VIVA-TECH IJRI
ANNUAL RESEARCH JOURNAL
VIVA-TECH INTERNATIONAL JOURNAL FOR RESEARCH AND INNOVATION
ANNUAL RESEARCH JOURNAL
ISSN(ONLINE): 2581-7280

A Survey on Multilingual Text to Image AI Generator

Ankit Chaudhary[1], Vinayak Mishra[2], Durvesh Kasar[3], Bhavika Thakur[4]
[1](*Department of Computer Engineering, VIVA Institute of Technology, Mumbai University, India)*
[2](*Department of Computer Engineering, VIVA Institute of Technology, Mumbai University, India)*
[3](*Department of Computer Engineering, VIVA Institute of Technology, Mumbai University, India)*
[4](*Department of Computer Engineering, VIVA Institute of Technology, Mumbai University, India)*

Abstract: *Multilingual text-to-image generation is a rapidly growing field in artificial intelligence, where models create images from text descriptions in multiple languages. This survey brings together key ideas and advancements in the field. It explores techniques like fine-tuning models to personalize image generation, using translation methods to handle different languages, and improving image quality with advanced technologies like stable diffusion and SDXL models. Applications of these models include education, webtoon creation, architecture, and artistic design, showing how they can be used in various industries. While the technology has made great progress, challenges remain, such as handling language diversity, reducing biases in training data, and managing the high computational demands of these models. This paper provides an overview of the current state of multilingual text-to-image generation, highlighting how it can create accurate and detailed images from text while discussing limitations and areas for improvement. By addressing these challenges, this survey aims to support the development of more inclusive and effective AI systems for creating visual content from text worldwide.*

Keywords *- Artificial Intelligence, Diffusion Models, Generative Models, Image Synthesis, Text-to-Image Generation.*

1. INTRODUCTION

Text-to-image generation has seen extraordinary advancements over the past few years, with models like Generative Adversarial Networks (GANs) and diffusion models leading the way in creating high-quality, detailed images from textual descriptions. These models have significantly improved the ability of AI systems to generate images that closely match the given text, opening up a wide range of applications in areas such as digital art, advertising, design, and education. As these models continue to evolve, there has been a growing emphasis on incorporating multilingual capabilities to ensure that AI systems can process text in various languages and generate accurate, contextually relevant images across diverse linguistic inputs. This multilingual approach has become particularly important in a globalized world, where the demand for inclusive, accessible, and culturally relevant AI tools is rapidly increasing.

Despite these advancements, several challenges remain. One of the main issues is ensuring that the generated images accurately reflect the intended meaning of text across different languages, as language-specific nuances and cultural differences can impact the effectiveness of image generation. Additionally, the computational complexity of training these models to handle multiple languages efficiently presents another significant hurdle. Furthermore, while the quality of images has improved, ensuring that the generated visuals are contextually accurate, especially in complex or abstract descriptions, remains a difficult task. Recent studies have proposed various solutions to address these challenges, including fine-tuning models for personalization, improving cross-linguistic translation methods, and leveraging techniques like stable diffusion models to enhance image realism and diversity.

This survey aims to provide a thorough overview of the current state of multilingual text-to-image generation, offering insights into the latest techniques, applications, and innovations in the field. It will also discuss the challenges that still need to be overcome, such as ensuring cross-linguistic accuracy and improving the efficiency of these models in resource-limited environments. Furthermore, the paper explores potential future

directions for research, focusing on how emerging technologies can address these challenges and push the boundaries of what is possible in multilingual image synthesis. By examining both the progress and the obstacles in this field, this survey highlights the transformative potential of multilingual text-to-image generation and its implications for the future of AI-driven content creation.

2. LITERATURE SURVEY

Nataniel Ruiz [1] concluded that large text-to-image models achieved a remarkable leap in the evolution of AI, enabling high-quality and diverse synthesis of images from a given text prompt. However, these models lack the ability to mimic the appearance of subjects in a given reference set and synthesize novel renditions of them in different contexts. In this work, we present a new approach for "personalization" of text-to-image diffusion models. Given as input just a few images of a subject, we fine-tune a pretrained text-to-image model such that it learns to bind a unique identifier with that specific subject. Once the subject is embedded in the output domain of the model, the unique identifier can be used to synthesize novel photorealistic images of the subject contextualized in different scenes. By leveraging the semantic prior embedded in the model with a new autogenous class-specific prior preservation loss, our technique enables synthesizing the subject in diverse scenes, poses, views and lighting conditions that do not appear in the reference images. We apply our technique to several previously-unassailable tasks, including subject recontextualization, text-guided view synthesis, and artistic rendering, all while preserving the subject's key features. We also provide a new dataset and evaluation protocol for this new task of subject-driven generation.

