Evaluating Gender Bias in Multilingual Multimodal AI Models: Insights from an Indian Context

Kshitish Ghate, Arjun Choudhry, Vanya Bannihatti Kumar
Language Technologies Institute, Carnegie Mellon University
{kghate, arjuncho, vbanniha}@cs.cmu.edu

Abstract

We evaluate gender biases in multilingual multimodal image and text models in two settings: text-to-image retrieval and text-to-image generation, to show that even seemingly gender-neutral traits generate biased results. We evaluate our framework in the context of people from India, working with two languages: English and Hindi. We work with frameworks built around mCLIP-based models to ensure a thorough evaluation of recent state-of-the-art models in the multilingual setting due to their potential for widespread applications. We analyze the results across 50 traits for retrieval and 8 traits for generation, showing that current multilingual multimodal models are biased towards men for most traits, and this problem is further exacerbated for lower-resource languages like Hindi. We further discuss potential reasons behind this observation, particularly stemming from the bias introduced by the pretraining datasets. Our code can be found here.

1 Introduction

In recent years, significant work has been done to ground image and language models together, to enable the ability to perform various downstream tasks like visual question answering, text-prompted image generation, and image captioning. These models typically involve merging image and text transformer architectures, making use of the contextual knowledge learned by these models during pretraining and reducing the model training cost and complexity. Models like `BLIP` (Li et al., 2022), `BLIP-2` (Li et al., 2023), and `CLIP` (Radford et al., 2021) are frequently used for various multimodal tasks, including dataset curation.

Recent models like `mBLIP` (Geigle et al., 2023), `mCLIP` (Chen et al., 2023a), `cross-lingual CLIP` (Carlsson et al., 2022) further build upon these to extend image-to-text tasks into a multi-lingual realm. However, these models are designed with language inclusivity in mind, with no prior evaluation of bias. Since inclusivity extends beyond just language inclusivity, this brings up the question, are these models really inclusive? As large-scale multimodal models become more integrated into global, multilingual contexts, it is essential to ensure fair representation.

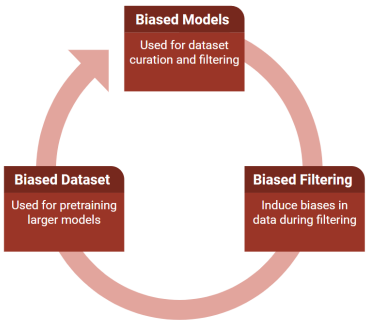

Figure 1: Bias amplification in large models.

There are also text-to-image diffusion models like `Imagen` (Saharia et al., 2022), `DALL-E 2` (Ramesh et al., 2022), and `Latent Diffusion` (Rombach et al., 2022a), which rely on pretrained unimodal text or image models that are extended to create their multimodal models. These models perform exceedingly well on metrics for predictive performance, but their bias evaluation was largely unexplored till recently. In this work, we analyze Stable Diffusion 2 (Rombach et al., 2022b) and Alt-Diffusion (Chen et al., 2022) for gender biases in generated images.

The bias evaluation of these models is extremely critical, since these models are further used for the curation of large-scale datasets used for various pretraining and fine-tuning tasks. E.g., the LAION-5B dataset (Schuhmann et al., 2022) was curated by extracting the data from Wikipedia and filtering using the CLIP model. It was used for training various large-scale models like BLIP-2. Since the CLIP model is biased, as shown

Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 338–350
August 16, 2024 ©2024 Association for Computational Linguistics

by (Wolfe and Caliskan, 2022), the LAION-5B dataset is likely to show biases, since the CLIP model was used to filter it, and these biases are further propagated to new models trained on the LAION-5B dataset. This process is called *Bias Amplification* (Hooker, 2021).

In this work, we take inspiration from the work by Wolfe and Caliskan (2022), and evaluate the gender biases portrayed by a mCLIP-based retrieval framework and mCLIP and CLIP-based multilingual text-to-image diffusion models. Based on the work done by Wolfe and Caliskan (2022), we felt the need for potentially region-specific work on evaluating gender biases in models, since traits and in-group words are likely to differ in the context of people across regions, countries, and continents. E.g., the trait `Indian`, is likely to lead to incoherent results for people from across the globe or in other regions like North America, while evaluating it for Indians in particular might portray gender biases different from the global trends for a given model. Thus, in this work, we explicitly work for the Indian context, and evaluate gender biases observed in text-to-image retrieval and generation models for English and Hindi prompts.

2 Related Work

This section builds upon prior studies to contextualize our analysis of gender bias within multilingual multimodal models. We highlight the unique contributions and limitations of existing methodologies in handling cultural and linguistic differences in AI systems.

Wolfe and Caliskan (2022) evaluated three SOTA image-to-text models `CLIP`, `SLIP` (Mu et al., 2021) and `BLIP` for biases associated with social and experimental psychology, particularly associated with equating the American identity as white. Upon running embedding association tests on the Chicago Face Database (Ma et al., 2015), they observed that White individuals had a higher association with collective in-group words compared to Asian, Latina/o, and Black individuals across all models. Certain phrases like *patriotism* and *born in America* were more associated with White individuals. This work introduced a new direction for the evaluation of multimodal models. However, it was restricted to monolingual models trained only in English.

