Revealing Gender Bias in Language Models through Fashion Image Captioning

María Villalba-Osés
University Institute for
Computing Research (IUII)
University of Alicante
maria.villalba@ua.es

Victoria Muñoz-García
University Institute for
Computing Research (IUII)
University of Alicante
victoria.munoz@ua.es

Juan Pablo Consuegra-Ayala
Digital Intelligence Centre
(CENID)
University of Alicante
juan.consuegra@ua.es

Abstract

Image captioning connects computer vision and natural language processing but remains prone to social biases. This study examines gender bias in ChatGPT, Copilot, and Grok by analyzing their descriptions of Spanish fashion images prompted without gender cues. We propose a methodology combining gender annotation, stereotype classification, and a manually curated dataset. Results show that GPT-4o and Grok frequently assign gender and reinforce stereotypes, while Copilot produces more neutral captions. Grok achieves the lowest error rate but still consistently attributes gender, even when cues are ambiguous. These findings underscore the need for bias-aware captioning strategies in multimodal systems.

1 Introduction

Automatically generating natural language descriptions of images has gained increasing attention due to its relevance in practical applications and its connection to two key artificial intelligence fields: computer vision and natural language processing. Image captioning plays a crucial role in accessibility and serves as a benchmark for multimodal reasoning (Zhao et al., 2021). However, as with many machine learning applications, it remains vulnerable to social biases—particularly those related to gender and race (Hirota et al., 2022; Bhargava and Forsyth, 2019)—which can distort descriptions, reinforce stereotypes, and potentially cause harm when portraying people. These biases typically stem from two sources: imbalanced datasets, which mirror societal stereotypes, and models themselves, which may amplify such patterns during training and inference. This paper proposes and validates a methodology to examine gender bias in image-to-text generation, focusing on Spanish. We analyze outputs from three large language models—ChatGPT, Copilot, and Grok—when prompted with fashion-related images and instructed to describe the depicted person without specifying gender.

The main contributions of this research can be summarized as follows:

- The development of a methodology for assessing gender representation in textual descriptions generated from gender-neutral prompts applied to fashion images.
- The introduction of an annotation framework capturing source gender, gender stereotype, stereotype class, and LLM-inferred gender to enable a more detailed analysis of gender bias in image-based text generation.
- The construction of a manually annotated dataset consisting of fashion images and model-generated descriptions.
- The evaluation of three state-of-the-art language models—ChatGPT, Copilot, and Grok—in the context of image-to-text task.

By integrating a multimodal approach combining text and image analysis, this study offers insights into how generative models interpret and depict gender.

The remainder of this paper is structured as follows: Section 2 reviews prior research on gender bias in multimodal models. Section 3 details the methodology, including dataset construction, prompt design, annotation, and evaluation. Section 4 presents the results and discussion, and Section 5 concludes with a summary and future research directions.

2 Related Work

Image captioning models have advanced in generating natural language descriptions of visual inputs

Proceedings of Recent Advances in Natural Language Processing,pages 1333–1340
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-155

but face a major limitation: the reproduction and amplification of social biases from their training data, raising concerns about fairness and stereotype reinforcement.

2.1 Image Captioning Models

Image captioning translates visual information into coherent textual descriptions, bridging computer vision and natural language processing. Early rule-based systems have evolved into neural architectures that combine visual encoders and language decoders (Ghandi et al., 2023; Stefanini et al., 2023), often enhanced with attention mechanisms to focus on semantically relevant regions (Al-Malla et al., 2022). Recent advances include controllable captioning (Zhao et al., 2024), which mitigates degeneration and improves coverage of less frequent concepts.

General-purpose large language models (LLMs) like ChatGPT, Copilot, and Grok now integrate multimodal capabilities, enabling image captioning despite not being explicitly trained for it. However, trained on large uncurated datasets (Birhane et al., 2021), these models behave as black boxes and frequently replicate social biases (Bansal et al., 2022; Cho et al., 2023), raising concerns about how people are represented in visual content.

