International Journal of Advances in
Soft Computing and Intelligent Systems (IJASCIS)
2024, Vol 03, Issue 02, 266-278
ISSN: 3048-4987

Science Transactions © 2024

Original Paper

# NOVEL TO ANIMATION: A LIGHT NOVEL BASED PHOTOGRAPHIC PROJECT

Dr. Deepali M Bongulwar[a*], Harshraj Ushire[b], Chinmay Sonsurkar[c], Sachin Naik[d], Yash Bagul[e]

[a] Dept. of CSE-AI, Brainware University, Kolkata, West Bengal, India,700125, deepalibongulwar@gmail.com

[b,c,d,e] Dept. of CSE-AI, PCET's and NMVPM's Nutan College of Engineering and Research, Pune, Maharashtra, India – 410507, harshrajushire4423@gmail.com , chinmaysonsurkar31@gmail.com, sachinnaik4292@gmail.com, yash07bagul@gmail.com

## ABSTRACT

The development of innovative technologies in a variety of industries has been greatly aided by artificial intelligence (AI). The anime industry is one such area where AI- powered solutions can be applied to produce animation in an organized and effective manner. In this study, we demonstrate a system that generates anime sequences based on text descriptions taken from relevant light novels by using the Stable Diffusion AI model. A lot of obstacles confront anime artists, including fitting in with weekly schedules, which can lead to health hazards and lack of sleep. With the use of AI-based image generation, our suggested approach will provide anime artists with a more accurate and efficient way to produce anime graphics from text descriptions. The user interface, database, picture production, and input processing are the four modules that make up the system. After receiving the text description of the intended scene, the input processing module preprocesses it to remove any characters that aren't needed. The Stable Diffusion AI model is used by the picture production module to produce a latent vector representation of the input text, which is subsequently decoded to produce an image. The created anime scene can be seen and text descriptions can be entered using an easy-to-use interface provided by the user interface module. The created anime images and the text descriptions that go with them are managed and stored by the database module. In comparison to current technologies, our suggested system achieves a higher degree of accuracy and efficiency. Using input scene descriptions and the Stable Diffusion AI model, our system can generate accurate and lifelike anime visuals from the corresponding light novel, saving time and increasing productivity.

## KEYWORDS

*Stable Diffusion, Variational Autoencoding, GAN (Generative Adversarial Network, Text-to-Image*

## 1. INTRODUCTION

Texts can successfully convey a variety of visual media, including pictures, paintings, and photos. But making them takes a great deal of effort and specialized knowledge. Therefore, a tool that can produce lifelike visuals from textual input might significantly improve people's capacity to produce complex visual content with ease. Furthermore, exact corrections and incremental improvements are made possible by the modify photos with natural language, which is essential for real-world applications. Using which were first created for density estimation problems, are now well-known for their ability to generate realistic features and finely detailed images. These models offer a viable foundation for producing visually appealing images from verbal descriptions by utilizing the principles of diffusion processes. The purpose of this research is to evaluate the efficacy of stable diffusion models in producing realistic and varied visual material when used for text-to-image generation in light novels. The model creates new scenes based on the light novel chapters by importing anime sequences and captions.

The structure of the paper is as follows: In Section 2, important approaches and techniques are highlighted as prior works on text-to-image generation are reviewed and discussed. Our method using the stable diffusion v1.5 model is described in Section 3. The results of our experiment are provided in Section 4. Section 5 wraps up the study's course and provides an overview of upcoming research projects

## 2. LITERATURE REVIEW

As this domain expands, there are numerous studies on text-to-image generation, and there are numerous implementations involving both generative models and diffusion models aimed at producing high quality image. The following are some of the papers thatwere examined in order to comprehend the text-to-image generating technique.

### 1.1 Image Labeling

By using publicly available image captioning datasets, Tao et al. [1]; Zhang et al. [2]; Ye et al. [4] GAN training with text conditioning. In Ramesh et al. [3], images are synthesized based on text using a generative autoregressive model trained with discrete latent coding building on van den Oord et al. [5].

### 1.2 Stable Diffusion Model

Meng et al. [6] found that diffusion models are capable of both inpainting specific capacity to areas of an image and doing so while accounting for a rough sketch or color set that represents the image. Similarly, Saharia et al. [7] observed that when diffusion models are directly trained for inpainting tasks, they can seamlessly fill in missing parts of an image without introducing edge artifacts. An approach that Guojun et al. [8] proposes is the use of word-level conditional batch normalization and dual encoders with triplet loss in order to improve the alignment of text and image.