Bhatt et al. (2022) offers an essential backdrop for the current research. This paper's comprehensive analysis of social disparities in India and their manifestation in NLP data and models lays the groundwork for understanding how cultural and linguistic diversity impacts AI fairness. The present study extends this understanding by applying these fairness considerations to the specific context of gender bias in multimodal models, thus filling a critical gap in the understanding of AI fairness in multilingual and multicultural settings. Saxon and Wang (2023) introduced the "Conceptual Coverage Across Languages" (CoCo-CroLa) technique, assessing the parity of generative text-to-image systems across languages. They focused on tangible nouns and their representation in image generations across various languages. Our approach is in line with their multilingual analysis but applies specifically to Hindi and the Indian context, providing a more targeted evaluation of biases in specific downstream tasks such as retrieval and generation. Ruggeri et al. (2023) conduct a multi-dimensional analysis of bias in vision-language models, focusing on gender, ethnicity, and age. Their study highlights the presence of harmful and stereotypical completions when subjects are input as images, which also perpetuate to downstream tasks, affecting minorities. Our work extends this by examining gender bias in generated images using AltDiffusion and Stable Diffusion 2, specifically comparing biases in English and Hindi prompts and considering the impact of language and cultural contexts, thereby broadening the scope to explore multilingual biases.

Wang et al. (2022a) examine multilingual fairness in multimodal models, focusing on equal treatment across languages. Their introduction of multilingual individual and group fairness concepts is pertinent to understanding gender biases in multilingual contexts. However, our study diverges by zooming in on gender bias outcomes in explicit downstream tasks, specifically within Indian demographics and incorporating Hindi, addressing a gap in Wang et al. (2022a)'s research. Chen et al. (2023b) evaluate the extensive capabilities of large-scale multilingual vision-language models in diverse tasks, such as object detection and video question answering. They also discuss bias-demographic parity in the proposed model, underscoring the significance of evaluating demographic disparities in AI systems. Our work adds a crucial layer to this conversation by explicitly

addressing gender biases, thereby contributing to a deeper understanding of the limitations and inherent biases in multilingual multimodal models. Wang et al. (2022b) in their study on FairCLIP introduced a novel two-step debiasing method for CLIP-based image retrieval, to find a balance between debiasing and performance. Concurrently, Kong et al. (2023) emphasized test-time fairness in image retrieval through Post-hoc Bias Mitigation, modifying outputs of pre-trained models for enhanced equity. We specifically derive our measures of gender bias from these works and apply them to a multilingual context.

3 Methodology

3.1 Gender Bias: In the context of multilingual multimodal models

We consider gender bias in the context of multilingual multimodal models to refer to the presence of unfair and undesirable associations, stereotypes, or imbalances related to gender within the model's understanding and generation of language and images across multiple languages and modalities. This bias can manifest in various ways and impact the model's performance, leading to unequal or inappropriate treatment of individuals based on their gender.

We highlight some key aspects which are responsible for the manifestation of gender bias in multilingual multimodal models:

- Language Bias: The model may exhibit bias in its understanding and generation of language across different languages. This bias can be reflected in the choice of words, phrases, or language structures that perpetuate stereotypes or favor one gender over another.
- Visual Bias: In multimodal models that process both text and images, gender bias can emerge in the interpretation and generation of visual information. This may include biased recognition of gender-related visual cues or the generation of biased visual content.
- Translation Bias: In multilingual models, translations of gendered terms or expressions may introduce bias if not handled appropriately. Translating from one language to another can sometimes result in the reinforcement of gender stereotypes or the loss of differences that are associated to gender identity.
- Training Data Bias: Bias in the training data used to train the model can significantly impact its performance. If the training data contains gender-related stereotypes or imbalances, the model is likely to learn and perpetuate those biases in its predictions and outputs.
- Cultural Sensitivity: Multilingual models should be sensitive to cultural differences related to gender norms and expectations. Failing to account for these differences may result in biased outputs that do not align with the diverse perspectives and expressions of gender across different cultures.

This study takes a step towards addressing gender bias in multilingual multimodal models by first quantifying it in the retrieval and generation settings, and showing how it can exacerbate for low-resource languages such as Hindi.

3.2 Measuring Gender Bias in Image Retrieval

We first focus on analysing gender bias in the text-to-image retrieval setting. We introduce a bias metric that aims to reflect the disparity in representation between male and female genders in the results of gender-neutral queries.

Let us consider a set of images V, where each image $v \in V$ is associated with a gender attribute $g(v)$, taking a value of $+1$ for male and -1 for female. For a query c, the retrieved set of images $V_{c,K}$ should ideally exhibit no gender bias, meaning that it should contain an equal number of male and female-associated images (Wang et al., 2021, 2022a). Following Wang et al. (2022b) and Kong et al. (2023), we define the gender bias in the retrieved image set is quantified as the normalized absolute difference in counts of each gender's images:

$$
\begin{aligned}
&AbsBias(V_{c,K}) \\
&= \frac{1}{K} \Bigg| \sum_{v \in V_{c,K}} \mathbb{1}\{g(v) = +1\} \qquad (1) \\
&- \sum_{v \in V_{c,K}} \mathbb{1}\{g(v) = -1\} \\
&= \frac{1}{K} \Big| \sum_{v \in V_{c,K}} g(v) \Big| \qquad (2)
\end{aligned}
$$

Here, $\mathbb{1}\{.\}$ is an indicator function, K is the number of top images considered.

To evaluate an image retrieval system across multiple queries, we can aggregate the bias scores over a collection of gender-neutral queries C. The aggregated bias metric, denoted as AbsBias@C, is the average of individual bias scores across all queries in C:

$$\text{AbsBias@}C = \frac{1}{|C|}\sum_{c\in C} B(V_{c,K}) = \frac{1}{|C|}\sum_{c\in C}\frac{1}{K}\left|\sum_{v\in V_{c,K}} g(v)\right| \quad (3)$$

We further extend our analysis to quantify how much more 1 gender is preferred in retrieval compared to another by defining MaleBias and MaleBias@C. These are simply the previously defined measures without applying the absolute operation.

