DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Explainable Photorealistic Image Synthesis Using Diffusion Model With Multi-Scale Feature Fusion

Sonal Fatangare[1], Premanand Ghadekar[2]

[1]Department of Computer Engineering, Vishwakarma Institute of Technology, (Savitribai Phule Pune University), Pune, Maharashtra, India, Email: sonalfatangare82@gmail.com
[2]Department of CSE-AIML, Vishwakarma Institute of Technology, (Savitribai Phule Pune University), Pune, Maharashtra, India

Abstract
Generative models, that generate images from scratch, have lately drawn a lot of attention. The diffusion models are especially popular for their training process and their excellent noise modelling capabilities. However, there is still challenging for conditional text image synthesis. In the proposed system the basic level architecture of U-Net is redesigned with attention and residual blocks for capturing complex features of images. Along with this, use of multi-scale feature fusion technique helps to handle images of different resolutions. Diffusion Transformer and cross-modal attention mechanisms enhance realism and coherence in image quality. The model generates the photorealistic image from the latent representation with VAE decoder. Furthermore, employment of Grad-CAM ensures the model is clearer and gives insights on what portions of the image the model highlighting to during its generation process. In comparison with both the basic diffusion model and GANs, the developed diffusion model improved significantly in photorealism, detail sharpness, and numbers like FID and PSNR to be much greater than those of both.

Keywords: Generative Models, Diffusion Models, Photorealistic Image Generation, Grad-CAM, U-Net Architecture

Introduction

Photorealistic image creation has observed great advancements over the recent years, especially on various fonts, such as digital art, advertising, and even virtual reality. Despite this, there has been only one technique that has remarkably stood out with attention grabbing qualities, and that technique is Generative Adversarial Networks (GANs), as it has been increasingly attracting researchers with immense potential. [1]. GANs work on two network methods with one as the generator which creating images and the other as discriminator grading those images, hence doing a competitive training process. But GANs have relevant challenges such as mode collapse in which the generator produces a restricts output diversity and unstable training dynamics makes them difficult in work. Furthermore, GANs tend to generate visually quite pleasing images but fail to capture fine details and textures often; therefore, in certain situations, the outputs will be less realistic.

However, diffusion models offer an interesting alternative. These models employ a series of steps to transform random Gaussian noise into something meaningful in order to generate images. This method has a great deal of promise because it generates a variety of images and is stable during the training process. Diffusion models can be improved, though. Sometimes, they fail to measure up to GANs in terms of resolution and photorealism, which reveals a great need for improvements that can bring the current diffusion models up to the level of these popular models. [6]. Although some recent studies have made a great improvement in the diffusion techniques, there is still a great need for effort in the improvements that leverage the strength of both methods while addressing their weaknesses.

This research provides a number of important contributions to image synthesis. This can be accomplished through an improved U-Net structure that encompasses two important contributions: attention mechanisms, which improve the feature extraction process for better information extraction, and residual blocks, which enable a more effective representation of feature maps. Additionally, the proposed approach also investigates an appropriate spatial fusion of multi-scale features in order to properly address resolutions and the extraction of fine details. The Diffusion Transformer Block is responsible for refining images through denoising with self-attention for detail preservation and incorporates text embeddings to guarantee semantic alignment during image synthesis. Cross-Modal Attention aligns textual and visual modalities for coherent outputs, leveraging Sinusoidal Timestep Embeddings and ROPE to ensure spatial and temporal consistency. Lastly, the proposed model also integrates classifier-free guidance to enhance the model's ability in producing images from textual prompts. Finally, the model

utilizes the Grad-CAM to gain insights on the attention mechanisms used in generating images. In doing so, the proposed solution tries to focus on building gap between diffusion models and GANs, offering a more robust option for high-quality synthesis while being interpretative.

Literature Review

A Despite remarkable achievements, challenges remain in realistic image generation, as demonstrated by prior studies. V. Amar et al. [1], demonstrates GAN-based framework for aligning Text to image for emphasizing advancement in image quality. Wang, Zheng et al. [2], developed a Diffusion GAN model for improving better convergence and image quality. The adaptive diffusion process and timestep dependent discriminator improve the stability and performance. StyleGAN-based improvements [3] enhanced visual quality through style-based latent manipulation, yet semantic alignment with text remained challenging. To address these limitations, diffusion models emerged as a dominant paradigm.