### 1.3 Contrastive Language-Image Pre-Training (Clip)

Image generation has previously been guided by CLIP. CLIP is used by Galatolo et al [9], Patashnik et al. [10], and Gal et al. [11] to direct GAN production in the directionof particular text prompts. In order to alter photos, Kim & Ye [12] use text prompts toadjust a diffusion model, aiming for a CLIP loss while recreating the DDIM latent of the original image, as suggested by Song et al. [13]. Furthermore, GAN models conditioned on perturbed CLIP image embeddings are trained by Zhou et al. [14], producing a model that can condition pictures on CLIP text embeddings.

### 1.4 SDXL: Improving Latent Diffusion Models For High-Resolution Image Synthesis

SDXL, a latent diffusion model developed by Dustin Podell et al.[19] The study presents SDXL, a sophisticated latent diffusion model intended for text-to-image conversion. Compared to earlier iterations, SDXL has a larger U-Net backbone with more attention blocks and cross-attention context via an extra text encoder. The goal of these upgrades is to improve performance. The study trains SDXL on many aspect ratios and explores various ways of conditioning. Moreover, image-to-image approaches are employed to enhance visual quality through the implementation of a refinement model. The results demonstrate that SDXL performs better than earlier Stable Diffusion models and holds its own against state-of-the-art image generators. To encourage transparency in the training and assessment of large models, the study places a strong emphasis on open research techniques and transparency.

### 1.5 Identified Gaps

We noticed that existing models pay little attention to particular genres or styles, like anime, and instead concentrate mainly on generic picture synthesis. This limits their ability to adjust to other artistic approaches, resulting in a gap in the specificity and versatility of the model. Moreover, existing research frequently employs datasets with a lack of diversity in text and visual styles, which can bias results toward recurring themes and possibly lower model efficacy for challenging prompts. To improve generalization skills, we use a more diverse dataset in our research to address this. In addition, we have noticed that there hasn't been much talk about the moral and legal ramifications of utilizing a variety of image and text sources. To address this, we make sure that our methodology complies with copyright laws. Additionally, the relevance cannot be effectively assessed by current evaluation methodologies. Linking generated visuals to textual cues, which has led us to develop more sophisticated assessment measures. Furthermore, we observe the underutilization of post-processing methods that could enhance image fidelity. Our study intends to address this shortcoming by integrating a unique image-to-image refinement model and establishing new benchmarks for image quality in text-to-image synthesis.

## 3. PROBLEM DEFINITION

Multimedia content creation—where textual material is quickly merged with visual media—has significantly increased in the modern digital landscape. Creating anime-style graphics straight from text descriptions is an interesting area of study in this subject. The goal of this project is to discuss the opportunities and problems related to text-to-image (anime) synthesis. Creating algorithms and models that can comprehend the subtleties and semantics found in written descriptions and convert them into visually appealing anime-style pictures is a major goal of this project. Making sure the created images faithfully capture the ideas, feelings, and minute details described in the text is the main goal. A crucial component of this project is gathering a varied dataset that contains textual descriptions and matching visuals in the manner of anime. Obtaining and annotating such material requires considerable attention to ethical and copyright consideration.

In order to improve the calibre and complexity of the created anime-style pictures, the research also investigates the integration of many modalities, including text, image, and maybe music. The goal is to build multimodal models that can efficiently integrate data from multiple sources to produce visually compelling stories that make sense in their context. Apart from picture synthesis, the initiative tackles issues about personalization and uniformity of style in anime art. It looks into ways to generate interactive and real-time images while considering ethical issues and biases present in the data and algorithms used. In addition, the initiative seeks to develop strong assessment metrics to thoroughly evaluate the generated pictures' accuracy and quality in relation to the original written descriptions. Additionally, great consideration is paid to accessibility, user experience, and the smooth integration of Text-to-Image (Anime) synthesis tools into current processes. In the end, this initiative aims to offer creative approaches and solutions that let producers easily turn written stories into visually appealing anime. The project is committed to addressing the ethical and societal obligations connected with the use of this transformative technology, in addition to its scientific breakthroughs.

