps taken during image generation using the diffusion model.
- Model ID: ("stabilityai/stable-diffusion-2"): Identifier for the specific pre-trained Stable Diffusion model.
- Image Size: ((400, 400)): Dimensions of the generated image.
- Guidance Scale: (9): A parameter controlling guidance influence in the image generation process. **Prompt Generation Parameters:**
- Model ID: ("gpt2"): Identifies the GPT-2 model used to generate prompts.
- Dataset Size: (6): Number of prompts to generate.
- Max Length: (12): Maximum length of the generated prompts.

3. Initializing Image Generation Model The code initializes the image generation model using the "StableDiffusionPipeline from-pre trained()" method:

- StableDiffusionPipeline: This is a class from the diffusers library that handles Stable Diffusion models. It loads a pre-trained Stable Diffusion model identified by the given model ID. The parameters include the model ID, torch data type, revision, authentication token, and guidance scale.
- to(CFG.device): Moves the initialized model to the specified device GPU for computation.

4. Image Generation Function: The code defines a generate-image() function to create images based on prompts:

- generate-image(): This function takes a prompt and the initialized image generation model as inputs. It uses the model to generate an image based on the provided prompt, controlling the number of inference steps, generator, and guidance scale. The generated image is then resized to the specified dimensions.

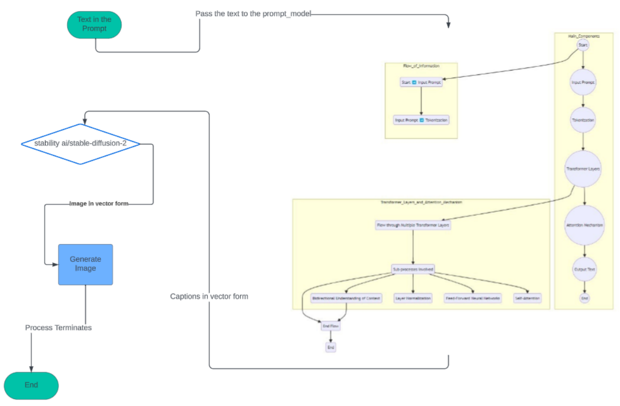


Figure 7. Flow chart of Text-To-Image Generation Model.

5. Executing Image Generation: The code executes the generate-image() function :

- It calls the generate-image() function with the prompt like for example "A Bride with bouquet" and the initialized image-gen-model and displays the result after diffusing layer by layer through multiple iterations.

The iterative diffusion process through multiple layers adds a dynamic element to the synthesis.

(2) TEXT-TO-IMAGE GENERATION FOR MULTILINGUAL LANGUAGE:

1. Package Installation and Library Imports

- The code starts by installing the googletrans package version 3.1.0a0. It then imports various libraries such as Translator from googletrans, Path from pathlib, tqdm, torch, pandas, numpy, StableDiffusionPipeline from diffusers, pipeline and set-seed from transformers, matplotlib.pyplot as plt, and cv2 for OpenCV.

Translation Function Using Google trans: There's a new function get-translation() added:

- get-translation(): This function takes a text and destination language as inputs. It utilizes the Translator class from the Google trans package to translate the given text into the specified destination language. The translated text is returned.

Configuration Class (CFG) and Model Initialization:

- The CFG class remains the same as in the previous code, holding configuration parameters. It initializes the Stable Diffusion model (image-gen-model) similarly to the previous code with the same parameters and settings.

Image Generation Function:

- The generate-image() function is unchanged from the previous code. It takes a prompt and the image generation model as inputs, generating an image based on the provided prompt using the configured model parameters.

Execution with Translation and Image Generation

- The code Calls get-translation() with a Telugu text and translates it into English.
- Uses the translated text as a prompt to generate an image using the initialized image-gen-model.

Summary of Changes:

- The primary divergence in this code compared to its predecessor is the switch in the translation package.
- The transformers library was previously employed, possibly leveraging a sophisticated model.
- In the current version, the code opts for the googletrans library, likely chosen for its simplicity, user-friendly interface, and possibly, better support for the specific translation requirements. This transition not only reflects adaptability but also hints at a preference for a lightweight and accessible solution in the language translation domain.
- Moreover, the adoption of the googletrans library suggests a strategic shift towards a more streamlined and accessible translation solution. The emphasis on user-friendly interactions and tailored support for specific translation needs underscores the code's responsiveness to practical considerations.

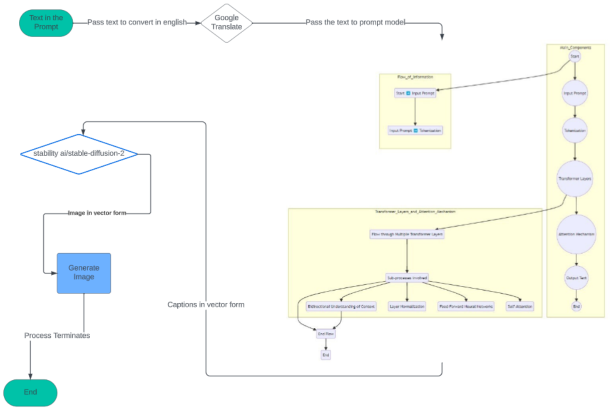


Figure 8. Flow chart of Multi-Lingual Text-To-Image Generation Model.