While early models had issues regarding stability, diffusion models developed as a stronger alternative by providing iterative denoising processes that enhance generation stability and realism. Ho et al. [4] proposed cascaded diffusion models for generating high-fidelity images at multiple resolutions and thus built the foundation of modern diffusion pipelines. Latent diffusion models [5],[6],[7] followed this work to develop the idea that compressing images to a latent space using VAEs explicitly allows efficient diffusion while preserving perceptual quality. These directly motivate using a VAE encoder–decoder in the proposed framework to get compact latent image representations for diffusion-based refinement.

Hence, semantic alignment between the text prompts and the generated images has been greatly updated due to the advent of strong text encoders and cross-modal learning approaches. Research by Radford et al. [8] and Raffel et al. [9] familiarized CLIP and T5, respectively, indicating the significance of strong semantic encodings for cross-modal-based learning approaches. Subsequent text-to-image models such as DALL·E [10] and Stable Diffusion [11] utilized the strong text encoders for the generation of the final image. Contemporarily, personalization approaches proposed by Hidalgo et al. [12] indicated that adjustment of the generation behavior based on specific domains was possible through the application of fine-tuning techniques of the diffusion model. Kumar et al. [13] have further explored the creative generation of tasks such as text-to-sketch conversion with a focus on semantic abstraction rather than photorealism.

From an architectural point of view, the role of U-Net backbone for diffusion models is already widely accepted for its efficacy in the representation and modeling of spatial hierarchies. However, unlike prior uses of the U-Net architecture with the simple skip connection method for feature integration, the current body of research underscores the use of multiscale feature integration for a better preservation of global structure and local details. Another corresponding body of research in the area of spatial modelling and rendering asserts the efficacy of multiscale representation for image synthesis [14],[15],[16]. In this regard, the proposed architecture incorporates an enhanced version of the U-Net architecture with an attention-based feature integration method.

Finally, beyond quantitative evaluation, recent works have highlighted the importance of explainability in vision-based AI systems. Grad-CAM and related XAI techniques have been successfully applied to interpret CNN and transformer decisions in sensitive domains [17],[18],[19]. Inspired by these efforts, the proposed work integrates Grad-CAM–based analysis as a post-hoc interpretability module to visualize salient regions in generated images, complementing standard metrics such as PSNR, SSIM, IS, and FID. Overall, existing literature motivates a unified framework that combines latent diffusion, transformer-based semantic guidance, multiscale U-Net fusion, and explainable evaluation forming the basis of the proposed methodology.

Methodology

The main reason for this technology is the growing need to create accurate, flexible, and meaning-filled images from written descriptions.