## 4. METHODOLOGY

### 4.1 Research Design

Our Text-to-Image (Anime) synthesis project is based on a multidisciplinary methodology that combines knowledge of deep learning, computer vision, natural language processing (NLP), and user-centered design. Our project has a well-organized workflow that starts with data collection. We ensure quality and legality by sourcing a variety of textual descriptions and anime-style visuals. To build a standardized dataset, preprocessing entails text tokenization, image scaling, and careful annotation. Semantic interpretation and understanding are central to our methodology. By utilizing sophisticated natural language processing algorithms, we are able to extract complex semantics from written descriptions, encompassing not just explicit information but also nuances of context and emotions. Following their integration, these text embeddings are combined with picture representations to create a multimodal input for the synthesis model. To create photos in the anime style, we utilize cutting-edge deep learning architectures, possibly using Variational Autoencoders (VAEs) or Generative Adversarial Networks (GANs). Using the multimodal input, our model creates visually appealing anime-style graphics. In order to satisfy user preferences, we also incorporate a style control module that enables authors to adjust the visual style—choosing from chibi, shonen, or shojo, among other options to perfection. A specialized module enables real-time and interactive generation capabilities, meeting the needs of applications where the creation of dynamic content is essential. In order to impact the creation process, users can offer real-time comments, which encourages user involvement and creativity.

## 4.2 LDM using Stable Diffusion

Our proposed system utilizes the Stable Diffusion Model, which consists of three key components: an Auto encoder, a U-Net, and a text encoder. Auto encoder (VAE): An encoder and a decoder make up the VAE model. A 512x512x3 image is transformed by the encoder during the latent diffusion training phase into a lower-dimensional latent representation, which is typically sized at 64x64x4 for the forward diffusion process. These encoded representations, known as latents, experience gradual noise addition at each training step. These latent representations act as inputs for the U-Net model. Transforming an image from dimensions (3, 512, 512) to a latent representation of dimensions (4, 64, 64) results in a substantial reduction in memory consumption by a factor of 48. This reduction in memory and computational requirements enables the rapid generation of 512 × 512 images on 16GB Colab GPUs. The VAE decoder converts the denoised latent representations produced during the reverse diffusion process back into pictures. To convert the denoised image into the real image during the inference stage, just the VAE decoder is required.

## 4.3 U-Net Architecture

U-Net: To forecast denoised image representation from noisy latents, the U-Net architecture is applied. The U-net receives the noisy latents as input and emits noise into the latents. We extract the true latent representation by deducting this noise from the noisy latents. Furthermore, a conditional model is used for guidance, taking the timestep and text embedding into account. Figure 3.1 presents the architecture of U- Net

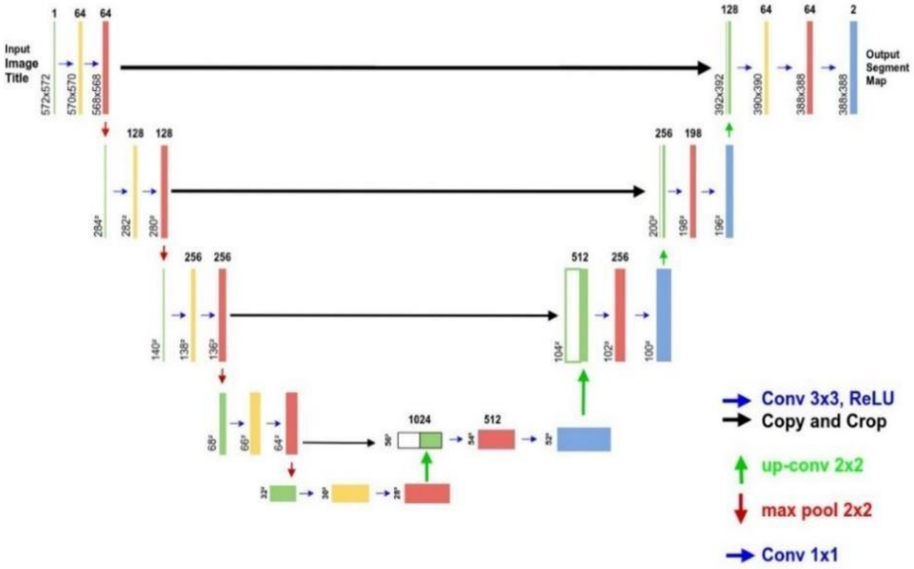


Figure 1. U-Net Architecture

The model design consists of a U-Net with a center block, 12 block encoders, and 12 block decoders connected by skip links. Of these twenty-five blocks, eight are used for convolution layers to do down- or up-sampling, while the other seventeen are main blocks that each contain two Vision Transformers (ViTs) and four ResNet layers. In order to reduce noise, the encoder compresses the image representation into a lower-resolution form, while the decoder reconstructs it into the original, higher-resolution image.

