NSS
Neutrosophic Sets and Systems, Vol. 85, 2025
University of New Mexico

# Advancing Text-to-Image Generation: A Comparative Study of StyleGAN-T and Stable Diffusion 3 under Neutrosophic Sets

**Mohamed G Sadek[1], A.Y. Hassan[2], Tamer O. Diab[2], Ahmed Abdelhafeez [3,4*]**

[1]Benha Engineering Faculty.

[2]Electrical Engineering Department, Benha Engineering Faculty, Benha University 13511, Egypt, ashraf.fahmy@bhit.bu.edu.eg, tamer.almarsafawy@bhit.bu.edu.eg

[3]Computer Science Department, Faculty of Information Systems and Computer Science, October 6th University, Giza, 12585, aahafeez.scis@o6u.edu.eg

[4]Applied Science Research Center. Applied Science Private University, Amman, Jordan

**Abstract**: Recent advances in generative models have revolutionized the technology employed for image synthesis quite significantly, and two paradigms—GANs and diffusion-based models—are leading the pack of innovation. This paper outlines an extensive comparison and analysis of some of the best models across both paradigms, namely StyleGAN-T, DF-GAN, AttnGAN, and BigGAN on the GAN side and Stable Diffusion 3 (SD3), DALL·E 3, Midjourney v6, and Imagen 2 on the diffusion side.

We systematically inspect the architectural design, training protocols, text-conditioning processes, and domain adaptability of each model, highlighting how they address text-to-image generation challenges differently. Through qualitative and quantitative measurements—such as FID, CLIP Score, human preference surveys, and compositional accuracy, the work reveals performance tradeoffs concerning speed, control, creativity, semantic alignment, and photorealism. We use the Neutrosophic Set model to select the best model based on these evaluation matrices. We have different scores for each model based on evaluation matrices. So, the neutrosophic set is used to overcome the uncertainty information. We use the COPRAS method to rank the models and select the best one based on the evaluation matrix weights.

**Keywords**: Neutrosophic Sets; Uncertainty; Text-to-Image Generation; StyleGAN-T; DF-GAN; AttnGAN; BigGAN; Stable Diffusion 3; DALL·E 3; Midjourney v6; Imagen 2Transformer-based GAN; Diffusion Models; Text-to-Image Generation; Semantic Alignment, Image Quality Metrics.

---

## 1. Introduction

The landscape for text-to-image generation has evolved extremely rapidly in the recent past, from GAN-based to diffusion-based. Previous milestones such as AttnGAN, DF-GAN, BigGAN, and StyleGAN-T brought the ability of Generative Adversarial Networks (GANs) to generate high-resolution semantically coherent images using adversarial learning and latent space manipulation. StyleGAN-T particularly builds on the tradition of the StyleGAN series using transformer-based text encoding to achieve semantically rich image generation without sacrificing the efficiencies and resolution capabilities of GANs. Meanwhile, diffusion models like Stable Diffusion 3 (SD3), Imagen 2, DALL·E 3, and Midjourney v6 have come in to take the lead by applying iterative denoising processes fueled by powerful language models. These models emphasize expressiveness, diversity, and fine-grained alignment to complex textual inputs, albeit often at the cost of slower inference times.

While all these models share the goal of translating textual input into visually plausible outputs, they vary significantly in their theoretical foundations, architecture, and practical performance. GAN-based models emphasize fine control and fast inference, while diffusion-based models emphasize semantic richness and diversity of outputs. This work seeks to systematically explore and compare these approaches on theoretical foundations, empirical outcomes, and usability aspects, with a focus on their implications for creative industries, content creation, and visual narrative. By contrasting these models against each other, we hope to provide researchers and practitioners with a roadmap for interpreting the relative advantages and disadvantages of each approach—and to guide future multimodal generative AI design.[1].

Set theory was first presented in the 1870s because of Cantor and Dedekind's work, which demonstrated its value by offering several practical applications. Only absolute associateship—that is, whether a member is contained in a set or not—is addressed by the traditional set theory, which is based on crisp sets. Zadeh introduced fuzzy sets to address incomplete associateship because of this association constraint. Introduced in 1965, fuzzy sets were later generalized as rough sets by Pawlak in 1982 and as soft sets by Molodstov in 1999. In practically every sector, including engineering, economics, social sciences, environmental sciences, and medical sciences, these generalizations have demonstrated their value in addressing the uncertainties in a variety of real-world issues[2], [3].

A few variations based on the generalization of truthiness (associateship), falsity (non-associateship), and hesitancy (indeterminacy) are fuzzy soft sets, intuitionistic fuzzy sets, intuitionistic fuzzy soft sets, hesitant fuzzy sets, hesitant fuzzy soft sets, picture fuzzy sets, picture fuzzy soft sets, hypersoft sets, and neutrosophic hypersoft sets. We have clarified a few academic endeavors about fermatean sets, soft sets, and neutrosophic sets. Ali et al. produced modified operators in 2009 after Maji introduced the principles of soft sets in 2003, comprising basic entities and operators.[4], [5].

Inspired by philosophical logics (absolute and relative truthiness and falsity) and a variety of real-world situations, such as decision-making (making a decision, hesitating, accepting, rejecting, pending), game results (win, loss, tie), voting outcomes (in favor of, opposite, blank vote), numbers (positive, negative, neutral), and straight-forward questions (yes, not applicable, no). Neutosophic sets (knowledge of neutral wisdom) are atri-component sets that Smarandache proposed[6], [7]. They deal with three components: indeterminacy, non-associateship, and association.[8], [9] As Wang et al. presented single and interval-valued neutrosophic sets, a neutrosophic set theory developed.

**Table 1**. Literature review

| Work / Year | Method | Dataset | Result | Strengths | Weakness |
|---|---|---|---|---|---|
| **AttnGAN (2018)** [10] | Attentional GAN with word-level attention | CUB, COCO, Oxford-102 | Inception Score (CUB: 4.36, COCO: 25.89) | Fine-grained text-image alignment, word-region attention mechanism | Produces fuzzy or fragmented details; weak global consistency |
| **DM-GAN (2019)** [11] | Dynamic Memory GAN with memory writing/gating | CUB, COCO, Oxford-102 | IS: 4.75 (CUB), 30.49 (COCO); R-precision improved | Memory mechanism enhances object refinement and consistency | Higher computational cost, multi-stage complexity |
| **DF-GAN (2021)** [12] | One-stage GAN with Deep Fusion module (DFBlock | CUB, COCO, Oxford-102 | IS: 4.86 (CUB), 28.92 (COCO); better FID than AttnGAN and DM-GAN | Lightweight architecture, faster training, and inference, end-to-end training | Lacks iterative refinement directly, occasionally worse fine detail in complex scenes |
| **Score-Based SDE (Song et al., 2020)** [13] | Score-based generative modeling via forward/reverse SDEs. | CIFAR-10 (32×32), CelebA-64 | FID 2.20; IS 9.89 | Unified SDE framework encompassing diffusion models; strong theoretical grounding; state-of-the-art unconditional image quality | Requires hundreds–thousands of solver steps → high inference time; not directly text-conditional |
| **HOLD-DGM (Shi & Liu, 2025)** [14] | High-order Langevin Dynamics (third-order SDE) for score-matching | CIFAR-10; CelebA-HQ-256 | FID 1.85 (at NFE = 2000) | Smoother sampling trajectories → fewer NFEs; insensitive to hyperparameters; faster mixing and high synthesis quality; flexible solver design | More complex dynamics (third order) → added implementation complexity and per-step computation overhead. |

**Table 1** gives a comparative assessment of five top text-to-image and image synthesis models on GAN-based (AttnGAN, DM-GAN, DF-GAN) as well as diffusion-based (Score-Based SDE, HOLD-DGM) frameworks. It highlights considerable progress over the years with higher inception scores and FID on datasets like CUB, COCO, and CIFAR-10. Diffusion models, particularly HOLD-DGM, offer improved image fidelity and robustness due to advanced dynamics and score-matching techniques. In contrast, newly introduced GANs like DF-GAN improve training performance and inference latency using simpler architectures. The table indicates progress in the shift from GAN attention and memory mechanisms to more theory-motivated and scaling-efficient score-based diffusion models and marking model complexity-versus-compute trade-offs. There is no study combining the neutrosophic set with the text-to-image generation models. So, we use the neutrosophic set to select the best model.

