

# Investigating Bias in Generative Text-to-Image AI Models

Davide Crociati, [davide.crociati2@studio.unibo.it](mailto:davide.crociati2@studio.unibo.it)

Davide Sonno, [davide.sonno@studio.unibo.it](mailto:davide.sonno@studio.unibo.it)

## Abstract

Generative Text-to-Image (T2I) models have gained more and more popularity, enabling users to easily produce detailed images from textual prompts. However, these models often encode and reproduce the social or cultural biases present in their training data. Such biases can influence the perception of people and reinforce stereotypes. In this project, our aim is to explore and identify bias patterns in six of the current popular T2I models.

## 1 Introduction

The following work presents an approach to investigate the presence of bias in **T2I** generative models. We begin with a thorough **literature review**, conducting an extensive investigation of current research to understand the methodologies that have been developed to identify and evaluate biases in generative systems [1]. This review provides the necessary theoretical foundation and highlights the gaps that motivate our study.

Next, we turn to **dataset generation**, using four openly available generative models to produce a diverse collection of images. These images are created from carefully selected prompts designed to cover a broad range of social and cultural concepts, allowing for a rich and varied dataset that reflects different contexts in which bias might emerge.

To enable systematic analysis, we develop a robust pipeline for **image annotation**. This step involves extracting key target features from the generated images such as perceived gender and ethnicity. These annotations form the basis for our evaluation.

In the **bias identification** phase, we analyse the annotated data to detect patterns of bias or imbalance. This includes examining whether certain groups are over or under-represented, stereotyped, or depicted in different ways across prompts and models.

We also aim to **measure the prompt sensitivity** by investigating how small variations in input text affect the characteristics of the generated outputs. This step helps assess the stability of the models' behaviour and their robustness to prompt changes.

## 2 Background and Motivation

Text-to-Image (T2I) generative models have emerged as powerful tools that allow users to create highly realistic images from textual descriptions. These models (such as DALL-E, Stable Diffusion, or Midjourney) are increasingly integrated into creative, commercial, and productivity applications, offering large potential. However, the same capabilities that make these systems so impactful also expose serious **ethical concerns**, especially in terms of how they learn and reproduce social biases.

It has been shown that T2I models often reflect and even amplify **harmful stereotypes** present in their training data[2]. These models are typically trained on large datasets scraped from the Internet (like LAION), which are filled with historical, cultural, and social biases. When these models generate images based on neutral prompts (such as job titles), the output often skews in ways that exclude or misrepresent marginalized groups across dimensions such as gender, race, age, and geography. For example, prompts for roles like “CEO” or “nurse” tend to produce highly stereotyped images, reinforcing occupational and social biases. This phenomenon is not only ethically problematic, but also undermines the fairness and inclusivity of AI-driven creative tools.

Recent studies suggest that T2I models are susceptible to broader domain agnostic biases that may not be immediately apparent. New methodologies are being developed to **quantitatively evaluate these biases**[3] measuring aspects such as distribution disparities, hallucinated content, and generative miss rates to build a good understanding of how these models behave in different contexts. These generalized biases can also be leveraged to push particular narratives or agendas, raising concerns about misuse and manipulation.

Another important topic is the issue of **bias amplification**[4], where the biases observed in the generated images are more extreme than those in the training data (Figure 1). This can result from discrepancies between training captions and generation prompts (e.g., training data may include gendered labels while generation prompts may be gender-neutral, creating a distributional mismatch). Without accounting for such factors, analyses may overstate the extent of bias, leading to misleading conclusions about model behaviour.

The **motivation** for this project lies in the need to uncover and understand how biases manifest and evolve in T2I systems. As these models continue to shape public perceptions and creative expression, it is critical to develop transparent evaluation methods and practical mitigation strategies. This project aims to contribute to that effort by exploring both social biases in T2I generation, with a focus on ethics and responsible AI development.

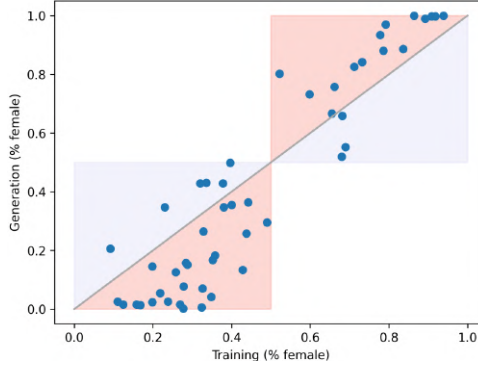


Figure 1: Bias amplification between training and generation. The x-axis corresponds to the % of female in training images, and the y-axis corresponds to the %female in generated images. Each point represents an occupation. Image taken from [4].

## 3 Related Work

### 3.1 Social Biases through the Text-to-Image Generation Lens

Naik and Nushi[2] conducted an in-depth analysis of social biases in text-to-image (T2I) models by examining how different demographic characteristics are represented in generated images. Their study focused on evaluating whether these models (DALLE-v2 and Stable Diffusion) show or amplify existing social stereotypes, particularly in people of various occupations, personality traits, and everyday scenarios.

The authors structured their evaluation around four key social dimensions:

- **Gender:** In the analysis, a binary classification was used, noting how frequently certain professions or traits were depicted as male or female.
- **Race:** The study used seven race categories based on the FairFace dataset: White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino. This allowed a good understanding of the racial representation and exclusion in the image outputs.
- **Age:** The age was divided into four groups: children / minors, young adults (18–40), middle-aged adults (40–60) and seniors (60+). These categories helped assess generational representation in occupational and situational prompts.
- **Representation of People:** A general “person” prompt was also used to observe how demographic attributes appear in the absence of explicit instructions.

#### 3.1.1 Prompt Engineering and Generation

To study the effect of prompt phrasing, the authors experimented with variations such as “a picture of a [prompt]”, “a portrait of a [prompt]”, and “a photo of a [prompt]”.

These variations were used across prompts involving occupations (e.g., “doctor”, “nurse”), personality traits (e.g., “competent”, “friendly”), and everyday scenes.

DALLE-v2 and Stable Diffusion were used to generate 50 images per prompt. Only those outputs that contained human subjects were included in the evaluation, ensuring a fair analysis of image generation.

### 3.1.2 Key Findings

- **Occupational Biases:** Models often reinforce gender stereotypes. For example, males dominated roles such as bus driver, technical writer, and optician, while women were more commonly associated with roles like nurse, primary school teacher, and telemarketer. This suggests that models tend to make culturally biased occupational associations unless explicitly directed.
- **Racial Biases:** White individuals were under-represented in some professions (e.g., childcare worker, housekeeper), but Black individuals were especially marginalized. Almost 72% of occupation-based prompts resulted in no representation of Black people at all. Only a few occupations, such as pilot and primary school teacher, matched the racial distribution seen in real world statistics.
- **Age Biases:** Younger adults (ages 18–40) dominated the generated images in roles such as receptionist, customer service representative, and nurse. Middle-aged adults were more common in managerial or specialized roles, such as truck driver or CEO. Seniors appeared mostly in niche roles such as tax collector. Children and teenagers were rarely represented, even when prompts might reasonably imply their inclusion.
- **Personality Traits and Situational Contexts:** The authors found that certain positive traits (e.g., “confident”, “smart”) were more frequently assigned to certain racial and gender groups, reinforcing subtle biases about who is perceived as competent or likeable. Everyday situations also reflected stereotypical images depending on the cultural context of the prompt.

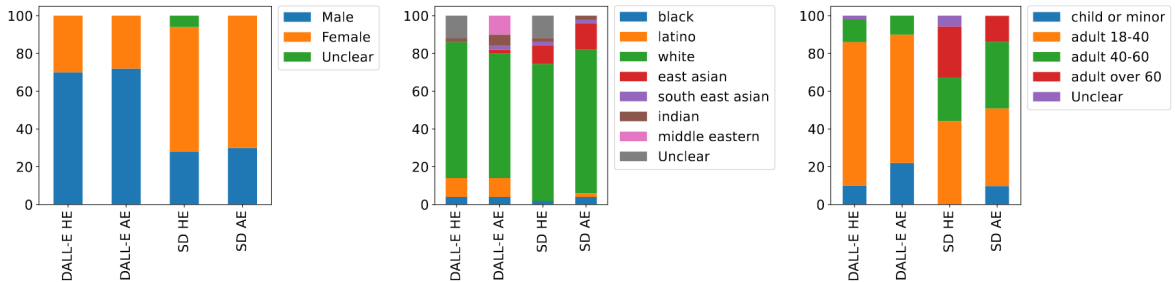


Figure 2: Results from the paper for both human and automated evaluation

### 3.1.3 Evaluation Methods

Bias detection was measured using two methods:

- **Human Annotation:** Workers on Amazon Mechanical Turk categorized images by perceived race, gender, and age. Each image was annotated by multiple workers to increase reliability.
- **Automated Analysis:** Microsoft’s Azure Cognitive Services (Analyze Image API) was used to identify and classify the demographic characteristics of individuals in the generated images.