Working Of GPT-2:

- The provided mind map outlines the core elements and sequential flow of processes involved in the operation of GPT-2, a language model. Beginning with the initiation point marked as "Start," it proceeds with the user-provided "Input Prompt," which undergoes "Tokenization" for the model's comprehension. The journey involves traversing through "Transformer Layers" and an "Attention Mechanism," where information flows through multiple layers, encompassing sub-processes like "Self-Attention," "Feed-Forward Neural Networks," and "Layer Normalization." This allows GPT-2 to gain a "Bidirectional Understanding of Context." Eventually, the processed input leads to the generation of "Output Text." Finally, the sequence culminates at the "End," marking the conclusion of the entire process within the GPT-2 working principle, capturing the intricate stages from initial input to final text generation.

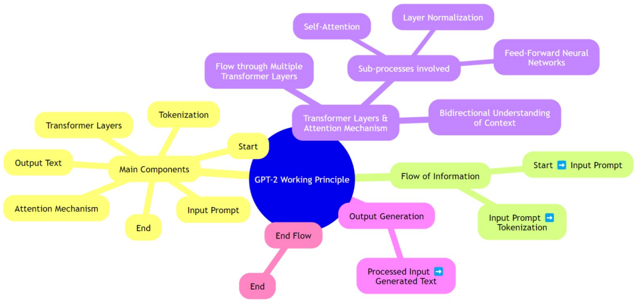


Figure 9. Mind Map of GPT-2

RESULT: The result presents a series of images generated through stable diffusion using distinct prompts.

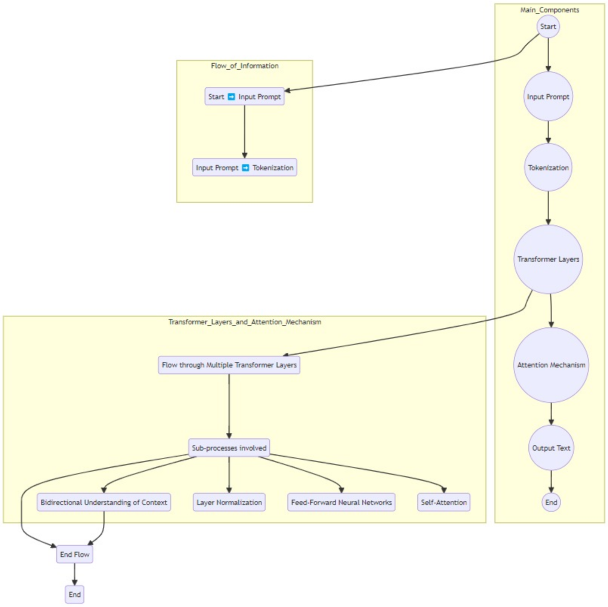
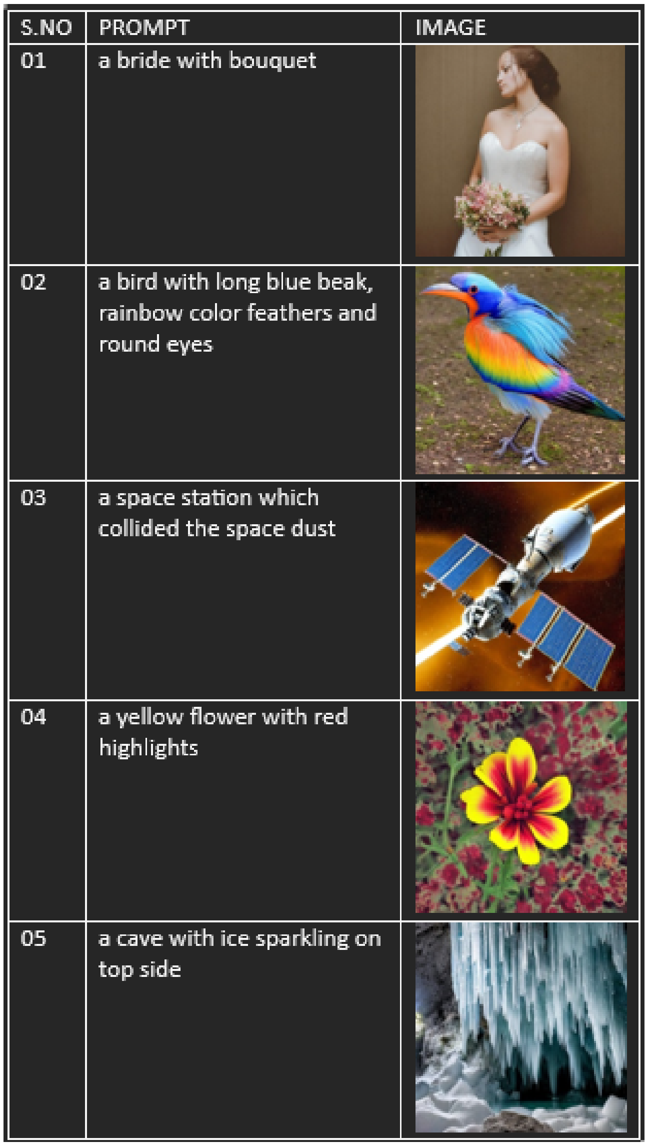


Figure 10. Flow chart of GPT-2.

