Vision and Language Models

Project to be submitted in partial fulfillment of the requirements for the degree

of

Bacholers Of Technolgy

in

AI and Data Science

by

Vidit Agrawal B21AI058
Renu Sankhla B21AI028

Under the guidance of

Dr. Mayank Vatsa

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY JODHPUR

SELF DECLARATION

We hereby declare that the research work presented in this project titled **"Vision and Language Models"** is a bonafide work of **Renu Sankhla (B21AI028) and Vidit Agrawal(B21AI058)**, carried out in the Department of Computer Science and Engineering, Indian Institute of Technology, Jodhpur under the supervision of **Dr. Mayank Vatsa** and that it has not been submitted elsewhere for a degree.

Renu Sankhla and Vidit Agrawal

Department of Computer Science and Engineering
Indian Institute of Technology, Jodhpur

Department of Computer Science and Engineering
Indian Institute of Technology, Jodhpur
Rajasthan - India

CERTIFICATE

This is to certify that we have examined the project entitled **Vision and Language Models**, submitted by **Vidit Agrawal(B21AI058) and Renu Sankhla(B21AI028)** undergraduate students of **Department of Computer Science and Engineering** in partial fulfillment for the award of the degree of Bacholers Of Technolgy. We hereby accord our approval of it as a study carried out and presented in a manner required for its acceptance in partial fulfillment for the Bachelor Degree for which it has been submitted. The project has fulfilled all the requirements as per the regulations of the Institute and has reached the standard needed for submission.

Dr. Mayank Vatsa

Department of Computer Science and Engineering
Indian Institute of Technology, Jodhpur

Place: Jodhpur
Date: 13-11-2024

ACKNOWLEDGEMENTS

For their extraordinary guidance and assistance, without which this research would not have been feasible, we would like to extend our appreciation to our guide and faculty supervisor, **Dr. Mayank Vatsa**. He has always encouraged us to do as much research as we can, read as many papers as we can, and test out as many concepts as we can. Whatever challenges we had while working on the project, he was always there to support us.

Finally, we would like to express our sincere gratitude to the Department of Computer Science and Engineering at IIT Jodhpur for providing us with the opportunity to undertake and complete this research. We would also like to extend our heartfelt thanks to PhD Scholar **Surbhi Mittal**, whose unwavering support and guidance constantly challenged us to push the boundaries of our knowledge and explore new dimensions in this research.

IIT Jodhpur
Date: 13-11-2024

ABSTRACT

This research investigates biases in text-to-image (TTI) models for key Indic languages primarily using the Devanagari script, which is widely spoken across India. It evaluates and compares the generative performance and cultural relevance of leading TTI models in these languages against their performance in English. Using the proposed IndicTTI benchmark, we assess the performance of 8 Indic languages with the Meta AI model for image generation.

Keywords: text-to-image generation, bias, MetaAI

Contents

0.1 Introduction

Text-to-image (TTI) models are a type of generative AI that creates images based on text descriptions. These models understand natural language input and attempt to visually represent it, producing images that reflect the details, context, and concepts conveyed in the prompt. TTI models have seen significant advancements, especially with the rise of diffusion models, and are used in various applications, from creative design to enhancing accessibility and personalized content. Artists, designers, and hobbyists use TTI models to produce unique images, concept art, or illustrations from text prompts. Despite these advancements, TTI models still face significant challenges, particularly in representing non-English languages and culturally specific content accurately. Language inclusivity remains a critical area for improvement, as most models are predominantly trained on English-language data, limiting their ability to interpret and visually represent prompts in other languages with equivalent detail and cultural nuance. Recognizing this gap, research into TTI models' language inclusivity has emerged as a priority, especially in multilingual and culturally diverse regions. Expanding TTI models' capabilities to support underrepresented languages and scripts such as those in Indic languages could unlock new applications and improve user experience across a broader spectrum. (1)

This report investigates the current state of TTI models (Meta AI) in handling multiple languages, emphasizing the Devanagari script and eight Indic languages, and offers insights into how inclusive TTI models can foster better cultural representation and accessibility in AI-generated imagery. We use the IndicTTI benchmark, an evaluation framework specifically designed to examine and quantify biases in text-to-image generation technologies. We began this research by working with 30 languages written in 10 different scripts, aiming to generate images based on 200 unique prompts per language that described various scenarios. However, we encountered significant limitations with 22 languages, primarily in non-Devanagari scripts, where less than 20% of prompts could be generated successfully. Given these challenges, we decided to focus our evaluation on languages written in the Devanagari script to ensure a meaningful and thorough assessment. From Fig 1, we observe that English achieved nearly 100% generation success, demonstrating strong model support for English prompts.

Figure 1: Prompt in (en): "The cover image shows an abstract green background" Images generated by Meta-AI when given equivalent prompts in the English and Kashmiri languages, highlight Cultural Stereotypes by representing an image of a Pandit and Om.

Figure 2: Prompt in (en): "The fabric is very dark green" Images generated by Meta-AI when given equivalent prompts in the English and Sindhi languages, highlight Cultural Stereotypes by representing an image of a Sitar and Silk sarees

Among the Devanagari-script languages, Bhojpuri, Awadhi, Hindi, Maithili, Kashmiri, and Sindhi showed a high generation rate, close to or above 80%. However, other Devanagari-script languages like Magahi and Dogri had lower generation success, falling between 40% to 70%. This variation highlights the disparity in model performance across different Indic languages, even within the same script.