2.2 Bias in Image Captioning

Bias in captioning systems often mirrors societal stereotypes encoded in training data (Hirota et al., 2022), with documented cases of racial (Zhao et al., 2021) and gender bias (Hirota et al., 2025). This is particularly evident in clothing descriptions, where garments, colors, or styles are stereotypically associated with specific genders. Imbalanced datasets and web-scraped pairs exacerbate the issue, leading to misgendering and reinforcing restrictive norms (Bhargava and Forsyth, 2019).

Efforts to address bias include fairness-aware captioning frameworks (Desai, 2024) and controllable captioning strategies (Zhao et al., 2024). However, multimodal LLMs inherit implicit biases from massive training corpora (Wang et al., 2021; Feng and Shah, 2022; Kay et al., 2015; Consuegra-Ayala et al., 2024), often defaulting to stereotypical associations based on visual cues like hairstyle or clothing. To counter adversarial manipulation and improve robustness, prompt-shielding frameworks such as PromptShield have been proposed (Jacob et al., 2025).

2.3 Our Contribution

In this work, we analyze these gender biases in the fashion domain, where appearance is central and the visual diversity should be preserved. Compared to previous approaches, our contribution is distinguished by adopting a transversal and multi-perspective approach within the Spanish context. Analyzing how captioning models interpret and describe clothing across genders helps reveal the extent to which generative AI perpetuates social stereotypes. Furthermore, it provides critical insights into how these systems could be refined to promote more inclusive and fair representations in multimodal applications.

3 Method

This section describes the methodology for evaluating gender bias in image-to-text generation. It covers dataset creation—including image selection, caption generation, and annotation—and bias quantification through gender attribution, disparity, and stereotype analysis.

3.1 Image-to-Text Dataset Creation

A dataset is constructed from a collection of source images. To mitigate potential memory-related biases, each interaction with each model is limited to a single request per chat session, and global memory across requests is disabled. For each image, a new chat is started and the page is refreshed before inserting the next image. If the language model does not explicitly mention any gender in the description, the instance is classified as “neutral” in the annotation.

A dataset is constructed from a collection of fashion images. To avoid memory-related biases, each interaction with a model is performed in isolation: a new chat session is initiated and refreshed for every image. If a model’s caption omit explicit gender references, the instance is annotated as *neutral*.

3.1.1 Source Images

The images are sourced from a Kaggle dataset for fashion classification[1], which mainly features Western-style clothing. We select a balanced subset of 50 images—25 from the “men” category and 25 from the “women” category—ensuring diversity in poses, styles, and contexts. The specific subset is

[1] https://www.kaggle.com/datasets/engrmwaqasniazi/mens-and-womens-images-for-fashion-classification

publicly available[2], and an example is shown in Figure 1 (see Supplementary Materials).

3.1.2 Text Generation

To generate captions, we prompt three LLMs: GPT-4o (DALL-E), Copilot (Microsoft Designer), and Grok (xAI). For each image, two captions are produced per model, yielding a total of 300 captions. For ChatGPT and Grok, we use the prompt: *"Haz una descripción de la persona que se muestra en la imagen"* ("Provide a description of the person shown in the image"). For Copilot, we adapt it to: *"Haz una descripción del hombre o mujer que se muestra en la imagen"* ("Provide a description of the man or woman shown in the image") because the neutral version consistently generates gender-neutral captions. This adjustment reveals a potential weakness in existing bias mitigation mechanisms, as small prompt variations can alter gender attribution. Despite this change, Copilot still produces the most neutral outputs.

Each of the 50 selected images is processed twice with each of the three language models, resulting in two independently generated captions per image and per model. In total, the dataset consists of 300 captions —100 generated by each model. Generating two captions per image allows us to observe possible variations in the models' outputs even under identical conditions, and helps identify potential inconsistencies or recurring patterns in gender attribution.

3.1.3 Data Organization

The image files used in the study are named sequentially from `01.jpg` to `50.jpg`. This naming allows for consistent referencing across the annotation and analysis processes, while maintaining the neutrality of the file identifiers.