Muhammad Ajmal [2] focused on developing methods to convert images into text across multiple languages to support literacy education. The study explores techniques for extracting descriptive text from visual content and translating it into various languages, making educational materials more accessible to diverse linguistic audiences. It discusses the integration of image recognition and natural language processing technologies to generate accurate and contextually relevant text. The paper highlights the potential impact on literacy education by providing multilingual resources and facilitating learning in different languages, thereby enhancing educational opportunities for learners around the world.

Mohammed Al-Yaari [3] introduced a novel image processing technique that combines denoising and enhancement through telegraph-diffusion operators. By integrating features of both diffusion and telegraph equations, this method achieves a stable balance between smoothing and detail preservation, overcoming limitations of traditional techniques. The approach demonstrates improved stability and quality in image processing, making it effective for applications requiring both noise reduction and feature enhancement. Experimental results highlight its superior performance and robustness, offering practical benefits for fields such as medical imaging, satellite imaging, and photography.

Lorenzo Papa [4] explored the application of stable diffusion models in generating highly realistic facial images and the subsequent challenges in detecting such synthetic faces. The study examines how stable diffusion techniques can produce high-quality, lifelike facial images that are indistinguishable from real ones, and it also addresses the difficulties in developing detection methods to identify these artificial faces. The research emphasizes the balance between enhancing the realism of generated faces and improving detection mechanisms to mitigate potential misuse. The findings contribute to understanding both the capabilities and limitations of stable diffusion in the context of facial image generation and detection.

Yaoyiran Li [5] presented a novel approach for generating images from text descriptions in multiple languages by leveraging translation techniques. The method enhances text-to-image generation models by first translating multilingual text inputs into a common language, which improves the model's ability to generate accurate images. This approach addresses challenges related to cross-linguistic discrepancies and ensures that the generated images faithfully represent the descriptions provided in various languages. The paper demonstrates that translation-enhanced techniques significantly improve the quality of images produced from diverse textual inputs.

Pedro Reviriego [6] introduced a method for creating high-quality images from text descriptions across a wide range of languages. It tackles the challenge of ensuring that text-to-image models perform effectively regardless of the language used in the input. The approach involves integrating advanced multilingual processing techniques to ensure that descriptions in any language are accurately translated and represented in the generated images. By addressing language-specific nuances and improving model performance across diverse linguistic inputs, the paper aims to make text-to-image generation more inclusive and effective for global applications.

Kyungho Yu [7] investigated the use of multilingual text-to-image models to create webtoons, a popular form of digital comics. The study explores how these models can generate detailed and contextually accurate webtoon panels from text descriptions in various languages. It focuses on overcoming challenges related to language diversity and ensuring that the generated images align with the narrative and artistic style of webtoons. By leveraging multilingual capabilities, the paper aims to enhance the accessibility and creativity of webtoon creation, allowing for broader and more inclusive content generation across different linguistic contexts.

Mr. R. Nanda Kumar [8] explored advancements in artificial intelligence techniques for generating images from textual descriptions. It reviews various AI models and methods, such as generative adversarial networks (GANs) and diffusion models, that translate written text into visual content. The paper highlights the progress in improving image quality, detail, and relevance to the provided text. Additionally, it discusses challenges such as ensuring the accuracy of generated images and addressing biases in training data. The study provides insights into the current capabilities and future directions of text-to-image generation technologies.

Aditi Singh [9] provided a comprehensive overview of current technologies and methodologies in generating visual content from textual descriptions using artificial intelligence. It reviews advancements in both text-to-image and text-to-video generation, detailing various AI models such as generative adversarial networks (GANs), diffusion models, and transformers. The paper compares these technologies in terms of their capabilities, performance, and applications, highlighting strengths and limitations. It also addresses challenges like maintaining coherence in generated content, handling diverse and complex inputs, and ensuring high-quality results. The survey aims to present a clear picture of the state-of-the-art in text-to-visual content generation and identify future research directions.

