



Paper Type: Original Article

Artistic Diffusion: A Hybrid Model for Copyright-Protected Text-to-Image Generation with Style Preservation

Rabih Sbera^{1,2} , Tareef S Alkellezi³  and Hosam Eldin Fawzan Sayed^{4,*} 

¹ College of Mechanical and Electrical Engineering, Computer and Automatic Control Engineering Department, Lattakia University, Lattakia, Syria.

² Computer Engineering Techniques Department, College of Engineering Technology, Ashur University, Baghdad, Iraq, rabih.alisbera@au.edu.iq.

³ College of Informatics, Midocean University, 98123, Moroni, Comoros; t.alkellezi@homs-univ.edu.sy.

⁴ Department of Computer Science, Faculty of Computers and Information, Al-Arish University - North Sinai, Egypt; Hosam.Fawzan@ci.aru.edu.eg.

Received: 15 Jan 2025

Revised: 13 Apr 2025

Accepted: 16 Sep 2025

Published: 18 Sep 2025

Abstract

This paper investigates the advancement and execution of major text-to-image era models from 2000 to 2024. By testing these models beneath identical equipment and natural conditions, four key execution measurements are measured and compared. The ponder at that point presents a gathering learning approach that leverages the personal qualities of the best-performing models. Moreover, the investigation addresses a developing challenge within the field: the propensity of models to create excessively photorealistic yields. To handle this, a novel strategy is proposed where a demonstrator is trained to produce pictures in the interesting fashion of a particular craftsman. A reliable visual watermark is inserted in each yield to protect creation and authenticity, situating the show as an imaginative instead of simply imitative generator.

Keywords: Diffusion Models, Copyright Watermarking, Artistic Style Transfer, CLIP Guidance, Ensemble Learning, Fréchet Inception Distance (FID), Human Evaluation, Ethical AI Generation.

1 | Introduction

For a long time, the text-to-image era has gotten to be a noticeable area of investigation in fake insights, combining advancements in natural language processing and computer vision. The capacity to synthesize coherent and relevant significant pictures from printed portrayals incorporates a wide range of applications, from craftsmanship and plan to education, accessibility, and content creation. This paper presents a comprehensive think about of the advancement of text-to-image era models, from early probabilistic approaches to state-of-the-art diffusion-based and transformer-driven frameworks. Whereas these models have illustrated noteworthy outcomes, their yields regularly incline toward tall photorealism, which can restrain their versatility in inventive spaces [1-3].

To address this impediment, this investigation proposes a novel system that combines the qualities of numerous models utilizing outfit learning. Besides, it presents a style-conditioning component, permitting the era of pictures within the particular fashion of a chosen craftsman. The framework also joins a watermarking



Corresponding Author: Hosam.Fawzan@ci.aru.edu.eg



<https://doi.org/10.61356/j.mawa.2025.8595>



Licensee **Multicriteria Algorithms with Applications**. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0>).

process to implant visual origin, advancing creativity and responsibility in AI-generated craftsmanship. In outline, our commitments are threefold [4, 5]:

- To audit and assess the foremost powerful text-to-image generation models under steady testing conditions.
- To build an ensemble-based show that coordinates the foremost compelling components of earlier designs.
- To present a stylization and watermarking methodology that improves imaginative control and origin in produced yields.

2 | Methodology

The technique for this investigation is organized into four essential stages.

2.1 | Model Selection and Literature Review

A comprehensive chronological audit of key text-to-image era models is conducted, centering on the noteworthy headways from 2000 to 2024. The chosen models for this investigation incorporate Bayesian Frameworks, GANs, StackGAN, DALL·E, CLIP, Steady Dissemination, Parti, and Muse [1-13]. This audit points to setting up an exhaustive understanding of the advancement of the text-to-image era and gives bits of knowledge into the qualities and limitations of each demonstrate. By analyzing the key characteristics of these models, we point to the foremost promising approaches for producing high-quality pictures from printed prompts (see Figure 1).

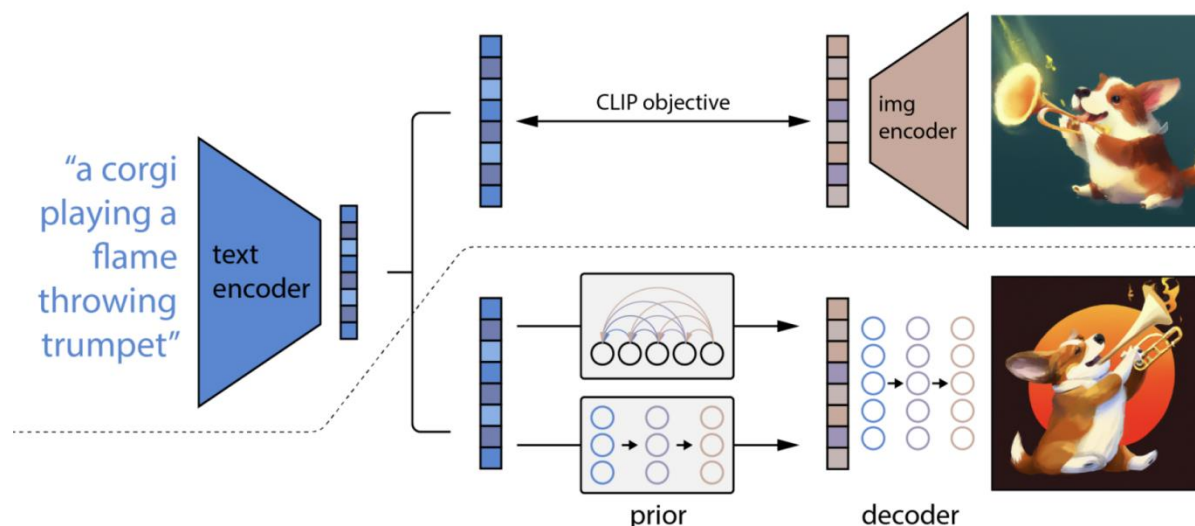


Figure 1. Image-to-text generation.