3.1.4 Annotation Procedure

Each of the 50 images and their captions are manually annotated to analyze gender attribution and stereotypes. Two expert annotators label the dataset, and a third resolves disagreements to ensure neutrality. The following annotations are recorded:

- **Source Gender:** Actual gender category of the person, based on dataset origin (*female* for images 01–25, *male* for 26–50).
- **Gender Stereotype:** Whether the person's appearance aligns with stereotypical gender associations, labeled as *male*, *female*, or *neutral*.
- **Stereotype Class:** If a stereotype is present, the visual basis (e.g., color, clothing type, hairstyle, footwear, accessories) is identified.
- **Caption Gender Attribution:** Gender assigned by the model in each caption, labeled as *male*, *female*, or *neutral*.

This annotation scheme supports direct comparison between source and described gender, and reveals how visual stereotypes influence gender attribution across models.

3.2 Bias Quantification Strategy

To evaluate gender bias in image-to-text generation, we design an analysis framework based on the annotated dataset introduced in Section 3.1. Each image is classified by source gender, visual stereotypes, stereotype class, and caption-inferred gender, as well as a higher-level concept called *situation type*, enabling a comparative analysis of model behavior under different stereotype dynamics.

3.2.1 Situation Categorization

Each image is classified into one of three *situation* types based on the relationship between source gender and visual stereotypes:

- **Aligned**: Source gender matches the stereotype (e.g., a male image with male-stereotyped cues).
- **Defiant**: Source gender contradicts the stereotype (e.g., a female image with male-stereotyped cues).
- **Non-stereotypical**: No clear gender association (stereotype annotation is neutral).

These categories form the basis for subsequent bias quantification (see Figure 2 in Supplementary Materials).

3.2.2 Analytical Framework

Our bias analysis considers two complementary scenarios reflecting key functions of LLMs in image-to-text tasks: (I) their reliability as gender annotators, and (II) their tendency to reproduce stereotypes when generating captions.

I. LLMs as Gender Annotators In the first scenario, we assess whether LLMs can consistently infer gender by comparing their outputs to the annotated source gender. The procedure, summarized

[2] https://www.kaggle.com/datasets/mariavillalbaoses/lphhuuuu

in Algorithm 1 (see Supplementary Materials), aggregates the two captions per image using decision rules that prioritize non-neutral attributions. Final outputs are classified as:

- **Match** (✓): Inferred gender matches source gender.
- **Mismatch** (×): Inferred gender differs from source gender.
- **Neutral** (∼): No gender is assigned.

This categorization allows us to evaluate alignment or divergence between model-generated captions and the annotated ground truth.

II. LLMs as Conversational Agents The second scenario, summarized in Algorithm 2 (see Supplementary Materials), examines stereotypes in captions when no explicit gender is prompted. For each image, the two caption-gender attributions are compared to the annotated stereotype label. Each caption is labeled as: *Stereotypical* if its inferred gender matches the stereotype, *Counter-stereotypical* if it contradicts the stereotype, or *Neutral* if no gender is attributed. This analysis reveals whether models reinforce, challenge, or bypass gendered associations present in visual cues.

This enables us to analyze the models' tendencies to reinforce, challenge, or bypass gendered expectations associated with the visual content.

3.2.3 Evaluation Protocol

For both analytical scenarios, we apply a consistent evaluation methodology at three levels:

Global Distributions: Frequency of each classification (e.g., match, mismatch, stereotypical) across situation types and source gender.

Directional Distributions: Analysis of gender mismatches and stereotype alignments (e.g., male source → female attribution), stratified by situation type, to detect asymmetric patterns.

Stereotype-Class Feature Analysis: Examination of how visual features (e.g., clothing, hairstyle, footwear, accessories) correlate with inferred gender and stereotype directionality. Details are provided in Supplementary Materials, Section A.

4 Experiments

This section presents the findings from our experimental evaluation. Section 4.1 summarizes the results, while Section 4.2 discusses their implications.

4.1 Results

The results are organized into two scenarios: (i) model performance in gender-annotation tasks (Section 4.1.1) and (ii) gender-related associations and stereotypes in captions (Section 4.1.2).

4.1.1 Performance in Gender-Annotation Tasks

Table 1 summarizes the performance of each model in gender-annotation tasks. Rows *aligned*, *defiant*, and *non-stereotypical* stand for situations in which the source gender and predefined stereotype (as observed in the source image) coincide, do not coincide, or the stereotype is neutral, respectively. Columns *Match*, *Mismatch*, and *Neutral* indicates the percentage of the number of times that the source and LLM-annotated gender match, mismatch or the annotation is neutral. The equivalent of that table presenting counts can be found in Supplementary Materials, Table 7.