These metric serves as a critical evaluation for fairness, providing a measure of the system's performance in offering balanced representations across genders.

3.3 Dataset

In this work, we use the Chicago Face Database (CFD) (Ma et al., 2015), which is a dataset of images used to study race and ethnicity in psychology. It includes 597 images of male and female images with self-identified race or ethnicity. The races and ethnicities included in the dataset are Asian, Black, Latina/o, and White. The dataset includes images with people portraying neutral, happy(open mouth), happy(closed mouth), angry, and fearful expressions. In line with previous works by Devos and Banaji (2005) and Wolfe and Caliskan (2022), we use only the images with neutral facial expressions in our experiments.

The training data used in the models we are evaluating our bias metrics on, tells a lot about the bias expressed by these models and hence understanding this training data is very important. For our analysis, we use the mCLIP model, a multilingual multimodal text-to-image model. The following is a description of datasets used to train mCLIP. The multilingual text encoder of this model is trained using the parallel text corpus MT6 which contains 120M parallel sentences between English and six languages and covers 12 language directions (Chi et al., 2021). The triangle cross-modal knowledge distillation is done using the CC3M dataset (Sharma et al., 2018). For the mCLIP+ variant, in addition to the MT6 dataset, the multilingual text encoder is trained with OPUS-100 dataset (Zhang et al., 2020) covering a total of 175M parallel sentences among 100 languages. The dataset used for the triangle cross-modal knowledge distillation of the mCLIP+ variant is TrTrain (CC12M), which is obtained by applying the translate-train method and translating the English captions of CC12M (Changpinyo et al., 2021).

3.4 Text-to-Image Retrieval

We employ a top-50 text-to-image retrieval approach using the mCLIP model to examine gender bias in response to gender-neutral trait queries. The process involves a pool of facial images taken from CFD, consisting of equal numbers of male and female individuals self-identified as Indian (N = 104). For each trait, deemed gender-neutral, the model retrieves the top 50 images that it associates most closely with the given trait. These traits are expressed in both English and Hindi, allowing us to explore potential disparities across languages. This method provides a comprehensive view of how the model perceives and associates gender with specific characteristics, offering insights into the inherent biases of multilingual multimodal AI systems.

To quantify the observed gender biases, we use a bias metric adapted from recent fairness studies in AI as introduced in Section 3.2. By aggregating these bias scores over a set of selected traits, we assess the overall gender bias exhibited by the model. Aggregating bias scores across multiple traits allows us to draw more generalized conclusions about the model's tendency towards gender bias in image retrieval tasks. We then compare the relative gender bias exhibited by the mCLIP model across the Hindi and English languages.

We select trait categories to represent 3 major characteristics of an individual: Identity (person and Indian), drawing from concepts in Caliskan et al. (2022) and specific to the Indian context; Status/Class (employed and business), drawing from concepts in Kozlowski et al. (2019); Attributes, a list of 50 attributes (25 highest valence and 25 lowest valence) taken from Warriner et al. (2013) and Caliskan et al. (2017).

We specifically choose single words without templates for this task following Saxon and Wang (2023) and Wang et al. (2022b) since our analysis showed template approaches can yield biased results due to choice of template (May et al., 2019).

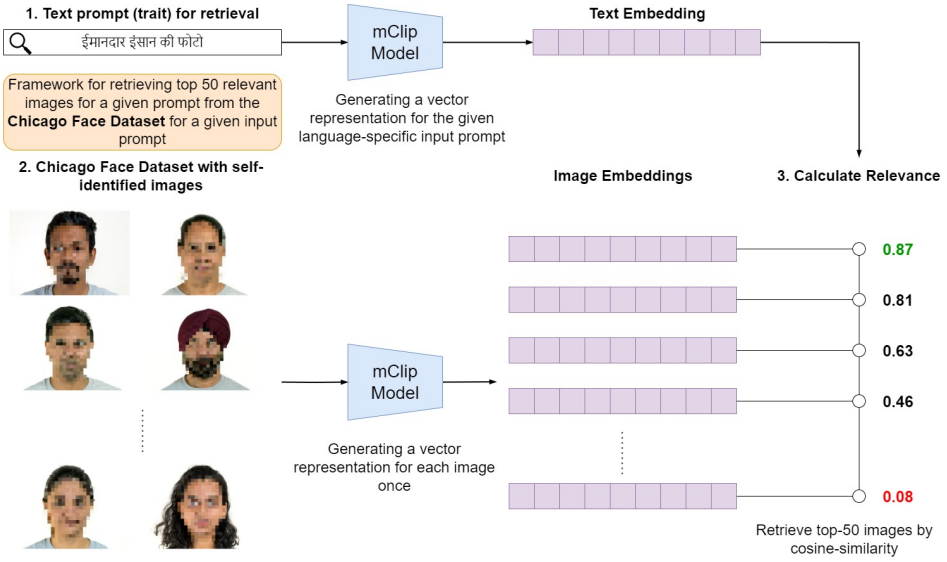

Figure 2: Our text-to-image retrieval pipeline for extracting images from the Chicago Face Dataset using traits as prompts in multilingual settings.

