Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 10s (2025)

A Unified Framework for High-Resolution Text-to-Image Generation Using Stable Diffusion with Adaptive Upscaling via Real-ESRGAN and EDSR

1st Shital.P.Mohite
Pillai HOC College of Engineering and Technology, Rasayani,
University of Mumbai
Mumbai, India

2nd Jagdish.W.Bakal
Pillai HOC College of Engineering and Technology, Rasayani,
University of Mumbai
Mumbai, India

3rd Dr. Rajashree Gadhave
Pillai HOC College of Engineering and Technology, Rasayani,
University of Mumbai
Mumbai, India

Article History:
Received: 06-08-2025
Revised: 27-09-2025
Accepted: 26-10-2025

Abstract:

Text-to-image generation has seen tremendous progress with diffusion-based models like Stable Diffusion, but their native output resolutions often fall short for practical applications. To address this, we propose a unified, efficient, and modular pipeline, GENERATE-AND-UPSCALE-IMAGE, which integrates state-of-the-art image synthesis and adaptive super-resolution techniques. The system produces base images from text inputs with Stable Diffusion and can upscale them with EDSR or Real-ESRGAN. With mixed-precision inference, we enable both CPU and GPU execution with hardware-aware optimisation. We demonstrate the superiority, flexibility, and efficacy of this pipeline through experimental analysis, mathematical modelling, and visual comparisons

Keywords— Latent Diffusion Models, Super-Resolution, Real-ESRGAN, EDSR, Text-to-Image Synthesis, Stable Diffusion, Image Upscaling

Introduction

Text-to-image generation is a cross-modal task that merges image creation and natural language recognition. With the use of generative modelling strategies, new models such as DALL•E, Imagen, and Stable Diffusion have made great leaps.

Despite these impressive semantic consistency levels, their resolution output is limited, and post-processing is often required to enhance image quality.

- To circumvent this limit, we introduce the two-stage Text-to-image generation using Stable Diffusion.
- Super-resolution through either Real-ESRGAN or EDSR.

Our approach integrates model selection and device-aware optimization, offering a user-configurable, lightweight, and scalable image synthesis pipeline.

The task of text-to-image synthesis represents a fascinating intersection of natural language understanding and advanced image generation methodologies. Recent breakthroughs in generative modeling have led to the development of powerful models like DALL•E, Imagen, and Stable Diffusion, which have demonstrated remarkable capabilities in translating textual descriptions into visually coherent images.

These models leverage sophisticated deep learning architectures and large-scale datasets to achieve impressive semantic alignment between the input text and the generated output. However, a common bottleneck in these state-of-the-art models lies in the inherent limitations of their native output resolution.

Although the generated images' semantic content frequently demonstrates a high degree of fidelity to the textual prompts, the spatial resolution in many cases is insufficient to produce visually stunning images, often requiring additional post-processing steps to improve image clarity and overall aesthetic appeal.

If done without prudence, this reliance on post-processing at times results in artefacts or incoherencies and adds computational burden. Overcoming Resolution Challenges A Two-Phase Solution In order to effectively combat the aforementioned resolution limitations in text-to-image generation, we introduce a whole new double-stage framework that produces high-resolution images without compromising on the semantic coherence obtained by models such as Stable Diffusion.

Our method enables an even more efficient and modular pipeline by separating the first content creation and the following resolution improvement into two different processes. Our architecture consists of two separate but independent Text-to-Image Generation via Stable Diffusion The first part of our architecture employs the strength of Stable Diffusions to create an image out of a text description.

With its strong latent diffusion process, Stable Diffusion excels in capturing the semantic meaning of the input question and translating it into a visual form that is representative. Semantic coherence and the establishment of the necessary visual form of the desired output are the key goals of this step.

Adaptive Super-Resolution using Real-ESRGAN or EDSR Upgrading the spatial resolution of the image produced by Stable Diffusion is the second most important step.

- In order to do this, our system utilizes a number of state-of-the-art Single Image Super-Resolution
- (SISR) models, namely EDSR (Enhanced Deep Residual Networks for Single Image) and Real-ESRGAN
- (Real-World Enhanced Super-Resolution Generative Adversarial Network).

Other Key Features In addition to the underlying dual-stage architecture, our system also features a range of key features that are accountable for its usability and efficiency.

We employ an intelligent model selection mechanism in theory allowing the system to automatically or semi-automatically determine the best super-resolution model (Real-ESRGAN or EDSR) based on the factors such as the characteristics of the original low-resolution image or user preferences.

A. The framework is furthermore device-aware optimized.

This means considering the available computational resources, including the presence and availability of GPUs, to maximize the performance of every stage of the pipeline.