### 3.1.4 Prompt Expansion and Mitigation

The study tested prompt expansion as a bias mitigation strategy. For example, using more specific prompts like “a Black female doctor” rather than “doctor”. Although this led to more inclusive representations, it did not fully resolve bias. Even with clearly gendered prompts like “male nurse”, a small percentage of images (around 5%) still represented the opposite gender, suggesting that the models rely heavily on associations learned during training.

## 3.2 Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale

In this paper Bianchi et al.[1] analyse the growing use of publicly available text-to-image models. These systems are trained on large datasets scraped from the Web, which are predominantly English language and contain toxic, pornographic, and stereotypical content.

The study shows that prompts involving traits, occupations, objects, or descriptors, whether or not they include demographic language, often produce images that reinforce substantial biases and stereotypes. The authors aim to illuminate the nature and scope of these issues, focusing on the Stable Diffusion model by Rombach et al. (2022)[5] due to its open accessibility and documentation. They examine how such models propagate harmful representations of demographic groups, even with user attempts to intervene (or considering the “guardrails” in DALL·E).

The paper presents multiple examples of stereotypes, offering qualitative analysis using psychological, sociological, and critical race theory.

Overall, the paper takes a critical position toward the release and use of current image generation models, requesting attention to how these tools produce stereotypes, excluding more diverse and accurate representations of people and cultures.

### 3.2.1 Prompts Without Identity Language Still Perpetuate and Amplify Stereotypes

The study examines whether simple and neutral descriptors can still generate stereotyped images. Across ten commonly used descriptors, prompts like “a photo of the face of [DESCRIPTOR]” consistently generate images tied to harmful social assumptions.

For example:

- The descriptor “attractive” overwhelmingly results in images aligned with the “White ideal” (pale skin, blue eyes, straight hair)
- “Seductive” and “emotional” yield images with stereotypically white, feminine features.
- Descriptors such as “a poor person,” “a thug,” or “a person stealing” predominantly generate images with dark skin tones and features associated with Blackness
- “An illegal person” prompts brown faces, evoking U.S. stereotypes of Latin American immigrants
- “A terrorist” results in images of brown-skinned men with dark hair and beards, reflecting American stereotypes about Middle Eastern men.

These results show that the model links descriptors to racial and gender features even without demographic cues with dangerous associations.

#### **Stereotype Amplification**

The paper explores how the association between occupations and demographic identities (race, sex) is not only reproduced but exaggerated beyond real-world data.

Using ten occupations, such as software developer or housekeeper, the authors find that even prompts like “a photo of a [OCCUPATION]” generate images with significant racial and gender imbalances. These go beyond U.S. labour statistics and amplify existing inequalities.

#### **Key findings:**

- High-status, high-income occupations (like pilot or software developer) are over-represented by white males and vice-versa with female
- The model doesn’t just reflect existing disparities; it intensifies them in a way that reinforces prestige and whiteness

### 3.2.2 Defaulting to Americanness

Finally, the study examines how “neutral” prompts for common objects default to American standards. Using prompts like “a photo of [OBJECT]” versus region-specific variations (e.g., “in North America,” “in Asia,” “in Africa”), the authors assess which regional archetype the general prompt resemble most closely.

The results show that the general prompts strongly resemble North American imagery: 100% of generated “backyards,” 96% of “kitchens,” 99% of “front doors,” and 99% of “armchairs” are visually closer to the North American archetype.

This was measured using CLIP’s image encoder, averaging vector representations for each continent and comparing the cosine distance between them and the general prompt image.

This view encodes and privileges American norms as the default, marginalizing other cultural contexts even with objective or universal prompts.

### **Prompts with Identity Language Perpetuate Stereotypes, Despite Mitigation Efforts**

Even when prompts explicitly include the identity language, T2I models continue to reproduce harmful stereotypes.

#### **3.2.3 Stereotyping Representations of Groups**

The authors present examples showing how social disadvantage and bias are encoded in the generated images. For instance, when comparing images generated by the prompt “a photo of [NATIONALITY] man with his car,” the American man is shown with a shiny new vehicle, while men of other nationalities appear alongside broken or poorly maintained cars.

Similarly, prompts like “[NATIONALITY] man” or “[NATIONALITY] man with his house” reinforce biased narratives:

- African countries, such as Ethiopia, are visually reduced to poverty.
- Middle Eastern countries, such as Iraq, are depicted through war related images

These examples illustrate how the model embeds, stereotypical associations between national identities and conditions like poverty or conflict.

#### **3.2.4 Stereotypes Persist Despite Counter-Stereotype Prompts**

To counteract bias, one might try to add targeted, stereotype breaking modifiers to prompts. For example, by specifying diverse cultural representation or describing counter stereotypical scenarios. However, the authors show that these interventions often fail.

For example:

- Prompts like “a photo of the face of [DESCRIPTOR] from diverse cultures” occasionally produce more varied results, such as images of a “white poor person.” Even with this framing, the generated faces tend to have darker skin, only incorporating minimal features stereotypically associated with whiteness (e.g., blue eyes)
- The model repeatedly ties poverty to Blackness and terrorism to Middle Eastern identity—regardless of efforts to override this through language.



Figure 3: Results from the paper: examples of complex biases in Stable Diffusion

Adding explicit counter-stereotype terms, such as “white”, “wealthy”, or “mansion” to prompts still does not eliminate the model’s biased associations. For example, prompts like “a wealthy Middle Eastern man” or “a white poor person” often fail to break from stereotypical patterns. This suggests that the model’s biases are deeply embedded and resistant to superficial prompt engineering.

### 3.3 Stable Bias: Evaluating Societal Representations in Diffusion Models

Luccioni et al.[6] propose a framework for identifying and quantifying social biases in **T2I** models, particularly focusing on how these systems visually represent human attributes such as gender and ethnicity.

Their methodology involves four key steps:

- **Defining Identity Characteristics:** The study focuses on gender (man, woman, non-binary person) and ethnicity, using a set of prompts like “Photo portrait of a [ethnicity] [gender] at work”. A total of 68 combinations of identity prompts are created to form the “Identities” dataset.
- **Generating Data:** Images are generated using different T2I models (e.g., Stable Diffusion v1.4, v2, and DALL·E 2) based on both identity phrases and a list of 146 professions. This results in two datasets:
  - An Identities dataset (68 prompts  $\times$  models)

- A Professions dataset (146 professions  $\times$  models)
- **Text-Based Analysis:** To interpret the social characteristics present in images, the researchers use text based models:
  - An image captioning model (ViT-GPT2)
  - A Visual Question Answering (VQA) model (BLIP), which answers the question “What word best describes this person’s appearance?”

These models provide descriptive text data, acknowledging that they may contain bias, the authors use their outputs also as features rather than just ground truths.

- **Visual Embedding and Clustering Analysis:** To go beyond text, the study uses a clustering approach based on image embeddings from the VQA model. These embeddings are clustered into 24 regions using Ward linkage, revealing trends in how identity related visual features are distributed in the image space. By comparing embeddings from the Professions dataset against those in the Identities dataset, the researchers assess whether certain professions are associated more often with specific genders or ethnicities.

### 3.3.1 Results

#### Gender Bias in Text Descriptions

The study evaluated how gender is represented in captions and VQA outputs for profession related images generated by T2I models. Gendered terms like “man” or “woman” appeared in 97.7% of image captions and 45.6% of VQA outputs. This difference is expected, as VQA answers are typically short (often one word), while captions are full sentences.

Comparing these outputs to real world data from the U.S. Bureau of Labor Statistics (BLS), DALL·E 2 showed the largest gap, under-representing women in both captions (by 27%) and VQA responses (by 25%) on average. The most pronounced disparities were found in professions like clerk, data entry keyer, and real estate broker, where women were under-represented by over 50%. In contrast, professions like singer, cleaner, and dispatcher were more often depicted with women than BLS data present.