3.5 Text-to-Image Generation

To further understand the bias variation across languages in different settings of multilingual multimodal models, we perform similar experiments in a "generation" setting as opposed to "retrieval". We mainly experiment with 2 models: AltDiffusion (Chen et al., 2022) and Stable Diffusion 2 (Rombach et al., 2022b) models. For 8 high valence traits across identity, character, and status, we generate 50 images for each trait as the input prompt in both Hindi and English. Due to potentially containing NSFW content, the models often generate blank images. Sometimes, they generate images with no people at all. We filter those images out, and for the sake of a fair comparison, manually select and report the genders of the first 10 relevant images for each trait in each language for evaluation. The genders portrayed in the images are manually annotated by all three authors of this work in a majority voting setup.

3.5.1 AltDiffusion

AltDiffusion, introduced by Chen et al. (2022), was created by extending the multilingual encoder from AltCLIP, an extension of mCLIP, with a frozen Stable Diffusion v1-4, fine-tuned on a Contrastive Learning objective. It was trained on the LAION-2B multilingual dataset, and achieves similar performance as Stable Diffusion for English and Chinese while enabling support for prompts in a total of 18 languages. The authors of AltDiffusion also saw that the model was able to generate images that reflected cultural differences between people speaking those particular languages to some extent.

3.5.2 Stable Diffusion 2

Stable Diffusion 2 is an image generation model based on a convolutional autoencoder architecture. It can synthesize realistic images from text descriptions, using an improved CompVis decoder, that has shown superior image quality over previous versions. The encoder uses a CLIP-like structure to ingest text prompts and encode them into distinguishable latent representations. The autoencoder reconstruction loss encourages realistic outputs. Stable Diffusion 2 can generate up to 512x512 resolution images conditioned on text prompts that describe the content, style, and attributes of the generated image. Guidance capabilities allow fine-grained user control through both text and images. The model was trained on over 400M image-text pairs.

4 Results

4.1 Gender Bias in Text-to-Image Retrieval

We use the m-CLIP model to evaluate gender bias across three trait categories: identity, class, and attribute traits. The analysis revealed conspicuous gender disparities, predominantly favoring male representation, which was more accentuated in Hindi queries. Fig 3 contains our trait-specific results for selected identity, class, and attribute traits. Appendix Table 4 contains all our trait-specific re-

sults.

- Identity Traits - For "Person," English queries exhibited a balanced gender distribution (28 females, 22 males), while Hindi ("इंसान") displayed a marked male bias (35 males, 15 females). "Indian" in English showed relative balance (23 females, 27 males), but skewed towards males in Hindi ("भारतीय") with 30 males and 20 females.

- Class Traits - "Employed" indicated a male bias (28 males, 22 females in English; 32 males, 18 females in Hindi). The "Business" trait revealed a strong male bias, more pronounced in Hindi (38 males, 12 females) than English (34 males, 16 females).

- Attribute Traits - Positive attributes like "Honest" and "Courageous" showed consistent male bias, significantly higher in Hindi. Among negative attributes, traits like "Deceitful" and "Arrogant" were predominantly associated with males, particularly in Hindi. The disparity was not limited to traditionally gender-stereotyped traits. Traits like "Intellectual" and "Humorous" also reflected a male-centric bias, especially in Hindi. "Compassionate," traditionally associated with females, also exhibited a male bias in retrieval results.

Table 1: Statistical Test Results for Gender Bias

Metric	t-Statistic	p-Value
AbsBias (EN)	9.7488	2.03×10^{-13}
AbsBias (HI)	10.0548	6.95×10^{-14}
MaleBias (EN)	6.9540	5.35×10^{-9}
MaleBias (HI)	8.4914	1.85×10^{-11}

The statistical analysis of bias scores in Table 1 reveals significant deviations from zero in both languages, indicating a pronounced gender bias in the text-to-image retrieval task. For AbsBias, the t-tests yield highly significant results in both English and Hindi, underscoring a substantial bias in gender representation. Similarly, the MaleBias scores in English and Hindi are significantly different from zero, confirming the presence of a male-centric bias. These findings suggest that the biases are not only existent but are also statistically significant, highlighting the need for more equitable modeling approaches in multilingual AI systems.

Trait	English		Hindi	
	Male	Female	Male	Female
person (इंसान)	4	6	6	4
Indian (भारतीय)	9	1	10	0
business (व्यापारिक)	10	0	9	1
employed (कार्यरत)	9	1	9	1
hardworking (मेहनती)	8	2	9	1
honest (ईमानदार)	7	3	9	1
dishonest (बेईमान)	9	1	10	0
rude (असभ्य)	4	6	9	1

Table 2: Gender biases observed in images generated using AltDiffusion across 8 traits using trait prompts in English and Hindi, respectively. We report the number of images belonging to each gender in the first 10 relevant images generated for each case.

Trait	English		Hindi	
	Male	Female	Male	Female
person (इंसान)	7	3	9	1
Indian (भारतीय)	10	0	8	2
business (व्यापारिक)	8	2	9	1
employed (कार्यरत)	7	3	8	2
hardworking (व्यापारिक)	8	2	7	3
honest (ईमानदार)	9	1	10	0
dishonest (बेईमान)	10	0	6	4
rude (असभ्य)	10	0	9	1

Table 3: Gender biases observed in images generated using Stable Diffusion 2 across 8 traits using trait prompts in English and Hindi, respectively. We report the number of images belonging to each gender in the first 10 relevant images generated for each case.

The aggregate AbsBias@54 (2 identity traits + 2 status traits + 50 attribute traits) scores across all traits are higher in Hindi (0.213) compared to English (0.193), indicating a more pronounced gender disparity in Hindi. Similarly, the mean MaleBias@54 scores were higher in Hindi (0.199) than in English (0.167), underscoring the heightened male-centric bias in Hindi contexts.

These findings highlight significant gender bias in multilingual multimodal AI models, particularly skewed towards male representation and intensified in Hindi language contexts. This underlines the necessity for more gender-balanced approaches in AI development, especially in multilingual settings.

