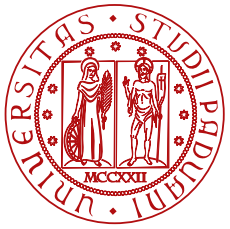

Università degli Studi di Padova

Dipartimento di Ingegneria dell'Informazione

Master Thesis in Computer Engineering

Gender and Racial Bias in Visual Question Answering Datasets

Master Candidate

Asifa Akter

Student ID 2013046

Supervisor

Prof. Antonio Roda

University of Padova

Academic Year
2023/2024

To my family
and friends

Abstract

In this thesis, we apply the methodology suggested by Hirota, Nakashima, and Garcia (2022) to examine gender and racial biases in Visual Question Answering (VQA) datasets. Four popular VQA datasets—Visual Genome, Visual7W, VQA 2.0, and OK-VQA—are the subject of our study. Every dataset undergoes a methodical analysis aimed at identifying and measuring biases related to race and gender. To enable a focused analysis of these biases, we use a rule-based method to find samples that specifically mention gender or race.

We divide questions into binary categories (men and women) for gender bias analysis based on terms that are specific to one gender over the other. Our analysis of all datasets shows that there is a notable gender representation gap, with questions about men being almost twice as common as those about women. This disparity is indicative of more general patterns found in the original COCO dataset, which provided the images for this collection.

In the same way, questions mentioning race and ethnicity are identified and examined to look at instances of racial bias. Underrepresentation trends and other biases in the datasets are brought to light by the comparative analysis.

Our work intends to further our knowledge of the biases present in VQA datasets and aid in the creation of AI models that are more equitable and inclusive. Our goal is to reduce negative stereotypes in AI systems and advance equity by recognizing and resolving these biases. To accomplish these objectives, this thesis emphasizes how crucial it is to critically assess AI datasets and to continuously enhance them.

To help make AI models more inclusive and equitable, we want to understand more about the biases present in VQA data. We aim to reduce negative preconceptions in AI and advance justice by identifying and addressing these prejudices. This thesis emphasizes the necessity of closely examining AI data sets and continuously enhancing them to achieve these aims.

Contents

List of Figures

List of Tables

List of Acronyms

AI Artificial Intelligence

API Application Programming Interface

BoW Bag of Words

CNN Convolutional Neural Network

COCO Common Objects in Context

CSV Comma Separated Values

CV Computer Vision

FN False Negative

FP False Positive

GAN Generative Adversarial Network

GPU Graphics Processing Unit

LLM Large Language Model

LSTM Long Short-Term Memory

ML Machine Learning

NLP Natural Language Processing

OK-VQA Outside Knowledge - Visual Question Answering

RNN Recurrent Neural Network

ROC Receiver Operating Characteristic

SGD Stochastic Gradient Descent

SVM Support Vector Machine

TN True Negative

TP True Positive

TPU Tensor Processing Unit

VQA Visual Question Answering

YOLO You Only Look Once

1 Introduction

1.1 Motivation

In contemporary society, the rapid increase of machine learning (ML) and artificial intelligence (AI) technologies in various facets of our lives has brought both unprecedented opportunities and daunting challenges. While these technologies promise to revolutionize industries, improve efficiency, and enhance decision-making processes, they also carry the risk of perpetuating biases and reinforcing existing inequalities. The issue of biased decision-making by ML algorithms represents a critical hurdle in the quest for building fair and just societies. Biases embedded within ML algorithms have far-reaching consequences, permeating domains as diverse as healthcare, criminal justice, finance, and employment. For instance, in healthcare, biased algorithms may inadvertently discriminate against certain demographic groups, leading to disparities in medical treatment and healthcare outcomes. Similarly, in the criminal justice system, biased algorithms can exacerbate existing racial disparities, resulting in unjust outcomes and perpetuating systemic inequalities. Addressing these biases and promoting fairness in decision-making processes are paramount objectives that underpin the ethical deployment of AI systems. Building trust in AI systems requires a concerted effort to mitigate biases and ensure equitable out-

comes for all individuals, regardless of race, gender, ethnicity, or socioeconomic status. Furthermore, fostering inclusivity and diversity within AI development is essential to safeguard against algorithmic discrimination and promote social justice. Consequently, the motivation behind this thesis is twofold: firstly, to illuminate the pervasive nature of biases within ML algorithms and their detrimental impact on society, and secondly, to explore innovative approaches and develop AI-driven applications that mitigate biases and promote unbiased decision-making. By addressing these challenges head-on, we can pave the way for the creation of AI systems that uphold ethical principles, foster inclusivity, and contribute to the realization of a more just and inclusive society for all.

1.2 Research Challenge

The research challenge inherent in addressing biased decision-making by ML algorithms is multifaceted and complex, spanning technical, ethical, and societal dimensions.

1.2.1 Technical Complexity

Developing effective techniques for detecting and mitigating biases within ML algorithms presents a significant technical challenge. Biases can manifest in various forms, including sampling bias, algorithmic bias, and societal bias, making them difficult to identify and mitigate. Moreover, the dynamic nature of data and the evolving nature of biases necessitate continuous monitoring and adaptation of algorithms, adding to the technical complexity.

1.2.2 Data Quality and Bias

A fundamental challenge lies in the quality and representations of the data used to train ML models. Biased training data can propagate and amplify existing biases, leading to biased predictions and decisions. Addressing data biases requires careful data collection, preprocessing, and augmentation techniques to ensure that ML models learn from diverse and representative datasets.

1.2.3 Interdisciplinary Collaboration

Mitigating biases in ML algorithms requires interdisciplinary collaboration between experts in ML, ethics, social sciences, and domain-specific knowledge. Understanding the socio-cultural context in which algorithms operate is essential for identifying and addressing biases effectively. Bridging the gap between technical expertise and ethical considerations is crucial for developing AI systems that uphold fairness and justice.

1.2.4 Evaluation and Validation

Evaluating the effectiveness and fairness of bias mitigation techniques poses a significant challenge. Traditional metrics for evaluating ML models, such as accuracy and precision, may not capture the nuances of bias and fairness. Developing comprehensive evaluation frameworks and metrics that account for various dimensions of bias, including disparate impact and fairness across demographic groups, is essential for robust assessment and validation of AI-driven applications.

1.2.5 Transparency and Accountability

Ensuring transparency and accountability in AI-driven applications is essential for building trust and fostering responsible deployment. However, achieving transparency in complex ML algorithms, such as deep neural networks, remains a challenge. Developing interpretable and explainable AI techniques that provide insights into algorithmic decision-making processes is critical for enhancing transparency and accountability.

1.2.6 Ethical Considerations

Ethical considerations permeate every stage of the AI development lifecycle, from data collection and model training to deployment and impact assessment. Balancing competing ethical principles, such as privacy, fairness, and utility,

requires careful deliberation and ethical reasoning. Moreover, addressing biases in AI systems raises ethical dilemmas, such as trade-offs between fairness and accuracy, necessitating ethical frameworks and guidelines to navigate these complexities.

1.3 Contribution

This thesis aims to contribute to the understanding and mitigation of gender and racial biases in Visual Question Answering (VQA) systems. Through the examination of well-known VQA datasets, including Visual Genome, Visual7W, VQA 2.0, and OK-VQA, the study demonstrates how racial and gender biases are present in the data and affect AI-powered decision-making. The thesis detects and quantifies these biases by repeating and expanding upon earlier investigations, applying techniques influenced by Hirota et al.

1.4 Chapter Outline

This thesis investigates the gender and racial bias of VQA datasets to ensure fairness, preventing AI models from perpetuating harmful stereotypes. The chapters are structured as follows.

Chapter 2: Background and Related Work

Chapter 2 provides a comprehensive overview of existing approaches and methodologies in addressing gender and racial bias in Visual Question Answering (VQA) datasets. We review relevant literature, highlight key studies on bias detection and mitigation, and discuss the gaps and limitations in current research. This chapter lays the groundwork for developing and evaluating strategies to reduce gender and racial bias in VQA systems, ensuring more ethical and accurate AI performance across diverse demographic groups.

Chapter 3: Experimental Evaluation

In Chapter 3, we present the methodology employed in this research. We detail the technical aspects, and procedures and apply this to the datasets to gather the results and evaluate them.

Chapter 4: Conclusion and future work

Chapter 4 serves as the concluding chapter of the thesis. We summarize the key findings, discuss the research implications, and provide a comprehensive conclusion. Additionally, we highlight the limitations of the study and suggest future research directions to further advance the field of gender and racial bias detection.

Finally, we will conclude this introduction chapter, which has provided a clear overview of the motivation, research challenges, and contributions of this thesis. We have addressed the critical need to investigate and mitigate gender and racial bias in the VQA dataset to reduce stereotypes in AI systems.

2

Background and Related Work

2.1 Objective

In this chapter, we provide a comprehensive overview of the background and related works that are central to Visual Question Answering (VQA) systems.We discuss key underlying topics such as image captioning techniques, filtering unanswered questions, and the role of manual review platforms in the early stages of VQA development. We also discuss previous studies that had a direct impact on this sector, highlighting how these investigations prepared the way for our investigation into bias mitigation and detection. Lastly, we summarise relevant publications that relate to our area of study and show how they handle the difficulties in developing just and accurate VQA systems.

2.2 Background

The field of Visual Question Answering (VQA) represents a significant stride toward achieving artificial intelligence systems capable of understanding and interpreting visual content in a manner akin to human cognition. By requiring a model to provide answers to questions about images, VQA tasks integrate the complexities of both visual perception and language understanding. How-

ever, as this field has advanced, it has also revealed underlying biases within its datasets, particularly concerning gender and racial representations. This chapter delves into the background necessary to understand these biases, their implications, and the related work aimed at addressing them.

2.3 Visual Question Answering (VQA)

Visual Question Answering (VQA) tasks involve presenting a model with an image accompanied by a question about that image, to which the model must provide an accurate answer. The complexity of VQA lies not only in the need for detailed visual understanding but also in comprehending the question's linguistic nuances and context. The pioneering work in VQA introduced datasets like VQA 1.0, followed by iterations such as VQA 2.0, which aimed to reduce language biases by including pairs of images with contrasting answers to the same question, thereby forcing models to rely more heavily on visual content rather than linguistic cues.[5]

Figure 2.1: Diagram of VQA. Given an image and a question about the image, a model answers the question.

2.4 Improving VQA Precision with Captioning

one of the greatest challenges in modern machine learning research is making decisions based on both visual and language content. These applications have to bridge the modality gap between the visual and language inputs in addition to addressing the difficulties of text and image understanding.To get this

application making the right decision also depends on the perfect annotation or captioning, The work of captioning images has become difficult due to the non-trivial modality gap.

a generative model that is based on a deep recurrent architecture that features recent advances in computer vision and machine translation which can be used to generate sentences describing an image. Given the training image, the model is trained to maximize the chance of the goal description sentence. Tests conducted on multiple datasets demonstrate the model's accuracy and the fluency of the language it picks up from picture descriptions alone.[22]

As we can see in the figure 2.2, the final system is giving the most accurate

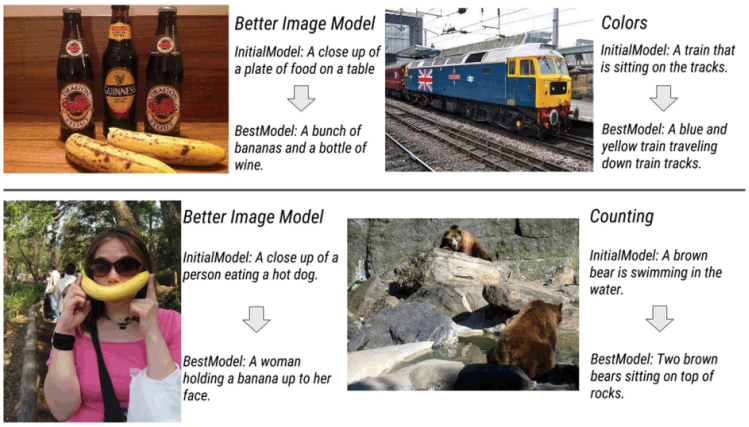

Figure 2.2: a collection of assessment photos, contrasting the captions produced by the initial model and the final model.[27]

answer. However, there is a possibility that it could be wrong in some cases. for that manual captioning is considered as an alternative approach. in this case, two main sources of societal bias in VQA datasets are identified, and mitigation techniques to reduce them. The under-representation of marginalized communities is the first issue. Since models trained on VQA datasets might learn to neglect minoritized features and lead to shortcut learning[4], we stress the significance of identifying and mitigating social bias in these datasets [4, 7, 8, 9, 12, 11, 13, 14, 15, 17, 26]. Another important factor that might affect the fairness and accuracy of models is the presence of dangerous or problematic samples in VQA datasets. We offer three simple fixes to this problem that can be put into

practice while taking annotation costs into account. By locating and reducing the impact of these samples, these methods seek to provide a more pristine and trustworthy dataset for VQA model training. Automatically filter to produce this, queries that are either too unclear to produce a trustworthy answer or lack enough context from the image can be flagged using rule-based algorithms or pre-trained models. By identifying these problematic questions early, we can reduce the likelihood of introducing noise into the dataset, thereby improving the overall quality of the data.

Detect Unanswerable Questions Automatically:

The first stage is to put automated mechanisms in place to screen datasets for questions that are unclear or cannot be answered. This can be accomplished by flagging questions that are either too unclear to Another important factor that might affect the fairness and accuracy of models is the presence of dangerous or problematic samples in VQA datasets. We offer three simple fixes to this problem that can be put into practice while taking annotation costs into account. By locating and reducing the impact of these samples, these methods seek to provide a more pristine and trustworthy dataset for VQA model training.

Automatically filter to Produce Include Moral Recommendations in the Annotation Procedure:

Annotators are better prepared to identify and address harmful content or bias in the questions or responses when ethical guidance is incorporated into the annotation process. Ensuring the dataset is inclusive and ethically conscious requires providing annotators with clear rules that assist them in identifying any social prejudices (e.g., racial or gender stereotypes) or improper information. By taking this step, the likelihood that bias will spread from the data to the VQA models is reduced.

Provide an Open Platform for User Comments:

In order to update the dataset continuously, it is essential to create an open platform where people can offer feedback. Dataset designers can ensure that the community actively contributes to maintaining the quality and fairness of the data by enabling users to identify faulty samples, suggest corrections, or offer insights into potential biases. By working together, we can find problems that automated systems or annotators would have overlooked, resulting in a more comprehensive and user-friendly dataset. By following these three steps—automatically screening unanswerable questions, embedding ethical considerations in the annotation procedure, and creating a manually inspecting platform—hazardous samples can be effectively managed, leading to a more ethical and high-quality VQA dataset which is very important in every aspect.