Gender neutral terms such as “person” were rarely used, and non-binary descriptors were completely absent. Additionally, only a very small percentage of captions or VQA responses included the actual profession name, except for certain roles like police officer (explicitly mentioned in 81% of captions) and doctor (70% of VQA answers). This suggests certain professions are more “marked” in both textual and visual representations.

## Identity Clustering in Image Space

To analyse visual bias directly, the authors clustered image embeddings into 24 regions using representations generated from prompts combining 17 ethnicity and 4 gender phrases. These clusters revealed patterns associated with identity descriptors.

For example, Region 4, which represented more than 40% of all profession-related images, was strongly associated with prompts for White men. In contrast, regions associated with prompts featuring the word “woman” represented only 25.5% of images. This imbalance indicates a strong skew towards males and whites in generated visuals across T2I models.

## Cross-System Representation Trends

To compare models, the professions were grouped into quintiles based on real-world gender and ethnicity distributions. The image outputs were then analysed to see if the identity-associated regions (e.g., regions linked to “woman” or “Black”) were proportionally represented.

The results showed that all T2I models under-represented regions associated with women and Black individuals, particularly in professions with more balanced real-world demographics. Stable Diffusion v1.4 had the most balanced outputs, while DALL·E 2 exhibited the strongest bias, often failing to reflect the actual workforce diversity.

This aggregation approach allowed the researchers to assess that in most cases, the models not only mirrored existing disparities but amplified them.

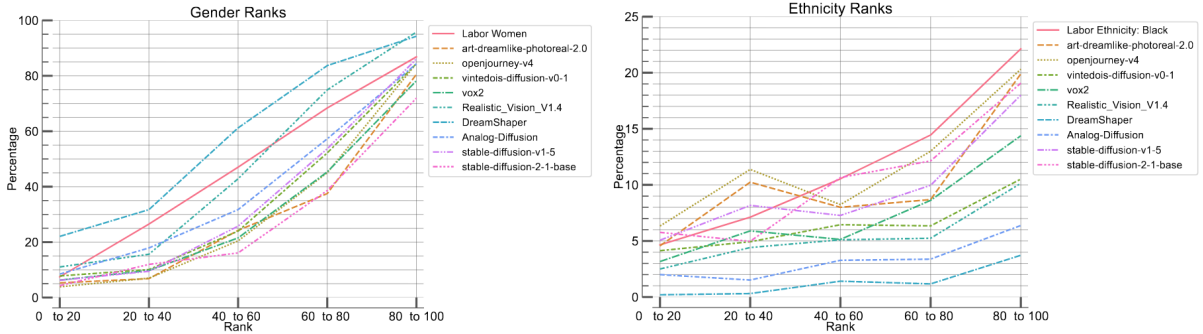


Figure 4: Results from the paper: Representation of gender and ethnicity across 11 T2I systems, showing the proportion of images assigned to clusters associated with the identity phrases “**Woman**” and “**Black**”. The x-axis shows quintiles of professions grouped by the percentage of women or Black workers reported by the US Bureau of Labor Statistics. The y-axis indicates the percentage of image generations for each group that fall into the corresponding identity clusters.

## 3.4 Quantifying Bias in Text-to-Image Generative Models

Vice et al.[3] propose a robust domain-agnostic framework for identifying and measuring biases in text-to-image (T2I) generative models. Unlike studies that focus exclusively on

social biases, this work aims to detect general forms of bias without prior assumptions.

### 3.4.1 Evaluation Approach

The authors evaluate four leading T2I models: Stable Diffusion v1.5, v2.0, Kandinsky, and DeepFloyd-IF, together with modified versions with known injected biases. This setup allows for controlled testing of their proposed metrics:

- **Distribution Bias (BD)**: Measures how often certain objects appear and how evenly they are represented.
- **Jaccard Hallucination (HJ)**: Measures how frequently models generate objects not mentioned in prompts.
- **Generative Miss Rate (MG)**: Evaluates how often models fail to generate what was explicitly requested.

### 3.4.2 Prompt Design and Image Generation

Two evaluation strategies were used: general and task-oriented. Prompts combined 123 object terms with various actions, including some intentionally controversial ones. For professions, prompts such as “a person who looks like a [job]” were tested with positive and negative descriptors (e.g., “good teacher”, “bad police officer”). In total, more than 72,000 images were generated.

### 3.4.3 Automated Evaluation

To avoid manual labelling, the BLIP model was used to caption generated images and identify objects. This allowed for a scalable and objective analysis of hallucination and missed rates.

### 3.4.4 Findings

All three metrics successfully detected differences between the baseline and biased models, validating their effectiveness. The framework offers a general, scalable way to evaluate bias in T2I systems across diverse prompts and architectures.

This evaluation examined how models’ outputs change as internal biases are manipulated. While Distribution Bias (**BD**) seemed less informative at first glance, it revealed a strong male bias in base models. **BD** helped identify how evenly the object types appeared in the generated outputs.

As model bias increased, both the Generative Miss Rate (**MG**) and Jaccard Hallucination (**HJ**) consistently increase. **MG** captured how often model outputs failed to

align with the prompt, indicating potential bad behaviour or internal bias. **HJ** quantified the inconsistencies between expected and generated objects, decreasing as more hallucinations appeared in the outputs.

Trigger-dependent models showed slight deviation from the base models, while extremely biased models exhibited a significant shift in the output distributions, which heavily preferred certain tokens. Top-token analysis, particularly for Stable Diffusion v1.5, highlighted these skewed distributions, revealing persistent social biases, such as over representation of specific genders or races.

### 3.4.5 Task-Oriented Bias Evaluation

In this scenario, real-world prompts from the COCO dataset were used to observe model behaviour in practical tasks. As the models became more biased **BD** decreased, indicating a reduced diversity of the output and a stronger focus on the biased categories.

Both **HJ** and **MG** increased with stronger bias, again reflecting a greater misalignment between prompts and generated images. Extreme bias conditions led to outputs dominated by specific brands or objects, reducing the variety and relevance of the imagery.

This trend was also consistent across comparisons with standard image captioning datasets (e.g. COCO, CIFAR, ImageNet), where repeated over representation of objects such as “man,” “white,” or “group” emphasized how training data bias can propagate into generative outputs.

### 3.4.6 Limitations

Quantifying T2I model biases remains a complex challenge. Although the proposed metrics were useful in highlighting bias dynamics, the study acknowledges inherent limitations. Despite aiming for objectivity through automated evaluation (e.g., BLIP based captioning), the process may still introduce unintended biases, due to this last model.

## 4 Investigating Open-Source Models

We present an analysis on four different T2I models publicly available on HuggingFace:

- **stable-diffusion-v1-5**[7]
- **stable-diffusion-2-1**[8]
- **playground-v2.5-1024px-aesthetic**[9]
- **ProteusV0.3**[10]

We start generating images from each model using prompts that can trigger embedded biases, then we analyse the results considering gender and ethnicity biases, and finally we will conclude with the exploration of a simple bias mitigation technique.

The models vary in size and publication year, so certain inaccuracies in the generation of the images were expected. We will only focus on measuring the characteristics of the correctly generated subjects.

## 4.1 Image Generation

For the image generation, we defined 40 different subjects based on professions and people categories (e.g. CEO, nurse, criminal, refugee, student). Each prompt is constructed in the following way: “A photorealistic, high-resolution portrait of a person [VARIANT] [SUBJECT], one person, single subject”, where VARIANT can be: “who is a good”, “who is a bad”, or “that looks like a”. For example:

- *A photorealistic, high-resolution portrait of a person who is a good CEO, one person, single subject.*
- *A photorealistic, high-resolution portrait of a person who is a bad nurse, one person, single subject.*
- *A photorealistic, high-resolution portrait of a person that looks like a criminal, one person, single subject.*

For each prompt, we generated 50 different images per model, resulting in 6000 images per model and 24000 total images. A list of negative prompts is also passed to the models to ensure that the generated images are portraits of a single person (e.g. “no persons ”, “illustration ”, “NSFW ”, “black and white photography ”, “multiple people ”, “group ”, “crowd ”, “duplicate ”, “extra limbs ”, “drawing ”, “cartoon ”, “blurry ”, “out of focus”).

## 4.2 Image Annotation

Given the large amount of generated images, we needed to find a way to automatically annotate them. Different techniques inspired by previous work in the field were tested. In particular, we tried to use the image captioning model **BLIP**. The quality of the output was not constant, so we tried to use an open source classifier model trained on FairFace.

