Multilingual Text-to-Image Generation: A Cross Lingual Synthesis Framework

Deepika A B,Tarun S, Ujwal N, Vikas S Hiremath, Vishal C Halkodu
Bengaluru, India

Abstract. This research investigates the challenges posed by the predominant focus on English language text-to-image generation (TTI) because of the lack of annotated image caption data in other languages. The resulting inequitable access to TTI technology in non-English-speaking regions motivates the research of multilingual TTI (mTTI) and the potential of neural machine translation (NMT) to facilitate its development. The study presents two main contributions. Firstly, a systematic empirical study employing a multilingual multi-modal encoder evaluates standard cross-lingual NLP methods applied to mTTI, including TRANSLATE TRAIN, TRANSLATE TEST, and ZERO-SHOT TRANSFER. Secondly, a novel parameter-efficient approach called Ensemble Adapter (ENSAD) is introduced, leveraging multilingual text knowledge within the mTTI framework to avoid the language gap and enhance mTTI performance. Additionally, the research addresses challenges associated with transformer-based TTI models, such as slow generation and complexity for high-resolution images. It proposes hierarchical transformers and local parallel autoregressive generation techniques to overcome these limitations. A 6B-parameter transformer pretrained with a cross-modal general language model (CogLM) and fine-tuned for fast super-resolution results in a new text-to-image system,denoted as It, which demonstrates competitive performance compared to the state-of-the-art DALL-E-2. Furthermore, It supports interactive text-guided editing on images, offering a versatile and efficient solution for text-to-image generation.

Keywords: Text-to-image generation, Multilingual TTI (mTTI), Neural machine translation (NMT), Cross-lingual NLP, Ensemble Adapter (ENSAD), Hierarchical transformers, Super- resolution, Transformer-based models, Cross-modal general language model (CogLM).

1 INTRODUCTION

The area of Text-to-Image Generation (TTI) has experienced significant advancements in recentyears, propelled by the emergence of large-scale pretrained transformers such as DALL-E and CogView. These models revolutionized the way images are generated from textual descriptions,providing more accurate and diverse outputs. However, one of the key challenges in TTI remains the language barrier, especially when dealing with multilingual contexts. Bridging languages through images is essential for enabling cross-lingual communication and understanding in many different contexts, including as visual storytelling, image captioning, and machine translation.

In this work, we suggest a Multilingual Text-to-Image Synthesis Approach that aims to deal with the language diversity issue in jobs involving the creation of images. By leveraging the potential of deep learning models and pretrained transformers, our method seeks to produce photos of superior quality from textual descriptions in multiple languages. Through a thorough review of existing literature and methodologies in the field of text-to-image generation, we identify the constraints of current approaches and introduce the motivation behind developing our approach, named CogView.

Our research focuses on enhancing the accuracy of semantic objects for generative text-to- image synthesis, aiming to improve the fidelity and diversity of produced images across different languages. By integrating multilingual capabilities into the Text-to-Image Diffusion Model, we strive to offer a thorough solution for bridging languages through images. The expected results of our approach are illustrated, demonstrating the possibility of multilingual text-to-image synthesis in enabling cross-lingual visual communication.

The paper helps to advance multilingual text-to-image synthesis and lays the starting point for additional research in cross-lingual image generation. Through addressing the language barrierin TTI tasks, our approach opens up new possibilities for enhancing

communication and understanding across diverse linguistic contexts.

2 RELATED WORK

Over the past few years, the area of text-to-image synthesis has witnessed significant advancements, driven by the exploration of various approaches and models aimed at improvingthe fidelity and relevance of pictures that were produced. The following works represent a variety of key contributions in this area:

2.1 Hinz et al. introduced the accuracy of Semantic Objects (SOA) metric as a fresh assessment measure for assessing the fidelity of produced pictures to textual descriptions [1]. Their work addresses the difficulty of accurately reflecting complex textual captions in synthesized pictures, which is often hindered by the inherent ambiguity and subjectivity of language. By explicitly modeling individual objects within images and introducing a new evaluation metric, Hinz and associates. provide a valuable tool for quantitatively assessing the alignment between textual descriptions and generated visual content.

2.2 Chang et al. proposed Maskgit, a model that leverages masking techniques for generative image transformation [2]. The Maskgit model integrates masking mechanisms into the image transformation process, allowing for improved control over the generation processand enhancing the realism of synthesized images. By selectively masking certain regions of the input image during the transformation process, Maskgit effectively captures and preserves important visual features while suppressing irrelevant details, leading to more visually appealing results.

