University of Sheffield

Beyond Sight: Generating Clothing Graphic Captions for Visually Impaired Users



Amnah Alluqmani

Supervisors: Dr Morgan Harvey and Dr Monica Paramita

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

in the

Information School

October 17, 2024

Declaration

All sentences or passages quoted in this document from other people's work have been specifically acknowledged by clear cross-referencing to author, work and page(s). Any illustrations that are not the work of the author of this report have been used with the explicit permission of the originator and are specifically acknowledged. I understand that failure to do this amounts to plagiarism and will be considered grounds for failure.

Name: Amnah Alluqmani

Date: October 17, 2024

Acknowledgements

This thesis would not be possible without the ongoing support and encouragement of my first supervisor, Dr. Morgan A Harvey. I not only thank you for your insightful guidance and support, but also for being more than a supervisor. A great thank you goes to my second supervisor, Dr. Monica L Paramita, for her insightful recommendations and advice.

I would also like to thank Dr. Alessandro Checco and Dr. Ziqi Zhang for their helpful discussions and insights in the first stage of this thesis during their time at the University of Sheffield. In addition, I want to thank Prof. Mark James Carman for his insightful discussions in the second stage of this thesis. I really appreciate your time and consideration in joining our meetings.

Warm thanks goes to my dear parents, and special thanks to my great husband, Abdulrahman. I will not forget your endless support and encouragement. Thank you for being there in my tough days, and thank you for your care and love.

To my little sons Ruwaid and Yazan, thank you for accompanying me on my journey overseas. Your warm hugs and smile were a huge stress reliever every day. I dedicate this work to you and my little baby.

To my beloved sisters, Dr. Hamida and Dr. Eman thank you for being my pillars of strength and for celebrating every milestone with me.

Finally, I want to extend my gratitude to all those who helped me to complete this thesis.

Abstract

People with visual impairment (VI) often face challenges when performing everyday tasks, of which clothes shopping is the most challenging. Many people with VI engage in online shopping to eliminate some of the challenges associated with physical shopping. However, buying clothes online presents many other limitations and barriers. More research is needed to address these challenges and propose solutions to enhance the shopping experiences of people with VI. Most research studies have relied on interviews alone, which provide only subjective, recall-biased information, and only few studies have offered a solution. Thus, we conducted two complementary studies, using both observational and interview approaches to fill the gap in understanding the behaviours of people with VI when selecting and purchasing clothes online. In addition, we built a VI-friendly clothing graphic captioning model, adopting the Large Language Model (LLM) approach to simulate the needs of people with VI and the Region of Interest (ROI) approach to generate focused, informative captions. We conducted multiple assessments, including a manual analysis of Ground Truth (GT) clothing image-text datasets and automatic and human evaluations of our proposed model. Our findings show that shopping websites have inaccurate, misleading and contradictory clothing descriptions and that people with VI rely mainly on (unreliable) search tools and check product descriptions by reviewing customer comments. Our findings also indicate that people with VI are hesitant to accept assistance from automated systems. Trust in such systems could be improved if researchers develop systems that better accommodate users' needs and preferences. The results of the evaluation of the GT captions showed that they had several limitations that disqualify them from consideration as gold standard captions. The evaluation of our model revealed that adopting LLMs and ROI approach generates captions that are the most informative, considering the needs of people with VI, compared with BLIP-2 (i.e., adopting a grid-based approach) and GT captions. Our model achieved average scores of 1.73 and 1.774 from evaluators with VI and sighted evaluators, respectively. In comparison, BLIP-2 and GT data obtained average scores of 1.460 and 1.440 from evaluators with VI and 1.565 and 1.671 from sighted evaluators, respectively. Furthermore, although our model was the most informative among the existing baseline models, it is difficult to maintain its accuracy, which is lower than those of the GT and BLIP-2 models. This thesis investigates the factors that affect the accuracy of captions and will be further addressed in future work.

Contents

1 Introduction **1**
- 1.1 Research Motivation . . . 1
- 1.2 Research Challenge . . . 5
- 1.3 Research Questions . . . 7
- 1.4 Aim and Objectives . . . 7
- 1.5 Project Procedures . . . 8
- 1.6 Contributions and Publications . . . 9
  - 1.6.1 Contributions . . . 9
  - 1.6.2 Publications . . . 9
- 1.7 Thesis Outline . . . 9

2 Literature Review **11**
- 2.1 Introduction . . . 11
- 2.2 Everyday Life Challenges Experienced by Visually Impaired People . . . 11
- 2.3 Shopping Challenges Experienced by Visually Impaired People . . . 13
  - 2.3.1 Offline Shopping Challenges . . . 13
  - 2.3.2 Online Shopping Challenges . . . 13
  - 2.3.3 Online Clothes Shopping Challenges . . . 14
- 2.4 VI-friendly Assistance Systems for Online Clothes Shopping . . . 15
- 2.5 Image Captioning . . . 17
  - 2.5.1 General Image Captioning Methods . . . 17
  - 2.5.2 Clothing Image Captioning Methods . . . 22
  - 2.5.3 Clothing Image–Text Datasets . . . 24
- 2.6 Clothing Image Representation For VI Users . . . 24
- 2.7 LLMs in Image Representation . . . 25
- 2.8 Image Captioning Evaluation . . . 27
- 2.9 Gap in the Literature . . . 28
- 2.10 Conclusion . . . 29

3 The Barriers to Online Clothing Websites **30**
- 3.1 Introduction . . . 30
- 3.2 Method . . . 31
  - 3.2.1 Participant Recruitment . . . 32
  - 3.2.2 Research Procedures . . . 33
  - 3.2.3 Study Involving Participants with Visual Impairments . . . 35

3.3 Data Analysis . . . 36
3.4 Findings . . . 37
3.4.1 Limitations on Clothing Websites . . . 38
3.4.2 Fashion Influences . . . 41
3.4.3 Shopping Malls . . . 42
3.4.4 Assistance Technology . . . 43
3.5 Discussion . . . 44
3.5.1 Design Implications . . . 46
3.6 Conclusions . . . 47

4 Building a VI-friendly Captioning Model 49
4.1 Introduction . . . 49
4.2 Model Architecture . . . 51
4.2.1 Graphic Recognition . . . 51
4.2.2 Graphic Description Generation . . . 54
4.3 Benchmark Datasets . . . 58
4.3.1 Clothing Datasets . . . 58
4.3.2 A Novel Clothing Graphics Dataset . . . 63
4.4 Research Procedures . . . 65
4.5 Experiments and Evaluation Methods . . . 66
4.5.1 Experimental setup . . . 66
4.5.2 Preliminary Analysis . . . 66
4.5.3 Automatic Evaluation . . . 66
4.5.4 Human Evaluation . . . 68
4.6 Findings . . . 75
4.6.1 Preliminary Analysis . . . 75
4.6.2 Automatic Evaluation . . . 82
4.6.3 Human Evaluation . . . 90
4.7 Discussion . . . 115
4.7.1 RQ 4.1: What are the limitations of ground truth captions of the image–text dataset considering VI people's needs? . . . 115
4.7.2 RQ 4.2: How informative are the captions generated by our model compared to those generated by existing models and ground truth data for VI users? . . . 116
4.7.3 RQ 4.3: How accurate are the captions extracted by our model compared to those extracted by existing models and ground truth data? . . . 118
4.8 Conclusion . . . 118
4.8.1 Limitations on the Human Evaluation Approach . . . 119

5 Conclusions and Future Work 121
5.1 Conclusions . . . 121
5.1.1 Thesis Summary . . . 121
5.1.2 Summary of Findings . . . 122
5.1.3 Summary of Contributions . . . 125
5.1.4 Significance of This Thesis . . . 125
5.2 Future Work . . . 126

5.2.1 Further Investigation 126
5.2.2 Possible Design Implications 126
5.2.3 Enhancements of the Clothing Graphic Dataset 127
5.2.4 Improvements of the Performance of the Proposed Model 129
5.2.5 Flexible VI-friendly Captioning Models 130

Appendices **146**

A Empirical Study Documents **147**

B Empirical Study Ethics **159**

C Interview Scripts **167**

D Human Evaluation Documents **171**

E Human Evaluation Ethics **177**

F More Results of The Graphic Description Generation **186**

List of Figures

1.1 Example of missing descriptive information, highlighting the absence of crucial details essential for VI people to visualise the item. 4
1.2 Example of unreliable product text descriptions. It shows an ambiguous or unclear colour name, leading to potential confusion for VI shoppers. 4
1.3 Amazon's search engine returns items that do not match the query. 5
1.4 Project stages overview. 8

2.1 Template-based framework. It is based on a predefined template. Object elements are extracted from the image and placed in the blank slots in the template to form a caption. The figure was adopted from Ming et al. (2022). 18
2.2 Retrieval-based framework. Captions are derived on the basis of the similarity of the prestored image caption pair. The similarity matching process is based on the overall scene and object similarity. This figure was adopted from Ordonez et al. (2011). 19
2.3 Encoder-decoder framework introduced by Vinyals et al. (2015). The CNN model encodes visual features and is fed to the RNN model to generate captions. 20
2.4 Visualisation of two visual encoding frameworks: grid and ROI based. Grid based (left): Visual features are extracted from the overall image. ROI-based (Right): Visual features are extracted from certain regions in the image. This figure was adopted from Stefanini et al. (2021). 21

3.1 Overview flow diagram of the research procedures: co-design with the Sheffield Royal Society for the Blind (SRSB), validity test, pilot studies, main studies, data analysis, and reporting the results. The research steps were implemented independently, but in the last step, the findings of both studies were merged as they were complementary. 35
3.2 Diagram of themes identified from Studies 1 and 2. 37
3.3 An example of using contradictory terms (i.e., casual and office wear) when describing clothing items online. 40

4.1 The clothing graphic description model contains two phases: graphic recognition and graphic description generation. 52
4.2 Examples of the results generated by FashionCLIP (Chia et al., 2022). 53
4.3 Examples of results generated by CLIPSeg (Lüddecke & Ecker, 2022). 54
4.4 Example of prompts and results for each component of the graphic description generation. 57