### 4.4 Text Encoder

The input prompt is transformed into an embedding space by the text encoder and sent into the U-Net. This is used as a guide for noisy latents while Unet is being trained for denoising. A simple transformer-based encoder is usually used for text encoding, which converts a sequence of input tokens into a sequence of latent text embeddings. Since Stable Diffusion makes use of CLIP, an already-existing text encoder, no new one needs to be trained. Text that matches the supplied text is generated by the text encoder.

### 4.5 Data Collection

Gathering an extensive and varied collection of anime-style photos from various sources and matching them with written explanations was the main objective. The development and training of machine learning models for text-to-image synthesis will be based on this dataset. carefully choosing resources that offer a variety of anime topics and styles, so as to guarantee a thorough portrayal. Online image banks, fan art sites, anime databases, and digital libraries that offered anime content for legal download were some of the sources used. Following to improve the model's capacity to generalize across various anime styles and descriptive complexity, make sure the dataset contains a wide range of characters, locales, and scenarios. When using data, make sure you have the required permissions. Only use data sources that offer

### 4.6 Data Pre-processing

We divide textual descriptions into distinct, manageable tokens using text tokenization, which we accomplish utilizing cutting-edge frameworks for natural language processing. To guarantee consistency throughout the dataset, this procedure entails removing punctuation and special characters and then changing all text to lowercase. We also remove stop words so that we may concentrate on the text's most informative sections. Using picture scaling algorithms, we bring all the photographs in the dataset's dimensions into uniformity. This entails adjusting images to a fixed resolution using methods like bilinear and bicubic interpolation, while maintaining the original aspect ratios of the images by suitable padding. This method preserves consistency and quality of images, which is important for training models. We add "Borutags"—metadata that contains essential details about every image—to our dataset to make it richer genre and feelings of the characters. We create these tags using semi-automated procedures or hand tagging, and we save the results in a manner separated by commas. The model can use these descriptive tags to improve specificity and accuracy in image generation from textual descriptions during training because this metadata is included into our database. Every one of these procedures is carried out with great care in order to guarantee that our data is complete, accurate, and consistent, which will provide a strong basis for our further study and model building.

### 4.7 Implementation

The first step in the training process is to compile the dataset. The next step is to label the photos. Following proper labeling, the images are fed into the latent diffusion model, which is trained using the predetermined configuration. Figure 3.2 shows our system's overall processing pipeline.

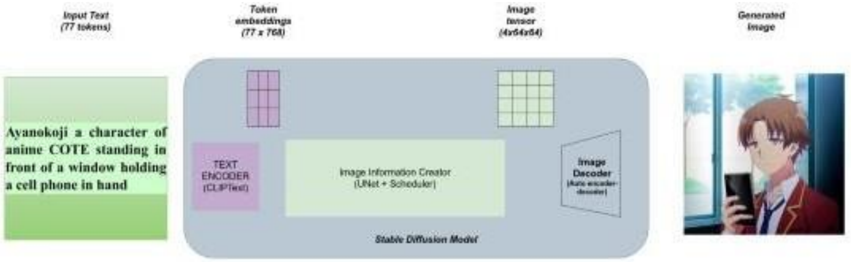


Figure 2. Overall system pipeline

## 5. RESULT AND DISCUSSION

In this part, we showcase and assess the outcomes achieved.

### 5.1 Setting Up the Environment

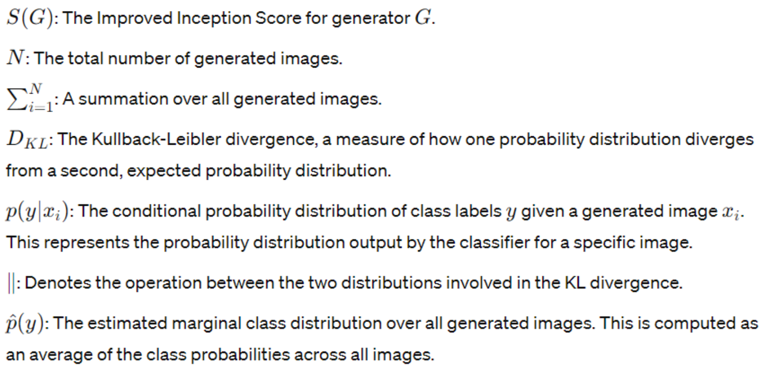
For the project execution, Python 3.8 was utilized, and Google Colab was employed to train the LDM model. The training of the model with our dataset on Google Colab lasted for 20 hours. In training the diffusion model, a personal computer with a Nivida 3060 GPU was utilized. The additional crucial parameters for LDM can be found in Table 1