2.3 Ding et al. introduced Cogview, a transformer-based model designed for text-to-image generation [3]. Building upon the success of transformer architectures in natural language processing tasks, Cogview demonstrates improved capabilities in mastering the text-to-image generation task. By leveraging attention mechanisms and self-attention mechanisms inherent in transformer models, Cogview is able to effectively capture and integrate textualsemantics into the image generation process, resulting in more coherent and contextually relevant visual outputs.

2.4 Dosovitskiy et al. proposed a transformer-based approach for image recognition, wherein images are represented as grids of words rather than traditional pixel-based representations [4]. By adopting a transformer architecture, Dosovitskiy et al. demonstrate the potential of transformer models in image-related tasks, achieving state-of-the-art results in image recognition. Their work highlights the versatility and effectiveness of transformer architectures in handling complex visual data and underscores the importance of exploring novel approaches in image synthesis tasks.

2.5 Gafni et al. introduced Make-ascene, a text-to-image generation method that incorporatesscene-based control mechanisms [5]. Unlike traditional text-to-image synthesis approaches that focus solely on generating images from textual descriptions, Make-asceneenables users to exert fine-grained control over the generated scenes. By introducing elements such as scene editing and text editing with anchor scenes, Make-ascene enhances the usability and practicality of text-guided image synthesis systems, opening up new possibilities for creative expression and storytelling through AI-generated images.

2.6 Nichol et al. presented GLIDE, a model of text-guided diffusion designed for creating and modifying photorealistic images [6]. By leveraging diffusion models as well as text guidance, GLIDE achieves high-quality results in image synthesis tasks, surpassing previous state-of-the-art methods regarding fidelity and realism. Their work demonstrates the efficiency of leveraging textual descriptions to direct the picture synthesis process, highlighting the potential of text-guided approaches in achieving photorealistic results in creation and manipulation of images tasks.

Collectively, these works represent significant contributions to the ongoing research efforts in text-to-image synthesis, offering novel methodologies, evaluation metrics, and models for generating realistic and contextually relevant images from textual descriptions. As the field continues to evolve, further exploration and innovation in this area are expected to yield even more sophisticated and capable text-to-image synthesis systems.

3 METHODOLOGY

The methodology for the research endeavor "Bridging Languages through Images: A Multilingual Text-to-Image Synthesis Approach" is a meticulously crafted framework that encompasses a series of interconnected steps aimed at revolutionizing Text-to-Image Generation (TTI) capabilities by transcending linguistic barriers and fostering cross-cultural visual communication. This comprehensive methodology delves into the intricacies of data collection, model architecture design, training strategies, ensemble techniques, evaluation metrics, and results analysis, offering a detailed roadmap for researchers to navigate the complex landscape of multilingual image synthesis.

3.1 Data Collection and Preprocessing:

The foundational stage of the methodology involves the meticulous curation of a diverse and inclusive dataset comprising image-text pairs in multiple languages, with a deliberate focus on languages beyond English to ensure representation and cultural diversity. The data undergoes rigorous preprocessing to standardize formats, eliminate noise, and enhance the quality of input data for training the multilingual TTI model. Advanced data augmentation techniques are employed to enrich the dataset, improve model generalization, and capture the nuances of diverse linguistic expressions.

3.2 Model Architecture Design and Innovation:

At the heart of the methodology lies the innovative design of a cutting-edge multilingual text-to-image synthesis model that transcends traditional language boundaries and embraces cultural diversity. The model architecture is meticulously crafted to incorporate a multilingual multi-modal encoder capable of processing diverse linguistic inputs and generating high-fidelity images across a spectrum of languages. A sophisticated hierarchical transformer architecture is adopted to facilitate seamless information flow, context preservation, and efficient image synthesis, particularly for tasks requiring intricate high-resolution generation.

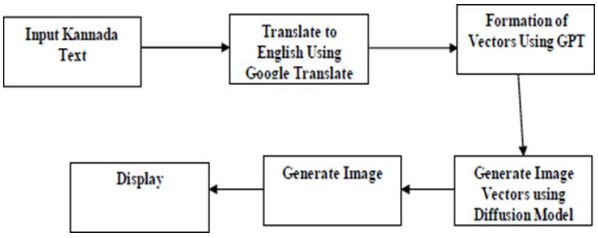


Figure 3.2: The Block diagram of Multi lingual Text to Image Generation

The system architecture as show in Figure 3.2 displays the process for generating images based on input Kannada text.