- The first prompt led to the creation of a visually enchanting image capturing the timeless elegance of a bride adorned with a delicate bouquet.
- Subsequently, the second prompt unfolded a captivating scene featuring a bird with a striking blue long beak, vibrant rainbow-colored feathers, and enchanting round eyes, all expertly rendered through stable diffusion.
- The third prompt steered the generative process towards the cosmic realm, resulting in an awe-inspiring representation of a space station colliding with the ethereal dance of space dust.
- Moving on, the fourth prompt yielded a visually striking image of a yellow flower embellished with vivid red highlights, showcasing the versatility and precision achievable through stable diffusion.
- Lastly, the fifth prompt guided the algorithm to craft a mesmerizing depiction of a cave, its top side adorned with sparkling ice, demonstrating the capability of stable diffusion to generate intricate and visually compelling scenes across a spectrum of imaginative prompts.

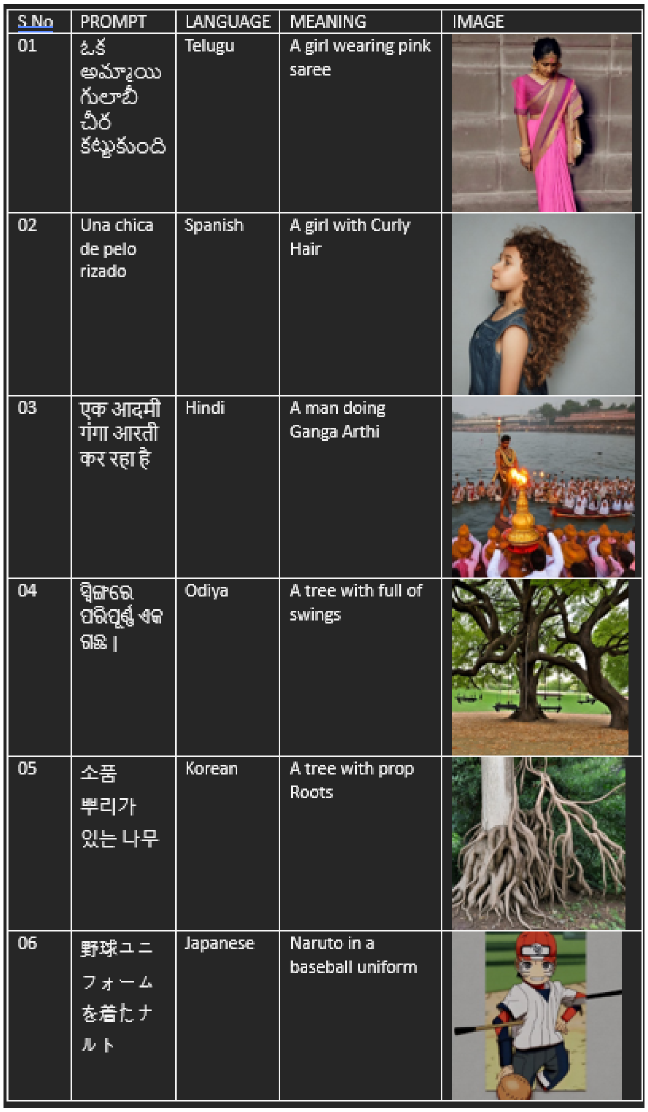
- In the meticulously curated table, each row is a nuanced narrative, weaving cultural richness through multilingual prompts. The table begins with a Telugu prompt, depicting a picturesque scene of "a girl wearing a pink saree." Transitioning to Spanish, the second row unfolds an image of "a girl with curly hair," while the Hindi prompt in the third row paints a vivid picture of "a man performing Ganga Aarti." The fourth row, adorned with an Odia prompt, captures the essence of "a tree with swings," and the Korean prompt in the fifth row adds a unique layer, describing "a tree with prop roots." Concluding with a touch of Japanese flair, the sixth row encapsulates the essence of "Naruto in a basketball uniform."



S.NO	PROMPT	IMAGE
01	a bride with bouquet	
02	a bird with long blue beak, rainbow color feathers and round eyes	
03	a space station which collided the space dust	
04	a yellow flower with red highlights	
05	a cave with ice sparkling on top side	

Figure 11. Results for Text-to-Image Generation when given a prompt.