4.5 Clothing images from the FashionGen dataset showing two different views of the same item. . . . 59
4.6 Two images from the FashionGen dataset with different appearances sharing the same description. . . . 59
4.7 Examples of clothing images from the DeepFashion multimodal dataset showing three different positions and the corresponding captions. . . . 60
4.8 Examples of description from the DeepFashion-MultiModal dataset with language misrepresentation problems. . . . 61
4.9 Examples of different terms used to describe the same clothing item in DeepFashion multimodal text descriptions. . . . 61
4.10 Examples of descriptions from the FACAD; using marketing-friendly terms to entice customers. . . . 62
4.11 Images and their ground-truth captions from FashionGen (left), DeepFashion-MultiModal(middle) and FACAD (right). . . . 63
4.12 An excerpt from the CGD. It shows an item from the FashionGen (top), DeepFashion-MultiModal (middle) and FACAD (bottom). CGD: Clothing Graphic Description dataset, FACAD: Fashion Captioning Dataset. . . . 64
4.13 Evaluation Methods. . . . 65
4.14 Human Evaluation Interface of an example of Question 1. . . . 73
4.15 Human Evaluation Interface of an example of Question 2. . . . 73
4.16 Human Evaluation Interface of an example of Question 3 and 4. . . . 74
4.17 Example of Attention Check Question 1. . . . 74
4.18 Example of Attention Check Question 2. . . . 74
4.19 Examples of results obtained from the graphic description generation component of the proposed model (including BLIP-2 captions, Automatic Questioning, Automatic Answering and Automatic Summarising) using images from the FashionGen (left) and FACAD (right) datasets. . . . 77
4.20 Examples of results obtained from the graphic description generation component of the proposed model (including BLIP-2 captions, Automatic Questioning, Automatic Answering ,and Automatic Summarising) using an image from the DeepFashion-MultiModal dataset. . . . 78
4.21 Example of results generated by the Automatic Questioning task fails to generate relevant questions about the inquired item due to the hallucination problem (highlighted in red boxes). . . . 79
4.22 Example of inaccurate descriptions generated from BLIP-2 affecting the quality of the Automatic Questioning results. . . . 80
4.23 Examples of the lack of graphic descriptions in the ground-truth captions of two items each from three fashion captioning datasets: FashionGen (top), FACAD (middle), and DeepFashion-MultiModal datasets (bottom). The ground-truth captions are in the boxes with dotted borders, and the descriptions generated from the proposed model are in the blue boxes. . . . 81
4.24 Classification results of the FashionCLIP model using the FashionGen dataset. 83
4.25 Examples of items incorrectly classified as non-graphic in the GT data (FashionGen). . . . 83

4.26 Example of item incorrectly classified as graphic clothing in the GT data (FashionGen). 84
4.27 Classification results of the FashionCLIP model using our classification dataset. 85
4.28 Example of items incorrectly classified as non-graphics by the FashionCLIP model. 86
4.29 Examples of two clothing items from the FashionGen dataset, comparing the rich, informative captions obtained from our model (middle) with captions generated from BLIP-2 (top) and GT captions (bottom). 88
4.30 Examples of two clothing items from the DeepFashion-MultiModal dataset, comparing the rich, informative captions obtained from our model (middle) with captions generated from BLIP-2 (top) and GT captions (bottom). . . . 89
4.31 Examples of two clothing items from the FACAD dataset, comparing the rich, informative captions obtained from our model (middle) with captions generated from BLIP-2 (top) and GT captions (bottom). 90
4.32 The results of the VI evaluators' assessment of informativeness, comparing GT, BLIP-2 and our model. Our model achieved the highest score in the 'Very Informative' category and is among the top-rated in the 'Informative' category. 92
4.33 Boxplot illustrating the average score given by VI evaluators' assessment of the informativeness of captions produced by GT, BLIP-2 and our model. . . . 92
4.34 Comparison of the results of our sighted evaluators' assessments of the informativeness of captions generated by GT, BLIP-2 and our model. Our model achieved the highest score in the 'Very Informative' category and was comparable to the other methods in the 'Informative' category. 94
4.35 Boxplot illustrating the average score given by sighted evaluators' assessing the informativeness of captions produced by GT, BLIP-2 and our model. . . . 95
4.36 An example of a caption for which VI and sighted evaluators disagreed. The former found it to be highly informative, while sighted evaluators did not. . . 97
4.37 Comparison of the results of our sighted evaluators' assessments of the accuracy of captions generated by GT, BLIP-2 and our model. 98
4.38 Boxplot illustrating the average score given by sighted evaluators' assessment of the accuracy of captions produced by GT, BLIP-2 and our model. 99
4.39 Example of a complex graphic that our model did a good job of describing. . 103
4.40 Example of a caption generated by our model, showing the failure to identify the type of the printed object (whether a printed text or picture). 104
4.41 Example of a caption generated from our model which shows lack of identification of object printed (text or picture printed). 105
4.42 Example of a caption generated from our model shows incorrectly describing colours of clothing graphic. 107
4.43 Example of a caption generated by our model includes unnecessary details. . 107
4.44 Example of a clothing caption generated by our model shows a hallucination problem. 109
4.45 Example of a clothing caption generated by our model shows inaccurately describing text printed. 110

4.46 Example of a clothing item and results obtained from the proposed model's components. The BLIP-2 caption model misrepresented objects in the graphic element without specifying if they are presented as text or a picture, as indicated in the red box. 111
4.47 Example of a clothing item and results obtained from the proposed model's components. The BLIP-2 caption model generated inaccurate information caused by the hallucination problem, as indicated in the red box. 112
4.48 Example of a clothing item and results obtained from the proposed model's components. It shows the limitations to providing sufficient descriptions of additional elements, as indicated in the red boxes. 113
4.49 Example of a clothing item and results obtained from the proposed model's components. It shows the limitations of the Automatic Answering task, which sometimes provided no answer or missed essential information, leading to inaccurate descriptions, as indicated in the red boxes. 114
4.50 Example of a clothing item and results obtained from the proposed model's components. It shows the case of the Automatic Answering task failing to generate accurate information, as indicated in the red box. 115

5.1 A low-resolution image from the CGD. 128
5.2 A clothing image from the CGD, showing text printed on the left sleeve. Although the item is shown in various angles, none of the images present the text clearly. 129

List of Tables

3.1 Participant demographic information. 33

4.1 Comparison of the three commonly used clothing image-text datasets, showing details of the year of publication, the number of items, image size and number of captions (Cap) included, the number of categories (Cat) such as dress and T-shirt and attributes (Att) such as colour and sleeves and the source. Details are obtained from the original sources. "/" denotes unavailable information. . 58
4.2 Classification results of the FashionCLIP model using FashionGen and our classification datasets. 85
4.3 Comparison of the results obtained by our model and BLIP-2 in a set of automatic image captioning metrics using three fashion captioning datasets. An asterisk (*) indicates that the P-value is less than 0.05, denoting statistical significance. 86
4.4 A summary of the statistical results indicating the number of evaluators assessing the informativeness and accuracy of each image captions. 91
4.5 Means and standard deviation (SD) results of the VI informativeness assessment for the three methods (i.e., GT, BLIP-2 and our model). 93
4.6 Krippendorff's Alpha Coefficient for VI Evaluators' Agreement When Judging the Informativeness of Captions Extracted by Each Method (GT, BLIP-2 and our model). 93
4.7 Mean and standard deviation (SD) results of the sighted informativeness assessment for the three methods (i.e., GT, BLIP-2 and our model). 95
4.8 Krippendorff's Alpha Coefficient for sighted Evaluators' Agreement When Judging the Informativeness of Captions Extracted by Each Method (GT, BLIP-2 and our model). 96
4.9 Mean and standard deviation (SD) results of the sighted accuracy assessment for the three methods (i.e., GT, BLIP-2 and our model). 99

C.1 Observation and Follow-up Interview Protocol for Study 1. "IQ" denotes interview questions, "RQ" denotes research questions. 168
C.2 Standalone Interview protocol matrix for study 2. "I" denotes interview questions, "RQ" denotes research questions. 169

Glossary

ANOVA Analysis of variance.

BLEU Bilingual Evaluation Understudy.

CGD Clothing Graphic Dataset.

CIDEr Consensus-Based Image Description Evaluation.

CNN Convolutional Neural Network.

CV Computer Vision.

FACAD Fashion Captioning Dataset.

GT Ground Truth.

HCI Human Computer Interaction.

LLM Large Language Model.

LM Language Model.

METEOR Metric for Evaluation of Translation with Explicit Ordering.

NIC Neural Image Caption.

NLG Natural Language Generation.

NLP Natural Language Processing.

RNN Recurrent Neural Network.

ROI Region of Interest.

ROUGE Recall-Oriented Understudy for Gisting Evaluation.

SPICE Semantic Propositional Image Caption Evaluation.

VLM Visual Language Model.

Chapter 1

Introduction

1.1 Research Motivation

According to the UK National Health Service (NHS) [1], the number of individuals with sight loss exceeds two million in the UK alone. Because blindness is often caused by lifestyle-related diseases (e.g., diabetes) (Zhang et al., 2008)—which are on the rise, and which become more prevalent as a population ages—it is expected that the relative incidence of visual impairment (VI) will continue to increase in the future [2].

Vision difficulties of varying severity affect millions of people worldwide. An inability to see faraway or extremely close objects can be corrected with eyeglasses or contact lenses. However, a number of vision problems (such as those that occur when the parts of the eye responsible for image processing are diseased or damaged) can result in partial or total vision loss that cannot be restored with medical treatment or corrective lenses (e.g., glasses or contact lenses) [3]. The Accessibility and Disability Resources Centre (ADRC) at the University of Cambridge divides visual impairment into two categories: partially sighted and blind. The term *partially sighted* is used to describe those with moderate vision loss, while *blind* is used to describe those who have experienced extreme vision loss. These conditions can be either congenital or acquired after birth [4].

Many people with VI face challenges that affect their quality of life and their ability to independently perform everyday tasks (Borges et al., 2020; Farzana et al., 2014; Karim, 2020; Roh et al., 2018; Szpiro et al., 2016). For instance, a cooking task that be completed by sighted individuals in a matter of minutes may take VI individuals much longer. This is due to the need for caution when carrying out essential tasks, such as identifying expiration dates, opening cans, and cutting fruits or vegetables. These tasks require visual ability or a high level of attention to prevent injuries or errors.

