International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Text-to-Image and Image-to-Image Generation Using Generative AI Diffusion Models

K. Vinay Kumar[1] G. Yogender[1] T. Uday Simhadri[1]
b201556@rgukt.ac.in b200116@rgukt.ac.in b201290@rgukt.ac.in
Department of Computer Science and Engineering
Rajiv Gandhi University of Knowledge Technologies, Basar, India

Dr. B. Venkat Raman

Assistant Professor
Department of Computer Science and Engineering
Rajiv Gandhi University of Knowledge Technologies, Basar, India
Email: bvraman@rgukt.ac.in

***Abstract*—Generative AI has revolutionized image synthesis through advanced deep learning models. Among these, diffusion models have demonstrated superior performance in generating high-quality and semantically aligned images from textual descriptions. However, existing systems face limitations in handling long and complex text prompts, maintaining structural consistency, and achieving efficient image transformations.**

This paper presents a diffusion-based framework for both text-to-image and image-to-image generation. The proposed system integrates segment-level text encoding, preference reweighting, and latent diffusion mechanisms to enhance semantic alignment and image quality. By leveraging models such as Stable Diffusion, CLIP embeddings, and Variational Autoencoders (VAE), the system efficiently generates realistic and context-aware images. The results demonstrate improved alignment between textual input and generated visuals, making the system suitable for real-world applications such as design, education, and creative industries.

***Index Terms*—Generative AI, Diffusion Models, Text-to-Image, Image-to-Image, Stable Diffusion, CLIP, VAE**

I. INTRODUCTION

A. Background and Motivation

The rapid advancement of Generative Artificial Intelligence (AI) has significantly transformed the field of image synthesis, enabling machines to generate high-quality visual content from textual descriptions. Text-to-image generation represents a crucial breakthrough in bridging the gap between natural language understanding and computer vision. This technology allows users to convert descriptive text prompts into realistic and semantically meaningful images, opening new possibilities in areas such as digital design, content creation, education, and entertainment.

Modern generative models, particularly diffusion-based architectures such as Stable Diffusion, have achieved remarkable success in producing high-resolution and visually coherent images. These models operate by iteratively refining random noise into meaningful images guided by textual embeddings. Additionally, image-to-image generation techniques further extend this capability by transforming existing images using textual guidance while preserving structural consistency.

Despite these advancements, existing systems face several limitations. Most text-to-image models struggle to accurately interpret long and complex textual prompts due to token length constraints in text encoders such as CLIP. As a result, generated images often fail to capture all relevant details described in the input. Furthermore, many models suffer from high computational requirements and may overfit to visually appealing features rather than maintaining strong semantic alignment with the input text.

To address these challenges, this project focuses on developing an enhanced diffusion-based framework for both text-to-image and image-to-image generation. The proposed approach incorporates segment-level text encoding to handle long prompts, preference reweighting to emphasize semantically important features, and latent diffusion techniques to improve computational efficiency. By leveraging advanced models such as Stable Diffusion, CLIP embeddings, and Variational Autoencoders (VAE), the system aims to generate high-quality, context-aware images with improved alignment between textual input and visual output.

B. Challenges in Text-to-Image and Image-to-Image Generation

Despite significant advancements in generative AI, text-to-image and image-to-image generation systems face several critical challenges. First, existing models struggle to effectively process long and complex textual prompts. Most text encoders, such as CLIP, have limitations in token length, which restrict their ability to capture detailed descriptions. As a result, generated images often fail to represent all elements present in multi-sentence or paragraph-level inputs.

Second, achieving accurate semantic alignment between textual input and generated images remains a major challenge. Many models tend to focus on visually dominant features such as colors or aesthetics rather than faithfully representing the intended meaning of the text. This leads to partial or incorrect interpretations of the prompt, especially when multiple objects, relationships, or styles are described.

Third, computational complexity is a significant limitation. High-quality image generation using diffusion models requires substantial computational resources, including powerful GPUs and large memory capacity. This makes real-time generation and deployment challenging, particularly in resource-constrained environments.

Fourth, maintaining structural consistency in image-to-image transformation tasks is difficult. While modifying an existing image based on a text prompt, the model must preserve important spatial and structural details while applying meaningful transformations. Many existing approaches either distort the original structure or fail to apply precise modifications.

