Bulletin of Electrical Engineering and Informatics
Vol. 15, No. 2, April 2026, pp. 1239~1254
ISSN: 2302-9285, DOI: 10.11591/eei.v15i2.9927

DeepFloyd-IF via diffusion and U-Net based cross-model attention for semantic coherence

Kowsalya Veilumuthu, Divya Chandrasekar, Sakthidevi Shunmugalingam Parvathi
Centre for Information Technology and Engineering, Manonmaniam Sundaranar University, Tirunelveli, India

Article Info

Article history:

Received Jan 22, 2025
Revised Dec 17, 2025
Accepted Feb 22, 2026

Keywords:

Cross-model attention
DeepFloyd-IF
Multi-head attention
Stable diffusion
U-Net

ABSTRACT

Text to image synthesis is getting harder in artificial intelligence, impacting gaming, advertising, and multimedia. The practical use of current Text to Image models is limited by the trade-off between semantic coherence and visual quality. To address this, this work presents stable diffusion cross-modal attention with multi-head attention (SD-CMA-MHA), a framework for the DeepFloyd-IF task. This combines stable diffusion with U-Net based cross-modal attention and multi-head attention (MHA) to improve DeepFloyd-IF, a standard for high quality image synthesis. This allows the model to capture subtle semantic relationships between text and images while dynamically focusing on relevant input features. Experiments on LAION-1.2B and MS-COCO datasets show that the model achieves 80% generation accuracy, 70% text-image alignment similarity and reduced divergence from real images, better than previous methods. This shows that SD-CMA-MHA improves semantic alignment and fidelity. The conclusion is that by enabling more reliable and context aware visual generation, this work not only bridges the gap between text and visual modalities but also has implications for creative industries, education and human-computer interaction.

Corresponding Author:

Kowsalya Veilumuthu
Centre for Information Technology and Engineering, Manonmaniam Sundaranar University
Tirunelveli, Tamil Nadu, India
Email: kowsalyaphd260@gmail.com

1. INTRODUCTION

DeepFloyd-IF is a text-to-image creation artificial intelligence (AI) that uses text to create amazing, photorealistic images. It has a multi-stage pipeline with super-resolution modules, diffusion models and frozen text encoders to increase clarity and depth. Unlike previous models it balances visual quality and semantic accuracy so the generated image looks exactly like the description. So, it is super useful for creative content in publishing, advertising and design. It also helps with educational tasks by allowing you to describe ideas directly from natural language.

This research looks at DeepFloyd-IF, a deep learning framework that can generate images from text descriptions that are visually pleasing and contextually correct. Image synthesis has moved away from traditional computer vision methods and towards modern deep learning methods which are becoming more and more important for publishing, education and the creative industries. Using diffusion based designs and large datasets DeepFloyd-IF can improve visual production for many workflows. Modelling the complex semantic relationships between text and images was hard for earlier prompt-to-image algorithms [1]-[3] which relied on human feature extraction and linear regression. This alignment was facilitated by deep learning, however generating high-fidelity outputs with increased resolution and fine details were still a heavy lift. For some time, we have been applying transformer-based approaches that work by self-attention

processes - this is a technology with advanced neural architecture to a great extent; however, both of them are tastefully competent and without fine detail fidelity and image synthesis that appears real. Further, early works [4] were primarily focused on benchmark performance and technology advances [5]-[7] with no mention of growing demand for automated, scalable image production in creative workflows. Applications associated with digital media content generation, marketing, and in education now require more efficiency and adaptability. In the meantime, the socio-communicative [8] issues and implications of text-to-image synthesis that are acknowledge ably and readily proliferating in practice, were neglected in these works.

By emphasizing on DeepFloyd-IF's capability to decode semantic cues from textual prompts and provide visually coherent, semantically matching, and high-resolution outputs, this study seeks to assess and improve the system's performance. Beyond prior approaches, DeepFloyd-IF exhibits the ability to convert natural language descriptions into realistic and detailed images while preserving both semantic accuracy and visual fidelity, as seen in Figure 1. Also, Figure 2 illustrates the wider societal advantages of this framework, stressing its usefulness in publishing, marketing, education, and social media by streamlining the generation of visual material and fostering creativity and communication. All of these goals work together to make DeepFloyd-IF a useful and flexible tool for a variety of real-world applications in addition to being a technically sound model.

Figure 1. Example for DeepFloyd-IF

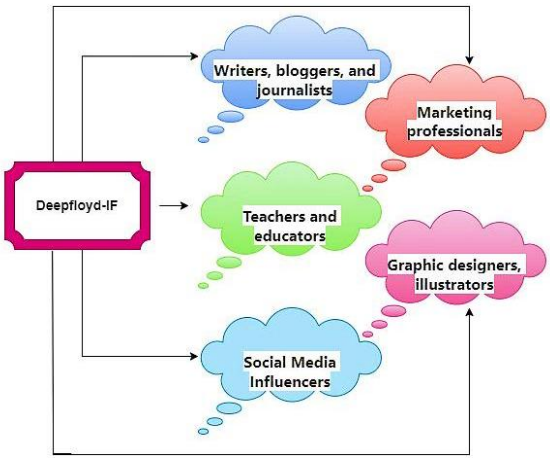

Figure 2. Societal benefits of DeepFloyd-IF

This study benefits marketers, educators, and content creators by demonstrating how DeepFloyd-IF streamlines picture generation for blogs, articles, presentations, and social media. Large-scale picture-text datasets and diffusion models have improved image production while successfully lowering artifacts found in older programs like Photoshop. The model captures both local and global information, maintains spatial features, and improves text–image alignment by combining T5 text encoders, U-Net architectures, and cross-modal attention. Together, these developments make it possible for text-to-image synthesis to be precise, effective, and scalable, solidifying DeepFloyd-IF as a versatile tool for innovation and cross-sector communication.

2. RELATED WORK

2.1. Transformer based image generation

Transformers excel in image generation tasks requiring comprehension of lengthy text relationships. Recent developments in text-to-image diffusion models with a variety of properties are highlighted in the examined studies. By incorporating cross-frame attention and volume rendering into U-Net, ViewDiff [9]

enhances visual quality and realism while highlighting 3D consistency. Text-to-image diffusion models [10] focus on visual concept-driven generation, improving semantic alignment between textual prompts and generated images. In contrast, disentangled diffusion [11] addresses model memorization issues by introducing cross-attention strategies that reduce overfitting and enhance generalization capability.

CogView [12], a 4B-parameter transformer with vector quantized variational autoencoder (VQ-VAE) tokenizer and 48-layer GPT, treats text and image tokens uniformly using separators, achieving state-of-the-art results on MS COCO and surpassing Fréchet inception distance (FID) benchmarks. StyleSwin [13], a transformer-based generative adversarial network (GAN), and swin transformer [14], with its hierarchical shifted-window design, replace convolutional architectures to enhance modeling efficiency, scalability, and consistency, demonstrating superior performance on datasets like large-scale scene understanding (LSUN) church, Flickr-faces-high-quality (FFHQ), and CelebFaces Attributes (CelebA-HQ). A multilevel process network incorporates text hierarchy to improve medical image segmentation [15]. This method effectively increases feature learning through contextual awareness and edges while providing a high level of accuracy across different imaging databases.