Among other activities, shopping is one of the most challenging. Research exploring the

[1] NHS. (n.d.). *Blindness and vision loss.* Retrieved from https://www.nhs.uk/conditions/vision-loss/. Accessed February 2024.

[2] SEE International. (n.d.). *Global Blindness Projected to Triple by 2050.* Retrieved from https://www.seeintl.org/blog/global-blindness-2050/. Accessed February 2024.

[3] KidsHealth. (n.d.). *Visual Impairment.* Retrieved from https://kidshealth.org/en/teens/visual-impairment.html. Accessed February 2024.

[4] patient. (n.d.). *Eye Problems in Babies.* Retrieved from https://patient.info/doctor/eye-problems-in-babies. Accessed February 2024.

impact of visual impairment on undertaking everyday life tasks revealed that 90% of participants found shopping to be one of the most difficult activities to perform independently (Jones et al., 2019). Study participants reported that they lacked a sense of reliability, control, and independence when shopping alone. Other related research (Jones et al., 2019) has revealed that, as a result, more than 25% of VI people avoid shopping altogether, while 40% shop with assistance. However, this dependence on others is an unsatisfying solution. Relying on someone to perform a task means waiting for the assistance to be available, and this sometimes means that VI people must schedule their day based on the availability of sighted assistance. Additionally, such a practice clearly places a burden on the sighted person. Even if a VI person overcomes the challenges involved in shopping at a store, they often find it difficult to identify product information. They commonly avoid asking questions of customer service representatives because they receive mostly unhelpful replies (Liu, Ding, et al., 2019). Although disability support policies provide VI people with personal assistance for activities such as shopping, these services have limitations, such as time restrictions on assistance and the possibility that one's personal assistant is unsupportive (Lee et al., 2020).

Online shopping has eliminated some of the challenges inherent in independent offline shopping, such as dressing oneself, travelling to a store, and identifying and reaching for products on shelves. As such, it can offer a greater sense of reliability and control, thereby improving VI people's quality of life (Williams, 2021). Accordingly, most VI individuals engage in online shopping (Stangl et al., 2020a). However, while this method is significantly preferred, VI people still find making online purchases to be among their most challenging tasks (Stangl et al., 2018).

Multiple challenges, as yet unresolved, are encountered when shopping for products in various categories online (Stangl et al., 2020a; Wang et al., 2021), including effectively navigating online stores, recognising images, and obtaining useful product information.

When shopping for clothing, the lack of product textual information in particular affects VI people's capacity to rely on such data to identify desired items. This is because clothing, unlike other shopping categories such as groceries or medicines, involves the identification of several visual features (such as colour, style, material, and size) that factor into making a selection. Clothing elements such as graphics (i.e., text or pictures), pockets, zips, or buttons, and the descriptions of each element, are also essential factors in making a selection.

Unfortunately, most online retailers rely predominantly on product images to provide detailed information, negatively impacting the reliability of textual data. This was highlighted by Wang et al. (2021), who interviewed 20 VI participants in order to identify the challenges that they experienced when seeking information online in general shopping categories such as home, clothing, and electronics. The study revealed that the most challenging aspect of shopping was the limited textual information to describe images, with the participants finding that product information was communicated primarily via images. While sighted users can easily identify product features and functionalities by looking at product images, VI users rely on alternative text (alt-text) to understand image content. However, in most cases the alt-text for product images is absent, lacks detail, or simply directs to image paths (Wang et al., 2021). This is typically sufficient for sighted people, but VI people require much more written information to make an informed selection (Liu, Ding, et al., 2019; Stangl et al., 2018; Wang et al., 2021), including objective information (e.g., colour, pattern, and graphics) and subjective information (e.g., clothing coordination and age appropriateness). While some

of this information can potentially be obtained from product reviews, their sheer volume, combined with VI people's inability to quickly scan and read content, makes this impractical.

Figures 1.1 and 1.2 show examples of the poor descriptions for online clothing products. The clothing item in Figure 1.1 is described by the text printed on it (i.e., 'I'm The Gamer Elf'). While sighted users can easily glean other missing information (such as the graphic printed on the t-shirt and the fact that red and green are also present) by looking at the product image, VI users struggle to identify this information independently. Inspecting the text accompanying the image in Figure 1.2 reveals an unclear colour name, where 'wine red' describes the colour of the gloves. This is difficult to determine without looking at the image. VI people have highlighted other such examples of mislabelled colour elements, such as blue being labelled 'surf the web' or 'Arctic blue' (Liu, Ding, et al., 2019; Wang et al., 2021; Williams, 2021).

Another limitation affecting the reliability of textual information is that, currently, textual descriptions are often written using marketing terms that are intended to attract and hold sighted people's attention, rather than to accurately describe the product. Another example of the limited textual information provided by online stores is a lack of descriptions of graphics on clothing items (Williams, 2021). Product descriptions sometimes do not identify the characters printed on clothing products (such as a woman, a man, an animal, or a specific character such as Mickey Mouse) assuming that these will be obvious to those looking at the image. Alternatively, while product descriptions might identify the characters in the images, they provide no description of what the characters are doing. This makes the shopping experience frustrating, as described in numerous studies (Stangl et al., 2018; Williams, 2021; Yang et al., 2015).

Amazon's search engine reveals a further limitation that VI people face when shopping online, and one which also impacts the reliability of textual data. Figure 1.3 shows example search results for the query 'red wool dress'. The resulting items are a poor match for the query, as the items in the first row of returned products are not dresses, and the first item returned is also not made of wool. While sighted users would notice these errors, VI users find it frustrating and time consuming to check every item to determine its relevance to their search.

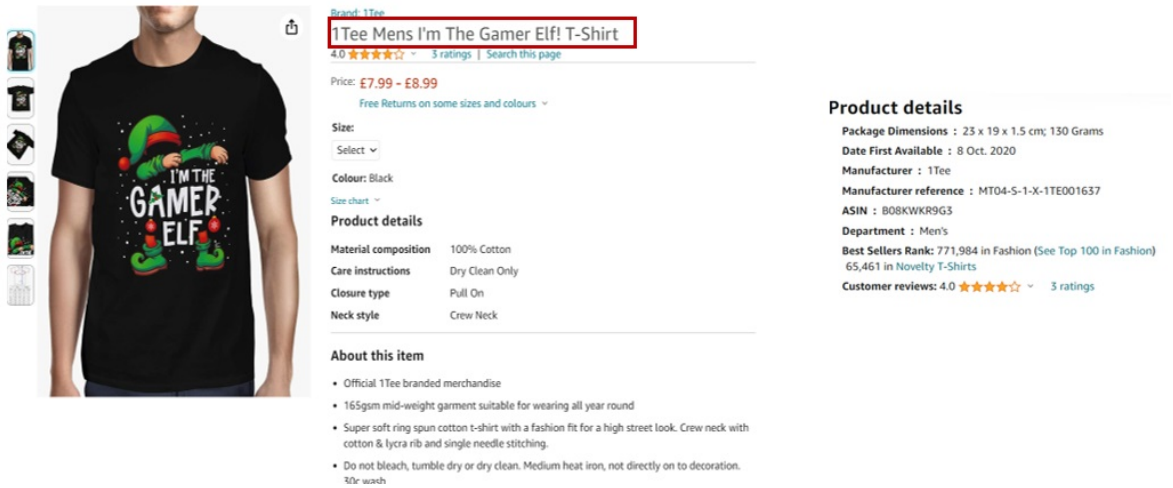


Figure 1.1: *Example of missing descriptive information, highlighting the absence of crucial details essential for VI people to visualise the item.*

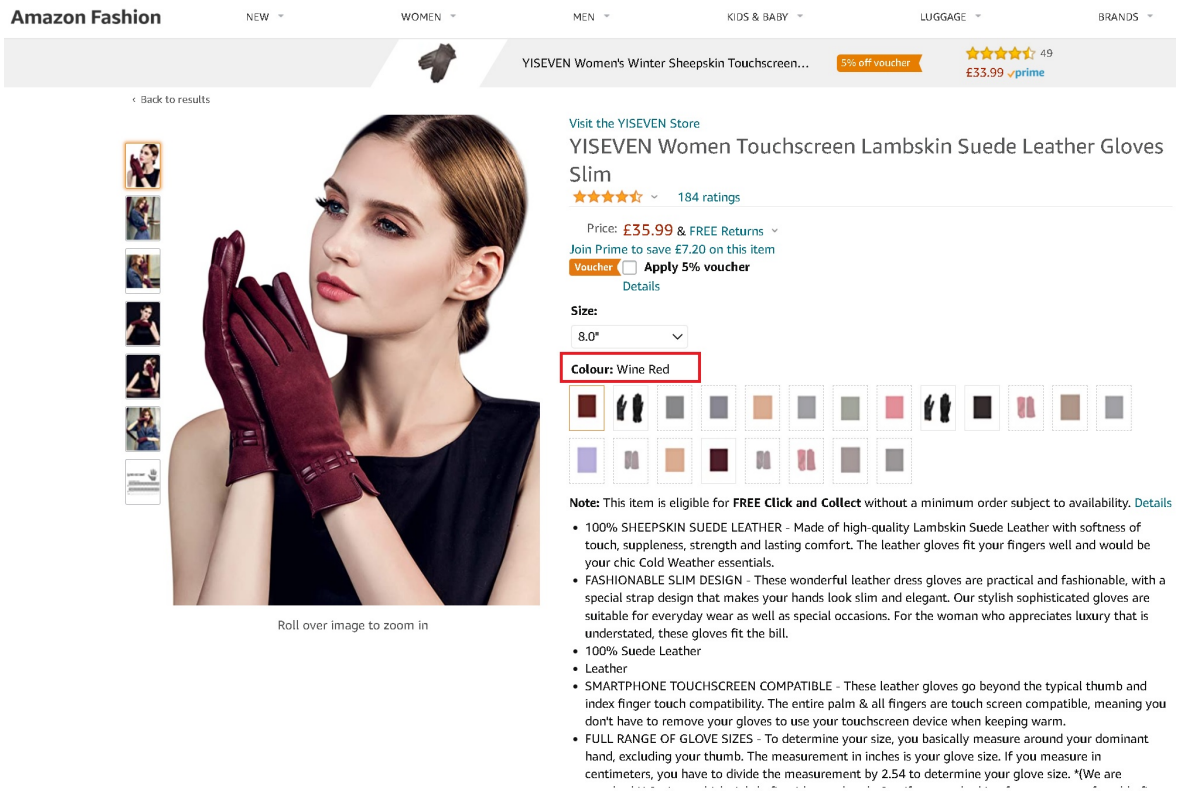


Figure 1.2: *Example of unreliable product text descriptions. It shows an ambiguous or unclear colour name, leading to potential confusion for VI shoppers.*