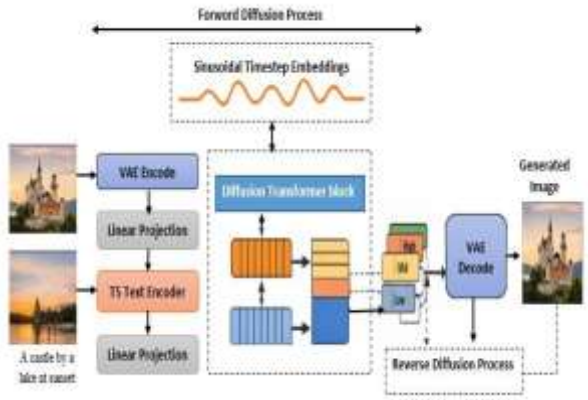

Fig. 1 The foundational idea behind the Proposed Methodology

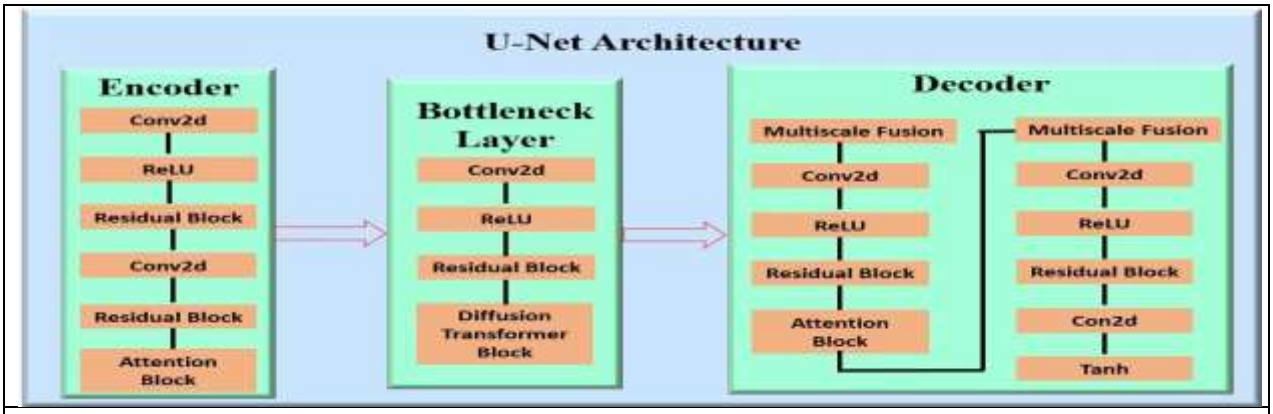

Fig. 2 Enhanced U-Net Architecture for Proposed Methodology

This improvement will greatly help various fields such as AI art, game development, visual storytelling, and human AI interaction. However, it is still very difficult to perfectly match generated images with the specific meanings of text descriptions.

Many current methods use late fusion, which means they combine image and text representations at the end. Unfortunately, this usually results in a weak connection between generated images and textual meanings.

This system will combine sophisticated methods such as U-Net, T5, VAE, and diffusion transformers to provide high-quality image synthesis with semantic consistency, preservation of details, and interpretability using Grad-CAM visualization.

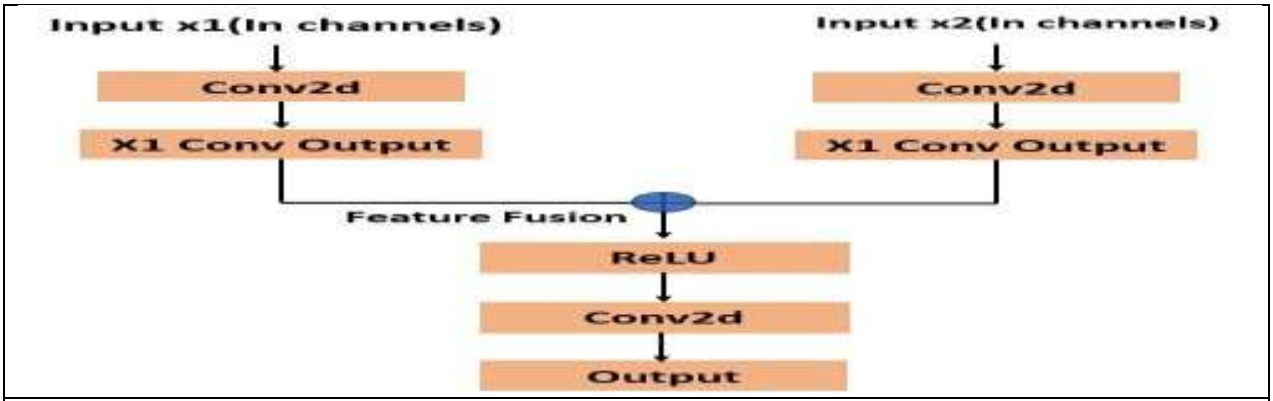

Fig. 3 Multiscale Fusion layer in U-Net Architecture

The multi-model approach in this application integrates U-Net and attention mechanism, VAE encoder, for extracting multilevel features that balances minute details and high-level semantic structures, diffusion-based transformer, which enables high-quality image generation. Semantic matching is achieved through cross-model attention between T5 text embeddings and latent image representations within the diffusion transformer blocks. Diffusion transformers enable iterative refinement, while Grad-CAM provides interpretability through visualization of attention focus for debugging and validation. This holistic integration ensures high performance

with highly preserved semantic and structural information. In Figure 1 & 2 shows the detailed representation of the proposed approach using the improved U-Net architecture. Figure 3 illustrates the flow from the two input feature maps to the convolutional layers, feature fusion, activation, and final output.