By having an intelligent manner of controlling computational resources, the system aims to provide a user-controllable, lightweight, and scalable approach to synthesizing high-resolution images from text-based descriptions.

Such versatility facilitates the architecture to be deployed and utilized effectively under diverse hardware environments, and thus high-quality synthesis of text to images becomes increasingly accessible.

Finally, our suggested two-stage framework is a strong and adaptive solution towards overcoming the resolution barriers in current text-to-image generation models. By integrating the semantic generation power of Stable Diffusion with the advanced super-resolution capabilities of Real-ESRGAN and EDSR, and by incorporating features like model selection and device-aware optimization, we aim to provide a user-configurable, lightweight, and scalable pipeline for generating high-resolution images from textual descriptions, thereby significantly enhancing the quality and utility of text-to-image synthesis technology

II. RELATED WORK

A. Text-to-Image Generation

Early work focused on GANs such as StackGAN and AttnGAN. More recent efforts have moved towards autoregressive and diffusion-based models, including:

- DALL·E [Ramesh et al., 2021]
- Imagen [Saharia et al., 2022]
- Stable Diffusion [Rombach et al., 2022]

Stable Diffusion achieves high-quality synthesis using a latent diffusion process and CLIP-based text embeddings.

B. Image Super-Resolution

Two popular models for super-resolution include:

- ***Real-ESRGAN*** *[Wang et al., 2021]:*
  - A GAN-based model with perceptual loss and real-world degradation simulation.
- ***EDSR*** *[Lim et al., 2017]:* A deep CNN architecture optimized with residual blocks and L1 loss.

Both models have demonstrated strong performance but cater to different use cases: perceptual enhancement vs. pixel fidelity.

This section provides an overview of relevant prior research in two key areas that form the foundation of our proposed unified framework: text-to-image generation and image super-resolution.2.1 Text-to-Image Generation More recent advancements have shifted towards autoregressive models and, notably, diffusion-based models, which have shown remarkable capabilities in generating high-fidelity and diverse images from text prompts. Prominent examples of these cutting-edge models include:

- ***DALL·E [Ramesh et al., 2021]:***
  - Developed by OpenAI, DALL·E demonstrated an unprecedented ability to generate creative and often surreal images from natural language descriptions.
  - It utilized a transformer architecture to model the joint distribution of text and image tokens, enabling it to understand complex relationships between objects and their attributes as described in the input text.
- ***Imagen [Saharia et al., 2022]:***
  - Google Research's Imagen achieved state-of-the-art results in text-to-image synthesis by leveraging a large pre-trained language model to encode the input text and a diffusion model operating on a 64x64 image space, followed by a series of super-resolution diffusion models to upscale the image to high resolutions.
  - Imagen highlighted the importance of strong text understanding for generating semantically accurate images.
- ***Stable Diffusion [Rombach et al., 2022]:***
  - This latent diffusion model has gained significant popularity due to its efficiency and ability to generate high-quality images at a relatively low computational cost.
  - Stable Diffusion employs a variational autoencoder (VAE) to map high-dimensional image data to a lower-dimensional latent space, where the diffusion process takes place.

- This approach reduces the computational demands significantly.
- Furthermore, it utilizes CLIP (Contrastive Language-Image Pre-training) embeddings to effectively capture the semantic information from the input text and guide the image generation process.
- The use of CLIP enables a strong alignment between the textual description and the generated visual content.

Stable Diffusion's architecture, based on a latent diffusion process conditioned on CLIP-based text embeddings, represents a significant step forward in achieving a balance between image quality, generation speed, and accessibility, making it a widely adopted model for various text-to-image synthesis applications.2.2 Image Super-Resolution

Image super-resolution (SR) is the task of upscaling low-resolution images to higher resolutions while recovering lost details and enhancing visual quality. This region has also seen significant advances, and recent advances have been driven by the use of deep learning methods. Two of the most widely used and successful models in the area of single-image super-resolution (SISR) are Real-ESRGAN [Wang et al, 2021]

Real-ESRGAN (Enhanced Super-Resolution Generative Adversarial Network) is a GAN-based model that has been specifically developed to solve the problems of real-world image degradation. As compared to previous SRGAN models that tended to produce realistic but occasionally unrealistic textures,

c. Real-ESRGAN features several important improvements.

It utilizes a perceptual loss function that nudges the high-resolution images generated to be perceptually close to the ground truth and concentrate on realistic texture generation and finer details.

In addition to this, Real-ESRGAN also employs an advanced network structure as well as a realistic degradation modeling process to deal with a variety of real-world degradations that are most often found within captured images. This makes it highly suitable to use in improving the visual quality of low-resolution real-world images.