Akanksha Singh [10] explored the application of deep learning techniques for creating images from textual descriptions. It examines various deep learning models, including generative adversarial networks (GANs) and diffusion models, that transform written text into detailed and contextually accurate images. The paper discusses advancements in model architectures, training strategies, and evaluation metrics, highlighting improvements in image quality and relevance. It also addresses challenges such as handling diverse text inputs and achieving high fidelity in generated visuals. The study provides an overview of current progress in text-to-image generation and identifies areas for future research and development.

Enjellina [11] reviewed the impact of AI image generation technologies on architecture. It explores how AI tools are used to create architectural designs and visualizations, highlighting their influence on creativity, efficiency, and design exploration. The paper discusses challenges such as ensuring accuracy, handling complex design specifications, and integrating AI outputs with architectural workflows. It also examines future prospects for AI in architecture, including potential advancements in generative models and their implications for architectural practice. The review provides insights into how AI can transform the architectural field while addressing current limitations and opportunities for innovation.

Fengxiang Bie [12] surveyed recent advancements in text-to-image generation facilitated by large-scale AI models. It explores how state-of-the-art models, such as large generative transformers and diffusion models, have revolutionized the field by improving the quality and diversity of generated images from textual descriptions. The paper discusses the benefits of large models, including enhanced detail, accuracy, and creative potential, while also addressing challenges like computational costs, model biases, and the need for extensive training data. It provides a comprehensive overview of current technologies, their impact on the field, and future research directions.

Dustin Podell [13] introduced SDXL, an enhanced approach to latent diffusion models aimed at generating high-resolution images. The study presents improvements in the model's architecture and training techniques to achieve finer details and greater accuracy in image synthesis. By optimizing the latent space and diffusion processes, SDXL addresses limitations of previous models, such as resolution constraints and image artifacts. The paper demonstrates that SDXL produces superior high-resolution images with enhanced fidelity and realism, setting a new standard for image synthesis using latent diffusion techniques.

Badhri Narayanan Suresh [14] introduced SDXL, a benchmark designed to evaluate the performance of text-to-image generation models within the MLPerf framework. It focuses on providing a standardized set of metrics and tests to assess the efficiency, accuracy, and scalability of different text-to-image models. The paper details how SDXL measures model performance in generating high-quality images from text inputs, addressing aspects such as inference speed, resource utilization, and model robustness. By establishing clear benchmarks, SDXL aims to drive improvements in text-to-image generation technologies and facilitate comparisons across various implementations.

Nimesh Bali Yadav [15] explored the use of artificial intelligence to create images based on textual descriptions. It reviews key techniques in text-to-image generation, such as generative adversarial networks (GANs) and diffusion models, focusing on how these methods translate written prompts into detailed visual representations. The paper highlights recent advancements in model performance, including improvements in image quality and coherence with the provided text. It also addresses challenges such as handling diverse text inputs and ensuring the accuracy of generated images. Overall, the study provides an overview of current AI capabilities in text-to-image generation and discusses future directions for the field.

2.1 Analysis Table

The following table presents the objective analysis of the research conducted.

2.1 Analysis Table

Title	**Technology Used**	**Advantages**	**Disadvantages**
SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. [1]	Latent diffusion models, UNet backbone, Attention blocks, Cross-attention, Text encoder and Refinement model.	Improved image quality, State-of-the-art performance, Open research and Novel conditioning schemes.	Computational resources, Ethical concerns and Limited control over output.
DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. [2]	Diffusion Models, Text-to-Image Synthesis, Rare-Token Identifiers and Class-Specific Prior Preservation Loss.	Personalization, Diverse Applications and Maintaining Diversity.	Computationally Intensive, Risk of Overfitting and Complexity in Fine-Tuning.
On the use of Stable Diffusion for creating realistic faces: from generation to detection. [3]	Stable Diffusion, Deep learning and Face detection and recognition.	Realistic face generation, Efficient generation and Customization.	Ethical concerns, Limitations in diversity and Computational requirements.
Text to Image Generation: Leaving no Language Behind. [4]	Text-to-image generators, Natural language processing and Machine learning.	Initial exploration of language-dependent performance and Identification of limitations.	Limited scope and Lack of quantitative analysis.
Image to Multilingual Text Conversion for Literacy Education. [5]	Image Processing, Optical Character Recognition, Natural Language Processing and Machine Learning.	Accessibility, Multilingualism, Efficiency and Customization.	Accuracy, Complexity, Cost and Data Dependency.
Stable denoising-enhancement of images by telegraph-diffusion operators. [6]	Telegraph-Diffusion (TeD) operators and Image processing.	Stable image enhancement, Improved image quality and Resolves the Perona-Malik paradox.	Computational complexity, Parameter tuning and Limited applicability.
Translation-enhanced multilingual text-to-image generation. [7]	Neural machine translation (NMT), Multilingual multi-modal encoder and Ensemble Adapter (EnsAd).	Improved multilingual text-to-image generation, Efficient use of resources and Consistent performance.	Dependency on translation quality, Data limitations and Computational complexity.
A Study on Generating Webtoons Using Multilingual Text-to-Image Models. [8]	Multilingual text-to-image models, Natural language processing, Image generation and Webtoon creation.	Efficient webtoon creation, Multilingual content and Creative possibilities.	Limited control over output, Ethical concerns and Computational resources.
TEXT TO IMAGE GENERATION USING AI. [9]	Text-to-image generation, GPT-3, Generative Adversarial Networks (GANs) and	Creative image generation, Versatility and High-quality images.	Limited control over output, Computational resources and Ethical concerns.

