Journal of Theoretical and Applied Information Technology
30st June 2026. Vol.104. No.12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

MCMC-GUIDED STABLE DIFFUSION FRAMEWORK FOR IMPROVED TEXT-TO-IMAGE GENERATION WITH ENHANCED SEMANTIC ALIGNMENT

DR RASWITHA BANDI[1], B RAVALI REDDY[2] ,KAMBHAM PRATAP JOSHI[3] ,
M VANIAH SUCHARITHA SANTOSH[4] ,SUCHITRA PATTABIRAMAN[5] ,DR J VAMSINATH[6]

[1]Assistant professor, Department of IT, VNRVJIET, Hyderabad, India
[2]Assistant professor, Department of CSE, MVSR ENGINEERING COLLEGE, Hyderabad, India
[3]Assistant professor, Department of CSE, VNRVJIET, Hyderabad, India
[4,5]Assistant professor, Department of CSE, MALLA REDDY UNIVERSITY, Hyderabad, India
[6]Senior Assistant Professor, Department of Computer Science & Engineering, Faculty of Science and Technology (IcfaiTech), ICFAI Foundation for Higher Education, Hyderabad, India
*raswitha_b@vnrvjiet.in.

ABSTRACT

Text-to-image generation has gained significant attention due to its ability to synthesize realistic images from natural language descriptions. Despite recent advancements in diffusion-based models, challenges such as inefficient sampling, poor semantic alignment, and lack of consistency remain unresolved. This paper proposes an enhanced Stable Diffusion framework integrated with Markov Chain Monte Carlo (MCMC) sampling to improve image quality and text-image alignment. The proposed approach leverages CLIP-based scoring to guide the sampling process, ensuring that generated images better correspond to the given textual prompts. Additionally, optimization strategies such as classifier-free guidance and LoRA-based fine-tuning are incorporated to further enhance performance. Experimental results on benchmark datasets demonstrate that the proposed method achieves improved CLIP scores, higher image sharpness, and better semantic consistency compared to baseline diffusion and GAN-based approaches. The findings indicate that integrating probabilistic sampling techniques significantly enhances the effectiveness of text-to-image generation models.

Keywords: *Stable Diffusion, MCMC Sampling, Text-to-Image Generation, CLIP, Image Synthesis*

1. INTRODUCTION

1.1 Background

Text-to-image generation has emerged as a critical area of research in artificial intelligence, enabling the transformation of textual descriptions into realistic images. Recent advancements in diffusion models have significantly improved image quality; however, these models still suffer from challenges such as inefficient sampling, lack of semantic alignment, and high computational cost.

1.2 Existing Methods

Existing approaches, including Generative Adversarial Networks (GANs) and diffusion-based models, have demonstrated promising results but often struggle with maintaining consistency between text and generated images. Furthermore, deterministic sampling strategies limit the exploration of the latent space, leading to suboptimal outputs.

1.3 Proposed Approach

To address these limitations, this work introduces an MCMC-guided Stable Diffusion framework that enhances sampling efficiency and improves text-image alignment. The proposed method integrates probabilistic sampling techniques with CLIP-based evaluation to guide the generation process toward more accurate and high-quality outputs.

1.4 Contributions

The key contributions of this work are as follows:

- Introduction of an MCMC-guided sampling strategy for diffusion models
- Improved semantic alignment using CLIP-based scoring

• Integration of optimization techniques such as LoRA and classifier-free guidance

• Comprehensive evaluation demonstrating improved performance over baseline models

1.5 Problem Statement

Despite the success of diffusion models in text-to-image generation, several challenges remain. Existing models often produce images with weak semantic alignment, inconsistent details, and inefficient sampling processes. Additionally, the reliance on deterministic sampling limits the ability to explore diverse outputs. Therefore, there is a need for an improved framework that enhances sampling efficiency while maintaining strong alignment between textual input and generated images.

2. LITERATURE REVIEW

Research in text-to-image synthesis using GANs has experienced significant growth over the past decade. Foundational studies, such as those by Hong et al. [10] and Gulrajani et al. [11], provide a comprehensive review of GANs, tracing the evolution of architectural advancements, training methodologies, and diverse application domains. Early analyses such as Dash et al. [12] further highlighted instability issues in GAN training, noting that mode collapse had already emerged as a core limitation even in early GAN variants.

