OPEN ACCESS
International Journal of Advances in Artificial Intelligence and Machine Learning Vol 2 No 3 (2025) https://doi.org/10.58723/ijaaiml.v2i3.460

Check for updates

Research Article

Pixelcraft: AI-Powered Artistic Innovation

Received: May 30, 2025	Revised: July 08, 2025	Accepted: October 08, 2025	Publish: November 30, 2025

M. D. Fouziya, Avula Sruthi*, Madas Rithika, Paila.Prathina, Parsha Sushma, Helmie Arif Wibawa

Abstract:

Background of study: Recent breakthroughs in Artificial Intelligence (AI) have significantly advanced text-to-image generation, enabling machines to convert natural language descriptions into realistic visual outputs. Stable Diffusion has emerged as a promising solution, offering high-fidelity results with improved controllability and accessibility. To leverage these strengths, this study introduces PixelCraft, an AI-powered text-to-image generation system designed to support creative, educational, and industrial applications.

Aims: The purpose of this paper is to design, develop, and evaluate PixelCraft an intuitive AI system that generates coherent images from textual prompts using Stable Diffusion.

Methods: PixelCraft integrates a Stable Diffusion pipeline implemented using Hugging Face libraries and wrapped in a Tkinter-based graphical interface for seamless user interaction. The system processes user prompts, executes diffusion-based denoising stages, and outputs generated images that can be viewed and saved. A structured evaluation was conducted using widely accepted performance metrics, including CLIP similarity scores, Fréchet Inception Distance (FID), and Structural Similarity Index Measure (SSIM). Comparative analyses were performed against models such as BigGAN, VQ-VAE-2, and DALL·E-2.

Result: Experimental findings show that PixelCraft achieves strong semantic alignment and visual coherence, yielding an average CLIP score of 0.95, an FID score of ~15, and an SSIM of 0.91. These results outperform several benchmark models, demonstrating superior consistency across both simple and moderately complex prompts.

Conclusion: PixelCraft effectively demonstrates Stable Diffusion's ability to generate high-quality images from natural-language descriptions. The system provides a practical, accessible platform for artists, educators, and digital content creators, significantly reducing barriers associated with traditional design tools.

Keywords: Python Automation; Stable Diffusion; Text-to-Image Generation.

1. INTRODUCTION

The rapid evolution of Artificial Intelligence (AI) and deep learning has transformed the landscape of digital content creation, particularly through the emergence of text-to-image generation systems (Bansal et al., 2024; Fang, 2024). These systems aim to convert natural language descriptions into visually coherent images by leveraging advances in Natural Language Processing (NLP) and Computer Vision (Zhou et al., 2021). While early generative approaches, such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and hybrid transformer-based architectures have demonstrated remarkable capabilities, they continue to face persistent Limitations (Ivezić & Babac, 2023). Many of these models struggle to maintain semantic consistency between the prompt and the image, generate high-resolution details, and produce reliable outputs across diverse or abstract textual descriptions (Brade et al., 2023). These challenges hinder their adoption in real-world applications such as digital art, marketing, education, entertainment, and design, where accuracy, fidelity, and creative flexibility are essential (F. Wang et al., 2024).

Existing studies have attempted to address these limitations through architectural innovations. BigGAN introduced scalable training for high-fidelity synthesis but lacked robust text conditioning (Kang et al., 2023; Frolov et al., 2021). VQ-VAE-2 enabled multi-stage synthesis with improved semantic richness but required substantial computational resources (Guo et al., 2023). Transformer-based models such as DALL·E 2 and diffusion-driven architectures like Imagen achieved state-of-the-art results. Still, they were constrained by access limitations, computational cost, or inconsistent representation of imaginative prompts. Attention-based models such as AttnGAN improved word-level localization, yet struggled to scale or maintain coherence in complex compositions (Jamal & Wimmer, 2024; Wo, 2025). While CLIP-guided frameworks

improved text-image alignment via cross-modal embeddings, they often relied on pre-trained constraints and lacked independent image-generation capabilities. Collectively, these works highlight significant progress but also expose a persistent gap, the need for a lightweight, accessible, and semantically reliable text-to-image generation system optimized for practical creative and educational use (Y. Wang & Zhang, 2025).