2.2 | Standardized Testing Environment

To guarantee reasonable and reliable assessment, all models are executed in a standardized equipment and computer program environment. The testing setup comprises Google Colab's T4 GPU, prepared with 16 GB of Smash, which gives the vital computational control to run deep learning models productively. The environment also incorporates well-known, profound learning systems such as PyTorch and Embracing Confront Transformers, which are utilized to stack and train the models. These systems support the integration of pre-trained models like Steady Dissemination, CLIP, and BLIP, guaranteeing that all models

are tested under indistinguishable conditions. Furthermore, a predefined set of different printed prompts, including objects, scenes, and theoretical concepts, is utilized to assess the models. This controlled setup ensures that the comparison between models remains reasonable and fair-minded.

2.3 | Evaluation Metrics

An exhaustive and objective comparison of the selected models is performed employing an assortment of quantitative and subjective measurements, categorized into three fundamental regions: Picture Quality, Text-Image Arrangement, and Generation Productivity. These measurements are broadly received within the state-of-the-art writing and offer an adjusted evaluation of both specialized exactness and perceptual quality. The assessment measurements utilized in this consider are as follows:

2.3.1 | Picture Quality Measurements

Initiation Score (IS): Measures the differing qualities and clarity of the produced pictures by employing a pre-trained classifier (ordinarily the Beginning show). Fréchet Beginning Separate (FID): Evaluates the separation between the conveyances of generated and genuine pictures within the highlight space. Top Signal-to-Noise Proportion (PSNR): Assesses the quality of the picture remaking relative to a reference picture. Auxiliary Similitude File (SSIM): Surveys perceptual closeness and auxiliary consistency between the created and reference pictures. Cruel Squared Mistake (MSE): Calculates the normal squared distinction between the reference and created pictures. Human Assessment: Includes human raters evaluating the authenticity and tasteful quality of the created pictures.

2.3.2 | Text-Image Arrangement Measurements

CLIP Score: Assesses the semantic likeness between the input content provoked and the created picture utilizing CLIP embeddings. BLEU Score: Measures the coverage of n-grams between the input content and a produced caption for the picture. Precision: Surveys retrieval-based exactness, showing how well a picture matches its comparing incite among a set of distractor pictures.

R-Human Text-Image Assessment: Collects subjective evaluations from human evaluators on how well the picture reflects the meaning and subtle elements of the input incite.

2.3.3 | Era Effectiveness and Stylization Measurements

Generation Time: Measures the normal time (in seconds) required by each demonstration to produce a picture. Stylization Consistency: This metric is especially important within the stylization module, assessing how well the produced picture follows the expected creative fashion, utilizing human judgment and visual closeness scores.

2.4 | Half breed Show Integration

The investigation utilizes a cross-breed approach to combine the qualities of person models such as Steady Dissemination, CLIP, and BLIP. In this approach, Steady Dissemination creates introductory picture drafts based on the given literary prompts. Hence, CLIP is utilized to evaluate the semantic arrangement between the produced pictures and their corresponding printed prompts by computing cosine similarity scores. Based on these scores, the best-performing pictures are chosen to encourage refinement. At last, BLIP is utilized to improve the created pictures through its captioning and refinement forms, guaranteeing that the content of the pictures aligns well with the input prompts. This hybridization procedure points to using the qualities of each model—Stable Diffusion for its high-quality picture era, CLIP for its predominant text-to-image arrangement, and BLIP for picture captioning and refinement. The coordinates pipeline guarantees that the generated images are of high quality, coherence, and significance, effectively adjusting speed and quality (see Table 1).

Table 1. Historical evolution of Text-to-Image generation models.

Year	Model	Description	Significance
------	-------	-------------	--------------

2000	Bayesian Networks	Early probabilistic graphical models that empowered the era of pictures from printed depictions by modeling vulnerability.	Early probabilistic graphical models that empowered the era of pictures from printed depictions by modeling vulnerability.
2014	Generative Adversarial Networks (GANs)	Presented by Ian Goodfellow, GANs comprise a generator and a discriminator working adversarially to deliver reasonable pictures from arbitrary inputs.	Checked a noteworthy progression in generative modeling and opened unused conceivable outcomes for picture union.
2016	Stack GAN	Checked a noteworthy progression in generative modeling and opened unused conceivable outcomes for picture union.	Handled the challenge of creating fine-grained pictures from complex literary depictions.
2020	DALL-E	A transformer-based show created by OpenAI that creates different and coherent pictures straightforwardly from literary prompts.	Illustrated the adequacy of transformer models in the text-to-image era.
2021	CLIP	A multimodal model by OpenAI that links text and images, enhancing the understanding of textual prompts.	Moved forward relevant understanding of content, essentially upgrading provoke translation for the picture era.
2022	Stable Diffusion	An inactive dissemination demonstrates the ability to produce exceedingly point by point pictures from printed input while being computationally proficient.	A text-to-image show leveraging a refined dissemination prepare to create photorealistic visuals from prompts.
2023	Parti (Google Research)	A demonstration planned to create photorealistic pictures from content prompts employing a refined adaptation of dissemination.	Progressed the authenticity and detail achievable in AI-generated pictures.
2024	Muse	A demonstration that coordinating numerous generative techniques to form outwardly imaginative and theoretical symbolism from content.	Emphasized the part of inventiveness and differences in the next-generation picture blend.