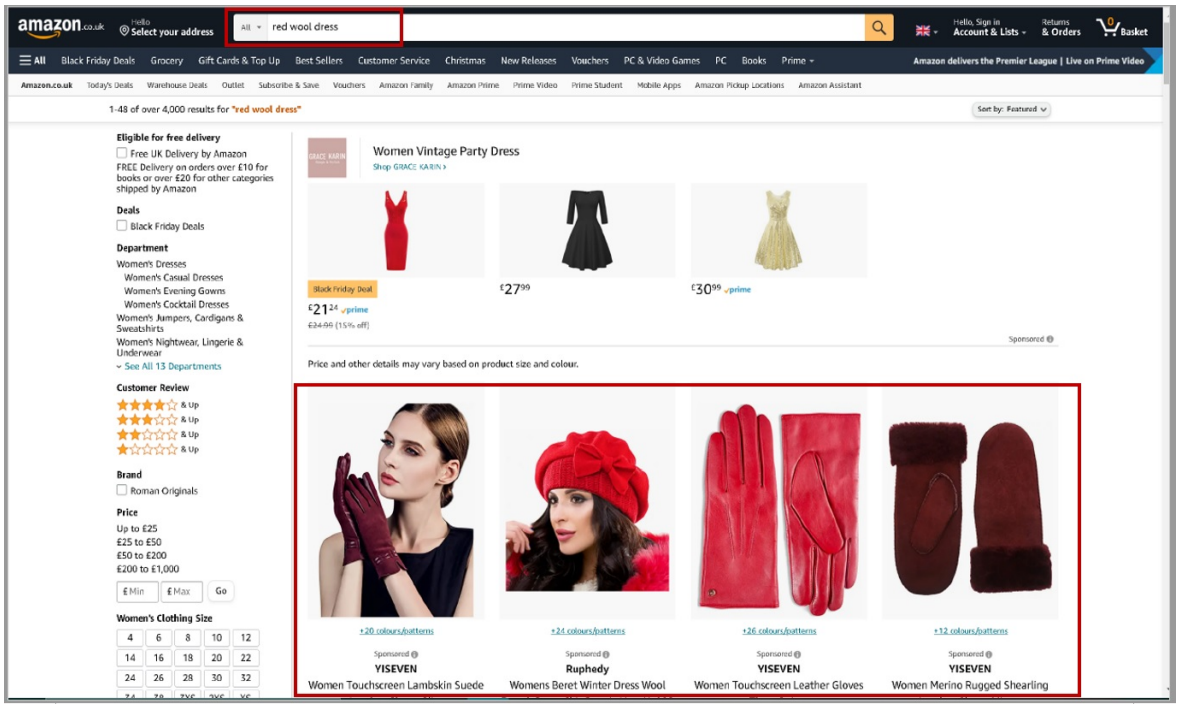


Figure 1.3: *Amazon's search engine returns items that do not match the query.*

1.2 Research Challenge

This section reviews the challenges and limitations outlined in the literature that examines VI people's experiences of purchasing clothing online and that attempts to propose solutions. It discusses these challenges under two main themes that align with the two main research areas of this thesis. The first is better understanding VI people's barriers when purchasing clothing online, and the second is developing automated, VI-friendly image captioning.

In regards to exploring VI people's experiences, only a limited number of studies have been published to date. These studies either investigate the challenges VI people face (Fuchs & Strauss, 2012; Tateno et al., 2020; Williams, 2021; Yang et al., 2015) or, from the provider side, highlight a lack of product information offered or how such information can only be found with difficulty (Stangl et al., 2018; Wang et al., 2021). None have yet combined the two to offer a comprehensive understanding of the barriers that VI people must overcome during clothes selection and coordination, nor have any explored VI people's behaviour when purchasing clothing online. Additionally, in regard to this last point, it was noted that research has largely been based on interview methods, leaving a research gap for observational studies.

This addition to the literature has thus been based on comprehensively exploring the limitations that VI people encounter when shopping online, and in doing so, removing limitations from the extant body of literature. This research has been motivated by the belief that exploring VI people's behaviours and interests when searching for, selecting, and coordinating clothing items online is essential to improve their quality of life. Beyond that, an additional goal has been to identify and synthesise possible solutions. To do so has required a detailed investigation of the elements of clothing captions that are most important for VI

individuals and the factors that influence them, such as celebrity or fashion style, when purchasing clothing. Exploring these aspects is crucial to providing better support for VI people. Furthermore, a point noted from the literature review was a lack of studies exploring how automated systems can be utilised to improve VI people's experiences of purchasing clothing online. In this regard, certain interesting aspects to consider have included VI people's trust in automated systems, the kinds of automated services they would like to have, and their fashion influences when purchasing or coordinating outfits. Considering these aspects has played a vital role in proposing suitable solutions.

When we consider potential assistance systems for VI people, those that provide better image captioning can be seen to hold high potential (Stangl et al., 2020b). Image captions provide useful descriptions of the details of images, which allow VI people to mentally visualise the contents of an image based solely on textual descriptions, and thus to make a decision based on those descriptions. They can also be readily included in web interfaces via HTML alt tags, which VI people are accustomed to using.

The typical approach to building an image captioning model relies on learning from existing datasets; this is problematic, however, due to such datasets' limitations. Previous research has shown that the ground truth (GT) captions of clothing datasets lack sufficient information on clothing attributes (Gao et al., 2020; Nguyen et al., 2021), indicating that such datasets may not necessarily contain gold-standard (or even good-quality) captions. It seems that analysing these datasets and exploring their limitations could be useful to guide the development of more advanced captioning models that efficiently capture the detailed elements of clothing images, though this has not previously been attempted in the literature. Providing a comprehensive analysis of the three widely used datasets' image captions (i.e., FashionGen (Sadeh et al., 2019), DeepFashion-MultiModal (Jiang et al., 2022) and FACAD (Yang et al., 2020)) whilst considering VI people's needs offers great potential to support the development of VI-friendly applications.

Up to now, most research on clothes image captioning has failed to consider VI people's needs if they are to successfully visualise the contents of images while relying solely on the captions generated. Surprisingly, only one study (Tateno et al., 2020) has addressed this challenge, and the researchers adopted a standard image captioning framework (i.e., encoder–decoder), which has been shown to have several limitations (Han, Zhang, et al., 2023; Ming et al., 2022; Stefanini et al., 2021; Zhuge et al., 2021). To overcome those limitations and thus further the efforts to develop VI-friendly clothing captions, there is scope now to adopt and leverage the recent development of LLM, applying those to simulate human behaviour (Aher et al., 2023) when generating VI-friendly captions.

Previous research has shown that VI people require detailed descriptions to visualise the contents of clothing images (Stangl et al., 2021). However, generating detailed captions is challenging, as the fine aspects of clothing images may be missed or misinterpreted by captioning models (Goenka et al., 2022; Zhuge et al., 2021). Recent research (Gao et al., 2020; Zhuge et al., 2021) has attempted to adopt Region of Interest (ROI), an approach which applies detection techniques to generate an object's fine-grained features. The adoption of this approach is inspired by its greater success in generating detailed descriptions in general image captioning (Anderson et al., 2018; Cornia et al., 2020; Kinghorn et al., 2018; Long et al., 2023; Luo et al., 2021; Shi et al., 2020). However, the findings reveal that the ROI approach is unsuccessful in generating detailed captions in the clothing domain. Instead, their

findings show that the grid-based approach, which extracts image contents by partitioning images into grids of equal size and mining visual information across all grids, is more suitable in the clothing domain. Their research, however, assesses the quality of the ROI approach using an existing clothing dataset, such as FashionGen, which fails to represent clothing attributes (Gao et al., 2020; Nguyen et al., 2021). In another attempt Li, Ye, et al. (2021) predefined salient clothing elements, including clothing colour, fabric and shape, and adopted an attribute detection technique. Their findings show that applying an early identification of clothing attributes leads to detailed captions with improved focus (Li, Ye, et al., 2021).

Existing research (Cai et al., 2022; Hacheme & Sayouti, 2021; Han, Zhu, et al., 2023; Moratelli et al., 2023a; Yang et al., 2020) has generated captions of clothing images that include basic elements such as their clothing category (e.g., a top, skirt, or coat) and the components of the garment (e.g., sleeves, a collar, or cuffs), as well as identifying some clothing features such as colours, and styles. However, the existing models frequently generate only a few words, which can be inadequate to represent the graphical aspects of an image (Tu, 2020; Yan, 2023), making them unsuitable for applications aimed at VI users. To our knowledge, no prior work has addressed how the descriptions of graphics in clothing images can be improved for this audience. With that in mind and considering the great success that has been achieved in adopting the ROI approach for captioning general images (Li et al., 2019; Long et al., 2023; Yin et al., 2019), it is pertinent to explore whether this approach, which extracts visual information from salient objects, may be superior in extracting information on fine clothing elements such as graphics.

1.3 Research Questions

In the previous section, we discussed the current research challenges and limitations. Here, we outline four main research questions (RQs) that build on these challenges and limitations and that we address in this thesis. Note that some of these main RQs will be split into sub-RQs later in the relevant sections of the thesis.

- RQ1: What challenges do VI people face when independently selecting and coordinating outfits while online shopping?
- RQ2: To what extent do VI people consider a specific celebrity or fashion style when selecting outfits?
- RQ3: What are the current VI-assistance systems, and what type of assistance do VI-people like to have to purchase clothing online?
- RQ4: How can we build an effective clothing graphic description model to generate useful descriptions for VI users?

1.4 Aim and Objectives

This thesis adopts the social model of disability Oliver (2013), which indicates that social barriers contribute to the challenges faced by people with visual impairment. Here we emphasise the need to focus on social change rather than concentrating solely on individual

impairments. The aim of this thesis is to explore the challenges faced by VI individuals when purchasing clothing online and propose a VI-friendly captioning model to enhance their online shopping experiences.