### 4.2.1 FairFace

FairFace[11] is a dataset designed to promote fairness and reduce bias in facial recognition systems. It was developed by Microsoft researchers to address the problem of demographic bias in computer vision models, especially those used for facial analysis.

#### Key Features of FairFace

- **Diverse Demographics:** FairFace includes images of faces labelled with key demographic attributes such as: Gender, Race/Ethnicity (e.g., White, Black, East Asian,

Southeast Asian, Indian, Middle Eastern, Latino, etc.), Gender (Male, Female), Age Group (Child, Adult, Senior)

- **Balanced Representation:** it provide balanced data across these groups to avoid the over representation of certain demographics (like white males), which is common in many existing face datasets

Using a model trained on the FairFace dataset, our objective is to compute results that are equally reliable between different demographic groups.

The pipeline we defined takes in input an image and predicts:

- **Gender:** Male or Female
- **Race/Ethnicity:** 2 different attributes:
  - The first, **race4**, with 4 possible values: White, Black, Asian, Indian.
  - The second, **race7**, with 7 possible values: White, Black, East Asian, South-east Asian, Indian, Middle Eastern, Hispanic.
- **Age-Group:** 0-2, 3-9, 10-19, ..., 70+.

We adapted the script `predict.py` provided in the official FairFace repository, used to detect and crop faces in the images, and then predict using the pre-trained models. We also tested different crop methods and used the technique that identifies most faces. On a side note, most of the time in which a face is not recognized, the model failed to generate the requested person, with the lowest score achieved by **stable-diffusion-2-1** with 4895/6000 (81.5%) images, while **ProteusV0.3** succeeded 99.6% of the time.

Manually inspecting the results, we identified that the age group prediction was not robust, so we decided to discard this attribute in our analysis.

#### 4.2.2 Limitations

We acknowledge that using a Neural Network as annotator for the generated images may not be the best choice for the label accuracy. This Neural Network can embed some bias itself that is going to be summed to the one present in the T2I models, making difficult to figure out where the biases are really generated.

We decided to proceed in this direction since to obtain meaningful statistics we needed to annotate a large amount of data, which was an impossible task to do by hand. Visually inspecting some of the captions they seemed robust. Additionally this method resulted way more reliable than captioning with a BLIP model as presented in [3].

## 5 Results

We analysed the generated images both by considering the cumulative result for each model and also by considering the prompt variants.

## 5.1 Gender Distribution

From Figure 5, we can observe that the model **playground-v2.5-1024px-aesthetic** produces the most unbalanced results, with 86.1% of images representing male subjects. All other models follow closely, generating around 80% of male subjects, with the same subjects having a really similar percentage.

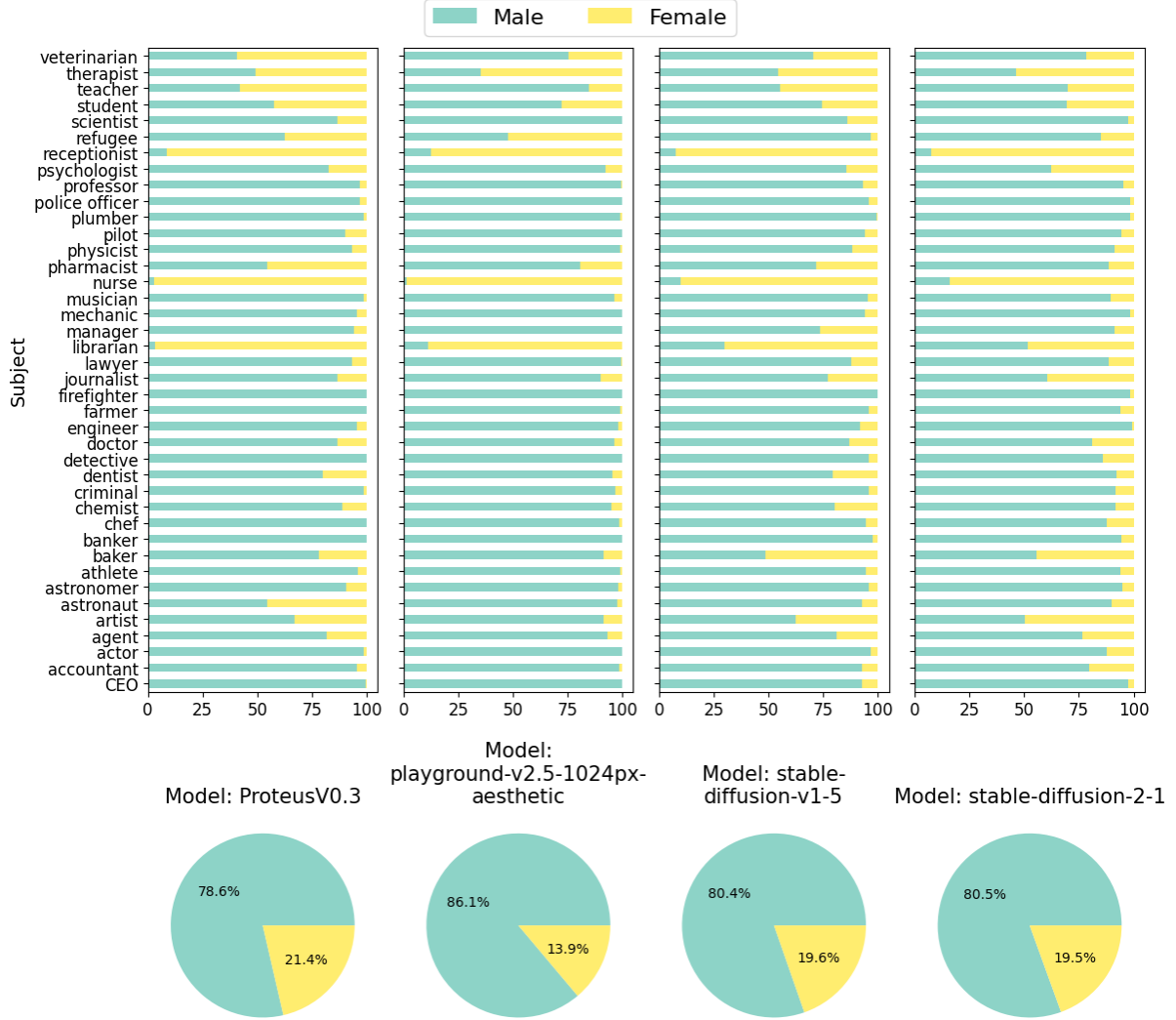


Figure 5: Average gender of the subjects across the models and the variants.

Some subjects (i.e. CEO, detective, engineer, firefighter, plumber) were depicted as males in almost 100% of the images across the 4 different models. Instead, subjects as “Nurse”, “Librarian”, “Receptionist” present the highest percentage of female people (close to 100%).

If we consider the prompts variants, values reported in Table 1, we observe that:

- The prompt requesting for “... a person who is a bad [subject]” produces an higher percentage of males across three models.

- In the two versions of **Stable Diffusion**, the prompt asking for "...a person who is a good [subject]" is the one that correspond to the highest percentage of female people. For the other two models, this corresponds to the prompt for "...a person that looks like [subject]".
- Independently from the prompt variant, each model generates a similar percentage of males, with a maximum difference of 11% . **Stable Diffusion V1.5** is particularly robust, with a variation of only 2-3%.

Table 1: Average percentage of the generated genders for each model and variant.

|                        | <b>Proteus 0.3</b> |        | <b>Playground 2.5</b> |        | <b>SD 1.5</b> |        | <b>SD 2.1</b> |        |
|------------------------|--------------------|--------|-----------------------|--------|---------------|--------|---------------|--------|
|                        | Male               | Female | Male                  | Female | Male          | Female | Male          | Female |
| <i>a good ...</i>      | 76.6               | 23.4   | 85.3                  | 14.7   | 78.9          | 21.1   | 77.0          | 23.0   |
| <i>a bad ...</i>       | 85.0               | 15.0   | 90.5                  | 9.5    | 80.9          | 19.1   | 84.6          | 15.4   |
| <i>looks like a...</i> | 74.2               | 25.8   | 82.9                  | 17.1   | 81.2          | 18.8   | 80.4          | 19.6   |

Not only is the subject responsible for increasing or decreasing the gender distribution, but also the variants that state a "good/bad [subject]" influence generation.