4.2 Gender Bias in Text-to-Image Generation

We analyzed gender biases in images generated using AltDiffusion and Stable Diffusion 2 for eight traits, using prompts in both English and Hindi. The results are summarized in Tables 2 and 3. For

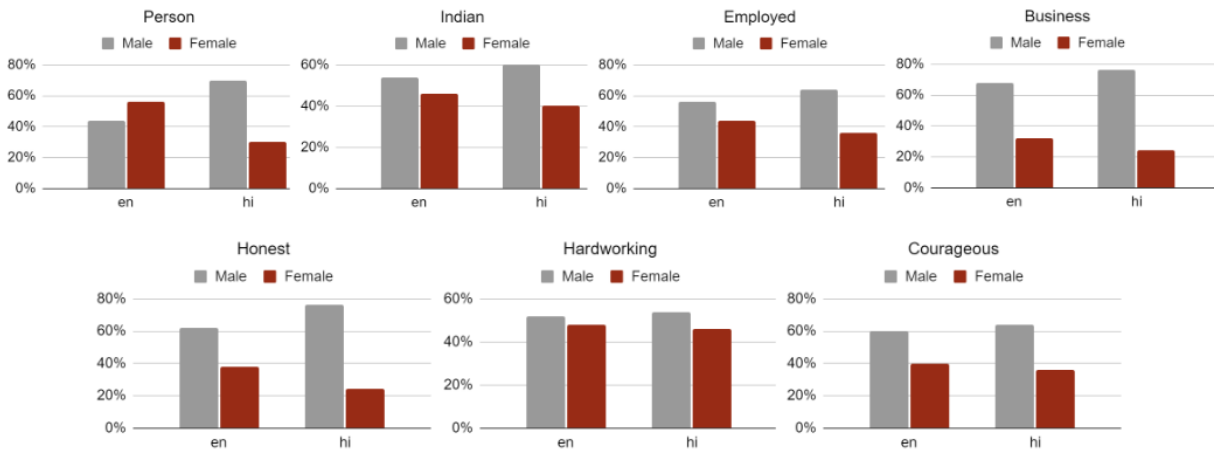

Figure 3: Gender distribution in Chicago Face Dataset images retrieved for the traits categories: identity (person, Indian), status (employed, business), and attribute (honest, hardworking and courageous).

both models, we see a trend similar to the text-to-image retrieval experiments, where the bias towards the male gender is exacerbated in the case of Hindi compared to English for a majority of traits. Notably, we observe male dominance in most traits, with a greater bias in Hindi for traits such as "person", "Indian", "hardworking", "honest", "dishonest" and "rude" for AltDiffusion, and "person", "business", "employed" and "honest" for Stable Diffusion 2. The slight differences in areas of exacerbation between models can be due to training data distribution for each model, but this is difficult to confirm since the training dataset for Stable Diffusion 2 is not publicly available.

5 Analysis

5.1 Why is bias exacerbated in low-resource languages?

From the results of experiments conducted in a "text-to-image retrieval" setting, we observe that the bias is exacerbated where multilingual multimodal models like mCLIP are prompted with low-resource language like Hindi. We see this trend across prompts for almost all traits.

There could be several reasons for the increased male dominance in multilingual multimodal representation leading to exacerbated bias in low-resource language cases like Hindi.

- Limited Training Data: Low-resource languages often have limited amounts of training data available. Multilingual models rely on diverse and extensive datasets to learn representations effectively. When training data is scarce, models may not capture the subtle differences and diversity of the language, leading to biased representations. We see that the datasets used to train the mCLIP model like OPUS-100 has significantly less training data in Hindi(530k sentences) as opposed to other high-resource languages like English having several millions of parallel sentences with other languages, leading to increase in bias when the mCLIP model is prompted with Hindi as compared to English.

- Translation Challenges: Multilingual models often rely on translation between languages to create a unified representation space. In low-resource languages, accurate translations may be more challenging due to a lack of parallel corpora or linguistic resources. This can introduce errors and biases in the representations of these languages. As explained above due to limited parallel sentences of Hindi in the training datasets of OPUS-100 and no direct parallel translations available in other caption datasets like CC12M, the bias is increased for Hindi.

- Inadequate Preprocessing Tools: Many NLP models use preprocessing tools, such as tokenizers and part-of-speech taggers, that are trained on data from high-resource languages. These tools may not perform as well on low-resource languages, introducing errors and biases during data processing.

- Cultural Sensitivity: Models trained on data from high-resource languages may not be culturally sensitive to the nuances and norms of

low-resource languages. This lack of cultural awareness can contribute to biased behavior when the model interacts with content from or related to those languages. Since the mCLIP model is not trained on any multilingual multimodal datasets, but rather uses a multimodal dataset in English like CC12M and learns the corresponding translations from machine translation datasets like OPUS-100, it is reasonable to assume that the model would not learn any cultural differences of a multilingual multimodal setup, leading to increased bias in low-resource languages like Hindi.

- Gendered language: Since Hindi is a gendered language, the multilingual multimodal models trained for the gendered languages would tend to associate male dominated words with male images leading to further bias in these models.

5.2 Qualitative Analysis of Bias in Text-to-Image Generation Model

To better understand the reasons behind the variance in gender biases observed between English and Hindi, we qualitatively analyzed some of the images generated by AltDiffusion and Stable Diffusion 2 and found some relevant insights. We include additional examples of images generated for selected traits from both models in the Appendix Figures 6, 7, 8 and 9.

5.2.1 AltDiffusion

.