I. Input Kannada Text:
The process begins with the user providing input in Kannada text, likelydescribing the content or scene for the image generation.

II. Translate to English Using Google Translate:
The Kannada text is then translated to English using Google Translate. This step is crucial if subsequent models or tools in the system work more effectively with English text.

Formation of Vectors Using GPT (Generative Pre-trained Transformer):
The translated English text is fed into a GPT model. GPT is a particular kind of transformer model that excels at natural language understanding and generation. It can convert the translated text into a vector representation, capturing the semantic meaning and the background of the input.

III. Generate Image Vectors using Diffusion Model:
The vector representation obtained from GPT is subsequently fed into a diffusion model. One kind of generative model that can produce images is the diffusion model. It utilizes the vector representation use the text to direct the creation of the images process, incorporating the semantic information from the text input into the image vectors.

IV. Generate Image:
The diffusion model produces the image vectors according to the input from GPT and other parameters. This process includes changing the vector information into pixel data tocreate an illustration of the of the described scene or content.

V. Display:
Finally, the the produced picture is displayed to the user. This could take the shape of a graphical user interface, a web application, or another display mechanism suitable for theintended user interaction.

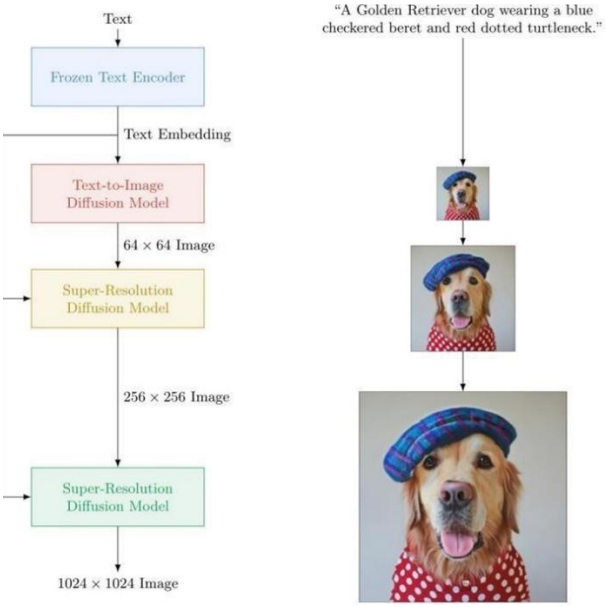
3.3 Training Strategy Implementation and Optimization:
The training strategy is meticulously crafted to leverage the power of pre-trained English TTI models as a foundational framework for extending TTI capabilities to diverse languages through the strategic application of cross-lingual transfer learning methodologies. The methodology explores a spectrum of innovative techniques, including TRANSLATE TRAIN, TRANSLATE TEST, and ZERO-SHOT TRANSFER, to adapt the prototype for new languages with minimal retraining. Fine-tuning procedures are optimized to strike a delicate balance between language-specific features and cross-lingual knowledge transfer, thereby enhancing the multilingual TTI model's performance and adaptability.

3.4 Ensemble Adapter (ENSAD) Integration and Parameter Efficiency:
A groundbreaking approach known as Ensemble Adapter (ENSAD) is introduced to the methodology as a pivotal mechanism for consolidating multilingual text knowledge within the TTI framework with unparalleled efficiency and efficacy. ENSAD acts as a parameter-efficient solution that intelligently weighs and integrates linguistic features from diverse languages, bridging the language gap and enhancing the model's multilingual capabilities. The ensemble technique not only strengthens the model's flexibility in a variety of linguistic contexts but also fosters cross-cultural understanding and promotes inclusive image synthesis practices.

3.5 Evaluation Metrics and Quality Assessment Framework:
The evaluation phase of the methodology is characterized by an exhaustive evaluation of the multilingual TTI model's performance using a diverse array of standard image generation metrics, including but not limited to Inception Score, Fréchet Inception Distance, and perceptual similarity measures. Furthermore to quantitative metrics, the methodology incorporates qualitative assessments through human evaluations to gauge the visual quality, linguistic coherence, and cultural relevance of the generated images across a myriad of languages and cultural contexts. The evaluation framework is intended to validate the model's effectiveness in producing culturally sensitive, visually appealing, and linguistically coherent images that resonate with diverse audiences.