Finally, models often suffer from overfitting to text-irrelevant visual features during training. This reduces the model's ability to generalize and weakens the connection between textual input and generated output.

To address these challenges, advanced techniques such as segment-level text encoding, preference reweighting, and latent diffusion architectures have been proposed. These ap-proaches improve the model's ability to handle long prompts, enhance semantic alignment, and reduce computational over-head, leading to more accurate and efficient image generation.

C. Related Work and Existing Approaches

Previous research in text-to-image generation has evolved significantly over the past decade, progressing from tradi-tional Generative Adversarial Networks (GANs) to modern diffusion-based architectures. Early approaches such as GAN-INT-CLS and StackGAN introduced the concept of gener-ating images from textual descriptions. While these models improved image resolution, they often suffered from unstable training, mode collapse, and limited diversity in generated outputs.

Subsequent models like AttnGAN incorporated attention mechanisms to align words with specific image regions, improving fine-grained detail generation. However, these models still struggled with maintaining global coherence and accurately capturing complex textual relationships.

More recent advancements introduced autoregressive and transformer-based models such as DALL·E, which demonstrated strong capabilities in generating creative and diverse images from text prompts. Despite their effectiveness, these models required significant computational resources and exhibited slower generation speeds.

The introduction of diffusion-based models, particularly Stable Diffusion, marked a major breakthrough in text-to-image generation. These models operate in a latent space, significantly improving computational efficiency while generating high-quality and realistic images. Stable Diffusion combines text encoders such as CLIP with latent diffusion processes to achieve strong alignment between text and generated visuals.

Recent research has further focused on improving semantic alignment and efficiency. Models such as PixArt-alpha and CogView3 have introduced diffusion transformers and relay diffusion techniques to enhance image quality and speed. Additionally, approaches like Hunyuan-DiT leverage multi-resolution diffusion transformers for better semantic understanding.

However, existing systems still face limitations in handling long and complex text prompts due to constraints in text encoders like CLIP. Many models fail to fully capture all elements described in detailed inputs, leading to incomplete or inaccurate image generation. Furthermore, there is limited exploration of techniques that explicitly address long-text alignment and preference optimization.

To overcome these challenges, recent approaches such as LongAlign propose segment-level text encoding and preference decomposition strategies. These methods aim to improve alignment between textual input and generated images by processing long prompts more effectively and emphasizing semantically relevant features during generation.

D. Proposed Work and Contributions

This project presents a comprehensive study on text-to-image and image-to-image generation using advanced generative AI diffusion models. The work is structured into two major phases to systematically enhance image generation quality, semantic alignment, and computational efficiency.

In the first phase, a baseline text-to-image generation system was implemented using latent diffusion models. The system leverages Stable Diffusion along with CLIP-based text encoders to convert textual prompts into visual representations. This phase focuses on understanding the core diffusion pipeline, including text encoding, latent space representation, and iterative denoising. The generated outputs were analyzed in terms of visual quality, coherence, and alignment with input prompts.

In the second phase, the study extends the baseline ap-proach by introducing an enhanced framework for handling complex and long textual inputs. A segment-level text en-coding mechanism was implemented, where long prompts are divided into smaller segments and processed individually. The resulting embeddings are combined using mean pooling to form a unified representation. Additionally, a preference reweighting strategy is applied to emphasize semantically important features while reducing the influence of irrelevant components. This improved representation is then used to guide the diffusion process for better alignment between text and generated images.

Furthermore, the project incorporates an image-to-image generation pipeline using latent diffusion techniques. In this approach, input images are encoded into latent space using Variational Autoencoders (VAE), followed by controlled noise injection and text-guided transformation. This enables precise and semantically guided image modifications while preserving structural consistency.

The key contributions of this work are as follows:

1) Implementation of a baseline text-to-image generation system using Stable Diffusion and CLIP embeddings.
2) Development of a segment-level text encoding approach to effectively handle long and complex textual prompts.

3) Introduction of a preference reweighting mechanism to improve semantic alignment between text and generated images.
4) Integration of latent diffusion-based image-to-image generation for controlled and structure-preserving transformations.
5) Comprehensive analysis of generated outputs in terms of quality, coherence, and prompt alignment.