2.5 Emergence of Bias in VQA Datasets

As VQA datasets have grown in size and complexity, they have inadvertently encapsulated societal biases present in the data collection process. These biases manifest as skewed representations of gender and race, where certain demographics are underrepresented or depicted in stereotypical contexts. For instance, questions related to men might significantly outnumber those related to women, and the depiction of activities or professions might reinforce outdated stereotypes.[3]

2.6 Implications of Bias

The presence of bias in VQA datasets is not merely a statistical anomaly; it has profound implications for the fairness and inclusivity of AI technologies. Models trained on biased data are likely to perpetuate and even amplify these biases in their outputs, leading to unfair or harmful outcomes. For instance, if a VQA system consistently associates certain professions or activities with one gender or race, it reinforces stereotypes and ignores the rich diversity of human

capabilities and interests.

2.7 Related work

Numerous studies have investigated biases in ML algorithms and proposed various approaches to mitigate them. Fairness-aware learning techniques, such as fairness constraints and regularization methods, aim to enforce fairness constraints during the training process to mitigate biases. Adversarial debiasing algorithms leverage adversarial training to learn fair representations of data by simultaneously optimizing for accuracy and fairness. Moreover, pre-processing techniques, such as reweighing and re-sampling, aim to mitigate biases in training data by adjusting sample weights or distributions.

Furthermore, researchers have explored the societal impact of biased decision-making in AI systems across different domains. In healthcare, biases in predictive models can lead to disparities in medical treatment and healthcare outcomes, exacerbating existing inequalities. In criminal justice, biased algorithms used for risk assessment and sentencing decisions can perpetuate racial disparities and injustices. Similarly, biases in hiring algorithms can lead to discriminatory practices and unequal opportunities in employment.

2.7.1 Overcoming Bias in Captioning Models

Burns et al. (2018) address a critical aspect of AI ethics by highlighting gender bias within image captioning models. Their investigation reveals how these models, driven by biased datasets, tend to perpetuate gender stereotypes by linking specific activities or objects to genders. For instance, the presence of a kitchen might trigger captions that unjustly associate females with cooking. To counteract this issue, Burns et al. propose a set of methodologies aimed at mitigating such biases. These include the development and utilization of balanced datasets that fairly represent genders across a wide range of activities and contexts. Moreover, they advocate for the integration of bias-awareness in the training process of models, ensuring that the systems are not only aware of the potential biases but are also equipped to actively avoid perpetuating

them. This pioneering work sets a foundational framework for future efforts in developing more equitable AI systems by emphasizing the importance of data diversity and model sensitivity to biases.[5]

2.7.2 Reducing Gender Bias Amplification

Zhao et al. (2017) delve into the complex issue of gender bias amplification in language models, a problem that arises when models reinforce and exaggerate existing stereotypes. They introduce an innovative approach to mitigate this issue by implementing corpus-level constraints. This method involves adjusting the output distribution of models to ensure a more equitable representation of genders. By doing so, Zhao and colleagues aim to diminish the models' propensity to reinforce societal stereotypes, thereby contributing to the development of more impartial and fair language processing systems. This work is pivotal as it not only identifies a significant challenge in AI but also offers a practical solution to address it, paving the way for future research in bias mitigation.[25]

2.7.3 Fairness in Computer Vision

Hendricks et al. (2018) expand upon the discourse around gender bias in visual technologies, specifically within visual captioning. Their work underscores the necessity of incorporating bias awareness throughout the model training process. They argue that achieving fairness in AI systems requires a multifaceted approach, including developing balanced datasets that accurately reflect the diversity of human experiences and identities. By fostering an environment where AI models are trained on diverse datasets, Hendricks and colleagues believe it is possible to significantly reduce the biases inherent in visual captioning systems. Their contributions are instrumental in guiding the AI community toward the development of more equitable technologies by demonstrating the critical role of thoughtful dataset compilation and model training practices.[16]

2.7.4 Intersectional Accuracy Disparities in Commercial Gender Classification

In their landmark study, Buolamwini and Gebru (2018) unveil the alarming biases present in commercial facial recognition systems, particularly those affecting gender and skin type. Their research demonstrates how these systems exhibit significant discrepancies in accuracy, disproportionately disadvantaging women of darker skin tones. By advocating for the inclusion of more diverse training datasets and promoting the use of intersectional evaluation frameworks, Buolamwini and Gebru highlight the urgent need for the AI field to address these disparities. Their work not only sheds light on the extent of bias within commercial AI applications but also sets a precedent for how such issues should be critically examined and resolved, emphasizing the importance of diversity and inclusion in AI development processes.[1]

2.7.5 Bias Mitigation Strategies in Visual Recognition

Wang et al. (2020) contribute to the ongoing efforts to combat bias in AI by proposing a series of strategies aimed at mitigating bias in visual recognition systems. Their approach is twofold, involving data resampling techniques to ensure a balanced representation of demographic groups within training datasets, and model regularization methods designed to minimize bias during the learning process. These strategies are intended to equalize performance across different groups, thereby fostering fairness and equity in AI outcomes. The significance of Wang and colleagues' work lies in its practical applicability; by offering tangible solutions for bias mitigation, they provide valuable tools for developers seeking to create more impartial AI systems.[28]

2.7.6 Elevating the Role of Image Understanding in VQA

Goyal et al. (2017) address the critical issue of language bias in Visual Question Answering (VQA) by introducing the VQA 2.0 dataset. This dataset is specifically designed to minimize language biases by ensuring that each ques-

tion is answerable with both "yes" and "no," depending on the accompanying image. This requirement compels VQA models to rely more heavily on visual information rather than linguistic cues, thus promoting a more genuine integration of visual and textual data. Goyal and colleagues' work represents a significant step towards reducing bias in VQA systems, underscoring the importance of carefully designed datasets in fostering models that truly understand and interpret visual content.[**reference7**]

2.7.7 Discovering Explicit Biases in VQA Models

Manjunatha et al. (2019) undertake a critical examination of explicit biases present in VQA models, with a particular focus on gender and racial stereotypes. Through innovative methodologies, they manage to trace these biases back to specific dataset instances, providing a detailed analysis of how and where biases are introduced into AI systems. By uncovering these explicit biases, Manjunatha and colleagues not only raise awareness about the depth of the problem but also lay the groundwork for developing targeted interventions to mitigate these biases. Their work highlights the necessity of continuous vigilance and proactive efforts to identify and address bias in AI, ensuring that these technologies serve all users equitably.[21]

2.8 Applications using (VQA) Systems

The ability of Visual Question Answering (VQA) systems to bridge the gap between natural language processing and computer vision has attracted a lot of interest. VQA systems allow machines to respond to questions based on photographs. These systems can be used for a variety of tasks, such as autonomous driving, human-computer interaction, and content moderation. Here is a thorough summary of several well-known individuals and systems that have made use of VQA technology, together with helpful details on its importance and applications.

1. IBM Watson Visual Recognition

One of the most well-known applications of VQA technology is IBM Watson's Visual Recognition service, which is a leading example of artificial intelligence. Users can upload photos to the system and ask precise queries about what's in them, such recognising scenes, identifying objects, or deciphering facial emotions. This technique is commonly utilised in retail and healthcare, where VQA may help with product identification and medical image diagnosis. In order to deliver precise answers based on visual content, IBM's VQA technology makes use of deep learning models that have been trained on big datasets.

2.Visual Dialog (AI2)

Visual Dialogue, a system created by the Allen Institute for AI (AI2), combines conversational AI and visual question answering (VQA) to provide an interactive platform where users can ask questions about images and receive dialogue-style responses from the AI. As opposed to normal VQA, which responds to one-time inquiries, Visual Dialogue retains information from prior exchanges, enabling a more sophisticated comprehension of the visual content. Applications like tutoring systems, in which a person interacts with an AI across numerous exchanges to comprehend complex visuals or diagrams, can benefit from this technology. It has the potential to be used in customer service as well, as users can converse with visual items.

3.Facebook AI Research (FAIR)

Significant progress has been achieved in VQA by Facebook's AI Research (FAIR) group, especially in applications pertaining to augmented reality, accessibility, and content moderation. Facebook utilises visual quality assurance (VQA) to examine photos and videos that are posted to its networks. This helps with tasks like identifying unsuitable content and interpreting the visual context of images. Facebook's accessibility features now include VQA integration, which enables visually challenged users to enquire about photos shared in their newsfeeds. When a user uploads a photograph and asks, "What is the person

in the picture doing?" for example, the system analyses the image and returns a response.

4.Google Lens

One example of how VQA is incorporated into technology aimed towards consumers is Google Lens. It enables users to snap images of scenes, objects, or text and ask queries about what they observe. For example, a user may snap a photo of a famous building and enquire, "What's the name of this building?" Google Lens analyses the image and returns precise results by using VQA techniques. There are many uses for this tool in daily life, such as product identification from photos when buying, text translation or landmark recognition when traveling, and textbook explanations of scientific processes or graphics when teaching. Google Lens's capacity to process a wide range of real-time questions in a variety of visual domains is key to its success.

5.Microsoft Azure Cognitive Services (Computer Vision API)

One of the computer vision APIs in Microsoft's Azure Cognitive Services package makes use of VQA technology. This API can answer specific queries regarding the content of photographs, describe scenarios, and automatically create captions for images. In terms of accessibility, this technology enables people who are blind or visually impaired to engage with images in ways that were not previously feasible. For example, users can now ask queries like, "Is there a person in this image?" or "What is in the top left corner of the picture?" Enterprise applications make extensive use of this approach, which helps companies to automatically tag, classify, and derive insights from visual data.

6.OpenAI's CLIP GPT-4 Vision

OpenAI has developed GPT-4 Vision and CLIP, two innovations that have revolutionised VQA. Without specialised training for VQA tasks, the model

CLIP (Contrastive Language-Image Pretraining) can answer questions about images because it can comprehend both text and images. This is furthered by OpenAI's GPT-4 Vision, which combines natural language processing and image comprehension to enable users to upload photos and pose intricate queries about them. For instance, someone might upload a picture of a mechanical gadget and enquire, "How does this work?" After analysing the image, GPT-4 Vision would offer an educational answer that took into account both visual and textual information. With uses ranging from technical help to teaching, this is a major advancement in multimodal AI.

7.Tesla's Autopilot System

To respond to inquiries about navigation, safety, and road conditions, Tesla's Autopilot system processes visual inputs from cameras all around the car in real-time. This is an example of VQA in action. Tesla's system must continuously "answer" implicit visual questions such as "Is there an obstacle in the path?" and "What is the speed limit?" even though it is incapable of processing spoken language queries. With the help of this real-time VQA, autonomous driving in Tesla vehicles is made possible, enabling safe navigation of challenging environments.

8.Amazon Rekognition

VQA-based sophisticated image and video analysis services are provided by Amazon Rekognition. It can respond to queries like "Is there a person in this image?" and "How many dogs are in this video?" and recognise objects, people, text, and activities in both photos and videos. Businesses frequently employ recognition for face verification, content restriction, and even security. Rekognition's VQA skills make it valuable in fields like law enforcement, where instantaneous processing and understanding of visual data is required to make choices.

9. Alibaba DAMO Academy

VQA systems have been studied at Alibaba's DAMO Academy, mostly for retail and e-commerce applications. Customers can post pictures of products and ask queries regarding them, like "What is the material of this dress?" or "Is the item offered in other colors?" thanks to their VQA technology. Alibaba's real-time query answering (VQA) technologies are seamlessly integrated into their e-commerce platforms, offering a flawless buying experience. The system is one of the most reliable VQA systems in the retail industry because it is extremely scalable and designed to handle a large catalog of products.

A growing number of industries, including healthcare (medical image diagnosis), accessibility (assisting people with vision impairments in navigating the environment), autonomous systems (self-driving automobiles), education, and customer service, are utilising VQA systems. The ability of machines to comprehend and reason about visual content will continue to grow as AI models advance, especially with the development of larger and more complex multimodal models like OpenAI's GPT-4 Vision. This will lead to more useful applications for VQA in daily life and business operations.

3

Experimental Evaluation

3.1 Introduction

In this thesis, we examine racial and gender bias in Visual Question Answering (VQA) datasets in detail. We are employing a technique described in 2022[10] by Hirota, Nakashima, and Garcia. They devised a method to determine whether these datasets include biases based on race or gender. To begin, we select samples from each dataset that contain explicit references to gender (such as questions regarding men or women) or race and ethnicity. We use a set of rules to accomplish this. Next, to detect bias, we compare several sets of these samples. In addition to examining sample size and response type, we physically inspect the data for any indications of potentially hazardous information.

3.2 VQA datasets

To look into the existence of racial and gender biases, we analyze four common Visual Question Answering (VQA) datasets in this thesis. These datasets, which are compiled in figure 3.5, each have distinct features regarding the number of images, questions, techniques of annotation, and forms of the responses. Despite these variations, the COCO (Common Objects in Context) dataset[19]

is the source of all images included in these datasets, giving our research a consistent visual foundation. The COCO dataset serves as the source of images for all of these datasets (Visual Genome, Visual7W, VQA 2.0, and OK-VQA), however, the quantity of images varies since each dataset chooses and annotates a different subset of the COCO dataset in accordance with its intended purpose. Here, we thoroughly explain each dataset, emphasizing its unique qualities and importance to our research of each. With its extensive collection of images tagged with relationships, object features, and region descriptions, the Visual Genome dataset provides a rich environment for analyzing biases in question-answer pairs and visual content. By concentrating on seven different "W" question types like what, where, when, who, why, how, and which Visual7W expands on the Visual Genome and adds further annotations, such as bounding boxes for objects, to enable a more in-depth investigation of relational and spatial biases. VQA 2.0 is a popular benchmark created to counteract biases by guaranteeing that every question is matched with comparable images that provide varying replies. This feature makes VQA 2.0 especially helpful when researching how various visual contexts influence responses. To answer the questions, OK-VQA requires external knowledge beyond what is visible in the photos. This challenges models by asking them to incorporate general world knowledge and offers a fresh viewpoint on potential biases that could result from combining external information with visual data. Our goal in examining these datasets is to identify any possible biases and help create AI models that are more just and equal.

Year	Dataset	Num. Images	Num. QA	QA Annotation	Answer Type
2016	Visual Genome [18]	108k	1.7M	Crowdsourcing	Open ended
2016	Visual7W [31]	47k	327k	Crowdsourcing	Multiple choice
2017	VQA 2.0 [6]	204k	1.1M	Crowdsourcing	Open ended
2019	OK-VQA [24]	14k	14k	Crowdsourcing	Open ended

Table 3.1: Summary of datasets used for visual question answering tasks.