2.1.1. Cross-modal attention mechanisms

Cross-attention manages uneven relationships between embedding patterns. Subject-driven diffusion models struggle with multi-concept images, prompting the development of textual localization models that use cross-attention to link visual representations and text tokens. Existing methods often yield unrealistic 3D objects due to limitations in fine-tuning or pretraining. Cross-frame attention layers added to U-Net frameworks address these issues by aligning spatial traits across views, trained on the CO3Dv2 dataset of multi-angle real-world images. To enhance prompt-to-image [16] synthesis fidelity, cross-modal contrastive GAN refines text-image information transfer using contrastive losses for intra- and inter-modality similarities. enhances feature extraction and cross-modal interactions by proposing a hybrid convolutional neural network (CNN)–transformer with non-local cross-modal attention for multimodal picture fusion.

2.2. U-Net architecture in image generation

By facilitating segmentation, correction, and transformation, U-Net contributes indirectly but significantly to text-to-image tasks. With a dual U-Net architecture, BootPIG [17] goes one step further. The reference U-Net uses reference pictures to extract features, while the Base U-Net adds reference self-attention (RSA) layers to enhance features. Despite its effectiveness, BootPIG has issues with model scalability, prompt compliance, and detail production; yet, its bootstrapping technique allows for zero-shot individualized picture generation in pretrained diffusion models. While pointing out synchronization problems, an uncertainty-driven edge prompt generation [18] network that will improve boundary precision in medical image segmentation by using uncertainty estimation techniques. This proposed approach it will show improved accuracy of segmentation, especially in ambiguous and low-contrast areas of images, with evidence of superior performance on benchmark datasets from the field of medical imaging. In addition to these, Refined U-Net framework [19] that incorporates discriminative probing and tuning techniques to enhance alignment, control, and overall generation quality, as well as down sampling, middle sampling, and up sampling for noise prediction and distortion minimization using L2 loss.

2.2.1. Diffusion-based models

Diffusion-based models [20] generate high-quality images by iteratively refining noise into coherent representations. RealComp introduced a dynamic balancer for realistic and well-composed images, addressing challenges with complex text prompts. Role-Playing Game enhanced prompt-to-image models by using multimodal large language models to divide tasks into subregions, enabling regional compositional synthesis. Advanced models like Stable Diffusion excel in text-image alignment but struggle with precise text rendering. The text diffuser addresses this by generating images based on prompts and refining layouts with transformer-extracted keywords. The MARIO-10M dataset supports text rendering, while CONPREDIFF employs self-denoising and context prediction for improved results. Performance is validated on datasets like CelebA-HQ, FFHQ, LSUN, and MS-COCO using zero-shot FID and CLIP scores.

2.3. Artificial intelligence-based image generation

AI-based image generation uses algorithms to create images from text, but faces challenges such as limited control, realism, and difficulty in capturing nuances. Despite advancements in creating aesthetically pleasing images, models struggle to faithfully convert complex written descriptions into coherent visuals. Over-reliance on prompt modifiers can limit creativity and quality.

Neural networks and sophisticated training techniques are used in deep learning techniques such as DeepFloyd-IF [20] to improve visual realism and variety. For text-to-image creation, GlyphControl [21] presents glyph-conditional control, which improves FID Contrastive language–image pretraining (CLIP), and

optical character recognition (OCR) accuracy without requiring model retraining. While HexaGen3D [22] extends diffusion capabilities to enable fast and diverse text-to-3D generation, SEGA [23] introduces semantic guidance that allows users to steer outputs along meaningful directions for precise and creative modifications. Zohra *et al.* [24] utilize linear regression techniques for facial picture quality estimation, enabling quantitative assessment of visual features. In addition, Paulin and Ivasic-Kos [25] examine synthetic dataset generation strategies that support broader computer vision applications by improving data availability and diversity.

3. EXPERIMENTAL DATASET

The experimental dataset used to assess the stable diffusion cross-modal attention with multi-head attention (SD-CMA-MHA) this model's performance in Prompt-to-Image generating tasks is described in this section.

3.1. Dataset description

The LAION-1.2B and MS-COCO datasets are used for prompt-to-image conversion. MS-COCO is a benchmark for image captioning, while LAION-1.2B is curated for DeepFloyd-IF tasks, providing a diverse collection of images and corresponding text for training and evaluation.

3.1.1. LAION-1.2 B dataset

This dataset includes 50,000 training and 10,000 testing high-resolution images with diverse prompts covering various subjects, settings, and items. The images are sourced from creative commons, photo repositories, and photographic sites.

3.1.2. MS COCO dataset

MS COCO contains over 200,000 captioned photos, with 150,000 training and 30,000 testing samples. It features diverse images of settings, items, and actions, each with multiple human-annotated captions for training and evaluation.

4. STABLE DIFFUSION CROSS-MODAL ATTENTION WITH MULTI-HEAD ATTENTION METHOD

This section presents crucial background information about variational autoencoders (VAE), attention mechanisms, the U-Net architecture, and stable diffusion performance, all of which are crucial to the design of the SD-CMA-MHA DeepFloyd-IF model and the understanding of its applications. Understanding these topics is the basis of forming strategies and understanding results.

4.1. Variational autoencoder

In machine learning, VAE are powerful generative models, commonly used for more mundane tasks such as prompt-to-image generation. VAE learns complex distributions over data to create new samples by combining neural networks and probabilistic modelling. Its objective is to learn a representation of the latent space that captures the variability and essential properties of the original data. The latent space is continuous and of low-dimensionality, allowing for spacing and interpolation in the latent space while making sampling efficient. VAE consists of two primary components, the encoder and the decoder. The general equation for the VAE is (1) and (2):

$$Z \sim q_{\emptyset}\left(\frac{Z}{X}\right) = \mathfrak{n}\left(\mu_{\emptyset}(X), \sigma_{\emptyset}^{2}(X)\right) \quad (1)$$

$$X_{Recon} \sim P_{\theta}\left(\frac{X}{Z}\right) \quad (2)$$

The encoder, parameterized by $\emptyset$, learns to estimate the posterior distribution of latent variables Z given input data X. In this case, the encoder is represented as $q_{\emptyset}\left(\frac{Z}{X}\right)$. The decoder, parameterized by θ, reconstructs the input data X from the latent representation Z. Its formula is $P_{\theta}\left(\frac{X}{Z}\right)$ the model is able to capture the underlying structure of the data in a latent space.

VAE has several of benefits for tasks involving prompt-to-image production.

- Latent space interpretability: VAE's continuous latent space enables meaningful interpolation and manipulation, allowing for varied and semantically relevant image generation from text prompts.

- Text-image connection: VAE's generative framework samples from the latent space to create high-quality images based on complex text-image relationships.
- Regularization: the VAE regularization term helps the model learn disentangled representations, improving generalization and robustness in DeepFloyd-IF tasks.

4.2. Attention mechanism

Attention mechanisms are crucial in deep learning tasks, allowing models to focus on important parts of input data for predictions. In prompt-to-image generation, attention aligns relevant text sections with corresponding image regions. It works by assigning weights to input elements based on their importance, with the weighted sum emphasizing the most relevant data for the task.