E. Organization of the Paper

The remainder of this paper is organized as follows. Section II describes the dataset, system requirements, and preprocessing techniques used for text and image inputs. Section III presents the proposed methodology, including the architecture of diffusion models, segment-level text encoding, and image-to-image generation pipeline. Section IV discusses the experimental results, generated outputs, and performance analysis in terms of image quality and semantic alignment. Finally, Section V concludes the paper and outlines potential future work in advancing generative AI systems for more efficient, scalable, and semantically accurate image generation.

II. DATASET AND SYSTEM OVERVIEW

A. Dataset Collection

For this project, a combination of publicly available image datasets and text–image paired data was utilized to support both text-to-image and image-to-image generation tasks. Since diffusion models require large-scale data for effective training and generalization, the system leverages pre-trained models such as Stable Diffusion, which are trained on extensive datasets containing millions of image–text pairs.

The text-to-image generation component relies on natural language prompts provided by users. These prompts describe visual scenes, objects, styles, and relationships, which are then converted into semantic embeddings using text encoders such as CLIP. Instead of manually collecting and annotating datasets, this project utilizes the knowledge already embedded in pre-trained diffusion models, significantly reducing the need for large-scale custom dataset preparation.

For image-to-image generation, input images are provided by the user and processed through a Variational Autoencoder (VAE) to obtain latent representations. These images may include real-world photographs, sketches, or synthetic visuals, enabling diverse transformation tasks guided by textual input.

Unlike traditional supervised learning approaches, diffusion-based models do not rely heavily on task-specific annotated datasets. Instead, they learn generalized representations from large-scale pretraining and adapt to specific tasks through fine-tuning or prompt engineering. This makes the system more flexible and scalable for various applications.

Additionally, preprocessing steps are applied to ensure input consistency. Text prompts are normalized and, in the case of long descriptions, segmented into smaller units to overcome token length limitations. For image inputs, resizing and normalization are performed before encoding into latent space. This structured data preparation improves the overall efficiency and quality of the generation process.

B. Prompt Design and Input Preparation

Unlike traditional supervised learning tasks that require manual annotation and labeling, generative AI diffusion models rely on text prompts and input images to guide the generation process. Therefore, instead of annotation, this project focuses on effective prompt design and input preparation.

For the text-to-image generation task, natural language prompts are used to describe the desired visual output. These prompts may include details such as objects, scenes, colors, styles, and relationships between elements. The quality and specificity of the prompt play a crucial role in determining the accuracy and realism of the generated image.

To improve the handling of complex and long textual inputs, prompts are segmented into smaller meaningful components. Each segment is processed independently using a text encoder such as CLIP to generate embeddings. These embeddings are then combined using mean pooling to form a unified representation. This approach helps overcome token length limitations and ensures that all key elements of the prompt are captured effectively.

Additionally, a preference reweighting mechanism is applied to emphasize semantically important features within the prompt. This allows the model to focus more on critical aspects of the description, such as primary objects or artistic styles, while reducing the influence of less relevant details.

For image-to-image generation, input images are prepared by resizing and normalizing them before encoding into latent space using a Variational Autoencoder (VAE). The transformation process is guided by both the input image and the corresponding text prompt, enabling controlled and meaningful modifications.

Overall, careful prompt design and structured input preparation serve as a crucial alternative to traditional annotation processes, ensuring better alignment between textual input and generated visual output.

C. Dataset Characteristics

The dataset and input configuration used in this project for text-to-image and image-to-image generation exhibit several important characteristics:

1) **Large-Scale Pretrained Data:**
The system leverages pre-trained diffusion models such as Stable Diffusion, which are trained on millions of image–text pairs. This large-scale training enables the model to learn rich visual and semantic representations without requiring task-specific dataset collection.
2) **Text–Image Pair Representation:**
The underlying training data consists of paired textual descriptions and corresponding images. This allows the model to establish strong relationships between language and visual content, enabling accurate image generation from natural language prompts.

3) **Diverse Visual Concepts:**
The pretrained dataset includes a wide range of objects, scenes, artistic styles, and compositions. This diversity allows the system to generate images across various domains such as landscapes, human activities, abstract art, and realistic scenarios.
4) **Variable Prompt Length:**
Input text prompts can vary from short phrases to long descriptive sentences. However, due to token length limitations in text encoders like CLIP, long prompts are segmented into smaller units to ensure effective processing and better semantic coverage.
5) **Multi-Modal Input Support:**
The system supports both text and image inputs. While text prompts guide the generation process, input images (in image-to-image tasks) provide structural and visual context, enabling controlled transformations.
6) **Latent Space Representation:**
Instead of operating directly on pixel space, the system uses latent representations obtained through Variational Autoencoders (VAE). This significantly reduces computational complexity and improves efficiency during image generation.