Table 1. LDM parameters and values

| Sr. No | Hyper Parameter | Value |
|---|---|---|
| 1 | Learning Rate | 1e-6 |
| 2 | Conditional Dropout | 0.10 |
| 3 | Clip Skip | 1 |
| 4 | Seed | -1 |

Formula for Inception Score: Inception score is the average of the divergences between the conditional distribution and the marginal distribution for every image in the dataset. The score becomes more stable and interpretable in terms of mutual information when the exponential function is removed and p^(y) is computed across the entire dataset instead of in batches. This reflects the decrease in uncertainty regarding an image's class when it is generated by the generator G.

$$\mathrm{S}(G) = \frac{1}{N}\sum_{i=1}^{N} D_{KL}(p(y|\mathbf{x}^{(i)} \parallel \hat{p}(y))$$



$S(G)$: The Improved Inception Score for generator $G$.

$N$: The total number of generated images.

$\sum_{i=1}^{N}$: A summation over all generated images.

$D_{KL}$: The Kullback-Leibler divergence, a measure of how one probability distribution diverges from a second, expected probability distribution.

$p(y|x_i)$: The conditional probability distribution of class labels $y$ given a generated image $x_i$. This represents the probability distribution output by the classifier for a specific image.

$\|$: Denotes the operation between the two distributions involved in the KL divergence.

$\hat{p}(y)$: The estimated marginal class distribution over all generated images. This is computed as an average of the class probabilities across all images.

The outcome of our model in terms of the state of the art's inception score is displayed in Table 2. The suggested model is found to operate well. Table 3 displays the final results.

Table 2. Comparison between State-of-Art and proposed model

| **Reference** | **Model** | **Inception Score** |
|---|---|---|
| Reed et al. [15] | GAN-INT-CLS | 2.67 ± 0.02 |
| Zhang et al.[16] | StackGAN++ | 3.25 ± 0.02 |
| Zhang et al. [17] | HDGAN | 3.47 ± 0.06 |
| Cai et al. [18] | DualAttn-GAN | 4.04 ± 0.01 |
| Proposed Method | LDM using Stable Diffusion v1.5 | 5.6 ± 0.07 |

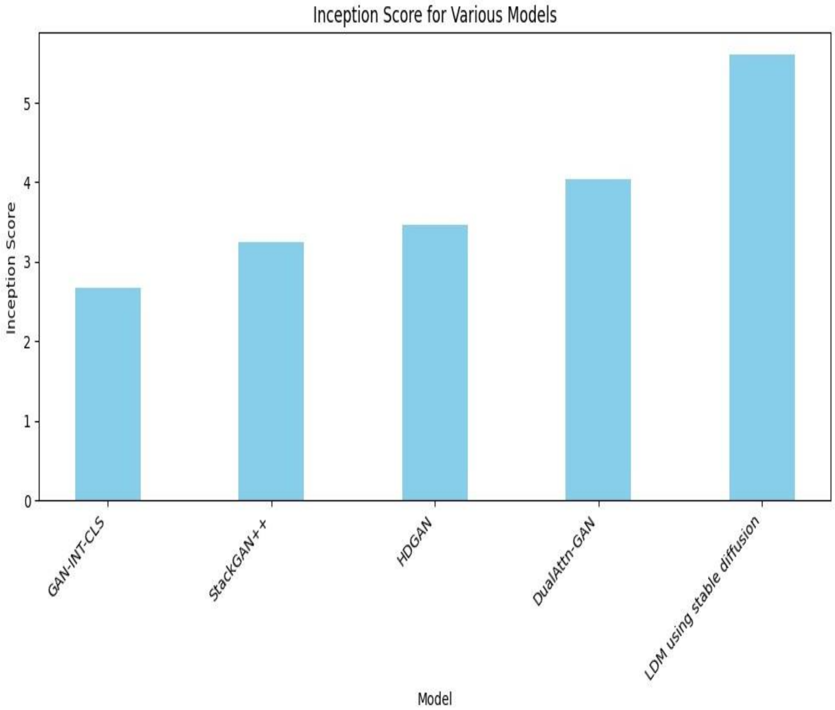
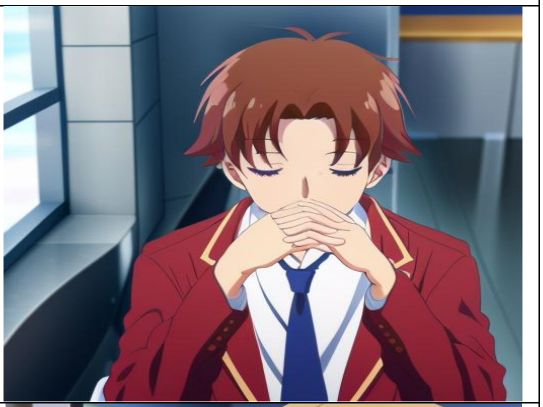
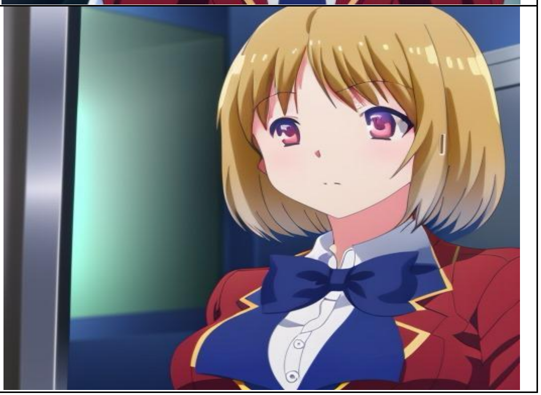
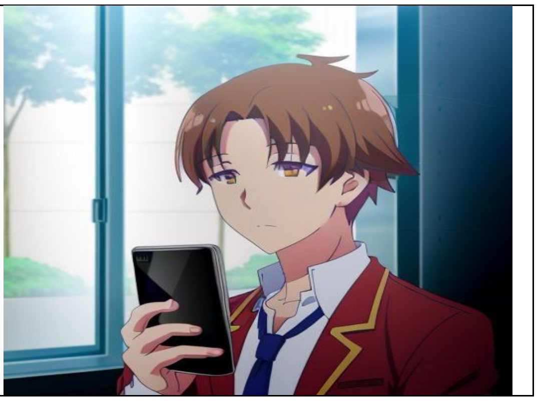