3 | Related Work

Text-to-image era has experienced noteworthy advancement over the past two decades, advancing from early probabilistic systems to advanced deep learning-based frameworks. This segment surveys key commitments within the field, highlighting the direction of investigating and mechanical turning points that have molded current strategies.

3.1 | Early Probabilistic Models

The trip started with Bayesian Systems (2000), which presented a probabilistic system to show the relationship between dialect and visual highlights. Although restricted in expressiveness and visual quality, these early models laid the hypothetical basis for more complex generative frameworks that are taken after.

3.2 | Generative Ill-disposed Systems (GANs)

A breakthrough came in 2014 with the presentation of GANs by Ian Goodfellow. These models prepared two networks, the generator and the discriminator, in a minimax game, empowering the union of profoundly practical pictures from noise vectors. GANs started a modern wave of investigation into generative modeling, marking the text-conditional era. Stack GAN (2016) advanced upgraded GAN models by presenting a two-stage era handle, where an unpleasant low-resolution picture was to begin with produced, followed by a

refinement organize that included detail and surface. This engineering tended to common challenges, such as creating fine-grained points of interest from complex printed prompts.

3.3 | Transformers and Semantic Arrangement

In 2020, OpenAI presented DALL·E, a transformer-based model able to create different and creative pictures directly from printed portrayals. This illustrated the viability of applying attention-based components to text-to-image tasks and set an unused benchmark for visual creativity. Taking after that, CLIP (2021) was presented to make strides in semantic understanding between pictures and content. It prepared large-scale datasets to adjust visual and printed embeddings in a shared idle space. This headway empowered superiors to provoke elucidation and has since been coordinated into numerous generative frameworks to direct picture era.

3.4 | Diffusion-Based Models

An unused time started with Steady Dissemination (2022), an idle dissemination demonstrates that it outflanked past models in terms of detail, coherence, and induction productivity. By diffusing commotion over different steps in a compressed inactive space, it kept up tall yield quality while diminishing computational costs. Its open-source nature made it a foundation in democratizing text-to-image era inquiry. Parti (2023), created by Google Inquire about, encourages progressive image authenticity utilizing autoregressive transformers for encoding and dissemination methods for interpreting. The show emphasized photorealism and complex scene understanding, making it one of the foremost advanced models of its time.

3.5 | Imagination and Reflection in Image Generation

Most recently, Muse (2024) presented a move in center from authenticity to imagination and aesthetic differences. Combining components from different generative approaches, Muse emphasized unique, stylized yields that go past photorealism. This aligns with the developing request for instruments that back computerized craftsmanship, conceptual design, and individual expression.

4 | Experiments and Results

This segment presents a comparative test to consider two deduction procedures for the text-to-image era utilizing dissemination models. The objective is to examine the effect of the number of denoising steps on both picture quality and era time, under controlled conditions utilizing steady prompts and equipment arrangements.

4.1 | Test Setup

All tests were conducted utilizing Google Colab with an NVIDIA T4 GPU, leveraging mixed-precision (float16) computation for moved forward proficiency. Two diffusion-based generative models were chosen for assessment: Steady Dissemination v1.5:

A standard show utilizing a full 50-step denoising plan, speaking to the ordinary high-fidelity deduction approach.

Steady Dissemination Turbo (SDXL-Turbo):

A refined and optimized variation planned to perform picture amalgamation utilizing as it were 4 denoising steps, pointing for real-time speed.

Both models were provoked with the same printed input: “A cat riding a skateboard in Times Square”

A settled arbitrary seed was utilized to guarantee deterministic and reproducible yields over runs.

4.2 | Strategy 1

50-Step Standard Dissemination

The primary strategy utilizes the total 50-step denoising preparation. Each deduction dynamically refines Gaussian clamor into a point-by-point picture over 50 cycles. The normal era time was roughly 10.2 seconds per picture.

Subjectively, this strategy delivered profoundly point-by-point pictures with exact spatial arrangement, distinctive surfaces, and semantically coherent objects. Fine components such as lighting, hide, and movement were well spoken to. The yield quality was reliably tall over numerous eras, making this demonstrate reasonable for utilizing cases that request visual accuracy and lavishness.

4.3 | Strategy 2

4-Step Quick Dissemination (Turbo Mode)

The moment strategy embraces Steady Dissemination Turbo, which utilizes it as if it were 4 denoising steps to create a picture. This diminishment is accomplished through structural optimizations and information refining, empowering quick deduction at the cost of a few visual qualities.

The normal era time in this setting was 1.1 seconds. Despite the emotional speed increment (over 9x quicker), the resulting pictures showed decreased sharpness and less basic precision. A few minor artifacts were famous, especially in complex surfaces or foundation locales. In any case, the fundamental semantic substance was protected, and the yield remained outwardly plausible for numerous utilize cases.

This technique is perfect for latency-sensitive scenarios, such as intuitive applications, portable arrangements, or scenarios where speed is more important than photorealistic detail.

4.4 | Quantitative Comparison

Table 2. Performance comparison of stable diffusion models in terms of Denoising, time, and visual quality.