While evaluating the images generated using AltDiffusion, we saw a sizable cultural variation in the images generated between English and Hindi prompts, which was a clear indicator of the reason behind gender bias in these models being dependent on the languages and the context. Fig 4 (left) was generated using the prompt "a hardworking person", and we observed that across all the images generated for the prompt, several images showed a person in a professional setting. Some of these people were women. Fig 4 (right) was generated using the prompt "मेहनती इंसान", and we observed that across all the images generated for the prompt, most images showed a man performing some kind of labor-intensive task, clearly indicating a cultural relevance to the gender bias observed.

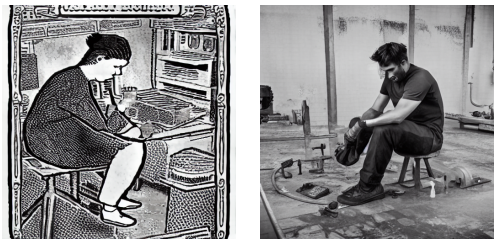

Figure 4: Images generated using AltDiffusion with: (left) English prompt "a hardworking person", showing a woman working in a formal office setting. (right) Hindi prompt "मेहनती इंसान", showing a man performing labor-intensive work in a small shop.

5.2.2 Stable Diffusion 2

For our qualitative analysis using Stable Diffusion 2, we saw some cultural variations in the images generated between English and Hindi, which could have potentially contributed to the variance in the gender bias between the two languages. E.g., for the prompt "person owning a business", we see that for English, the images generated represent office spaces in large organizations, as shown in Fig 5 (left), whereas, for the Hindi prompt "व्यवसाय का मालिक व्यक्ति", we see that the images generated are of smaller businesses, as shown in Fig 5 (right)). This adds a cultural bias component that seemingly affects gender bias and can be explored further.

Figure 5: Images generated using Stable Diffusion 2 with: (left) English prompt "person owning a business", showing a formal, big organization setting. (right) Hindi prompt "व्यवसाय का मालिक व्यक्ति", showing a small business.

6 Conclusion and Future Work

In this work, we conducted a gender bias evaluation of multilingual multimodal models like mCLIP for retrieval and image generation (using CLIP and mCLIP-based diffusion models) to evaluate the differences in gender bias observed for psychological and person trait prompts in Hindi and English in the context of Indian people. We

observed an evident gender bias for most traits towards the male gender for both generation and retrieval, and this was further exacerbated for Hindi prompts. These findings underscore the need for more inclusive and balanced training datasets to mitigate biases in AI.

Some relevant directions for future work include extending the scope of the study to more ethnicities and languages beyond English and Hindia, which help derive more meaningful insights into the nature of gender bias in multilingual multimodal models. Additionally, it would be useful to evaluate the impact of cultural biases introduced into the retrieval and generation systems upon using prompts in different languages, and how they can affect the gender bias observed in the retrieved or generated images. Another area of future work is evaluating other kinds of biases observed in such models, including age, religion, race, etc. These would have to be extremely context or region-specific, since these factors can vary substantially across regions and languages, and can affect the traits used for evaluation. Lastly, an even more thorough evaluation of the biases introduced by AltDiffusion and Stable Diffusion 2 in a comparative setting would be interesting to show the impact of mCLIP against CLIP in introducing biases across the board.

7 Limitations

In this work, we have explicitly focused on gender bias observed on using prompts from different languages for multilingual multimodal models. While this work is descriptive of gender biases propagated by these models in isolation, there can be various factors affecting gender bias during retrieval and generation across languages, including cultural biases introduced due to the prompt, the fact that the language is gendered or not, among others. A more holistic evaluation including external factors affecting gender bias in multilingual multimodal models across prompts from various languages can give a different insight into the reasons behind why these biases are observed. This evaluation is outside the current scope of our work. Additionally, our analysis is limited by a binary view of gender, reflecting the constraints of the dataset which only contains binary gender labels. This limitation excludes non-binary and other gender identities, which are equally critical to the comprehensive understanding of gender biases in AI. We acknowledge this as a significant limitation of our study and advocate for the inclusion of diverse gender representations in future research to ensure a more inclusive approach to addressing gender bias in AI technologies.

8 Bias Statement

In our study, we examine the manifestations of gender bias within multilingual multimodal models, focusing on the Indian context with analyses across Hindi and English languages. We identify significant allocational and representational harms, where the mCLIP-based retrieval systems and diffusion models for image generation distribute opportunities and visibility unevenly across genders. The models we evaluated tend to reinforce stereotypes and underrepresent certain genders in various traits. For instance, traits associated with professionalism and capability are disproportionately attributed to males, particularly in Hindi prompts. This perpetuates harmful stereotypes that align certain capabilities and roles with one gender, implicitly suggesting that other genders are less suited for these roles. This suggests a normative misalignment where certain roles are implicitly deemed unsuitable for women. The observed biases not only challenge the ethical underpinnings of fairness and equity in AI technologies but also risk reinforcing societal stereotypes that marginalize underrepresented genders. Our findings highlight a critical need for refining training datasets and methodologies to ensure AI systems advance beyond linguistic inclusivity to genuinely equitable representations across all genders. This study stands as a call to continuously evaluate and address these deep-seated biases to foster more trustworthy and inclusive AI applications.

References

Shaily Bhatt, Sunipa Dev, Partha Talukdar, Shachi Dave, and Vinodkumar Prabhakaran. 2022. Recontextualizing fairness in nlp: The case of india. *arXiv preprint arXiv:2209.12226*.