1.1 Input Processing Model

Input processing plays an important role in establishing the connection between the text prompts and the images. In this input processing, it is utilizing the combination of the pretrained T5 encoder [9] for Text understanding. The input image is encoded into a compact latent space using a pretrained Variational autoencoder (VAE) [20]. The outputs go through linear projection layers, hence transforming the latent space representations to enable the model for its process of generating the images. Another context introduced is the Sinusoidal Timestep Embeddings, which assist in the encoding of the temporal evolution of the diffusion process, hence facilitating the iterated denoising [21]. These embeddings assist the model in the refinement of features in a coherent fashion at every step, hence facilitating a accurate transition from the randomness to the final visual.

Let $x \in \mathbb{R}^{H\times W\times C}$ denotes real and t denotes the corresponding textual input. The generative process consists of the following stages:

$$p(\hat{x} \mid t) \quad (1)$$

Where $\hat{x}$ is the synthesized image semantically aligned with t.

The input image is encoded into a compact concise latent space via pre-trained VAE:

$$z_0 \sim q_\varphi(z|x) = N(\mu_\varphi(x), \sigma_\varphi^2(x)) \quad (2)$$

The reparameterization trick yields:

$$z_0 = \mu_\varphi(x) + \sigma_\varphi(x) \odot \epsilon, \qquad \epsilon \sim N(0, I) \quad (3)$$

This latent space enables efficient diffusion and multiscale processing.

The textual prompt is encoded using a pre-trained T5 encoder:

$$e_t = \mathrm{T5}_{enc}(t) \quad (4)$$

where $e_t \in \mathbb{R}^{L\times d_t}$ captures contextual and semantic dependencies across the input sequence. To align heterogeneous modalities, linear projections are applied:

$$\tilde{z}_0 = W_z z_0, \tilde{e}_t = W_t e_t \quad (5)$$

where $W_z \in \mathbb{R}^{d\times d_z}$ and $W_t \in \mathbb{R}^{d\times d_t}$.

A forward diffusion process slowly increases random noise into the latent space:

$$z_t = \sqrt{\alpha_t}\tilde{z}_0 + \sqrt{1-\alpha_t}\epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, I) \quad (6)$$

with $t = 1,2,\dots,T$ and Sample Gaussian noise $\epsilon_t \sim \mathcal{N}(0, I)$

Sinusoidal timestep embeddings τ_t are added:

$$h_t = z_t + \tau_t \quad (7)$$

1.2 Input Processing Model

Multimodal fusion takes textual and visual information to form the required latent representation semantically aligned with the input text. Double-Stream Processing is the framework under which data is processed; it consists of two streams handling the input in parallel: one associated to the textual information and the other to the latent information.

Latent Stream: The noisy vector created during the previous stage must now pass through the latent linear projection layer to prepare the noise for integration with the textual information of interest. During this, the Text Stream computes the text embeddings that are further allowed to pass through a text linear projection to align with latent features. It allows the fusion of features from both the text and visual representations without conflict. The Diffusion Transformer is the critical module here which enables the potential of having cross-modal attention between the two streams at both semantic and visual levels. For reverse diffusion with multiscale fusion calculated in equation (8), (9), (10), (11) and (12). Extract multiscale latent features:

$$\{z_t^{(s)}\}_{s=1}^{S} \leftarrow \mathcal{D}_s(h_t) \quad (8)$$

Each scale is Project scale-specific features:

$$\hat{z}_t^{(s)} = W_s z_t^{(s)} \quad (9)$$

Scale-wise attention weights are computed:

$$\alpha_s = \frac{\exp\left(g(\hat{z}_t^{(s)})\right)}{\sum_{k=1}^{S} \exp\left(g(\hat{z}_t^{(k)})\right)} \quad (10)$$