	Contrastive Learning.		
A Survey of AI Text-to-Image and AI Text-to-Video Generators. [10]	Deep learning, Natural language processing (NLP), Neural networks, Data preprocessing and Evaluation metrics.	Comprehensive overview, Invaluable resource and Identifies future research directions.	Limited to a survey, Does not provide practical guidance and May not cover all existing techniques.
Text-to-Image Generation Using Deep Learning. [11]	Deep learning, Generative Adversarial Networks (GANs), Natural Language Processing (NLP) and Python programming language.	Creative image generation, Versatility and High-quality images.	Limited control over output, Computational resources and Ethical concerns.
A Review of AI Image Generator: Influences, Challenges, and Future Prospects for Architectural Field. [12]	Artificial Intelligence (AI), Image generation, Deep learning and Generative Adversarial Networks (GANs).	Increased efficiency, Enhanced creativity, Improved communication and Reduced costs.	Lack of human creativity, Ethical concerns, Technical limitations and Dependency on data.
RenAIssance: A Survey into AI Text-to-Image Generation in the Era of Large Model. [13]	Text-to-image generation, Generative Adversarial Networks (GANs), Diffusion models, Large language models and Machine learning.	Comprehensive overview, Identifies key trends and Discusses future directions.	Focus on recent advancements, Limited to a survey and May not cover all existing techniques.
SDXL: An MLPerf Inference benchmark for text-to-image generation. [14]	Text-to-image generation, MLPerf Inference and SDXL.	Standardized benchmark and Focus on inference.	Limited scopePotential for bias and Complexity.
Generation of Images from Text Using AI. [15]	Artificial Intelligence (AI), Generative Adversarial Networks (GANs) and Natural Language Processing (NLP).	Creativity, Efficiency and Versatility.	Limited control, Ethical concerns and Computational resources.

3. PROPOSED SYSTEM

The system takes a text prompt in any language and creates a high-quality image based on it. First, it detects the language and translates the text to English if needed. The text is improved for clarity and converted into a format the AI model (SDXL) can understand. The SDXL model then generates an image from the text. The image is enhanced for better quality, and a description of the image is created and translated back into the user's language. Finally, the user gets the image and its description through an easy-to-use, multilingual interface.

3.1 Algorithm

Step-1: Input Text
Step-2: Text Preprocessing
Step-3: Text Encoding

Step-4: SDXL Model
Step-5: Image Generation
Step-6: Image Postprocessing
Step-7: Output Image.

4. CONCLUSION

Multilingual text-to-image generation has made impressive progress, allowing AI systems to create high-quality and contextually accurate images from text descriptions in multiple languages. These advancements, driven by models like GANs and diffusion models, have opened up many possibilities in areas like art, education, design, and entertainment. However, challenges still exist, such as ensuring that images are generated accurately across different languages, understanding language-specific nuances, and improving the efficiency of these models. While recent innovations like personalized fine-tuning, translation-enhanced methods, and stable diffusion models have made strides in addressing these challenges, more research is needed to further refine the technology and make it more efficient and accessible for diverse global contexts. As the field continues to evolve, multilingual text-to-image generation has the potential to break down language barriers, create more inclusive AI systems, and enable new creative possibilities across industries, making it a valuable tool for the future of AI-driven content creation.