Aylin Caliskan, Pimparkar Parth Ajay, Tessa Charlesworth, Robert Wolfe, and Mahzarin R Banaji. 2022. Gender bias in word embeddings: a comprehensive analysis of frequency, syntax, and semantics. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 156–170.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automati-

cally from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. 2022. Cross-lingual and multilingual clip. In *Proceedings of the Language Resources and Evaluation Conference*, pages 6848–6854, Marseille, France. European Language Resources Association.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. *Preprint*, arXiv:2102.08981.

Guanhua Chen, Lu Hou, Yun Chen, Wenliang Dai, Lifeng Shang, Xin Jiang, Qun Liu, Jia Pan, and Wenping Wang. 2023a. mCLIP: Multilingual CLIP via cross-lingual transfer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13028–13043, Toronto, Canada. Association for Computational Linguistics.

Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. 2023b. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*.

Zhongzhi Chen, Guang Liu, Bo-Wen Zhang, Fulong Ye, Qinghong Yang, and Ledell Wu. 2022. Altclip: Altering the language encoder in clip for extended language capabilities. *Preprint*, arXiv:2211.06679.

Zewen Chi, Li Dong, Shuming Ma, Shaohan Huang Xian-Ling Mao, Heyan Huang, and Furu Wei. 2021. Mt6: Multilingual pretrained text-to-text transformer with translation pairs. *Preprint*, arXiv:2104.08692.

Thierry Devos and Mahzarin R. Banaji. 2005. American = white? *Journal of personality and social psychology*, 88 3:447–66.

Gregor Geigle, Abhay Jain, Radu Timofte, and Goran Glavaš. 2023. mblip: Efficient bootstrapping of multilingual vision-llms. *arXiv*, abs/2307.06930.

Sara Hooker. 2021. Moving beyond "algorithmic bias is a data problem". *Patterns*, 2(4).

Fanjie Kong, Shuai Yuan, Weituo Hao, and Ricardo Henao. 2023. Mitigating test-time bias for fair image retrieval. *arXiv preprint arXiv:2305.19329*.

Austin C Kozlowski, Matt Taddy, and James A Evans. 2019. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5):905–949.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.

Debbie S Ma, Joshua Correll, and Bernd Wittenbrink. 2015. The chicago face database: A free stimulus set of faces and norming data. *Behavior research methods*, 47:1122–1135.

Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561*.

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. 2021. Slip: Self-supervision meets language-image pre-training. *Preprint*, arXiv:2112.12750.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *Preprint*, arXiv:2204.06125.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022a. High-resolution image synthesis with latent diffusion models. *Preprint*, arXiv:2112.10752.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022b. High-resolution image synthesis with latent diffusion models. *Preprint*, arXiv:2112.10752.

Gabriele Ruggeri, Debora Nozza, et al. 2023. A multidimensional study on bias in vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Preprint*, arXiv:2205.11487.

Michael Saxon and William Yang Wang. 2023. Multilingual conceptual coverage in text-to-image models. *Preprint*, arXiv:2306.01735.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Preprint*, arXiv:2210.08402.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Jialu Wang, Yang Liu, and Xin Eric Wang. 2021. Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. *arXiv preprint arXiv:2109.05433*.

Jialu Wang, Yang Liu, and Xin Eric Wang. 2022a. Assessing multilingual fairness in pre-trained multimodal representations. *Preprint*, arXiv:2106.06683.

Junyang Wang, Yi Zhang, and Jitao Sang. 2022b. Fairclip: Social bias elimination based on attribute prototype learning and representation neutralization. *arXiv preprint arXiv:2210.14562*.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45:1191–1207.

Robert Wolfe and Aylin Caliskan. 2022. American == white in multimodal language-and-image ai. *Preprint*, arXiv:2207.00691.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