3.2.1 Visual Genome

Visual Genome dataset has more than 108K photos, with an Average number of distinct objects per image, along with their characteristics and relationships to other objects. By "average of objects," we mean the typical number of labeled objects identified within each image. For example, each image might have objects like a "tree," "dog," or "car," and across all images, they calculate the mean number of such objects. In addition, they canonicalize to WordNet synsets the noun phrases, objects, attributes (adjectives), and relationships that may be discovered in region descriptions and question-answer pairs. This means that various keywords are standardised by mapping them to a common reference in WordNet, a lexical database that creates synsets (sets of synonyms) out of related words. To ensure that changes in language are related to the same notion, for example, various terms like "car" and "automobile" are transformed to the same synset. This makes it easier for AI systems to process and interpret the data. it includes a vast number of images with extensive annotations. Each image contains many detailed descriptions, including numerous objects and the relationships between them, making it highly informative for training AI models. The region descriptions, objects, characteristics, relationships, region graphs, scene graphs, and question-answer pairs are the seven primary components that make up the Visual Genome dataset. Figure 3.1 shows examples of each component for a single image.

3.2.2 Visual7W

A total of 1.3 million human-generated multiple-choice responses based on 47,000 COCO images and 327,000 QA pairs make up Visual7W. Each question has multiple-choice answers with object-level justifications and various candidate options; there is only one right answer per question. The seven Ws—What, Where, When, Who, Why, How, and Which—are the first words of each query. Wordiness or conjecture should be avoided when creating clear, brief sets of question-answer, according to instructions given to annotators. Other annotators then confirm these pairs to make sure the ordinary individual can correctly

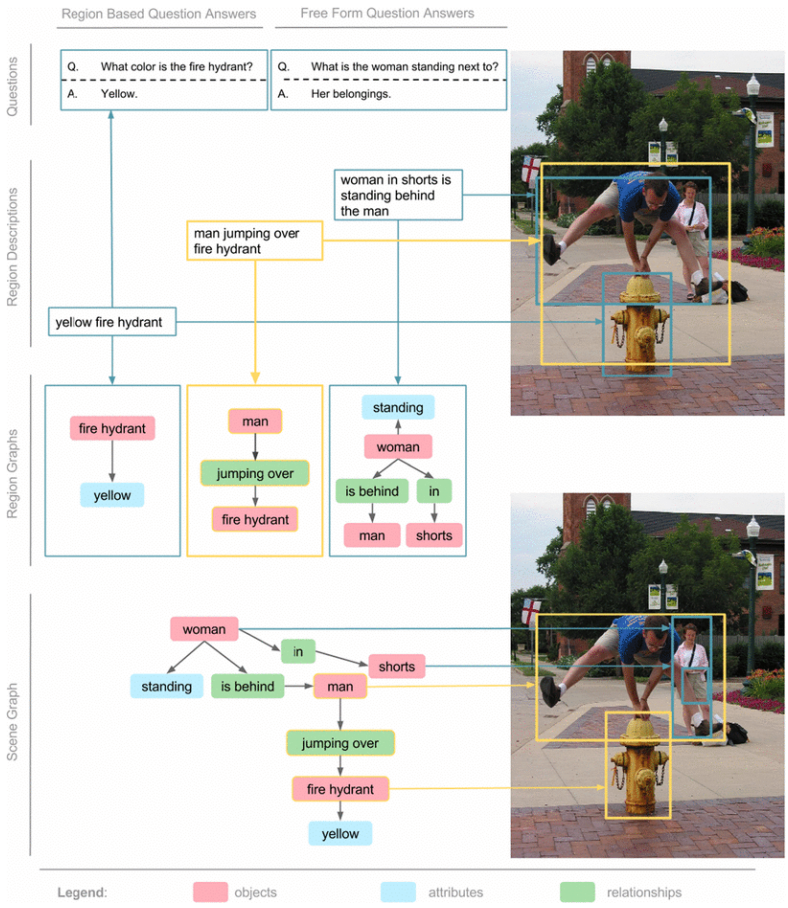

Figure 3.1: A representation of the Visual Genome dataset

answer them. This thorough annotation method guarantees the dataset's dependability and usefulness for Visual Question Answering (VQA) research, enabling in-depth examination of how various question kinds are handled by VQA models. examples from the Visual7W[**30**] dataset are displayed in the figure 3.2, highlighting some of its distinctive features. Questions beginning with one of the seven Ws—What, Where, When, Who, Why, How, and Which—are matched with each image.

3.2.3 VQA 2.0

One of the most popular datasets for Visual Question Answering (VQA) research is VQA 2.0[6], which was created to address and lessen biases seen in

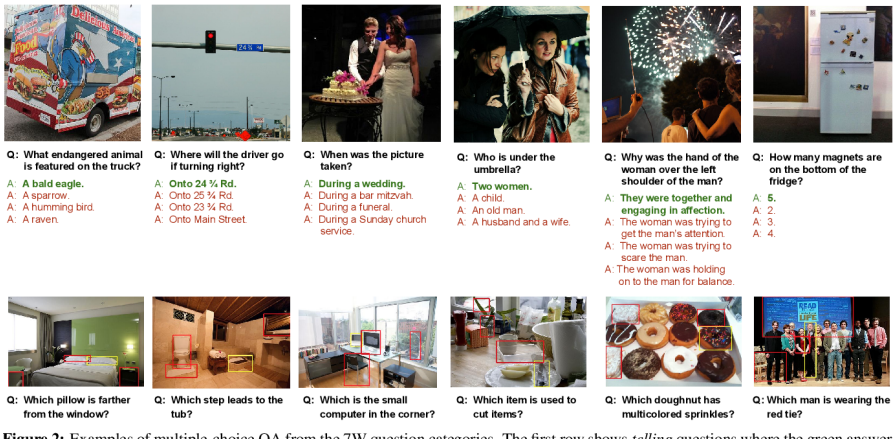

Figure 3.2: Examples of multiple-choice QA from the 7W question categories.

prior iterations. More than 204,000 photos from the COCO dataset are linked with more than 1.1 million queries. By ensuring that every question is linked to pairs of comparable images that provide distinct answers, VQA 2.0 aims to increase the robustness and fairness of VQA models. This configuration lessens answer biases, which occur when a model guesses an answer based more on question patterns than on visual content. VQA 2.0 contains a wide variety of question kinds, such as yes/no, numerical, and other sorts that call for concise, targeted responses. The variety of these queries necessitates that models comprehend and interpret a wide range of visual elements, from scenes and objects to actions and characteristics. The questions and answers are created by human annotators, who make sure they are realistic and diverse.

This dataset makes models rely more on visual understanding than language shortcuts, which makes it especially useful for researching biases. Through the examination of VQA 2.0, scholars can enhance their evaluation of VQA systems' efficacy and pinpoint domains where prejudices may continue to exist, ultimately assisting in the creation of more just AI models.

3.2.4 OK-VQA

A distinct Visual Question-answering dataset called OK-VQA (Outside Knowledge VQA) is intended to test models by forcing them to use external knowl-

Figure 3.3: Examples of multiple-choice QA from the 7W question categories.: Random examples from VQA 2.0 dataset.[29]

edge that extends beyond what is apparent in the images. Unlike typical VQA datasets, OK-VQA contains questions that require additional general knowledge or common sense to answer, rather than being able to be answered merely based on the visual content of the image.

About 14,000 questions make up the dataset, which was assembled from more than 14,000 photos taken from the COCO dataset. Every question is designed so that providing an accurate response necessitates the use of outside knowledge, which can range from historical details and cultural allusions to common sense. For example, queries may concern the purpose of an object or the importance of a scene, which cannot be deduced from the picture alone. These questions and responses are thoughtfully crafted by annotators, who make sure they are varied and pertinent. Because of this extra layer of complexity, OK-VQA is especially useful for extending the capabilities of VQA models and promoting the creation of systems that can combine large amounts of external knowledge with visual input.

Researchers can investigate how successfully VQA models generalise and

apply a larger understanding to visual content by utilising OK-VQA. As a result, this dataset is essential to improving AI's comprehension and interpretation of challenging real-world situations.

Figure 3.4: Examples of multiple-choice QA from the 7W question categories.: Random examples from VQA 2.0 dataset.[23]

3.3 Methodology

I study racial and gender bias in Visual Question Answering (VQA) datasets in my thesis by carefully adhering to the technique described by Hirota, Nakashima, and Garcia (2022). They offer important insights into the representation of gender and race in these datasets by using a methodical approach to analyse biases found in VQA systems.

This process begins with the selection of samples from each dataset that specifically address gender (questions for women and men) or race and ethnicity (questions for racial groups). Utilising a rule-based methodology, I ascertain and categorise these samples to guarantee that the examination concentrates on inquiries that are specifically associated with racial and gender biases in VQA.

Then, using the Hirota paper's referenced earlier research as a guide, I perform a thorough investigation of gender bias in the VQA datasets. I categorize

samples according to gender-specific terms found in the questions using a binary classification of gender that divides categories into men and women. This classification makes a more in-depth analysis of the gender representation in the datasets possible.

The results of this gender bias study are presented, offering information on how questions about women and men are distributed throughout the datasets. As the Hirota study points out, this approach frequently shows differences in gender representation, with some genders being marginalized or underrepresented in the VQA datasets. Using the Hirota paper's methodology as a guide, I also expand the research to look into racial bias within the datasets. Finding bias tendencies entails locating samples that specifically address race and ethnicity and conducting a comparative study.

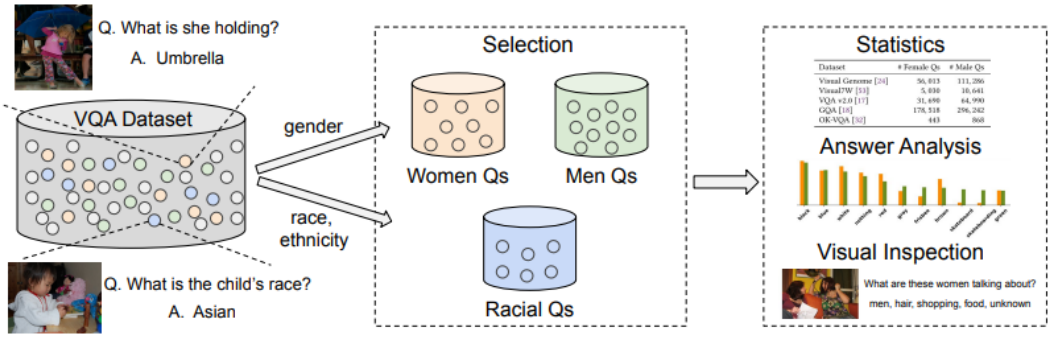

Figure 3.5: Overview of the methodology to detect gender and racial bias

3.3.1 Methodology for gender bias

To respond to enquiries about images by utilising both visual and textual data, the field of Visual Question Answering (VQA) in computer vision and natural language processing has grown in importance in recent years. Nevertheless, VQA datasets and systems are prone to gender bias, just like many other areas of artificial intelligence. In VQA systems, gender bias has the potential to reinforce prejudices, marginalise specific groups, and provide immoral or erroneous AI system results. Our goal in this work is to examine gender bias in VQA datasets, with an emphasis on questions that specifically mention gender.

Binary Classification of Gender Using NLP Techniques

We use a binary gender categorisation approach, which divides people into two groups based on gender: men and women, to examine gender bias in Visual Question Answering (VQA) datasets. Although it has limitations, the traditional binary perspective of gender serves as the foundation for this classification, which is consistent with the methodology used in earlier VQA domain research [21, 59]. In this case, the categorisation does not represent the gender identities of the individuals involved; rather, it represents the perceived gender as determined by the VQA dataset annotators using the language signals included in the questions. This study aids in highlighting the possible biases in the language used by annotators as well as how they formulate their enquiries.

Rule-Based Classification Through Linguistic Parameters

Our method is based on locating particular gendered phrases in the queries related to the pictures. There are two separate categories for these terms: one for women and one for men. Words like "woman," "girl," "female," and pronouns like "she" and "her" are used in reference to the women category. Similar terminology and pronouns are used for the men category, such as "man," "boy," and "male," as well as "he" and "him." When creating enquiries about images based on visual clues, annotators utilise these phrases. For instance, because it specifically utilises the word "woman," a question like "What is the woman doing in the image?" would be categorised as a "women question." Similar to the previous example, a question like "What is the man holding?" would be classified as a "men question" since it uses the word "man."

We can use fundamental Natural Language Processing (NLP) techniques, including Bag of Words (BoW), to perform this classification. In order to create a vocabulary of all the unique terms in a given dataset and encode each question as a vector of word frequencies, or binary values, corresponding to the presence or absence of those words, BoW is a straightforward text representation technique. To achieve our goal, we employ the Bag of Words model to identify any instances of gendered language in each question. In this instance, the BoW

model functions as follows:

- **Tokenisation:** To prevent case sensitivity, divide each question into discrete words, or tokens, and change every word to lowercase. Establish a lexicon that includes any term that can be gendered from the "men" and "women" categories. This is simply a list of terms that has been predetermined based on linguistic clues.
- **Binary Encoding:** Create a vector for each question, with each element denoting whether a word from the lexicon is present or absent.
  - The matching vector element is set to 1 if a word from the "women" list appears in the question.
  - Similarly, words from the "men" list would also result in an entry.
- **Categorisation:** Predicted based on the generated vectors:
  - If the question exclusively contains terms from the women list, then it should be categorised as a "women question."
  - If the question contains solely terms from the men list, it should be categorised as a "men question."
  - If a term appears in neither list or in both lists, then the query should be marked as "excluded."

Gender Bias and Perceived Gender

It is crucial to note that our research is predicated on the gender that VQA annotators assigned to the subjects, not on the gender identities of the people portrayed in the pictures. This distinction is important because VQA datasets are annotated by humans, who frequently base their conclusions about the gender of the subjects in the photographs on visual cues like attire, hairstyles, or other stereotyped markers of gender. Therefore, rather than accurately representing gender diversity, the gender bias found in these datasets may be a reflection of cultural assumptions and biases. Relying on perceived gender can also result in the misrepresentation of some groups or the exclusion of others. For instance, people who don't fit into stereotypical gender roles could have their gender

incorrectly reflected in the dataset or be misclassified altogether. As a result, prejudices and negative stereotypes may be strengthened in AI systems that use these datasets for evaluation and training.

Excluding Non-Gendered Questions

We have removed from our analysis a large number of questions from VQA datasets that do not contain any gendered phrases. Generally speaking, these queries could be along the lines of "What is the person holding?" or "Where is the person standing?" These questions don't immediately contribute to the examination of gender bias because they don't make any explicit mention of gender. Nonetheless, it is important to remember that prejudices could exist even in the absence of directly gendered language. For instance, some tasks or behaviors might be more frequently linked to one gender than the other, giving rise to implicit gender presumptions that our rule-based method cannot capture. That being said, we restrict our analysis to questions that make explicit reference to gender. This gives us a better understanding of how gender is represented in VQA datasets by enabling us to concentrate on questions where the annotator's gender perception is explicitly communicated.

Gender Bias Findings Across Four VQA Datasets

Across four Visual Question Answering (VQA) datasets—Visual Genome, Visual7W, VQA 2.0, and OK-VQA, our study reveals a substantial gender bias. We found different patterns in the representation of gendered roles, acts, and activities by classifying questions according to the usage of gendered terminology. The representation of gender and its consequences for AI systems trained on these datasets are the main topics of our discussion below, which covers our findings across all datasets.