Additionally, languages in scripts other than Devanagari, such as Gurmukhi, Kannada, and Bengali, showed minimal generation success, often below 20%, highlighting a significant gap in the model's ability to handle prompts in these scripts. Languages like Arabic, Malayalam, and Sinhala had even lower generation rates, indicating substantial limitations in the model's support for these languages. This analysis emphasizes the current limitations in Meta AI model support for non-Devanagari scripts and highlights the need for targeted improvements in the Meta AI model to accommodate linguistic diversity more effectively across scripts and languages.

Figure 3: Prompt in (en): A rug made with black yarn and yellow stitchs. Images generated by Meta AI when given equivalent prompts in the English and Sindhi languages highlighting the tendency of the model to generate incorrectly.

Figure 4: Prompt in (en): Three men are standing in the street with headphones. Images generated by Meta AI when given equivalent prompts in the English and Awadhi languages highlighting the tendency of the model to generate incorrectly.

Figure 5: Prompt (en): Two men standing next to each other with a remote control. Images generated by Meta AI with equivalent prompts in English and Awadhi demonstrate the model's tendency to misinterpret and miss the semantic meaning, as the images fail to depict two men holding a remote control.

0.2 Motivation

This project is motivated by the need to understand how text-to-image models handle different languages and whether they generate images that are culturally relevant and appropriate. Currently, these models are primarily designed for the English language, often overlooking the unique cultural elements of other languages. This can lead to images that are not only inaccurate or culturally insensitive but may also reinforce misleading stereotypes that do not align with the prompt's intent.

Our goal is to identify insights into how these models (Mainly Meta AI) perform across diverse languages and cultural contexts. By doing so, we aim to highlight areas where improvements are needed to make AI-generated images more meaningful and accurate for speakers of various languages. Ultimately, this project seeks to promote fairness and inclusivity in AI, ensuring that text-to-image technology can be relevant and accessible to all, regardless of language or cultural background.This approach will contribute to more inclusive AI systems that respect and celebrate the richness of global cultures in AI-generated content.

0.3 Related Work

The papers TAM GAN: Tamil Text to Naturalistic Image Synthesis Using Conventional Deep Adversarial Networks, Text to Image Generation: Leaving no Language Behind, and Navigating Text-to-Image Generative Bias across Indic Languages collectively address key challenges in multilingual text-to-image synthesis (2). TAM GAN focuses on generating culturally accurate images from Tamil text, highlighting the need for GANs to adapt to non-Latin scripts, ensuring culturally relevant and meaningful visual outputs. Reviriego and Merino-Gómez explore the performance degradation in text-to-image models when handling languages other than English, emphasizing the need for improvements to ensure consistent performance across diverse linguistic inputs. Mittal et al.'s research takes a deeper dive into biases within text-to-image models for Indic languages, evaluating 30 languages spoken by over 1.4 billion people. They introduce the IndicTTI benchmark to assess and compare the generative performance of popular models, revealing significant gaps in cultural relevance and support for Indic languages.

Together, these studies highlight the urgent need for text-to-image models to be more inclusive, ensuring accessibility and fairness for speakers of non-English and lesser-represented languages, while also preserving cultural diversity. This combined perspective informs our research on improving multilingual and culturally sensitive text-to-image generation.

0.4 Methodology — Benchmark Design

In this work, we propose a benchmark on Indian languages. Our benchmark comprises eight languages evaluated using the Meta AI open-source text-to-image generation model, with six evaluation metrics that cover aspects of correctness and representation. The languages included in this study are shown in Figure 7.

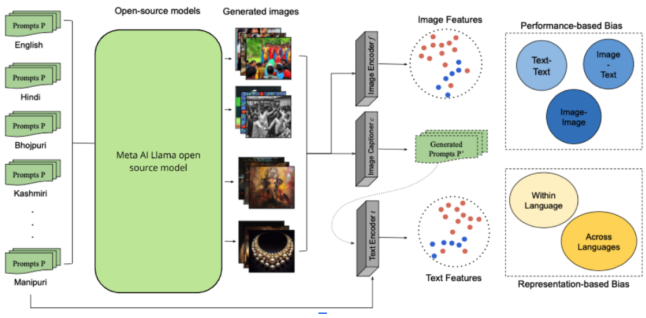

Figure 6: Pipeline for the generation and evaluation of the benchmark

0.4.1 Indic Languages and Prompts

In our research, we study the performance of text-to-image (TTI) models in eight languages. As shown in Figure 7, the Meta AI Llama model predominantly generates images in the Devanagari script compared to other Indian languages like Telugu and Urdu. The languages we work with include Bhojpuri, Awadhi, Hindi, Chhattisgarhi, Kashmiri, Magahi, and Sindhi, in addition to English. All these languages are represented in the Devanagari script. Our dataset contains seven Indian languages along with English, for which captions have already been created. For each language, we have 69 prompts, all translated from English to the respective Indian languages using the Meta NLLB 200 translator (3).