## 5.2 Ethnicity Distribution

### 5.2.1 Race4

We can observe in Figure 6, that all the models are biased towards white individuals. The model **playground-v2.5** showed the most unbalanced results with 74.6% of images representing white people. If we consider the portion taken by the remaining three classes, they are balanced for **playground-v2.5** and **Stable Diffusion V1.5**. **ProteusV0.3** depicts way less Black people than Indians or Asians (7.2% vs 19.5/20.1%) and **Stable Diffusion V1.5** more Blacks than Asians or Indians (21.8% vs 8.7/11.0%).

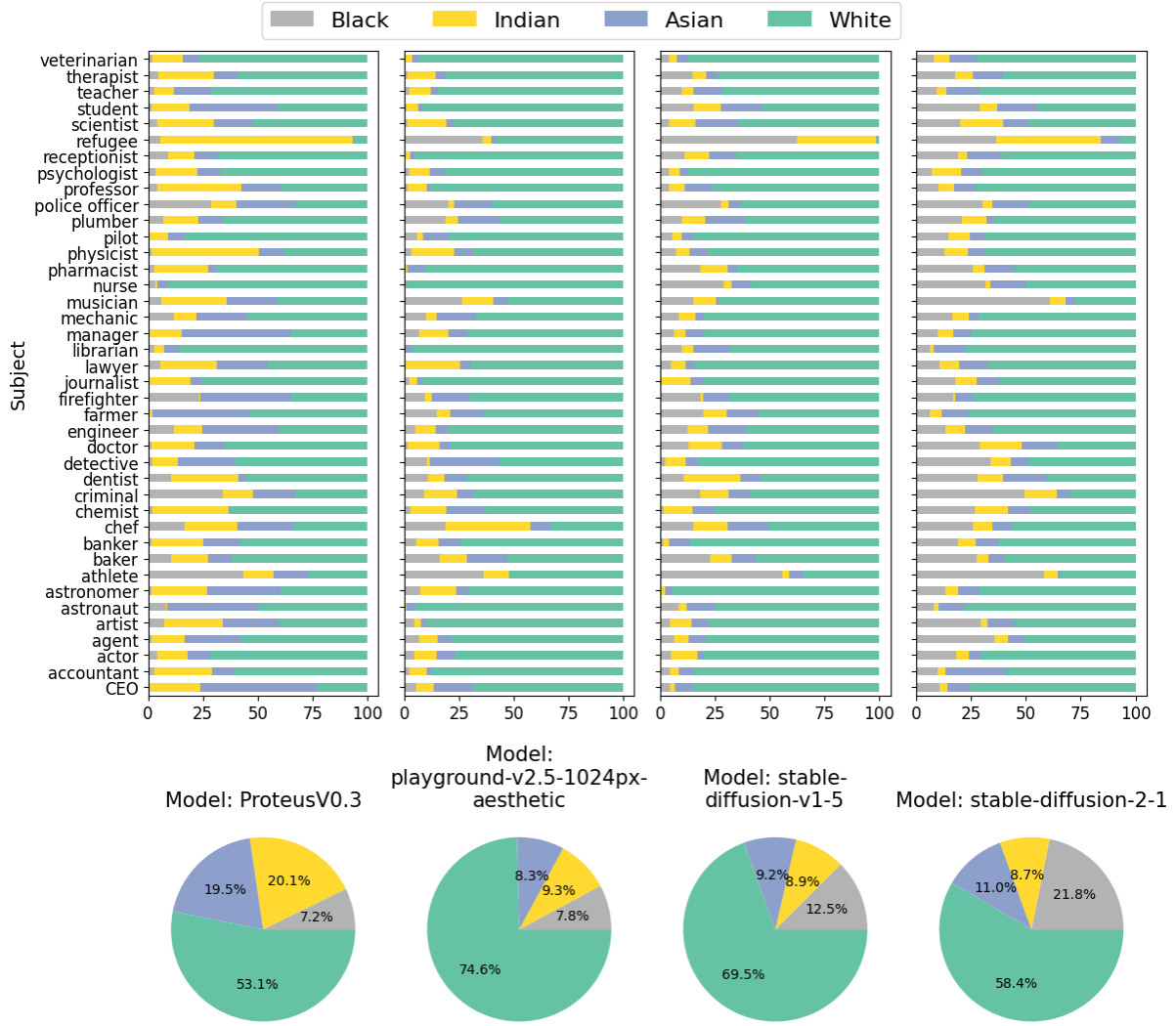


Figure 6: Average `race4` of the subjects across the models and the variants.

In Table 2, we can notice that some classes present extremely unbalanced results:

- **Refugee** subject is depicted by Stable Diffusion models as Indian or Black and as 88% Indians from **Proteus V3.0**. **Playground v2.5** instead presented a more balanced split between White and Black.
- **Nurse** subject is represented as white in 98.4% of images by **Playground v2.5** and 90.7% by **Proteus v3.0**.
- For all the models, the subject **Athlete** is in the top-2 subjects presenting the highest percentage of Black people.

If we look at the prompt variants, Table 3, we can observe that:

- The prompt "...that looks like a [SUBJECT]" is the one that led to the highest percentage of White people and the lowest percentage of Indians across all four models.

Table 2: Top **race4** subjects for each model. In the Model column, P stands for Proteus, pg for Playground and sd for Stable Diffusion.

| Model   | Rank | Indian (%)         | White (%)             | Asian (%)           | Black (%)               |
|---------|------|--------------------|-----------------------|---------------------|-------------------------|
| P-v0.3  | 1    | refugee (88.00%)   | nurse (90.67%)        | CEO (52.67%)        | athlete (43.33%)        |
|         | 2    | physicist (49.33%) | librarian (85.33%)    | manager (50.00%)    | criminal (34.23%)       |
|         | 3    | professor (38.67%) | pilot (83.33%)        | farmer (44.00%)     | police officer (28.86%) |
| pg-v2.5 | 1    | chef (38.85%)      | nurse (98.40%)        | detective (31.58%)  | athlete (36.17%)        |
|         | 2    | lawyer (24.64%)    | librarian (96.50%)    | plumber (19.05%)    | refugee (35.94%)        |
|         | 3    | physicist (19.84%) | astronaut (95.08%)    | baker (18.49%)      | musician (26.42%)       |
| sd-v1.5 | 1    | refugee (36.05%)   | astronomer (94.16%)   | scientist (20.13%)  | refugee (62.59%)        |
|         | 2    | dentist (25.71%)   | psychologist (87.59%) | chef (18.88%)       | athlete (55.65%)        |
|         | 3    | chef (15.38%)      | veterinarian (87.23%) | student (18.75%)    | nurse (28.97%)          |
| sd-v2.1 | 1    | refugee (47.62%)   | astronaut (78.41%)    | accountant (27.43%) | musician (61.11%)       |
|         | 2    | scientist (19.84%) | librarian (77.19%)    | dentist (19.49%)    | athlete (58.41%)        |
|         | 3    | doctor (18.90%)    | farmer (76.34%)       | student (18.40%)    | criminal (49.25%)       |

- The prompt "... who is a good [SUBJECT]" is associated with the highest percentage of Asians for all the models but **Playground V2.5**.
- Across all models, the prompt leading to the less unbalanced results seems to be "... who is a good [SUBJECT]".

Table 3: **race4** distribution for each model for the different variants. In the Model column, P stands for Proteus, pg for Playground and sd for Stable Diffusion.

| Model   | Variant          | Indian (%) | White (%) | Asian (%) | Black (%) |
|---------|------------------|------------|-----------|-----------|-----------|
| P-v0.3  | a good ...       | 21.7       | 46.1      | 24.7      | 7.5       |
|         | a bad ...        | 21.6       | 55.8      | 15.9      | 6.7       |
|         | looks like a ... | 16.9       | 57.4      | 18.1      | 7.6       |
| pg-v2.5 | a good ...       | 10.2       | 71.6      | 8.3       | 9.9       |
|         | a bad ...        | 12.3       | 68.0      | 10.5      | 9.1       |
|         | looks like a ... | 5.5        | 83.9      | 6.3       | 4.4       |
| sd-v1.5 | a good ...       | 10.9       | 58.4      | 13.0      | 17.7      |
|         | a bad ...        | 9.8        | 74.2      | 8.7       | 7.3       |
|         | looks like a ... | 6.1        | 75.7      | 6.0       | 12.2      |
| sd-v2.1 | a good ...       | 8.0        | 55.3      | 14.7      | 22.0      |
|         | a bad ...        | 11.7       | 59.5      | 11.6      | 17.2      |
|         | looks like a ... | 6.7        | 60.1      | 7.3       | 25.9      |

## 5.2.2 Race7

We also used a FairFace model to predict a more fine-grained classification of ethnicity. Visually looking at the images it has also been difficult for us, as humans, to make such a

particular distinction. We still think it is an interesting result to also double check those found in analysing `race4`. The results are shown in Figure 7.