The key objectives of this thesis are:

- To conduct an observational study to understand VI people's behaviour while shopping for clothes online
- To perform semi-structured interviews to explore challenges and barriers in online clothing shopping for VI people
- To develop a novel model to improve the online experience of VI people when shopping for clothes online
- To implement an evaluation to critically evaluate the proposed system and identify further improvements

1.5 Project Procedures

The guidance proposed by van Miltenburg (2023) for conducting a successful project to develop an Natural Language Generation (NLG) system inspired the current project's procedure, which follows the recommended essential steps beginning with first understanding the audience through developing and evaluating the proposed system. Thus, this project was designed in three stages as shown in Figure 1.4. The first stage is to understand VI individuals' behaviour and barriers when purchasing clothes online. The second stage develops a user-friendly model for VI individuals, which is assessed in stage three. Stage one is presented in Chapter 3, followed by a discussion on stages two and three in Chapter 4.

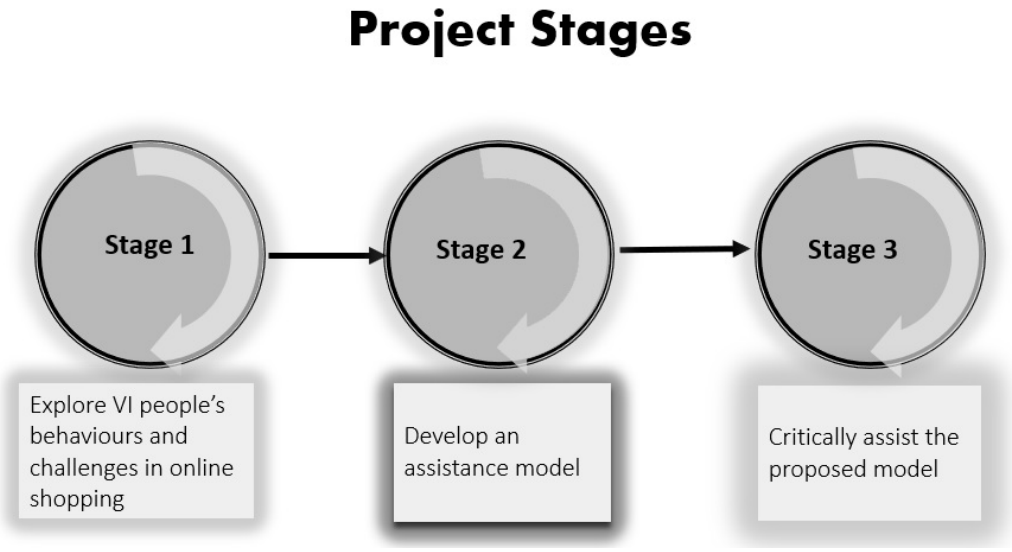


Figure 1.4: *Project stages overview.*

1.6 Contributions and Publications

1.6.1 Contributions

This thesis presents a comprehensive understanding of VI people's challenges in purchasing clothing online. It comprehensively analyses three existing clothing image-text benchmark datasets and assesses how to develop a possible VI-friendly captioning model. We contribute to the existing body of knowledge in the following ways:

1. **Showing that conducting a task observation and standalone interview studies produces a comprehensive understanding of the problem:** We show that conducting task observations and standalone interviews produces an in-depth understanding of participants' behaviours and challenges in purchasing clothing online. This has resulted in several implications for designing VI-friendly assistance systems to enhance VI people's online clothes shopping experiences.

2. **Assessing three fashion captioning datasets from blind people's perspectives:** We provide a comprehensive analysis of three commonly used fashion captioning datasets (i.e., FashionGen (Sadeh et al., 2019), DeepFashion-MultiModal (Jiang et al., 2022) and FACAD (Yang et al., 2020), considering VI users' needs.

3. **Creating a clothing graphic dataset (CGD):** We create the first specific clothing graphic dataset as a subset of three common fashion datasets (i.e., FashionGen, DeepFashion-MultiModal and FACAD). This dataset facilitates the evaluation of our proposed model.

4. **Showing that LLMs can be successfully adapted to simulate VI people's needs:** We show that LLMs can be leveraged to generate VI-friendly captions.

5. **Showing that an ROI-based approach can generate informative descriptions of graphical elements:** We demonstrate that the ROI-based approach can be used to generate rich descriptions of graphics.

1.6.2 Publications

Alluqmani, A., Harvey, M. A., & Zhang, Z. (2023, July). The Barriers to Online Clothing Websites for Visually Impaired People: An Interview and Observation Approach to Understanding Needs. In Proceedings of the 2023 ACM Designing Interactive Systems Conference (pp. 753-764).

1.7 Thesis Outline

This thesis is organised as follows:

In Chapter 1, we discuss the motivations behind this doctoral thesis and the existing research challenge. We also present our aims and objectives and discuss the procedures we followed to conduct this project. We then briefly introduce our contributions, which are discussed in detail in later chapters, and present the publication outcomes of this work. We end the chapter by introducing this thesis outline.

In Chapter 2, we review the literature in several areas related to this project. This chapter covers the literature exploring the challenges VI people face in everyday life activities. We also review the literature offering VI-friendly assistance systems for online clothes shopping and the literature on image captioning, with a special focus on the literature on clothing. Additionally, we review the literature on extracting clothing information from images to assist VI people. Furthermore, we review the literature on adopting LLMs to generate image representations. Finally, we review the literature on image captioning evaluation metrics. We end this chapter by discussing gaps in the literature.

In Chapter 3, we present our approach of conducting two complementary studies (observations and interviews) to explore VI people's challenges and needs. Specifically, we explore VI people's behaviours and barriers when shopping for clothes online. We also discuss how VI people could be better supported in online clothes selection and coordination. In addition, we discuss how automation could be adapted to enhance this demographic's shopping experience. This work was accepted as a full paper and presented at the ACM Designing Interactive Systems (DIS) conference in 2023.

In Chapter 4, we discuss our approach to comprehensively analysing the benchmark clothing image-text datasets. We assess whether these datasets meet the needs of VI individuals and whether they are suitable for use in applications that attempt to serve this population. Furthermore, we present our new clothing graphic dataset (CGD), which includes several clothing graphic items. We also present our model, adopting an ROI-based approach and LLMs to generate VI-friendly captions for clothing graphics. Furthermore, we discuss our evaluation approach, which includes automatic and human assessments with VI and sighted individuals. We compare captions generated from our model with those generated from the BLIP-2 captioning model and ground truth (GT) captions. This work is being prepared for submission to a journal.

In Chapter 5, we summarise several areas of this doctoral thesis. We provide a summary of the thesis chapters and a summary of the findings outlined in this thesis. In addition, we provide a brief discussion of the contributions of this work. Finally, in this chapter, we discuss ways in which this project can be expanded in the future.

Chapter 2

Literature Review

2.1 Introduction

In this Chapter, we review research that addresses VI people's everyday challenges (Section 2.2). In Section 2.3, we explore VI people's challenges in shopping environments (i.e., online and offline) in particular. Then, we explore more closely related research that explores VI people's challenges in purchasing clothing online.

Furthermore, in Section 2.4, we review research proposing VI-friendly assistance systems for online clothes shopping, which has a relevant role in our project. Reviewing this area allows us to examine current assistance systems and address their limitations. Next, we review image captioning, including both general image captioning and clothing captioning in Section 2.5. Due to a lack of research into clothing image captioning for VI users, in Section 2.6, we consider research in the broader area of clothing image representation for VI users.

In the following Section 2.7, we review large language models and explore the potential for adopting such models in image representation. Investigation of this area plays a significant role in building a VI-friendly captioning model. We end this chapter by addressing the image captioning evaluation research in Section 2.8.

2.2 Everyday Life Challenges Experienced by Visually Impaired People

VI people face a multitude of challenges when performing their everyday tasks. Independently navigating indoor and outdoor spaces, for instance, encounter numerous obstacles. One study (Jeamwatthanachai et al., 2019) attempted to understand blind people's experiences and behaviours in navigating indoor places, including universities, hospitals, malls, and airports. Researchers collaborated with 30 VI people by conducting questionnaires and interviews. The study revealed several difficulties faced by VI people especially when navigating unfamiliar places. In particular, VI people find it difficult to estimate distance and to find their way in such places. To overcome such difficulties, they require assistance when visiting places for the first time. For instance, they seek assistance in identifying objects such as furniture and performing certain tasks, such as hanging items, which can be challenging

when using a white cane. Moreover, in a large or crowded place, it can be very difficult for VI people to discover the area or navigate comfortably (Williams et al., 2013).

In most cases, however, navigating outdoors is less secure than indoor environments. The challenges of both environments have been investigated by Williams et al. (2013). They interviewed 30 VI participants from the United States to explore their challenges. The participants highlighted a number of difficult situations that threaten their safety while navigating outdoors. For example, they regularly encounter danger while crossing streets. This task is dangerous because not all drivers follow traffic rules, and sometimes road construction requires a change in route, making recall of the previous layout of little use. As a result, many participants had experienced injuries or near misses when crossing roads. Weather also affects the ability of VI people to navigate outdoors. For example, extreme wind and snow might render their white canes and assistance gadgets less effective. Sunlight is also a factor that affects blind people's behaviour while navigating outdoors since they have higher light sensitivity (Stangl et al., 2018).

In both indoor and outdoor environments, VI people might find changing service levels challenging, including going up or down stairs, stepping onto curbs, and going between blocks. They find detecting stairs particularly challenging, especially going down stairs and detecting the first and last stair (Zhao et al., 2018). Additionally, safety in public is a concern for VI people. For example, asking a stranger to help them reach home or stores can be risky. Being blind or partially sighted might put them in a situation where they ask support from an untrustworthy person. On the other hand, carrying an expensive gadget in public, such as a new phone or navigation system, could put them in danger as well (Williams et al., 2013).