- Here a table with five columns was created, consisting of prompts in various languages, a corresponding image generated by a multilingual model, the English translation of the prompt, and an image produced by a text-to-image model. The prompts included sentences about a soldier shooting another soldier (in German and English), a group selling panipuri with a smile (in Hindi and English), a girl eating chicken biryani (in



S.No	PROMPT	LANGUAGE	MEANING	IMAGE
01	ఒక అమ్మాయి గులాబీ చీర కట్టుకుంది	Telugu	A girl wearing pink saree	
02	Una chica de pelo rizado	Spanish	A girl with Curly Hair	
03	एक आदमी गंगा आरती कर रहा है	Hindi	A man doing Ganga Arthi	
04	ସ୍ୱିଙ୍ଗରେ ପରିପୂର୍ଣ୍ଣ ଏକ ଗଛ ।	Odiya	A tree with full of swings	
05	소품 뿌리가 있는 나무	Korean	A tree with prop Roots	
06	野球ユニフォームを着たナルト	Japanese	Naruto in a baseball uniform	

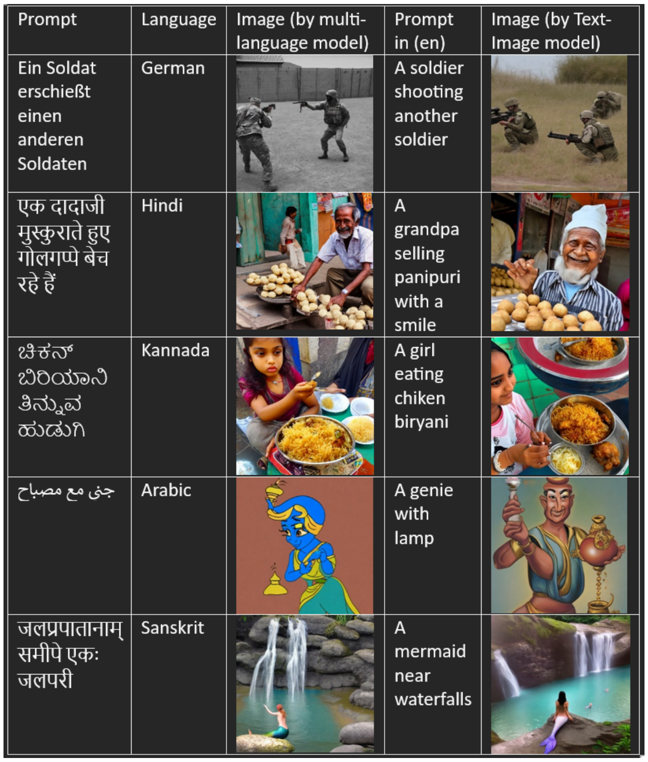
Figure 12. Results for Text-to-Image Generation when given a prompt in MultiLingual.

Kannada and English), a genie with a lamp (in Arabic and English), and a mermaid near waterfalls (in Sanskrit and English). Stable diffusion was employed to convert these multilingual prompts into diverse and meaningful images, showcasing the model’s ability to capture and express the same concept across different languages.

V MODEL EVALUATION

[1] Unlocking Precision in Stable Diffusion: The Superiority of IS and CLIP Metrics Over Conventional Accuracy and Precision Assessments:

- In the realm of stable diffusion, the choice of metrics such as Inception Score (IS) and CLIP (Contrastive Language-Image Pre-training) plays a pivotal role in assessing model accuracy. Unlike traditional metrics like accuracy and precision, IS and CLIP offers a more nuanced evaluation that aligns with the complex nature of diffusion processes. IS evaluates the quality and diversity of generated images, capturing the model’s ability to produce visually appealing and varied results. CLIP, on the other hand, introduces a semantic understanding by measuring the alignment between textual prompts and generated images, ensuring not just visual fidelity but also conceptual coherence.



Prompt	Language	Image (by multi-language model)	Prompt in (en)	Image (by Text-Image model)
Ein Soldat erschießt einen anderen Soldaten	German		A soldier shooting another soldier	
एक दादाजी मुस्कुराते हुए गोलगप्पे बेच रहे हैं	Hindi		A grandpa selling panipuri with a smile	
ಚಿಕನ್ ಬಿರಿಯಾನಿ ತಿನ್ನುವ ಹುಡುಗಿ	Kannada		A girl eating chiken biryani	
جني مع مصباح	Arabic		A genie with lamp	
जलप्रपातानाम् समीपे एकः जलपरी	Sanskrit		A mermaid near waterfalls	

Figure 13. Results for Text-to-Image Generation when given a prompt in MultiLingual with different Results.

- The decision to favor IS and CLIP over conventional metrics stems from the intricate challenges associated with stable diffusion. Accuracy and precision, although fundamental in many contexts, may fall short in encapsulating the holistic performance of diffusion models, particularly when dealing with diverse and intricate data patterns. IS and CLIP provide a more comprehensive lens, enabling a nuanced evaluation that aligns with the multi-faceted objectives of stable diffusion, where visual fidelity and semantic coherence are paramount[19].