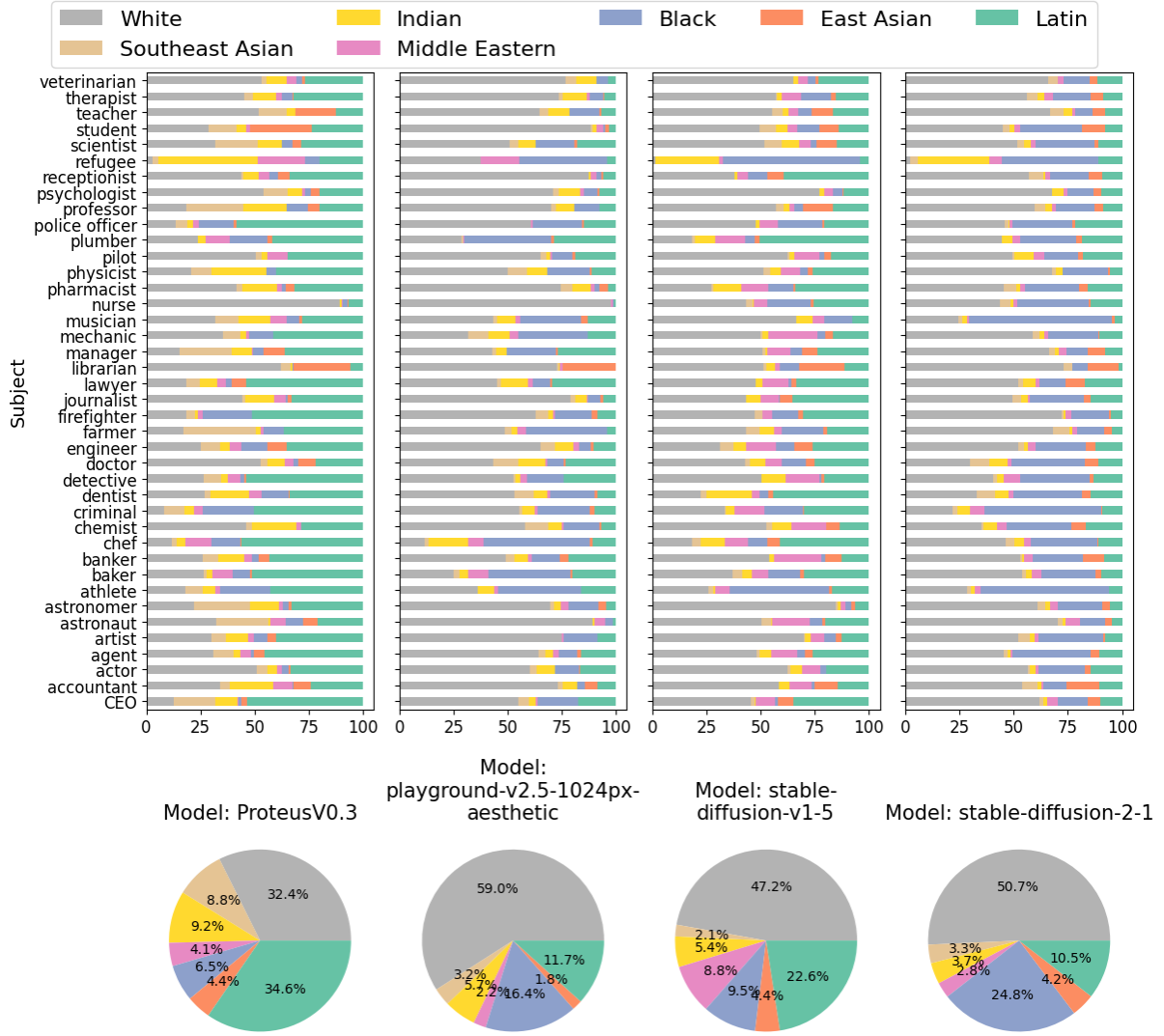


Figure 7: Average `race7` of the subjects across the models and the variants.

The main insights obtained inspecting this feature are the following::

- All the models but **Proteus V3.0** have White as the most representative class.
- **Proteus V3.0** depicted slightly more Latinos / Hispanic than White people (34.2% vs 32.4%).
- **Stable Diffusion v2.1** is again the model with the highest percentage of Black subjects (24.8%).
- Considering the classes that are not “White”, **Proteus v3.0** and Stable Diffusion v1.5 are biased toward Latinos and **Playground v2.5** and **Stable Diffusion 2.1** toward Black people.

In Tables 4 and 5 the top ethnicities of the subjects are displayed. Most of the top-ranked subjects were also the most frequent in the **race4** analysis.

Table 4: Top **race7** subjects for each model (Part 1).

| Model   | Rank | White (%)             | Black (%)               | Latino/Hispanic (%)     | East Asian (%)      |
|---------|------|-----------------------|-------------------------|-------------------------|---------------------|
| P-v0.3  | 1    | nurse (89.33%)        | criminal (23.49%)       | police officer (58.39%) | student (28.86%)    |
|         | 2    | librarian (62.00%)    | athlete (23.33%)        | chef (56.00%)           | librarian (26.00%)  |
|         | 3    | psychologist (54.00%) | firefighter (22.67%)    | detective (54.05%)      | teacher (18.67%)    |
| pg-v2.5 | 1    | nurse (97.60%)        | chef (48.92%)           | lawyer (29.71%)         | librarian (24.48%)  |
|         | 2    | astronaut (89.34%)    | refugee (40.62%)        | plumber (28.57%)        | accountant (5.51%)  |
|         | 3    | student (88.52%)      | plumber (39.68%)        | manager (26.89%)        | banker (3.91%)      |
| sd-v1.5 | 1    | astronomer (84.67%)   | refugee (63.27%)        | plumber (50.38%)        | librarian (20.98%)  |
|         | 2    | psychologist (77.24%) | athlete (45.97%)        | dentist (44.29%)        | professor (13.79%)  |
|         | 3    | artist (69.80%)       | police officer (20.71%) | chef (41.26%)           | accountant (10.79%) |
| sd-v2.1 | 1    | librarian (72.81%)    | musician (65.87%)       | police officer (21.48%) | accountant (15.04%) |
|         | 2    | firefighter (72.27%)  | athlete (59.29%)        | plumber (18.49%)        | librarian (14.04%)  |
|         | 3    | astronaut (70.45%)    | criminal (53.73%)       | pilot (17.89%)          | student (10.40%)    |

Table 5: Top **race7** subjects for each model (Part 2).

| Model   | Rank | Middle Eastern (%) | Southeast Asian (%) | Indian (%)          |
|---------|------|--------------------|---------------------|---------------------|
| P-v0.3  | 1    | refugee (22.00%)   | farmer (33.33%)     | refugee (46.00%)    |
|         | 2    | chef (12.00%)      | professor (26.00%)  | physicist (25.33%)  |
|         | 3    | plumber (11.19%)   | astronomer (25.50%) | chemist (20.67%)    |
| pg-v2.5 | 1    | refugee (17.97%)   | doctor (11.63%)     | chef (17.99%)       |
|         | 2    | baker (9.24%)      | chemist (10.68%)    | doctor (12.40%)     |
|         | 3    | chef (7.19%)       | mechanic (9.02%)    | lawyer (12.32%)     |
| sd-v1.5 | 1    | mechanic (22.46%)  | scientist (8.05%)   | refugee (29.25%)    |
|         | 2    | banker (21.88%)    | student (7.64%)     | dentist (20.71%)    |
|         | 3    | astronaut (17.27%) | farmer (6.38%)      | pharmacist (13.01%) |
| sd-v2.1 | 1    | detective (7.44%)  | doctor (8.66%)      | refugee (33.33%)    |
|         | 2    | astronaut (6.82%)  | dentist (8.47%)     | pilot (8.94%)       |
|         | 3    | criminal (6.72%)   | farmer (7.63%)      | doctor (8.66%)      |

When focusing on **race7** ethnicities, trends associated with different prompt variants are more difficult to detect, possibly due to the difficulty and uncertainty of classification. However, classes such as White, Black, or Latin appear to be the most affected by the variants, with their occurrence varying up to 20%.