VI individuals also face challenges in education, including struggles with mobility. One study (Otyola et al., 2017) investigated VI students' challenges in two Ugandan universities, Makerere and Kyambogo. To investigate the challenges from multiple points of view, researchers collected data from 200 participants in various roles at both universities. The researchers used questionnaires to collect data from lecturers, administrators, and non-visually impaired students and interviewed 50 VI students. The study revealed that VI students faced challenges in traveling to their university, navigating inside the university, and accessing their lectures on time. These challenges affected their learning progress because they often arrived late. As another education example, VI people often face difficulties in their reading and writing processes. There is a general lack of Braille language textbooks, and they may struggle to read printed books and information on white boards, challenges impacting the quality of their education (Bualar, 2018; Sahasrabudhe & Palvia, 2013). Often, VI students write slowly and because their braille machines make a great deal of noise, blind students might feel self-conscious in class (Otyola et al., 2017). Some learning subjects are inherently difficult for VI people; for example, it is difficult for them to conduct experiments in physics, chemistry, and biology (Sahasrabudhe & Palvia, 2013).

Other challenging tasks encountered in everyday life include cooking and shopping. Cooking is a challenging task for VI people as they have to find and identify ingredients, organise items in the kitchen, and throw out expired items. They also have to manage hazards when cutting food, dealing with hot items, and lifting heavy objects (Li, Dorst, et al., 2021).

The process of shopping represents one of the most challenging tasks (Burton, 2011; Lee et al., 2020) because it involves difficulties before, during, and after the shopping. Before going to stores, VI people face a number of tasks, including preparing a list (Yuan et al.,

2019), getting dressed, traveling to the store, and dealing with obstacles in both outdoor and indoor environments (Szpiro et al., 2016). Once they reach the store, they encounter difficulties navigating the store and identifying products (Liu, Ding, et al., 2019), and, upon arriving home, they have other difficult tasks, such as identifying products and organising items (Yuan et al., 2019).

These difficulties lead to numerous mental health challenges. Brown and Barrett (2011) studied the effects of visual impairment on quality of life, revealing that higher levels of visual impairment are linked to higher depressive symptoms. It also leads to lower degrees of life satisfaction (Habib & Irshad, 2018).

2.3 Shopping Challenges Experienced by Visually Impaired People

2.3.1 Offline Shopping Challenges

Several studies have investigated the challenges VI people face when purchasing offline. Lee et al. (2020) interviewed eight participants from the VI community about their product selection process, including the difficulties they face in confirming that they are selecting their desired product. The study revealed that the amount of product information affected their ability to shop efficiently offline. For instance, when purchasing fresh food, the expiration date could be difficult to locate on the product label. Such problems affected the selection of suitable products without assistance. The lack of clearly visible information on expiration dates and other information results in an unpleasant shopping experience for VI people. They have to either seek customer service assistance, which involves waiting some time for a customer service representative to become available, or restrict their shopping options to certain familiar products. Ultimately, as confirmed previously (Park et al., 2020), a lack of product information makes it nearly impossible for VI people to find and compare items without assistance.

Lack of information is not the only factor that impacts shopping self-sufficiency for VI people. Navigating inside shopping malls can be a stressful task (Lee et al., 2020), especially when stores are poorly organised. Examples of prohibitive organisation include poorly laid-out department divisions, labels that lack colour consistency, very narrow aisles (Yu et al., 2015) and items in the wrong department. Further, stores often change their layouts, which can confuse VI people, who must often rely on their spatial memory from past visits (Jones et al., 2019). When trying on clothes, VI people need to manage such challenges as closing the fitting room door; finding hooks for canes, purses or coats; looking for a mirror, and managing poor lighting. The payment process is also worth noting, as VI people often experience difficulty reading receipts, estimating their total before cashing out and using credit cards (Lee et al., 2020; Yu et al., 2015). The limitations of offline shopping have led VI people to prefer online shopping as an alternative. However, online shopping presents several limitations.

2.3.2 Online Shopping Challenges

VI people report that online shopping can give them a sense of reliability and control and is generally the preferred mode of shopping for clothes compared to the offline equiva-

lent (Williams, 2021). Nevertheless, online shopping comes with its own obstacles, including a lack of accessibility features in online spaces (Stangl et al., 2018).

Table accessibility is an often-cited example of an online accessibility limitation. Web designers often use tables to present information in a compact manner; however, when screen readers are employed, a number of challenges arise, including the necessity for a good memory to remember each table cell's content and its relationship to the columns and rows. The more complex the table, the more difficult it is for screen reader users to associate cells with their columns and rows (Whitney, 2020).

Several guidelines for iOS[1], Android[2] and web content[3] have been proposed to improve online accessibility for VI people. Unfortunately, most websites fail to adhere to such key VI accessibility criteria such as simple headings, high contrast ratios and consistent navigation. A study by Sohaib and Kang (2017) tested the accessibility of Australian business to consumer (B2C) e-commerce websites. Thirty B2C e-commerce websites were selected from the top sites in Australia. The selected websites were evaluated using A-Checker, an open-source web accessibility assessment system. The observed websites exhibited several accessibility problems that negatively impacted the experiences of VI visitors and were found to not adhere to even a basic level of web content accessibility criteria. Examples of such problems were pages without logical heading structures, pages missing alt-text, and pages with poor foreground and background colour contrast.

Seeking information also presents a challenge to VI online shoppers. One study compared the abilities of blind people and sighted individuals in accomplishing tasks online. The researchers (Bigham et al., 2017) found that blind people take roughly twice as long as sighted people to complete an identical task. Even screen readers are often insufficient in helping a VI person make a purchasing decision. This is because of accessibility problems that are hard for the screen reader to handle. Such limitations include the generalisation of product descriptions (Fuchs & Strauss, 2012) and insufficient details. The use of ambiguous language to describe products, such as nonstandard colour names, can further inhibit VI people from recognising key product details (Kaufman-Scarborough & Childers, 2009; Wang et al., 2021). Further, online retailers rely on images to attract consumers instead of descriptive phrases (Yang et al., 2015). In some cases, product images rather than product descriptions contain the crucial information, or product images are not associated with alt-tags, which would allow screen readers to describe the content of the image to a VI shopper (Sohaib & Kang, 2017). Despite the challenges associated with online shopping, purchasing clothing, in particular, introduces further difficulties to VI shoppers.

2.3.3 Online Clothes Shopping Challenges

Shopping online for clothes presents a set of specific additional challenges for VI users. For example, clothing websites often lack sufficient information on size, colour, material, outfit coordination and age appropriateness (Burton et al., 2012; Liu, Ding, et al., 2019; Williams,

[1]Apple Developer. (n.d.).*Building accessible apps.* Retrieved from https://developer.apple.com/accessibility/ios/. Accessed February 2024.

[2]Android Developer. (n.d.). *Build accessible apps.* Retrieved from https://developer.android.com/guide/topics/ui/accessibility. Accessed February 2024.

[3]W3C. (18 May 2021). *Blindness and vision loss.* Retrieved from https://www.w3.org/TR/WCAG10/. Accessed February 2024.

2021). Yang et al. (2015) interviewed nine VI people about their challenges in purchasing clothes online. The participants noted that most retailers rely on images to present the features of their clothing and often provide no alternative or alt-tag descriptions. Williams (2021) reported that blind people are unhappy with how difficult it is to obtain clothing information online, such as colour, pattern, graphics and fit. The participants in both studies highlighted the challenge of understanding textual clothing descriptions. Some retailers employ complex words to name items or describe product details, such as using unfamiliar or culture-dependent colour names like "teal", "sky blue" and "royal blue" (Wang et al., 2021; Williams, 2021).

Wang et al. (2021) investigated the difficulties faced by VI shoppers when seeking clothing information online. They found a general lack of useful visual descriptions and a heavy reliance on images to provide information. To overcome these limitations, most VI people seek assistance from family or friends when shopping online. However, some seek the assistance of online services. An example of such services is Aira[4] A monthly membership fee to Aira enables users to interact with an assistance agent for various activities that are difficult due to visual impairment. These activities include instructing the assistance agent to get a cab or submit a job application. Aira was introduced in (Williams, 2021) as a way of seeking help when shopping online. VI shoppers connect with the application, and the assistance agent accesses their computer through Team Viewer. Then, VI shoppers ask the assistance agent to describe the clothing item and ask questions, including if a different item matches.

Liu, Ding, et al. (2019) conducted an in-depth study of how blind people shop online. They interviewed 20 blind people (face-to-face or by telephone), aged 17 to 48. The participants described wanting to be 'normal' and to not stand out in mismatched clothing items. They wanted to be treated like sighted people but faced challenges in selecting and matching clothes online. According to Williams (2021), VI shoppers look for information to help them decide whether a clothing item suits their style, personality and age. They also seek information to help them identify whether a pattern, colour or style will match another clothing item (Stangl et al., 2018). Like any shopper, VI people seek information about the latest fashion trends (Burton et al., 2012) but face difficulties viewing other people's outfits to enrich their knowledge of fashion. Exploring VI people's influences, such as whether they follow specific fashion styles or celebrities when selecting clothing, is essential to providing insight into better support for VI people. This vast number of challenges encountered in clothing shopping websites demonstrates the need for VI-friendly assistance systems.

2.4 VI-friendly Assistance Systems for Online Clothes Shopping

Researchers have proposed assistance systems to improve VI people's online clothing purchase experiences, but the current efforts have noticeable limitations. A major limitation is that little work has been done to mitigate the lack of textual descriptions in online shopping. Moreover, the proposed assistance systems are not supported with sufficient details about the clothing.

Yang et al. (2015) manually designed a website prototype to guide the visually impaired

[4]Aira. https://www.airahome.com/en-gb. Accessed February 2024.

in clothes shopping. They classified clothing according to six attributes: style, sleeve length, material, fashion, colour and price. Each category was designed to include only nine items to make shopping easier for the visually impaired by lessening the number of things to remember (e.g., when a screen reader is used). When a user selects an item, detailed information appears, including seller information, price and a description. However, this design has multiple limitations. First, it is difficult to standardise the design to all online retailers, and the visually impaired require extensive training to use the new system. Second, generalising this design to clothing websites would involve extensive lists, as there is an enormous number of clothing categories. Lastly, the proposed design presents only the available information, which creates a number of limitations (Wang et al., 2021; Williams, 2021), including a lack of information (such as descriptions of graphics, buttons and pockets) and the use of unfamiliar terms to describe colours, such as "teal", "sky blue" and "royal blue". Ten VI individuals were recruited to assess the proposed design by browsing items and summarising their features, after which they were interviewed to discuss their comments.