Figure 3. Bar chart of Inception Score

Table 3. Final Result

| **Text Input** | **Output** |
|---|---|
| Ayanokoji in red jacket and a tie sitting in front of a window with his hands folded in front of his face. |  |
| Kushida with blonde hair wearing a red jacket and a blue bow tie is standing in front of a mirror. |  |

| | |
|---|---|
| Ayanokoji standing in front of a window holding a cell phone in one hand with a tree in the background. |  |

## 6. CONCLUSION

The study focuses on the application of latent diffusion models (LDMs) to enhance the efficiency of training and sampling in de-noising diffusion models while maintaining high image quality standards. Through the integration of a tailored cross-attention conditioning mechanism tailored to particular tasks, the suggested approach shows promise in surpassing current methods across a range of conditional picture synthesis challenges.

Compared to Generative Adversarial Networks (GANs), LDMs may sample sequentially a little more slowly, but they still have several computational benefits. These benefits, along with the technique's capacity to preserve image quality, make it a formidable rival in the conditional image synthesis space. But it's crucial to be aware of any potential drawbacks, particularly when working on jobs like background refining that need for exact pixel accuracy. The writers hope that additional improvements will assist in overcoming these restrictions and greatly enhancing background refinement outcomes with their super-resolution models. The objective is to improve coherence in synthesized visuals and human perception of image quality by fine-tuning the alignment between textual descriptions and generated images.

This study provides guidance for future developments in text-to-image generation by emphasizing the possibility of boosting background features to match textual descriptions smoothly, which would improve overall image quality and increase viewers' immersive experience. The will to push the limits of artificial creativity drives the research team's unwavering commitment to excellence in text-guided image synthesis.

## 7. FUTURE SCOPE

Enhanced Realism: Produce images that are very similar to real photographs by painstakingly refining minute details, textures, and subtle aspects. This will raise the realism of computer-generated images to previously unheard-of levels. Leveraging generative model improvements, including the incorporation of state-of-the-art approaches like higher-resolution methodology and more sophisticated architectural designs, can help attain this ambitious aim of heightened realism. The goal is to close the gap between artificial and genuine pictures by pushing the boundaries of realism and bringing in a new era of visually compelling image synthesis.

Multimodal Comprehension: Advances in creating intricate models that can process information from several modalities, such as audio or extra contextual cues, in addition to textual descriptions. This all-encompassing method of picture creation has the potential to improve content and context comprehension, resulting in more comprehensive and contextually aware image creation. The aim of text-guided image synthesis is to achieve new expressive and creative levels by embracing the synergistic integration of many modalities.