As noted by Salimans et al. [13], "one of the main failure modes for GANs is when the generator collapses to a parameter setting where it emits the same output for many inputs, and our proposed minibatch discrimination helps prevent this collapse." These works encompass various generative model architectures and demonstrate the application of GANs to challenges such as image inpainting, semi- supervised learning, and multimodal generation. Subsequent research has of focused on text-to- image synthesis, highlighting prominent models like StackGAN and AttnGAN [14], [15]. Ku & Lee demonstrated that integrating a dedicated regressor to better learn text-conditional vectors significantly improves the alignment between generated images and their corresponding textual descriptions [16].

While these studies elucidate the technical underpinnings and state-of-the-art performance specific architectures, they often exhibit a limited scope. A prevalent limitation is the insufficient exploration of mode collapse, which remains a critical obstacle in addressing the diversity issue of text-conditioned images. However, even models that enhanced textual correspondence and reconstruction quality, such as those presented by Wang et al. [17], did not explicitly address the underlying causes of mode collapse, leaving diversity issues largely unresolved.

Furthermore, there is a paucity of discussion regarding the causes or remedies for collapse within these frameworks [18]. This lack of targeted analysis indicates a gap in the literature, necessitating data to substantiate the presence and interventions for mode collapse in the context of text-to-image GANS.

Recent studies have explored various approaches for text-to-image generation, including GAN-based models and diffusion-based frameworks. While GANs provide high-quality outputs, they often suffer from training instability and mode collapse. Diffusion models, on the other hand, offer more stable training but face challenges related to computational cost and sampling inefficiency.

Several techniques such as CLIP-guided generation, classifier-free guidance, and LoRA-based optimization have been proposed to improve performance. However, these methods still lack effective mechanisms for exploring the latent space efficiently. The proposed work addresses this gap by incorporating MCMC-based probabilistic sampling into the diffusion process.

3. RESEARCH METHODOLOGY AND EXECUTION PROTOCOL

This study uses the MS-COCO dataset [23], which is a cornerstone in the field of computer vision. The dataset comprises about 330,000 images, with more than 1.5 million human-annotated captions. Each image in the dataset is accompanied by five unique text descriptions, which focus on particular aspects of the image, thus providing a comprehensive set of prompts that are useful in testing the image generation from text models. For the current study, a set of captions from the COCO validation set was selected, with the aim of using them as a JSON prompt list in the image generation process, with about 2,000 to 5,000 captions used in the study. The use of a wide range of prompts allows us to evaluate the level of generality of the models as well as the level of correspondence between the text prompts and the images that are generated.

MS-COCO consists of three sections: training, validation, and test as shown in Table 1[3][7]. The training set contains 118,287 images. There are about 591,435 captions in the training set. This set is used to pretrain a model named Stable Diffusion [8]. In our paper, we don't use this set to train a model.

Instead, we use a pre-trained model. The validation set contains 5,000 images. There are about 25,000 captions in the validation set. This set is used to generate prompts and fine-tune techniques. The test set contains 5,000 images. There are about 25,000 captions in the test set as shown in Table 1[3][7]. This set is used to test the performance of all the techniques. Prompts are obtained from a dataset named test. This dataset contains images. MS-COCO consists of three sections: training, validation, and test. This division helps to maintain all performance evaluations unbiased. Outputs obtained from all techniques depend on the generalization of a technique.

Table: 1 Coco Dataset Train - Test Split Table

Split	Images	Captions	Purpose
Train (COCO 2017)	118,287 (subset 60,000 used)	~591,435	LoRA fine tuning
Validation (COCO 2017)	5,000	~25,000	Prompt selection & evaluation
Test (COCO 2017)	5,000	~25,000	Final evaluation & comparison

The methodology adopted in the present study offers a complete framework for assessing different text-to-image generation strategies through a well-structured multi-stage framework that includes prompt preparation, generation, evaluation, and visualization of results. Overall, the methodology adopted in the present study aims at systematically comparing different text-to-image generation techniques based on the stable diffusion framework. Initially, text prompts are selected from the MS COCO caption dataset, which ensures that the text prompts prepared are diverse and descriptive in nature. These text prompts are then utilized for generating images through different text-to-image generation techniques.

Once the images are generated, different quantitative metrics are computed to assess the quality of the results obtained. These images along with their metrics are stored, and then the results are visualized through a research dashboard. The study employs a pre-trained Stable Diffusion model provided in the Hugging Face Diffusers library as a baseline. It generates images directly from text prompts without any additional optimization tricks. The generation process begins with converting the text prompts into embeddings, followed by adding random latent noise, denoising the latent representation repeatedly, and finally decoding it into an image. This baseline serves as a reference to evaluate the effectiveness of new generation techniques. To enhance the adaptability of the model, Low-Rank Adaptation.