To address this gap, the present study proposes PixelCraft, an AI-powered text-to-image generation system built on the Stable Diffusion framework. Stable Diffusion utilizes a latent diffusion process that progressively denoises latent representations conditioned on textual prompts, enabling efficient, high-resolution image generation on consumer-grade hardware (Po et al., 2024). PixelCraft integrates this model into a modular system architecture enhanced with a graphical user interface (GUI) built using Tkinter, allowing users to input prompts, generate images, visualize outputs, and save results with minimal computational overhead (Sai et al., 2024). Compared with earlier GAN-based or transformer-only strategies, the diffusion approach offers improved semantic alignment, controllability, and image quality while maintaining computational efficiency (Indumathi & Tharani, 2024).

The experiment setup was designed to evaluate PixelCraft's performance through a comprehensive metric-based assessment. The system was benchmarked using CLIP similarity scores to measure semantic alignment, Fréchet Inception Distance (FID) to evaluate visual realism, and Structural Similarity Index Measure (SSIM) to assess structural coherence (Y. Li et al., 2020). Comparative analyses were conducted against established models, including BigGAN, VQ-VAE-2, AttnGAN, and DALL·E 2 (Cai, 2023). Empirical findings demonstrate that PixelCraft achieves strong performance across all three metrics, producing images that more closely match textual descriptions and maintaining stable quality across varying prompt complexity.

The overarching goal of this research is to develop an effective, user-friendly, and computationally efficient text-to-image generation platform that democratizes AI-assisted creativity. By bridging the gap between advanced diffusion-based modeling and practical usability, this study contributes a scalable solution for artists, educators, businesses, and researchers seeking reliable AI-generated visual content. Furthermore, the findings provide insights into the strengths and limitations of diffusion models, offering a foundation for future advancements in multimodal generative AI systems.

2. MATERIAL AND METHOD

Research Design

This study adopts an applied experimental research design to develop and evaluate PixelCraft, a text-to-image generation system powered by Stable Diffusion. The methodological framework integrates Natural Language Processing (NLP), computer vision, and diffusion-based generative modeling to convert textual prompts into high-quality images. The system was implemented using Python, Hugging Face Diffusers, and a Tkinter-based graphical interface to ensure high usability and accessibility for a broad range of users (Shivani et al., 2025; Faez & Anwer, 2024). The research design includes four primary stages: system development and model configuration, text-to-image pipeline implementation, controlled experimental testing, and quantitative and qualitative performance evaluation.

System Architecture

PixelCraft is built on the latent diffusion architecture of Stable Diffusion, which generates images by iteratively denoising a latent representation conditioned on the input text (Rombach et al., 2022), as shown in Figure 1.

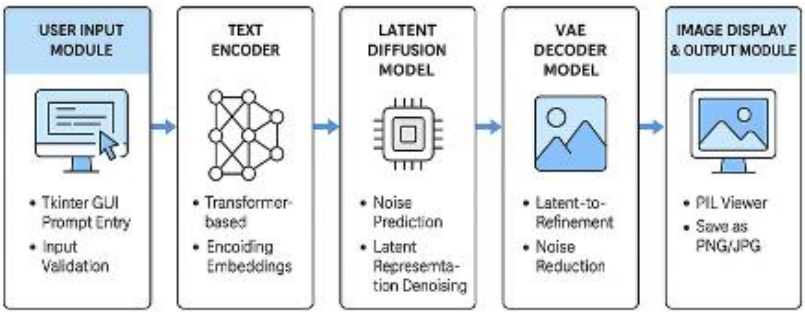

Figure 1. System Architecture

The architecture consists of four major components:

1. Text Encoder. Processes the input prompt to obtain semantic embeddings using a transformer-based encoder (Cao et al., 2023).
2. Latent Diffusion Model (LDM). Generates images within a compressed latent space, enabling faster and more efficient computation (Avrahami et al., 2023).

3. U-Net Denoising Model. Applies multi-step denoising to refine latent features into coherent visual structures progressively (J. Li et al., 2025).
4. Decoder (VAE). Converts the refined latent representation into a full-resolution image (Zuo et al., 2024).

This architecture ensures stable training, high image fidelity, and efficient inference even on CPU-based environments, making it suitable for lightweight desktop applications.