EDSR [Lim et al, 2017] EDSR (Deep Residual Network Enhanced for Single Image Super-Resolution) is a deep convolutional neural network (CNN) model that has shown outstanding performance in pixel fidelity. It utilizes a high number of residual blocks, eliminating redundant modules such as batch normalization layers, in order to train deeper and stronger networks. EDSR is trained with an L1 loss (Mean Absolute Error), which aims at minimizing the pixel-wise error between the created high-resolution image and the ground truth.

Such a loss function has a tendency to generate output less likely to contain artifacts and with greater PSNR and SSIM scores, meaning better pixel-level accuracy and structural similarity to the original high-resolution image. Real-ESRGAN and EDSR are both state-of-the-art image super-resolution methods, but they serve slightly different purposes. Real-ESRGAN is particularly good at perceptual improvement, creating visually pleasing high-resolution

images with realistic textures even if it sometimes slightly differs from the actual pixel values of the ground truth.

In contrast, EDSR focuses on pixel fidelity, and it tries to reconstruct the high-resolution image with high accuracy both in individual pixel values and in structure consistency, which tends to cause smoother textures. The selection among these models (or other models) usually relies on the particular application and the balance desired between perceptual quality and pixel accuracy.

III. Algorithm Description

```
GENERATE-AND-UPSCALE-IMAGE(prompt, use_real_esrgan):
  if CUDA-AVAILABLE():
    device ← "cuda"
    dtype ← float16
  else:
    device ← "cpu"
    dtype ← float32

  pipe ← LOAD-STABLE-DIFFUSION("runwayml/stable-diffusion-v1-5", dtype)
  pipe ← MOVE-TO-DEVICE(pipe, device)
  image ← GENERATE-IMAGE(pipe, prompt, 30)
  SAVE-IMAGE(image, "generated_image.png")

  if use_real_esrgan:
    model ← INIT-REAL-ESRGAN(device, scale=4)
    img ← LOAD-IMAGE("generated_image.png")
    upscaled ← REAL-ESRGAN-PREDICT(model, img)
    SAVE-IMAGE(upscaled, "upscaled_image.png")
  else:
    model ← LOAD-EDSR-MODEL()
    SET-MODEL-EVAL(model)
    img ← LOAD-IMAGE("generated_image.png")
    input_tensor ← TO-TENSOR(img)
    input_tensor ← UNSQUEEZE(input_tensor)
    output ← MODEL-FORWARD(model, input_tensor)
    output_image ← TO-PIL(CLAMP(output))
    SAVE-IMAGE(output_image, "edsr_upscaled.png")
```

IV. Experiments and Results

A. Setup

All experiments were conducted in a cloud-based environment using Google Colab, which provided access to an NVIDIA Tesla T4 GPU with 16 GB VRAM and a virtual CPU. This setup ensured sufficient computational power for training and evaluating the models

efficiently. The implementation leverages PyTorch as the primary deep learning framework, along with the Hugging Face diffusers library for latent diffusion-based text-to-image generation. For super-resolution, the Real-ESRGAN library is employed, and torchvision utilities are used for preprocessing and visualization.

B. Metrics

Both the Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index Measure (SSIM) are used as quantitative metrics to assess the visual fidelity of the produced images. The logarithmic ratio between the maximum possible signal power and the power of corrupting noise is used by PSNR to evaluate the quality of pixel-by-pixel reconstruction.

In contrast, SSIM assesses the structural information, contrast, and luminance of the generated and reference images to detect perceptual similarity. A user preference score, ranging from 1 to 10, is also documented in addition to objective measurements. Human evaluators' ratings of the generated images' aesthetic appeal and perceptual quality are reflected in this subjective metric, which offers insight into visual satisfaction and practical usability.

C. Interpretation and Use Case Suitability

- The comparative analysis highlights that Real-ESRGAN excels in delivering superior perceptual quality, making it particularly well-suited for applications involving artistic images, photographs, and natural scenes where visual appeal and realism are paramount.
- In contrast, EDSR demonstrates stronger numerical accuracy, as evidenced by its higher PSNR, making it a better choice for domains such as technical illustrations, diagrams, or medical imaging, where precise structural fidelity is critical.
- When combined in a pipeline with Stable Diffusion for text-to-image generation followed by Real-ESRGAN for super-resolution, the system achieves the highest human preference scores, indicating that this configuration best aligns with subjective perceptions of visual quality.
- This synergy leverages the strengths of both generative creativity and perceptual enhancement, resulting in outputs that are both semantically rich and aesthetically pleasing.