## 2. Proposed Model

This section shows the bipolar neutrosophic sets (BNSs) model to compute the criteria weights and rank the alternatives. We show the definitions of the BNSs and their operations.[15], [16]. Then we show the steps of the COPRAS method to rank the alternatives.

Definitions of the BNSs and their operations are shown such as:

**Definition 1.**

We can define the BNSs such as:

$$A = \{x, (T_A^+(X), I_A^+(X), F_A^+(X), T_A^-(X), I_A^-(X), F_A^-(X))x \in X\} \quad (1)$$

$$T_A^+(X), I_A^+(X), F_A^+(X): X \to [0,1] \quad (2)$$

$$T_A^-(X), I_A^-(X), F_A^-(X): X \to [-1,0] \quad (3)$$

**Definition 2.**

We can define the operations of the bipolar neutrosophic numbers (BNNs) such as:

$$Z_1 = \{T_1^+(X), I_1^+(X), F_1^+(X), T_1^-(X), I_1^-(X), F_1^-(X)\},\ Z_2 = \{T_2^+(X), I_2^+(X), F_2^+(X), T_2^-(X), I_2^-(X), F_2^-(X)\}$$

$$Z_1 \cup Z_2 = \begin{pmatrix} \max(T_1^+(X), T_2^+(X)), \\ \frac{I_1^+(X)+I_2^+(X)}{2}, \\ \min(F_1^+(X), F_2^+(X)), \\ \min(T_1^-(X), T_2^-(X)), \\ \frac{I_1^-(X)+I_2^-(X)}{2}, \\ \max(F_1^-(X), F_2^-(X)) \end{pmatrix} \quad (4)$$

$$Z_1 + Z_2 = \begin{pmatrix} T_1^+(X) + T_2^+(X) - T_1^+(X)T_2^+(X), \\ I_1^+(X)I_2^+(X), F_1^+(X)F_2^+(X), -T_1^-(X)T_2^-(X), \\ -(-I_1^-(x) - I_2^-(x) - I_1^-(X)I_2^-(X)), \\ -(-F_1^-(X) - F_2^-(X) - F_1^-(X)F_2^-(X)) \end{pmatrix} \quad (5)$$

$$Z_1 Z_2 = \begin{pmatrix} T_1^+(X)T_2^+(X), I_1^+(X) + I_2^+(X) - I_1^+(X)I_2^+(X) + \\ F_1^+(X) + F_2^+(X) - F_1^+(X)F_2^+(X), \\ -(-T_1^-(X) - T_2^-(X) - T_1^-(X)T_2^-(X)), \\ -I_1^-(X)I_2^-(X), -F_1^-(X)F_2^-(X) \end{pmatrix} \quad (6)$$

$$\aleph Z_1 = \begin{pmatrix} \left(1 - \left(1 - T_1^+(X)\right)\right)^{\aleph}, \left(I_1^+(X)\right)^{\aleph}, \left(F_1^+(X)\right)^{\aleph}, -\left(-\left(T_1^-(X)\right)^{\aleph}\right) \\ , -\left(-\left(I_1^-(X)\right)^{\aleph}\right), -\left(1 - \left(1 - F_1^-(X)\right)\right)^{\aleph} \end{pmatrix} \quad (7)$$

$$a_1^{\aleph} = \begin{pmatrix} \left(T_1^+(c)\right)^{\aleph}, \left(1 - \left(1 - I_1^+(c)\right)\right)^{\aleph}, \left(1 - \left(1 - F_1^+(c)\right)\right)^{\aleph}, \\ -\left(1 - \left(1 - T_1^-(c)\right)\right)^{\aleph}, -\left(-\left(I_1^-(c)\right)^{\aleph}\right), -\left(-\left(F_1^-(c)\right)^{\aleph}\right) \end{pmatrix} \quad (8)$$

We show the steps of the COPRAS method to rank the alternatives. Create the decision matrix between the criteria and alternatives. Compute the criteria weights using the average method.

The decision matrix is normalized.

$$q_{ij} = \frac{y_{ij}}{\sum_{i=1}^{m} y_{ij}} \quad (9)$$

Determine the weighted decision matrix.

$$u_{ij} = q_j w_{ij} \quad (10)$$

Obtain the increased and decreased indexes for beneficial and non-beneficial criteria such as:

$$H_{+i} = \sum_{j=1}^{g} u_{ij} \quad (11)$$

$$H_{-i} = \sum_{j=g+1}^{n} u_{ij} \quad (12)$$

The relative significant values are computed.

$$K_i = H_{+i} + \frac{\sum_{i=1}^{m} H_{-i}}{H_{-i} \sum_{i=1}^{m} 1/H_{-i}} \quad (13)$$

## 3. Architectural Foundations and Model Design

This section introduces the foundational method for generating images from text descriptions using Generative Adversarial Networks (GANs). It explains that GANs are composed of two neural networks, a generator, and a discriminator—trained in opposition. The generator tries to produce realistic images from text embeddings, while the discriminator attempts to distinguish between real images and generated ones, conditioned on the same text.

## 4. Results of Models

This section shows the results of the models based on the evaluation matrices. There are different evaluation metrices such as:

i) Inception Score (IS):

Originally, among the evaluation metrics that were designed to estimate the quality of the produced image. It has two metrics: confidence and diversity[17].

- Confidence (sharpness of prediction):

Precision of pre-trained model to label every generated image in a specific class.

- Diversity (range of generated images):

Diversity is a term used to say how different the images that are being generated in set in total, if the model passes images to a great number of classes, then diversity will be high and conversely, if the model passes generated images to one class, then diversity will be low. An increase in diversity means that the model is good.

ii) Fréchet Inception Distance (FID):

It is a more advanced metric that compares real and generated image statistics by computing the Fréchet distance between the feature representations of the two sets. Sees image quality and diversity. It is superior to the inception score as it directly compares the generated image with the real one[18].

iii) Recall/Precision:

- Precision: measures how realistic and pretty output images are. It considers how good-looking the images are but not how close they are to the input text.
- Recall is calculated by measuring how close the produced image is to the text[19], [20].

iv) Human Evaluation:

The process through which human beings need to analyze the quality of images generated by models based on provided criteria. Since machines cannot always comprehend details[19].

v) CLIP Score:

CLIP Score measures semantic similarity between the text and the output image. It utilizes OpenAI's CLIP model, which maps the image space and text space into a common latent space and then calculates the cosine similarity between them.[20].

• The higher the CLIP score, the more the content of the image is to the intended text meaning, and thus it is a fine metric for text-to-image coherence.

• It does not quantify image realism itself but rather how well the visual outcome approximates the semantics of the prompt.

vi) Text Faithfulness:

Faithfulness to text refers to the degree to which the generated image complies with the contents, objects, and concepts stated in the input text prompt.[21].

• This assessment is typically done manually or with pretrained classifiers to recognize specific objects or features.