Fused latent representation:

$$z_t^{\text{fusion}} = \sum_{s=1}^{S} \alpha_s \cdot \hat{z}_t^{(s)} \quad (11)$$

Residual integration:

$$h_t = z_t + \tau z_t^{\text{ms}} = h_t + z_t^{\text{fusion}}{}_t \quad (12)$$

The Cross-Modal Attention mechanism is responsible for ensuring that the textual embeddings are properly aligned with the latent features to enable a good complementarity of the two modalities. The Sinusoidal Timestep Embeddings are founded on the idea of preserving the temporal alignment during the subsequent iterations for the iterative refinement process. To prevent the spatial inter-relations among features in the latent space, free Positional Embeddings are also used. This will then make sure that images coming from the generation of the model will have both appropriate contextual and spatial coherence.

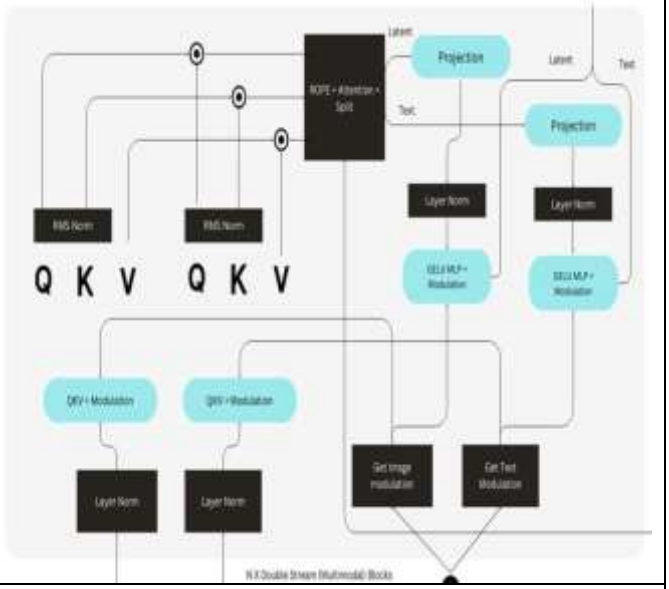

Fig. 4 The model structure of Multimodal Block

The latent representation from this fusion process then forms the basis for the iterative refinement. The Figure 4 shows the Multimodal block structure.

1.3 Iterative Refinement

The iterative refinement stage is the core mechanism behind the model to iteratively improve the superiority of the created visuals. A sequence of processes in multiple timesteps can refine the latent representation, which improves both the semantic and visual properties of the image. Single-Stream Blocks initiate the process of refinement by progressively processing the latent space. Within these blocks, the mechanism of self-attention permits the model to attention on important parts in the latent space so as to be able to improve the most important features for this particular image. Thus, through this attention mechanism, the model is able to zoom into important areas for refinement while being sure that important details are highlighted.

Additionally, ROPE improves spatial contextual understanding of the latent features so that the model may better generate realistic and coherent spatial relationships. In order to refine the generated features, the model uses Modulation Blocks. These blocks are designed to improve certain aspects of the image, and there are two types: the first is Image Modulation and the second is Text Modulation. Image Modulation brings forward the texture, color, or fine details to make the output image realistic and of high quality.

Text Modulation ensures that the generated semantic alignment of the visual features is such that the semantically lost meaning is not found in the generated image. The refinement process also uses Non-Linear Transformations by Gated Linear Unit, GLU. These transformations inject complex non-linear dynamics into the model, thus gain deeper understanding intricate patterns and further empower it to better capture finer details.

At each diffusion timestep, the latent representation is refined via transformer blocks:

Cross-modal attention

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (13)$$

where:

$Q = W_Q z_t^{ms}, K = W_K \tilde{e}_t, V = W_V \tilde{e}_t$

Denoising update

$$\hat{\epsilon}_t = f_\theta(z_t^{ms}, \tilde{e}_t, t) \quad (14)$$

Reverse diffusion step:

$$z_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(z_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\hat{\epsilon}_t\right) + \sigma_t \epsilon \quad (15)$$

Finally, Layer Normalization guarantees computational stability over timesteps, preventing potential problems that may arise in the iterative refinement. The Diffusion Scheduler governs the entire refinement process, which noise denoises the latent representation step-by-step so that with each iteration, it generates an image progressively clearer by removing noise step-by-step

1.4 Image Reconstruction

After the iterative refinement, the latent refined representation passes through a Variational Autoencoder Decoder. VAE decoder plays the key role in reconstructing latent features into the high resolution of an image.