3.6 Results Analysis, Interpretation, and Implications:
The culminating phase of the methodology involves an in-depth examination of the trial findings to assess the effectiveness of the suggested approach in bridging languages through images and fostering cross-cultural visual communication. Comparative analysesbetween the multilingual TTI model and language-specific models are conducted to elucidate the benefits of the multilingual framework in promoting equitable access to TTI technology and advancing cross-cultural understanding. The implications of the findings are covered in a nuanced manner, highlighting the model's performance, scalability, and potential applications in diverse linguistic settings, thereby paving the direction of upcoming research projects and technological developments in the realm of multilingualimage synthesis

4 RESULTS



1. Text
A text is prompted by the user describing the image they wish to generate

2. Text-to-Text Embedding:
Obtain a high-dimensional vector representation (text embedding) capturing the semanticcontent of the input textual description. This can be accomplished with a pre-trained language modellike GPT or through other text embedding techniques.
3. Text Embedding to 64x64 Image:
Use the text embedding to generate a 64x64 pixel image that visually represents the content as stated in the text. A diffusion model, like the one put out in the architecture, leverages thetext embedding to direct the production of a lower-resolution image.

4. 64x64 Image to 256x256 Image:
Upscale the 64x64 image to a higher resolution of 256x256 pixels while preserving and enhancing details. specifics. Make use of super-resolution methods like generative adversarial networks(GANs) or convolutional neural networks (CNNs), to make this happen upscaling.

5. 256x256 Image to 1024x1024 Image:
Further upscale the 256x256 image to a higher resolution of 1024x1024 pixels, maintaining clarity and enhancing finer details. Apply additional super-resolution techniques to achieve the desired image resolution.
• Super-Resolution Text-to-Image Diffusion Model:
1. Text-to-Text Embedding:
Same as above—obtain a high-dimensional vector representation (text embedding).

2. Text Embedding to 64x64 Image:
Generate a 64x64 pixel image using the text embedding as a guide, incorporating details fromthe text. Employ a diffusion model suitable for super-resolution tasks.

3. 64x64 Image to 256x256 Image (Super-Resolution):
Leverage super-resolution diffusion techniques to upscale the 64x64 image to 256x256 pixels, emphasizing higher image fidelity. Utilize advanced super-resolution model capable of preserving and enhancing image details during upscaling

5 CONCLUSION AND FUTURE WORK

Conclusion:
In conclusion, this research underscores the transformative potential of multilingual Text-to-Image Generation (mTTI) as a means to bridge linguistic disparities and democratize access to TTI technology across diverse language communities. By leveraging cross-lingual transfer learning techniques from Natural Language Processing and neural machine translation (NMT) , the study proposes a pragmatic approach to extend the capabilities of TTI models beyond English, mitigating the risks of technological exclusivity and inequitable access. Through empirical investigation and theoretical analysis, the research lays the groundwork for future advancements in mTTI, emphasizing the importance of collaboration, innovation, and interdisciplinary research in addressing the complex challenges posed by linguistic variations inTTI systems.

Future Work:
Looking ahead, several avenues for future research emerge from this study. Firstly, further empirical evaluation and refinement of multilingual TTI models are warranted to assess their performance across diverse languages and cultural contexts. Additionally, investigating new methods for data augmentation, domain adaptation, and fine-tuning of pre-trained models could enhance the resilience and broader applicability capabilities of mTTI systems. Moreover, investigating the incorporation of of user-generated content and user feedback mechanisms into mTTI frameworks can foster user engagement and strengthen the relevance and applicability ofgenerated images. Lastly, addressing ethical, legal, and societal implications, such as bias and representation issues, in multilingual TTI development warrants careful consideration and interdisciplinary collaboration. By pursuing these avenues, future research endeavors can continue to push the boundaries of mTTI technology and contribute to a more inclusive and equitable digital landscape.