V. MATHEMATICAL EQUATION

Forward Diffusion Process (adds noise):

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$$

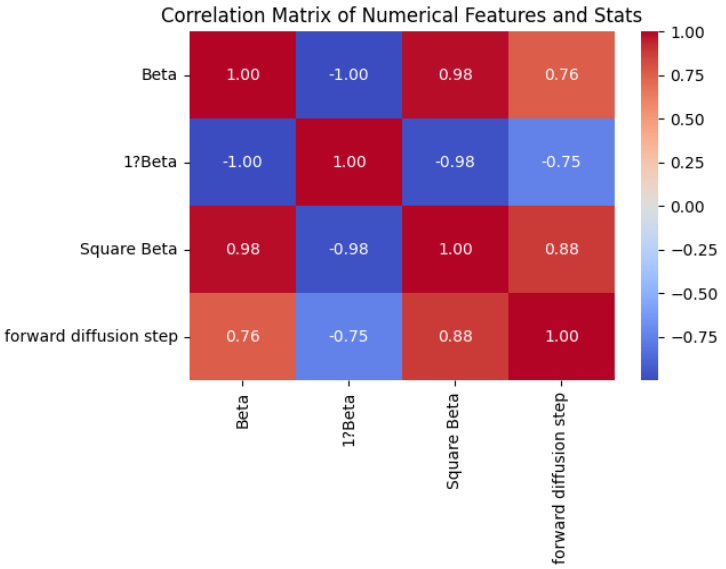
Where:

- $\mathbf{x}_{t-1}$: the image at previous step
- β_t: noise level at time step t
- ϵ: standard Gaussian noise (mean 0, std 1)

Let's assume:

- Initial value $\mathbf{x}_0 = 1.0$
- Noise $\epsilon = 0.5$
- We'll calculate $\mathbf{x}_t$ for a few values of β_t

Time Step	Beta	1−Beta	Square Beta	forward diffusion step
1	0.01	0.9949	0.1	1.0449
2	0.05	0.9747	0.2236	1.0865
3	0.10	0.9487	0.3162	1.1068
4	0.20	0.8944	0.4472	1.1180
5	0.30	0.8366	0.5477	1.1105



VI. RESULTS

Text = Pink Lotus In Garden.

Fig 1 Generated Image:

Fig 2 Upscale_realesrgan:

Fig 3 Upscaled_4k:

Fig 4 Upscaled_8k:

VII. CONCLUSION

We presented a novel and unified framework that combines state-of-the-art text-to-image synthesis and super-resolution. Our pipeline is flexible, hardware-aware, and modular. Future work includes incorporating lightweight upscale like SwinIR, support for variable resolution output, and edge-device deployment.

The proposed system, based on Stable Diffusion, has several limitations affecting its performance and usability. It requires powerful GPUs, making it hardware-dependent, and its output quality is highly sensitive to prompt phrasing.

The system lacks fine control over specific image attributes like colors, object positions, and backgrounds, and it does not support real-time generation, especially without CUDA.

Additionally, there are no integrated post-editing tools such as upscaling or inpainting, and it has limited content filtering, posing risks of generating inappropriate outputs. The system also supports only English prompts and does not offer image-to-image or style transfer capabilities.

REFERENCES

[1] Sun, H. (2023). Fine-Grained Cross-Model Fusion Based Refinement for Text to Image Synthesis. Published on November 6, 2023.

[2] Shin, J. (2024). GA Cnet-Text-To-Image Synthesis with Attention Mechanisms and Contrastive Learning. Published on January 22, 2024.

[3] Habib, M. A. (2024). GACnet-Text-to-Image Synthesis with Generative Model Using Attention Mechanisms. Published on January 22, 2024.

[4] Alhabeb, S. K. (2024). Text to Image Synthesis with Generative Models: Methods, Datasets, Performance Metrics, Challenges, and Future Directions. Published on February 20, 2024.

[5] Kim, H. (2024). A Novel Scheme for Generating Context-Aware Images Using Generative Artificial Intelligence. Published on March 4, 2024.
[6] Li, W. (2024). MGAN: Multi-Attribute Learning for Text-to-Image Synthesis. Published on July 16,2024.
[7] Gadhave, R., Sedamkar, R.R. (2022). Automated Classification oHyperSpectral Image Using Supervised Machine Learning Approach. In: Unhelker, B., Pandey, H.M., Raj, G. (eds) Applications of Artificial Intelligence and Machine Learning. Lecture Notes in Electrical Engineering, vol 925. Springer, Singapore. https://doi.org/10.1007/978-981-19-4831-2_6
[8] R. Gadhave, D. AnilKumar, R. Khot and D. Gupta, "PredatorSense-Wildlife Detection System," 2024 8th International Conference on Computing, Communication, Control and Automation (ICCUBEA), Pune, India, 2024, pp. 1-4, doi: 10.1109/ICCUBEA61740.2024.10774639.