4.2.1. Calculating attention scores

A similarity function, which gauges how close each component of the input data is to the model's present state, is usually used to calculate the attention scores. The dot product is one often used similarity function (3):

$$\text{Attention Score}\ (\mathrm{h_t}, \mathrm{x_i}) = \mathrm{h_t^T x_i} \tag{3}$$

Where x_i is each piece of the input data (e.g., the word embeddings of the text description or the features of the input picture), and $\mathrm{h_t^T}$ is the model's current state (e.g., the hidden state of a recurrent neural network).

4.2.2. Attention weight

Attention weights specify how much emphasis should be placed on each component of the input sequence while producing the output in the context of attention processes. The similarity scores between each input piece and the model's current state are used to determine these weights. These similarity scores are often transformed into probabilities using the SoftMax function, which guarantees that the attention weights add up to one. The equation that ensues may be used to determine the attention weight $\propto_{\mathrm{ij}}$ for each input element x_i:

$$\propto_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{N} exp(e_{ik})} \tag{4}$$

Here, N is the total number of input elements, and e_{ij} reflects the similarity score between the input element x_i and the model's current state h_i.

4.2.3. Context vector

The context vector, or weighted sum of the input items, is generated by taking the attention weights into account while computing the weights. This process yields the context vector, or C_t. The information from the input sequence that is thought to be most important for producing the output at the current time step is represented by this context vector. The context vector C_t can be calculated using (5):

$$C_t = \sum_{i=1}^{N} \propto_{ij} \cdot X_i \tag{5}$$

The attention weight for the input element X_i is represented by α ij in this case, where $\propto_{ij}$ denotes the input feature at location X_i.

4.3. Traditional U-Net architecture

Because of its accuracy and efficiency, the traditional U-Net CNN architecture is frequently employed for picture segmentation. It can capture both fine spatial information and semantic context because to its encoder–decoder design with skip links. In order to recover hierarchical features like textures, forms, and borders, the encoder down samples the input. The decoder then uses transposed convolutions to up sample these features in order to restore spatial resolution. Skip connections preserve crucial information like edges and contours during reconstruction by sending low-level encoder characteristics straight to the decoder. U-Net captures both global structure and local details to enable extremely accurate segmentation. Beyond segmentation, its application in image synthesis, satellite imaging, agriculture, biological analysis, and contemporary text-to-image models is made possible by its capacity to strike a compromise between accuracy and context. Diffusion processes, residual blocks, and attention methods are examples of improvements that improve U-Net's ability to handle complicated, high-resolution images. U-Net is a crucial architecture in computer vision and generative modelling because of its extensibility, resilience, and adaptability.

4.4. Stable diffusion

Another powerful model for creating images is stable diffusion, which is often referred to as denoising diffusion probabilistic models. Stable diffusion models, in contrast to conventional generative models, produce images by continuously improving a noise tensor in the latent space by the use of a diffusion process. To get the desired picture, the noise is first added to the original noise tensor and then progressively reduced in intensity until it converges. In stable diffusion models, the diffusion process may be expressed mathematically as (6):

$$Z_T = Z_O + \sigma_T \cdot \in \tag{6}$$

Where σ_T the noise level at time step T is, $\in$ is Gaussian noise, Z_T is the final noise tensor, and Z_O is the starting noise tensor.
For every x, y, and z, the produced picture is y, and the generator function is y.

$$X_T = f(Z_T) \tag{7}$$

Where f the generating is function and X_T is the produced picture.
Stable diffusion models have proven their efficacy in a number of picture production applications, such as unconditional image generation, image inpainting, and image super-resolution.

By mapping natural language cues to visual concepts, stable diffusion enables text-to-image synthesis, enabling users to produce high-quality images from in-depth descriptions. It is a very versatile generative model that may be applied to tasks such as style transfer, image modification, and domain adaptation. Stable diffusion is possible even on consumer-grade technology since it operates in latent space instead of pixel space, which speeds up creation, lowers computing cost, and preserves strong image fidelity. Because it is open-source, it has been more widely adopted, customized, and improved by the community.

5. DEEPFLOYD-IF MODEL

The SD-CMA-MHA DeepFloyd-IF model seeks to generate high-fidelity images from textual prompts by utilizing the complementary effects of stable diffusion with the U-Net architecture. The model can successfully bridging of the semantic gap between text and images is achieved through a synergistic fusion of two key components: the hierarchical feature extraction capabilities of U-Net and the controlled creation process of stable diffusion. This results in outputs that are both visually attractive and contextually meaningful. An overview of the SD-CMA-MHA model is provided in this part, along with information on its main elements and procedure.

5.1. Stable diffusion cross-modal attention with multi-head attention

The SD-CMA-MHA strategy creates a unique DeepFloyd-IF framework by combining U-Net architecture with stable diffusion. The encoder, the decoder, and the steady diffusion module make up its three primary parts. The textual prompts are processed by the encoder to extract hierarchical characteristics, which are then encoded into a latent space representation. Based on the U-Net design, the decoder gradually creates the matching image by using the latent representation as input. By using repeated diffusion processes to modify the latent representation, the stable diffusion module makes sure that the generation process is under control and results in high-quality Images. The final images are produced by feeding the stable diffusion module's outputs into the decoder. The creation of realistic visuals that accurately represent the meanings provided by the textual prompts is made possible by this model design. In the following sections, delve deeper into the implementation details and experimental results of the SD-CMA-MHA model.

5.2. Architectural design

For high-fidelity prompt-to-image generation, an efficient architecture must be designed. The main elements of the SD-CMA-MHA design are described in this section along with their interrelationships. Coherent and pertinent picture synthesis is made possible by the careful blending of text and image modalities in every component, from cross-modal attention to integrating steady diffusion inside the U-Net architecture.

Three separate phases make up the SD-CMA-MHA SD-CMA-MHA model process, which is shown in Figure 3 and was created to make prompt-to-image production easier. The design emphasizes the integration of multi-head attention (MHA), cross-modal attention (CMA), and stable diffusion in a U-Net framework, introducing numerous innovative components to improve the generation process.
Stage 1: variational autoencoders-encoder (down-sampling)

The VAE-encoder serves as the initial stage of the SD-CMA-MHA model, primarily focusing on encoding both the input image and prompt text into meaningful latent representations. Through tokenization and embedding processes, the input image and prompt undergo preprocessing to extract essential features. Image embedding (IE) and text embedding (TE) are generated, followed by concatenation (Concat) to fuse the features of the image and prompt. This concatenated result is then passed through the CMA mechanism, aligning the features of the image and prompt to facilitate the generation of a coherent image relevant to the given prompt. Introducing a rectified linear unit (ReLU) layer followed by a 1D convolution operation enhances novelty by promoting non-linearity and feature refinement within the generated content. The ReLU activation allows the model to capture complex patterns and variations, enhancing the diversity and uniqueness of the generated outputs. Subsequently, the 1D convolution operation further refines the feature representation, smoothing out noise and enhancing the clarity of the generated content. Together, these operations effectively enhance novelty by promoting richer, more varied, and visually appealing results, ultimately improving the overall quality of the generated content.

$$Concat(IE, TE) \rightarrow CMA(Concat) \rightarrow ReLU(CMA) \rightarrow Conv1D \quad (8)$$