Appendix
A Multilingual Text-to-Image Synthesis Approach." It includes details on data collection methods, encompassing sources and preprocessing techniques utilized for textual and image data in multiple

languages. Additionally, it elucidates the architecture of the multilingual text-to-image synthesis model, providing insights into neural network layers, activation functions, and specialized components incorporated. The experimental setup section delineates the hardware, software, and configurations employed for modeltraining and evaluation, while the evaluation metrics subsection elucidates the criteria utilized to assess the model's performance, encompassing standard metrics like BLEU score, SSIM, and Perceptual Distance, alongside any custom metrics devised. Results are thoroughly presented, incorporating tables, graphs, and qualitative analysis of generated images across different languages. Furthermore, additional examples of synthesized images are provided to further illustrate the model's capabilities.Implementation details shed light on the practical aspects of deploying the multilingual text-to-image synthesis approach, including code snippets, libraries, and custom functions. Ethical considerations are also addressed, encompassing issues of data privacy, bias, and cultural sensitivity. Lastly, potential avenues for future research and enhancements to the proposed approach are discussed, suggesting directions for further exploration and experimentation.

REFERENCES

[1] T. Hinz, S. Heinrich and S. Wermter, "Semantic Object Accuracy for Generative Textto- Image Synthesis," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 3, pp. 1552-1565, 1 March 2022, doi: 10.1109/TPAMI.2020.3021209.

[2] H. Chang, H. Zhang, L. Jiang, C. Liu, and W. T. Freeman. Maskgit: Masked generative image transformer. arXiv preprint arXiv:2202.04200, 2022.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[4] M. Ding, Z. Yang, W. Hong, W. Zheng, C. Zhou, D. Yin, J. Lin, X. Zou, Z. Shao, H. Yang, et al. Cogview: Mastering text-to-image generation via transformers. Advances in Neural Information Processing Systems, 34, 2021.

[5] C. Dong, C. C. Loy, K. He, and X. Tang. Learning a deep convolutional network for image super-resolution. In European conference on computer vision, pages 184–199. Springer, 2014.

[6] Z. Du, Y. Qian, X. Liu, M. Ding, J. Qiu, Z. Yang, and J. Tang. All nlp tasks are generation tasks: A general pretraining framework. arXiv preprint arXiv:2103.10360, 2021.

[7] P. Esser, R. Rombach, and B. Ommer. Taming transformers for high-resolution image synthesis. arXiv preprint arXiv:2012.09841, 2020.

[8] O. Gafni, A. Polyak, O. Ashual, S. Sheynin, D. Parikh, and Y. Taigman. Make-ascene: Scene-based text-to-image generation with human priors. arXiv preprint arXiv:2203.13131, 2022.

[9] M. Ghazvininejad, O. Levy, Y. Liu, and L. Zettlemoyer. Mask-predict: Parallel decoding of conditional masked language models. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 6112–6121, 2019.

[10] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. arXiv preprint arXiv:1406.2661, 2014. Bridging Languages through Images: A Multilingual Text-to-Image Synthesis Approach Dept of ISE 2023-2024 24

[11] K. Grace, J- Salvatier, A. Dafoe, B. Zhang, and O. Evans, Viewpoint: When will AI exceed human performance? Evidence from AI experts. Journal of Artificial Intelligence Research, 62, 729–754, 2019. [12] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation. In International Conference on Machine Learning, pages 8821–8831. PMLR, 2021.

[12] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 2022.

[13] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen. Glide: Towards photorealistic image generation and editing with text guided diffusion models. arXiv preprint arXiv:2112.10741, 2021.

[14] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. Kamyar, S. Ghasemipour, B.

Karagol, S. Sara Mahdavi, R. Gontijo-Lopes, T. Salimans, J. Ho, D. J Fleet, and M. Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. arXiv preprint arXiv:2205.11487, 2022.

[15] J. Yu, Y. Xu, J. Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku Y. Yang, B. Ayan, B. Hutchinson, W. Wei, Z. Parekh, X. Li, H. Zhang, J. Baldridge and Y. Wu Yonghui. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation, arXiv preprint arXiv:2206.10789, 2022.

[16] D. Ming Ding et al. Cogview: Mastering text-to-image generation via transformers. Advances in Neural Information Processing Systems, 34, 2021.

[17] B. Dayma, S. Patil, P. Cuenca, K. Saifullah, T. Abraham, P. Le Khac, L. Melas, R. Ghosh. DALL·E Mini, https://github.com/borisdayma/dalle-mini, 2021.

[18] Marta R. Costa-jussà et al, No Language Left Behind: Scaling Human-Centered Machine Translation, https: //arxiv.org/abs/2207.04672.