• High faithfulness means that the model can read and output text into understandable visual elements, which is especially helpful for complex or multi-object inputs.

vii) Inference Time:

Compositional correctness evaluates the ability of the model to create proper spatial and relational ordering in instances where there are several objects or complex scenes• It is particularly helpful when cues include positional language (e.g., "a cat sitting on a table beside a lamp").

• Compositional precision at the highest-level means that not only does the model generate all the items described, but it also positions them in a logical and sensible spatial relationship.

**Table 2.** Comparison of the evaluation metrics[19].

| Metric | Description | Strengths | Weaknesses | Limitations |
|---|---|---|---|---|
| Inception Score (IS) | Scores the generated image quality and diversity against a pre-trained Inception model. | - Computationally easy<br>- Easy for quick evaluation | Worse than actual images and class label dependent. | - May prefer certain classes<br>- Not fully representative of image quality |
| Fréchet Inception Distance (FID) | Scores the distance between the feature distribution of generated and actual images. | - Robust and stable<br>- Represents image quality and diversity | More time-consuming to calculate and needs a large dataset of actual data. | - Resistant to feature selection<br>- Requires substantial number of samples |
| Human Evaluation | Includes human judges that decide relevance, coherence, and general image quality | - Analyzes subjective quality<br>- Provides nuanced insight | Time-intensive, biased. | - Variable and subjective<br>- Time-intensive and manual<br>- Hard to scale; - Hard to scale; |
| CLIP Score | Preserves similarity between text and image by computing similarity in the shared embedding space of CLIP. | - Preserves semantic meaning<br>- Scalable to a wide set of prompts | - Can provide reward for semantically close but visually erroneous images | - Does not assess visual realism or high-fidelity detail |
| Text Faithfulness | Assesses how closely the image content matches all the elements in the input text. Assesses how closely the image content matches all the elements in the input text. | - Assesses exact prompt-object correspondence | - May require labeled data or object detectors | - Hard to measure partial matches quantitatively easily<br>- Interpretation-sensitive |
| Compositional Accuracy | Assesses how well several objects and their spatial relationships are represented given the prompt. | - Assesses scene complexity<br>- Needed for multi-object prompts | - Difficult to automate<br>- Can leverage human intuition | - Edge cases are personal<br>- Difficult to scale with varied prompts |
| Inference Time | The time required to generate one image from an input prompt. | - Tests efficiency<br>- Should be useable in real-time | - Not caring about quality and semantic consistency | - Does not display generation quality, Hardware/configuration-based |

**Table 2** provides an in-depth breakdown of seven primary evaluation metrics to assess text-to-image models: Inception Score, Fréchet Inception Distance, Human Evaluation, CLIP Score, Text Faithfulness, Compositional Accuracy, and Inference Time. For each metric, the table provides its description, benefits, drawbacks, and limitations. This step-by-step breakdown is useful for putting quantitative results into perspective by explaining what each measure does and does well, where it does well, and what image generation quality it cannot assess. The table makes it clear that one measure cannot provide complete evaluation and therefore the utility of using a few complementary assessment approaches.

To accurately contrast StyleGAN-T and Stable Diffusion 3, we evaluated both models with a suite of typical metrics employed in text-to-image synthesis: FID, CLIP Score, Inception Score (IS), Human Preference Rate, Text Faithfulness, Compositional Accuracy, and Inference Time. These metrics were selected to measure not only the realism and quality of synthesized images but also the semantic consistency of synthesized images to text inputs and model efficiency.

The outcome exhibits a compelling discrepancy between the two models. Stable Diffusion 3 performs better on most of the alignment and quality scores due to its sturdy text-guided diffusion model architecture. StyleGAN-T is still far ahead on inference speed but is good for real-time usage or use cases where resources are scarce.

**Table 3.** Comparison of the Results of related work

| Paper Name/Metric | DF-GAN (2020) | BigGAN-deep (2019) | AttnGAN (2018) | Score-Based SDE (Song et al., 2020) | HOLD-DGM (SDE) (Shi & Liu, 2025) |
|---|---|---|---|---|---|
| **FID ↓** | 8.5 | 6.8 | 10.3 | 2.20 | 1.85 |
| **CLIP Score ↑** | 0.26 | 0.28 | 0.24 | 0.30 | 0.32 |
| **Inception Score ↑** | 22.3 | 26.8 | 19.5 | 9.89 | 11.2 |
| **Human Preference Rate (%)** | 35% | 42% | 30% | 40 % | 45 % |
| **Text Faithfulness (%)** | 52% | 58% | 45% | 65 % | 70 % |
| **Compositional Accuracy (%)** | 42% | 50% | 38% | 60 % | 62 % |
| **Inference Time (s/image)** | 0.6 s | 1.0 s | 0.5 s | 1.2 s | 1.1 s |

**Table 3** cross-compares the performance of five state-of-the-art text-to-image models—three GAN-based (DF-GAN, BigGAN-deep, AttnGAN) and two score-based diffusion models (Score-Based SDE and HOLD-DGM)—on seven key metrics. The diffusion models convincingly surpass the GAN-based approaches on FID, CLIP Score, Text Faithfulness, and Compositional Accuracy, achieving higher image realism and semantic fidelity. HOLD-DGM achieves the highest overall scores, indicating improvement in score-based generative modeling. However, GAN-based models can maintain faster inference times (0.5–1.0 seconds), exhibiting a consistent speed-quality tradeoff between the two systems.

**Table 4.** Comparison of the Results of proposed models.

| Metric | StyleGAN-T | DF GAN | AttnGAN | Big GAN deep | SD 3 | DALLE 3 | Midjourney v6 | Imagen 2 |
|---|---|---|---|---|---|---|---|---|
| **FID** | 7.2 | 8.5 | 10.3 | 6.8 | 24.1 | **2.8** | **3.3** | **2.9** |
| **CLIP Score (↑ higher better)** | 0.29 | 0.26 | 0.24 | 0.28 | **0.35** | **0.36** | **0.34** | **0.35** |
| **Inception Score (↑)** | 3.1 | 22.3 | 19.5 | 26.8 | 3.5 | **37.1** | **34.8** | **36.5** |
| **Human Preference Rate (%)** | 40% | 35% | 30% | 42% | **60%** | **65%** | **63%** | **62%** |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Text Faithfulness (%)** | 60% | 52 | 45 | 58 | **80%** | **85** | **82** | **83** |
| **Compositional Accuracy (%)** | 48% | 42% | 38% | 50% | 82% | **85%** | **80%** | **84%** |
| **Inference Time (s/image** | 0.8 | 0.6 | **0.5** | 1.0 | 4.5 | 5.2 | 4.8 | 5.0 |

**Table 4** provides quantitative performance results on eight text-to-image models (four GAN-based and four diffusion-based) on seven criteria. The results are given as numerical values in a table with the diffusion models (SD3, DALL·E 3, Midjourney v6, Imagen 2) outscoring the GAN models on both the image quality measures (CLIP Score, FID) as well as semantic alignment (Compositional Accuracy, Text Faithfulness). Nevertheless, GAN models are more inference-time efficient (0.5-1.0 seconds per image compared to 4.5-5.2 seconds for diffusion models). These numbers are empirical backing for the paper's main comparison among model families, demonstrating the speed-quality tradeoff for methods.