(LoRA) is incorporated into the generation process. LoRA incorporates low-rank matrices into the attention layers of the diffusion network to enable efficient fine-tuning with a reduced number of trainable parameters, computational cost, and adaptability to specific generation tasks. Classifier-Free Guidance (CFG) is used in diffusion sampling to increase the degree of guidance provided by the text prompt on the result. The model generates conditional predictions based on the prompt and unconditional predictions based on the prompt itself and combines them using a guidance scale. The guidance scale increases prompt alignment;

however, if it is set too high, it can have a detrimental impact on image diversity. To further refine semantic alignment, reranking using CLIP is used. For each prompt, several images are generated and embedded using the image encoder of CLIP, and the prompt is embedded using the text encoder of CLIP. The image is selected based on the highest cosine similarity between image and text embeddings. A seed search strategy is utilized as a strategy for dealing with the random impacts of seed initialization in diffusion models. Various seed values are tried for every prompt, and images are created for every seed value, then scored to select the best seed value that produces the best result, thus increasing the likelihood of receiving high-quality results.

Journal of Theoretical and Applied Information Technology
30st June 2026. Vol.104. No.12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

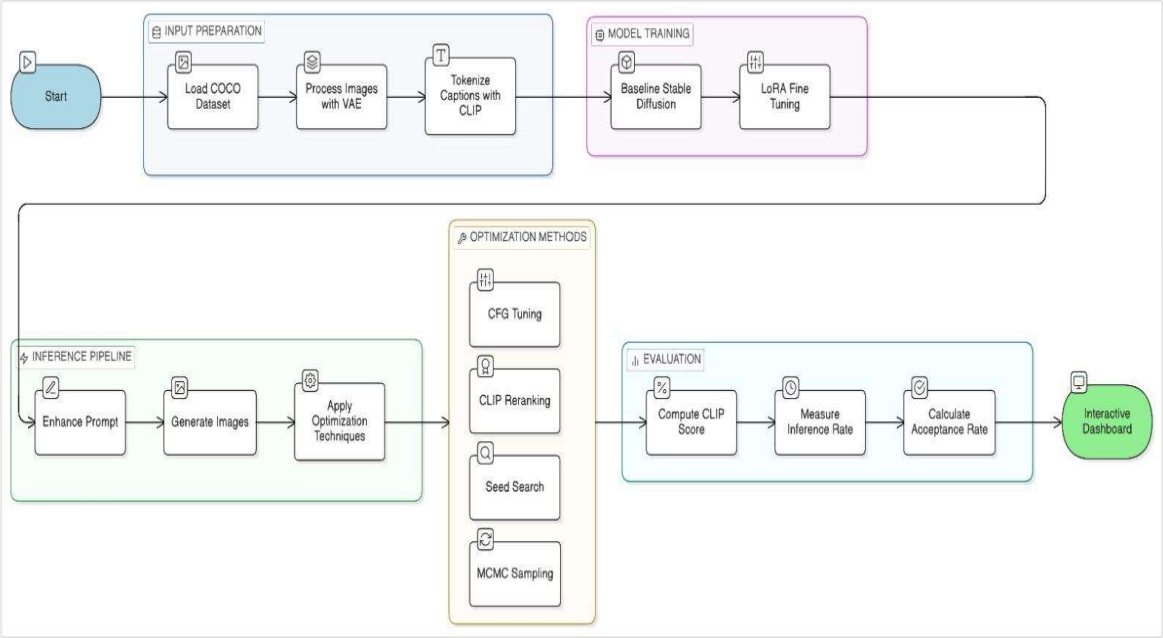

Fig (1). The Pipeline Of Proposed Method

To further refine the sampling process, Markov Chain Monte Carlo (MCMC) sampling is incorporated, whereby images are created sequentially and assessed based on their quality, such as the CLIP score and sharpness of the image, and an acceptance probability determines whether a new image replaces the previous image in the sampling process, effectively exploring the image space efficiently and progressively increasing image quality, represented by the acceptance rate. Classifier-Free Guidance (CFG) is used in diffusion sampling to increase the degree of guidance provided by the text prompt on the result. The model generates conditional predictions based on the prompt and unconditional predictions based on the prompt itself and combines them using a guidance scale. The guidance scale increases prompt alignment; however, if it is set too high, it can have a detrimental impact on image diversity. To further refine semantic alignment, reranking CLIP is used. For each prompt, several images are generated and embedded using the image encoder of CLIP, and the prompt is embedded using the text encoder of CLIP. The image is selected based on the highest cosine similarity between image and text embeddings. A seed search strategy is utilized as a strategy for dealing with the random impacts of seed initialization in diffusion models. Various seed values are tried for every prompt, and images are created for every seed value, then scored to select the best seed value that produces the best result, thus increasing the likelihood of receiving high-quality results. To further refine the sampling process, Markov Chain Monte Carlo (MCMC) sampling is incorporated, whereby