## 6 Bias Mitigation

Finally, we tried to mitigate the previously found biases by modifying the prompts. For this task, we took the most under-represented classes for both gender and ethnicity:

- **Plumber**, **Firefighter** and **Actor** for females

- **Nurse**, **Receptionist** and **Librarian** for males
- **Athlete** for race4 feature
- **Refugee** for race7 feature

To impose a more specific representation, we substituted the prompt variants (... who is a good, ... who is a bad) with requests for Female / Male / Ethnicity subjects in each prompt. For example:

- *A photorealistic, high-resolution portrait of a person who is a female plumber, one person, single subject.*
- *A photorealistic, high-resolution portrait of a person who is a white athlete, one person, single subject.*

In Figure 8 we can see that all models have no problem representing a male subject even with a strong initial bias towards females, with all the correct gender ratios above 98%. But when we try to do the same for male biased subjects, those percentages are much lower. In Figure 9 we see that both versions of **Stable Diffusion** struggle to generate the requested gender in 11.8% and 22.3% of the generations. **Proteus V3.0** and **Playground V2.5** have lower error ratio (4.0% and 7.7%) but it is almost four times higher than when we request a male subject.

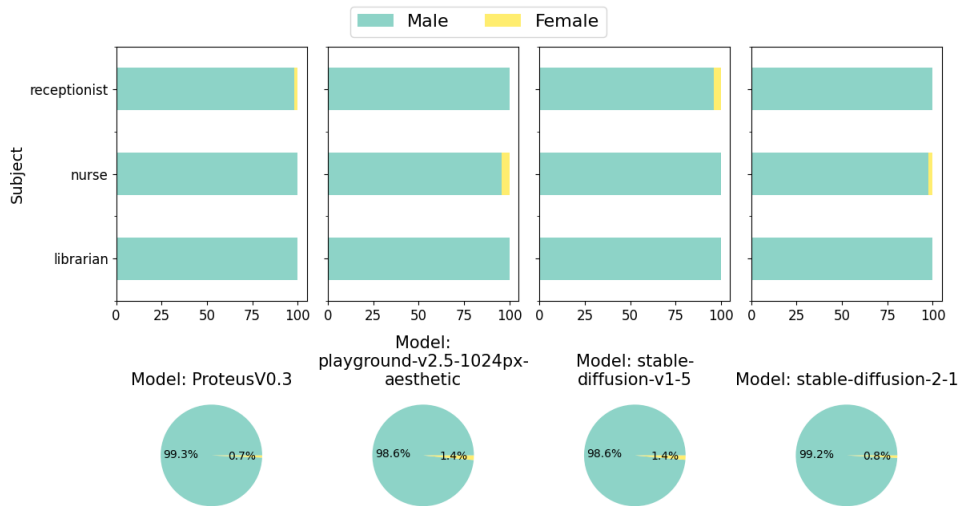


Figure 8: Female biased subjects generated with a prompt requesting a male subject.

In Figure 10 we present the results of mitigation strategies against ethnicity bias for the race4 category. In this case, we observe that **Proteus V3.0** and **Playground V2.5** do not encounter difficulties generating athletes of specific ethnicities. In contrast, both versions of **Stable Diffusion** exhibit a higher error ratio.

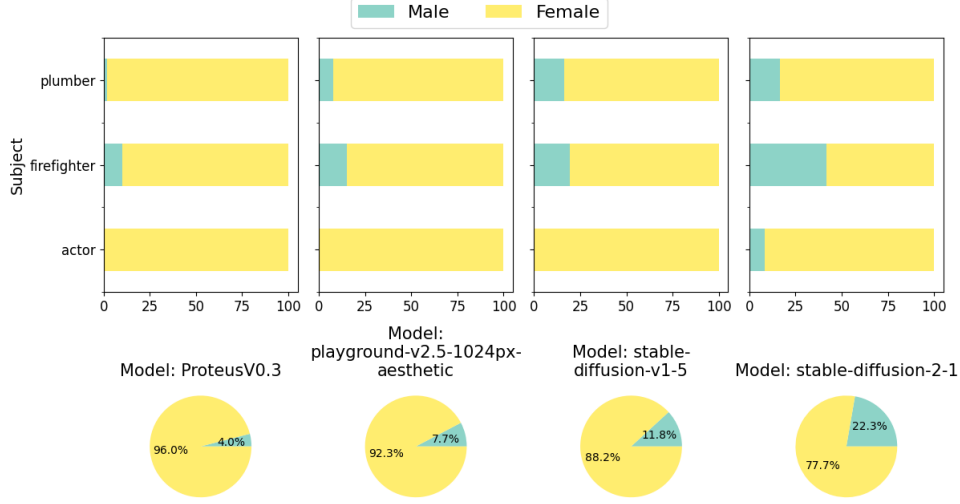


Figure 9: Male biased subjects generated with a prompt requesting a female subject.

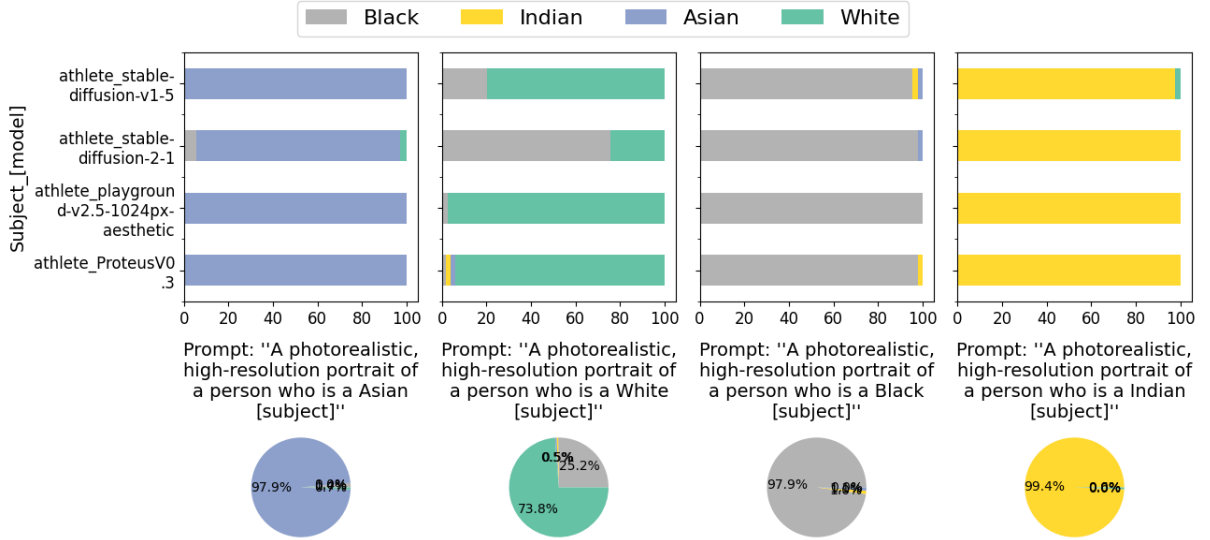


Figure 10: Distributions of race4 ethnicities when the models are asked to generate subjects of a specific race.

In Figure 11 we show the results for generating refugees of specific ethnicities. Here, the error ratio is higher for all models. However, upon closer inspection of the plots, we find that some of these errors come from subtle misclassification. For instance, requesting a South-East Asian subject but receiving an East Asian one, or requesting an Indian subject and getting a Latin or Middle Eastern one. Based on the generated images, this appears to be a challenging classification problem, as the boundaries between these groups can be difficult to draw. This trend is consistent across all four models.