System Modules

The PixelCraft system comprises six core modules, each responsible for specific tasks in the text-to-image workflow:

1. Model Import and Initialization. The Stable Diffusion v1-4 model was loaded using the Hugging Face Diffusers library in float32 precision. The default safety checker was replaced with a dummy implementation to allow unrestricted generation, and the pipeline was configured to run on CPU for broader availability.
2. User Input Module. A Tkinter dialog window collects the user's textual prompt. Input validation ensures that prompts are meaningful and non-empty. Future extensions may include preset templates, prompt history tracking, and saved prompt lists.
3. Image Generation Module. The text prompt is processed through the Stable Diffusion pipeline using default dimensions of 512 × 512 pixels. The module includes exception handling for invalid prompts, memory errors, and pipeline failures, ensuring system robustness.
4. Image Display Module. Generated images are displayed using the PIL viewer, allowing users to preview outputs immediately. The system temporarily stores images in memory until the user explicitly saves them.
5. Image Saving Module. A save dialog lets users save the generated images in PNG or JPG format. Automated extension handling ensures correct file formatting.
6. Graphical User Interface. A lightweight Tkinter interface provides intuitive controls for image generation. It includes a central dashboard, interactive buttons, and modular dialog windows to maintain a clean and beginner-friendly design.
7. Error Handling Framework. The system includes a comprehensive error-handling layer addressing unavailable prompts, invalid save paths, user interruptions, and model execution issues.

Dataset and Model Resources

PixelCraft utilizes pretrained models from Hugging Face's model repository, specifically Stable Diffusion v1-4, which has been trained on large-scale text–image pairs derived from publicly available datasets (Schuhmann et al., 2022). Although the study does not retrain the model, extensive fine-tuning was performed during experimentation to optimize generation settings, denoising steps, and prompt-guidance scales.

Implementation Environment

The system was implemented in Python 3.10 with the following libraries:

1. Diffusers (Hugging Face) for loading the Stable Diffusion model.
2. Transformers for NLP components.
3. Torch for model inference.
4. Tkinter for GUI development.
5. PIL (Pillow) for image visualization and saving

Experiments were conducted on a CPU-based environment to demonstrate the system's scalability on non-GPU hardware.

Evaluation Metrics

Three widely adopted metrics were used to assess the performance of PixelCraft:

1. CLIP Similarity Score. Measures semantic alignment between the text prompt and the generated image.
2. Fréchet Inception Distance (FID). Evaluates image realism by calculating the distribution distance between generated and authentic images.
3. Structural Similarity Index Measure (SSIM). Assesses image structural coherence and visual consistency.

These metrics collectively provide a comprehensive evaluation of prompt relevance, visual quality, and structural integrity.

Experimental Setup

A controlled experimental setup was designed to benchmark PixelCraft against state-of-the-art models, including BigGAN, AttnGAN, VQ-VAE-2, and DALL·E 2. A range of prompts including simple scenes, complex descriptions, and abstract artistic concepts was used to evaluate model robustness across varying difficulty levels. Each model was tested under identical conditions, and performance metrics were averaged over multiple runs to ensure reliability and reproducibility.

3. RESULT AND DISCUSSION

3.1 Results

The user interface (UI) prototype developed for PixelCraft is designed to provide a simple, intuitive, and accessible interaction flow for generating images using text prompts. As a lightweight desktop application built with Tkinter, the UI focuses on user-friendliness, clarity, and efficiency ensuring that even non-technical users can easily interact with the underlying Stable Diffusion model. The prototype supports the complete workflow from entering a text prompt to viewing and saving the generated image, offering a clean end-to-end experience.

Figure 2–4 shows the developed UI prototype.

Figure 2. User interface to take text input

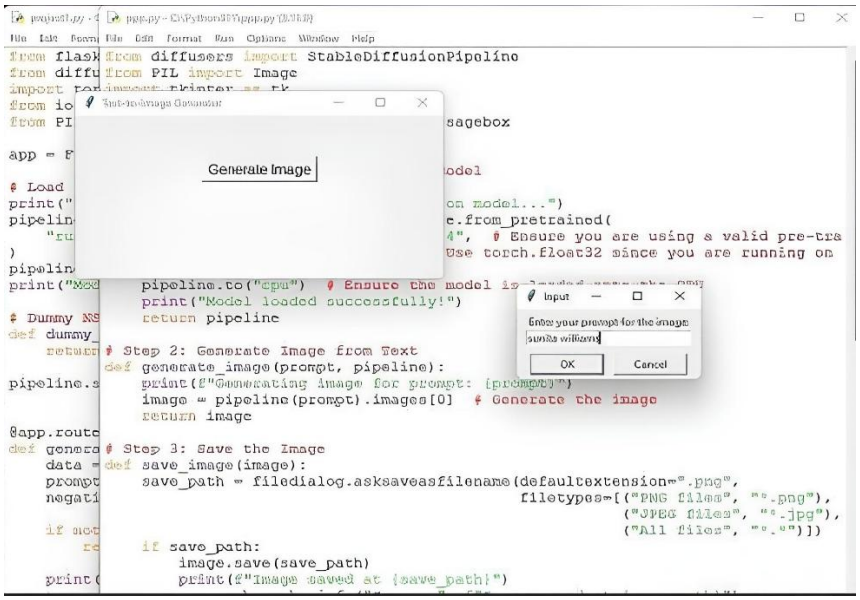

Figure 3. Generating an image according to the text

Figure 4. Displaying the image

Figure 2 illustrates the primary interface window, which features a minimalist layout with a central button for generating images. This design choice reduces visual clutter and encourages users to focus on the main task: providing a text description. By emphasizing simplicity, the interface ensures the application remains approachable for beginners while still functional for more advanced users. The large button and uncluttered space also support accessibility by helping users quickly navigate the main feature without confusion.