images are created sequentially and assessed based on their quality, such as the CLIP score and sharpness of the image, and an acceptance probability determines whether a new image replaces the previous image in the sampling process, effectively exploring the image space efficiently and progressively increasing image quality, represented by the acceptance rate. Lastly, we compare the performance of each of the generation strategies using a wide range of metrics. We compared the CLIP Score for text-image alignment, Fréchet Inception Distance (FID) for the realism of the images, sharpness of the images for clarity, entropy for the complexity of the images, inference time for the computational cost of the images, and the acceptance rate for the dynamics of the MCMC. Overall, we have a wide range of metrics for the quality and the efficiency of the images.

3.1 Stable Diffusion Background

Stable Diffusion [8] is a type of latent diffusion model (LDM) [7] that aims to efficiently generate high-resolution images based on text prompts. Unlike traditional diffusion models, Stable Diffusion [8] does not directly perform the diffusion process in pixel space [7], thus reducing the computation required without compromising the
quality of the generated images [7]. The Stable Diffusion model [8] structure includes three main

parts, namely, the Variational Autoencoder (VAE), UNet [7] Denoising Network, and the text encoder, which uses the CLIP model as shown in Fig. (2)[7][8]. The Variational Autoencoder (VAE) is in charge of compressing the images into a lower-dimensional representation. This allows the diffusion process to take place in the latent space as opposed to the pixel space.

The encoding process can be formulated as follows:

$$z = Encoder(x) \quad (1)$$

Where,

x: x is the image, and

z: z is the representation of the image in the latent space.

The decoding process is the process of reconstructing the image from the latent space and is calculated using Eq. (2). The decoding process can be formulated by using Eq. (1) as follows:

$$x\hat{} = Decoder(z) \quad (2)$$

where:

$x\hat{}$ is the reconstructed image.

The process of operating in the latent space reduces the memory and computational complexity. The main part of the diffusion model consists of the denoising network, which is composed of the UNet [7] as shown in Fig. (3)[7][8]. It predicts the noise present in the representation at each step of the diffusion process. In the training process, noise is progressively added to the image using the forward diffusion process, which is then predicted by the UNet [7]. In the generation process, the reverse diffusion process is performed, which progressively denoises the noise to produce an image.

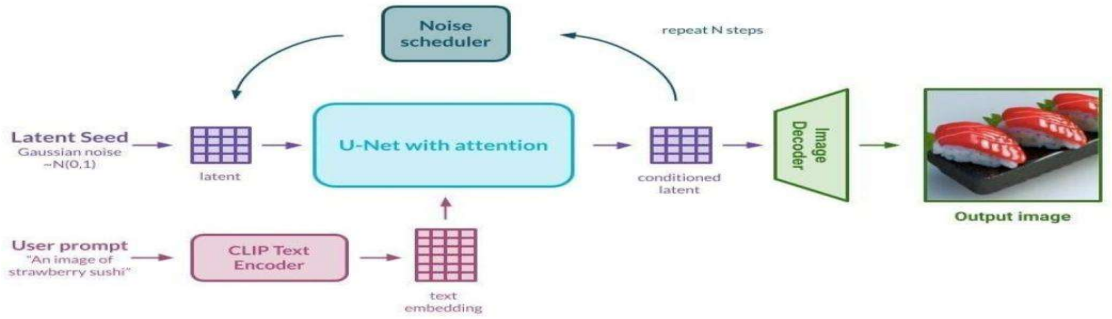

Fig (3): Diffusion Latent Model With U-Net And CLIP Text Conditioning

3.2 Metropolis-Hastings Algorithm

One of the most frequently used MCMC algorithms is the Metropolis-Hastings algorithm [22]. The acceptance probability is given by the following

equation: $$\alpha = min(1, exp(-(E\_new - E\_current) / T)) \quad (4)$$

where:

E_new = energy of the candidate sample

E_current = energy of the current sample

T = temperature parameter

If the candidate sample has lower energy, it is more likely to be accepted.