1. Visual Genome

Analysis of the Visual Genome dataset is given below:
Number of Men Questions: 109,652

Number of Women Questions: 55,808
Male-to-Female Ratio(MOW): 1.9
Total Gendered Questions: 165,460
Total Questions in Dataset: 11,445,322
Gendered Question Ratio: 11.45

Men are referenced almost twice as frequently as women, according to the male-to-female ratio of 1.9, which suggests a significant gender imbalance. This disparity highlights social preconceptions and raises questions about potential biases that AI systems trained on this dataset may pick up, particularly concerning gender role assignment.

2. Visual7W

Analysis of the Visual7W dataset is given below:
Number of Men Questions: 23,877
Number of Women Questions: 11,781
Male-to-Female Ratio(MOW): 2.0
Total Gendered Questions: 35,658
Total Questions in Dataset: 327,936
Gendered Question Ratio: 10.87

With a male-to-female ratio of 2.0, or twice as many references to men as to women, Visual7W shows a bigger gender gap. Similar to Visual Genome, men were asked questions regarding activities that are usually associated with men, such as being outdoors and having physical strength, while women were more often asked questions about appearance and taking care of others. The roles that society expects of men and women are reflected in this pattern of biased gender association, which runs the risk of using AI systems to reinforce these biases.

3. VQA 2.0

Analysis of the VQA 2.0 dataset is given below:
Number of Men Questions: 64,479
Number of Women Questions: 33,643
Male-to-Female Ratio(MOW): 1.9
Total Gendered Questions: 98,122
Total Questions in Dataset: 658,111
Gendered Question Ratio: 14.91

At 1.9 male to female, VQA 2.0 exhibits gender bias in line with the earlier datasets. The dataset exhibits a little greater percentage of gendered questions (14.91although the correlation between males and physical or technical jobs and women and caregiving or appearance-related tasks remains. Because of this bias, AI systems may process or respond to queries about gender-specific activities in a way that favors traditional gender roles.

4. OK-VQA

Analysis of the OK-VQA dataset is given below:
Number of Men Questions: 840
Number of Women Questions: 457
Male-to-Female Ratio: 1.8
Total Gendered Questions: 1,297
Total Questions in Dataset: 14,055
Gendered Question Ratio: 9.23

A smaller sample with a lower overall gendered question ratio of 9.23is presented by OK-VQA. The male-to-female ratio of 1.8 indicates that men are mentioned more frequently than women, indicating that bias still exists. Gender stereotypes are maintained in this dataset as well; men are more frequently linked to technical and outdoor activities, while women are more frequently associated with caregiving and queries about looks.

The important metrics for every dataset are compiled in the table below. Finding out how often questions in these databases refer to men versus women and if certain behaviours, positions, and activities are excessively connected with one gender are the main goals. Concerns regarding the fairness of AI systems trained on these datasets are raised by our analysis, which shows a significant pattern of gender bias that resembles larger cultural preconceptions.

Dataset	**Num.Men Qs**	**Num.Women Qs**	**MOW**	**Num. Gender Qs**	**Total Qs**	**Ratio (%)**
Visual Genome	109,652	55,808	1.9	165460	1.7M	11.45%
Visual7W	23,877	11,781	2.0	35,658	327K	10.87%
VQA 2.0	64,479	33,643	1.9	98,122	658K	14.91%
OK-VQA	840	457	1.8	1,297	14k	9.23%

Table 3.2: Statistics of women/men questions (Women Qs/Men Qs) in VQA datasets. MoW is the number of men questions over the number of women questions. Ratio is the number of gender questions (Num. Women Qs + Num. Men Qs) over the total number of questions (Num. Total Qs)

Comparison of result

The experiment's outcomes, which are shown in the table above 3.2, closely resemble the findings of the original study that our goal was to recreate. Although there are significant variances in the proportion of questions classified as belonging to men or women, the differences are mostly explained by the particular keywords we used to identify questions that are relevant to gender. In our method, we classified queries according to gender by using a specified set of keywords. The scope and specificity of the keywords employed may be the cause of the modest variation in the results, especially in the number of questions asked of men and women. It is feasible that the classification of the

questions could produce different results if a larger or more broad list of keywords were used. The total ratio of males to women questions is still rather close to what was reported in the original research, even with these little differences. The broad pattern's consistency supports both the validity of our methodology and the initial study's conclusions. It also emphasises how crucial it is to carefully choose terms that appropriately reflect the relevant gender groups because it shows how sensitive the findings are to keyword selection. In the future, modifying the keyword selection procedure might improve the analysis even more and possibly result in a more accurate classification of questions based on gender. But even taking into account these factors, our findings essentially confirm previous research's findings, showing that women's under-representation in VQA datasets is a persistent problem that holds true in various experimental configurations.

Male-Dominated Representation in VQA Datasets

Men are regularly referenced in more queries than women in all four datasets. In the Visual Genome dataset, for example, there are 55,808 questions about women and 109,652 questions about men. This means that the male-to-female ratio is 1.9. In other words, men are mentioned almost twice as frequently as women. In a similar vein, the Visual7W dataset reveals an even more pronounced difference, with men cited twice as frequently as women (a ratio of 2.0). This pattern holds for both OK-VQA and VQA 2.0, where the male-to-female ratios are 1.8 and 1.9, respectively. These figures show a notable gender gap that may have an impact on how AI systems understand and react to gender-related queries. When dealing with female-related content, a model trained on these datasets is more likely to come across and become accustomed to scenarios involving men, which could result in biased predictions and reactions.

Gendered Question Distribution

Within each dataset, the percentage of gendered questions is a crucial component of the research. Questions that specifically mention men or women, like "What is the man doing?" or "What is the woman holding?" are considered gendered. Among the four datasets, the VQA 2.0 dataset had the greatest percentage of gendered questions at 14.91OK-VQA (9.23%), Visual7W (10.87%), and Visual Genome (11.45%) come next.

The presumptions found in these questions are alarming, even though the overall percentage of gendered questions seems low in comparison to the total number of questions. Gender-specific questions often reinforce conventional roles by associating males with physical prowess, outdoor pursuits, and technical skills, and women with caregiving, household chores, and looks. The stereotyping in these datasets is a reflection of societal standards and can reinforce prejudices in the systems they assist in training.

Pseudo-Algorithm: Gender Bias in Question-Answer Pairs

First, the general question counter, the men's and women's question counters, and the pseudo-algorithm for analysing gender bias in question-answer pairings are initialised to zero. To store the answers to the questions pertaining to men and women, respectively, two empty lists are constructed. Two distinct lists include the definitions of gender-specific keywords: `men_keywords` containing terms like 'man', 'boy', 'he', 'his', etc., and `women_keywords` including terms such as 'woman', 'girl', 'she', 'her', etc.

Next, the algorithm goes over every element in the dataset one by one. It verifies whether a list of question-answer pairs is present for every item. We examine every question-answer pair to see if it has any keywords related to males or women. The matching answer is appended to `men_answers` list and the men counter is incremented if the question contains keywords relating to men. Con-

versely, if the question contains female-related keywords, the women counter is incremented, and the corresponding answer is appended to the `women_answers` list.

Following this, the algorithm calculates the frequency of each answer using counters: `men_answer_counter` for answers in men-related questions and `women_answer_counter` for answers in women-related questions. It then computes the ratio r as

$$r = \frac{\text{men\_count}}{\text{women\_count}}$$

handling cases where there are no women-related questions to avoid division by zero.

The bias score (BS) for each answer is calculated using the formula:

$$\text{BS}(a, q_m) = \frac{c(a, q_m)}{c(a, q_m) + r \cdot c(a, q_w)}$$

where $c(a, q_m)$ represents the frequency of the answer a in men-related questions, $c(a, q_w)$ represents the frequency of the answer a in women-related questions, and r is the ratio of men to women questions.

Both the `men_answer_counter` and the `women_answer_counter` provide the top 20 most often provided answers. The bias score is calculated for each top response in questions of males after retrieving the count of the same response in questions pertaining to women. For further examination, the outcomes which include the solution, its count, and the bias score—are saved. The count in questions of males is obtained for every most popular response in questions about women and the bias score is adjusted using:

$$\text{BS}(a, q_w) = 1 - \text{BS}(a, q_m)$$

This adjusted bias score is also stored for subsequent analysis.

Reinforcement of Stereotypes

We find that jobs and activities associated with women are disproportionately linked to tasks traditionally performed by women across all datasets. enquiries pertaining to caregiving, cooking, or personal attractiveness, for example, are typically aimed at women, whereas enquiries about physical strength, outdoor activity, or technical jobs are more likely to be directed toward men. For instance, in the Visual Genome dataset, women are typically mentioned in questions about caring for others or household chores. In contrast, men are usually mentioned in questions about manual labour or outdoor activities.

There may be significant ramifications to this pattern of skewed correlations between gender and particular roles. When dealing with users, AI systems educated on these datasets may unintentionally reinforce negative stereotypes by assuming that men are more likely to conduct physically demanding activities and women are more likely to take on caregiving responsibilities. This is especially concerning for applications where the persistence of these biases can sway responses or outcomes, such as virtual assistants or automated decision-making tools.

Comparison of the result frequents answer

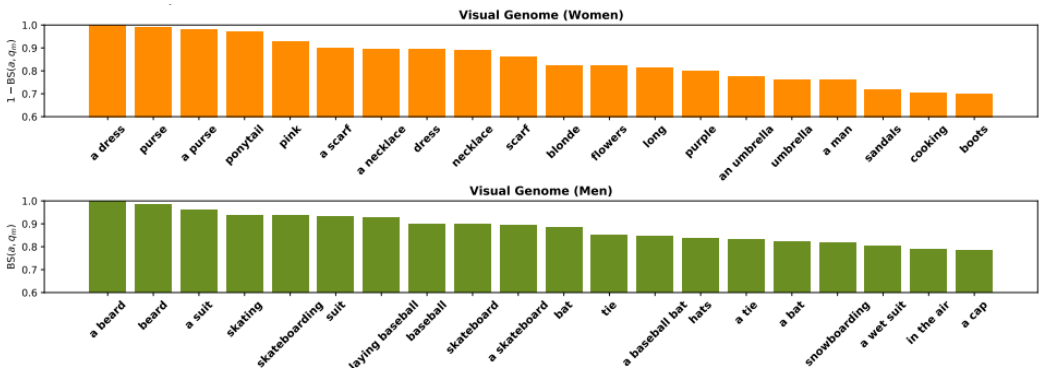

Figure 3.6: Top 20 frequent answer for men and women related question from research paper

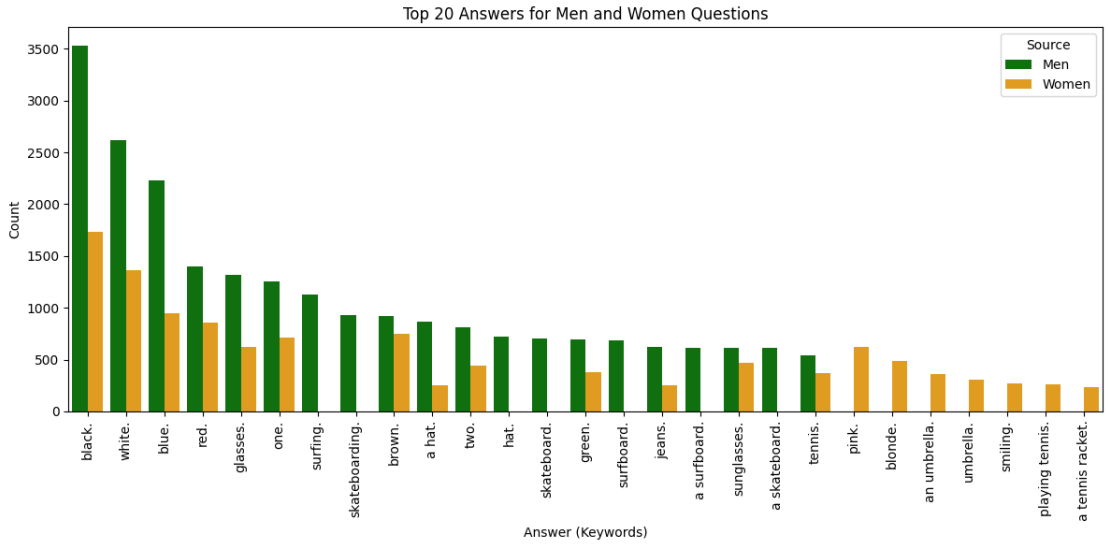

Figure 3.7: Top 20 frequent answer for men and women related question-experimented

In order to identify any potential gender bias, we thoroughly examined the top 20 responses to questions pertaining to both men and women in the Visual Genome dataset. Our main goal was to assess the dataset's handling of gender-specific enquiries and look for any indications of biassed responses. In order to measure this, we focused on the recurrent themes in responses to questions pertaining to women and men and created a bias score for each gender category. We sought to identify any gender inequalities or stereotypes by looking at the top 20 frequently provided responses. After executing this analysis, we discovered that questions about men and women produced distinct answers, many of which reflected prevalent gender roles or preconceptions. Through the comparison of answer distributions and the identification of potential areas where gendered assumptions may impact the dataset's output, the bias score enabled us to evaluate the extent of this inequality. It's interesting to note that our analysis's findings closely match those in the study that we set out to duplicate which is given in 3.6 and 3.7. This closeness indicates the efficacy and dependability of the methodology we employed, which included the rule-based method for detecting gender-related questions and computing bias ratings. The results of the prior study and our work are consistent, which supports the

validity of the methods we used. The consistency of the results further suggests that the keywords we chose and the way we categorised questions about gender were suitable. We do observe, however, that the exact keywords employed in the analysis may have had a small impact on the proportion of enquiries pertaining to males versus women. Although more focused or broader keyword lists may produce different results, overall the gender distribution of responses is consistent between the two studies. This shows that systematic analysis can reveal hidden biases and assist guide changes in VQA datasets, further validating the validity of the bias detection approaches used here.

3.3.2 Conclusion from the Differences in Answers Between Men and Women Questions

Several patterns that offer interesting information on gender-answer correlations are shown by examining the top 20 frequently provided responses to questions pertaining to men and women in the dataset. These patterns may be a reflection of underlying stereotypes or social norms.