[2] Process of finding Clip Score: Two stable diffusion pipelines are initialized using pre-trained models. The first pipeline is loaded from "CompVis/stable-diffusion-v1-4," and the second from "stabilityai/stable-diffusion-2."

- Image Generation:For each model, images are generated based on a set of diverse prompts.
- CLIP Score Calculation: The clip-score-fn function is defined using the CLIP model from the "openai/clip-vit-base-patch16" checkpoint. This function calculates the CLIP score, which measures the semantic alignment between generated images and textual prompts.
- Image Pre-processing:The generated images are preprocessed by converting pixel values to the uint8 format and adjusting their shape to fit the CLIP model's input requirements.
- CLIP Score Computation: CLIP scores are computed for the images generated by each stable diffusion model using the clip-score-fn function. The resulting scores represent how well the generated images align semantically with the given prompts.
- Results Display: The CLIP scores are then printed for each model, providing a quantitative measure of the model's ability to capture the intended meaning of the prompts.
- Seed-Controlled Randomness: To showcase the impact of seed-controlled randomness, the script sets a random seed before generating images with both models. This allows for a controlled comparison of the models' outputs under the same initial conditions. The utilization of a consistent seed ensures that variations observed in the generated images are attributed to the inherent characteristics of each model, facilitating a more precise evaluation of their performance. This controlled environment enhances the reliability of the comparative analysis, providing valuable insights into the nuanced behaviors and strengths of Stable Diffusion and other models in response to varying textual inputs.

[3] Result:

- CLIP Score with CompVis/stable-diffusion-V1-4: 34.8414
- CLIP Score with stabilityai/stable-diffusion-2: 35.5463

[4]Process of finding Inception Score:

- Preprocessing and Model Initialization: The code begins by defining a function, calculate-inception-score, which takes a set of generated images as input.A series of preprocessing transformations are applied to each image using the torchvision library. These include resizing to (299, 299) pixels, converting to a PyTorch tensor, and normalization.The InceptionV3 model is loaded from torchvision, with pretrained weights, and moved to the CUDA device.
- Image Preprocessing: The generated images, initially in NumPy array format, are converted to PIL images. These images are then preprocessed using the defined transformations to align with the input requirements of the InceptionV3 model.

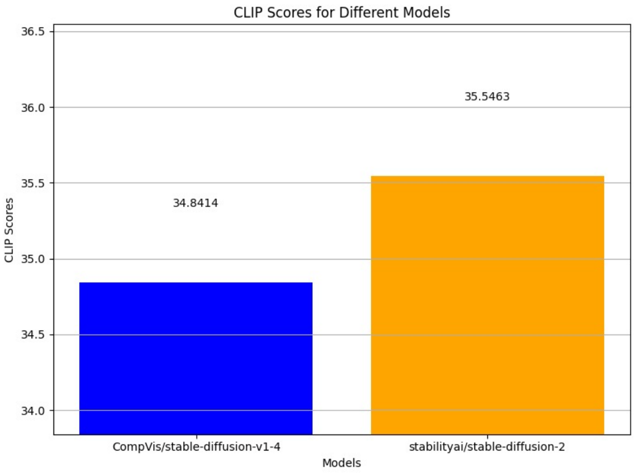


Figure 14. Clip Score for different models

- Inference and Probability Calculation: The preprocessed images are fed into the InceptionV3 model to obtain logits (raw scores) for each class. A softmax operation is applied to convert logits into class probabilities, facilitating a more interpretable representation of the model's confidence in each classification. This probability calculation step is essential for discerning the likelihood of different classes, enabling a nuanced understanding of the model's decision-making process. The resulting probability distributions serve as a foundation for further analysis and interpretation of the image classifications.
- Marginal and Conditional Distributions: The code calculates the marginal distribution by averaging the predicted probabilities across all generated images. This represents the distribution of classes across the entire set of images.The conditional distribution is the distribution of classes for each individual image in the generated set.
- KL Divergence Calculation: For each image, the code computes the Kullback-Leibler (KL) divergence between its conditional distribution and the marginal distribution. KL divergence measures the difference between two probability distributions, providing a quantitative assessment of the extent to which the conditional distribution diverges from the overall distribution. This analysis is crucial for evaluating the specificity of the model's predictions for individual images in relation to the broader dataset. The KL divergence values serve as informative metrics to gauge the model's focus on particular patterns or characteristics within the dataset, contributing to a comprehensive

understanding of its performance and decision-making dynamics

- Exponentiation of KL Divergence: The computed KL divergences are exponentiated (converted to the power of e) to emphasize the effect of large divergences. This exponential transformation amplifies the significance of divergence values, allowing for a more pronounced distinction between images with subtle differences in their conditional and marginal distributions. By exponentiating the KL divergences, the analysis accentuates instances where the model exhibits notable deviations in its predictive focus, enabling a finer-grained examination of the impact of these divergences on the overall interpretation of image classifications.
- Inception Score Calculation: The Inception Score is calculated as the average of the exponentiated KL divergences across all images. It provides a single scalar value that quantifies the quality and diversity of the generated images. Higher Inception Scores indicate better image quality and diversity.
- Results Display: Finally, the Inception Scores for the two sets of generated images, images and images-2, are printed.