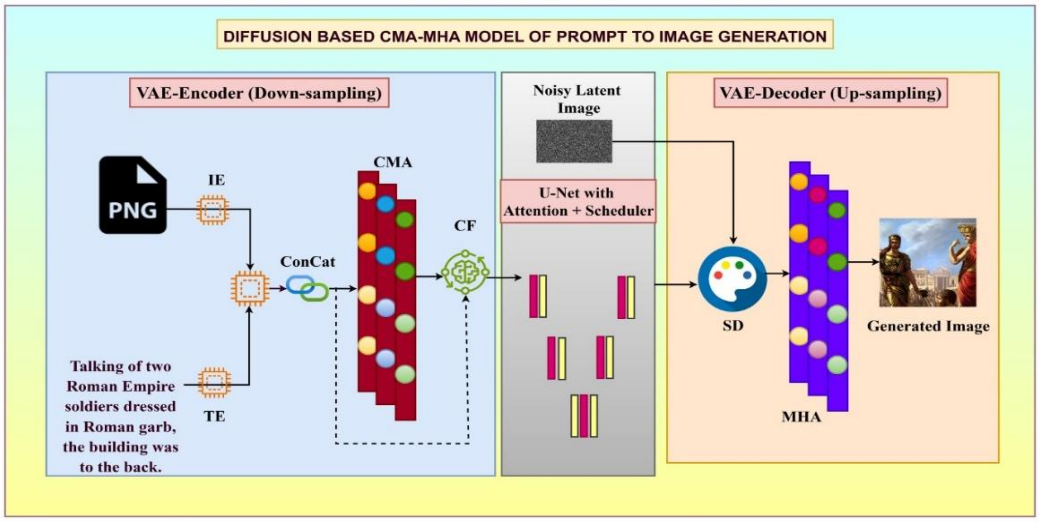

Figure 3. Framework of SD-CMA-MHA

Stage 2: U-Net with attention mechanism

In the second stage, the processed feature representation from the VAE-Encoder is fed into a U-Net architecture enhanced with attention mechanisms. This stage aims to refine the generated noisy image from the previous stage and extract more detailed features. The incorporation of attention mechanisms, such as ResNet and attention layers, enables the model to focus on relevant regions of the image and capture intricate details effectively. These mechanisms contribute to improving the coherence and quality of the generated images.

$$\text{VAE} - \text{Encoder} = \text{FE} \quad (9)$$

$$\text{RB(FE)} = \text{FE} + \text{Convolutional Layers} \quad (10)$$

$$\text{Attention (FE)} = \text{Attentional}_{\text{FR}} \quad (11)$$

$$\text{U} - \text{Net Output} = \text{Attention}\left(\text{RB (FE)}\right) \quad (12)$$

- The processed feature representation from the VAE-encoder is represented by FE.
- The residual block, or RB, is made up of convolutional layers that process the feature representation that was input.
- The attention mechanism that is applied to the feature representation $Attentional_{FR}$ and concentrates on pertinent areas of the image is referred to as attention.
- The output is the improved feature representation that is produced when the residual block is subjected to attention processes.

Stage 3: diffusion and decoder

The final stage of the SD-CMA-MHA model involves Stable Diffusion-V 1.5 (SD V'1.5) and decoder operations. Here, the generated noisy image from the U-Net with attention stage undergoes denoising through stable diffusion, which iteratively refines the image at each time-step. This denoised image is then subjected to MHA, introducing additional attention processes to enhance feature extraction and refinement. Finally, the output of the MHA mechanism is used to generate the final image based on the user prompt, completing the DeepFloyd-IF process.

$$SD(Noisy\ Image) \rightarrow MHA(SD) \rightarrow Generated\ Image \tag{13}$$

The integration of stable diffusion, cross-modal attention, and MHA within the U-Net architecture enhances DeepFloyd-IF. This approach aligns text and image features, captures intricate details, and generates high-quality images, ensuring robustness and adaptability in DeepFloyd-IF tasks.

The three stages of DeepFloyd-IF collectively ensure high-quality text-to-image generation. Stage 1 encodes and aligns text–image features through VAE-Encoder, cross-modal attention, and refinement operations. Stage 2 leverages U-Net with attention and residual blocks to enhance feature details, while stage 3 applies stable diffusion and MHA to iteratively denoise and refine, producing coherent, visually appealing final images.

5.2.1. Cross-modal attention and multi-head attention

CMA and MHA are critical components in the SD-CMA-MHA SD-CMA-MHA enabling integration of data from multiple modalities and capturing complex feature correlations. Let's examine their equations and operations in more detail:

a. Cross-modal attention

CMA aligns features like prompt and visual embeddings from several modalities to promote efficient integration and comprehension. According to the SD-CMA-MHA model, CMA combines the characteristics from both modalities by calculating attention weights between the text and picture features. The CMA mechanism may be expressed mathematically as follows: Based on word embeddings query Q and picture embeddings key-value pairs K and V, the attention weights α are calculated as (14):

$$\propto = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{14}$$

Where the dimensionality of the key vectors is represented by the expression d_k. The attended picture features are then computed using these attention weights Att(V):

$$Att(V) = \propto \tag{15}$$

Finally, the original text features and attended picture features are concatenated to produce the final aligned features (AF):

$$AF = Concat\left(Q, Att(V)\right) \tag{16}$$

b. MHA

This technique expands on the concept of attention by enabling the model to concentrate on many input components at the same time. It does this by concurrently computing many sets of attention weights, each of which concentrates on a distinct feature of the input. MHA is used in the SD-CMA-MHA model to increase the efficiency of the attention mechanism. MHA may be expressed mathematically as follows: Considering the query Q, key-value combinations K and V, and the lots of attention heads h, MHA calculates many sets of attention weights $\propto_i$and attended values Att_i(V) in (17) and (18):

$$\propto_i = Softmax\left(\frac{QW_i^K(K^T)}{\sqrt{d_k}}\right) \tag{17}$$

$$Att_i(V) = \propto_i V \tag{18}$$

When W_i^k the learnable parameters unique to the i[th] attention head is represented. The final output is then obtained by concatenating and linearly transforming the outputs of each attention head:

$$Final_{Out} = Concat\big(Att_1(V), Att_2(V), \ldots., Att_h(V)\big)W^o \quad (19)$$

Where W^o is the output transformation's learnable parameter matrix.

The proposed model uses CMA to align features in text and images, and MHA to improve the attention mechanism by enabling the model to attend to numerous input characteristics at once. The model may more efficiently collect and make use of the intricate interactions between features from many modalities by incorporating these processes, which improves the coherence and quality of the produced images.

5.2.2. Integration of U-Net and stable diffusion

Textual prompts and latent-space Gaussian noise are used as inputs in the SD-CMA-MHA model's U-Net (Figure 4), with ResNet and attention modules improving each layer. By concentrating on pertinent features and refining noisy images, this approach aids the model in generating outputs that are more precise and cohesive.

$$Res(x) = x + Conv(x) \quad (20)$$

Where Conv stands for the convolutional operation performed inside the residual block and x represents the input feature map.

$$Attention\ Weights: \alpha = softmax(QK^T) \quad (21)$$

$$Attended\ Features: Att(V) = \alpha V \quad (22)$$

Where the query, key, and value matrices are indicated by the symbols Q, K, and V, respectively. The attended features $Att(V)$ capture the weighted sum of values depending on these attention weights, and the attention weights describe the importance of each key to the query.