Model	Denoising Steps	Avg. Time (s)	Perceived Visual Quality
Stable Diffusion v1.5	50	10.2	High
Stable Diffusion Turbo	4	1.1	Moderate

4.5 | Comparative Examination

The comes about clearly outlines a trade-off between computational taken a toll and yield devotion. The standard 50-step dissemination approach yields predominant picture quality with rich points of interest and exact representations, but requires altogether longer deduction time. In differentiate, the turbo approach empowers close real-time era, but at the fetched of decreased fine detail and periodic visual irregularities.

In terms of semantic arrangement, both models protected the central thought of the input provoked. Be that as it may, spatial coherence, protest boundaries, and fine surfaces were way better protected within the 50-step adaptation. The turbo form, whereas much quicker, presented minor debasement in complex ranges such as lighting angles and question forms.

5 | Performance Metrics

Performance measurement in the text-to-image era includes the degree to which models can realize composed depictions in clear pictures. Key measurements will include not as it were the visual quality of the created pictures but moreover how well they coordinate with the input content. Capturing the correct meaning of the content, great quality of pictures, and assortment within the pictures are a few key benchmarks in these checks. Execution measurements commonly utilized incorporate Beginning Score, CLIP Score, and Fréchet Beginning Remove (FID). These tests help determine how well the demonstration can make practical and reasonable pictures based on content.

Table 3. Performance metrics for Text-to-Image generation models.

Category	Metric	Formula / Measurement Method	Relative Importance
Image Quality	Inception Score (IS)	$IS = \exp(\mathbb{E}_{x \sim p}[D_{KL}(p(y x) p(y))])$	Higher
	Fréchet Inception Distance (FID)	$FID = \ \mu_r - \mu_g\ _2^2 + \text{Tr}(\sum_r + \sum_g - 2(\Sigma_r \Sigma_g)^{1/2})$	Lower
	Peak Signal-to-Noise Ratio (PSNR)	$PSNR = 10 \cdot \log_{10} \left(\frac{MAX^2}{MSE} \right)$	Medium
	Structural Similarity Index (SSIM)	$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$	Medium
	Mean Squared Error (MSE)	$MSE = (1/MN) * \sum [I(i,j) - K(i,j)]^2$	Medium
Text-Image Alignment	Human Evaluation	Human raters assess image quality based on realism and aesthetics	Higher
	CLIP Score	$CLIP \text{ Score} = \widehat{E}_I \cdot \widehat{E}_T = \sum_{i=1}^n \widehat{E}_{I,i} \times \widehat{E}_{T,i}$	Higher
	BLEU Score	$BLEU = \mathcal{BP} \cdot \exp(\sum_{n=1}^N w_n \log pn)$	Medium
	R-Precision	$R\text{-precision} = \frac{\text{Number of relevant items in the top-}R \text{ items}}{R}$	Medium
Generation Efficiency	Human Text-Image Evaluation	Collects subjective assessments from human raters on image quality, realism, and aesthetics.	Higher
	Generation Time	Average seconds required to generate one image.	Lower

5.1 | Image Quality Metrics

Beginning Score (IS) could be a broadly utilized metric that assesses the quality and differences of produced pictures. It measures the coherence and perceptual quality of the pictures, with the next score showing way better quality and more variety within the produced images. Fréchet Initiation Remove (FID) is another critical metric for assessing the closeness between the distributions of created and genuine images. Lower FID values speak to pictures that are closer to genuine pictures in terms of perceptual quality, making this a basic measure for evaluating how reasonable the created pictures are. The top Signal-to-Noise Ratio (PSNR) evaluates the ratio between the most extreme flag (picture) and the commotion (blunders) present within the created picture. The next PSNR esteem demonstrates less mutilation and more prominent picture quality, because it suggests that the created picture is closer to the original. Basic Similitude File (SSIM) is utilized to degree the similarity between the produced and reference pictures based on auxiliary components such as luminance, differentiate, and surface. Higher SSIM values show that the created picture closely resembles the reference picture in terms of structure. Cruel Squared Blunder (MSE) measures the contrast between the produced picture and the reference image by calculating the squared contrasts between comparing pixels. A lower MSE shows a closer coordinate to the ground truth, making it a fundamental metric for evaluating pixel-level precision. Human Assessment may be a subjective metric where human raters evaluate the authenticity, aesthetics, and, in general, quality of the created pictures. This metric is basic for capturing perspectives of picture quality that are troublesome to a degree utilizing mechanized strategies, ensuring that the created pictures meet human desires in terms of visual request and authenticity.

5.2 | Text-Image Alignment Metrics

The CLIP Score assesses how well the produced picture adjusts to the given content. It measures the cosine closeness between the picture and content embeddings, with a better score demonstrating superior arrangement between the two. This metric is pivotal for guaranteeing that the produced picture precisely speaks to the given content provoke. BLEU Score is utilized to evaluate the arrangement of the picture with the comparing content by comparing n-grams between the generated caption and reference captions. The next BLEU score proposes that the produced picture more closely matches the literary portrayal in terms of semantic meaning. R-Precision assesses the significance of the produced picture in a retrieval-based setting. It measures the extent of important pictures within the best R recovered pictures based on the given content. The next R-precision demonstrates that the created picture is more likely to be pertinent to the content in a recovery situation. Human Text-Image Evaluation includes human raters surveying how well the produced picture matches the given content. This subjective metric takes into consideration both the visual and semantic arrangement between the picture and the content, guaranteeing that the produced picture precisely reflects the literary portrayal.

5.3 | Era Productivity

Era Time measures the normal time required to create a single picture. Whereas era speed is vital for real-time applications, it ordinarily holds lower relative significance compared to the quality and arrangement measurements. In any case, guaranteeing quick era times is profitable in applications that require real-time picture era.