[5] RESULT:

- Inception Score for images: 3.500929832458496
- Inception Score for images-2: 3.6218223571777344

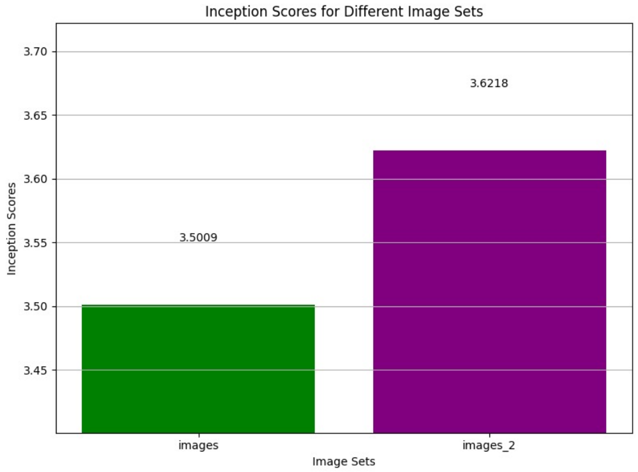


Figure 15. Inception Score for different models

In conclusion, the model is performing relatively well compared to other similar models. But MSE(mean square error), PPL score, SSIM(Structural Similarity Index) and PSNR score (Peak- signal to noise ratio) are not used here even though sd-pipelines for the models to be evaluated are gained. Because the images set formed from a few prompts for each model have different sizes it is difficult to test the measures for this, i.e., images set size = (6,512,512,3) and images-2 of our model has size = (6,578,578,12).

VI Comparative Analysis: Stable Diffusion vs. Other Text-to-Image Synthesis Models

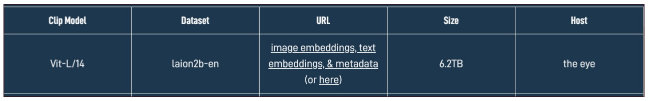
Aspect	Stable Diffusion	Other Models
Quality and Detail in Outputs	Excels in generating high-quality images with intricate details, emphasizing stability and coherence.	cGANs, BigGAN, StyleGAN/StyleGAN2 produce high-quality images but may not capture finer details as effectively.
Handling Ambiguity in Text Descriptions	Demonstrates better capabilities in interpreting and filling gaps within ambiguous descriptions.	cGANs and BigGAN may struggle with ambiguous descriptions, while StyleGAN/StyleGAN2 may not handle them as effectively.
Diversity in Generated Images	Offers broader diversity in generated images from the same text input, exploring the data distribution for varied outputs.	cGANs, BigGAN, and StyleGAN/StyleGAN2 also produce diverse outputs, but Stable Diffusion inherently promotes a wider spectrum of image variations.
Stability and Training Efficiency	More stable training dynamics and less prone to issues like mode collapse.	cGANs and BigGAN may face challenges in training stability, and StyleGAN/StyleGAN2 may encounter mode collapse under certain conditions.
Flexibility in Content Generation	Exhibits flexibility in creating a wide range of styles and abstractions, valuable for creative applications.	cGANs, BigGAN, and StyleGAN/StyleGAN2 are versatile, but Stable Diffusion's emphasis on controlled transformation enhances its flexibility.
Control Over Generation Process	Provides better control over the image generation process, allowing for more intuitive steering towards desired characteristics.	cGANs and BigGAN offer control, but Stable Diffusion's approach enhances precision in steering the generation process.
Efficiency in Scaling and Adaptation	Generally more efficient in scaling and adapting to different datasets and domains.	cGANs, BigGAN, and StyleGAN/StyleGAN2 may require more careful tuning and adaptation for different scenarios.
Energy Efficiency	Some studies suggest potential for greater energy efficiency compared to GANs, especially at scale.	cGANs, BigGAN, and StyleGAN/StyleGAN2 have made strides in efficiency, but Stable Diffusion's potential adds to its appeal.

In summary, Stable Diffusion stands out for its emphasis on stability, detailed image generation, and adaptability, making it a compelling choice in text-to-image synthesis. While other models like cGANs, BigGAN, CLIP, and StyleGAN/StyleGAN2 have their strengths, the specific advantages of Stable Diffusion position it as a promising alternative in various applications. The choice among these models should consider the specific requirements and objectives of the task at hand.

VII Curated Dataset: A Comprehensive Overview of the Text-to-Image Synthesis Benchmark Dataset

Dataset columns: The following features are present in the dataset :

- URL: the image url, millions of domains are covered
- TEXT: captions, in english for en, other languages for multi and nolang
- WIDTH: picture width
- HEIGHT: picture height
- LANGUAGE: the language of the sample, only for laion2B-multi, computed using cld3
- similarity: cosine between text and image ViT-B/32 embeddings, clip for en, mclip for multi and nolang
- pwatermark: probability of being a watermarked image, computed using our watermark detector
- punsafe: probability of being an unsafe image, computed using our clip based detector pwatermark and punsafe are available either as individual collections that must be joined with the hash of url+text, either as prejoined collections.