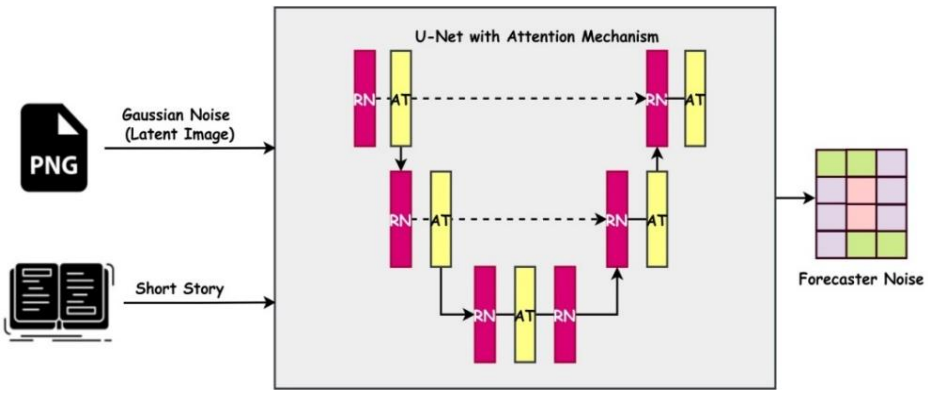

Figure 4. U-Net architecture of SD-CMA-MHA

In the U-Net architecture, the input undergoes a series of modifications, refining noisy images to produce clearer representations. Ultimately, the U-Net block forecasts noise images that capture the structure of the input prompt, enabling the model to generate high-quality, prompt-aligned visuals. In the Stable Diffusion workflow, as seen in Figure 5, both forward and backward diffusion processes are used to create relevant images. The noisy image generated at each stage is refined through multiple diffusion rounds, progressively improving the clarity and accuracy of the output. The forward diffusion process can be expressed logically as (23):

$$x_{t+1} = Diffuse(x_t) \quad (23)$$

Where the noisy picture is represented by $x_t$and the image following the diffusion process is represented by x_{t+1} at time-step $t + 1$. The process of adding noise and diffusing it throughout the image is represented by the diffuse function.

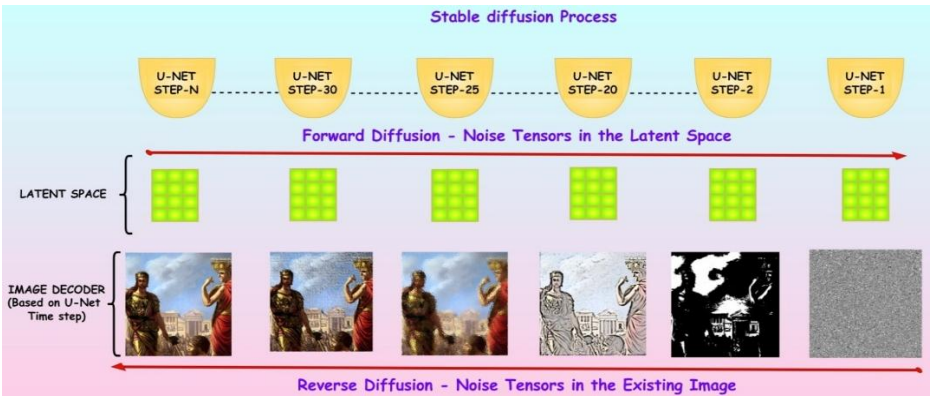

Figure 5. Pipeline of stable diffusion (SD V'1.5)

The created noisy picture is then decoded in the stable diffusion. The goal of this reverse diffusion procedure is to recover the context and semantic content that the user prompt gave when denoising the resulting image. The reverse diffusion process can be stated analytically as (24):

$$x_{reconstructed} = ReverseDiffuse(x_{noisy}) \quad (24)$$

Where the produced noisy picture is represented by x_{noisy} and the recovered meaningful image is represented by $x_{reconstructed}$. The reverse diffusion process, which seeks to retrieve the original meaningful content and denoised the noisy image, is represented by the reverse diffuse function.

Stable diffusion involves iteratively generating noisy images through forward diffusion in the U-Net architecture, followed by reverse diffusion to reconstruct meaningful, denoised images. This process ensures the generated visuals accurately reflect the context and semantics of the user's prompt.

6. RESULTS AND ANALYSIS

This section includes an evaluation of the SD-CMA-MHA model as well as the findings of our trials. The performance analysis using quantitative metrics like CLIP and FID scores to produce high-quality Images from textual cues were made. Furthermore, evaluated the produced images were compared with the ground truth photographs to do qualitative assessments.

6.1. Evaluation metrics

The evaluation analysis of the SD-CMA-MHA models using a variety of evaluation metrics including accuracy, diversity in novelty (DINO), CLIP-I (CLIP visual similarity), FID, human preference score (HPS), and image rewards. These metrics provide comprehensive insights into the quality; diversity, alignment, and preference of the generated images are described.

6.1.1. Accuracy

Evaluates how accurately the resulting photos match the real-world images. It is determined by dividing the total number of photos by the ratio of correctly created images.

$$Accuracy = \frac{Number\ of\ Correctly\ Generated\ Images}{Total\ Number\ of\ images} \times 100 \quad (25)$$

A higher accuracy score shows that there is a stronger match between the generated images and the written descriptions that go with them, indicating that the model can comprehend and convert textual cues into meaningful visual material.

Table 1 compares image-generation accuracy across stable diffusion and DeepFloyd variants on the LAION-1.2B dataset. The proposed SD-CMA-MHA model achieved 80% accuracy, outperforming DeepFloyd IF-I-M (40%) and IF-I-L (70%), and matching IF-I-XL (80%), confirming superior semantic alignment and image quality at larger scales. All models were evaluated in a zero-shot setting using official checkpoints. SD-CMA-MHA (1.48 B parameters) was trained for 250k steps using AdamW (LR=1×10^{-4}, cosine decay), batch size 64, on 8×A100 80 GB GPUs for 92 hours with seed 42, ensuring a consistent and reproducible evaluation protocol.

Table 1. Image size analysis of LAION-1.2B

Dataset	Models (SD-CMA-MHA)	Accuracy (%)
LAION-1.2B	Stable diffusion	80
	DeepFloyd (IF-I-M)	40
	DeepFloyd (IF-I-L)	70
	DeepFloyd (IF-I-XL)	80

6.1.2. Diversity in novelty score

DINO assesses the diversity and novelty of generated content. It evaluates how unique or different the generated outputs are from each other. Higher DINO scores indicate greater diversity and novelty in the generated content, ensuring a wider range of outputs.

$$DINO = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{K_i}\sum_{j=1}^{K_i} Distance\left(I_i, I_{ij}\right) \quad (26)$$

Where N is the total number of images created. K_i is the quantity of photos in the collection that are comparable to image I_i. The distance metric between an image I_i and its j-th comparable image I_{ij} is called $Distance\left(I_i, I_{ij}\right)$.