3.3 Proposed MCMC-Guided Stable Diffusion

The suggested framework will utilize MCMC sampling [22] in the diffusion generation process in the following way: It will start by generating an initial image using Stable Diffusion [8]. It will then compute CLIP similarity between the prompt and generated images [3]. It will then generate a new candidate image using a different seed. It uses a CLIP-based energy function to evaluate this new candidate. It will then accept or reject this new candidate image using an MCMC acceptance rule. It will then repeat this process a number of iterations. It will then return the best image from this process. This iterative process allows the model to refine generated images and explore alternative latent representations.

Energy-Based Guidance Function:

In order to measure the semantic alignment between images and input prompts, the proposed method makes use of an energy-based function.

The proposed energy-based function is as follows:

$$\boldsymbol{E(image, text) = -CLIP(image, text)} \quad (5)$$

where:

CLIP represents the cosine similarity between the image and text embeddings.

Lower energy values correspond to stronger semantic alignment between the generated image and the input prompt. Therefore, the MCMC [22] process attempts to minimize the energy function.

3.4 Algorithm

MCMC-Guided Stable Diffusion Sampling
Input: Text prompt P, number of MCMC iterations N
Output: Generated image $I*$

1) Encode the text prompt P using the CLIP text encoder to obtain text embedding ep.
2) Generate an initial image $I0$ using the Stable Diffusion model.
3) Compute the initial energy using Eq.(5):

$$E0 = -CLIP(P, I0)$$

4) for k = 1 to N do
a) Generate a candidate image Ik using Stable Diffusion with a new random seed.
b) Compute the candidate energy using Eq. (5):

$$Ek = -CLIP(P, Ik)$$

c) Compute the acceptance probability using Eq. (4):

$$\alpha = min(1, exp(-(Ek - Ek - 1)))$$

d) Sample $u \sim Uniform$ (0,1)
e) if $u < \alpha$ then
Accept Ik as the current sample.
Set $Ek - 1 = Ek$
f) else
Reject Ik and keep the previous sample.
5) end for
6) Return the best accepted image $I*$.

The MCMC sampling process for the Stable Diffusion model [8] is illustrated in Algorithm. This process begins by using the CLIP text encoder to encode the prompt and capture its semantic meaning. An initial image is then generated using the Stable Diffusion model, and its semantic similarity to the prompt is measured using the CLIP energy function. For each step of the MCMC process [22], a new candidate image is generated by modifying the random seed. We measure the goodness of the new candidate image relative to the prompt using the same energy function. Finally, we accept or reject the new candidate image using the Metropolis-Hastings criterion, where images with lower energy (i.e., more aligned semantics) are more likely to be accepted. If the candidate image has survived the test, it will replace the old image in the Markov chain; otherwise, we will retain the old image. This loop-style sampling process will allow the system to "wander" across various regions in the latent space and continually improve the quality of the image. After we have completed the designated number of iterations, we will obtain the best image generated, which will be the final result of the entire MCMC-assisted diffusion process.