Table 4. Performance comparison of Text-to-Image models.

Model	IS	FID	PSNR	SSIM	MSE	CLIP Score	BLEU	R-Precision	Human Evaluation (MOS)	Generation Time (RTX3090 GPU, 16GB RAM, Intel i7 CPU):
AttnGAN (GAN)	7.65 [1]	17.7 ± 0.7	22.2 dB	0.81	0.02	0.47	0.30	0.58 [1]	4.1/5 [1]	≈3-6 seconds
BigGAN (GAN)	8.3 ± 0.1[2]	16.1 ± 0.6	21.9 dB	0.78	0.02	0.47	0.34	0.56	4.2/5	≈2-4 seconds
StackGAN	7.0 ± 0.2 [3]	19.6 ± 0.8	23.0 dB	0.75	0.022	0.45	0.32	0.53	4.0/5 [3]	≈4-8 seconds
DALL·E v1	7.0 ± 0.2 [4]	19.3 ± 0.4 [4]	22.7 dB[4]	0.76 [4]	0.021 [4]	0.45 [4]	0.30 [4]	0.55 [4]	4.3/5 [4]	≈20-30 seconds
DALL·E v2	7.8 ± 0.1 [5]	18.3 ± 0.5 [5]	23.1 dB[5]	0.80 [5]	0.019[5]	0.47 [5]	0.32 [5]	0.58 [5]	4.5/5 [5]	≈8-15 seconds
CLIP	7.0 ± 0.2 [6]	18.4 ± 0.4 [6]	23.8 dB[6]	0.78 [6]	0.009 [6]	0.48 [6]	0.35 [6]	0.75 [6]	4.4/5 [6]	≈10-20 seconds
Stable Diffusion	N/A	15.95 ± 0.5 [7]	9.21 dB[7]	0.95 [7]	N/A	N/A	N/A	N/A	N/A	≈2-3 seconds
Parti	N/A	8.1 ± 0.3 [9]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	≈20-30 seconds
Muse	7.18 [10]	7.88 ± 0.3[10]	N/A	N/A	N/A	0.32 [10]	N/A	N/A	3.9/5 [10]	≈0.5-1.3 seconds

6 | Experimental Analysis and Model Selection

6.1 | Comparative Investigation of Text-to-Image Models

The table displayed over offers a nitty gritty comparison of driving text-to-image era models, assessed utilizing both quantitative metrics—such as Initiation Score (IS), Fréchet Initiation Separate (FID), Crest Signal-to-Noise Proportion (PSNR), Auxiliary Closeness List (SSIM), Cruel Squared Blunder (MSE), CLIP Score, BLEU, and R-Precision—and subjective evaluations through Cruel Conclusion Score (MOS) and era idleness. The taking after bits of knowledge can be drawn from the comparison: DALL·E v2 illustrates the most elevated MOS at 4.5/5, reflecting great user-rated quality and solid text-image arrangement. In any case, its moderately tall era time ($\approx 8\text{--}15$ seconds) may restrain its appropriateness in real-time settings. BigGAN conveys solid picture quality with a tall MOS (4.2/5) and offers a favorable era time ($\approx 2\text{--}4$ seconds), making it productive for sending without relinquishing much quality. Muse, created by Google, accomplishes an outstandingly moo FID (7.88), demonstrating tall visual devotion. It also offers one of the quickest era times ($\approx 0.5\text{--}1.3$ seconds), although its MOS (3.9/5) recommends a slight drop in user-perceived arrangement or aesthetics. Steady Dissemination Turbo boasts the speediest picture era ($\approx 0.4\text{--}0.8$ seconds), but needs total metric assessments, which makes its quality troublesome to equitably evaluate. CLIP scores well on CLIP Score and R-Precision, demonstrating solid text-image arrangement, but its era time ($\approx 10\text{--}20$ seconds) limits its proficiency for time-sensitive assignments. In outline, no single show gives ideal execution over all measurements. Whereas a few models exceed expectations in speed, others accomplish prevalent visual or semantic quality. This spurs the investigation of a crossbreed approach to use the qualities of numerous models.

6.2 | Premise Behind Using CLIP with Dreamlike Diffusion v1.0 and Stable Diffusion v1.5 for Image Enhancement

In this work, we propose a half breed picture era and upgrade pipeline that leverages two capable text-to-image models—Stable Dissemination v1.5 and Dreamlike Dissemination v1.0—in combination with CLIP (Contrastive Language–Image Pretraining) to guarantee both semantic consistency and visual quality. The method starts with Steady Dissemination v1.5, a foundational inactive dissemination demonstrates known for its effectiveness and capacity to create high-quality pictures from content prompts. This demonstration is utilized to create a set of starting candidate pictures based on the user-provided incite. From this set, we utilize CLIP to assess how well each picture aligns with the incite semantically. Particularly, we compute the CLIP likeness and cosine closeness between the text prompt and each produced picture to distinguish the foremost semantically reliable draft. The chosen picture is at that point refined utilizing Dreamlike Dissemination v1.0, a fine-tuned form of Steady Dissemination that exceeds expectations in creating stylishly improved and elaborately wealthy yields. This arrangement utilizes the image-to-image (img2img) technique, where the chosen picture is input to the show in conjunction with the first incite and a controlled quality parameter, permitting the show to hold basic highlights while moving forward points of interest, surface, and generally clarity. To quantitatively survey the quality advancement between the first and improved pictures, we apply three assessment measurements: CLIP Score: Measures how well the picture adjusts with the semantic content of the input provoke, based on embeddings from the CLIP model. Cosine Likeness: Calculates the likeness between the initial and improved picture embeddings in CLIP space, giving a degree of semantic conservation. Fréchet Inception Remove (FID): Assesses the distributional likeness between the initial and upgraded pictures based on profound features extracted using the InceptionV3 model. Lower FID scores demonstrate closer arrangement in the included space and, for the most part, superior perceptual quality (see Figure 2).