Language Code	Name	Family	Script	Sub-family	#Native Speaker
bho_deva	Bhojpuri	Indo-Aryan	Devanagari	Northern Indo-Aryan	*
awa_deva	Awadhi	Indo-Aryan	Devanagari	Northern Indo-Aryan	2.52 M
hin_deva	Hindi	Indo-Aryan	Devanagari	Central Indo-Aryan	528.3 M
hne_deva	Chhattisgarhi	Indo-Aryan	Devanagari	Northern Indo-Aryan	13 M
kas_deva	Kashmiri	Indo-Aryan	Devanagari	Indo-Aryan	*
snd_deva	Sindhi	Indo-Aryan	Devanagari	North Western Indo-Aryan	2.7 M
mag_deva	Magahi	Indo-Aryan	Devanagari	Northern Indo-Aryan	13.5 M

Figure 7: Languages

0.4.2 TTI Models and Generated Images

We utilize the Meta AI open-source Llama model for the benchmark. Using this model, we generate four images per prompt for the selected 69 prompts across the eight languages. This results in a total of 2,208 images, with 276 images for each language.The size of the images is 1280 x 1280 pixels.

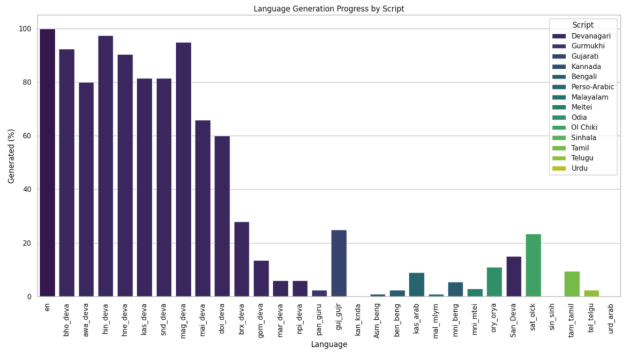

Figure 8: Generated images using Meta AI for most Indian languages

We can see from Figure 8 that scripts such as Assamese, Arabic, Bengali, Malayalam, Gujarati, Kannada, Oriya, and Tamil have a generation rate for the 200 prompts below 25 percent. This highlights the limited performance of the model in generating images for these languages, suggesting a need for improved representation and language-specific adaptations in multilingual generative models to better support these diverse scripts.

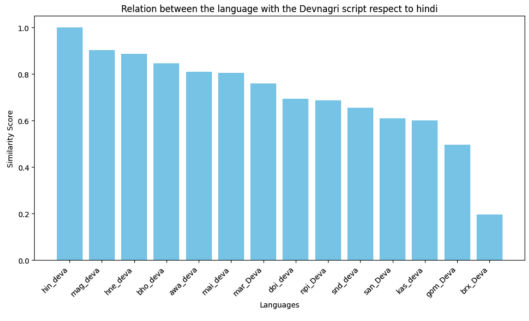

Figure 9: Relation between the languages within the Devanagari script

We observe that languages in the Devanagari script achieve high generation rates of about 85-90 percent, suggesting that the Meta model has been trained more extensively on this script. Additionally, when we compare the similarity of sentence

embeddings for prompts in Devanagari-script languages, as shown in Figure 9, we find that languages with high similarity scores, such as Magahi , Chhattisgarhi , Bhojpuri , Awadhi, and Maithili , exhibit a positive correlation with generation rate. This indicates that the model recognizes keywords in these languages that closely align with Hindi vocabulary, allowing it to generate contextually relevant images based on these shared linguistic features. This analysis suggests that the model leverages Hindi semantics to interpret and generate relevant content for related languages in the Devanagari script. By identifying shared linguistic features, it effectively captures meaning across these languages, enhancing generation accuracy.

0.4.3 Evaluation Methodology

For evaluation, all experiments are conducted on Linux-based systems using Python libraries, specifically the PyTorch library. To extract rich semantic text and image features for the evaluation metrics, we utilize various modules of the BLIP-2 model (4). The image encoder of the BLIP-2 model serves as the image feature extractor f for computing the CLGC, IGC, SCAL, SCWL, and DWL metrics.

Additionally, for the CLGC metric, we leverage the image-captioning capabilities of the BLIP-2 captioner c to generate captions for the produced images. To extract rich textual features, we employ the Sentence-Former model. For the LGC metric, we require image-text features, which are extracted from BLIP-2 using the LAVIS library, as discussed in the experiments section.

0.4.4 Prompt Selection

Prompt selection is crucial for investigating generative models, as the prompts themselves directly influence the diversity, relevance, and cultural sensitivity of the images generated. By carefully selecting prompts, we can examine how well a model interprets various cultural symbols, regional relevance, and linguistic understanding across different languages. For example, using culturally specific prompts like a traditional festival scene allows us to see if the model accurately captures elements unique to that culture. Diverse prompts also help us assess the model's inclusivity, showing how it represents different groups, traditions, and settings. This approach ensures a thorough evaluation, allowing us to highlight the model's strengths and pinpoint areas for improvement to better serve a globally diverse audience.

This prompt selection analysis highlights a broad spectrum of themes that are essential for examining cultural awareness and diversity in multilingual image generation models. Each category, ranging from representations of people and activities to elements of nature, art, and design, reflects aspects that are culturally sensitive and may vary greatly across different languages.