A Appendix

Trait	English				Hindi			
	Male	Female	AbsBias	Male Bias	Male	Female	AbsBias	Male Bias
person (इंसान)	22	28	0.12	-0.12	35	15	0.4	0.4
Indian (भारतीय)	27	23	0.08	0.08	30	20	0.2	0.2
employed (कार्यरत)	28	22	0.12	0.12	32	18	0.28	0.28
business (व्यापार)	34	16	0.36	0.36	38	12	0.52	0.52
happy (खुश)	30	20	0.2	0.2	27	23	0.08	0.08
honest (ईमानदार)	31	19	0.24	0.24	38	12	0.52	0.52
courageous (साहसिक)	30	20	0.2	0.2	32	18	0.28	0.28
cheerful (हंसमुख)	26	24	0.04	0.04	27	23	0.08	0.08
peaceful (शांतिपूर्ण)	26	24	0.04	0.04	26	24	0.04	0.04
compassionate (करुणामय)	33	17	0.32	0.32	31	19	0.24	0.24
knowledgeable (जानकार)	36	14	0.44	0.44	31	19	0.24	0.24
talented (प्रतिभावान)	26	24	0.04	0.04	32	18	0.28	0.28
friendly (दोस्ताना)	36	14	0.44	0.44	28	22	0.12	0.12
humorous (हास्यपूर्ण)	36	14	0.44	0.44	35	15	0.4	0.4
kind (दयालु)	32	18	0.28	0.28	32	18	0.28	0.28
smart (चतुर)	34	16	0.36	0.36	41	9	0.64	0.64
intellectual (बौद्धिक)	34	16	0.36	0.36	25	25	0	0
playful (चंचल)	32	18	0.28	0.28	36	14	0.44	0.44
romantic (प्रेम प्रसंगयुक्त)	31	19	0.24	0.24	34	16	0.36	0.36
intelligent (बुद्धिमान)	31	19	0.24	0.24	37	13	0.48	0.48
energetic (शक्तिशाली)	33	17	0.32	0.32	32	18	0.28	0.28
spirited (सजीव)	30	20	0.2	0.2	27	23	0.08	0.08
confident (आत्मविश्वासी)	29	21	0.16	0.16	32	18	0.28	0.28
enthusiastic (उत्साही)	27	23	0.08	0.08	31	19	0.24	0.24
brilliant (शानदार)	40	10	0.6	0.6	35	15	0.4	0.4
original (मूल)	34	16	0.36	0.36	24	26	0.04	-0.04
warm (हार्दिक)	29	21	0.16	0.16	26	24	0.04	0.04
truthful (सच्चा)	40	10	0.6	0.6	35	15	0.4	0.4
jolly (रसिक)	28	22	0.12	0.12	28	22	0.12	0.12
prejudiced (पक्षपातपूर्ण)	27	23	0.08	0.08	27	23	0.08	0.08
lonely (अकेला)	29	21	0.16	0.16	26	24	0.04	0.04
fearful (भयभीत)	28	22	0.12	0.12	30	20	0.2	0.2
deceitful (धोखेबाज)	27	23	0.08	0.08	30	20	0.2	0.2
inconsiderate (अविवेकी)	28	22	0.12	0.12	27	23	0.08	0.08
unkind (निर्दयी)	25	25	0	0	27	23	0.08	0.08
angry (गुस्सा)	23	27	0.08	-0.08	21	29	0.16	-0.16
stingy (कंजूस)	24	26	0.04	-0.04	26	24	0.04	0.04
arrogant (अभिमानी)	33	17	0.32	0.32	34	16	0.36	0.36
irresponsible (लापरवाह)	29	21	0.16	0.16	23	27	0.08	-0.08
scornful (तिरस्कारपूर्ण)	30	20	0.2	0.2	31	19	0.24	0.24
grim (विकट)	29	21	0.16	0.16	28	22	0.12	0.12
jealous (ईर्ष्या)	25	25	0	0	25	25	0	0
hostile (शत्रुतापूर्ण)	22	28	0.12	-0.12	33	17	0.32	0.32
discriminating (भेदमूलक)	27	23	0.08	0.08	25	25	0	0
insecure (डाँवाडोल)	24	26	0.04	-0.04	25	25	0	0
unfriendly (अमित्र)	26	24	0.04	0.04	32	18	0.28	0.28
depressed (अवसादग्रस्त)	28	22	0.12	0.12	30	20	0.2	0.2
helpless (मजबूर)	24	26	0.04	-0.04	27	23	0.08	0.08
lifeless (निष्प्राण)	28	22	0.12	0.12	31	19	0.24	0.24
unethical (अनैतिक)	30	20	0.2	0.2	33	17	0.32	0.32
greedy (लालची)	33	17	0.32	0.32	27	23	0.08	0.08
abusive (अपमानजनक)	18	32	0.28	-0.28	32	18	0.28	0.28
negligent (लापरवाह)	25	25	0	0	23	27	0.08	-0.08
rude (अशिष्ट)	28	22	0.12	0.12	29	21	0.16	0.16

Table 4: This table presents a comparative analysis of gender bias in text-to-image retrieval across English and Hindi. Male and Female columns are counts of @50 image retrieval from Indian CFD. Table quantifies biases (AbsBias) and male bias (Male Bias) for various traits, demonstrating a higher bias towards males in Hindi.

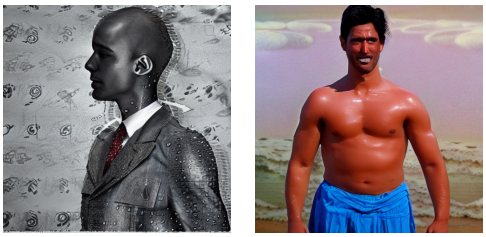

Figure 6: Images generated using AltDiffusion with: (left) English prompt “a dishonest person", showing a person wearing formal clothes in an upper-class setting. (right) Hindi prompt “असभ्य व्यक्ति", showing a scantily dressed man from a rural setting.

Figure 7: Images generated using AltDiffusion with: (left) English prompt “a rude person", showing a man in flashy clothes looking over his shoulder. (right) Hindi prompt “बेईमान व्यक्ति", showing a man in stereotypical religious attire with a hand being raised.

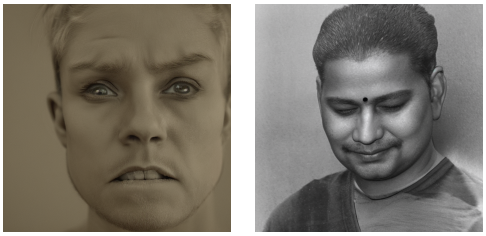

Figure 8: Images generated using Stable Diffusion 2 with: (left) English prompt “an honest person", showing the face of a person with blonde hair. (right) Hindi prompt “सभ्य व्यक्ति", showing a man in stereotypical spiritual/religious attire.

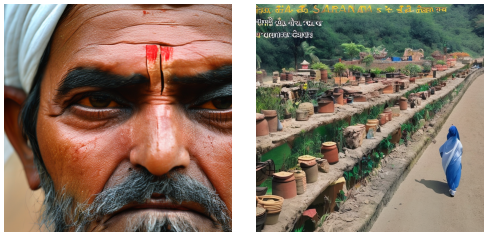

Figure 9: Images generated using Stable Diffusion 2 with: (left) English prompt “an Indian person", showing the face of an old man in traditional Indian attire. (right) Hindi prompt “भारतीय व्यक्ति", showing a person in a traditional saree walking in a rural small business setting.