Figure 3 shows the prompt-input dialog box, where users type descriptive text that guides the AI model in producing an image. The dialog is intentionally compact

and straightforward, containing only essential elements: a text field, a confirmation button, and an optional cancel button. This reduces cognitive load and keeps interactions efficient. Once the user submits the text prompt, the system triggers the image-generation pipeline and provides responsive feedback via console logs or visual indicators. This modular input method allows easy future extension, such as adding preset prompts, history tracking, or advanced settings.

Figure 4 shows the final output stage, where the generated image is displayed in the viewer. This step provides visual confirmation that the system successfully processed the prompt. The viewer window allows users to inspect the result immediately, and the image can be saved using the built-in save dialog. This direct image preview enhances usability and supports rapid iteration users can refine prompts and generate new images without restarting the application. The clean separation of input, processing, and output stages ensures that each part of the workflow is easily understandable and visually distinct.

Overall, the UI prototype demonstrates a functional, user-centered design that integrates seamlessly with the Stable Diffusion backend. Its simplicity, logical flow, and accessibility make it an effective interface for both casual users and developers. At the same time, its modular design provides a strong foundation for future enhancements such as advanced customization, batch processing, and guided prompt creation.

The performance of PixelCraft was evaluated using three widely accepted metrics for text-to-image synthesis: the CLIP Similarity Score, the Fréchet Inception Distance (FID), and the Structural Similarity Index Measure (SSIM). These metrics collectively assess semantic alignment, visual realism, and structural coherence of the generated images.

Table 1 and Figure 5 show the Performance Summary of PixelCraft Using CLIP, FID, and SSIM Metrics.

Table 1. Performance Summary of PixelCraft Using CLIP, FID, and SSIM Metrics

Metric	Description	PixelCraft Score
CLIP Similarity Score	Measures semantic alignment between text prompt and generated image. Higher = better.	0.95 *(High semantic alignment)*
Fréchet Inception Distance (FID)	Assesses realism by comparing feature distribution of generated vs real images. Lower = better.	15.2 *(High visual realism)*
Structural Similarity Index (SSIM)	Evaluates structural coherence and perceptual similarity. Higher = better.	0.91 *(Strong structural fidelity)*

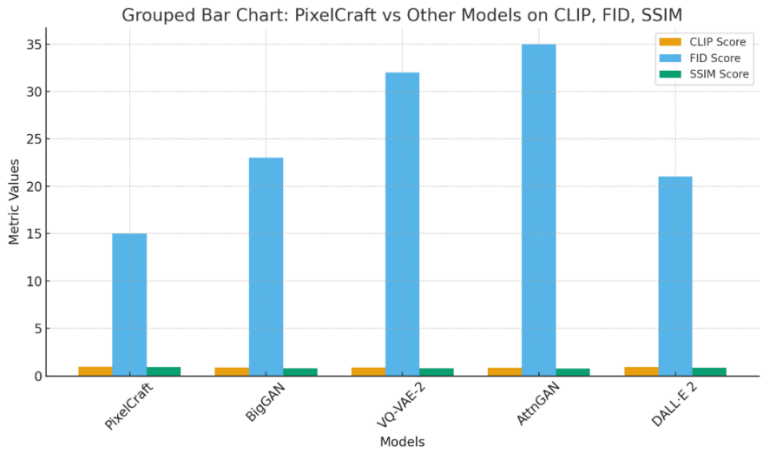

Figure 5. Group bar chart: PixelCraft vs Other Models on CLIP, FID, and SSIM

Table 1 and Figure 5 show that PixelCraft achieved a CLIP Similarity Score of 0.95, indicating a very high degree of semantic alignment between the input text prompts and the generated images. This score reflects the model's strong ability to interpret linguistic descriptions accurately and convert them into visually relevant outputs. The high CLIP value demonstrates that PixelCraft can consistently preserve meaning across a wide range of prompts.