For a more detailed view, Table 2 subdivides results by source gender and stereotype type, enabling analysis of specific combinations, such as male subjects with female-coded stereotypes.

Finally, Table 3 classifies results by visual features (e.g., *prenda* (garment), *pelo* (hair), *zapatos* (shoes)), showing how specific attributes influence inferred gender.

4.1.2 Stereotypes in Image Caption Tasks

Table 4 summarizes model performance in identifying gender-related stereotypes in captions. Rows *aligned*, *defiant*, and *non-stereotypical* indicate whether the source gender matches, contradicts, or is unrelated to traditional stereotypes. Columns *Stereotypical*, *Counter-stereotypical*, and *Neutral* show the percentage of captions that reinforce, challenge, or avoid gender associations. Because two captions are generated per image, percentages may exceed 100% when captions fall into different categories. Full count data are provided in Supplementary Materials.

Table 5 further analyzes aligned, defiant, and non-stereotypical situations by source gender and associated stereotype, showing where stereotypical, counter-stereotypical, or neutral annotations occur across gender combinations.

Table 6 shows results classified by visual features where an element associated with one gender appears on a person of another. Columns *Feature*, *Looks-Like*, and *Source* indicate the triggering element, its associated stereotype, and the person's

Category	Count	GPT-4o			Copilot			Grok		
		✓	×	∼	✓	×	∼	✓	×	∼
total	50	88%	2%	10%	44%	6%	50%	90%	0%	10%
source-female	25	92%	4%	4%	36%	0%	64%	92%	0%	8%
source-male	25	84%	0%	16%	52%	12%	36%	88%	0%	12%
aligned	43	95.35%	0%	4.65%	48.84%	2.33%	48.84%	95.35%	0%	4.65%
defiant	4	0%	25%	75%	0%	25%	75%	25%	0%	75%
non-stereotypical	3	100%	0%	0%	33.33%	33.33%	33.33%	100%	0%	0%

Table 1: Summary of each LLM performance (%) in gender-annotation tasks, where "✓", "×", and "∼" stand for "Match", "Mismatch" and "Neutral", respectively.

Category			Count	GPT-4o			Copilot			Grok		
Source	Stereotype	Situation		✓	×	∼	✓	×	∼	✓	×	∼
female	female	aligned	23	95.65%	0%	4.35%	39.13%	0%	60.87%	95.65%	0%	4.35%
male	male	aligned	20	95%	0%	5%	60%	5%	35%	95%	0%	5%
female	male	defiant	1	0%	100%	0%	0%	0%	100%	0%	0%	100%
male	female	defiant	3	0%	0%	100%	0%	33.33%	66.67%	33.33%	0%	66.67%
female	neutral	non-stereotypical	1	100%	0%	0%	0%	0%	100%	100%	0%	0%
male	neutral	non-stereotypical	2	100%	0%	0%	50%	50%	0%	100%	0%	0%

Table 2: LLM performance (%) in gender annotation, by source and visually implied gender, where "✓", "×", and "∼" stand for "Match", "Mismatch" and "Neutral", respectively.

actual gender.

4.2 Discussion

The discussion is organized into two parts: Section 4.2.1 examines model performance in gender-annotation tasks, while Section 4.2.2 explores the presence of stereotypes in generated image captions. For supporting statistical comparisons, see Section 4.2.3.

4.2.1 Performance in Gender-Annotation Tasks

As shown in Table 1, the mismatch rate is consistently higher in *defiant* scenarios. This is expected, as these cases involve individuals whose clothing or appearance contradicts traditional gender stereotypes, making gender identification more challenging. An exception is observed in Grok, which does not produce any mismatched annotations across scenarios. Across the three models, *defiant* situations also show a higher tendency to produce neutral outputs. While neutral outputs are desirable for general-purpose chatbots to avoid reinforcing stereotypes, they can be problematic for automatic gender annotation tasks, as they imply a refusal to assign gender. Regarding the rate of neutral responses, Copilot shows a greater overall tendency to produce neutral captions across all types of situations. In contrast, GPT-4o and Grok mostly restrict neutral outputs to defiant scenarios; in other cases, they tend to provide explicit gender annotations.