D. Input Representation and Visualization

- **Text Prompt and Generated Output Distribution**
In this study, instead of traditional dataset splits such as training, validation, and testing, the system relies on pre-trained diffusion models and evaluates performance through qualitative analysis of generated outputs. The input to the system consists of natural language prompts of varying complexity, which describe objects, scenes, and artistic styles.
These prompts are processed using a text encoder to generate semantic embeddings, which guide the image generation process. The effectiveness of the system is evaluated based on how accurately the generated images reflect the input descriptions.
- **Intermediate Generation Stages:**
The diffusion process begins with random noise and iteratively refines it into a meaningful image using the guidance of text embeddings. Each step progressively reduces noise while enhancing structural and semantic details.
- **Image-to-Image Transformation Visualization:** For image-to-image generation, an input image is transformed based on a given text prompt. The model preserves the structural elements of the original image while applying semantic modifications guided by the textual input.
- **Prompt Diversity and Output Distribution:**
The system processes a wide range of text prompts varying in length, complexity, and semantic detail. These prompts may describe simple objects, complex scenes, or artistic styles. The diversity in input prompts allows the model to generate a variety of images across different domains.
Unlike traditional classification datasets, there is no concept of class distribution (such as toxic vs non-toxic). Instead, the effectiveness of the model is evaluated based on how well the generated images align with the input prompts. A diverse set of prompts ensures that the model is tested across multiple scenarios, improving its robustness and generalization capability.

Fig. 1. Sample Prompts and Corresponding Generated Outputs

- **Frequent Visual Concepts and Feature Representation:**
Figure X illustrates commonly generated visual elements such as objects (e.g., animals, vehicles), environments (e.g., landscapes, urban scenes), and artistic styles (e.g., watercolor, realistic, cartoon). These recurring visual concepts indicate the model's ability to learn and reproduce patterns from large-scale pretraining data.
However, the presence of specific visual elements alone does not guarantee accurate generation, as proper spatial relationships, context, and style interpretation are equally important. This highlights the importance of semantic alignment between text prompts and generated images.
The analysis of frequently generated features helps in understanding the strengths and limitations of the model. It also demonstrates the effectiveness of techniques such as segment-level encoding and preference reweighting in improving the representation of complex prompts and generating more coherent visual outputs.

III. Methodologies and Model Framework

A. Overall Framework

The primary objective of this project is to develop an automated system for text-to-image and image-to-image generation using advanced diffusion-based generative AI models. The methodology is structured into two major phases. The

first phase focuses on implementing a baseline text-to-image generation system using latent diffusion models. The second phase enhances the system by introducing segment-level text encoding and preference reweighting techniques to improve semantic alignment and handle long textual prompts effectively.

The complete workflow of the proposed system consists of the following stages:

1) Input text prompt or input image with text guidance
2) Text preprocessing and segmentation (for long prompts)
3) Text encoding using CLIP to generate embeddings
4) Latent space representation using Variational Autoencoder (VAE)
5) Diffusion-based image generation through iterative denoising
6) Image decoding and output generation

B. Input Processing and Preprocessing

Preprocessing plays an important role in improving the quality of generated outputs. Unlike traditional text classification tasks, generative models require structured input preparation rather than heavy data cleaning.

- Normalization of text prompts (removal of unnecessary symbols)
- Segmentation of long prompts into smaller meaningful components
- Preservation of semantic keywords (objects, styles, relationships)
- Conversion of text into embeddings using CLIP encoder
- Image resizing and normalization for image-to-image tasks

For handling long textual descriptions, the input prompt is divided into multiple segments. Each segment is encoded independently, and the resulting embeddings are combined using mean pooling to form a unified representation. This approach helps overcome token length limitations and ensures better coverage of complex prompts.

C. Diffusion-Based Generation Process

The core of the system is based on diffusion models, which generate images by iteratively refining random noise into meaningful visual content. The process consists of two main stages: forward diffusion and reverse diffusion.

Forward Diffusion:
In this stage, noise is gradually added to an image until it becomes completely random.

Reverse Diffusion:
The model learns to remove noise step by step, guided by text embeddings, to generate a coherent image.