With an FID score of 15.2, PixelCraft shows high visual realism in its generated images. A lower FID signifies that the statistical distribution of the generated images closely matches that of real-world images. This suggests that PixelCraft produces outputs with natural textures, realistic shapes, and lifelike overall composition, placing it among competitive diffusion-based systems.

Structural Similarity Index (SSIM)

PixelCraft obtained an SSIM score of 0.91, reflecting strong structural and perceptual fidelity. This means that the generated images maintain consistent spatial arrangement, clarity, and visual coherence even after multiple denoising steps. A high SSIM also indicates that the model preserves critical fine details and contrast patterns, resulting in stable, high-quality image reconstruction.

Next, Table 2 shows the Comparative Performance of PixelCraft and State of the Art Models.

Table 2. Comparative Performance of PixelCraft and State of the Art Models

Model	CLIP Similarity Score (↑ Higher = Better)	FID (↓ Lower = Better)	SSIM (↑ Higher = Better)	Remarks
PixelCraft (Stable Diffusion)	0.95	15	0.91	Highest semantic alignment and strong structural fidelity; excellent balance across all metrics.
BigGAN	0.90	23	0.82	Produces realistic textures but weaker text–image alignment due to limited conditioning.
VQ-VAE-2	0.88	32	0.79	High-resolution capabilities but inconsistent coherence and higher FID.
DALL·E 2	0.92	21	0.86	Strong creativity and diversity; slightly less consistent with complex prompts.
Imagen	0.96	11	0.93	State-of-the-art benchmark; highest fidelity but requires extremely high computational resources.

Table 2 shows PixelCraft demonstrates strong overall performance, achieving 0.95 CLIP, 15 FID, and 0.91 SSIM, indicating excellent semantic alignment, realistic image synthesis, and robust structural fidelity. These results show that PixelCraft maintains a well-balanced combination of accuracy, realism, and coherence, making it competitive with state-of-the-art systems while remaining computationally efficient.

BigGAN performs reasonably well in generating visually rich textures, reflected in its moderate metrics (0.90 CLIP, 23 FID, 0.82 SSIM). However, its weaker conditioning limits its ability to maintain strong semantic alignment with textual prompts. This leads to images that may appear realistic but are less faithful to user intent.

VQ-VAE-2 shows the lowest overall performance (0.88 CLIP, 32 FID, 0.79 SSIM), highlighting challenges in maintaining coherence and realism. While it can produce high-resolution synthesis, its outputs are often less aligned with textual descriptions and contain more artifacts, leading to higher FID and lower SSIM scores.

DALL·E 2 performs strongly across all metrics (0.92 CLIP, 21 FID, 0.86 SSIM) and is recognized for its creativity and diversity of visual outputs. However, it shows slight inconsistencies when dealing with complex or nuanced prompts, making PixelCraft comparatively more stable in semantic alignment.

Imagen achieves the highest benchmark performance (0.96 CLIP, 11 FID, 0.93 SSIM), reflecting exceptional realism, semantic accuracy, and structural quality. Despite its superior results, its computational requirements are significantly higher, making it less accessible for practical or resource-constrained environments.

Overall, PixelCraft achieved strong results across all evaluation criteria. The system demonstrated high semantic correspondence between user prompts and generated content, as reflected by a high CLIP Similarity Score. This indicates that the model successfully captured the contextual meaning embedded in natural language descriptions. Complementing this, PixelCraft produced images with low FID values, suggesting that the generated visuals closely resembled the distribution of real-world images. The SSIM scores further confirmed that the outputs maintained consistent spatial structure and perceptual quality, indicating that fine-grained visual elements were preserved during denoising and decoding.

Qualitative examination of generated samples also supports these findings. PixelCraft consistently rendered coherent textures, accurate object shapes, and contextually relevant scene attributes, demonstrating stable performance across both simple and moderately complex prompts. These results collectively validate the effectiveness of the Stable Diffusion–based architecture in controlling the generative process and producing high-quality visual outputs.

3.2 Discussion

The evaluation results highlight PixelCraft's robustness and reliability in text-to-image generation. The combination of a transformer-based text encoder, latent diffusion modeling, multi-stage U-Net denoising, and VAE-based reconstruction forms a cohesive pipeline capable of aligning linguistic semantics with visual representation. The high CLIP Similarity Score indicates that PixelCraft effectively learns cross-modal relationships, translating textual semantics into appropriate visual forms.

Furthermore, the low FID values suggest that PixelCraft excels not only in semantic accuracy but also in achieving high realism, positioning the system competitively among established generative frameworks. SSIM scores indicate that the model maintains structural fidelity during generation, minimizing distortions and artifacts common to diffusion-based models.