Precise Control: By implementing systems that enable the specification of precise features, styles, or elements within textual inputs, you can offer users hitherto unheard-of levels of control over created images. This move toward fine control makes it possible to create an extremely customized and individualized image synthesis experience that meets each user's unique preferences and artistic aspirations. The goal is to enable users to realize their artistic dreams with remarkable accuracy and authenticity by giving them creative autonomy.

Interactive Generation: Rethink the way that images are created by providing interactive and iterative workflows that allow users to take an active role in the model, providing input on preliminary findings and directing further iterations toward more satisfying results. Human intuition and machine intelligence come together to co-create visually captivating pictures in this symbiotic relationship that is fostered by the dynamic and cooperative interaction between users and the model. The objective is to democratize the picture synthesis process by embracing the concepts of interactivity and adaptability and enabling people to participate in a collaborative process of creation and discovery.

Experiment with new areas of cross-domain image generation, where textual descriptions are translated into various artistic styles, historical eras, or combined elements from many genres, in order to broaden the scope of text-to-image synthesis. This bold project breaks through conventional barriers to provide an infinite potential for creativity where the mind is free. The goal is to inspire new forms of visual storytelling and expression while enhancing the creative environment through the exploration of several artistic fields.

## REFERENCES

[1] Tao, M., Tang, H., Wu, F., Jing, X.Y., Bao, B.K. and Xu, C., 2022. Df-gan: A simple and effective baseline for text-to-image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 16515-16525).

[2] Zhang, H., Koh, J.Y., Baldridge, J., Lee, H. and Yang, Y., 2021. Cross-modal contrastive learning for text- to-image generation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 833-842).

[3] Ra mesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M. and Sutskever, I., 2021, July. Zero-shot text-to-image generation. In International Conference on Machine Learning (pp. 8821-8831). PMLR.

[4] Ye, H., Yang, X., Takac, M., Sunderraman, R. and Ji, S., 2021. Improving text-to-image synthesis using contrastive learning. arXiv preprint arXiv:2107.02423.

[5] Oord, A.V.D., Vinyals, O. and Kavukcuoglu, K., 2017. Neural discrete representation learning. arXiv preprint arXiv:1711.00937.

[6] Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y. and Ermon, S., 2021. Sdedit: Guided image

synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073.

[7] Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D. and Norouzi, M., 2022, July. Palette: Image-to-image diffusion models. In ACM SIGGRAPH 2022 Conference Proceedings (pp. 1-10).

[8] Yin, G., Liu, B., Sheng, L., Yu, N., Wang, X. and Shao, J., 2019. Semantics disentangling for text-to-image generation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2327- 2336).

[9] Galatolo, F.A., Cimino, M.G. and Vaglini, G., 2021. Generating images from caption and vice versa via clip-guided generative latent space search. arXiv preprint arXiv:2102.01645.

[10] Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D. and Lischinski, D., 2021. Styleclip: Text-driven manipulation of stylegan imagery. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 2085-2094).

[11] Gal, R., Patashnik, O., Maron, H., Chechik, G. and Cohen-Or, D., 2021. Stylegan-nada: Clip-guided domain adaptation of image generators. arXiv preprint arXiv:2108.00946.

[12] Kim, G., Kwon, T. and Ye, J.C., 2022. Diffusionclip: Text-guided diffusion models for robust image manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2426-2435).

[13] Song, J., Meng, C. and Ermon, S., 2020. Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502.

[14] Zhou, Y., Zhang, R., Chen, C., Li, C., Tensmeyer, C., Yu, T., Gu, J., Xu, J. and Sun, T., 2022. Towards language-free training for text-to-image generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 17907-17917).

[15] Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B. and Lee, H., 2016, June. Generative adversarial text to image synthesis. In International conference on machine learning (pp. 1060-1069). PMLR.

[16] Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X. and Metaxas, D.N., 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In Proceedings of the IEEE international conference on computer vision (pp. 5907 -5915).

[17] Zhang, Z., Xie, Y. and Yang, L., 2018. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6199-6208).

[18] Cai, Y., Wang, X., Yu, Z., Li, F., Xu, P., Li, Y. and Li, L., 2019. Dualattn-GAN: Text to image synthesis with dual attentional generative adversarial network. IEEE Access, 7, pp.183706 -183716.

[19] Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J. and Rombach,

[20] R., 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.019

[21] Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M. and Aberman, K., 2023. Dreambooth: Fine

[22] tuning text-to-image diffusion models for subject-driven generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 22500-22510).

[23] Kingma, D.P. and Welling, M., 2013. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

[24] Diederik, P.K., 2014. Adam: A method for stochastic optimization. (No Title).