After completing the reverse diffusion:

$$\hat{z}_0 = z_{t=0} \quad (16)$$

The final image is reconstructed using the VAE decoder:

$$\hat{x} = \mathrm{VAE}_{dec}(\hat{z}_0) \quad (17)$$

This is done to ensure that the output is not in any way inconsistent with respect to the text input, while retaining all the minute details. In this process, the decoder uses the fine-tuned latent and maps it to a high-resolution image in such a way that the entire text content and all the visual details are smoothly incorporated into the output. Thus, the subtlety of the details captured by the VAE decoder for image creation is also encompassed. This includes textures, lighting, and other minute details that have an impact on the photorealism of the output image as compared to the input text.

Results and Discussion

The proposed system measures the quality of model against both the original diffusion model and GANs with standard metrics. Namely, we compare Fréchet Inception Distance (FID), Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM), Inception Score (IS) along with training time. Table 1 yields that our model reflects that low FID means the ability to produce realistic and diverse outputs. And High PSNR means the generated images are close to the ground truth in terms of pixel values, High SSIM means the generated images are visually coherent and preserve important structural details. and High IS Score means the images are not only realistic but also varied across different categories. Here we found that the proposed model creates clearer and more realistic images as compared to Style GAN and Diffusion model.

Table 1. Comparison of advanced frameworks.

Model	FID	PSNR	SSIM	IS
Diffusion Model [10]	25.3	28.4	0.81	8.6 ± 0.4
Style GAN [3]	21.8	29.1	0.85	7.2 ± 0.5
Proposed Model	16.7	32.5	0.92	9.8 ± 0.3

For measuring the quality of the proposed models, FID, PSNR, SSIM as well as IS score have been used in this work. Equation (18), (19), (20), (21) and (22) are compute these scores

FID (Fréchet Inception Distance): It compares the statistics of created images to input images from a pre-trained Inception-v3.

$$\mathrm{FID} = \| \mu_x - \mu_y \|_2^2 + \mathrm{Tr}(\Sigma_x + \Sigma_y - 2\sqrt{(\Sigma_x \Sigma_y)}) \quad (18)$$

Where, μ_x, Σ_x are the mean and covariance for input images. μ_y, Σ_y are mean and covariance for created images. $\|.\|_2$ represents the norm which measures the length or magnitude of vectors in Euclidean space. Tr(.) signifies the trace of a matrix.

PSNR (Peak Signal-to-Noise Ratio): PSNR measures the quality of the reconstructed image compared to the original image. Let $x \in \mathbb{R}^{H \times W}$ be the real image.$\hat{x} \in \mathbb{R}^{H \times W}$ be the generated image. Then the Mean Squared Error (MSE) calculated by

$$\text{MSE (x, } \hat{x}) = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}(x_{ij} - \hat{x}_{ij})^2 \qquad (19)$$

$$\text{PSNR}(x, \hat{x}) = 10log_{10}\left(\frac{L^2}{\text{MSE}(x,\hat{x})}\right) \qquad (20)$$

where $L$is the maximum pixel value.

SSIM (Structural Similarity Index): SSIM assesses image similarity in terms of luminance, contrast, and structure.

$$SSIM(x, \hat{x}) = \frac{(2\,\mu_x\mu_{\hat{x}} + C1)(2\sigma_{x\hat{x}} + C2)}{(\mu_x^2 + \mu_{\hat{x}}^2 + C1)(\sigma_x^2 + \sigma_{\hat{x}}^2 + C2)} \qquad (21)$$

Where; $\mu_x\mu_{\hat{x}}$ Mean intensities of images x and $\hat{x}$. $\sigma_x^2, \sigma_{\hat{x}}^2$ Variances of x and $\hat{x}$. $\sigma_{x\hat{x}}$ Covariance between x and $\hat{x}$. C1, C2 Small constants to stabilize the division

IS (Inception Score): The Inception Score evaluates the realism and the variety of the generated images.

$$IS = \exp\left(\mathbb{E}_{\hat{x}}[D_{KL}(p(y|\hat{x})||p(y))]\right) \qquad (22)$$

where; $D_{KL}(p||q) = \sum_y p(y)log\frac{p(y)}{q(y)}$

For a better understanding of the results, different performance metrics like FID, PSNR, SSIM and IS score are calculated both for “laion400” and “Flickr 300” datasets.