6.1.3. CLIP-I score (visual similarity)

Using visual cues that the CLIP model extracts, CLIP-I calculates the degree to which created images resemble reference images. Greater visual similarity and improved alignment with the reference images are indicated by higher scores.

$$\text{CLIP} - \text{I} = \text{Cosine}_{\text{Similarity}}(\text{F (Text)}, \text{G(Image)} \quad (27)$$

Where F(Text) represents the feature embedding of the generated text, G(Image) represents the feature embedding of the generated image.

DINO and CLIP-I scores for various models on the LAION-1.2B dataset are compared in Figure 6. Our SD-CMA-MHA model outperforms Re-Imagen [26], ELITE [27], BLIP-Diffusion [28], and Subject-Diffusion [29] with a DINO score of 70.2 and a CLIP-I score of 71.2. Stronger alignment between text prompts and generated images is reflected in the higher CLIP-I score, demonstrating that SD-CMA-MHA model mention it as 29 in x axis for reference, the generates outputs that are diversified, visually cohesive, and contextually accurate.

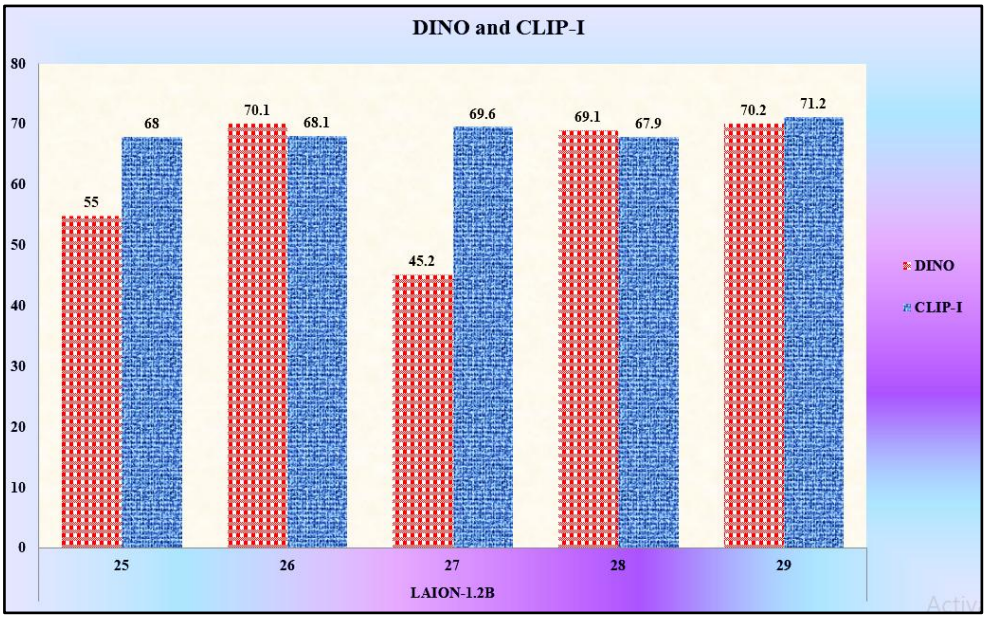

Figure 6. DINO and CLIP-I performance in LAION-1.2B

6.1.4. Fréchet inception distance score

FID score measures the similarity between two sets of images, often used to evaluate the quality of generated images compared to real images. Lower FID scores indicate better agreement between the distributions, suggesting that the generated images closely resemble the real images in terms of visual features.

$$FID = || \mu_1 - \mu_2||^2 + Tr\ (\sum_1 + \sum_2 - 2\ (\sum 1 \sum 2)^{0.5}) \quad (28)$$

Where $\mu_1 - \mu_2$ are the mean feature vectors of real and generated image, $\sum 1 \sum 2$ are the covariance matrices of real and generated image, Tr denotes the trace (sum of diagonal elements)

Figure 7 shows how effectively various text-to-image models align with prompts and generate high-quality images by comparing FID scores across eight zero-shot evaluation phases on the MS-COCO dataset. The FID scores of all models are normalized and represented using reference values DPM [30], DPM++ [31], Meng *et al.* [32], SnapFusion [33], and proposed model as 34 for reference in figure to provide a clear comparative visualization scale. Better fidelity is indicated by lower FID ratings, and at step 8, the suggested SD-CMA-MHA model outperforms previous stages and rival baselines with an astounding score of 29. Interestingly, it routinely outperforms SD-v1.5, which produces higher (worse) FID values but requires more processing. The efficacy and superiority of the SD-CMA-MHA framework in precise, efficient text-to-image synthesis are demonstrated by the figure, which displays a consistent improvement in our model's performance over the assessment steps.

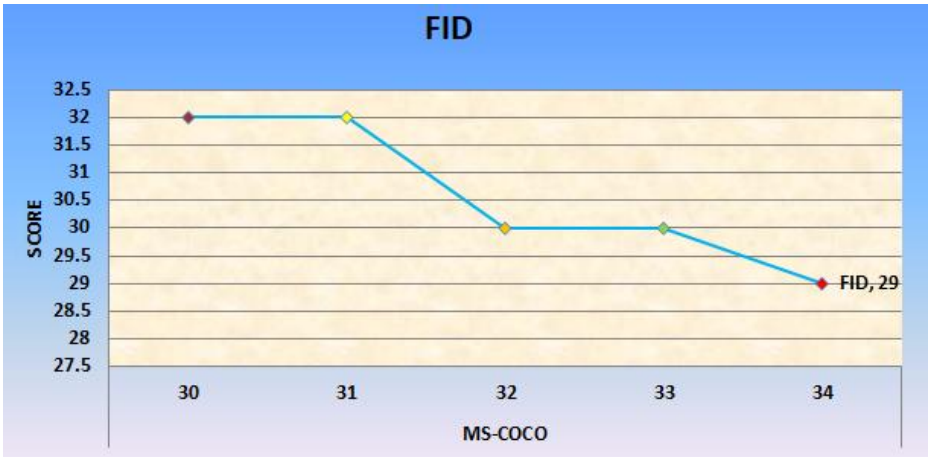

Figure 7. FID score analysis for eight steps in MS-COCO

The metrics were normalized by substituting a defined text-image consistency score (TICS) calculated using CLIP ViT-L/14 for the ambiguous "Accuracy." Fixed procedures are followed by CLIP-I, DINO, and FID (CLIP ViT-H/14, DINOv2-giant, 50k samples with Inception-V3 embeddings). 42 raters provided their HPSs using randomized A/B tests, yielding κ=0.71. All results include 95% CIs with paired t-tests ($p<0.05$).

HPS and image rewards:

- HPS reflects the overall preference of human evaluators for the generated images. It considers factors such as visual quality, coherence with textual prompts, and aesthetic appeal. Human evaluators rank or rate the generated images based on their preference, assigning scores or rankings to each image. The average score or ranking for all images is calculated to obtain the overall HPS score.
- Image rewards quantify the perceived quality or value of generated images, considering factors such as visual fidelity, diversity, and novelty. Assign a numerical reward value to each generated image based on its perceived quality, with higher values indicating better quality. The total image rewards is the sum of reward values assigned to all generated images.

When it comes to quality, coherence, and appeal, HPS represents the general preference of human judges for created images. This is enhanced by image rewards, which quantitatively measure faithfulness, diversity, and novelty; greater values denote higher-quality images.