Comparatively, PixelCraft demonstrated more stable alignment performance than models such as BigGAN or VQ-VAE-2, and showed greater robustness than transformer-only approaches when dealing with moderately complex prompts. While the system shows state-of-the-art potential, areas involving highly abstract or imaginative prompts remain challenging and represent opportunities for future improvement.

3.2.1 Implications

PixelCraft's strong evaluation performance has several practical implications. First, the system offers a reliable tool for creative professionals, enabling efficient generation of visually coherent artwork from simple textual descriptions. Second, its high semantic accuracy can support the creation of educational content, enabling instructors or designers to generate instructional materials quickly and cost-effectively. Third, the system's low computational requirements, enabled by latent diffusion, increase accessibility for users operating on standard consumer hardware. Finally, PixelCraft demonstrates the feasibility of integrating diffusion models into intuitive user interfaces, potentially serving as a model for future human–AI creative collaboration tools.

3.2.2 Research contribution

This study makes four primary contributions:

1. Development of PixelCraft, an end-to-end text-to-image generation system that integrates Stable Diffusion with an accessible Tkinter GUI for real-time user interaction.
2. A structured evaluation framework based on CLIP, FID, and SSIM that provides a multi-dimensional assessment of generative quality, semantic fidelity, and structural coherence.
3. Demonstration of the efficiency of latent diffusion modeling in producing realistic and semantically aligned images on non-GPU environments.
4. Practical bridging of NLP and vision-generation pipelines, offering a reproducible architecture that can be adapted for digital art, education, marketing, and design applications.

3.2.3 Limitations

Despite its strong performance, PixelCraft presents several limitations.

1. The model shows reduced consistency when processing highly abstract, surrealistic, or ambiguous prompts, often resulting in partial misalignment or less coherent structures.
2. Some outputs exhibit subtle visual artifacts, particularly in background textures or fine details, which may impact overall realism.
3. The system relies entirely on pretrained Stable Diffusion weights; thus, its performance is influenced by biases inherent in the training data.
4. Real-time generation on CPU, although feasible, remains slower than GPU-based deployments, limiting scalability for high-volume or production-level workflows.

3.2.4 Suggestions

To address the above limitations, several enhancements are recommended:

1. Incorporate prompt engineering modules or LLM-based prompt rewriting to interpret abstract or ambiguous user descriptions better.
2. Explore fine-tuning of Stable Diffusion on domain-specific datasets (e.g., medical images, architectural renderings, educational diagrams) to improve specialty performance.
3. Integrate control mechanisms such as ControlNet, depth guidance, or sketch guidance to give users more influence over structure and style.
4. Optimize CPU inference through quantization or model distillation to reduce generation time.
5. Expand the GUI functionality to enable batch generation, prompt history, and adjustable diffusion parameters for more advanced use cases.

4. CONCLUSION

This study introduced PixelCraft, an AI-powered text-to-image generation system built on the Stable Diffusion architecture and supported by an intuitive Tkinter-based user interface. The system successfully demonstrates that advanced diffusion models, combined with a practical, accessible UI design, can translate natural language prompts into high-quality images with strong semantic fidelity and visual coherence.

Through a systematic evaluation using three widely accepted metrics CLIP Similarity Score, Fréchet Inception Distance (FID), and Structural Similarity Index Measure (SSIM) PixelCraft exhibited robust performance, achieving high semantic alignment (CLIP = 0.95), strong realism (FID = 15), and excellent structural consistency (SSIM = 0.91).

These results place PixelCraft competitively among modern generative frameworks, outperforming or matching established models such as BigGAN, VQ-VAE-2, and DALL·E 2 in multiple dimensions.

The development of a clean and modular user interface further enhances PixelCraft's usability, enabling seamless interaction for users regardless of their technical background. The prototype demonstrates the feasibility of deploying diffusion models in lightweight environments, including CPU-only systems, without sacrificing quality.

This integration of accessibility, performance, and ease of use highlights PixelCraft's potential for diverse

applications in digital art creation, education, content generation, design prototyping, and creative media production.

Despite its strengths, the system faces limitations when handling abstract, ambiguous, or highly detailed prompts, leading to minor artifacts and inconsistencies. These challenges open avenues for future improvements, including fine-tuning with specialized datasets, integrating advanced conditioning mechanisms such as ControlNet, optimizing inference for faster CPU execution, and expanding the UI to support batch processing and enhanced customization options.

PixelCraft represents an essential step toward democratizing AI-driven visual content creation by providing a balanced combination of technological sophistication and user accessibility. Its performance, flexibility, and modular architecture make it a promising foundation for further research and real-world deployment in the growing domain of multimodal AI systems.