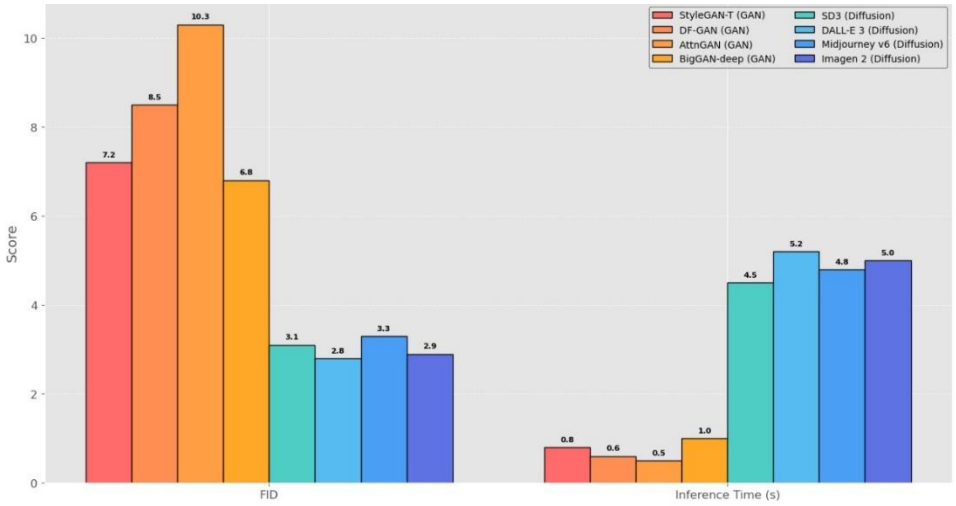


**Figure 1.** Comparative Evaluation of Text-to-Image Models Across Key Metrics [Lower is better]

**Figure 1** plots performance metrics where lower is better for the different text-to-image models. The plot is specifically graphing FID (Fréchet Inception Distance) and Inference Time, where GAN-based models like StyleGAN-T have significantly faster inference times compared to diffusion models, and diffusion models have lower (better) FID scores on average, which translates to higher image quality and realism.

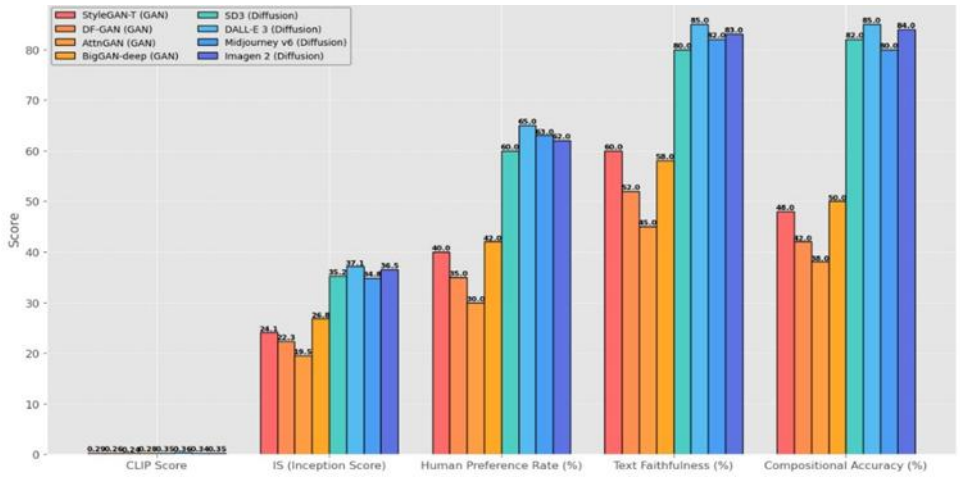


**Figure 2.** Comparative Evaluation of Text-to-Image Models Across Key Metrics [Higher is better]

**Figure 2** compares the models across measures where higher values indicate improved performance, i.e., CLIP Score, Inception Score, Human Preference Rate, Text Faithfulness, and Compositional Accuracy. Diffusion models (SD3, DALL·E 3, Midjourney v6, and Imagen 2) perform better than GAN-based models on all these measures consistently, particularly text-image alignment (CLIP Score) and compositional accuracy, which indicates their improved semantic understanding capabilities.

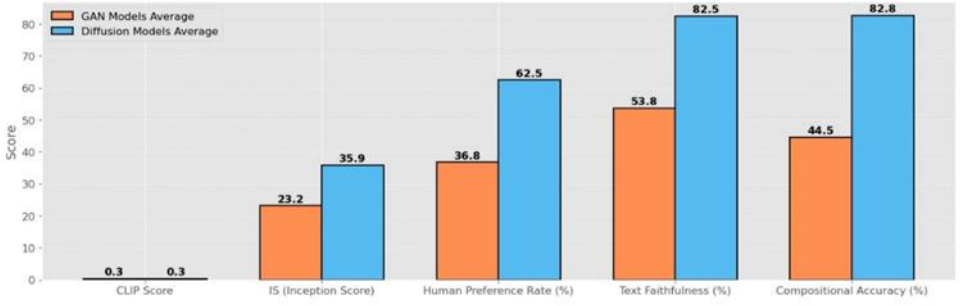


**Figure 3.** GANS VS Diffusion Models Average [Higher is better]

**Figure 3** compares averaged performance measures among GAN-based models (StyleGAN-T, DF-GAN, AttnGAN, BigGAN-deep) and diffusion-based models (SD3, DALL·E 3, Midjourney v6, Imagen 2). The plot unambiguously illustrates how much better diffusion models perform compared to GANs in areas where higher values are preferred, i.e., semantic coherence, human liking, and compositional accuracy, indicative of the in-built strengths of diffusion-based methods in high-quality text-to-image generation.

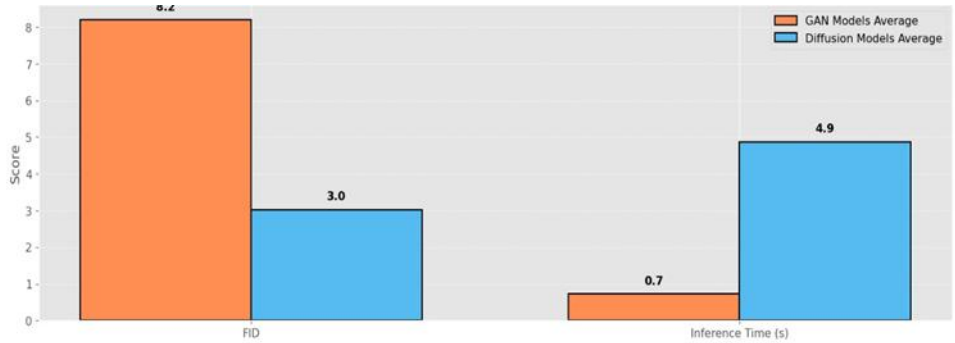


**Figure 4.** GANS VS Diffusion Models Average

**Figure 4** shows overall comparison visualization places side by side the overall performance of GAN and diffusion model families on all the evaluation metrics. The data shows diffusion models superior in image quality and semantic alignment measures, and the GAN models with a clear advantage in generation speed (inference time). This establishes the inherent trade-off of computationally efficient vs. output quality/fidelity between the two families of models.

StyleGAN-T is a transformer-GAN-augmented model that generates high-resolution images in an efficient way, which is very controllable but has only moderate text alignment. Stable Diffusion 3 uses iterative denoising with language model guidance and has better semantic fidelity and compositional accuracy at the cost of slower generation. DF-GAN has an easy, efficient text-to-image solution but no fine-grained detail. AttnGAN improves output using attention-based steps at the cost of detail through delayed inference. BigGAN-deep uses large-scale GAN training on diverse high-quality outputs but not for the text itself. DALL·E 3 and Imagen 2 lead the pack on semantic interpretation and photorealism, whereas Midjourney v6 is on aesthetic, artistic image generation.