0.4.4.1 Categories of Diverse Prompts Taken

People and Activities Representations of individuals, attire, and actions.

- "A man in winter clothes and scarf walking down the street."
- "A woman is sitting at her desk using a laptop."
- "A man and woman dancing together in the studio."

Home and Interior Decor Household items, furniture, or home settings.

- "The dining room table is set with white chairs."
- "A living room with red couches and large pictures."

Food and Utensils Food items or kitchen-related scenarios.

- "The vegan meal is on display in this poster."
- "A table with candles and blue dishes on it."
- "Four bowls are shown in different colors."

Nature, Outdoors, and Animals Elements related to outdoor settings, plants, and animals.

- "Soccer players are on the field during a game."
- "An old bicycle parked in front of a house."
- "An upside down bicycle is in the middle of a garage."
- "A baseball field with trees and a fenced-in area."

Objects and Accessories Individual items or accessories, which may have cultural significance.

- "A necklace made with pearls and gold."
- "A bottle of whiskey with the label Teeling and Sky."

Buildings and Urban Scenes Buildings, architectural elements, and public spaces.

- "A large group of houses with many red roofs."
- "A building with scaffolding and construction materials on it."

Art and Design Artistic representations, paintings, and patterns.

- "A painting with an image of a boy playing chess."
- "An intricate flower design in black and white."
- "The fabric is very dark green."
- "A rug made with black yarn and yellow stitches."

Technology and Office Settings Technology-related items or work environments.

- "Business people in a meeting with the word accounting."
- "The charity fund website homepage."

0.5 Experiments and Result

In the IndicTTI benchmark, we focus on two aspects of TTI model evaluation: correctness and representation. Correctness refers to the ability to measure the semantic faithfulness with which the model generates the images for the corresponding prompt. On the other hand, representation refers to the measurement of the diversity of generated images within and across the different languages for the given prompts.

We adopt the following notation to describe the metrics and the evaluation pipeline:

- Set of languages L
- Set of prompts $P = \{p_1, p_2, \ldots, p_k\}$
- Corresponding representative images $Q = \{q_1, q_2, \ldots, q_k\}$
- Generative model g

For each prompt in P, we generate n images, resulting in a total of $n \cdot |P| \cdot |L|$ images for a given generative model g. The set of images generated by g for a prompt p_k in language $l \in L$ is denoted as $I_{p_k,l}$.

$$I_{p_k,l} = \{i_{p_k,l,r}\}_{r=0}^{n}$$

Figure 10: Evaluation Pipeline

where $I_{p_k,l,r}$ denotes the image generated using g as $g(p_k, l)$.

Across all the metrics, we compute high-level semantic features of images (using feature extractor f) and text (using feature extractor t) for effective comparison using the similarity function ϕ.

0.5.1 Correctness-based Metrics

These three metrics comprehensively evaluate the generated images across text-text, image-image, and image-text spaces. A higher correctness value indicates greater accuracy of the generative model in producing images that are faithful to the text.

$$p'_k = c(i_{p_k,l,r})$$

Figure 11: Cyclic Language-Grounded Correctness Process

Cyclic Language-Grounded Correctness (CLGC): This metric evaluates the generated images I in the text-text space. This is achieved in a cyclic manner through generating captions over I using a caption generator c.

where p'_k denotes the generated caption for the r-th image in the set $I_{p_k,l}$. Then, with p_k and p'_k as prompts in the English language for the language l, we obtain:

$$CLGC(l) = \frac{1}{n|P|} \sum_{p_k}^{P} \sum_{r=0}^{n} \phi[\, t(p_k),\ t(p'_k)\,]$$

Figure 12: Comparison of Generated Captions

Image-Grounded Correctness (IGC): For this metric, the correctness of the generated images is evaluated in the image-image space.

$$IGC(l) = \frac{1}{n|P|} \sum_{p_k}^{P} \sum_{r=0}^{n} \phi[\, f(q_k),\ f(i_{p_k})\,]$$

Figure 13: Image-Grounded Correctness Evaluation

where the ground-truth image q_k is compared to each of the n generated images for the prompt p_k.

Language-Grounded Correctness (LGC): This metric evaluates correctness in the image-text space with p_k denoting the prompt in the English language.

0.5.2 Representation-based Metrics

In this section, we discuss the three metrics utilized for measuring the representativeness of concepts across the languages. We propose the Self-Consistency Across Languages metric to capture the bias in representation across the different languages.

$$LGC(l) = \frac{1}{n|P|} \sum_{p_k}^{P} \sum_{r=0}^{n} \phi[\ t(p_k),\ f(i_{p_k})\ ]$$

Figure 14: Language-Grounded Correctness Analysis

Self-Consistency Across Languages (SCAL): This metric captures the variation in the generated images for the same prompt across the different languages.

Navigating Text-to-Image Generative Bias across Indic Languages, for a language pair (l_a, l_b).

$$SCAL(l_a, l_b, p_k) = \frac{1}{{}^{n}C_2} \sum_{u=0}^{n} \sum_{v=0}^{n} \phi[\ f(i_{p_k,l_a,u}),\ f(i_{p_k,l_b,v})\ ]; u \neq v$$

$$SCAL(l_a, l_b) = \frac{1}{|P|} \sum_{p_k}^{P} SCAL(l_a, l_b, p_k)$$

Figure 15: Self-Consistency Across Languages

The lower the value of SCAL, the less consistent the generations are across the different languages. Conversely, a higher value of SCAL would demonstrate high consistency between concepts across the different languages.