When examining incorrect annotations (mismatches), Grok stands out by making no errors, with all outputs being either matches or neutral. GPT-4o limits its mistakes to defiant scenarios only, whereas Copilot exhibits mismatches across all three situations, indicating less reliable performance in gender annotation. Performance also varies when considering the gender of the person in the source image. GPT-4o shows a higher mismatch rate for female images, whereas Copilot makes more errors in male images. Grok maintains consistent behavior, with no mismatches across gender groups. In summary, it is observed that all three models showed poor performance as gender annotators in defiant scenarios, highlighting the challenges of the images that contradict traditional gender stereotypes.

As shown in Table 2, Copilot shows the greatest gender variation in *aligned* cases, generating more neutral responses for female images and matches for male ones. GPT-4o and Grok show minimal variation. All mismatches in aligned scenarios occur with male images. In *defiant* situations, GPT-4o produces mismatches only for female images with male stereotypes and neutral responses in the

Category			Count	GPT-4o			Copilot			Grok		
Feature	Looks-Like	Source		✓	×	∼	✓	×	∼	✓	×	∼
prenda	male	female	2	50%	50%	0%	0%	0%	100%	50%	0%	50%
prenda	female	male	3	0%	0%	100%	0%	33.33%	66.67%	33.33%	0%	66.67%
pelo	male	female	6	83.33%	0%	16.67%	16.67%	0%	83.33%	83.33%	0%	16.67%
pelo	female	male	1	100%	0%	0%	0%	100%	0%	100%	0%	0%
zapatos	male	female	1	0%	100%	0%	0%	0%	100%	0%	0%	100%
zapatos	female	male	2	0%	0%	100%	0%	50%	50%	0%	0%	100%

Table 3: LLM performance (%) in gender annotation, classified by visual features, where “✓”, “×”, and “∼” stand for “Match”, “Mismatch” and “Neutral”, respectively.

Category	Count	GPT-4o			Copilot			Grok		
		S	CS	N	S	CS	N	S	CS	N
total	50	90%	0%	16%	50%	2%	62%	88%	2%	12%
source-female	25	96%	0%	16%	36%	0%	76%	92%	0%	12%
source-male	25	84%	0%	16%	64%	4%	48%	84%	4%	12%
aligned	43	95.35%	0%	6.98%	51.16%	2.33%	58.14%	95.35%	0%	6.98%
defiant	4	25%	0%	100%	25%	0%	100%	0%	25%	75%
non-stereotypical	3	100%	0%	33.33%	66.67%	0%	66.67%	100%	0%	0%

Table 4: Summary of found stereotypes for each LLM in image captioning tasks, where “S”, “CS”, and “N” stand for “Stereotypical”, “Counter-stereotypical” and “Neutral”, respectively.

reverse case. Copilot and Grok follow similar patterns, favoring neutral responses for female defiant cases, though differences are minor.

As shown in Table 3, GPT-4o and Copilot consistently produce neutral responses when subjects wear stereotypically opposite-gender clothing – GPT-4o for males and Copilot for females. Regarding hairstyles, GPT-4o and Grok are generally accurate, while Copilot leans toward neutral for females and mismatches for males. In footwear, Copilot outputs neutral for females with male-coded shoes, while GPT-4o mismatches; the opposite occurs for males with female-coded shoes.

4.2.2 Stereotypes in Image Caption Tasks

As shown in Table 4, models differ in handling gender stereotypes. GPT-4o does not produce counter-stereotypical responses, while Grok shows the highest proportion, specifically in defiant situations. GPT-4o and Grok lean toward stereotypical outputs, whereas Copilot favors neutral captions, particularly for female subjects.

To assess bias in conversational contexts, we examine *defiant* and *non-stereotypical* situations, where annotation errors reveal inherent bias. GPT-4o and Copilot mostly return neutral or stereotypical responses, while Grok is the only model generating counter-stereotypical outputs, though this does not imply better or worse performance.