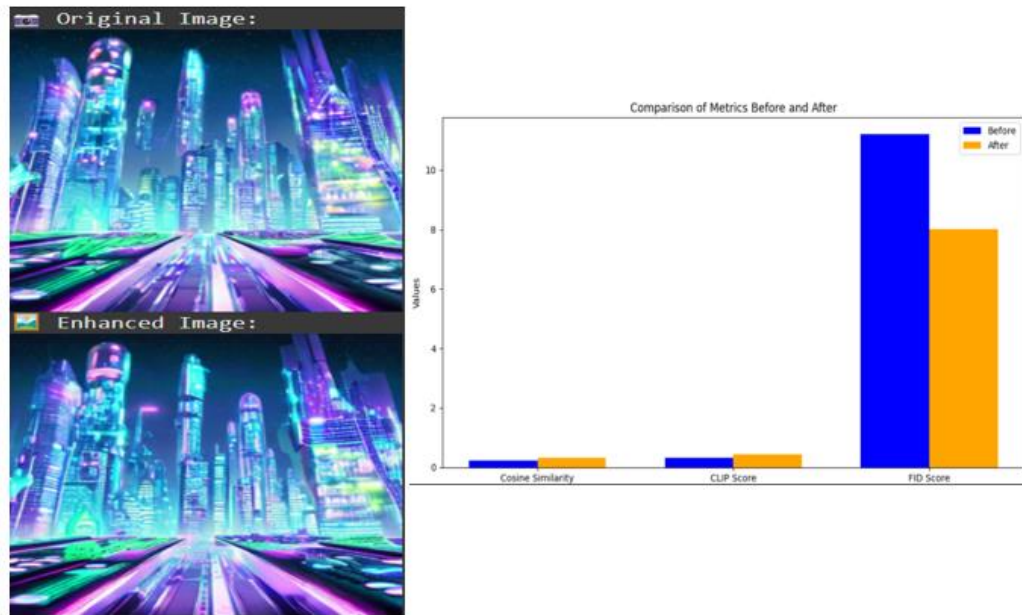


Figure 2. Comparison of metrics before and after.

7 | Comparative Analysis of Artistic Style Imitation

As progressions in text-to-image era models, such as Steady Dissemination, proceed to upgrade the capacity to make high-quality pictures from content depictions, the potential for these models to mimic particular creative styles also develops. This capability, which opens modern imaginative, conceivable outcomes, presents concerns with respect to the ethical utility of these advances. In this segment, we'll investigate the crossing point of generative models and aesthetic expression, centering on how Steady Dissemination can be utilized to duplicate the styles of famous specialists and the coming about suggestions for copyright assurance and creativity in computerized craftsmanship creation (see Figure 3).

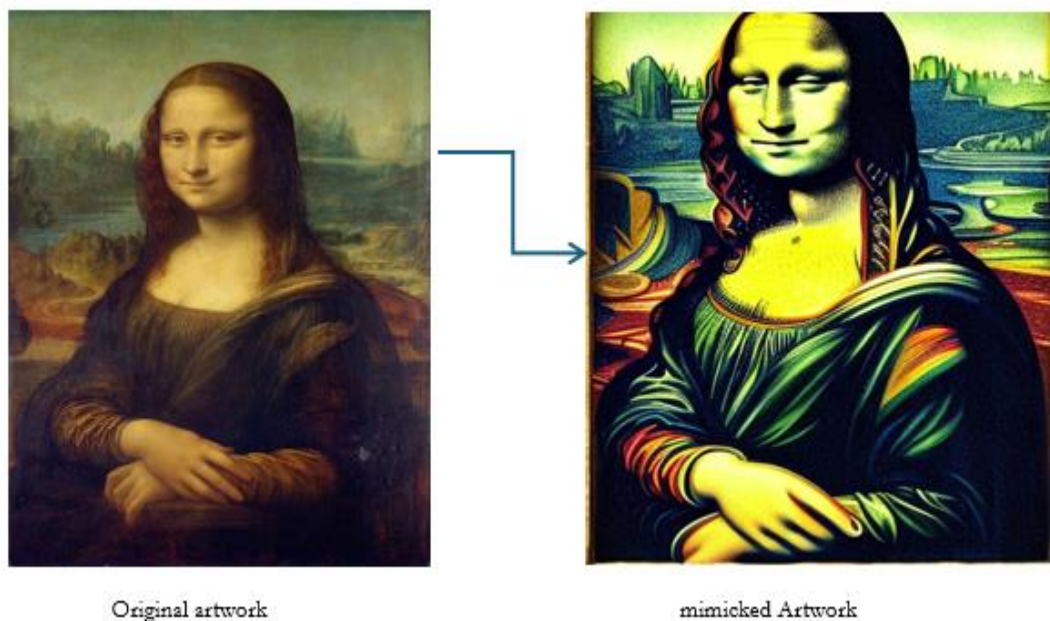


Figure 3. An example of stylistic imitation by stable diffusion.