D. Image-to-Image Generation Pipeline

In addition to text-to-image generation, the system supports image-to-image transformation. In this approach:

1) Input image is encoded into latent space using VAE
2) Controlled noise is added to the latent representation
3) Text embeddings guide the transformation process

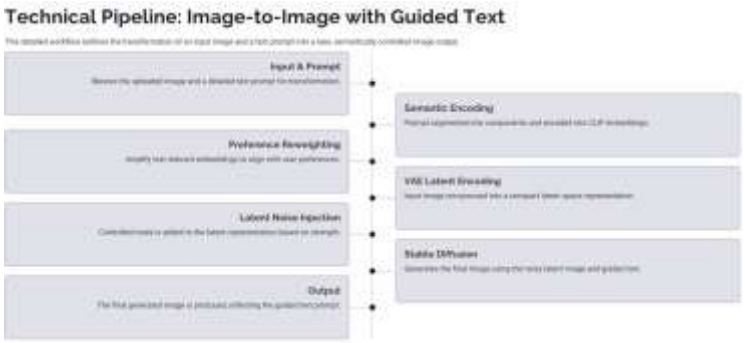

Fig. 2. Proposed Diffusion-Based Generation Framework

4) Diffusion model generates a modified output image

This pipeline ensures that the generated image maintains structural consistency while incorporating semantic changes specified in the text prompt.

E. Preference Reweighting Mechanism

To improve semantic alignment, a preference reweighting strategy is applied to text embeddings. This mechanism emphasizes important features such as key objects and styles while reducing the influence of less relevant details.

This enhances the model's ability to accurately interpret complex prompts and generate visually coherent outputs.

F. Model Training and Fine-Tuning

The system utilizes pre-trained diffusion models such as Stable Diffusion, which are trained on large-scale datasets. Instead of training from scratch, the model is fine-tuned using prompt-based adjustments and optimization techniques.

Fine-tuning focuses on:

- Improving prompt-to-image alignment
- Enhancing visual quality
- Reducing noise artifacts

G. Core Generative Models

Unlike traditional machine learning and NLP-based classification models, this project utilizes advanced generative AI models based on diffusion architectures. These models learn to generate images from textual descriptions by capturing deep semantic relationships between language and visual content.

*1) **CLIP (Contrastive Language–Image Pretraining)**:*
CLIP is a multimodal model that learns a joint representation of text and images. It plays a crucial role in converting textual prompts into meaningful embeddings that guide the image generation process.

In this project:

- Text prompts are encoded into high-dimensional embeddings using CLIP.
- These embeddings capture semantic relationships between words and visual concepts.
- The embeddings are used to guide the diffusion model during image generation.

CLIP enables strong alignment between textual input and generated images by ensuring that the visual output corresponds closely to the semantic meaning of the prompt.

*2) **Stable Diffusion Model**:* Stable Diffusion is a latent diffusion model that generates high-quality images efficiently. Unlike traditional pixel-based models, it operates in a compressed latent space, reducing computational complexity.

In this project:

- Pre-trained Stable Diffusion is used for image generation.
- The model iteratively refines random noise into a meaningful image.
- Text embeddings from CLIP guide the denoising process.

Stable Diffusion provides a balance between image quality and computational efficiency, making it suitable for real-world applications.

*3) **Variational Autoencoder (VAE)**:* The Variational Autoencoder is used to encode images into latent representations and decode them back into pixel space.

- Input images are compressed into latent space using VAE encoder.
- Generated latent representations are decoded into final images.
- Reduces memory usage and improves processing speed.

VAE enables efficient manipulation of images during both text-to-image and image-to-image tasks.

H. Enhanced Diffusion Framework

*1) **Segment-Level Text Encoding**:* To handle long and complex prompts, the input text is divided into smaller segments. Each segment is encoded independently using CLIP, and the resulting embeddings are combined using mean pooling.

- Overcomes token length limitations
- Captures all important elements of long prompts
- Improves semantic coverage

*2) **Preference Reweighting Mechanism**:* A preference reweighting strategy is applied to emphasize important features in the text embeddings.

- Assigns higher weights to key objects and styles
- Reduces influence of irrelevant features
- Improves alignment between text and generated images

I. Image-to-Image Transformation Model

In addition to text-to-image generation, the system supports image-to-image transformation using latent diffusion.