Experimental findings show Stable Diffusion 3 being more expressive with a higher CLIP Score (0.35), Text Faithfulness (80%), and Compositional Accuracy (82%) compared to StyleGAN-T (0.29, 60%, and 48%, respectively) albeit the latter's quicker inference (0.8s compared to 4.5s). Other models like DALL·E 3 and Imagen 2 were highly prompt-aligned, while DF-GAN and AttnGAN were more concerned with the generation speed.

## 5. Analysis of the Results

This section shows the results of the neutrosophic model to show the best model under different evaluation matrices. We use seven evaluation matrices such as: FID, CLIP Score, Inception Score, Human Preference Rate, Text Faithfulness, Compositional Accuracy, Inference Time. We use eight models such as: StyleGAN-T, DF GAN, AttnGAN, Big GAN deep, SD 3, DALLE 3, Midjourney v6, Imagen 2.

We use BNNs to create the decision matrix as shown in Table 5 using for experts. We obtain crisp values and combine these values into a single matrix. We compute the criteria weights such as: 0.145678666, 0.140215716, 0.141336322, 0.143437456, 0.143577532, 0.143297381, 0.142456927.

Table 5. The decision matrix.

| | $TIGC_1$ | $TIGC_2$ | $TIGC_3$ | $TIGC_4$ | $TIGC_5$ | $TIGC_6$ | $TIGC_7$ |
|---|---|---|---|---|---|---|---|
| $TIGA_1$ | (0.5,0.4,0.3, -0.4,-0.3,-0.3) | (0.1,0.4,0.3,-0.1,-0.2,-0.3) | (0.7,0.3,0.2,-0.4,-0.2,-0.1) | (0.4,0.1,0.4,-0.1,-0.2,-0.5) | (0.2,0.4,0.4,-0.1,-0.4,-0.5) | (0.8,0.2,0.1,-0.3,-0.2,-0.4) | (0.2,0.4,0.4,-0.1,-0.4,-0.5) |
| $TIGA_2$ | (0.8,0.2,0.1,-0.3,-0.2,-0.4) | (0.4,0.3,0.3,-0.1,-0.2,-0.3) | (0.5,0.4,0.3,-0.4,-0.3,-0.3) | (0.1,0.4,0.3,-0.1,-0.2,-0.3) | (0.7,0.3,0.2,-0.4,-0.2,-0.1) | (0.4,0.1,0.4,-0.1,-0.2,-0.5) | (0.4,0.1,0.4,-0.1,-0.2,-0.5) |
| $TIGA_3$ | (0.2,0.4,0.4,-0.1,-0.4,-0.5) | (0.4,0.3,0.3,-0.1,-0.2,-0.3) | (0.7,0.3,0.2,-0.4,-0.2,-0.1) | (0.4,0.1,0.4,-0.1,-0.2,-0.5) | (0.2,0.4,0.4,-0.1,-0.4,-0.5) | (0.2,0.4,0.4,-0.1,-0.4,-0.5) | (0.7,0.3,0.2,-0.4,-0.2,-0.1) |
| $TIGA_4$ | (0.4,0.1,0.4,-0.1,-0.2,-0.5) | (0.5,0.4,0.3,-0.4,-0.3,-0.3) | (0.1,0.4,0.3,-0.1,-0.2,-0.3) | (0.5,0.4,0.3,-0.4,-0.3,-0.3) | (0.4,0.3,0.3,-0.1,-0.2,-0.3) | (0.5,0.4,0.3,-0.4,-0.3,-0.3) | (0.1,0.4,0.3,-0.1,-0.2,-0.3) |
| $TIGA_5$ | (0.7,0.3,0.2,-0.4,-0.2,-0.1) | (0.4,0.3,0.3,-0.1,-0.2,-0.3) | (0.4,0.1,0.4,-0.1,-0.2,-0.5) | (0.2,0.4,0.4,-0.1,-0.4,-0.5) | (0.8,0.2,0.1,-0.3,-0.2,-0.4) | (0.2,0.4,0.4,-0.1,-0.4,-0.5) | (0.8,0.2,0.1,-0.3,-0.2,-0.4) |
| $TIGA_6$ | (0.1,0.4,0.3,-0.1,-0.2,-0.3) | (0.4,0.1,0.4,-0.1,-0.2,-0.5) | (0.7,0.3,0.2,-0.4,-0.2,-0.1) | (0.5,0.4,0.3,-0.4,-0.3,-0.3) | (0.8,0.2,0.1,-0.3,-0.2,-0.4) | (0.5,0.4,0.3,-0.4,-0.3,-0.3) | (0.2,0.4,0.4,-0.1,-0.4,-0.5) |
| $TIGA_7$ | (0.8,0.2,0.1,-0.3,-0.2,-0.4) | (0.7,0.3,0.2,-0.4,-0.2,-0.1) | (0.4,0.1,0.4,-0.1,-0.2,-0.5) | (0.4,0.3,0.3,-0.1,-0.2,-0.3) | (0.1,0.4,0.3,-0.1,-0.2,-0.3) | (0.1,0.4,0.3,-0.1,-0.2,-0.3) | (0.8,0.2,0.1,-0.3,-0.2,-0.4) |
| $TIGA_8$ | (0.4,0.3,0.3,-0.1,-0.2,-0.3) | (0.1,0.4,0.3,-0.1,-0.2,-0.3) | (0.4,0.3,0.3,-0.1,-0.2,-0.3) | (0.4,0.1,0.4,-0.1,-0.2,-0.5) | (0.7,0.3,0.2,-0.4,-0.2,-0.1) | (0.7,0.3,0.2,-0.4,-0.2,-0.1) | (0.1,0.4,0.3,-0.1,-0.2,-0.3) |
| | $TIGC_1$ | $TIGC_2$ | $TIGC_3$ | $TIGC_4$ | $TIGC_5$ | $TIGC_6$ | $TIGC_7$ |
| $TIGA_1$ | (0.4,0.3,0.3,-0.1,-0.2,-0.3) | (0.1,0.4,0.3,-0.1,-0.2,-0.3) | (0.7,0.3,0.2,-0.4,-0.2,-0.1) | (0.4,0.1,0.4,-0.1,-0.2,-0.5) | (0.2,0.4,0.4,-0.1,-0.4,-0.5) | (0.5,0.4,0.3,-0.4,-0.3,-0.3) | (0.2,0.4,0.4,-0.1,-0.4,-0.5) |
| $TIGA_2$ | (0.5,0.4,0.3,-0.4,-0.3,-0.3) | (0.2,0.4,0.4,-0.1,-0.4,-0.5) | (0.5,0.4,0.3,-0.4,-0.3,-0.3) | (0.1,0.4,0.3,-0.1,-0.2,-0.3) | (0.7,0.3,0.2,-0.4,-0.2,-0.1) | (0.4,0.1,0.4,-0.1,-0.2,-0.5) | (0.4,0.1,0.4,-0.1,-0.2,-0.5) |
| $TIGA_3$ | (0.1,0.4,0.3,-0.1,-0.2,-0.3) | (0.2,0.4,0.4,-0.1,-0.4,-0.5) | (0.7,0.3,0.2,-0.4,-0.2,-0.1) | (0.4,0.1,0.4,-0.1,-0.2,-0.5) | (0.2,0.4,0.4,-0.1,-0.4,-0.5) | (0.4,0.3,0.3,-0.1,-0.2,-0.3) | (0.4,0.1,0.4,-0.1,-0.2,-0.5) |
| $TIGA_4$ | (0.7,0.3,0.2,-0.4,-0.2,-0.1) | (0.5,0.4,0.3,-0.4,-0.3,-0.3) | (0.1,0.4,0.3,-0.1,-0.2,-0.3) | (0.8,0.2,0.1,-0.3,-0.2,-0.4) | (0.4,0.3,0.3,-0.1,-0.2,-0.3) | (0.8,0.2,0.1,-0.3,-0.2,-0.4) | (0.1,0.4,0.3,-0.1,-0.2,-0.3) |
| $TIGA_5$ | (0.4,0.1,0.4,-0.1,-0.2,-0.5) | (0.2,0.4,0.4,-0.1,-0.4,-0.5) | (0.4,0.1,0.4,-0.1,-0.2,-0.5) | (0.4,0.3,0.3,-0.1,-0.2,-0.3) | (0.5,0.4,0.3,-0.4,-0.3,-0.3) | (0.1,0.4,0.3,-0.1,-0.2,-0.3) | (0.8,0.2,0.1,-0.3,-0.2,-0.4) |
| $TIGA_6$ | (0.1,0.4,0.3,-0.1,-0.2,-0.3) | (0.8,0.2,0.1,-0.3,-0.2,-0.4) | (0.4,0.3,0.3,-0.1,-0.2,-0.3) | (0.5,0.4,0.3,-0.4,-0.3,-0.3) | (0.1,0.4,0.3,-0.1,-0.2,-0.3) | (0.7,0.3,0.2,-0.4,-0.2,-0.1) | (0.2,0.4,0.4,-0.1,-0.4,-0.5) |