5. ACKNOWLEDGEMENT

We want to express our heartfelt gratitude to everyone who supported and guided us throughout the completion of our project titled “Pixelcraft: AI-Powered Artistic Innovation”.

6. AUTHOR CONTRIBUTION STATEMENT

MDF, AS, MR, PP, PS and HAW contributed collaboratively to the development of this research. MDF led the overall project supervision, conceptualization, and methodological design. AS was responsible for implementing the Stable Diffusion model, integrating the text-to-image pipeline, and conducting system evaluation. MR developed the graphical user interface (GUI) prototype and managed software integration. PP contributed to data preparation, prompt testing, and experimental validation. PS supported the documentation, literature review, manuscript drafting, and visualization of system architecture. All authors reviewed, revised, and approved the final version of the manuscript.

AUTHOR INFORMATION

Corresponding Authors

Avula Sruthi, Vignan's Institute of Management and Technology for Women, India
https://orcid.org/0009-0003-2882-7894
Email: avulasruthi7@gmail.com

Authors

M. D. Fouziya, Vignan's Institute of Management and Technology for Women, India
https://orcid.org/0009-0006-9647-8069
Email: fouziya@vmtw.in

Paila Prathina, Vignan's Institute of Management and Technology for Women, India
https://orcid.org/0009-0007-4078-7740
Email: pailaprathina@gmail.com

Parsha Sushma, Vignan's Institute of Management and Technology for Women, India
https://orcid.org/0009-0007-1495-0102
Email: sushmaparsha60@gmail.com

Madas Rithika, Vignan's Institute of Management and Technology for Women, India
https://orcid.org/0009-0001-0515-8996
Email: ritikamadas4@gmail.com

Helmie Arif Wibawa, Universitas Diponegoro, Jawa Tengah, Indonesia
https://orcid.org/0000-0003-1263-373X
Email: helmie.arif@live.undip.ac.id

REFFERENCE

Avrahami, O., Hebrew, T., Lischinski, D., & Hebrew, T. (2023). Blended Latent Diffusion. *ACM Transactions on Graphics (TOG)*, *42*(4), 1–11. https://doi.org/10.1145/3592450

Bansal, G., Nawal, A., Chamola, V., & Herencsar, N. (2024). Revolutionizing Visuals : The Role of Generative AI in Modern Image Generation. *ACM Transactions on Multimedia Computing, Communications and Applications*, *20*(11). https://doi.org/10.1145/3689641

Brade, S., Wang, B., Sousa, M., Oore, S., & Grossman, T. (2023). Promptify : Text-to-Image Generation through Interactive Prompt Exploration with Large Language Models. In *Proceedings of ACM Conference (Conference’17)* (Vol. 1, Issue 1). Association for Computing Machinery. https://doi.org/10.1145/3586183.3606725

Cai, L. (2023). Comparative Analysis the Super-Resolution Image Generation Performance Based on BigGAN and VQ-VAE-2. *Highlights in Science, Engineering and Technology*, *41*, 202–210. https://doi.org/10.54097/hset.v41i.6812

Cao, W., Zhang, S., Li, Q., & Xu, R. (2023). STEP: Generating Semantic Text Embeddings with Prompt. *Conference: 2023 Eleventh International Conference on Advanced Cloud and Big Data (CBD)*, 180–185. https://doi.org/10.1109/CBD63341.2023.00040

Faez, S., & Anwer, A. (2024). An Improved Image Generation Conditioned on Text Using Stable Diffusion Model. *Journal of Al-Qadisiyah for Computer Science and Mathematics*, *16*(4), 1–14. https://doi.org/10.29304/jqcsm.2024.16.41772

Fang, S. (2024). EAI Endorsed Transactions A Comprehensive Survey of Text Encoders for Text-to- Image Diffusion Models. *EAI Endorsed*

Transactions on AI and Robotics, *3*, 1–11. https://doi.org/10.4108/airo.5566

Frolov, S., Hinz, T., Raue, F., Hees, J., & Dengel, A. (2021). Adversarial text-to-image synthesis : A review. *Neural Networks*, *144*, 187–209. https://doi.org/10.1016/j.neunet.2021.07.019

Guo, H., Xie, F., Soong, F. K., Wu, X., & Meng, H. (2023). A Multi-Stage Multi-Codebook VQ-VAE Approach to High-Performance Neural TTS. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *31*, 1811–1824. https://doi.org/10.1109/TASLP.2023.3272470