1. Common Answers with Similar Frequencies

Certain responses, such "black," "standing," or "walking," typically come up in response to enquiries about both men and women. This implies that some visual traits are perceived and explained comparably for both genders, particularly when the issue is posed in a neutral context (e.g., dress colour, activities that apply to both). The responses exhibit comparatively balanced bias scores (BS) in the vicinity of 0.5, signifying that there is no significant correlation between the observed phenomena and gender. Also like "Sitting" and "walking" are common answers to enquiries about both men and women. For instance, the answer ratio for "sitting" would be

$$\text{Ratio} = \frac{50}{25+1} = \frac{50}{26} \approx 1.92$$

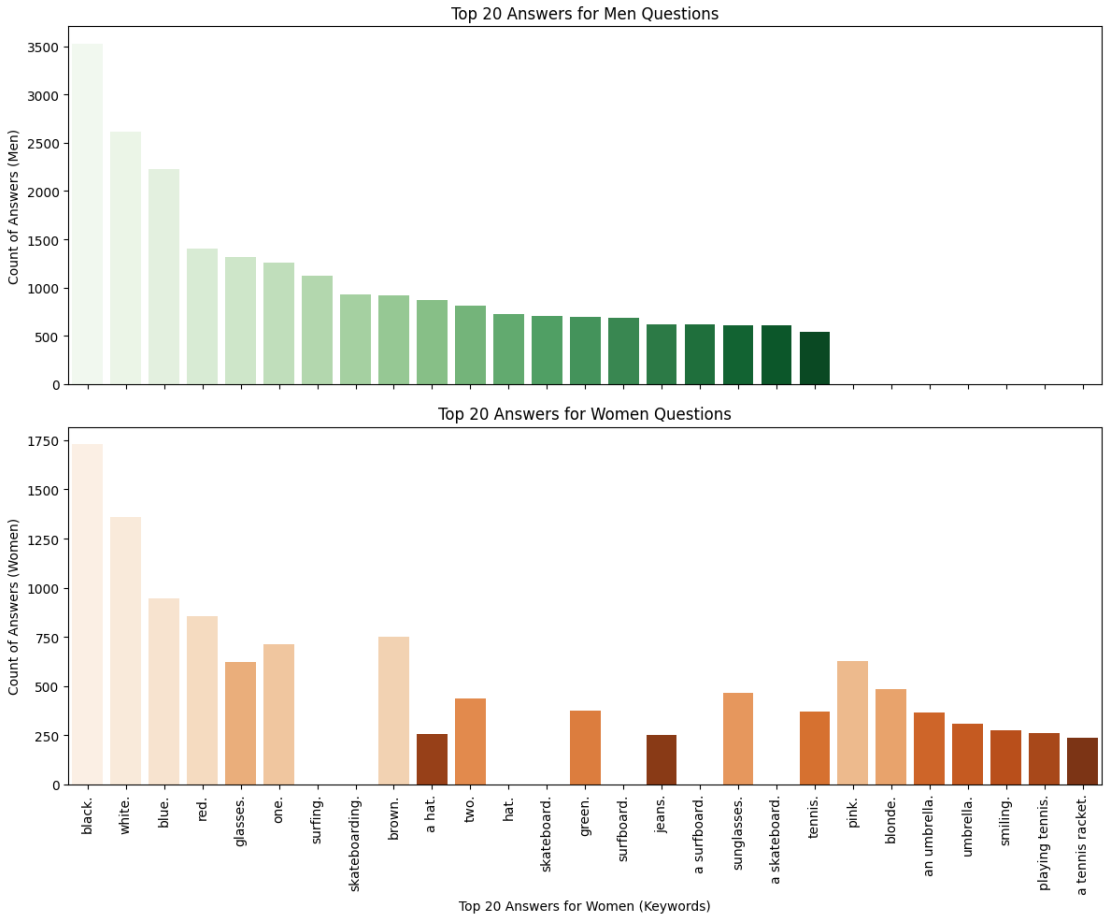

Figure 3.8: Top 20 frequent answer for men and women related question (Visual Genome)

if the men count was 50 and the women count was 25. This means that "sitting" is almost twice as likely to appear in queries about men.

2.Answers Biased Toward Men

A lot of the responses are biassed towards questions pertaining to guys. In questions regarding men, for instance, responses like "running," "backpack," and "cap" tend to occur more frequently. The majority of these responses centre

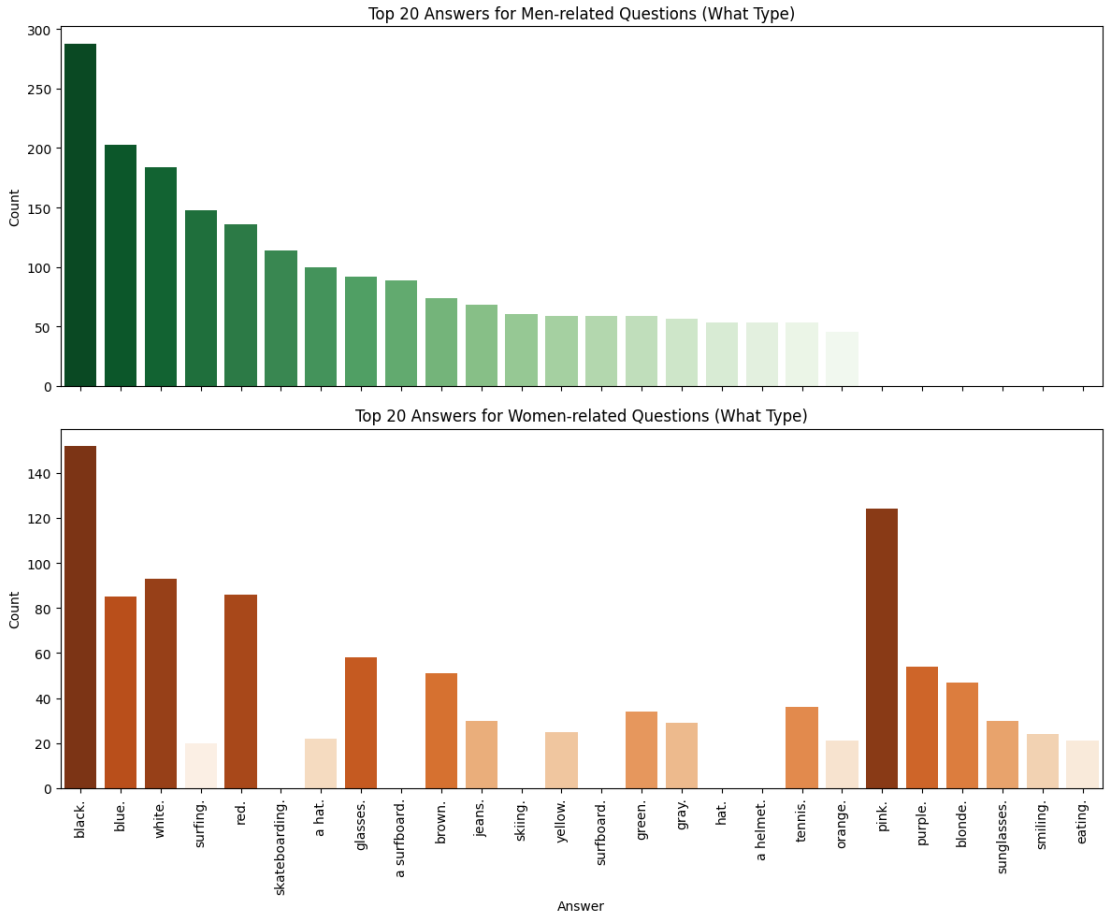

Figure 3.9: Top 20 frequent answer for men and women related question (Visual7w)

around activities and accessories, indicating that males are more often depicted in situations that require action or with items that are generally connected to mobility or usefulness. Sports-related responses like "playing football," "Skateboard", "running," or "holding a ball" are also largely seen in queries pertaining to men. This is consistent with prevailing social norms that men participate in sports and physical activities at a higher rate than women. These responses have bias scores that are closer to 1, indicating a strong association with queries about men. The focus on sports and physical exercise speaks volumes about how men are viewed through the perspective of physical appearance and activity, which feeds into popular conceptions of masculinity and athleticism. answers like

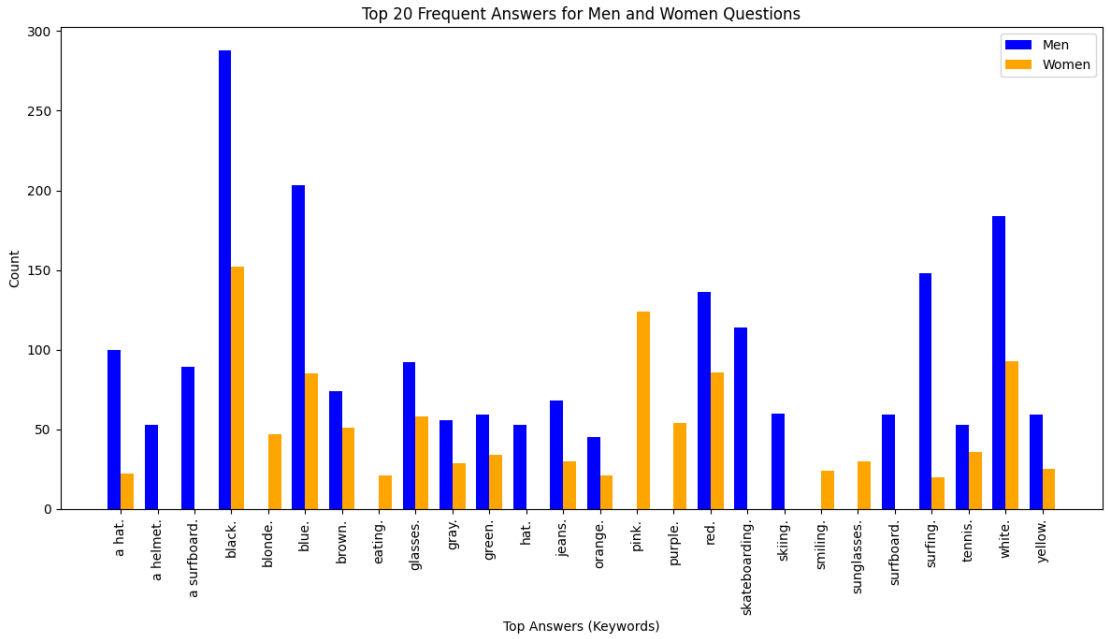

Figure 3.10: Comparison of top answer between Men and Women Questions (Visual7w)

"shopping" or "childcare" may be more typical in queries about women, whereas some answers[31], like "football" or "car" for questions concerning males, may be unusual to topics regarding men only.The absence of responses from women on sports-related topics draws attention to a difference that is a reflection of the societal perception that sports are more closely linked to men.

4. Observations on Visual Bias

Answers skewed towards women typically concentrate more on looks and accessories, whereas answers skewed towards men are frequently linked to acts and activities. This signifies a pattern in which women are usually defined in terms of their appearance or the object they are holding, while men are usually connected with movement or activity. This observation is consistent with broader tendencies in society, where women are frequently perceived through

the filter of looks or passive positions, whereas males are perceived as more action-oriented.

5. Implications for Gender Stereotypes

Gender stereotypes may be reflected in the differences in the sorts and frequency of responses given to enquiries by men and women. Women are typically portrayed in roles that emphasise caring, attractiveness, or accessories (e.g., "holding," "wearing a dress"), whereas men are more commonly connected with physical actions or roles that involve movement and practicality (e.g., "running," "wearing a cap"). This kind of visual bias in descriptive language has the ability to affect how people view the roles of men and women in various contexts by reinforcing gender roles and stereotypes. Different patterns of gender bias in the description of visual aspects are revealed by analysing frequently given responses to queries pertaining to men and women. Responses that are biassed towards men tend to focus on action and accessories like hats and bags, whereas answers that are biassed towards women tend to focus on attractiveness and passive behaviours like holding objects. Addressing gender discrimination in AI systems and visual representation requires a knowledge of these biases, which may be reflected in these discrepancies. To neutralize gendered descriptions and prevent the reinforcement of prejudices in future datasets and applications, addressing such biases would require deliberate effort.

There are clear trends in how men and women are portrayed when the frequency and bias scores of the answers in this dataset are examined. those about men typically highlight sports and physical activities, which are conventionally linked with masculinity, whereas those relating to women typically center around personal or everyday situations. These patterns present important concerns regarding the inclusion and portrayal of multiple gender identities in visual question-answering frameworks, in addition to reflecting current societal standards. It is imperative to address these biases to create more egalitarian

Figure 3.11: Examples of gender stereotypes in VQA 2.0[6] (above), OK-VQA[24] (middle), and Visual Genome[18] (below)

datasets that correctly reflect the whole range of human experience.

3.3.3 Methodology For Racial Bias

The methods used for the gender bias study will be applied to the analysis of the racial biases in the Visual Question Answering (VQA) datasets in this section. In particular, to find and measure racial bias in question-answer pairs, we will examine four VQA datasets. Questions that specifically mention race, ethnicity, or nationality will be extracted using a string-matching method. To identify relevant question-answer pairings and create a racial sample subset, an identified set of keywords will be employed. To find any differences in representation among various racial and ethnic groups, we will examine the frequency and distribution of responses pertaining to race. This research will

shed light on whether specific demographic groups—like people who identify as White, Black, Asian, or Hispanic—are over- or under-represented in the datasets. Conducting such an inquiry is crucial in recognising the degree of racial bias that exists in VQA datasets and its possible influence on model performance, specifically for members of excluded communities.

Racial Bias Findings Across Four VQA Datasets

Our analysis identifies a considerable racial bias across four Visual Question Answering (VQA) datasets: OK-VQA, Visual Genome, Visual7W, and VQA 2.0. By examining questions that specifically mention race, ethnicity, or nationality, we were able to identify significant patterns in the representation of different racial and ethnic groupings. Through the use of racial terminology to categorise questions, we were able to identify differences in the representation of various demographic groups. The main focus of our discussion below, where we describe our findings across all datasets, is how race is represented and what that means for AI systems trained on these datasets.

1. Viaual Genome

Analysis of the Visual Genome dataset is given below:
Num. Racial Qs: 658
Num. Total Qs: 1.7M
Ratio (%): 0.038%
Out of 1.7 million questions, 658 are related to race, according to the Visual Genome dataset study. This yields a ratio of roughly 0.038%, suggesting that racial enquiries make up a relatively minor fraction of the dataset as a whole.

This low percentage would indicate that there was less emphasis on racial issues in the questions that were asked, which could be a reflection of more general patterns in data collection or the kinds of questions that are given priority in visual datasets. Comprehending the question types' distribution is essential to guaranteeing equitable and varied representation in AI models and mitigating potential biases resulting from marginalized viewpoints in training data.

2. Visual7w

Analysis of the Visual7W dataset is given below:
Num. Racial Qs: 93
Num. Total Qs: 327k
Ratio (%): 0.03%
Based on the Visual7W dataset analysis, there are 93 questions about race out of 327,000 total questions, or roughly 0.03% of the total questions. This suggests that racial queries make up a very small portion of the whole dataset.

3. VQA2.0

Analysis of the VQA2.0 dataset is given below:
Num. Racial Qs: 510
Num. Total Qs: 658k

Ratio (%): 0.08%
Based on the dataset analysis, there are 510 questions about race out of 658,000 total questions, or around 0.08% of the total. Compared to some other datasets, this proportion shows a little higher representation of racial enquiries; however, it still indicates that such questions represent a relatively small percentage of the overall content.