6.1.5. Stable diffusion cross-modal attention with multi-head attention model generated image

The LAION-1.2B and MS COCO datasets were used to train and assess the SD-CMA-MHA model, which produces contextually appropriate images in response to text cues. The variety of the LAION-1.2B dataset enabled the prompt "Magical Forest with Unicorn" in Figure 8, allowing the model to realistically depict imaginative features such as magical illumination, woodland textures, and the unicorn's form.

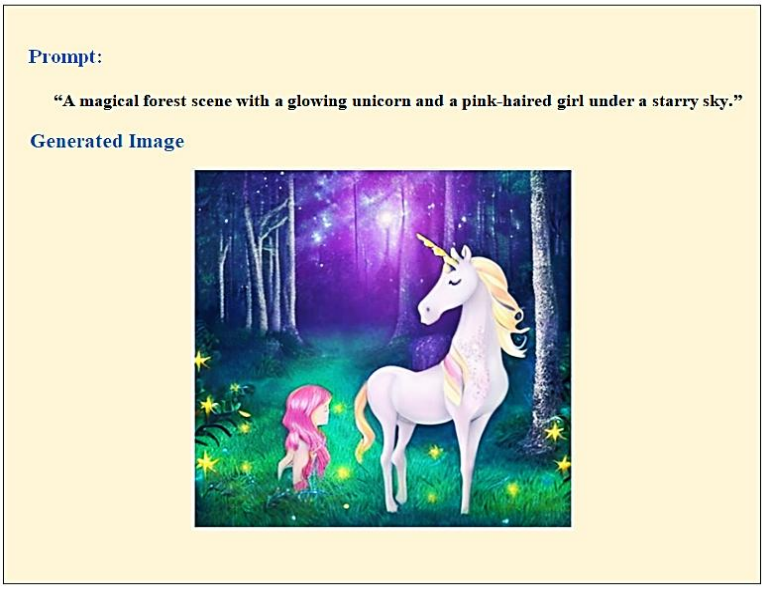

Figure 8. SD-CMA-MHA model generated image of magical forest with Unicorn

The model was able to synthesis architectural characteristics, clothing, and symbolic features that resembled Roman heritage by using the rich captioned samples from the MS COCO dataset for the question "Roman Empires" in Figure 9. Key textual aspects were matched with their visual representations by the cross-modal attention mechanism, whereas resolution and detail were maintained via U-Net and diffusion processes. Collectively, these findings demonstrate how the model may combine visual quality and semantic correctness for both creative and historical tasks.

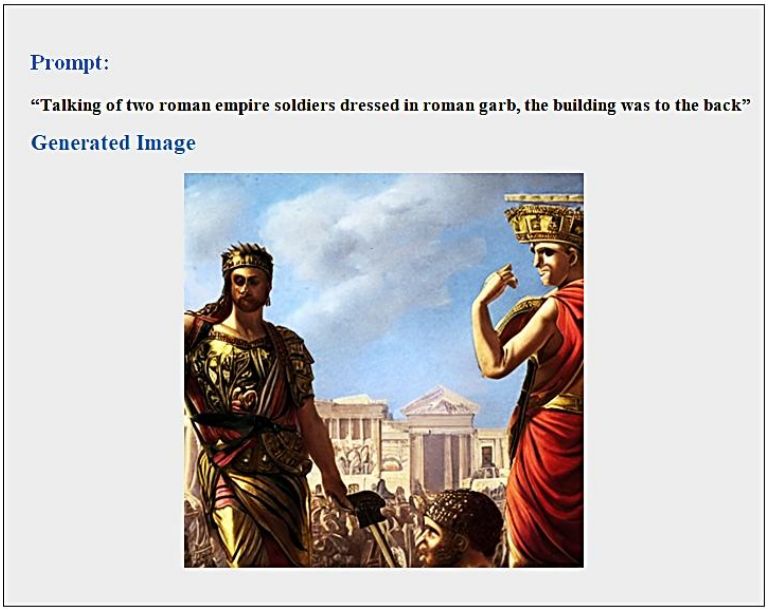

Figure 9. SD-CMA-MHA model generated image of Roman Empires

7. CONCLUSION

Text-to-image creation is improving mainly to the SD-CMA-MHA framework, which combines MHA, cross-modal attention, and stable diffusion to achieve excellent visual fidelity and semantic accuracy. The model achieves 80% accuracy in realistic outputs and a 70% similarity score for text–image alignment with less deviation from genuine images, according to experimental evaluations on the LAION-1.2B and MS-COCO datasets. These quantitative gains outperform those of previous methods. The present research is innovative in that it combines diffusion processes and cross-modal attention in a way that allows the model to capture intricate semantic linkages while maintaining resolution consistency and detail. Beyond these findings, the study highlights the technological robustness and versatility of the framework by offering a scalable and flexible approach that may be used to more general multimodal challenges. As a significant advancement in the field of text-to-image synthesis, these results not only confirm the framework's effectiveness but also highlight its usefulness in fields including education, the creative industries, and human–computer interaction. Future research will investigate how to improve generation quality even further by integrating new developments in machine learning and natural language processing, diversifying datasets, and improving architecture.

ACKNOWLEDGMENTS

This work was supported by the Manonmaniam Sundaranar University, Centre for Information Technology and Engineering, Tirunelveli, Tamil Nadu, India.

FUNDING INFORMATION

Authors state no funding involved

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Kowsalya Veilumuthu	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Divya Chandrasekar						✓		✓	✓	✓	✓	✓		
Sakthidevi Shunmugalingam	✓		✓	✓			✓			✓	✓		✓	✓
Parvathi														

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

REFERENCES

[1] N. Kumari, B. Zhang, S. -Y. Wang, E. Shechtman, R. Zhang, and J. -Y. Zhu, "Ablating Concepts in Text-to-Image Diffusion Models," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, 2023, pp. 22634-22645, doi: 10.1109/ICCV51070.2023.02074.

[2] P. Cao, F. Zhou, Q. Song, and L. Yang, "Controllable generation with text-to-image diffusion models: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 48, no. 4, pp. 4771-4791, Apr. 2026, doi: 10.1109/TPAMI.2025.3646548.

[3] Y. Tewel *et al.*, "Training-free consistent text-to-image generation," *ACM Transactions on Graphics (TOG)*, 2024, doi: 10.48550/arXiv.2402.03286.

[4] L. Yang, Z. Yu, C. Meng, M. Xu, S. Ermon, and B. Cui, "Mastering Text-to-Image Diffusion: Recaptioning, Planning, and Generating with Multimodal LLMs," in *Proceedings of the 41st International Conference on Machine Learning (ICML)*, vol. 235, 2024, pp. 55648–55679.

[5] J. Chen *et al.*, "TextDiffuser: Diffusion models as text painters," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[6] Y. Liu *et al.*, "Cross-modal generative semantic communications for mobile AIGC: Joint semantic encoding and prompt engineering," *IEEE Transactions on Mobile Computing*, vol. 23, no. 12, pp. 14871–14888, 2024, doi: 10.1109/TMC.2024.3449645.

[7] S. Gu *et al.*, "Vector quantized diffusion model for text-to-image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[8] L. Yang *et al.*, "Improving diffusion-based image synthesis with context prediction," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[9] L. Höllein *et al.*, "ViewDiff: 3D-consistent image generation with text-to-image models," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2024, pp. 5043-5052, doi: 10.1109/CVPR52733.2024.00482.