Clip Model	Dataset	URL	Size	Host
Vit-L/14	laion2b-en	image embeddings, text embeddings, & metadata (or here)	6.2TB	the eye

Figure 16. Dataset: LAION-2B

The acquisition pipeline can be split into three major components:

- Distributed processing of petabyte-scale Common Crawl dataset, which produces a collection of matching URLs and captions during the preprocessing phase. This meticulous process ensures a robust foundation for subsequent analysis. The curated dataset forms a rich resource for investigating patterns, trends, and insights within the vast expanse of online content.
- The distributed download of images based on shuffled data to pick a correct distribution of URLs, to avoid too heavy request loads on single websites
- Few GPU node post-processing of the data, which is much lighter and can be run in a few days, producing the final dataset.

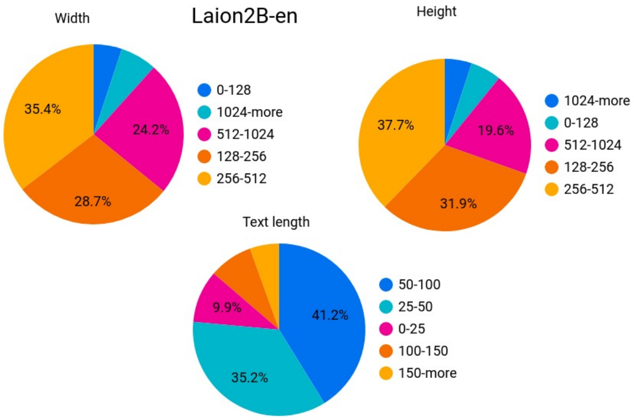


Figure 17. Image Distribution: LAION-2B

VIII Future scope

1. Realistic and Trustworthy: - Examine how cutting-edge ,Stable Diffusion to produce images with more fidelity and realism using generative models like GANs, with the goal of yields results that are quite similar to actual photos.

2. Controllable Generation: - Develop techniques to provide users with finer control over image generation, enabling manipulation of style, lighting, composition, and object placement within the scene, offering a more customizable experience.

3. Domain-Specific Adaptation: - Investigate tailoring Stable Diffusion for specific domains, such as medical imaging, fashion, or architecture, to enhance performance and address unique requirements in these specialized fields.

4. Creative Applications: - Explore unconventional applications of Stable Diffusion, including real-time storyboarding, interactive storytelling, and dynamic concept art generation, unlocking new avenues for creative expression.

5. Accessibility and Usability: - Prioritize improvements in the accessibility and usability of Stable Diffusion, making it more user-friendly for non-technical individuals and fostering broader adoption across diverse user bases.

6. Ethical Considerations: - Address ethical implications associated with text-to-image generation, including potential biases, copyright concerns, and impacts on human creativity, to ensure responsible and ethical use of the technology.

7. Combining with Other Modalities: - Explore the integration of Stable Diffusion with other modalities like natural language processing (NLP) and audio generation to create multimodal experiences that blend text, images, and sound seamlessly.

8. Large-Scale Datasets and Training: - Investigate the utilization of large-scale datasets and advanced training techniques to further enhance the capabilities of Stable Diffusion, enabling it to handle a wider range of prompts with increased expressiveness.

9. Hardware Optimization and Efficiency: - Explore techniques for hardware optimization and efficiency to make Stable Diffusion more computationally efficient, facilitating real-time generation and deployment on edge devices.

10. Open-Source Collaboration: - Foster open-source collaboration and community development to accelerate progress in text-to-image generation, encouraging the sharing of knowledge, tools, and advancements to maximize the collective impact of the research community.

By focusing on these research directions, the potential for Stable Diffusion in text-to-image generation is poised to bring about advancements that enhance creative expression, transform artistic processes, and push the boundaries of human-computer interaction.

IX Conclusion

In conclusion, this work presents a comprehensive exploration of Stable Diffusion from both human and AI perspectives. The proposed prompt engineering method showcases its efficacy in generating highly realistic face images. The examination of deep learning architectures reveals promising results in detecting and generalizing across a diverse population of over 600 individuals, despite some limitations in generalization. Notably, AI-based models exhibit superior performance compared to non-expert human users in preliminary experiments, highlighting their potential for accurate estimations on realistic and fake samples. The text-to-image generation model outlined in the provided information utilizes Stable Diffusion, employing packages such as ‘diffusers‘ and ‘transformers‘. We define a configuration class (‘CFG‘) to manage parameters and initialize the image generation model. A translation function is introduced, incorporating the ‘googletrans‘ package for multilingual capabilities. The results showcase visually compelling images generated from diverse prompts, underscoring the versatility of Stable Diffusion. Notable changes include our shift from the ‘transformers‘ library to the lightweight ‘googletrans‘ for translation, reflecting our preference for simplicity and adaptability. The implementation highlights the potential for creative applications and artistic expression, with suggestions for future refinement in optimization, exploration of advanced generative models, and consideration of ethical implications.