4. OK-VQA

Analysis of the OK-VQA dataset is given below:
Num. Racial Qs: 30
Num. Total Qs: 14k
Ratio (%): 0.21%
According to the OK-VQA dataset study, there are 30 racial questions out of 14,000 total questions or roughly 0.21% of the total. This proportion shows that racial enquiries are only slightly represented in the dataset.
Racial questions represent a very small portion of all questions in all datasets, which emphasises the continued difficulties in obtaining complete diversity.

Intentional data gathering tactics that prioritise different views are necessary to develop more equitable and successful AI systems, as evidenced by the under-representation of these enquiries. Overcoming these gaps is critical to enhancing AI systems' comprehension and management of diverse racial social concerns. The important metrics for every dataset are compiled in the table below. Finding out how often questions in these databases refer to race or ethnicity. Racial

Dataset Name	Racial Questions	Total Questions	Ratio (%)
Visual Genome[18]	658	1.7M	0.038%
Visual7W[31]	93	327k	0.03%
VQA2.0[6] 4	510	658k	0.08%
OK-VQA[24]	30	14k	0.21%

Table 3.3: Summary of Racial Questions in Various Datasets

terms are present in most databases. 3.3 displays the number of racial samples for each dataset. Although there are racial questions in every dataset examined, the proportion of racial questions varies. Specifically, there are comparatively few racial questions in Visual7W. This can be explained by the fact that samples that were deemed inappropriate or unnecessary were eliminated throughout the annotation process, and annotators for Visual7W were provided clear guidance on how to generate visually grounded queries. In contrast, datasets like VQA 2.0, Visual Genome, and OK-VQA show a higher proportion of racial questions. However, as the absolute number of racial samples in OK-VQA is relatively small, the focus of our analysis is primarily on VQA 2.0 and Visual Genome, where there is a more substantial representation of racial questions. These datasets provide a richer source for studying the treatment of racial and ethnic groups in visual recognition tasks.

A methodical technique was used to find questions pertaining to race and ethnicity in the VQA dataset. The procedures involved are described below:

Methodology: Analyzing Racial Bias in Visual Genome Dataset

The dataset used in this study is the *Visual Genome Question-Answer (QA) dataset*[18]. This dataset contains pairs of images, questions, and corresponding

answers.

Racial and Ethnicity-Related Question Identification

The complete dataset is represented as $D = \{(q_i, a_i)\}_{i=1}^{N}$ in which q_i represents a question and a_i represents its matching answer. Every question in this dataset, denoted by q_i, is a natural language query linked to an image, and its associated answer is represented by a_i. Only a portion of the questions will specifically mention race, ethnicity, or nationality, but the questions and answers are still related to the content of the images.

We establish a new subset $D_r \subset D$, which contains only the question-answer pairs that are relevant to these concepts, in order to identify the subset of question-answer pairings where the question q_i addresses race or ethnicity. The *racial sample subset* is the name we give to this subgroup. Within this category, enquiries can be related to racial or ethnic characteristics, for example, enquiring about the race of an individual depicted in the picture.

String Matching for Racial/Ethnic Questions

We use a *string-matching technique* based on *keyword detection* to identify questions that mention race or ethnicity. This method filters the pertinent questions using a predetermined list of phrases pertaining to race and ethnicity.This stage essentially uses keyword matching along with basic natural language processing (NLP) approaches to categorise questions as about race. This type of rule-based filtering-based text classification We establish a set of keywords K_q that contain words that specifically mention nationality, race, or ethnicity. These words have been chosen with care to encompass a wide variety of racial and ethnic characteristics. The following terms are included in the set K_q for this study:

$$K_q = \{\text{"race"}, \text{"ethnicity"}\}$$

Subset Definition Based on Keyword Detection

We determine if the question q_i has any of the keywords from the set K_q for each question-answer pair (q_i, a_i) in the original dataset D. Tokenisation, a *natural language processing (NLP)* approach that divides the query into discrete words or phrases, can be used to carry out this operation. A question-answer pair is deemed relevant to race or ethnicity if any token in the question q_i matches one of the keywords in K_q.

The racial sample subset D_r has the following mathematical definition:

$$D_r = \{(q_i, a_i) \mid q_i \cap K_q \neq \emptyset\}$$

In this equation:

- D_r is the subset of question-answer pairs where the question references race or ethnicity.
- $q_i \cap K_q \neq \emptyset$ indicates that the intersection of the words in question q_i and the set K_q is non-empty, meaning at least one keyword is found in the question.

Implementation Details

1. **Tokenisation**: A list of words, or tokens, is created from each question q_i. This is a typical NLP preprocessing phase where sentences are broken down into their individual words.
2. **Keyword Matching**: We determine whether any tokens in each tokenised query match a keyword in K_q. A set-based membership check or a search algorithm can be used to accomplish this.
3. **Subset Construction**: The matching procedure adds the matching question-answer pairs (q_i, a_i) to the subset D_r after determining which questions are relevant. We use this subset for additional analysis.

Purpose of Subset Identification

Determining the subset D_r is essential to understanding the Visual Genome dataset's representation of race and ethnicity. We can examine the nature and distribution of responses about race and ethnicity by separating these questions. In order to identify any underlying biases that can affect machine learning models trained on this data, this subset offers insights about the wide range of representation in the dataset.

To summarise, all question-answer pairings pertaining to race or ethnicity are filtered using the string-matching technique, which is based on keyword identification. Because it allows us to concentrate on the portion of the data that is directly related to demographic representation, this approach is essential for the downstream study of racial bias.

Racial/Ethnicity-Related Answer Detection

In machine learning terminology, this step involves *label filtering* for answers that reference race, ethnicity, or nationality. From the subset D_r, we filter out irrelevant answers, such as those that are binary responses ("yes", "no") or describe colors and objects (e.g., "red", "blue", "building"). Removing irrelevant answers (such as "yes" and "no") is essential to ensure that only meaningful demographic information is retained. This is an application of data preprocessing in machine learning, improving data quality for further analysis. We create a set K_{ignore} that contains these *irrelevant terms*:

$$K_{\text{ignore}} = \{\text{"yes"}, \text{"no"}, \text{"red"}, \text{"green"}, \ldots\}$$

Next, we construct a set K_a, containing *racial/ethnic terms* found in the dataset and literature:

$$K_a = \{\text{"white"}, \text{"black"}, \text{"caucasian"}, \text{"african"}, \text{"asian"}, \text{"hispanic"}, \ldots\}$$

For each question-answer pair $(q_i, a_i) \in D_r$, if $a_i \cap K_a \neq \emptyset$ and $a_i \cap K_{\text{ignore}} = \emptyset$, we retain the answer as relevant to race or ethnicity.

The filtered racial-related answers can be represented as:

$$D_f = \{a_i \mid a_i \cap K_a \neq \emptyset,\ a_i \cap K_{\text{ignore}} = \emptyset\}$$

Frequency Analysis of Answers

Once the relevant answers are identified, we compute the *frequency distribution* of each racial or ethnicity-related answer $a_i \in D_f$. This distribution, represented as a histogram, shows the most common racial demographic groups present in the dataset. The *Bag-of-Words (BoW)* model, a key technique for text analysis in NLP, is used to calculate the frequencies of racial/ethnic keywords. This model counts the instances of specific words or phrases. For each answer, we count its occurrences:

$$f(a) = \sum_{i=1}^{|D_f|} \mathbb{1}(a_i = a)$$

where $\mathbb{1}(\cdot)$ is the indicator function, and $f(a)$ gives the frequency of each answer $a \in K_a$. By doing these actions, we may investigate the representation of various racial and ethnic groups in the Visual Genome dataset and identify any potential biases in the information. This mismatch may cause *biased performance* across various demographic groups in downstream machine learning models, especially those applied for tasks like image captioning or facial recognition. We select the *top-10 most frequent answers* for further analysis.

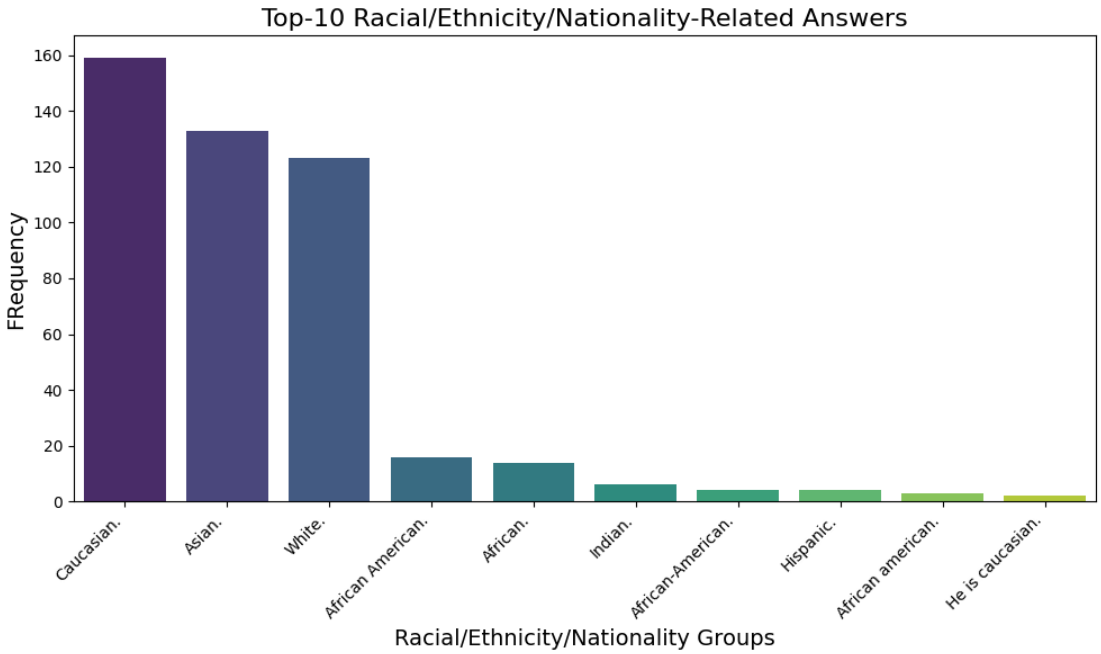

Figure 3.12: Top-10 Racial/Ethnicity/Nationality-Related Answers(VISUAL GENOME[18]

The top ten responses to racial questions from the Visual Genome and Visual7W datasets are shown in Figures 3.10 and 3.10. There is a notable demographic inequality in the responses, according to an analysis of these datasets. remarkably, the group most frequently mentioned is related to White people and includes phrases like "White" and "Caucasian." References to Asian groups, especially those with terminology like "Asian" and "Chinese," immediately follow this.
On the other hand, terminology related to Black people, including "Black," "African American," and "African," as well as phrases related to Hispanic people, such "Hispanic," are much less common. This pattern suggests that there is a significant under-representation of people with darker skin tones in the examined samples.

These differences in representation are not specific to these datasets; comparable patterns have been observed in a number of different computer vision ap-

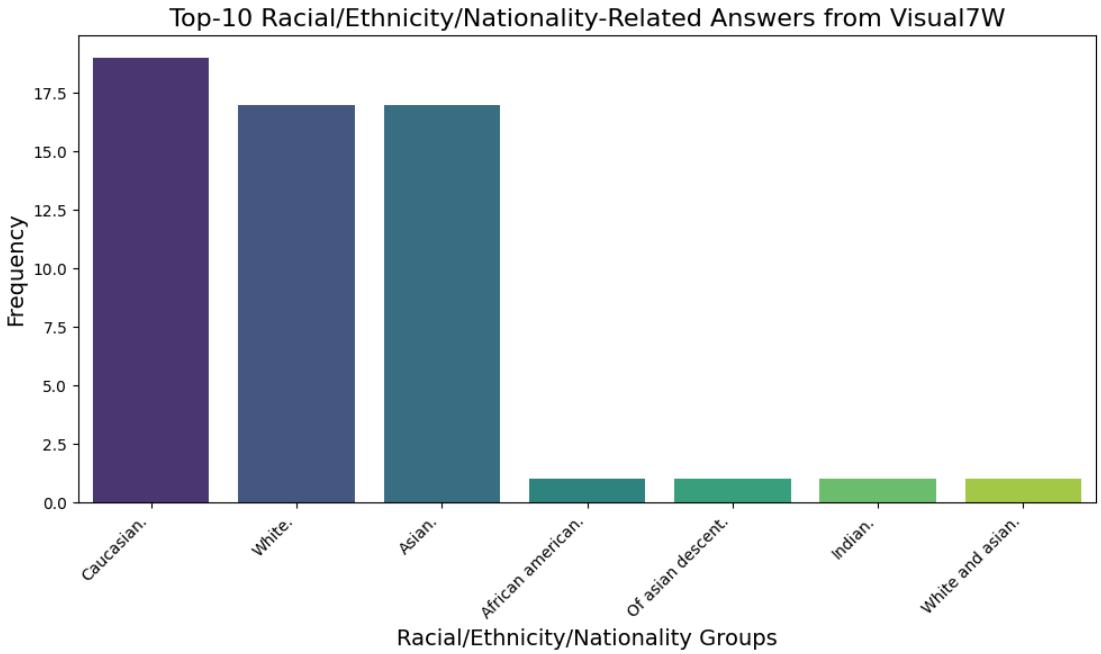

Figure 3.13: Top-10 Racial/Ethnicity/Nationality-Related Answers(VISUAL7w[31]

plications. According to research on facial recognition, for instance, 79.6% of participants in the IJB-A dataset have lighter skin tones [2]; Wang et al., 2019). This pattern is also seen in image captioning[**inproceedings**, 30]databases, where people with darker skin tones are frequently under-represented.

This population mismatch has far-reaching consequences. When analysing photos of people with darker skin tones, models trained on datasets that primarily include lighter-skinned individuals may perform badly, producing biassed results and feeding current stereotypes. In order to create just and accurate machine learning systems that can generalise across a variety of populations, these inequities must be addressed.