In non-stereotypical cases, Grok often defaults to stereotypical responses, assigning gender even with ambiguous cues, while Copilot shows a more balanced mix of neutral and stereotypical outputs.

As shown in Table 5, in defiant situations, GPT-4o outputs stereotypical responses only for *female* images and neutral ones for *male*. Copilot maintains similar neutral rates across genders, with all stereotypical outputs tied to female images. Grok shows slight variation, leaning toward counter-stereotypical responses for male images.

As shown in Table 6, GPT-4o consistently produces stereotypical responses when male subjects exhibit female-associated features, although it occasionally remains neutral. In contrast, Copilot tends to provide neutral descriptions, particularly for female subjects with male-associated traits. Both GPT-4o and Grok label male subjects with stereotypically female hairstyles as stereotypical, whereas Copilot defaults to neutral–illustrating its tendency to avoid gender inference under ambiguity, unlike GPT-4o and Grok.

4.2.3 Statistical Analysis

A chi-squared test revealed a significant difference in caption alignment distributions (*Match*, *Mismatch*, *Neutral*) across models, $\chi^2(4) = 35.49$,

Category			Count	GPT-4o			Copilot			Grok		
Source	Stereotype	Situation		S	CS	N	S	CS	N	S	CS	N
female	female	aligned	23	95.65%	0%	8.70%	39.13%	0%	73.91%	95.65%	0%	8.70%
male	male	aligned	20	95%	0%	5%	65%	5%	40%	95%	0%	5%
female	male	defiant	1	100%	0%	0%	0%	0%	100%	0%	0%	100%
male	female	defiant	3	0%	0%	100%	33.33%	0%	100%	0%	33.33%	66.67%
female	neutral	non-stereotypical	1	100%	0%	0%	0%	0%	100%	100%	0%	0%
male	neutral	non-stereotypical	2	100%	0%	0%	100%	0%	50%	100%	0%	0%

Table 5: LLM stereotype summary in captioning, by source and visually implied gender, where "S", "CS", and "N" stand for "Stereotypical", "Counter-stereotypical" and "Neutral", respectively.

Category			Count	GPT-4o			Copilot			Grok		
Feature	Looks-Like	Source		S	CS	N	S	CS	N	S	CS	N
prenda	male	female	2	100%	0%	100%	0%	0%	100%	50%	0%	50%
prenda	female	male	3	0%	0%	100%	33.33%	0%	100%	0%	33.33%	66.67%
pelo	male	female	6	83.33%	0%	33.33%	16.67%	0%	83.33%	83.33%	0%	33.33%
pelo	female	male	1	100%	0%	0%	100%	0%	100%	100%	0%	0%
zapatos	male	female	1	100%	0%	100%	0%	0%	100%	0%	0%	100%
zapatos	female	male	2	0%	0%	100%	50%	0%	100%	0%	0%	100%

Table 6: LLM stereotype summary in captioning, classified by visual features, where "S", "CS", and "N" stand for "Stereotypical", "Counter-stereotypical" and "Neutral", respectively.

$p < .000001$. Pairwise comparisons showed significant differences between GPT-4o and Copilot ($\chi^2(2) = 21.67$, $p < .00002$), and between Copilot and Grok ($\chi^2(2) = 24.23$, $p < .00001$), but not between GPT-4o and Grok ($\chi^2(2) = 1.01$, $p = .60$).

Similarly, stereotype-related framing (*Stereotypical*, *Counter-stereotypical*, *Neutral*) differed significantly across models, $\chi^2(4) = 32.21$, $p < .000002$. GPT-4o and Copilot ($\chi^2(2) = 20.16$, $p < .00005$), as well as Copilot and Grok ($\chi^2(2) = 21.86$, $p < .00002$), showed significant distributional differences, while GPT-4o and Grok did not ($\chi^2(2) = 1.26, p = .53$).

These results suggest that GPT-4o and Grok behave similarly on both alignment and stereotype framing tasks, whereas Copilot differs significantly from both.