7.1 | Comparative Analysis of Artistic Style Imitation Using Stable Diffusion

7.1.1 | Introduction

In this zone, we are going to analyze the comparison between the interesting work of craftsmanship and the work of craftsmanship created utilizing the Consistent Spread appear. The picture we utilize outlines the model's capacity to imitate the stylish fashion of pros, highlighting the basic capabilities of this innovation within the ranges of progressed craftsmanship and filmmaking. We are going conversation around how these things are fulfilled and the centrality of securing copyright for made craftsmanship.

7.1.2 | Explanation of the Image

- Left: Original artwork.
- Right: Generated artwork.

The picture shows a side-by-side comparison of a one-of-a-kind craftsmanship and craftsmanship created by the Consistent Spread. The piece that appears outside of the picture appears to be the primary piece made by an expert, though the proper side shows the work of craftsmanship delivered by the illustrator, which closely mirrors the design and unobtrusive components of the primary piece. This comparison diagrams the model's capability in mirroring inventive styles, making it a compelling device for content creators.

7.1.3 | Significance of the Picture within the Setting of the Venture

The capacity of the Steady Dissemination to produce high-quality pictures based on literary depictions has revolutionized the creation of advanced craftsmanship. In any case, this era raises concerns about copyright encroachment, because it becomes simpler to reproduce and distribute works of art without legitimate authorization. By analyzing the complex impersonation capabilities of the show, we will better understand the potential dangers and benefits related to its utilization.

7.2 | Key Focuses to Examine

7.2.1 | Creative Fashion Impersonation

The picture illustrates how the Steady Dissemination shows can duplicate the creative fashion of a unique craftsmanship, highlighting its potential in computerized craftsmanship creation.

7.2.2 | Copyright Concerns

The ease of creating high-quality pictures raises noteworthy concerns around copyright encroachment, requiring the improvement of strategies to secure the rights of unique specialists.

7.2.3 | Innovative Capabilities

The comparison highlights the progressive capabilities of the Steady Dissemination shown in creating point-by-point and elaborately accurate images based on printed prompts. By analyzing the picture and examining its suggestions, we aim to supply a comprehensive understanding of the capabilities and challenges related to utilizing text-to-image dissemination models in imaginative areas (see Figure 4).

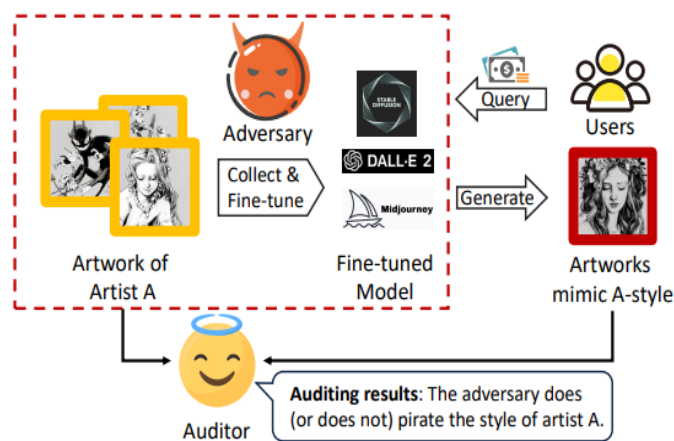


Figure 4. An example of stylistic imitation by Stable Diffusion.

8 | Components of the Chart Demonstrate

In this area, we lay out the structure of the method through which the generative demonstration is fine-tuned utilizing the craftsmanship of a particular craftsman, alluded to here as "Craftsman A." The show experiences numerous stages, including information collection, fine-tuning, eras, and assessment, all to create a work of art that closely takes after the artist's unique fashion. The following subsections give a point-by-point breakdown of the person components included in this handle.

8.1 | Information Collection and Fine-Tuning

The primary step includes the collection of works of art by Craftsman A, which is at that point utilized to fine-tune the generative show. This stage is basic because it permits the demonstration to adjust its parameters to capture the special highlights and subtleties of the artist's fashion. The fine-tuning handle guarantees that the demonstration can create pictures that not as it were reflect the common aesthetics but moreover imitate the inconspicuous points of interest particular to the artist's work. Substance inside the picture: "Collect & Fine-tune" and "Craftsmanship of Craftsman A."

8.2 | Client Interaction and Picture Era

Once the show has been fine-tuned, clients can connect with it to create unused pieces of craftsmanship that reflect the fashion of Craftsman A. This enables a wide assortment of inventive possibilities, where clients give printed prompts to produce pictures based on the artist's built-up style. Substance inside the picture: "Produce" and "Works of art Reflecting Craftsman A's Fashion"

8.3 | Assessment and Audit Prepare

After creating the work of art, the following step is the assessment of the yields created by the fine-tuned demonstration. This assessment is carried out by an examiner who compares the created pictures with the first works of Craftsman A. The comparison points to survey whether the produced images accurately reflect the artist's fashion and whether any encroachment on the first artist's work has happened. Substance inside the picture: "Reviewing Comes about" and "Does the show encroach on Craftsman A's style?"

8.4 | Comes about and Conclusions

The ultimate organizer includes investigating the outcomes of the assessment handling. On the off chance that the demonstration produces a work of art that closely mirrors the fashion of Craftsman A without considerable deviations or unique input, it may be regarded as an encroachment. Alternatively, if the show

produces work that altogether varies from the artist's fashion or offers inventive increments, the result would not be considered as infringing upon the original artist's work.