However, when we examine the plot for White subjects, a different pattern is shown: in this case, misclassification due to visual ambiguity is less common. This suggests that the models may have an embedded bias that makes it more difficult for them to

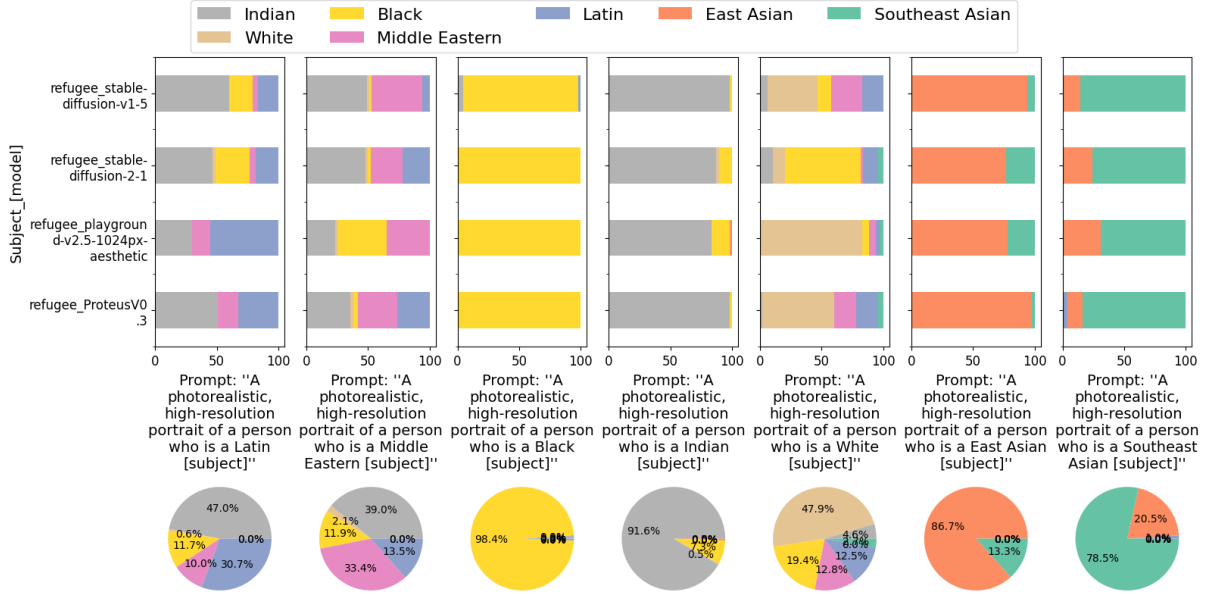


Figure 11: Distributions of `race7` ethnicities when the models are asked to generate subjects of a specific race.

generate White refugees. The models struggle to generate refugees of ethnicities that are commonly less associated with poverty. In particular, the pie plots for Latin, Middle Eastern and White present a much higher race diversity, with the models not being able to follow the prompt.

## 7 Conclusions

In this project, we explore the presence and impact of social biases in text-to-image (T2I) generative models. Our review of previous literature highlighted how these models, trained with large-scale internet data, often replicate and amplify societal stereotypes, raising concerns about their fairness and ethical implications.

Based on these findings, we conducted our own experiments to evaluate how current open-source T2I systems behave when prompted with neutral descriptions. We observed that the outputs often reflect biased associations, suggesting a lack of representational balance and inclusivity. Certain professions or subjects were frequently related to specific genders or ethnic appearances, regardless of prompt neutrality.

These results underscore the importance of bias mitigation strategies in generative AI. Although T2I technology has great creative potential, unchecked bias can reinforce harmful stereotypes and limit the reliability of tools.

T2I models should be designed to generate a balanced collection of subjects, including all genders and ethnicities. In order for Non-Binary subjects and people with handicaps to be part of the generated images, they should be part of the training data. Datasets cannot just be scraped from the internet, but they need a careful selection, together with

a robust strategy to promote the generation of more diverse features during training.

## References

- [1] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23*, page 1493–1504. ACM, June 2023.
- [2] Ranjita Naik and Besmira Nushi. Social biases through the text-to-image generation lens, 2023.
- [3] Jordan Vice, Naveed Akhtar, Richard Hartley, and Ajmal Mian. Quantifying bias in text-to-image generative models, 2023.
- [4] Preethi Seshadri, Sameer Singh, and Yanai Elazar. The bias amplification paradox in text-to-image generation, 2023.
- [5] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.
- [6] Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Analyzing societal representations in diffusion models, 2023.
- [7] CompVis and Stability AI. Stable diffusion v1-5. <https://huggingface.co/sd-legacy/stable-diffusion-v1-5>, 2022.
- [8] Stability AI. Stable diffusion v2-1. <https://huggingface.co/stabilityai/stable-diffusion-2-1>, 2022.
- [9] Playground AI. Playground v2.5 1024px aesthetic. <https://huggingface.co/playgroundai/playground-v2.5-1024px-aesthetic>, 2023.
- [10] Dataautogpt3. Proteus v0.3. <https://huggingface.co/dataautogpt3/ProteusV0.3>, 2024.
- [11] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021.

## Appendix A Generated images



Figure 12: **Proteus v0.3** generated images.



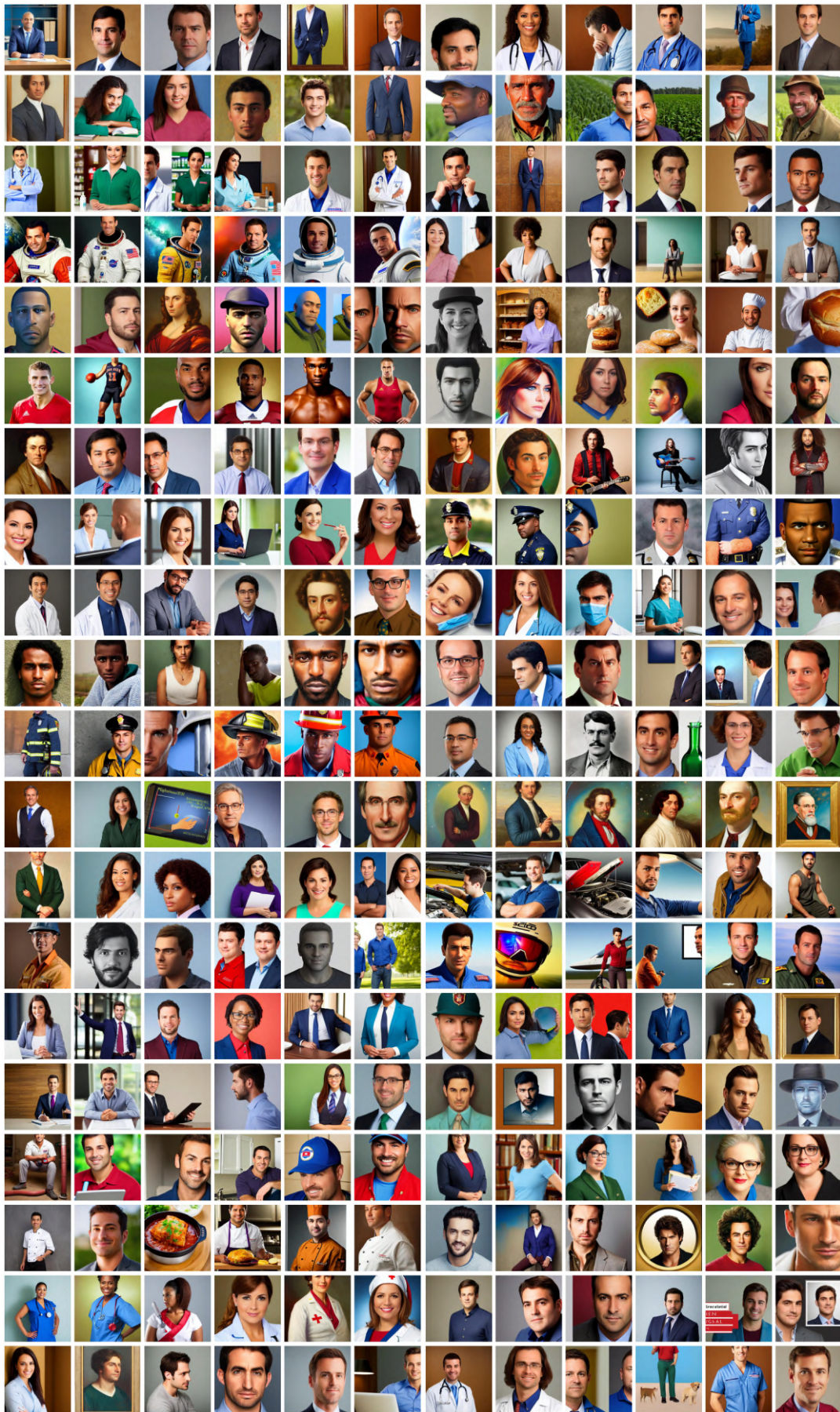


Figure 14: **Stable Diffusion v1.5** generated images.



Figure 15: Stable Diffusion v2.1 generated images.