Another study (Stangl et al., 2018) focused on redesigning clothing webpages. The researchers proposed converting the clothing data on the webpage into structured data. They also suggested adopting speech features that respond to specific queries, such as price, colour, material and other details. The proposed technique, called BrowseWithMe, involves various steps. First, product information (such as price and material) is extracted from text descriptions using Natural Language Processing (NLP). Other data, such as colour and clothing category, are obtained by analysing product images using Computer Vision (CV). The second component implements the voice command using the Chrome browser's built-in voice recognition and dictation API. Interviews were conducted with eight VI individuals to determine whether BrowseWithMe mitigates some of the challenges of online shopping. Study participants were asked to access three shopping websites (ASOS, H&M and Forever 21) and to use BrowseWithMe to extract certain information, including price, colour and material. Each participant was then asked to complete a survey to rate their experience. Another assessment was conducted with sighted participants to evaluate the accuracy of the generated descriptions in relation to the associated images. Sighted participants were shown a text description (either from the existing alt-text or the BrowseWithMe descriptions) and were asked to choose the image that best matched the description from among five images.

To overcome the limited product information provided by online retailers, another development Wang et al. (2021) proposed a new web search page called Revamp. This system analyses and summarises customer reviews to enrich the existing product information. It then sorts and presents customer reviews of specific products into positive and negative reviews. It also allows users to ask questions about products using either text or voice. Eight blind participants were recruited to assess how Revamp affected the shopping process. Participants were asked to browse six products—water bottle, men's tee, women's dress, bed sheet, Bluetooth speaker and chair cushion—using the original Amazon webpage. Specifically, they were asked to obtain certain product information: colour, logo, size and shape. The participants were then asked to repeat the same tasks using Revamp to see if it improved their experience compared to the original Amazon website. Upon completion of these tasks, the participants were asked to fill out a survey in which they rated their experience using Revamp. They were also asked to explain the reasoning behind their ratings. Tateno et al. (2020) also attempted to address the limited clothing information found on online websites. They built a dataset

limited to tops that includes ten features, such as colour, print, decoration, fasteners and sleeves. Product information is extracted from product images to provide better image captions. The proposed system provides informative product captions that include at least three to five clothing features. Researchers qualitatively assessed the accuracy of the generated captions by considering successful and failed example descriptions.

While the proposed systems discussed above are crucial for enhancing the online shopping experience of VI individuals, the generation of rich data from product images remains an essential and high-demanded technology among VI users (Stangl et al., 2020b). Furthermore, generating automated VI-friendly image captions opens the opportunity for further advancements built upon such caption, such as using the generated data to develop voice assistance systems, or to support outfit recommendation systems, resulting in more informed and tailored recommendations for VI users.

2.5 Image Captioning

2.5.1 General Image Captioning Methods

Image captioning is the process of converting an image into text. It involves extracting and classifying attributes from an image and generating useful captions. Image captioning techniques are classified into two types (Hossain et al., 2019): traditional and deep machine learning captioning methods.

Traditional Image Captioning Techniques

Traditional image captioning techniques are either template or retrieval-based. In the template-based technique (for the framework, see Figure 2.1), a caption template is predefined with blank slots. A sentence template suggested by Yang, Teo, et al. (2011) is structured as follows: (Nouns-Verbs-Scenes-Prepositions). Different attributes, actions, or objects are then extracted from the images and placed in the blank spaces (Lebret et al., 2015; Mitchell et al., 2012; Ushiku et al., 2015).

In template-based image captioning, certain visual elements (e.g., objects, scene, and actions) are specified and detected using object detection, classification, or prediction models (e.g., support vector machine [SVM] (Lin et al., 2007) or hidden Markov model (Rahman et al., 2020)). Then, the detected visual elements are assembled according to grammar rules (Ushiku et al., 2012) or a predefined template (Mitchell et al., 2012).

Farhadi et al. (2010) used SVM to detect three elements (objects, actions, and scene). Yang, Teo, et al. (2011) detected quadruplet elements (nouns, verbs, sentences, and prepositions). On the other hand, Krishnamoorthy et al. (2013) focused on subject, verb, and object information.

The template-based image captioning technique can generate grammatically correct captions, but it is necessarily highly restrictive in format and cannot generate variable-length captions (Hossain et al., 2019).

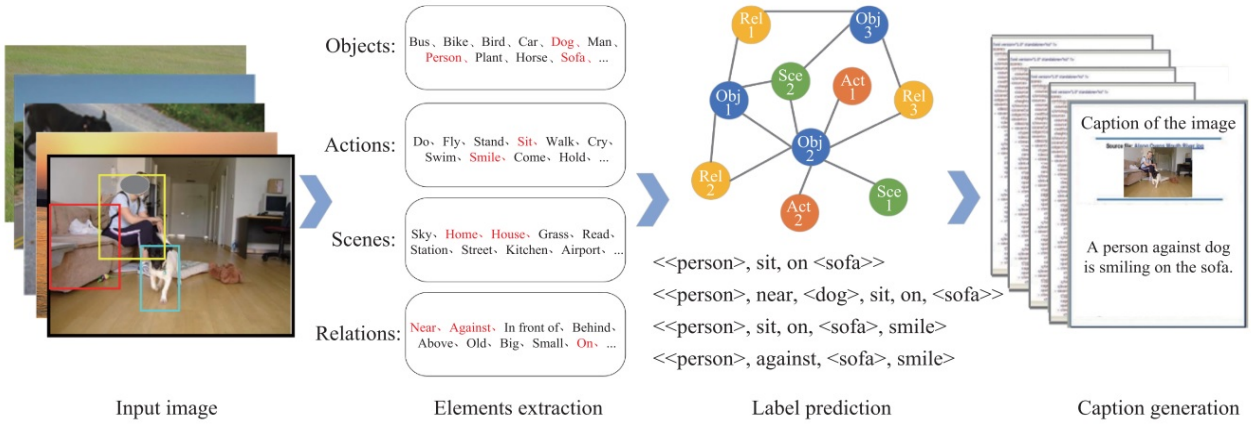


Figure 2.1: *Template-based framework. It is based on a predefined template. Object elements are extracted from the image and placed in the blank slots in the template to form a caption. The figure was adopted from Ming et al. (2022).*

The retrieval-based technique retrieves captions from a set of prespecified sentence pools (Gupta et al., 2012; Kuznetsova et al., 2014; Mason & Charniak, 2014). A corpus of image caption pairs is stored. To generate a caption for a query image, the similarity between the target image and the pooled images is calculated, and the caption of the image most similar to the target image is used for image captioning or a caption composed of retrieved sentences. Various methods are adopted when applying the similarity comparison.

Figure 2.2 shows an example of a retrieval-based image captioning technique proposed by Ordonez et al. (2011). They identify the similarity between the query image and previously stored image caption pairs in two steps: First, they identified the global features of the query image and selected the top images similar to the query image. This step identifies the similarity based on the overall scene. In the second step, they detected salient objects in the query image to generate more specific and matching captions and to match features such as colour. Gupta et al. (2012) extracted phrases from the image-stored caption pairs and selected the most similar images based on the phrase. The visual concept discovery framework, proposed by Sun et al. (2015), extracts and filters unigram and dependency unigrams on the basis of visual discrimination. These extracted data were grouped into concept clusters based on visual and semantic similarities. The downside of the retrieval-based technique is that it can generate captions that are irrelevant to the images (Bai & An, 2018).

Traditional image captioning methods, either retrieval and template based, were adopted in the early phase of image captioning. However, each method category has its own limitations. In the case of new combinations of images and novel scenes, the retrieval-based approach leads to captioning unrelated of the image content. Although the template-based method provides relatively more relevant image descriptions, because of the nature of employing a predefined template, the resulting captions are less like human descriptions. However, recent research has mainly relied on deep-learning approaches.

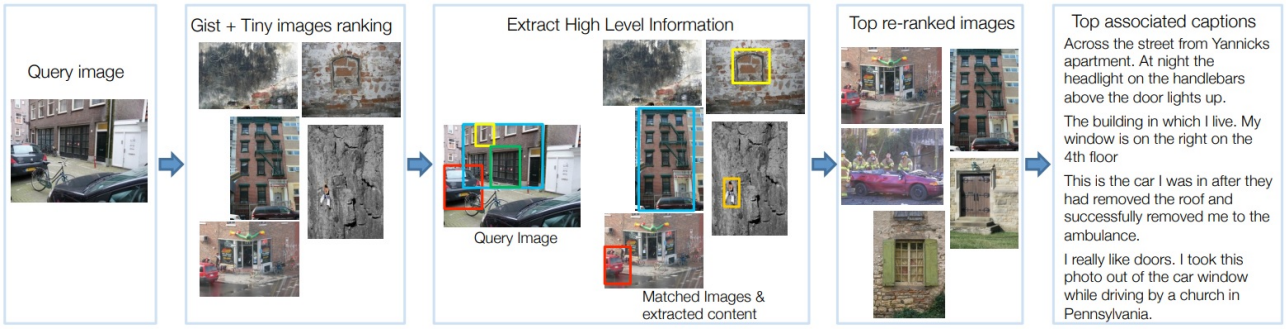


Figure 2.2: *Retrieval-based framework. Captions are derived on the basis of the similarity of the prestored image caption pair. The similarity matching process is based on the overall scene and object similarity. This figure was adopted from Ordonez et al. (2011).*

Deep learning Image Captioning

Deep machine learning captioning techniques overcome the image captioning limitations of traditional techniques (Bai & An, 2018). One strength of deep learning-based captioning techniques is that feature extraction and selection are automatically implemented by performing a training process using a large training dataset and not determined heuristically by humans (Mazurowski et al., 2019). Moreover, when generating image captions, these techniques do not rely on preexisting sentences (Sharma et al., 2020).

In the early phase of using deep learning techniques, retrieval- and template-based methods were augmented by neural networks to automatically extract and select features instead of utilising hard-coded and hand-engineered features. This enhanced version of retrieval and template techniques has been effectively used in the literature (Lebret et al., 2015; Socher et al., 2014; Yan & Mikolajczyk, 2015). Lebret et al. (2015) proposed a model that can extract phrases from an image and use these phrases to generate image descriptions. Their model was trained to learn the metrics between image representation and the phrases used to describe images. (Karpathy et al., 2014) adopted a retrieval-based method using neural networks to retrieve images (given a sentence query) and retrieve sentences (given an image). A dataset contains images and the corresponding sentence descriptions used to train the model. The model learned the association between image fragments (i.e. objects and attributes) and sentence fragments (using a dependency tree relation).