Racial-stereotypical example

To determine their potential for harm, we manually examined racial samples from each dataset in our analysis. In order to investigate racial bias, we also performed an intersectional analysis utilising over 300 examples of questions

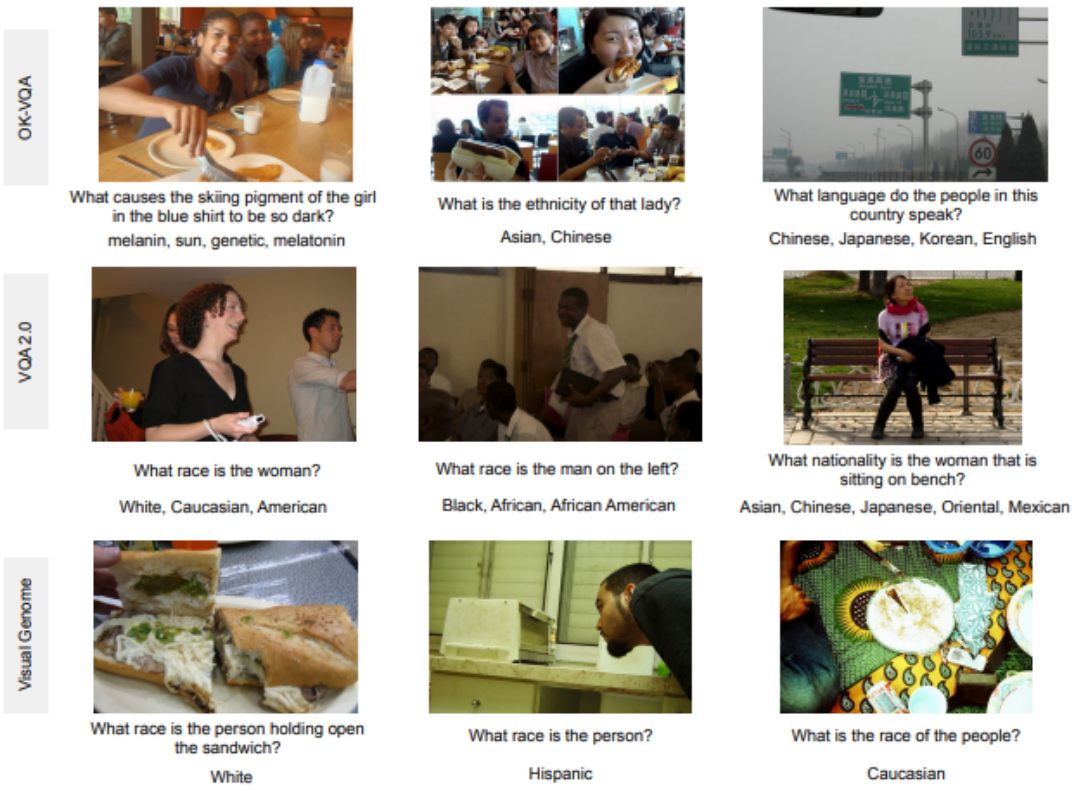

Figure 3.14: Racial samples in OK-VQA[23] (above), VQA 2.0[6](middle), and Visual Genome[18] (below).

regarding men and women in each dataset. Based on this, we were able to distinguish between two main types of racially biassed content:
(1) discriminating samples based on race and
(2) judgement samples that are biassed. various datasets, including VQA 2.0, Visual Genome, and OK-VQA, were found to exhibit these biases; Figure 10 provides various examples. Content that discriminates against certain races was one common instance of racial bias we came across. For example, the query "What causes the skiing pigment of the girl in the blue shirt to be so dark?" in Figure 10's top-left example suggests that lighter skin is the standard, exposing an underlying prejudice that links light skin to a default standard. Another troubling instance can be found in 3.13, second row, right-hand section. In this instance, a question concerning a woman's nationality is asked, and among the

answers is "Oriental"—a term that, although problematic and out-of-date, is nonetheless used 23 times in the VQA 2.0 dataset. It is significant to remember that this term was formally eliminated from federal terminology in the United States in 2016 [11]. Biassed judgement samples were the second kind of bias that we discovered. These samples frequently took the form of questions without any visual background to support the responses that were given. In many instances, assumptions about someone's race, ethnicity, or country were drawn without any consideration for the photograph itself. For instance, in Figure 10, the bottom-left image merely depicts a person's finger tips; nonetheless, despite the lack of any obvious visual cue suggesting race, the answer selected is "White." In a similar vein, the bottom-right image in the same figure depicts a scenario in which a biassed decision is reached in the absence of any hard proof from the visual content. The aforementioned instances highlight the wider issue of racial bias in vision-language models, as questions asked and responses produced may reflect implicit and antiquated social beliefs. In order to address this problem and promote a more inclusive attitude towards AI-based systems, an extensive review of these datasets is necessary to reduce any harmful or biassed content.

3.3.4 Applied methodologies on LLAVA dataset

We are now adding the LLaVA (Large Language and Vision Assistant)[20] dataset to our analysis of gender and racial bias. We seek to determine whether comparable patterns of gender bias appear in this dataset by using the same methodology as earlier datasets. In order to do this analysis, questions will be categorised according to gendered terminology (such as "woman," "man," "she," or "he"), and the ratio of questions pertaining to men to women as well as the overall percentage of gendered questions relative to the entire number of questions will be determined. This investigation aims to ascertain whether gender biases that could reinforce societal preconceptions are present in the LLaVA dataset as well as other VQA datasets. This comparison allows us to assess how the LLaVA dataset's gender representation stacks up against other VQA datasets and consider implications for AI models trained on it. We hope to provide a more comprehensive understanding of gender bias in VQA datasets and investigate if biases seen in earlier datasets can be addressed in more recent ones by including the LLaVA dataset into our analysis. The results of this analysis will also be useful in detecting any enduring trends of gender bias and emphasise the significance of creating more inclusive and balanced datasets to enhance the fairness and precision of AI systems.

LLAVA Dataset

With state-of-the-art performance on 11 benchmarks, the LLaVA-1.5 (Large Language-and-Vision Assistant) dataset offers a state-of-the-art breakthrough in the field of multimodal AI. Using only a single 8-A100 node, LLaVA-1.5 trains very quickly—in less than a day—using publicly available data with very little modification from its predecessor. This efficiency not only demonstrates how well-designed the model is, but it also puts it ahead of competing approaches that use far larger billion-scale datasets. To accomplish comprehensive visual and language understanding, LLaVA is an innovative end-to-end trained large multimodal model that combines Vicuna, a large language model (LLM), with a vision encoder. With this architecture, LLaVA can perform exceptionally well in

many jobs; its remarkable conversation skills are particularly noteworthy, since they bear resemblance to the multimodal version of GPT-4. Because of its architecture, LLaVA can handle extremely complicated tasks including both language and picture data, and it has set a new record for AI performance on Science QA, where it scores a state-of-the-art accuracy of 92.53LLaVA's method of instruction tweaking is one of its main innovations. Although instruction tweaking has been shown to enhance large language models' (LLMs') zero-shot capabilities in the language domain, multimodal AI has not made much use of it. LLaVA leverages machine-generated multimodal instruction-following data to make notable advancements in this field. The LLaVA team generated multimodal language-image instruction-following data using GPT-4, an unprecedented step. The secret of LLaVA's comprehension and reaction to intricate image-based instructions is this innovative method. The LLaVA model is designed to perform well in a wide range of linguistic and visual activities. Its ability to process and react to multimodal inputs smoothly is made possible by the integration of a visual encoder with an LLM. Preliminary tests show that LLaVA performs exceptionally well on multimodal conversation tasks, frequently matching or surpassing multimodal GPT-4 scores. When tested on artificial multimodal instruction-following datasets, LLaVA shows a noteworthy 85.1% relative score above GPT-4 in some scenarios, indicating its resilience and versatility. Furthermore, LLaVA's open-source nature is a noteworthy advancement for the scientific community. To encourage more research and advancement in the multimodal field, the model's coding and the GPT-4-generated visual instruction tuning data are publicly accessible from the creators. Because of this openness, researchers and developers may work together more easily and take advantage of LLaVA's capabilities for a variety of applications, such as language processing and visual understanding. In conclusion, LLaVA-1.5 represents a major step forward in multimodal AI, combining state-of-the-art performance with efficient training and open-source accessibility. Its groundbreaking use of GPT-4 to generate instruction-following data, along with its impressive results in tasks that combine vision and language, position it as a leading model in the field, pushing the boundaries of what AI can achieve in understanding and interacting with the world through both images and text.

We engage with language-only GPT-4 based on the COCO dataset, and we gather a total of 158K unique language-image instruction-following samples: 58K in chats, 23K in comprehensive description, and 77K in complex reasoning. the chart is given below: An assessment collection including thirty previously

Table 3.4: LLaVA Dataset Files Overview

Data File Name	**File Size**	**Sample Size**
conversation_58k.json	126 MB	58K
detail_23k.json	20.5 MB	23K
complex_reasoning_77k.json	79.6 MB	77K

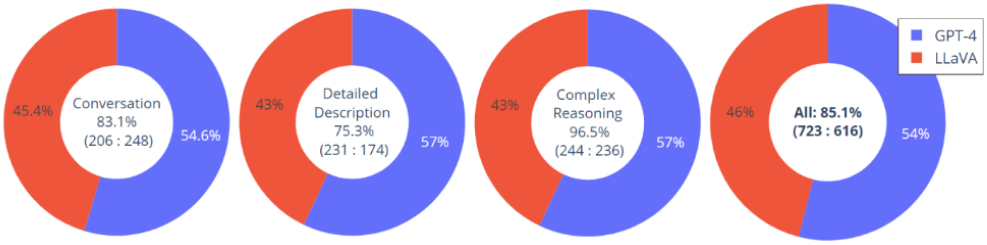

Figure 3.15: Evaluation performance of LLaVA and GPT-4

unreleased photos is created, with each image linked to one of three categories of instructions: dialogue, in-depth explanation, or sophisticated reasoning. This results in 90 new language-image instructions. We test LLaVA and GPT-4 on these, then use GPT-4 to provide a score of 1 to 10 to each respondent's response. Each type's relative score and total score are given. Overall, LLaVA outperforms GPT-4 by 85.1% relative score, demonstrating the effectiveness of the suggested self-instruct approach in multimodal environments.

GENDER BIAS IN LLAVA DATASET

We examined gender-related questions in the LLaVA dataset using the same natural language processing method (NLP) as Hirota's research work. Through the use of a rule-based methodology that categorises questions according to gendered terminology (e.g., "he," "she," "man," "woman"), we were able to discern trends in the gender representation within the dataset. The findings show

patterns that are consistent with those seen in previous VQA datasets, with a clear skewedness in the jobs and activities that are linked to men and women. Similar to other studies, men were more typically linked to physical strength and technical skills, while women were more usually linked to questions about caring for others and household duties. These results highlight the enduring gender biases in several datasets, asking questions on the perpetuation of social stereotypes in AI models trained on these datasets.

Dataset	Men Ques	Women Ques	(MOW)	Gendered Questions	Total Question	Ratio
LLAVA	28,996	17,650	1.64	46,646	361,411	12.91%
Visual Genome	109,652	55,808	1.9	165,460	1.7M	11.4%
Visual7W	23,877	11,781	2.0	35,658	327K	10.8%
VQA 2.0	64,479	33,643	1.9	98,122	658K	14.9%
OK-VQA	840	457	1.8	1,297	14K	9.23%

Table 3.5: Comparison of Gender-Based Question Distribution Across Datasets

Gender Disparity

The LLAVA dataset shows a male-dominant trend in question framing, with a men-to-women question ratio (MOW) of 1.64. According to this ratio, there are around one question about women for every 1.64 questions about men. This degree of discrepancy implies that although LLAVA exhibits some gender bias, it is not as much as in some other datasets, including Visual7W (which has a MOW of 2.0) and VQA 2.0 (which displays a ratio of 1.9). In Visual7W, for instance, the preponderance of sports-related queries, such as "What sports equipment is the man using?" illustrates the tendency to concentrate on issues pertaining to men. On the other hand, enquiries aimed at women tend to revolve around domestic or less dynamic situations, such as "What is the woman cooking?" Stereotypes and conventional gender roles are strengthened by this discrepancy. Moreover, the percentage of questions about gender in LLAVA is 12.91%, lower than the 14.9% seen in VQA 2.0 but similar to Visual Genome (11.4%) and Visual7W (10.8%). This placement implies that, in comparison to some VQA datasets, LLAVA may employ a more balanced approach to question development, potentially mitigating the impact of gender bias. Although there is still work to be done to

achieve full gender neutrality in AI training data, LLAVA represents an effort to create diversity in question framing by focussing on a wider range of topics and minimizing stereotyped meanings.

Patterns of Question Framing

Gender stereotypes are reinforced by the fact that more questions about men than women are asked in all datasets. This pattern is mirrored in the LLAVA dataset, which is consistent with the observations found in the VQA datasets. Interestingly, the more extreme ratios shown in Visual7W and other datasets suggest a greater propensity to give priority to enquiries linked to men, which may reinforce biases in society.

Racial Bias in LLAVA Dataset

Using a natural language processing (NLP) method, we examined racial-related questions in the LLaVA dataset, concentrating on four crucial terms: racial, ethnic, racism, and ethnicity. We identified possible racial bias in the dataset by classifying questions based on these particular terms using a rule-based approach akin to Hirotas study.

Our results show trends that are in line with what has been seen in other vision and language datasets. The phrases "ethnicity" and "ethnic" were frequently used in questions related to cultural identity and background. In the meantime, words like "racism" and "racial" were usually connected to societal problems, indicating biases that related to discrimination based on race and ethnicity.

These findings demonstrate that, despite the possibility of some direct racial prejudice, the dataset still includes questions that highlight racial and ethnic differences, pointing to the possible existence of underlying social stereotypes. These kinds of questions are common across many datasets, which emphasises how critical it is to address these biases because AI models trained on such data may perpetuate or reinforce existing racial assumptions and inequality.

The table shows the difference across all the dataset: The percentage of racial

Dataset Name	Racial Questions	Total Questions	Ratio (%)
LLAVA Dataset[20]	137	361,411	0.038
Visual Genome [**18**]	658	1.7M	0.038
Visual7W [**30**]	93	327k	0.03
VQA2.0 [**6**]	4,510	658k	0.08
OK-VQA [**24**]	30	14k	0.21

Table 3.6: Comparison of Racial Question Ratios Across Datasets

questions in your dataset (0.038%) is comparable to that of Visual7W (0.03%) and Visual Genome (0.038%), indicating that these datasets, which are mainly concerned with visual comprehension, have comparatively few questions about race or ethnicity.

The fact that the ratios are constant across all datasets indicates that questions referring to race or ethnicity are included in nearly all of the datasets, including those that concentrate on visual components. Even though these questions are not frequently asked, their existence implies that racial bias—however subtle—occurs in a variety of datasets.

But comparing your dataset to the greater ratio VQA2.0 (0.08%) and OK-VQA (0.21%) makes it clear that some datasets have a higher likelihood of having racial questions in them. This might be because the enquiries are more open-ended or knowledge-based, and themes connected to society, culture, or identity come up more regularly.

This comparison shows that although racial questions are present in most vision-language datasets, their level of engagement differs depending on the dataset's aim and design. Because OK-VQA and VQA2.0 feature more difficult, real-world reasoning questions, it is possible that these higher ratios reflect a greater frequency of interaction with racial and social notions.

In summary, the data shows that while racial questions are included in all datasets, the ratio is typically low unless the dataset is intended for tasks that are more socially conscious or open-ended.