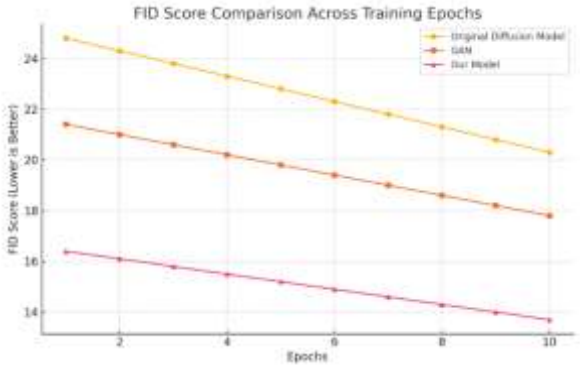

Fig. 5 FID Score comparison for trending models

FID, PSNR, SSIM and IS score of all the trending models and proposed model are showed in Figure 5,6 and 7 respectively. These scores illustrate proposed model excels in generating realistic, and wide range images while preserving structural and pixel-level details.

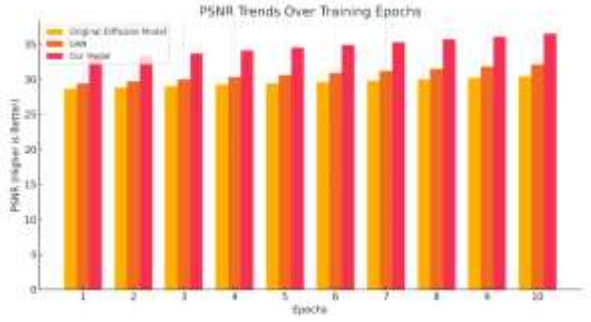

Fig 6 .PSNR Score comparison for trending models

Fig 7. SSIM Score comparison for trending models

In the Figure 8, comparison of generated images by all the three models is done and it is demonstrated that the proposed model is better than the Stable diffusion model and Style GAN model. Grad-CAM offers explainability by indicating which regions of the input affect the output, assisting in attention model understanding and error detection when reconstruction is unsuccessful. It confirms text-image alignment in cross-modal tasks and provides attention heatmaps to enhance training by pointing out areas that require improvement.

Fig 8. Qualitative Comparison of between Stable Diffusion, Style GAN and the Proposed Model

Conclusion

The proposed new model shows significant improvements over the original diffusion model and GANs for photorealistic images produced. Embedding an improved U-Net architecture with attention mechanisms and residual blocks along with multi-scale feature fusion and perceptual loss has helped in effective capture of details and textures. Classifier-free guidance finally ensures better align ability with textual prompts and results in excellent quality images. Quantitative metrics indicate the superiority of our model with best possible FID scores and higher values of PSNR and SSIM, which can be said to be images of better quality, closer to reality. This quantitative quality of the qualitative output is sharper and has better detail, making our approach a strong one for high-quality synthesis of images across very different applications.

Currently, the system being proposed is based on static image generation, but future research based on this idea can include video generation as well. Lastly, a more complex form of explainability, such as causal interpretability, rather than Grad-CAM, may provide more in-depth information about how the model makes decisions.