Self-Consistency Within Language (SCWL):

$$SCWL(l, p_k) = \frac{1}{{}^{n}C_2} \sum_{u=0}^{n} \sum_{v=0}^{n} \phi[\ f(i_{p_k,u}),\ f(i_{p_k,v})\ ]; u \neq v$$

$$SCWL(l) = \frac{1}{|P|} \sum_{p_k}^{P} SCWL(l, p_k)$$

Figure 16: Self-Consistency Within Language

This metric is computed within the images generated for a particular prompt for a given language. A high self-consistency showcases good consistency between the semantic content of the generated images.

Distinctiveness Within Language (DWL): This metric depicts the diversity across the images generated for the different prompts p_a and p_b.

A higher distinctiveness score showcases the capability of the model to generate diverse images with varying prompts.

$$SWL(l) = 1 - \frac{1}{|P|C_2} \sum_{p_a}^{P} \sum_{p_b}^{P} \phi[\ f(i_{p_a}),\ f(i_{p_b})\ ]; a \neq b$$

Figure 17: Distinctiveness Within Language

0.6 Benchmark Results and Analysis

Correctness Bias:

- **Distinctiveness Within Language (DWL):** For all languages, DWL values are approximately 86%, indicating high diversity across images generated for different prompts.
- **Cyclic Language-Grounded Correctness (CLGC):** For the CLGC metric, values across languages are generally less than 40%, indicating moderate effectiveness in cyclic consistency between original and generated captions.
- **Image-Grounded Correctness (IGC):** Higher IGC metric values across all language prompts suggest strong semantic similarity between generated images and ground-truth images.

Representativeness Bias:

- **Language-Grounded Correctness (LGC):** LGC values across languages are below 40%, highlighting limitations in language-grounded representational consistency.
- **Self-Consistency Across Languages (SCAL):** SCAL values around 45% for all languages indicate moderate consistency in generated image representations across languages.
- **Self-Consistency Within Language (SCWL):** SCWL values are approximately 35%, suggesting moderate consistency within language-specific generated images for the same prompt.

Following our analysis of the metric values, we compare our results with those obtained from DALL-E 3 (1), as illustrated in the figure below. All values are scaled to a range between 0 and 1, where values closer to 1 indicate better performance, while

Metric	en	mag_deva	hin_deva	hne_deva	bho_deva	awa_deva	kas_deva	snd_deva
CLGC	0.217425	0.205875	0.214609	0.211941	0.216652	0.215193	0.20262	0.199751
IGCM	0.933867	0.909251	0.907719	0.904264	0.907868	0.91491	0.892946	0.894935
LGCM	0.359177	0.324843	0.326614	0.326486	0.314481	0.298674	0.290383	0.273467

Correctness -based matric result using Meta AI

Metric	en	mag_deva	hin_deva	hne_deva	bho_deva	awa_deva	kas_deva	snd_deva
CLGC	0.657	0.6042	0.6416	0.5918	0.6075	0.5737	0.4745	0.5678
IGCM	0.4852	0.4417	0.4486	0.4403	0.4422	0.4488	0.4404	0.4218
LGCM	0.3306	0.2939	0.3157	0.3002	0.2999	0.2891	0.2309	0.1992

Correctness -based matric result using Dalle-3

Figure 18: Correctness -based matric result using Meta AI and Dalle-3

values near 0 reflect poorer performance. Our findings indicate that the Language-Grounded Correctness (LGC) metric yields results comparable to those of DALL-E 3. However, in terms of the Image-Grounded Correctness (IGC) metric, our Meta AI model achieves an impressive average score of 90%, while DALL-E 3's performance is nearly half that of our model. Additionally, in the Cyclic Language-Grounded Correctness (CLGC) metric, DALL-E 3 demonstrates a higher value than our model.

0.7 Analysis

In this section, we qualitatively analyze the images generated by the Meta AI across the different languages. Based on our observations in the previous section, we explore the generated images and study the alignment between the quantified metrics and generated images. As in the previous section, we study two aspects of bias: correctness-based and representation-based. We observe that the images generated by the Meta AI model are particularly sensitive to the Sindhi language. In many instances, the generated images, especially those involving human figures, are depicted as cartoon characters, which highlights potential limitations in the model's capacity to authentically capture cultural and linguistic diversity in its image generation process. Examples of this phenomenon can be seen in Figure 19. For instance, when prompted to generate an image of a baseball field with trees and fenced areas, the model failed to include humans on the field. Instead, it captured only the keyword "baseball" and generated an animated cartoon character holding a baseball, demonstrating the model's struggle to properly interpret the full context of the prompt. Additionally, taking the example of the prompt "Black and white photograph of a man standing on a cliff overlooking mountains," the model generated an image of a monk on the mountains, which reflects a cultural stereotype. This suggests that the model associates mountains with monks or Buddhists, overlooking the possibility of a diverse range of individuals, such as mountaineers or tourists, who might also be found in such settings. This highlights how the model may be relying on cultural assumptions or biases, leading to a narrow and potentially inaccurate representation of the scene based on its interpretation of the prompt.Refer Figure 20. Similarly, for the prompt "Four bowls are shown in different colors," the model adds cultural context by filling the bowls with specific regional dishes. While this may reflect certain food associations, it can introduce unintended biases, misrepresenting the diversity of food choices and overemphasizing cultural stereotypes. This highlights the model's tendency to incorporate additional context based on its training data. Similarly when analyzing the images generated with the help of the prompts in the Awadhi language we came across some examples of misinterpretation such as the prompt "the cover for it in defense of housing." in English typically conveys a message related to a poster cover that displays housing rights. However, when the same prompt is given