4

Conclusions and Future Works

This chapter includes a reflection on the research findings, a discussion of the limits, and suggestions for future directions. This chapter offers a thorough summary of the accomplishments, the motivations behind them, and the consequences of the outcomes. Additionally, we assess our methodology's shortcomings critically and offer recommendations for enhancements that can raise the standard and equity of next Visual Question Answering (VQA) models. We also speculate about how better algorithmic results may result from improvements in dataset balance.

4.1 Summary of the Research

This thesis' major goal was to investigate racial and gender biases in Visual Question Answering (VQA) datasets, an area where fairness is vital given how much AI is being used in many different industries. In particular, we concentrated on four well-known datasets: Visual Genome, Visual7W, VQA 2.0, and OK-VQA. Our objective was to identify and measure biases associated with gender and race.

The requirement to detect and reduce bias in AI models developed using these datasets served as the driving force behind our effort. To duplicate their

methodology and identify biases in racial and gender representations inside VQA datasets, we implemented the approaches presented by [10]. As part of this methodology, questions pertaining to men and women as well as those addressing racial or ethnic groups were classified according to a set of rules. In order to evaluate the representation gaps in these datasets, we compared and analysed questions pertaining to gender and race.

We repeated findings from Hirota's work, such as Table 3.2, which displays the distribution of questions related to men and women. This made it possible for us to see that our findings agreed with those of [10]. For instance, questions about men were substantially more common than ones about women in all four datasets, supporting the finding that there is a gender representation gap. For example, Visual Genome had a male-to-female ratio of 1.9, meaning that there were nearly twice as many questions about males as there were about women. In the OK-VQA, VQA 2.0, and Visual7W datasets, a comparable pattern was seen. the table is given below.

Dataset	**Num.Men Qs**	**Num.Women Qs**	**MOW**	**Num. Gender Qs**	**Total Qs**	**Ratio (%)**
Visual Genome	109,652	55,808	1.9	165460	1.7M	11.45%
Visual7W	23,877	11,781	2.0	35,658	327K	10.87%
VQA 2.0	64,479	33,643	1.9	98,122	658K	14.91%
OK-VQA	840	457	1.8	1,297	14k	9.23%

In terms of racial bias, our findings concurred with those of Hirota. In addition to illustrating the under-representation of several demographic groups in these datasets, questions related to particular racial or ethnic groups were rare and frequently presented in a restricted number of conditions. If societal prejudices are not addressed, AI models may reinforce them, as evidenced by these biases. In Table 3.3, which displays the distribution of questions related to racial. table is given below,

Dataset Name	**Racial Questions**	**Total Questions**	**Ratio (%)**
Visual Genome[18]	658	1.7M	0.038%
Visual7W[31]	93	327k	0.03%
VQA2.0[6] 4	510	658k	0.08%
OK-VQA[24]	30	14k	0.21%

In order to guarantee that AI models are just equal and inclusive, our goal

was to further our understanding of how these biases appear in VQA datasets and to draw attention to the necessity of bias mitigation techniques.

4.2 Limitations

Although this study makes a substantial contribution to our understanding of the biases present in VQA datasets, it is important to recognise that it has several limitations.

4.2.1 Limited Dataset Scope

The limited number of datasets we used is one of our work's main weaknesses. While we have examined four primary VQA datasets (Visual Genome, Visual7W, VQA 2.0, and OK-VQA), numerous additional datasets might be incorporated for a more comprehensive examination. Larger datasets, like AI2D or GQA, for example, may offer more information about how biases increase in size. Incorporating a wider range of datasets could potentially uncover previously unseen bias patterns.

4.2.2 Bias Detection Methodology

Although it has limitations, the rule-based classification approach we utilised to detect racial and gender bias is efficient. Our analysis is limited since, for example, the binary gender classification method does not take non-binary or gender nonconforming people into consideration. Furthermore, while subtler types of stereotyping and implicit biases may still be present in VQA datasets, these are not picked up by this approach. More sophisticated natural language processing (NLP) methods, including neural networks, may be used in future research to more accurately identify implicit bias.

4.2.3 Hypothetical Imbalance in Data

The imbalance of the datasets itself is a major drawback. The gender distribution was skewed, as our results demonstrate, with questions linked with men being more common in all datasets. This raises the theory that algorithms trained on an unbalanced dataset are unable to produce findings that are impartial or entirely correct. An AI system trained mostly on male-centric data, for instance, may overfit to questions related to men and underperform when exposed to data centred on women, producing biassed results. According to this theory, the quality and harmony of the data used to train algorithms may have an impact on their performance, even for the best-performing ones.

4.2.4 Generalization Beyond VQA

The exclusive emphasis on VQA datasets is another drawback. Although our results apply to VQA systems, bias may manifest differently in other domains of artificial intelligence, such as speech recognition or sentiment analysis. A more comprehensive understanding of the universal problem of bias across AI systems would result from broadening the scope of this investigation to cover different AI jobs.

4.3 Future Work

Several areas for further research are presented, building on the foundation established by this thesis.

4.3.1 Larger and More Diverse Datasets

An important goal for future work is to apply the analysis to more extensive and varied datasets. More thorough findings would be obtained by include other datasets such as GQA, AI2D, or custom-curated datasets that reflect a wider range of demographic groupings. Furthermore, research in the future might assess datasets that go beyond those that are restricted to visual question

responding, including text-based datasets or multimodal datasets that integrate language and vision.

4.3.2 Advanced Bias Detection Techniques

More advanced techniques can improve our rule-based system for identifying racial and gender bias. For instance, using deep learning-based natural language processing (NLP) models, like transformers, may make it possible to identify more subtle instances of bias, like variations in language that are specific to gender or implicit links between particular demographic groups and particular types of behaviour. These models could potentially identify intersectional biases, which are discrimination that combine more than one kind, like gender and race.

4.3.3 Real-Time Bias Monitoring

The creation of real-time bias monitoring systems is among the most exciting fields for future study. As AI algorithms process new data, they might be designed to continuously detect and correct for bias. This will minimise the possibility of damaging or out-of-date preconceptions being reinforced by allowing AI systems to learn from a variety of data streams. In high-stakes fields like criminal justice or healthcare, where biassed decision-making might have serious consequences, real-time bias mitigation may be essential.

4.3.4 Hypothesis Testing on Algorithm Performance

As we previously suggested, an unbalanced dataset causes AI systems to perform slightly less than ideal. This theory could be thoroughly tested in the future by comparing the results of algorithms trained on balanced and unbalanced datasets. Such trials could statistically evaluate the impact of dataset balance on model outcomes using a variety of fairness indicators, such as statistical parity, disproportional impact, and equal opportunity.

4.3.5 Interdisciplinary Collaboration for Ethical AI

One more critical component of future effort will be encouraging interdisciplinary cooperation. Collaboration between AI researchers, ethicists, and social scientists is necessary to eliminate bias at all stages of AI development, from designing algorithms and creating datasets to deploying models. This kind of cooperation could result in the development of more ethically and inclusively managed datasets as well as more openly and transparently designed algorithms.

4.4 Final Remarks

To sum up, this thesis has clarified the racial and gender biases seen in VQA datasets and how they affect AI models. We presented a thorough examination of how these biases arise and can be identified by expanding Hirota's methodology to numerous datasets. Our results support the presence of significant biases and highlight the need for more investigation into methods for bias prevention and detection.

However, this is only the first phase of the research. AI bias must be addressed with persistent work, creative solutions, and a dedication to justice. The AI community may get closer to developing models that accurately capture the diversity and complexity of the actual world by embracing bigger datasets, more sophisticated detection techniques, and real-time monitoring systems. By doing this, we can guarantee that AI systems treat every person fairly, irrespective of their gender, colour, or origin.

References

[1] Joy Buolamwini and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Ed. by Sorelle A. Friedler and Christo Wilson. Vol. 81. Proceedings of Machine Learning Research. PMLR, 2018, pp. 77–91. URL: https://proceedings.mlr.press/v81/buolamwini18a.html.

[2] Joy Buolamwini and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Ed. by Sorelle A. Friedler and Christo Wilson. Vol. 81. Proceedings of Machine Learning Research. PMLR, 23–24 Feb 2018, pp. 77–91. URL: https://proceedings.mlr.press/v81/buolamwini18a.html.

[3] Kaylee Burns et al. "Women also Snowboard: Overcoming Bias in Captioning Models". In: *CoRR* abs/1803.09797 (2018). arXiv: 1803.09797. URL: http://arxiv.org/abs/1803.09797.

[4] Robert Geirhos et al. "Shortcut learning in deep neural networks". In: *Nature Machine Intelligence* 2.11 (Nov. 2020), pp. 665–673. ISSN: 2522-5839. DOI: 10.1038/s42256-020-00257-z. URL: http://dx.doi.org/10.1038/s42256-020-00257-z.

[5] Yash Goyal et al. "Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering". In: *CoRR* abs/1612.00837 (2016). arXiv: 1612.00837. URL: http://arxiv.org/abs/1612.00837.

[6] Yash Goyal et al. "Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering". In: *CoRR* abs/1612.00837 (2016). arXiv: 1612.00837. URL: http://arxiv.org/abs/1612.00837.

[7] Yash Goyal et al. *Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering*. 2017. arXiv: 1612.00837 [cs.CV]. URL: https://arxiv.org/abs/1612.00837.

[8] Danna Gurari et al. *VizWiz Grand Challenge: Answering Visual Questions from Blind People*. 2018. arXiv: 1802.08218 [cs.CV]. URL: https://arxiv.org/abs/1802.08218.

[9] Alex Hanna et al. "Towards a critical race methodology in algorithmic fairness". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. ACM, Jan. 2020. DOI: 10.1145/3351095.3372826. URL: http://dx.doi.org/10.1145/3351095.3372826.

[10] Yusuke Hirota, Yuta Nakashima, and Noa Garcia. "Gender and Racial Bias in Visual Question Answering Datasets". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. ACM, June 2022. DOI: 10.1145/3531146.3533184. URL: http://dx.doi.org/10.1145/3531146.3533184.

[11] Yusuke Hirota, Yuta Nakashima, and Noa Garcia. *Quantifying Societal Bias Amplification in Image Captioning*. 2022. arXiv: 2203.15395 [cs.CV]. URL: https://arxiv.org/abs/2203.15395.

[12] Yusuke Hirota et al. "Visual Question Answering With Textual Representations for Images". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*. Oct. 2021, pp. 3154–3157.

[13] Zhicheng Huang et al. *Seeing Out of tHe bOx: End-to-End Pre-training for Vision-Language Representation Learning*. 2021. arXiv: 2104.03135 [cs.CV]. URL: https://arxiv.org/abs/2104.03135.

[14] Drew A. Hudson and Christopher D. Manning. *GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering*. 2019. arXiv: 1902.09506 [cs.CL]. URL: https://arxiv.org/abs/1902.09506.

[15] Huaizu Jiang et al. *In Defense of Grid Features for Visual Question Answering*. 2020. arXiv: 2001.03615 [cs.CV]. URL: https://arxiv.org/abs/2001.03615.

[16] Jungseock Joo and Kimmo Kärkkäinen. "Gender slopes: Counterfactual fairness for computer vision models by attribute manipulation". In: *Proceedings of the 2nd international workshop on fairness, accountability, transparency and ethics in multimedia*. 2020, pp. 1–5.

[17] Corentin Kervadec et al. *Roses Are Red, Violets Are Blue... but Should Vqa Expect Them To?* 2021. arXiv: 2006.05121 [cs.CV]. URL: https://arxiv.org/abs/2006.05121.

[18] Ranjay Krishna et al. "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations". In: *CoRR* abs/1602.07332 (2016). arXiv: 1602.07332. URL: http://arxiv.org/abs/1602.07332.

[19] Tsung-Yi Lin et al. *Microsoft COCO: Common Objects in Context*. 2015. arXiv: 1405.0312 [cs.CV].

[20] Haotian Liu et al. "Visual Instruction Tuning". In: *NeurIPS*. 2023.

[21] Varun Manjunatha, Nirat Saini, and Larry S. Davis. "Explicit Bias Discovery in Visual Question Answering Models". In: *CoRR* abs/1811.07789 (2018). arXiv: 1811.07789. URL: http://arxiv.org/abs/1811.07789.

[22] Kenneth Marino et al. *OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge*. 2019. arXiv: 1906.00067 [cs.CV]. URL: https://arxiv.org/abs/1906.00067.

[23] Kenneth Marino et al. "OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[24] Kenneth Marino et al. "OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge". In: *CoRR* abs/1906.00067 (2019). arXiv: 1906.00067. URL: http://arxiv.org/abs/1906.00067.

[25] Marie-Francine Moens et al., eds. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021. URL: https://aclanthology.org/2021.emnlp-main.0.

[26] Hao Tan and Mohit Bansal. *LXMERT: Learning Cross-Modality Encoder Representations from Transformers*. 2019. arXiv: 1908.07490 [cs.CL]. URL: https://arxiv.org/abs/1908.07490.

[27] O. Vinyals et al. "Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge". In: *IEEE Transactions on Pattern Analysis amp; Machine Intelligence* 39.04 (Apr. 2017), pp. 652–663. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2016.2587640.

[28] Zeyu Wang et al. "Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation". In: *CoRR* abs/1911.11834 (2019). arXiv: 1911.11834. URL: http://arxiv.org/abs/1911.11834.

[29] Peng Zhang et al. "Yin and Yang: Balancing and Answering Binary Visual Questions". In: *CoRR* abs/1511.05099 (2015). arXiv: 1511.05099. URL: http://arxiv.org/abs/1511.05099.

[30] Dora Zhao, Angelina Wang, and Olga Russakovsky. "Understanding and Evaluating Racial Biases in Image Captioning". In: *CoRR* abs/2106.08503 (2021). arXiv: 2106.08503. URL: https://arxiv.org/abs/2106.08503.

[31] Yuke Zhu et al. "Visual7W: Grounded Question Answering in Images". In: *CoRR* abs/1511.03416 (2015). arXiv: 1511.03416. URL: http://arxiv.org/abs/1511.03416.

Acknowledgments

My sincere appreciation goes out to everyone who helped me along the way while I worked on this thesis. First and foremost, I want to express my sincere gratitude to Prof. Antonio Roda, my supervisor, for all of his help, support, and patience. His insightful criticism and recommendations really influenced this work and motivated me to keep getting better. His advice and assistance helped me in improving the research. A particular thank you to the University of Padova for providing the sources and setting necessary for this research. The community of researchers, the facilities, and the sources have all been crucial to finishing this job. I will always be grateful to my family, especially my husband and parents, for their unfailing faith in me. Their unwavering love, support, and comprehension got me through all of the difficulties I was going through at the time—both personal and intellectual.

Lastly, I would want to express my gratitude to all the writers and scholars whose writings influenced mine. One persistent source of motivation has been the academic community's commitment to expanding knowledge in the field of artificial intelligence.

I'm grateful to everyone.

Asifa AKter