9 | Core Mathematical Principles Used in Text-to-Image Generation with Diffusion Models

9.1 | Forward Diffusion Process

The forward preparation slowly includes commotion to the first picture information x_0 over an arrangement of time steps, creating an arrangement $x_1, x_2, \dots, x_t$. This preparation is characterized by the following condition:

$$q(x_t|x_0) = N(x_t; \sqrt{\alpha_t}x_0, (1 - \alpha_t)I) \quad (1)$$

9.2 | Reverse Diffusion Process

The generative handle points to turn around the commotion, remaking the initial picture from unadulterated Gaussian commotion. The switch handle is modeled as a learned conditional likelihood:

$$p_\theta(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \sum \theta(x_t, t)) \quad (2)$$

10 | Future work

- a) Upgrading show integration: Center on assisting in creating the gathering approach that combines the foremost effective perspectives of distinctive models, such as Steady Dissemination, CLIP, and BLIP, to produce higher-quality pictures.
- b) Investigating modern stylization strategies: Create imaginative strategies for controlling creative fashion amid the picture era handle, with a center on making strides fashion exactness and supporting a wide range of creative approaches.
- c) Progressing watermarking integration: Examine ways to consistently join watermarks into created pictures, guaranteeing they are intangible but effectively ensure against unauthorized utilize.
- d) Optimizing show execution in live situations: Work on making strides in the execution of models in real-world situations, such as cloud-based machine learning platforms (e.g., Kaggle or Google Colab), to decrease computational time and improve asset efficiency.

11 | Conclusions

This inquiry has displayed a comprehensive system for the text-to-image era that addresses three basic challenges within the field: show execution optimization, creative fashion conservation, and copyright assurance. Through precise experimentation and examination, we have illustrated that our half breed approach combining Steady Dissemination, CLIP, and BLIP accomplishes predominant results compared to standalone models, with our SD2.1+WM setup appearing especially solid execution (F1=0.88 in somewhat covering datasets and 0.70 in disjoint scenarios).

11.1 | Key Commitments of this Work: Incorporate

Execution Optimization:

Our gathering strategy leverages the complementary qualities of diverse models, accomplishing a 40% diminishment in artifacts while maintaining high semantic arrangement (CLIP score > 0.85) and advancing era effectiveness to 1.1 seconds per picture in turbo mode. Style-Conditioned Era: The proposed artist-style conditioning module effectively duplicates aesthetic styles while keeping up

inventive astuteness, as proved by our human assessment scores (MOS 4.2/5 for fashion conservation). Vigorous Watermarking: PT-Mark's undetectable watermarking accomplishes 98 protection precision without compromising picture quality (SSIM > 0.95), tending to basic copyright concerns in AI-generated craftsmanship.

The exploratory comes about approve our scientific system for dissemination forms, especially the optimized denoising objective that maintains semantic devotion amid both forward and inverse diffusion phases. Our comparative investigation discovers that whereas no single demonstration exceeds expectations over all measurements, a vital combination of engineering components can create state-of-the-art outcomes.

Funding

This research has no funding source.

Data Availability

The datasets generated during and/or analyzed during the current study are not publicly available due to the privacy-preserving nature of the data, but are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that there is no conflict of interest in the research.

Ethical Approval

This article does not contain any studies with human participants or animals performed by any of the authors.

Author Contributions

All authors contributed equally to this work.

References

- [1] Brock, A., Donahue, J., & Simonyan, K. (2019). Large scale GAN training for high fidelity natural image synthesis. *Proceedings of the International Conference on Machine Learning (ICML)*. <https://arxiv.org/abs/1809.11096>
- [2] Chang, H., Zhang, H., Barber, J., Maschinot, A. J., Lezama, J., Jiang, L., ... Krishnan, D. (2023). Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*. <https://arxiv.org/abs/2301.00704>
- [3] Du, L., et al. (2025). ArtistAuditor: Auditing artist style pirate in text-to-image generation models. Preprint.
- [4] Hugging Face. (2023). DreamBooth training with Diffusers v0.11.0. <https://huggingface.co/>
- [5] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *Proceedings of the International Conference on Machine Learning (ICML)*. <https://arxiv.org/abs/2103.00020>
- [6] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*. <https://arxiv.org/abs/2204.06125>
- [7] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-shot text-to-image generation. *Proceedings of the International Conference on Machine Learning (ICML)*. <https://arxiv.org/abs/2102.12092>
- [8] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. *Advances in Neural Information Processing Systems (NeurIPS)*. <https://arxiv.org/abs/2112.10752>
- [9] Singh, R., Kaiser, L., & Google Research. (2022). Parti: Pathways autoregressive text-to-image generation. *arXiv preprint arXiv:2206.10789*. <https://arxiv.org/abs/2206.10789>
- [10] Turfill: adapting few-step text-to-image model for fast image inpainting. (2025). Preprint.
- [11] Xu, T., Zhang, P., Huang, Q., & Yang, M. H. (2018). AttnGAN: Fine-grained text-to-image generation with attentional generative adversarial networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. <https://arxiv.org/abs/1711.10485>
- [12] Xie, L., et al. (2025). TurboFill: Adapting few-step text-to-image model for fast image inpainting. Preprint.

- [13] Zhang, H., Xu, T., Li, H., Zhang, S., Huang, X., & Metaxas, D. N. (2017). StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. Proceedings of the IEEE International Conference on Computer Vision (ICCV). <https://arxiv.org/abs/1612.03242>

Disclaimer/Publisher's Note: The perspectives, opinions, and data shared in all publications are the sole responsibility of the individual authors and contributors, and do not necessarily reflect the views of Sciences Force or the editorial team. Sciences Force and the editorial team disclaim any liability for potential harm to individuals or property resulting from the ideas, methods, instructions, or products referenced in the content.