Figure 19: Prompts in English : 1. Two men standing next to each other with a remote control. 2. A baseball field with trees and a fenced-in area.

in Awadhi, the model generates an image depicting a group of Indians holding hammers and hockey sticks. This misinterpretation highlights the model's failure to grasp the semantic meaning of the prompt in Awadhi, instead associating it with a more aggressive or confrontational scenario. The model struggles to accurately capture the intended context and cultural nuances, leading to a mismatch between the input and the generated image.Refer Figure 21. A similar example of cultural stereotyping can be seen in Figure 22, where the prompt "The legs and feet of a man in black running shoes" is provided. When given in English, the model successfully generates an accurate image that matches the prompt. However, when the same prompt is given in Awadhi, the model generates an image of worn-out shoes in a slushy and muddy environment. This shift in the generated image indicates that the model may be associating the Awadhi language with a stereotype of poor or worn-out footwear, reflecting an underlying bias in its image generation process. This suggests that the model struggles to dissociate certain cultural contexts from the literal meaning of the prompt. Similarly, stereotypes related to housing and infrastructure appear when examining the prompt "A large group of houses with many red roofs." For languages like Hindi, Maghai, and Chhattisgarhi, the generated images depict rooftops resembling temple and monastery structures, indicating a cultural bias. This suggests the model associates these languages with specific architectural styles, reinforcing stereotypes rather than generating neutral representations of housing. Additionally, some images also feature large slums, further highlighting a bias toward portraying impoverished living conditions, which may not accurately reflect the diversity of housing in these regions.Refer Figure 23.

We also observe in the cultural and artistic representations that flowers and other art forms are depicted in a way that reflects traditional Indian motifs in the Indic prompts. These images often emphasize elements such as intricate floral patterns,

Figure 20: Prompts in English : 1. Black and white photograph of man standing on a cliff overlooking mountains. 2. Four bowls are shown in different colors.

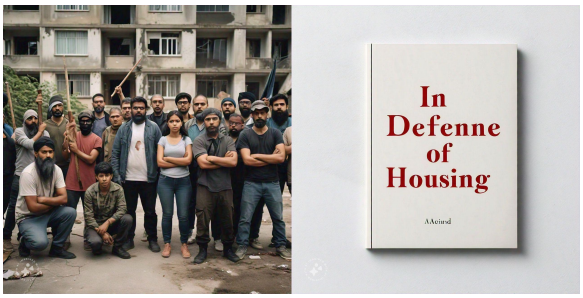

Figure 21: Prompt in English : the cover for it in defense of housing.

Figure 22: Prompt in English :The legs and feet of a man in black running shoes

Figure 23: Prompt in English : A large group of houses with many red roofs

Figure 24: Cultural Representation in different Languages

religious symbols, and classical art forms, which are deeply rooted in Indian culture. This highlights how the model tends to incorporate culturally specific aesthetics, yet it may also reflect an overemphasis on certain cultural symbols while overlooking the diversity of artistic expressions across different regions and communities.

In examining the primary objective of studying cultural representation, it becomes evident that, despite prompting for Indian contexts, many generated images still predominantly depict Western or American characteristics. This trend suggests that the model has an underlying bias towards Western-centric representations, limiting its ability to accurately reflect the cultural context intended in Indian-themed prompts. Such instances highlight an important gap in the model's capability to adapt to cultural nuances in image generation across diverse languages and contexts. Refer to the Figure 24 for getting to know about the cultural representation in the images.

Lastly, we observe that some images have been incorrectly generated due to limitations in the language translation model, NLLB (No Language Left Behind), which resulted in inaccurate translations of prompts. These translation inaccuracies directly impacted the image generation process, leading to mismatched or irrelevant outputs. For example, prompts in languages like Awadhi or Magahi, when translated by NLLB, may have been misunderstood by the model, causing it to generate images that did not align with the intended context or meaning of the original prompt. This highlights how translation errors can exacerbate the bias and misrepresentation in the generated images, further distorting the cultural context and leading to inaccurate or misleading visual depictions. Refer Figure

Figure 25: Incorrect Image generation due to Wrong Translation like baseball getting translated to basketball and green tracksuit getting translated to green tea

Figure 26: Prompt in English: An older woman is smiling while holding a remote. Incorrect representation of Indian men instead of women

Figure 27: Prompt in English: A man in winter clothes and scarf walking down the street. Images generated by Meta AI help in highlighting the tendency of the model to generate incorrectly

0.8 Conclusion and Future Direction

In this research, we evaluated the performance of text-to-image (TTI) generation models, specifically Meta AI's open-source model, on Indic languages primarily written in the Devanagari script(5). Through the IndicTTI benchmark, we highlighted significant discrepancies in generative success and representational accuracy across languages, emphasizing the current limitations of TTI models in supporting non-English prompts with cultural specificity and linguistic inclusivity. While languages like Hindi, Bhojpuri, and Awadhi achieved a high generation success rate, others, even within the Devanagari script, showed marked variability in generation quality. Furthermore, the representational metrics underscored biases and inconsistencies across both language diversity and cultural depiction, revealing moderate self-consistency and low language-grounded correctness.