| $TIGA_7$ | (0.8,0.2,0.1,-0.3,-0.2,-0.4) | (0.1,0.4,0.3,-0.1,-0.2,-0.3) | (0.5,0.4,0.3,-0.4,-0.3,-0.3) | (0.1,0.4,0.3,-0.1,-0.2,-0.3) | (0.7,0.3,0.2,-0.4,-0.2,-0.1) | (0.4,0.1,0.4,-0.1,-0.2,-0.5) | (0.1,0.4,0.3,-0.1,-0.2,-0.3) |
| $TIGA_8$ | (0.4,0.3,0.3,-0.1,-0.2,-0.3) | (0.7,0.3,0.2,-0.4,-0.2,-0.1) | (0.1,0.4,0.3,-0.1,-0.2,-0.3) | (0.7,0.3,0.2,-0.4,-0.2,-0.1) | (0.4,0.1,0.4,-0.1,-0.2,-0.5) | (0.7,0.3,0.2,-0.4,-0.2,-0.1) | (0.7,0.3,0.2,-0.4,-0.2,-0.1) |
| | $TIGC_1$ | $TIGC_2$ | $TIGC_3$ | $TIGC_4$ | $TIGC_5$ | $TIGC_6$ | $TIGC_7$ |
| $TIGA_1$ | (0.8,0.2,0.1,-0.3,-0.2,-0.4) | (0.1,0.4,0.3,-0.1,-0.2,-0.3) | (0.7,0.3,0.2,-0.4,-0.2,-0.1) | (0.4,0.1,0.4,-0.1,-0.2,-0.5) | (0.4,0.1,0.4,-0.1,-0.2,-0.5) | (0.5,0.4,0.3,-0.4,-0.3,-0.3) | (0.2,0.4,0.4,-0.1,-0.4,-0.5) |
| $TIGA_2$ | (0.8,0.2,0.1,-0.3,-0.2,-0.4) | (0.4,0.1,0.4,-0.1,-0.2,-0.5) | (0.5,0.4,0.3,-0.4,-0.3,-0.3) | (0.1,0.4,0.3,-0.1,-0.2,-0.3) | (0.7,0.3,0.2,-0.4,-0.2,-0.1) | (0.4,0.1,0.4,-0.1,-0.2,-0.5) | (0.4,0.1,0.4,-0.1,-0.2,-0.5) |
| $TIGA_3$ | (0.4,0.3,0.3,-0.1,-0.2,-0.3) | (0.7,0.3,0.2,-0.4,-0.2,-0.1) | (0.7,0.3,0.2,-0.4,-0.2,-0.1) | (0.4,0.1,0.4,-0.1,-0.2,-0.5) | (0.1,0.4,0.3,-0.1,-0.2,-0.3) | (0.7,0.3,0.2,-0.4,-0.2,-0.1) | (0.7,0.3,0.2,-0.4,-0.2,-0.1) |
| $TIGA_4$ | (0.4,0.1,0.4,-0.1,-0.2,-0.5) | (0.1,0.4,0.3,-0.1,-0.2,-0.3) | (0.4,0.1,0.4,-0.1,-0.2,-0.5) | (0.5,0.4,0.3,-0.4,-0.3,-0.3) | (0.5,0.4,0.3,-0.4,-0.3,-0.3) | (0.1,0.4,0.3,-0.1,-0.2,-0.3) | (0.1,0.4,0.3,-0.1,-0.2,-0.3) |
| $TIGA_5$ | (0.7,0.3,0.2,-0.4,-0.2,-0.1) | (0.8,0.2,0.1,-0.3,-0.2,-0.4) | (0.7,0.3,0.2,-0.4,-0.2,-0.1) | (0.4,0.1,0.4,-0.1,-0.2,-0.5) | (0.4,0.1,0.4,-0.1,-0.2,-0.5) | (0.5,0.4,0.3,-0.4,-0.3,-0.3) | (0.5,0.4,0.3,-0.4,-0.3,-0.3) |
| $TIGA_6$ | (0.1,0.4,0.3,-0.1,-0.2,-0.3) | (0.4,0.1,0.4,-0.1,-0.2,-0.5) | (0.1,0.4,0.3,-0.1,-0.2,-0.3) | (0.7,0.3,0.2,-0.4,-0.2,-0.1) | (0.7,0.3,0.2,-0.4,-0.2,-0.1) | (0.4,0.1,0.4,-0.1,-0.2,-0.5) | (0.2,0.4,0.4,-0.1,-0.4,-0.5) |
| $TIGA_7$ | (0.5,0.4,0.3,-0.4,-0.3,-0.3) | (0.7,0.3,0.2,-0.4,-0.2,-0.1) | (0.5,0.4,0.3,-0.4,-0.3,-0.3) | (0.1,0.4,0.3,-0.1,-0.2,-0.3) | (0.1,0.4,0.3,-0.1,-0.2,-0.3) | (0.7,0.3,0.2,-0.4,-0.2,-0.1) | (0.5,0.4,0.3,-0.4,-0.3,-0.3) |
| $TIGA_8$ | (0.2,0.4,0.4,-0.1,-0.4,-0.5) | (0.1,0.4,0.3,-0.1,-0.2,-0.3) | (0.2,0.4,0.4,-0.1,-0.4,-0.5) | (0.8,0.2,0.1,-0.3,-0.2,-0.4) | (0.8,0.2,0.1,-0.3,-0.2,-0.4) | (0.1,0.4,0.3,-0.1,-0.2,-0.3) | (0.1,0.4,0.3,-0.1,-0.2,-0.3) |
| | $TIGC_1$ | $TIGC_2$ | $TIGC_3$ | $TIGC_4$ | $TIGC_5$ | $TIGC_6$ | $TIGC_7$ |
| $TIGA_1$ | (0.2,0.4,0.4,-0.1,-0.4,-0.5) | (0.5,0.4,0.3,-0.4,-0.3,-0.3) | (0.2,0.4,0.4,-0.1,-0.4,-0.5) | (0.4,0.3,0.3,-0.1,-0.2,-0.3) | (0.8,0.2,0.1,-0.3,-0.2,-0.4) | (0.2,0.4,0.4,-0.1,-0.4,-0.5) | (0.2,0.4,0.4,-0.1,-0.4,-0.5) |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $TIGA_2$ | (0.5,0.4,0.3,-0.4,-0.3,-0.3) | (0.1,0.4,0.3,-0.1,-0.2,-0.3) | (0.4,0.1,0.4,-0.1,-0.2,-0.5) | (0.4,0.3,0.3,-0.1,-0.2,-0.3) | (0.1,0.4,0.3,-0.1,-0.2,-0.3) | (0.4,0.1,0.4,-0.1,-0.2,-0.5) | (0.5,0.4,0.3,-0.4,-0.3,-0.3) |
| $TIGA_3$ | (0.4,0.3,0.3,-0.1,-0.2,-0.3) | (0.7,0.3,0.2,-0.4,-0.2,-0.1) | (0.7,0.3,0.2,-0.4,-0.2,-0.1) | (0.4,0.1,0.4,-0.1,-0.2,-0.5) | (0.7,0.3,0.2,-0.4,-0.2,-0.1) | (0.4,0.1,0.4,-0.1,-0.2,-0.5) | (0.2,0.4,0.4,-0.1,-0.4,-0.5) |
| $TIGA_4$ | (0.4,0.1,0.4,-0.1,-0.2,-0.5) | (0.4,0.1,0.4,-0.1,-0.2,-0.5) | (0.1,0.4,0.3,-0.1,-0.2,-0.3) | (0.7,0.3,0.2,-0.4,-0.2,-0.1) | (0.4,0.1,0.4,-0.1,-0.2,-0.5) | (0.1,0.4,0.3,-0.1,-0.2,-0.3) | (0.4,0.1,0.4,-0.1,-0.2,-0.5) |
| $TIGA_5$ | (0.7,0.3,0.2,-0.4,-0.2,-0.1) | (0.4,0.3,0.3,-0.1,-0.2,-0.3) | (0.8,0.2,0.1,-0.3,-0.2,-0.4) | (0.1,0.4,0.3,-0.1,-0.2,-0.3) | (0.4,0.3,0.3,-0.1,-0.2,-0.3) | (0.5,0.4,0.3,-0.4,-0.3,-0.3) | (0.7,0.3,0.2,-0.4,-0.2,-0.1) |
| $TIGA_6$ | (0.1,0.4,0.3,-0.1,-0.2,-0.3) | (0.5,0.4,0.3,-0.4,-0.3,-0.3) | (0.2,0.4,0.4,-0.1,-0.4,-0.5) | (0.5,0.4,0.3,-0.4,-0.3,-0.3) | (0.5,0.4,0.3,-0.4,-0.3,-0.3) | (0.2,0.4,0.4,-0.1,-0.4,-0.5) | (0.1,0.4,0.3,-0.1,-0.2,-0.3) |
| $TIGA_7$ | (0.8,0.2,0.1,-0.3,-0.2,-0.4) | (0.4,0.3,0.3,-0.1,-0.2,-0.3) | (0.4,0.1,0.4,-0.1,-0.2,-0.5) | (0.4,0.3,0.3,-0.1,-0.2,-0.3) | (0.4,0.3,0.3,-0.1,-0.2,-0.3) | (0.2,0.4,0.4,-0.1,-0.4,-0.5) | (0.5,0.4,0.3,-0.4,-0.3,-0.3) |
| $TIGA_8$ | (0.4,0.3,0.3,-0.1,-0.2,-0.3) | (0.5,0.4,0.3,-0.4,-0.3,-0.3) | (0.2,0.4,0.4,-0.1,-0.4,-0.5) | (0.5,0.4,0.3,-0.4,-0.3,-0.3) | (0.5,0.4,0.3,-0.4,-0.3,-0.3) | (0.4,0.1,0.4,-0.1,-0.2,-0.5) | (0.2,0.4,0.4,-0.1,-0.4,-0.5) |