5 Conclusions

This paper proposes a structured evaluation protocol to study gender bias in image-to-text generation, analyzing outputs from ChatGPT, Copilot, and Grok when describing fashion-related images prompted without gender cues. Through this approach, our findings indicate that:

1. GPT-4o and Grok generate more gendered and stereotypical descriptions, while Copilot produces a higher proportion of neutral prompts.
2. Grok shows the lowest mismatch rate in gender attribution but tends to consistently assign gender even when visual cues are unclear.
3. GPT-4o and Grok reinforce gender stereotypes more prominently, while Copilot exhibits a more conservative, neutral approach in ambiguous cases.

Key contributions of this work include: (1) a reproducible methodology for gender bias analysis in image captioning; (2) a detailed annotation strategy; (3) a manually curated dataset for future research; and (4) an evaluation of three state-of-the-art LLMs under gender-annotation and image captioning tasks. We expect this work lays the foundation for fairer image captioning systems and guides future annotation and multimodal evaluations practices by highlighting critical points where implicit biases emerge and must be addressed.

Future research may extend this study by balancing the number of the class-level within the dataset. While gender representation was controlled at the source level—comprising an equal number of male and female subjects (25 each)—the distribution across classes remained unbalanced. Furthermore, the investigation should be broadened to encompass additional protected attributes, such as race or ethnic bias, to provide a more comprehensive assessment of bias in fashion imagery.

Acknowledgments

This research is funded by the projects "NL4DISMIS: Natural Language Technologies for dealing with dis- and misinformation (CIPROM/2021/021)" by the Generalitat Valenciana and the VIVES: "Pla de Tecnologies de la Llengua per al valencià" (2022/TL22/00215334) from the Strategic Project for Economic Recovery and Transformation (PERTE), as well as by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA.

References

Muhammad Abdelhadie Al-Malla, Assef Jafar, and Nada Ghneim. 2022. Image captioning model using attention and object features to mimic human image understanding. *Journal of Big Data*, 9(1):20.

Hritik Bansal, Da Yin, Masoud Monajatipoor, and Kai-Wei Chang. 2022. How well can text-to-image generative models understand ethical natural language interventions? *arXiv preprint arXiv:2210.15230*.

Shruti Bhargava and David Forsyth. 2019. Exposing and correcting the gender bias in image captioning datasets and models. *arXiv preprint arXiv:1912.00578*.

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*.

Jaemin Cho, Abhay Zala, and Mohit Bansal. 2023. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3043–3054.

Juan Pablo Consuegra-Ayala, Iván Martínez-Murillo, Elena Lloret, Paloma Moreda, and Manuel Palomar. 2024. A multifaceted approach to detect gender biases in natural language generation. *Knowl. Based Syst.*, 303:112367.

Shubhang Desai. 2024. Fair attention-based image captioning. *arXiv:2401.17910v3*.

Yunhe Feng and Chirag Shah. 2022. Has ceo gender bias really been fixed? adversarial attacking and improving gender fairness in image search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11882–11890.

Taraneh Ghandi, Hamidreza Pourreza, and Hamidreza Mahyar. 2023. Deep learning approaches on image captioning: A review. *ACM Comput. Surv.*, 56(3).

Yusuke Hirota, Yuta Nakashima, and Noa Garcia. 2022. Quantifying societal bias amplification in image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13450–13459.

Yusuke Hirota, Yuta Nakasshima, and Noa Garcia. 2025. Societal bias in image captioning: Identifying and measuring bias amplification. *IEICE Transactions on Information and Systems*, page 2024EDP7116.

Dennis Jacob, Hend Alzahrani, Zhanhao Hu, Basel Alomair, and David Wagner. 2025. Promptshield: Deployable detection for prompt injection attacks.

Matthew Kay, Cynthia Matuszek, and Sean A Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd annual acm conference on human factors in computing systems*, pages 3819–3828.

Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. 2023. From show to tell: A survey on deep learning-based image captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):539–559.

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*.

Dora Zhao, Angelina Wang, and Olga Russakovsky. 2021. Understanding and evaluating racial biases in image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14830–14840.

Yuzhong Zhao, Yue Liu, Zonghao Guo, Weijia Wu, Chen Gong, Fang Wan, and Qixiang Ye. 2024. Controlcap: Controllable region-level captioning.