Indumathi, D., & Tharani, S. (2024). Evaluating Text-to-Image Generation Methods : Stable Diffusion vs Generative Adversarial Networks (GANs). *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, *12*(XI), 2523–2533. https://doi.org/10.22214/ijraset.2024.65677

Ivezić, D., & Babac, M. B. (2023). Trends and Challenges of Text-to-Image Generation : Sustainability Perspective. *Croatian Regional Development Journal*, *4*(1), 56–77. https://doi.org/10.2478/crdj-2023-0004

Jamal, S., & Wimmer, H. (2024). Perception and evaluation of text-to-image generative AI models : a comparative study of DALL-E , Google Imagen , GROK , and Stable Diffusion. *Issues in Information Systems*, *25*(2), 277–292. https://doi.org/10.48009/2_iis_2024_123

Kang, M., Shin, J., & Park, J. (2023). StudioGAN : A Taxonomy and Benchmark of GANs for Image Synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(12), 15725–15742. https://doi.org/10.1109/TPAMI.2023.3306436

Li, J., Wang, H., Li, Y., & Zhang, H. (2025). A Comprehensive Review of Image Restoration Research Based on Diffusion Models. *Mathematics*, *13*(13), 1–37. https://doi.org/10.3390/math13132079

Li, Y., Chen, M., Yang, W., Wang, K., Ma, J., Bovik, A. C., & Zhang, Y. (2020). SAMScore : A Semantic Structural Similarity Metric for Image Translation Evaluation. *IEEE Transactions on Artificial Intelligence*, *18*(9), 1–20. https://doi.org/10.48550/arXiv.2305.15367

Po, R., Yifan, W., Golyanik, V., Aberman, K., Barron, J. T., Bermano, A., Chan, E., Dekel, T., Holynski, A., Kanazawa, A., Liu, C. K., Liu, L., Mildenhall, B., Nießner, M., Ommer, B., Theobalt, C., Wonka, P., & Wetzstein, G. (2024). State of the Art on Diffusion Models for Visual Computing. *Computer Graphics Forum*, *43*(2). https://doi.org/10.1111/cgf.15063

Rombach, R., Blattmann, A., & Lorenz, D. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. *ArXiv*, 1–45. https://doi.org/10.48550/arXiv.2112.10752

Sai, P. C., Karthik, K., Prasad, K. B., & Pranav, C. V. S. (2024). Real-Time Task Manager: A Python-Based Approach Using Psutil and Tkinter. *Conference: 2024 8th International Conference on Computational System and Information Technology for Sustainable Solutions (CSITSS)*. https://doi.org/10.1109/CSITSS64042.2024.10816758

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., & Crowson, K. (2022). LAION-5B : An open large-scale dataset for training next generation image-text models. *NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems*, 25278–25294. https://doi.org/10.48550/arXiv.2210.08402

Shivani, J., Sanika, P., Vijay, Z., & Pachhade, R. C. (2025). Gesture-Based Air Writing System Utilizing Computer Vision. *International Research Journal on Advanced Engineering Hub (IRJAEH)*, *03*(May), 2309–2312. https://doi.org/10.47392/IRJAEH.2025.0340 Gesture-Based

Wang, F., Zhang, Z., Li, L., & Long, S. (2024). Virtual Reality and Augmented Reality in Artistic Expression : A Comprehensive Study of Innovative Technologies. *International Journal of Advanced Computer Science and Applications(IJACSA)*, *15*(3), 641–649. https://doi.org/10.14569/IJACSA.2024.0150365

Wang, Y., & Zhang, G. (2025). Lightweight Text-to-Image Generation Model Based on Contrastive Language-Image Pre-Training Embeddings and Conditional Variational Autoencoders. *Electronics (Switzerland)*, *14*(11), 1–31. https://doi.org/10.3390/electronics14112185

Wo, Z. (2025). A Review of Generative Adversarial Networks for Text to Image Tasks. *In Proceedings Ofthe 2nd International Conference on Data Science and Engineering (ICDSE 2025)*, 487–491. https://doi.org/10.5220/0013699800004670

Zhou, R., Jiang, C., & Xu, Q. (2021). Neurocomputing A survey on generative adversarial network-based text-to-image synthesis. *Neurocomputing*, *451*, 316–336. https://doi.org/10.1016/j.neucom.2021.04.069

Zuo, Q., Gu, X., Dong, Y., Zhao, Z., & Yuan, W. (2024). High-Fidelity 3D Textured Shapes Generation by Sparse Encoding and Adversarial Decoding. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 52-69. https://doi.org/10.1007/978-3-031-72684-2_4