Furthermore, the study uncovers the disentanglement capability inherent in the stable diffusion model. This discovery leads to the development of a simple and lightweight disentanglement algorithm, showcasing its effectiveness in style matching and content preservation with minimal parameter optimization. With only 50 parameters optimized, the proposed method surpasses sophisticated baselines, demonstrating a generalizable disentanglement ability without the need for extensive fine-tuning on image editing tasks. Looking ahead, future developments aim to deepen the analysis of human performance, explore diverse generative techniques, and enhance detector robustness against new and unseen manipulations through innovative protocols. Overall, the findings underscore the promising synergy between Stable Diffusion and AI, offering avenues for advancements in realistic image generation and disentanglement tasks.

Limitations:

- The model does not achieve perfect photorealism .It cannot render legible text.It does not perform well on more difficult tasks which involve compositionality, such as rendering an image corresponding to “A red cube on top of a blue sphere”
- Faces and people in general may not be generated properly.
- The model was trained mainly with English captions and will not work as well in other languages.
- The autoencoding part of the model is lossy.It was trained on a subset of the large-scale dataset LAION-5B, which contains adult, violent and sexual content. To partially mitigate this, we have filtered the dataset using LAION's NFSW detector[18].

Bias:

- While the capabilities of image generation models are impressive, they can also reinforce or exacerbate social biases. Stable Diffusion was primarily trained on subsets of LAION-2B(en), which consists of images that are limited to English descriptions. Texts and images from communities and cultures that use other languages are likely to be insufficiently accounted for. This affects the overall output of the model, as white and western cultures are often set as the default. Further, the ability of the model to generate content with non-English prompts is significantly worse than with English-language prompts. Stable Diffusion v2 mirrors and exacerbates biases to such a degree that viewer discretion must be advised irrespective of the input or its intent[18].

X References:

1. Mathesul, Shubham, B. Ganesh, and Ayush. Rambhad, AttnGAN: Realistic Text-to-Image Synthesis with Attentional Generative Adversarial networks in IFIP Conference on Human-Computer Interaction, Springer International Publishing, 2021.
2. X. Tao, P. Zhang, Q. Hu, Z. Han, G. Zhe, H. Xiaolei, and H. Xiaodong,"AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks" in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.
3. S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, Generative Adversarial Text to Image Synthesis in University of Michigan, Ann Arbor, MI,USA (UMICH.EDU) and Max Planck Institute for Informatics, Saarbrucken,Germany (MPI-INF.MPG.DE)2016

4. Pengchuan Zhang2, Qiuyuan Huang2,Han Zhang3, Zhe Gan4, Image generation from text using AttnGAN Lehigh University, Microsoft Research, Rutgers University, Duke University, JD AI Research 2019.
5. H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks 2017
6. Alec. Radford, Luke. Metz, and Soumith.C, Self-Guided Feature Extraction using Deep Convolutional Generative Adversarial Networks 2016.
7. Kate Loginova Attention GAN in Natural Language Processing 2018
8. Ayan Sengupta Text to image generation Using Deep Convolution Generative Adversarial Networks (DCGANs), Coget Labs 2016
9. L. S. Hanne, R. Kundana, R. Thirukkumaran, Y. V. Parvatikar and K. Madhura, Text-To-Image Synthesis Using Modified GANs 2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), Chennai, India, 2022.
10. L. Papa, L. Faiella, L. Corvitto, L. Maiano, and I. Amerini, On the use of Stable Diffusion for creating realistic faces: from generation to detection Sapienza University of Rome, Italy,2023.
11. Q. Wu, Y. Liu, H. Zhao, A. Kale, T. Bui, T. Yu, Z. Lin, Y. Zhang, and S. Chang, Uncovering the Disentanglement Capability in Text-to-Image Diffusion Models UC, Santa Barbara, Adobe Research, MIT-IBM Watson AI Lab,qiucheng, yujianliu,2022.
12. Richard Gall Working principles of GAN'S 2018
13. Sai Rajeswar, Sandeep Subramanian, Francis Dutil, Christopher Pal , Aaron Courville, Adversarial Generation of Natural Language2017
14. J. Huang, M. Johanes, F. C. Kim, C. Doumpioti, and G.-C. Holz, On GANs, NLP and Architecture: Combining Human and Machine Intelligences for the Generation and Evaluation of Meaningful Designs2021
15. Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient Estimation of Word Representations in Vector Space In ICLR, September 2013.
16. Dong, A. The Enactment of Design through Language 2007
17. Ungureanu, L.-C. and T. Hartmann. Analysing Frequent Natural Language Expressions from Design Conversations Design Studies 2021
18. Hong, Susung, et al. Improving sample quality of diffusion models using self-attention guidance 2023.
19. Kim, Gwanghyun, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation 2022.

———————————