[10] T. Rahman *et al.*, "Visual concept-driven image generation with text-to-image diffusion model," *arXiv preprint*, 2024, doi: 10.48550/arXiv.2402.11487.

[11] J. Ren *et al.*, "Unveiling and mitigating memorization in text-to-image diffusion models through cross attention," in *European Conference on Computer Vision*, 2024, pp. pp 340–356, doi: 10.1007/978-3-031-72980-5_20.

[12] M. Ding *et al.*, "CogView: Mastering text-to-image generation via transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 19822–19835, 2021.

[13] B. Zhang *et al.*, "StyleSwin: Transformer-based generative adversarial network for high-resolution image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 2022, pp. 11294-11304, doi: 10.1109/CVPR52688.2022.01102.

[14] Z. Liu *et al.*, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, QC, Canada, 2021, pp. 9992-10002, doi: 10.1109/ICCV48922.2021.00986.

[15] X. Han, Q. Chen, Z. Xie, X. Li, and H. Yang, "Multiscale progressive text prompt network for medical image segmentation," *Computers & Graphics*, vol. 116, pp. 262–274, 2023, doi: 10.1016/j.cag.2023.08.030.

[16] Y. Yuan, J. Wu, Z. Jing, H. Leung, and H. Pan, "Multimodal image fusion based on hybrid convolutional neural network-transformer and non-local cross-modal attention," *arXiv preprint*, 2022, doi: 10.48550/arXiv.2210.09847.

[17] S. Purushwalkam, A. Gokul, S. Joty, and N. Naik, "BootPIG: Bootstrapping zero-shot personalized image generation capabilities in pretrained diffusion models," in *European Conference on Computer Vision*, 2024, pp. 252–269, doi: 10.1007/978-3-031-91907-7_15.

[18] J. Zhao *et al*., "Uncertainty-Driven Edge Prompt Generation Network for Medical Image Segmentation," in *IEEE Transactions on Medical Imaging*, vol. 44, no. 10, pp. 3950-3961, Oct. 2025, doi: 10.1109/TMI.2025.3535478.

[19] L. Qu, W. Wang, Y. Li, H. Zhang, L. Nie, and T. -S. Chua, "Discriminative probing and tuning for text-to-image generation," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2024, pp. 7434-7444, doi: 10.1109/CVPR52733.2024.00710.

[20] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, 2023, pp. 22500-22510, doi: 10.1109/CVPR52729.2023.02155.

[21] Y. Yang *et al.*, "GlyphControl: Glyph conditional control for visual text generation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[22] A. Mercier *et al.*, "HexaGen3D: Stable Diffusion is just one step away from fast and diverse text-to-3D generation," *arXiv preprint*, 2024, doi: 10.48550/arXiv.2401.07727.

[23] M. Brack, F. Friedrich, D. Hintersdorf, L. Struppek, P. Schramowski, and K. Kersting, "SEGA: Instructing text-to-image models using semantic guidance," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[24] F. T. Zohra, A. D. Gavrilov, O. Z. Duran, and M. Gavrilova, "A linear regression model for estimating facial image quality," in *2017 IEEE 16th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC)*, Oxford, UK, 2017, pp. 130-138, doi: 10.1109/ICCI-CC.2017.8109741.

[25] G. Paulin and M. Ivašić-Kos, "Review and analysis of synthetic dataset generation methods and techniques for application in computer vision," *Artificial Intelligence Review*, vol. 56, pp. 9221–9265, 2023, doi: 10.1007/s10462-022-10358-3.

[26] W. Chen, H. Hu, C. Saharia, and W. W. Cohen, "Re-imagen: Retrieval-augmented text-to-image generator," *arXiv preprint*, 2022, doi: 10.48550/arXiv.2209.14491.

[27] Y. Wei *et al.*, "ELITE: Encoding visual concepts into textual embeddings for customized text-to-image generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15943–15953, 2023, doi: 10.1109/ICCV51070.2023.01463.

[28] D. Li, J. Li, and S. Hoi, "BLIP-Diffusion: Pre-trained subject representation for controllable text-to-image generation and editing," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, pp. 30146–30166, 2023.

[29] J. Ma, J. Liang, C. Chen, and H. Lu, "Subject-Diffusion: Open domain personalized text-to-image generation without test-time fine-tuning," in *ACM SIGGRAPH Conference Papers*, 2024, pp. 1–12, doi: 10.1145/3641519.3657469.

[30] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps," in *NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems*, vol. 35, pp. 5775–5787, 2022.

[31] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models," *Machine Intelligence Research*, vol. 22, pp. 730–751, 2025, doi: 10.1007/s11633-025-1562-4.

[32] C. Meng, R. Rombach, R. Gao, D. Kingma, S. Ermon, J. Ho, and T. Salimans, "On distillation of guided diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 14297–14306, doi: 10.1109/CVPR52729.2023.01372.

[33] Y. Li *et al.* "Snapfusion: Text-to-image diffusion model on mobile devices within two seconds," in *NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023, pp. 20662-20678.

BIOGRAPHIES OF AUTHORS

Kowsalya Veilumuthu is a full-time Research Scholar at the Centre for Information Technology and Engineering, Manonmaniam Sundaranar University. She completed her postgraduate studies (M.Sc. in Information Technology) at the same university in 2021, following her undergraduate degree (B.Sc. in Information Technology) from Rosemary College of Arts and Science, affiliated with Manonmaniam Sundaranar University, in 2019. Between 2022 and 2024, she authored three books indexed in Scopus. As a reviewer for Elsevier Publications, she has reviewed 14 research papers, with credits added to her ORCID by the editors. Her research interests lie in the fields of deep learning and sentiment analysis. She can be contacted at email: kowsalyaphd260@gmail.com.

Dr. Divya Chandrasekar an Assistant Professor at the Centre for Information Technology and Engineering within Manonmaniam Sundaranar University, has made significant contributions to the field. She holds a Ph.D. from the same university and has authored more research papers in International or National journals or Proceedings or Books. Her actively participates in scholarly activities, serving as a reviewer for international journals and being part of editorial boards. Her current research interests include data analytics, cyber security, nanodevices and low power VLSI circuits wireless sensor networks, and communication networks. With a strong background in engineering, she continues to impact the academic community through her research and teaching. She was awarded the Young Scientists Fellowship by TNSCST. She can be contacted at email: cdivyame@gmail.com.

Sakthidevi Shunmugalingam Parvathi is a full-time Research Scholar at the Centre for Information Technology and Engineering, Manonmaniam Sundaranar University. She completed her postgraduate studies (M.Sc. in Information Technology) at the same university in 2023, following her undergraduate degree (B.Sc. in Information Technology) from Sri Ram Nallamani Yadava College of Arts and Science, affiliated with Manonmaniam Sundaranar University, in 2021. She authored one book indexed in Scopus. Also, she authored two journal papers indexed in Scopus. As a reviewer for Elsevier and Springer Publications, she has reviewed 27 research papers, with credits added to her ORCID by the editors. Her research interests lie in the fields of deep learning and artificial intelligence. She can be contacted at email: spsakthidevi2000@gmail.com.