The decision matrix is normalized using eq. (9) as shown in Table 6.

Determine the weighted decision matrix using eq. (10) as shown in Table 7.

Obtain the increased and decreased indexes for beneficial and non-beneficial criteria using eqs. (11 and 12).

The relative significant values are computed using eq. (13) as shown in Table 8.

Table 6. The normalized decision matrix.

| | $TIGC_1$ | $TIGC_2$ | $TIGC_3$ | $TIGC_4$ | $TIGC_5$ | $TIGC_6$ | $TIGC_7$ |
|---|---|---|---|---|---|---|---|
| $TIGA_1$ | 0.171875 | 0.149606 | 0.165563 | 0.177922 | 0.177461 | 0.168394 | 0.165803 |
| $TIGA_2$ | 0.177083 | 0.166667 | 0.165563 | 0.150649 | 0.156736 | 0.181347 | 0.17487 |
| $TIGA_3$ | 0.161458 | 0.165354 | 0.164238 | 0.181818 | 0.159326 | 0.168394 | 0.167098 |
| $TIGA_4$ | 0.177083 | 0.161417 | 0.157616 | 0.167532 | 0.167098 | 0.160622 | 0.154145 |
| $TIGA_5$ | 0.166667 | 0.175853 | 0.184106 | 0.164935 | 0.17487 | 0.15544 | 0.177461 |
| $TIGA_6$ | 0.145833 | 0.181102 | 0.162914 | 0.157143 | 0.164508 | 0.165803 | 0.160622 |
| $TIGA_7$ | 0.1875 | 0.160105 | 0.172185 | 0.155844 | 0.154145 | 0.163212 | 0.163212 |
| $TIGA_8$ | 0.166667 | 0.153543 | 0.164238 | 0.174026 | 0.173575 | 0.161917 | 0.154145 |

Table 7. The weighted normalized decision matrix.

| | $TIGC_1$ | $TIGC_2$ | $TIGC_3$ | $TIGC_4$ | $TIGC_5$ | $TIGC_6$ | $TIGC_7$ |
|---|---|---|---|---|---|---|---|
| $TIGA_1$ | 0.025039 | 0.020977 | 0.0234 | 0.025521 | 0.025479 | 0.02413 | 0.02362 |
| $TIGA_2$ | 0.025797 | 0.023369 | 0.0234 | 0.021609 | 0.022504 | 0.025987 | 0.024912 |
| $TIGA_3$ | 0.023521 | 0.023185 | 0.023213 | 0.02608 | 0.022876 | 0.02413 | 0.023804 |
| $TIGA_4$ | 0.025797 | 0.022633 | 0.022277 | 0.02403 | 0.023992 | 0.023017 | 0.021959 |
| $TIGA_5$ | 0.02428 | 0.024657 | 0.026021 | 0.023658 | 0.025107 | 0.022274 | 0.025281 |
| $TIGA_6$ | 0.021245 | 0.025393 | 0.023026 | 0.02254 | 0.02362 | 0.023759 | 0.022882 |
| $TIGA_7$ | 0.027315 | 0.022449 | 0.024336 | 0.022354 | 0.022132 | 0.023388 | 0.023251 |
| $TIGA_8$ | 0.02428 | 0.021529 | 0.023213 | 0.024962 | 0.024921 | 0.023202 | 0.021959 |

Table 8. The relative significant values.

| | Significant values |
|---|---|
| $TIGA_1$ | 0.166998 |
| $TIGA_2$ | 0.162438 |
| $TIGA_3$ | 0.168311 |
| $TIGA_4$ | 0.164336 |
| $TIGA_5$ | 0.168343 |
| $TIGA_6$ | 0.170705 |
| $TIGA_7$ | 0.160358 |
| $TIGA_8$ | 0.167802 |

We show the DALLE 3 model is the best based on the evaluation matrices.