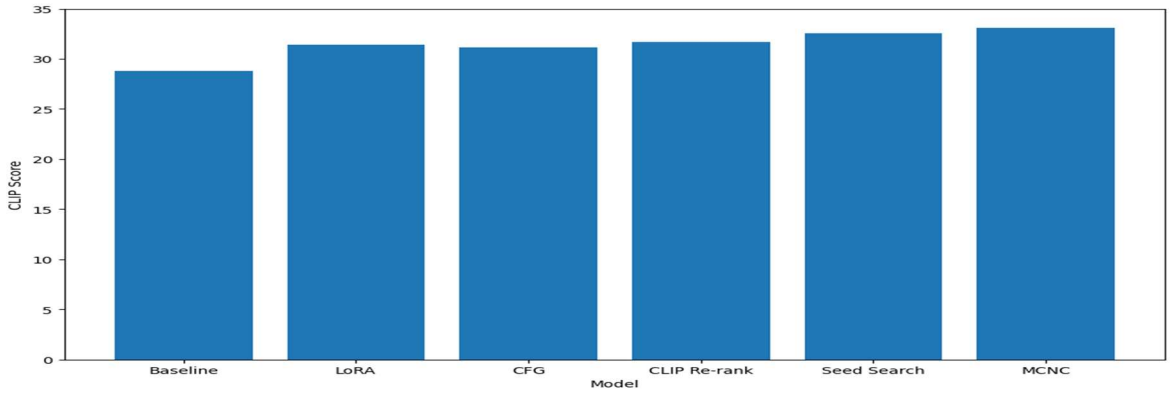

Fig (4). The Comparison Of CLIP Score Of All The Methods That Were Evaluated

4. RESULT AND DISCUSSION

4.1 Accuracy

Various quantitative measures were used to assess the performance of the proposed and comparative methods of text-to-image generation, such as CLIP Score, sharpness, entropy, and inference time. Prompts of MS-COCO were used to conduct the experiments to evaluate image quality and semantic alignment. A semantic similarity of the generated images and input text prompts was measured using the CLIP Score. In Fig. (4), Table. (2)[3][7] the obtained results, the proposed MCMC-guided [22] Stable Diffusion model had the highest CLIP Score of 33.062, then the seed search [7][8] (32.546), and lastly, CLIP reranking [3] (31.645). The Stable Diffusion model [8] used as the baseline had the lowest CLIP Score (28.779) which means that it is less aligned to text descriptions. Sharpness, entropy was studied in terms of image quality. The LoRA-

based solution [21] had the most sharpness (2041.89) to give very detailed images, but the MCMC approach also reached the competitive sharpness (1601.83) with the highest entropy (7.91) to indicate more content in images with more different colors.

4.2 Performance

The performance comparison of different text-to-image synthesis approaches is presented in Table. (2)[3][7]. The results show that diffusion-based methods significantly outperform GAN-based

approaches in terms of stability, image quality, and semantic alignment. However, GAN models demonstrate faster inference speed compared to diffusion-based methods. Among the evaluated diffusion-based techniques, methods incorporating multiple sampling strategies, such as CLIP reranking [3], seed search [7][8], and MCMC [22], show improved performance compared to the baseline Stable Diffusion model [8]. Specifically, the MCMC-guided approach demonstrates the best overall performance by achieving the highest CLIP Score and improved entropy, indicating better semantic alignment and diversity. However, these improvements come at the cost of increased computational time. As shown in the Fig.(5),Table.(2)[3][7] the inference time for CLIP reranking[3], seed search[7][8], MCMC[22], Baseline(SD), LoRA [21] and CFG [20] methods were **29.14 s**, **72.92 s**, **80.41 s**, **7.95s, 7.22 s** and **7.22 s** respectively, which are significantly higher compared to the baseline approach.

Table.(2). Comparative Performance Evaluation Of Text-To-Image Generation Methods

Method	CLIP Score	Sharpness	Entropy	Inference Time(s)
Baseline (SD)	28.77	1105.6	7.81	7.95
LoRA	31.43	2041.89	7.46	7.22
CFG	31.14	951.9	7.68	7.22
CLIP Rerank	31.64	1504.14	7.86	29.14
Seed Search	32.54	1014.4	7.68	72.92
MCMC	33.06	1601.83	7.91	80.41

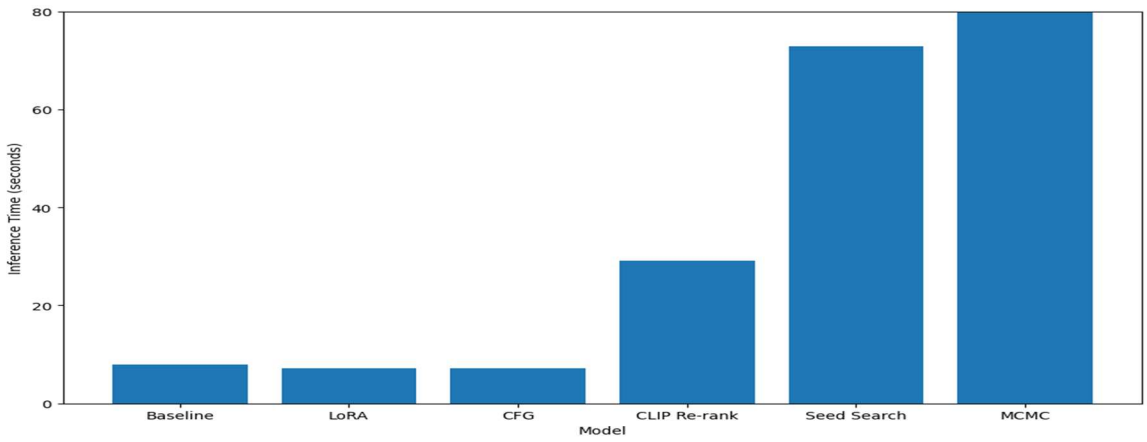

Fig. (5) Comparative Performance Evaluation Of Text-To-Image Generation Methods

This paper among other things compares the performance of different sampling methods in enhancing text to picture generation. The findings show that diffusion models have a higher quality [7][8] in images and semantic compatibility with GAN-based models [5][6] as shown in the Table. (3)[5][7][8]. This has been made more achievable by the iterative mechanism of denoising of diffusion models that allows more detailed and realistic images to be generated. The method with the best semantic alignment was the MCMC-guided sampling approach, as shown by CLIP Score. The reason is that the MCMC [22] framework facilitates the exploration of the latent space in a probabilistic way, which can enable the model to update candidate images with the help of CLIP-based energy functions. On the same note, performance is also enhanced in seed search [7][8] as the best latent initializations are picked during search.