- Input image is encoded into latent space using VAE
- Controlled noise is added to latent representation
- Text embeddings guide the transformation process
- Output image is generated through diffusion decoding

This approach preserves structural consistency while enabling semantic modifications based on textual input.

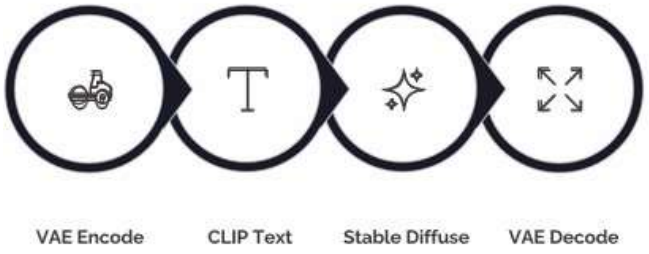

Fig. 3. Diffusion-Based Text-to-Image and Image-to-Image Framework

J. Drawbacks of Existing Systems

Existing text-to-image and image-to-image generation systems, particularly early GAN-based and basic diffusion models, suffer from several limitations:

- **Limited Handling of Long Text Prompts:** Most models rely on text encoders such as CLIP, which have token length limitations. This restricts their ability to process long and complex descriptions, leading to incomplete or inaccurate image generation.
- **Weak Semantic Alignment:** Many models fail to fully capture relationships between multiple objects, attributes, and styles described in the input prompt. As a result, generated images may not accurately reflect the intended meaning.
- **High Computational Cost:** Diffusion models require multiple iterative denoising steps, making them computationally expensive and difficult to deploy in real-time environments.
- **Poor Structural Consistency in Image Transformation:** In image-to-image tasks, existing models often distort the original structure of the input image while applying transformations.
- **Overfitting to Visual Features:** Models sometimes prioritize visual aesthetics over semantic correctness, reducing the alignment between text and generated images.

K. Advantages of Proposed System

The proposed system enhances the generative framework by integrating advanced diffusion techniques along with improved text representation strategies.

- **Improved Long-Text Understanding:** Segment-level text encoding allows the system to process long and complex prompts effectively by dividing them into manageable components.
- **Better Semantic Alignment:** Preference reweighting emphasizes important features in the text embeddings, improving the correspondence between textual input and generated images.
- **Efficient Latent Space Processing:** The use of latent diffusion models reduces computational complexity while maintaining high-quality output.
- **Structural Preservation in Image-to-Image Tasks:** The VAE-based latent representation ensures that important

structural features of the input image are retained during transformation.

- **Scalability and Flexibility:** The system leverages pre-trained models, enabling efficient adaptation to various domains without requiring large-scale task-specific datasets.

IV. Experimental Results and Analysis

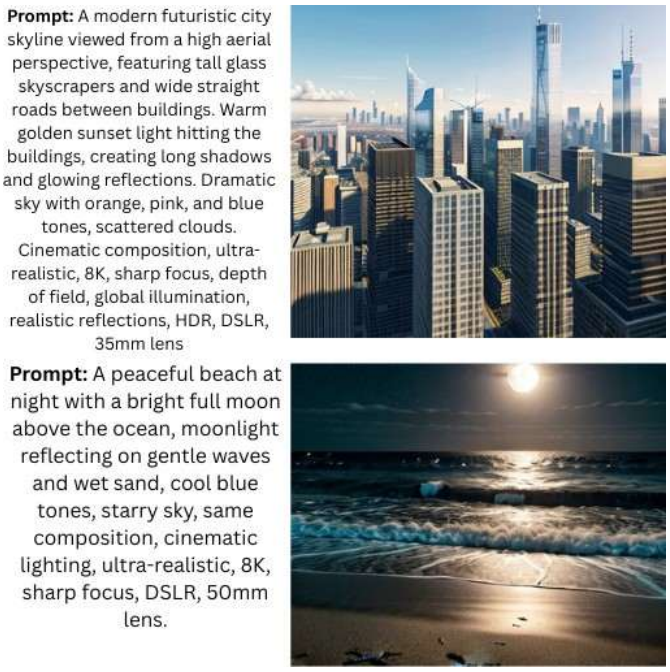

Fig. 4. Generated Image Quality and Prompt Alignment Analysis

A. Evaluation Strategy

Unlike traditional machine learning models, evaluation of generative AI systems is primarily qualitative and based on visual inspection and semantic alignment. The generated images are analyzed based on the following criteria:

- **Visual Quality:** Clarity, realism, and absence of artifacts in generated images.
- **Semantic Alignment:** Degree to which the generated image matches the input text prompt.
- **Diversity:** Ability of the model to generate varied outputs for different prompts.
- **Structural Consistency:** Preservation of key elements in image-to-image transformations.