## 6. Challenges and Limitations

The landscape of text-to-image generation is changing very quickly, and there are nonetheless a few central challenges that persist:

- Quality of Generated Images: Although newer models have worked towards the realism of the generated images, there still exist some situations where generated images may be of inferior quality or fidelity. These issues range from blurriness, artifacts, and insufficiency of detail, and can contribute to usability loss of generated images.
- Evaluation Metrics: The issue of measuring generated images still eludes us. Existing measures such as Inception Score (IS) and Fréchet Inception Distance (FID) have their shortcomings and do not accurately capture the qualitative aspect of images. Human judgment, while useful, is unreliable and variable, and one cannot infer model performance.
- Understanding Context: The current models may not be able to understand context or subtleties in written definitions. This can lead to image misinterpretation, with the generated image failing to carry the intended meaning as planned by the text.
- Scalability and Efficiency: Most advanced models need enormous computational resources to train and make inferences, and this may inhibit their availability and usability in real-time systems.

To counter these challenges, some of the potential directions to explore and create are:

- Enhanced Model Architectures: Future work can be directed toward creating new architectures with hybrid strategies by combining the strategies of different models (e.g., transformers, GANs, and attention mechanisms) to enhance the quality and diversity of generated images.

- Advanced Evaluation Techniques: Enhanced evaluation techniques, both qualitative and quantitative, are required. More recent metrics with the ability to measure improved subjective quality of output images and coherence of text description would be beneficial.
- Incorporating Contextual Understanding: Experiments can examine in what way the contextual knowledge of models and the text richness can be increased. It can be achieved by utilizing more diverse quantities of data or by having better natural language processing such that the text is better represented.
- Data Efficiency: Explore ways in which models can be efficiently trained from scarce data, i.e., few-shot or zero-shot learning strategies, for the generalization and popularization of text-to-image technology.
- User-Controlled Generation: Subsequent work may include allowing users greater control of generation. Models can generate closer-to-user taste and specification through the introduction of user-specified parameters or feedback loops.
- Cross-Modal Learning: Inspiration from cross-domain methods, such as vision-language pre-training, can be used to enhance coupling between the text and image modalities to support better understanding and generation.

## 7. Conclusions

This comparison analysis of text-to-image generation has revealed inherent trade-offs between GAN-based models like StyleGAN-T and diffusion models like Stable Diffusion 3. As our analysis demonstrates, while GAN architectures are superior in computational efficiency and generation speed (with inference rates of 0.8s compared to 4.5s for diffusion models), they never beat diffusion models on semantic fidelity, compositional correctness, and image quality overall. The quantitative metrics show diffusion-based approaches scoring significantly higher in CLIP testing (0.35 vs 0.29), text faithfulness (80% vs 60%), and compositional accuracy (82% vs 48%), demonstrating their superior ability to translate complex text descriptions into well-formed visual forms.

The evolutionary history from previous GAN systems such as AttnGAN and DF-GAN to current diffusion-based models reflects the direction of the field toward favoring semantic comprehension and visual consistency, even at higher computational costs. This evolution mirrors the evolving generative AI priorities toward models capable of interpreting and visualizing human complex instructions more effectively.

In real-world scenarios, our findings suggest that model choice must be determined by use-case needs: GAN-based models remain appropriate for real-time and budget-conscious scenarios, while diffusion models become central when semantic accuracy and photorealism are at top priority. Future research needs to focus on filling the gap by developing such hybrid architectures

that combine the efficiency of GANs with the semantic interpretability of diffusion models, through knowledge distillation techniques or architectural designs that ensure inference speed without sacrificing image quality and prompt alignment. As text-to-image technology continues to advance, surmounting age-old challenges in evaluation metrics, contextual comprehension, and computational efficiency will continue to be essential to developing systems that can reliably translate human creative intention into a visual form for a broad range of domains and applications.

We use the neutrosophic set model to evaluate different text-to-image generation models based on different evaluation metrices. We use eight models and seven evaluation matrices. The bipolar neutrosophic set is used to overcome the uncertainty information. We compute the evaluation matrices weights using the average method. The models are ranked using the COPRAS method. The results show the DALLE 3 model is the best under different evaluation matrices.

**References**

[1] G. Yin, B. Liu, L. Sheng, N. Yu, X. Wang, and J. Shao, "Semantics disentangling for text-to-image generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2327–2336.

[2] L. A. Zadeh, G. J. Klir, and B. Yuan, *Fuzzy sets, fuzzy logic, and fuzzy systems: selected papers*, vol. 6. World scientific, 1996.

[3] L. A. Zadeh, "Fuzzy sets," *Inf. Control*, vol. 8, no. 3, pp. 338–353, 1965.

[4] T. Fujita, "General plithogenic soft rough expert mixed graphs," *Algorithms*, vol. 55, pp. 63–105, 2024.

[5] D. Xu, X. Zhang, and H. Feng, "Generalized fuzzy soft sets theory-based novel hybrid ensemble credit scoring model," *Int. J. Financ. Econ.*, vol. 24, no. 2, pp. 903–921, 2019.

[6] F. Smarandache, *Neutrosophic precalculus and neutrosophic calculus: neutrosophic applications*. Infinite Study, 2015.

[7] F. Smarandache, *A unifying field in logics: neutrosophic logic. Neutrosophy, neutrosophic set, neutrosophic probability: neutrsophic logic. Neutrosophy, neutrosophic set, neutrosophic probability*. Infinite Study, 2005.

[8] F. Smarandache, *Introduction to neutrosophic measure, neutrosophic integral, and neutrosophic probability*. Infinite Study, 2013.

[9] F. Smarandache, "A unifying field in Logics: Neutrosophic Logic.," in *Philosophy*, American Research Press, 1999, pp. 1–141.

[10] Z. Shi and R. Liu, "Generative Modelling with High-Order Langevin Dynamics," *arXiv Prepr. arXiv2404.12814*, 2024.

[11] Y. Wu, J. Donahue, D. Balduzzi, K. Simonyan, and T. Lillicrap, "Logan: Latent

optimisation for generative adversarial networks," *arXiv Prepr. arXiv1912.00953*, 2019.

[12] M. Lee and J. Seok, "Score-guided generative adversarial networks," *Axioms*, vol. 11, no. 12, p. 701, 2022.

[13] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," *arXiv Prepr. arXiv1809.11096*, 2018.

[14] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv Prepr. arXiv2011.13456*, 2020.

[15] I. Deli, M. Ali, and F. Smarandache, "Bipolar neutrosophic sets and their application based on multi-criteria decision making problems," in *2015 International conference on advanced mechatronic systems (ICAMechS)*, Ieee, 2015, pp. 249–254.

[16] V. Ulucay, I. Deli, and M. Şahin, "Similarity measures of bipolar neutrosophic sets and their application to multiple criteria decision making," *Neural Comput. Appl.*, vol. 29, pp. 739–748, 2018.

[17] B. Ma, Z. Zong, G. Song, H. Li, and Y. Liu, "Exploring the role of large language models in prompt encoding for diffusion models," *arXiv Prepr. arXiv2406.11831*, 2024.

[18] M. Zhu, P. Pan, W. Chen, and Y. Yang, "Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5802–5810.

[19] M. Otani *et al.*, "Toward verifiable and reproducible human evaluation for text-to-image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14277–14286.

[20] Z. Tan *et al.*, "An empirical study and analysis of text-to-image generation using large language model-powered textual representation," in *European Conference on Computer Vision*, Springer, 2024, pp. 472–489.

[21] M. Liu *et al.*, "Llm4gen: Leveraging semantic representation of llms for text-to-image generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, pp. 5523–5531.

Received: Nov. 20, 2024. Accepted: May 9, 2025