References

1. Amar, V., Sonu, N., & Shyan, H. (2023). Text-to-Image Generator using GANs. 2023 International Conference on Integrated Intelligence and Communication Systems (ICIICS), 1–4. https://doi.org/10.1109/iciics59993.2023.10421307
2. Wang, Z., Zheng, H., He, P., Chen, W., & Zhou, M. (2022). Diffusion-GAN: Training GANs with Diffusion. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2206.02262

3. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2019). Analyzing and improving the image quality of StyleGAN. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1912.04958
4. Ho, J., Saharia, C., Chan, W., Fleet, D. J., Norouzi, M., & Salimans, T. (2021). Cascaded diffusion models for high fidelity image generation. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2106.15282
5. Rauniyar, A., Raj, A., Kumar, A., Kandu, A. K., Singh, A., & Gupta, A. (2023). Text to Image Generator with Latent Diffusion Models. 2023 International Conference on Computational Intelligence, Communication Technology and Networking (CICTN), 144–148. https://doi.org/10.1109/cictn57981.2023.10140348 .
6. R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer (2022) "High-Resolution Image Synthesis with Latent Diffusion Models," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, pp. 10674-10685, https://doi.org/10.1109/CVPR52688.2022.01042
7. Zhang, Jinjin, Qiuyu Huang, Junjie Liu, Xiefan Guo, and Di Huang. "Diffusion-4k: Ultra-high-resolution image synthesis with latent diffusion models." In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 23464-23473. 2025. 10.1109/CVPR52734.2025.02185
8. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2103.00020
9. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research, 21(140), 1-67. https://doi.org/10.48550/arXiv.1910.10683
10. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-Shot Text-to-Image Generation. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2102.12092
11. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., & Rombach, R. (2023). SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2307.01952
12. Hidalgo, R., Salah, N., Jetty, R. C., Jetty, A., & Varde, A. S. (2024). Personalizing Text-to-Image diffusion models by Fine-Tuning Classification for AI applications. In Lecture notes in networks and systems (pp. 642–658). https://doi.org/10.1007/978-3-031-47721-8_44
13. Kumar, P., Pandi, S. S., Kumaragurubaran, T., & Chiranjeevi, V. R. (2024). Computer Vision and creative Content generation: Text-to-Sketch conversion. 2022 International Conference on Communication, Computing and Internet of Things (IC3IoT), 1–6. https://doi.org/10.1109/ic3iot60841.2024.10550294
14. Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196. https://doi.org/10.48550/arXiv.1710.10196
15. Yan, Q. (2023). Efficient Screen Space Ambient Occlusion Generation via Residual network. 2023 4th International Conference on Computer Vision, Image and Deep Learning (CVIDL), 8, 494–498. https://doi.org/10.1109/cvidl58838.2023.10165719
16. Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2021). Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM, 65(1), 99-106. https://doi.org/10.48550/arXiv.2003.08934
17. Raghavan, K., B, S., & V, K. (2023). Attention guided grad-CAM: an improved explainable artificial intelligence model for infrared breast cancer detection. Multimedia Tools and Applications, 83(19), 57551–57578. https://doi.org/10.1007/s11042-023-17776-7
18. Hossain, M. Z., Zaman, F. U., & Islam, M. R. (2023). Advancing AI-Generated Image Detection: Enhanced Accuracy through CNN and Vision Transformer Models with Explainable AI Insights. 2023 26th International Conference on Computer and Information Technology (ICCIT), 1–6. https://doi.org/10.1109/iccit60459.2023.10440990
19. Jadon, R., Gollapalli, V. S. T., Srinivasan, K., Chauhan, G. S., & Budda, R. (2025). Interpretable AI for Skin Lesion Detection: Enhancing Diagnostic Accuracy with CNN and Score-CAM in IoMT Systems. Journal of Ubiquitous Computing and Communication Technologies, 7(1), 1–18. https://doi.org/10.36548/jucct.2025.1.001
20. Shen, D., Celikyilmaz, A., Zhang, Y., Chen, L., Wang, X., Gao, J., & Carin, L. (2019, July). Towards generating long and coherent text with multi-level latent variable models. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 2079-2089).
21. Deshmukh , J. .Y, Bhokare , P., Malunjkar, P., Shenkar, A., & Thorat , S. (2025). AI-Based Approach using Generative Adversarial Network for Interior Design System. International Journal on Advanced Computer Engineering and Communication Technology, 14(1), 428–431. https://doi.org/10.65521/ijacect.v14i1.560