Limitation: One major limitation of this research is the focus on Devanagari-script languages, due to low generative success rates in other scripts like Gurmukhi, Kannada, and Bengali. As a result, our findings may not generalize to all Indic scripts, limiting the scope of our conclusions on language inclusivity. Additionally, while we evaluated several correctness and representation metrics, these measures do not capture all aspects of cultural relevance or user perception, particularly in assessing the nuanced depiction of regional elements. Lastly, the limited dataset of 69 prompts per language may restrict the variability and comprehensiveness of the evaluation, potentially affecting the model's performance assessment.

Future Directions : Future work can address these limitations by expanding the dataset to include more diverse and contextually rich prompts across multiple Indic languages and scripts beyond Devanagari. Incorporating a broader array of TTI models and comparing their performance will also provide a more comprehensive view of the state of language inclusivity in TTI technologies.

lang	distinctiveness
en	0.8423029780387878
mag_Deva	0.8707159161567688
hin_Deva	0.874266505241394
hne_Deva	0.8696539402008057
bho_Deva	0.8713070750236511
awa_Deva	0.87724369764328
kas_Deva	0.8575316071510315
snd_Deva	0.8752766251564026

Figure 28: DWL

Cyclic Language-Grounded Correctness

key	en	mag_Deva	hin_Deva	hne_Deva	bho_Deva	awa_Deva	kas_Deva	snd_Deva
000070697	0.1388	0.1396	0.278	0.1467	0.1724	0.1747	0.141	0.158
000143885	0.0644	0.1268	0.0565	0.0747	0.1352	0.075	0.0677	0.0412
000193984	0.4444	0.4818	0.4243	0.4677	0.4417	0.4714	0.4527	0.486
000135378	0.2275	0.1975	0.1576	0.1221	0.2123	0.3221	0.1882	0.1637
000052001	0.3581	0.1786	0.2312	0.2232	0.1971	0.2043	0.2087	0.2139
000147209	0.109	0.1045	0.1063	0.0417	0.1285	0.065	0.0469	0.1313
000000362	0.2614	0.251	0.2553	0.2044	0.256	0.1829	0.2698	0.2875
000095797	0.2208	0.2183	0.3507	0.2516	0.2392	0.2939	0.2901	0.2895
000123434	0.2456	0.2252	0.2014	0.2887	0.2328	0.2042	0.2212	0.2601
000030486	0.2062	0.2091	0.1951	0.2073	0.2069	0.1763	0.2037	0.2066

Figure 29: CLGC

0.9 Evaluation metrics results

Image grounded correctness

key	en	mag_Deva	hin_Deva	hne_Deva	bho_Deva	awa_Deva	kas_Deva	snd_Deva
000282477	0.9463	0.9492	0.9634	0.9556	0.9541	0.939	0.9551	0.9502
000110094	0.9946	0.9717	0.9878	0.9922	0.9849	0.9736	0.9834	0.9341
000124175	0.9727	0.9448	0.96	0.9614	0.957	0.9565	0.96	0.959
000012521	0.9478	0.9341	0.9341	0.8965	0.9307	0.9341	0.9287	0.915
000052001	0.7998	0.7158	0.7461	0.7427	0.7461	0.7236	0.7241	0.6812
000070697	0.9746	0.8735	0.8071	0.9512	0.9312	0.9619	0.9526	0.9258
000193984	0.9199	0.8887	0.8408	0.8184	0.8867	0.8545	0.9023	0.9258
000170556	0.8159	0.918	0.8252	0.8564	0.8525	0.8647	0.8579	0.8354
000043499	0.9766	0.9492	0.9663	0.9541	0.958	0.9409	0.9546	0.9648
000102725	0.9746	0.9473	0.9834	0.9717	0.9736	0.9424	0.9785	0.9746

Figure 30: IGC

Language grounded correctness

key	en	mag_Deva	hin_Deva	hne_Deva	bho_Deva	awa_Deva	kas_Deva	snd_Deva
000195285	0.4014	0.4054	0.3875	0.3865	0.2351	0.3784	0.2299	0.3376
000070697	0.3907	0.2656	0.4371	0.2702	0.2634	0.1899	0.2275	0.2513
000034998	0.3374	0.3279	0.3206	0.3262	0.3283	0.3318	0.3142	0.3284
000104792	0.4349	0.4138	0.4188	0.4214	0.4268	0.4156	0.2549	0.4051
000093847	0.3907	0.391	0.3706	0.3405	0.3073	0.27	0.3311	0.319
000061146	0.3592	0.3591	0.3783	0.3781	0.3428	0.3757	0.3184	0.3664
000145354	0.337	0.3831	0.2221	0.2975	0.3481	0.3544	0.3867	0.2144
000053783	0.3123	0.2678	0.1044	0.3293	0.303	0.3366	0.294	0.2342
000150454	0.4224	0.3637	0.3847	0.3725	0.3418	0.1562	0.3398	0.3674
000082205	0.2767	0.273	0.2803	0.2715	0.3033	0.0537	0.2562	0.1973