Table. (3). Comparison Between GAN And Diffusion Models

Feature	GAN	Diffusion
Stability	Low	High
Quality	Medium	High
Alignment	Weak	Strong
Speed	Fast	Slow

The method that used LoRA [21] generated high sharpness values and high detailed images, whereas CFG [20] generated better guidance as it generated, but not much of an improvement over the probabilistic ones as shown in the Table. (2)[3][7]. These findings indicate that although guidance techniques are more effective in conditioning, probabilistic sampling methods are more effective in exploration and refinement.

4.3 Metric Analysis

The metrics of evaluation also offer additional results about the performance of the various models. The CLIP Score validates that semantic alignment between text and generated images is enhanced with the help of probabilistic techniques, including MCMC and seed search [7][8]. The values of the entropy refer to the fact that MCMC is able to produce more diverse outputs than other methods. Diffusion-based methods incur much more inference time than GANs because they are iterative in nature. The MCMC technique also consumes more computation time owing to repetitive sampling and assessment techniques. The compromised quality of the image and alignment notwithstanding, the better quality of the image and alignment warrants the extra cost of computation in applications where accuracy matters. Overall, the findings indicate that optimization of the sampling process in diffusion models can greatly increase the text-to-image generation results without changing the underlying model architecture.

5. CONCLUSION AND FUTURE WORK

In this study, we present a comprehensive study on the improvement of the performance of the text-to-image synthesis models by enhancing the sampling strategies of the diffusion models. Although the Stable Diffusion model [8] has shown impressive results in the synthesis of high-quality images, the sampling strategy does not always guarantee the best semantic alignment with complex input texts. To overcome the limitations of the Stable Diffusion model, we have explored different optimization strategies such as LoRA adaptation, classifier-free guidance [20], CLIP reranking [3], seed search [7][8], and a probabilistic MCMC-guided sampling strategy. In the proposed approach, the MCMC-guided strategy was introduced with a probabilistic refinement strategy that allows for the exploration of the latent space more effectively. The CLIP energy function was utilized to evaluate the generated images and select those that show the best alignment with the input text. This allows the system to move beyond the deterministic approach and improve the quality of the generated image iteratively. The experimental results on the MS-COCO dataset show that the proposed approach with the probabilistic methods such as seed search [7][8] and MCMC-guided sampling achieved higher

CLIP score values compared to the baseline Stable Diffusion model [8] for image synthesis. Moreover, the proposed approach achieved competitive results with the baseline approach for image quality metrics such as sharpness and entropy with a manageable increase in computational cost. This study emphasizes the importance of the sampling strategy in the diffusion models and highlights the potential of the proposed approach for improving the performance of the image synthesis models by the inclusion of a probabilistic approach, such as the MCMC strategy, without the need for retraining the models or making changes to the underlying architecture of the models.

Author Contributions:

Kambham Pratap Joshi and Ravali Reddy conceptualized and designed the study, conducted data collection, and participated in data analysis and interpretation. Dr Raswitha Bandi contributed to the development of the educational media, oversaw the implementation of the intervention, and contributed to manuscript writing and revisions. Vaniah Sucharitha Santosh, Suchitra Pattabiraman and Dr J Vamsinath assisted with data analysis and interpretation and provided critical feedback on the manuscript. All authors reviewed and approved the final version of the manuscript and agreed to be responsible for all aspects of the work, ensuring integrity and accuracy.

REFERENCES:

[1] S. K. Alhabeeb and A. A. Al-Sharabi, "Text-to-image synthesis with generative models: Methods and challenges," *IEEE Access*, vol. 12, pp. 24412–24429, 2024.

[2] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, H. Laga, and M. Bennamoun, "Text-to-image synthesis for improved image captioning," *IEEE Access*, vol. 9, pp. 64918–64934, 2021.

[3] A. Radford et al., "Learning transferable visual models from natural language supervision," in Proc. ICML, 2021.