REFERENCES

[1] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein and Kfir Aberman, “DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation”, Computer Vision and Pattern Recognition, 2023.

[2] Muhammad Ajmal, Farooq Ahmad, AM Martinez-Enriquez and Mudasser Naseer, "Image to Multilingual Text Conversion for Literacy Education", 17th IEEE International Conference on Machine Learning and Applications, 2018.

[3] Mohammed Al-Yaari , Hasan Alkahtani, Vadim Ratner and Yehoshua Y. Zeevi, “Stable denoising-enhancement of images by telegraph-diffusion operators”, IEEE, 2013.

[4] Lorenzo Papa, Lorenzo Faiella, Luca Corvitto, Luca Maiano and Irene Amerini, “On the use of Stable Diffusion for creating realistic faces: from generation to detection”, 11th International Workshop on Biometrics and Forensics, 2023.

[5] Yaoyiran Li, Ching-Yun Chang, Stephen Rawls, Ivan Vulić and Anna Korhonen, “Translation-Enhanced Multilingual Text-to-Image Generation”, Computation and Language, 2023.

[6] Pedro Reviriego and Elena Merino Gómez, “Text to Image Generation: Leaving no Language Behind”, Computation and Language, 2022.

[7] Kyungho Yu, Hyoungju Kim, Jeongin Kim,nChanjun Chun and Pankoo Kim, “A Study on Generating Webtoons Using Multilingual Text-to-Image Models”, Applied Sciences, 2023.

[8] Mr.R. Nanda Kumar, Manoj Kumar M, Hari Hara Sudhan V and Santhosh R, "TEXT TO IMAGE GENERATION USING AI", International Journal of Creative Research Thoughts, vol 11, issue 5, May 2023.

[9] Aditi Singh, "A Survey of AI Text-to-Image and AI Text-to-Video Generators", Computer Vision and Pattern Recognition, 2023.

[10] Akanksha Singh, Sonam Anekar, Ritika Shenoy and Prof. Sainath Patil, "Text to Image using Deep Learning", International Journal of Engineering Research & Technology, vol 10, issue 4, April 2021.

[11] Enjellina, Eleonora Vilgia Putri Beyan and Anastasya Gisela Cinintya Rossy, “Review of AI Image Generator: Influences, Challenges, and Future Prospects for Architectural Field”, Journal of Artificial Intelligence in Architecture, 2023.

[12] Fengxiang Bie, Yibo Yang, Zhongzhu Zhou, Adam Ghanem, Minjia Zhang, Zhewei Yao, Xiaoxia Wu, Connor Holmes, Pareesa Golnari, David A. Clifton, Yuxiong He, Dacheng Tao and Shuaiwen Leon Song, "RenAIssance: A Survey into AI Text-to-Image Generation in the Era of Large Model", Computer Vision and Pattern Recognition, 2023.

[13] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna and Robin Rombach, "SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis", Computer Vision and Pattern Recognition, 2023.

[14] Badhri Narayanan Suresh(chair), Ahmad Kiswani, Ashwin Nanjappa, Itay Hubara, Michal Szutenberg, Rachitha Prem Seelin, Vijaya Singh and Yiheng Zhang, “SDXL: An MLPerf Inference benchmark for text-to-image generation”, MLCommons, 2024.

[15] Nimesh Bali Yadav, Aryan Sinha, Mohit Jain and Aman Agrawal, "Generation of Images from Text Using AI", International Journal of Engineering and Manufacturing, 2024.

[16] M. Ozaki, Y. Adachi, Y. Iwahori, and N. Ishii, “Application of fuzzy theory to writer recognition of Chinese characters,” *International Journal of Modelling and Simulation*, 18(2), 1998, pp. 112-116.

[17] R.E. Moore, Interval analysis (Englewood Cliffs, NJ: Prentice-Hall, 1966).

[18] P.O. Bishop, Neurophysiology of binocular vision, in Houseman (Ed.), Handbook of physiology, 4 (New York: Springer-Verilog, 1970) pp. 342-366.

[19] D.S. Chan, Theory and implementation of multidimensional discrete systems for signal processing, doctoral diss., Massachusetts Institute of Technology, Cambridge, MA, 1978.

[20] W.J. Book, “Modelling design and control of flexible manipulator arms: A tutorial review,” 29th IEEE Conf. on Decision and Control, San Francisco, CA, 1990, pp. 500-506.