[25] Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B., 2022. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10684-10695).

## Authors

Dr. Deepali Makarand Bongulwar is presently working as Asst. Prof. at CSE-AI Department of Brainware University, Kolkatta. She has completed her PhD. from Sri Satya Sai University of Technology and Medical sciences, Sehore, Madhya Pradesh, India in August 2023. She has received a B.E. degree in Electronics Engineering and an M.E. degree in Electronics with Specialization in Computer Technology from SGGSCoE & T, Nanded, Maharashtra, India from Dr Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra, India in 1997 and Swami Ramanand Teerth Marathwada University, Nanded, Maharashtra, India in 1999 respectively. She has nearly nine years of teaching experience in Mumbai University, one year in DBATU, Lonere University and two years in other colleges and two years of Industry experience. She is a Life Member of InSc International Publishers (IIP). She is a reviewer of the InSc-International Journal of Electronics, Electrical and Communication Engineering (IJEECE), and the International Journal of Science, Technology and Society (IJSTS). She has reviewed papers from Elsevier’s Journal and Web of Science Journal. As a TPC member of International Conference, she has reviewed many papers. She has received the “Research Excellence Award 2022” by InSc (Institute of Scholars) International Publisher. She has exceptional contributions as a Mentor in the Smart India Hackathon”, 2023. She has published more than six research papers in National/International Journals and conferences, one Book, one Book chapter, one patent and one copyright.

She has demonstrated expertise in a wide area of advanced technological domains, including Deep Learning, Machine Learning, Computer Vision, Image Processing, Cyber Security, and Medical Imaging. Currently focused on pioneering work with Convolutional Neural Networks, Large Language Models (LLMs), Generative Adversarial Network (GAN), text-to-image conversion technologies, vision transformers, and stable diffusion algorithms.

She is passionate about integrating theoretical knowledge with practical applications to foster innovation and drive technological advancement, and committed to mentoring the next generation of engineers and researchers while continually pushing the boundaries of what’s possible in computer science.

Harshraj Ushire is a B-Tech graduate in Computer Science and Engineering with a specialization in Artificial Intelligence from Nutan College of Engineering and Research, Pune. Combining his technical expertise with a passion for storytelling, Ushire has embarked on a groundbreaking project titled "Novel to Animation: A Light Novel-Based Photographic Project." This innovative endeavour merges the narrative depth of light novels with the visual artistry of photography, creating a novel hybrid that bridges literature and digital media. Ushire’s project exemplifies his ability to integrate advanced technological skills with creative expression, offering a fresh perspective on how stories can be represented and experienced. His work stands out for its originality and its exploration of new frontiers in both art and technology.

Chinmay Sonsurkar recently completed a B-Tech degree in Computer Science and Engineering with a focus on Artificial Intelligence from Nutan College of Engineering and Research, Pune. He is the author of the research paper "Novel to Animation: A Light Novel-Based

Photographic Project," which explores the intersection of narrative techniques and visual media in transforming light novels into animated formats. Chinmay’s academic background includes extensive study in both computer science and artificial intelligence, providing a strong foundation for his exploration of innovative approaches to multimedia projects. He is particularly interested in the application of AI technologies in creative fields and how they can enhance storytelling and animation processes. He aims to further contribute to the fields of AI and multimedia through continued research and development.

Sachin Naik is a B-Tech graduate in Computer Science and Engineering with a focus on Artificial Intelligence from Nutan College of Engineering and Research, Pune. His academic background in technology and AI has significantly influenced his creative endeavors. He is the creator of "Novel to Animation: A Light Novel-Based Photographic Project," a pioneering initiative that seamlessly integrates the narrative richness of light novels with the visual storytelling of photography. This project exemplifies Naik's unique ability to blend technical acumen with artistic vision, offering a novel approach to storytelling that bridges literature and visual art. His work is noted for its innovative use of technology and its fresh perspective on narrative expression.

Yash Bagul is a B-Tech graduate in Computer Science and Engineering with a specialization in Artificial Intelligence from Nutan College of Engineering and Research, Pune. Drawing on his technical expertise and creative flair, Bagul has developed a unique project titled "Novel to Animation: A Light Novel-Based Photographic Project." This innovative project fuses the compelling narrative elements of light novels with the artistic dimension of photography, creating a distinctive form of visual storytelling. Bagul’s work is distinguished by its integration of advanced AI techniques with creative expression, showcasing a fresh approach to narrative visualization and digital media. His contributions to the field highlight a convergence of technology and art, reflecting his commitment to exploring new boundaries in storytelling**.**