B. Observations

The experimental results demonstrate that the proposed system produces high-quality images with improved alignment to input prompts. The use of segment-level encoding significantly enhances the handling of long and complex descriptions. Additionally, preference reweighting improves the representation of important visual features, resulting in more coherent and context-aware outputs.

The image-to-image generation pipeline successfully maintains structural consistency while applying meaningful transformations guided by textual input. Overall, the system shows improved performance compared to conventional approaches in terms of quality, accuracy, and flexibility.

C. Evaluation Metrics

Unlike traditional classification tasks, text-to-image and image-to-image generation are evaluated using qualitative and perceptual metrics. Since the output is visual, performance is measured based on image quality, semantic alignment, and diversity rather than discrete class labels.

1) ***Visual Quality****:* Visual quality measures the realism, clarity, and overall aesthetic appeal of generated images. High-quality images should be free from noise artifacts and distortions while maintaining sharp details and coherent structure.

2) ***Semantic Alignment****:* Semantic alignment evaluates how accurately the generated image reflects the input text prompt. A well-aligned output should correctly represent objects, relationships, styles, and contextual details described in the prompt.

3) ***CLIP Score****:* CLIP score is a widely used metric that measures the similarity between text prompts and generated images. It computes the cosine similarity between text and image embeddings produced by the CLIP model.

$$\text{CLIP Score} = \cos(E_{text}, E_{image})$$

A higher CLIP score indicates better alignment between the generated image and the input prompt.

4) ***Fre´chet Inception Distance (FID)****:* FID is used to evaluate the quality of generated images by comparing their distribution with real images.

$$FID = ||\mu_r - \mu_g||^2 + Tr(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2})$$

where μ_r, Σ_r represent the mean and covariance of real images, and μ_g, Σ_g represent those of generated images.

Lower FID scores indicate higher similarity between generated and real images, implying better quality.

5) ***Diversity****:* Diversity measures the model's ability to generate varied outputs for different prompts or variations of the same prompt. A good generative model should avoid producing repetitive or similar images.

6) ***Structural Consistency****:* This metric is particularly important for image-to-image generation. It evaluates how well the model preserves the structural elements of the input image while applying semantic modifications.

7) ***Human Evaluation****:* Human evaluation involves manual inspection of generated images to assess realism, relevance, and creativity. This is often considered the most reliable metric for generative tasks.

By utilizing these evaluation metrics, the performance of the proposed system is assessed in terms of image quality, semantic accuracy, and generation diversity. These metrics provide a comprehensive understanding of the effectiveness of diffusion-based models in generating high-quality and context-aware images.

V. Future Scope

A. Potential Enhancements for the Proposed System

*1) **Integration of Advanced Diffusion Models:**:* Future work can explore the implementation of more advanced diffusion architectures such as Diffusion Transformers (DiT) and next-generation Stable Diffusion models. These models can improve image quality, generation speed, and semantic alignment with complex text prompts.

*2) **Handling Long and Complex Prompts:**:* Although segment-level encoding improves prompt understanding, further research can focus on integrating large language models (LLMs) to enhance semantic comprehension of long and detailed textual inputs.

*3) **Real-Time Image Generation:**:* Current diffusion models are computationally intensive. Future improvements can include model optimization techniques such as model compression, quantization, and efficient sampling methods to enable real-time image generation on edge devices.

*4) **Multi-Modal Input Expansion:**:* The system can be extended to support multiple input modalities such as sketches, voice descriptions, and video frames. This would allow users to interact with the model more intuitively and enable richer content generation.

*5) **Video and 3D Content Generation:**:* Future research can extend the system from static image generation to video-to-video and text-to-video generation. Additionally, 3D scene generation can be explored for applications in virtual reality (VR) and gaming.

*6) **Fine-Grained Image Editing:**:* Enhancements can be made to enable localized editing, where specific regions of an image can be modified based on textual instructions while preserving the rest of the image.