Figure 31: LGC

SCAL

key	en-mag_Deva	en-hin_Deva	en-hne_Deva	en-bho_Deva	en-awa_Deva	en-kas_Deva	en-snd_Deva	mag_Deva-hin_Deva	mag_Deva-hne_Deva	mag_Deva-bho_D
000193984	0.46240234375	0.440673828125	0.439697265625	0.455322265625	0.453857421875	0.45654296875	0.4609375	0.457275390625	0.4638671875	0.465087890625
000075124	0.3994140625	0.4624023437 5	0.450195312 5	0.4658203125	0.466796875	0.462646484375	0.46533203125	0.406494140625	0.40478515625	0.4052734375
000112012	0.492431640625	0.487060546875	0.483642578125	0.48291015625	0.489501953125	0.437255859375	0.488525390625	0.485107421875	0.484619140625	0.48486328125
000004336	0.470458984375	0.479736328125	0.472412109375	0.475830078125	0.4775390625	0.4765625	0.472900390625	0.478271484375	0.490478515625	0.48291015625
000093465	0.4482421875	0.4716796875	0.473388671875	0.467041015625	0.465576171875	0.46923828125	0.4609375	0.44873046875	0.445556640625	0.443359375
000045281	0.4384765625	0.4140625	0.45458984375	0.421142578125	0.4482421875	0.354736328125	0.29052734375	0.423828125	0.44873046875	0.41943359375
000165385	0.388916015625	0.4482421875	0.400634765625	0.383056640625	0.45751953125	0.46484375	0.42236328125	0.40625	0.457275390625	0.48291015625
000171429	0.482421875	0.484619140625	0.484375	0.484130859375	0.483642578125	0.432373046875	0.482421875	0.4794921875	0.47802734375	0.47900390625
000093691	0.42822265625	0.406494140625	0.404052734375	0.417724609375	0.420654296875	0.408447265625	0.402587890625	0.440185546875	0.4521484375	0.43310546875
000053783	0.336669921875	0.367919921875	0.352294921875	0.366455078125	0.347412109375	0.365966796875	0.371826171875	0.37939453125	0.346923828125	0.36181640625

Figure 32: SCAL

SCWL

key	en	mag_Deva	hin_Deva	hne_Deva	bho_Deva	awa_Deva	kas_Deva	snd_Deva
000035590	0.336517333984375	0.3354034423828125	0.2974090576171875	0.3048858642578125	0.3305816650390625	0.3332977294921875	0.3390350341796875	0.29425048828125
000110094	0.371063232421875	0.3592529296875	0.3713836669921875	0.3708343505859375	0.3716888427734375	0.3582916259765625	0.3684539794921875	0.36431884765625
000030486	0.368560791015625	0.365753173828125	0.3570098876953125	0.3689117431640625	0.367919921875	0.3508911328125	0.3664398193359375	0.364166259765625
000141227	0.336578369140625	0.322479248046875	0.3618011474609375	0.364349365234375	0.3525543212890625	0.352813720703125	0.361297607421875	0.3308868408203125
000002185	0.3713836669921875	0.3700408935546875	0.37261962890625	0.37249755859375	0.3722381591796875	0.37200927734375	0.3646240234375	0.3578338623046875
000124175	0.3572845458984375	0.3660430908203125	0.3681640625	0.3697967529296875	0.3669891357421875	0.3652801513671875	0.3657989501953125	0.3647003173828125
000180443	0.31243896484375	0.326904296875	0.31170654296875	0.3100738525390625	0.3328094482421875	0.3544769287109375	0.33502197265625	0.336639404296875
000150454	0.35009765625	0.36322021484375	0.332061767578125	0.3214569091796875	0.345428466796875	0.3609466552734375	0.3138885498046875	0.340423583984375
000104803	0.3650512695312 5	0.366363525390625	0.3661346435546875	0.3644561767578125	0.3677368164062 5	0.3572235107421875	0.3698120117187 5	0.3657684326171875
000190885	0.3131561279296875	0.3413238525390625	0.3363800048828125	0.358245849609375	0.3477935791015625	0.352630615234375	0.3082733154296875	0.350982666015625

Figure 33: SCWL

Bibliography

[1] S. Mittal, A. Sudan, M. Vatsa, R. Singh, T. Glaser, and T. Hassner, "Navigating text-to-image generative bias across indic languages," 2024. [Online]. Available: https://arxiv.org/abs/2408.00283

[2] D. M and K. A, "Tam gan: Tamil text to naturalistic image synthesis using conventional deep adversarial networks," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 22, no. 5, May 2023. [Online]. Available: https://doi.org/10.1145/3584019

[3] N. Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, and J. Wang, "No language left behind: Scaling human-centered machine translation," 2022. [Online]. Available: https://arxiv.org/abs/2207.04672

[4] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," 2023. [Online]. Available: https://arxiv.org/abs/2301.12597

[5] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, and A. Letman, "The llama 3 herd of models," 2024. [Online]. Available: https://arxiv.org/abs/2407.21783