[4] D. Jin, Q. Yu, L. Yu, and M. Qi, "SAW-GAN: Multi-granularity text fusion generative adversarial networks," *Knowl. -Based Syst.*, vol. 294, p. 111795, 2024.

[5] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2014, pp. 2672–2680.

[6] S. Frolov, T. Hinz, F. Raue, J. Hees, and A. Dengel, "Adversarial text-to-image synthesis: A review," *Neural Netw.*, vol. 144, pp. 187–209, 2021.

[7] J. Ho, A. Jain, and P. Abbeel, "Denoising stable diffusion probabilistic models," in *Proc. NeurIPS*, 2020.

[8] R. Rombach *et al.*, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022.

[9] D. Liu, L. Y. Wu, B. Li, Y. Zhao, Z. Ge, and J. Zhang, "T-Person-GAN: Text-to-person image generation with identity-consistency and manifold mix-up," *Expert Syst. Appl.*, vol. 288, p. 128178, 2025.

[10] Y. X. Tan, J. Y. Lim, K. M. Lim, and C. P. Lee, "LAP-GAN: Label augmentation with perceptual loss for self-supervised text-to-image synthesis," *Expert Syst. Appl.*, vol. 296, p. 129005, 2026.

[11] Y. Hou, W. Zhang, Z. Zhu, and H. Yu, "Language-vision matching for text-to-image synthesis with context-aware GAN," *Expert Syst. Appl.*, vol. 255, p. 124615, 2024.

[12] D. Jin, G. Li, Q. Yu, L. Yu, J. Cui, and M. Qi, "GMF-GAN: Gradual multi-granularity semantic fusion GAN for text-to-image synthesis," *Digit. Signal Process.*, vol. 140, p. 104105, 2023.

[13] Y. Cai *et al.*, "DualAttn-GAN: Text-to-image synthesis with dual attentional generative adversarial network," *IEEE Access*, vol. 7, pp. 183706–183719, 2019.

[14] Z. Zhang and L. Schomaker, "DiverGAN: An efficient and effective single-stage framework for diverse text-to-image generation," *Neurocomputing*, vol. 473, pp. 182–198, 2022.

[15] H. Zhang, H. Zhu, S. Yang, and W. Li, "DGattGAN: Cooperative up-sampling based dual generator attentional GAN," *IEEE Access*, vol. 9, pp. 29584–29600, 2021.

[16] H. Lin, Q. Chen, C. Liu, and J. Hu, "TDG-Diff: Advancing customized text-to-image synthesis with two-stage diffusion guidance," *Comput. Graph.*, vol. 124, p. 103889, 2024.

[17] F. Barrientos-Espillco, G. Pajares, J. A. Lopez-Orozco, and E. Besada-Portas, "Customization of text-to-image diffusion models by fine-tuning for synthetic image generation," *Expert Syst. Appl.*, vol. 287, p. 128169, 2025.

[18] Y. Wang, R. Liu, X. Xie, L. Wang, Z. Yi, and R. Ma, "DP-Adapter: Dual-pathway adapter for boosting fidelity and text consistency," *Graph. Models*, vol. 141, p. 101292, 2025.

[19] B. Huang and H. Xie, "PromptNavi: Text-

to-image generation through interactive prompt visual exploration," *Comput. Graph.*, vol. 132, p. 104417, 2025.

[20] J. Ho and T. Salimans, "Classifier-free diffusion guidance," in *Proc. NeurIPS Workshops*, 2021.

[21] E. J. Hu *et al.*, "LoRA: Low-rank adaptation of large language models," in *Proc. ICLR*, 2022.

[22] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.

[23] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. ECCV*, 2014, pp. 740–755.

[24] R. Gopalakrishnan, S. Naveen, S. Kalathil, and P. V. Sudeep, "Self-attention-based text encoder for enhancing DMGAN," *IEEE Access*, vol. 13, pp. 125442–125456, 2025.

[25] M.A. Habib *et al.*, "GACNet: Text-to-image synthesis using attention mechanisms with contrastive learning," *IEEE Access*, vol. 12, pp. 9572–9592, 2024.

[26] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2021.

[27] A. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in Proc. Int. Conf. Mach. Learn. (ICML), 2021.

[28] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in Proc. Int. Conf. Learn. Representations (ICLR), 2021.

[29] K. Crowson, S. Biderman, D. Hall, et al., "VQGAN-CLIP: Open domain image generation and editing with natural language guidance," arXiv preprint arXiv:2204.08583, 2022.

[30] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in Proc. Int. Conf. Learn. Representations (ICLR), 2019.