*7) **Bias Reduction and Ethical AI:**:* Future work can focus on improving dataset diversity and reducing biases in generated outputs. Ensuring fairness, safety, and ethical use of generative AI systems is an important research direction.

*8) **Explainable AI for Generative Models:**:* Incorporating explainability techniques such as attention visualization and feature attribution can improve transparency and help users understand how the model generates images based on text prompts.

*9) **Deployment and Integration:**:* The system can be deployed as a web application, API service, or integrated into creative design tools. This would enable practical usage in industries such as digital art, marketing, education, and content creation.

The future scope of this project involves enhancing diffusion-based generative models to achieve better semantic understanding, faster generation, and broader application capabilities. By integrating advanced architectures, optimizing performance, and expanding to multi-modal and real-time systems, the proposed framework can evolve into a powerful and scalable solution for next-generation AI-driven content creation.

VI. Conclusion

This project presents a comprehensive study on text-to-image and image-to-image generation using advanced diffusion-based generative AI models. In the first phase, a baseline system was implemented using Stable Diffusion and CLIP-based text encoders to generate images from natural language prompts. The results demonstrated that diffusion models are highly effective in producing realistic and high-quality images with strong visual coherence.

The study further explored the limitations of existing systems, particularly in handling long and complex textual inputs and maintaining semantic alignment. To address these challenges, the proposed approach introduced segment-level text encoding and preference reweighting techniques. These enhancements improved the model's ability to capture detailed semantic information and generate images that more accurately reflect the input prompts.

Additionally, the integration of image-to-image generation using latent diffusion and Variational Autoencoders (VAE) enabled controlled and structure-preserving transformations. This approach ensures that generated outputs maintain both visual consistency and semantic relevance.

Overall, the proposed system demonstrates improved performance in terms of image quality, semantic alignment, and flexibility compared to conventional approaches. The project highlights the potential of diffusion-based generative models in bridging the gap between natural language and visual content, contributing to the development of scalable and efficient AI-driven content generation systems for real-world applications.

References

[1] R. Rombach et al., "High-Resolution Image Synthesis with Latent Diffusion Models," Proceedings of CVPR, 2022.
[2] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," ICML, 2021.
[3] A. Ramesh et al., "Zero-Shot Text-to-Image Generation," ICML, 2021.
[4] T. Xu et al., "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks," CVPR, 2018.
[5] H. Zhang et al., "StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks," ICCV, 2017.
[6] J. Chen et al., "PixArt-alpha: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis," 2023.
[7] W. Zheng et al., "CogView3: Finer and Faster Text-to-Image Generation via Relay Diffusion," 2024.
[8] Z. Li et al., "Hunyuan-DiT: Multi-Resolution Diffusion Transformer for Text-to-Image Generation," 2024.
[9] L. Liu et al., "LongAlign: Segment-Level Encoding and Preference Decomposition for Long Text-to-Image Generation," 2025.
[10] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," ICLR, 2014.
[11] J. Ho et al., "Denoising Diffusion Probabilistic Models," NeurIPS, 2020.
[12] R. Rombach et al., "Latent Diffusion Models for Efficient Image Generation," CVPR, 2022.
[13] K. Crowson et al., "CLIP-Guided Diffusion for Image Generation," 2022.
[14] Stability AI, "Stable Diffusion 3: Multimodal Diffusion Transformer," 2024.
[15] W. Peebles and S. Xie, "Scalable Diffusion Models with Transformers," ICCV, 2023.
[16] C. Saharia et al., "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding," NeurIPS, 2022.
[17] A. Nichol et al., "GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models," ICML, 2022.

[18] P. Ramesh et al., "Hierarchical Text-Conditional Image Generation with CLIP Latents," 2022.
[19] Stability AI, "SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis," 2023.
[20] L. Zhang et al., "Adding Conditional Control to Text-to-Image Diffusion Models," 2023.
[21] N. Ruiz et al., "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation," 2022.
[22] E. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," ICLR, 2022.
[23] A. Hertz et al., "Style Aligned Image Generation via Shared Attention," 2023.
[24] Y. Fan et al., "Reinforcement Learning for Fine-Tuning Diffusion Models," 2024.
[25] K. Baltrus˘aitis et al., "Multimodal Machine Learning: A Survey and Taxonomy," IEEE TPAMI, 2019.
[26] I. Goodfellow et al., "Generative Adversarial Nets," NeurIPS, 2014.