The encoder-decoder framework is commonly used in deep learning image captioning (Cho et al., 2014; Li, Ye, et al., 2021; Sutskever et al., 2014; Tateno et al., 2020; Yang et al., 2020). A neural network method encodes the input image, and then a Recurrent Neural Network (RNN) generates captions. Several works in image captioning have employed an encoder-decoder-based architecture using a Convolutional Neural Network (CNN) and RNN (Devlin et al., 2015; Gan et al., 2017; Mao et al., 2014; Vinyals et al., 2015). The CNN model is an encoder to extract visual features from images, while the RNN model decodes these features.

The initial aim of the encoder-decoder framework was to generate overall textual descriptions representing query images. It was first introduced by Vinyals et al. (2015). Their model is called a Neural Image Caption (NIC) model and is depicted in 2.3. They adopted the pre-trained CNN model to extract global visual features and feed these features to a Language

Model (LM). Although global image captioning demonstrates simplicity, the major problem with this approach is that it treats significant and non-significant regions equally, making it unsuitable for specific or fine-grained descriptions (Ming et al., 2022; Stefanini et al., 2021).

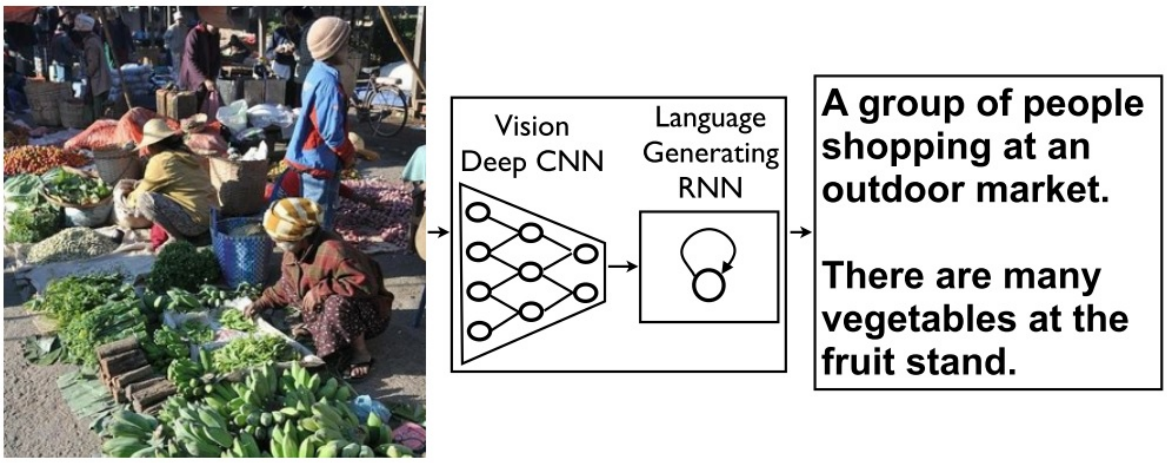


Figure 2.3: *Encoder-decoder framework introduced by Vinyals et al. (2015). The CNN model encodes visual features and is fed to the RNN model to generate captions.*

To extract fine-grain information from images, several approaches leverage the additive attention mechanism introduced in the machine translation field. In the additive attention mechanism, rather than using a fixed-size vector to represent the source data (e.g., image) and relying on a fixed-length representation, it proposes forcing the decoder to pay attention to a certain part of the encoding sequence.

Additive attention is mechanised as adopted in image encoding in grid (Lu et al., 2017; Xu et al., 2015; You et al., 2016) and ROI approaches (Anderson et al., 2018; Long et al., 2023; Pedersoli et al., 2017). Rather than considering the overall visual features alone of the image, this technique enables the model to focus on a specific region of the image when generating captions. The main aim is to mimic the natural human visual system from two points of view: top-down and bottom-up (Anderson et al., 2018). Figure 2.4 presents the related works in two main frameworks: grid and ROI based.

Grid-based Approach The top-down attention mechanism, which is grid based, partitions images into grids of equal size and extracts visual information across all grids. In this approach, which has been employed in a substantial amount of image captioning work, attention is applied to grids based on attention weights (Lu et al., 2017; Xu et al., 2015).

Several approaches have been introduced to enhance the level of granularity in visual representations (Lu et al., 2017; Xu et al., 2015; Zhang et al., 2021). Xu et al. (2015) introduced a new method that extracts visual features using the pre-trained VGGnet model and then adopts additive attention (Bahdanau et al., 2014) to compute weights to each grid, indicating the significant level of each grid when generating captions. The decision of calculating weights and where to look in each step of the caption generation is made in either a hard or soft attention mechanism. In the hard attention mechanism, the model

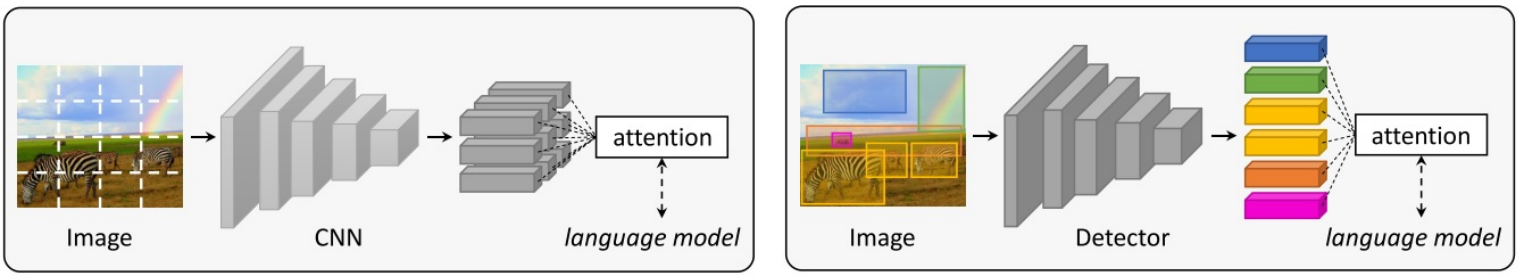


Figure 2.4: *Visualisation of two visual encoding frameworks: grid and ROI based. Grid based (left): Visual features are extracted from the overall image. ROI-based (Right): Visual features are extracted from certain regions in the image. This figure was adopted from Stefanini et al. (2021).*

focuses on certain regions. In the soft attention mechanism, caption generation is based on a weighted sum of all image regions. Instead of looking into the image to determine where to look, in the model constructed by Xu et al. (2015), Lu et al. (2017) introduced the spatial attention method, which automatically decides when to look and where. Their model tackles the drawback of the model by Xu et al. in identifying visual elements in every step of generating captions, which diminishes the effectiveness of incorporating the visual elements. In reality, some words do not have corresponding visual information such as 'of' and 'a'. Spatial attention decides when to rely on visual elements and on the LM in each step of generating the subsequent word. Further models have been proposed to extract visual features that adopt a global grid framework while augmenting other explicit representations to generate more precise captions (Wu et al., 2016; Zhang et al., 2021).

Although adopting additive attention using grid-based representation outperforms other conventional encoder-decoder image captioning frameworks (Lu et al., 2017; Xu et al., 2015), it fails to generate specific and accurate descriptions (Ming et al., 2022). Moreover, the grid-based framework does not consider the specific content of images, nor does it consider objects that may fall across the regions (Anderson et al., 2018).

Region of Interest-based Approach The ROI was proposed to overcome the limitations of the grid-based approach. It applies detection techniques such as Faster R-CNN (Ren et al., 2015) to generate fine-grain features of objects. The Faster R-CNN object detection model identifies objects presented in the image and localises them with bounding boxes. In the first stage, the Region Proposal Network generates object proposals in which each object is identified with a bounding box. In the second stage, the ROI is pooled to extract features for each proposed bounding box. This framework has achieved significant success and is commonly adopted in image caption generation (Anderson et al., 2018; Cornia et al., 2020; Kinghorn et al., 2018; Long et al., 2023; Luo et al., 2021; Shi et al., 2020).

Pioneering work was introduced by Anderson et al. (2018). The authors proposed a model that combines two mechanisms: top-down (grid based) and bottom-up (ROI based). They used the Resnet-101 CNN model in conjunction with the Faster R-CNN model. The model obtained high performance in identifying objects, object attributes, and the relationships between objects compared with the same model with the top-down mechanism alone. The success of the ROI-based approach has encouraged researchers to incorporate it with

supportive techniques to enhance accuracy and generate diverse captions. For example, Fei (2022) proposed a new model that evaluates the salience object and suggests regions that influence the output performance most. The evaluation was adopted in three steps: First, learnable mask perturbation is performed to identify important image regions by using a mask perturbation technique, which plays a vital role in evaluating the regions that are most effective in generating useful captions. Second, the attention-aligned network is used to adjust the attention weights based on the identified most important regions. Finally, the training and inference step is carried out to fine-tune the the model to adopt the aligned attention weights. By contrast, Du et al. (2023) proposed the application of object analysis in the early phase of visual feature extractions. Their proposed analysis adopts two techniques: the PoolNet model (Liu, Hou, et al., 2019) to extract object salience and the context-aware model (Goferman et al., 2011) to extract background saliency.

The adoption of the ROI has been shown to be successful in enhancing image captioning, considering the representations of salient regions, including humans and objects and their attributes (Kinghorn et al., 2018). It is also a significant technique in generating dense captions (Li et al., 2019; Long et al., 2023; Yin et al., 2019).

LLM-based Image Captioning

Recent research has exploited the emerging abilities of LLMs (Wei, Tay, et al., 2022) with the aim of enhancing image captioning approaches and generating more accurate and richer captions. Wang et al. (2023), harnessing the user alignment capability of LLMs, have proposed the Caption AnyThing model to generate captions of user-selected elements in images. The captions are controlled by users' preferences, such as preferred language, sentiment (i.e., positive or negative) and factuality (i.e., imaginary or factual). They adopt LLMs together with zero-shot models, including the Segment Anything model (Kirillov et al., 2023) and the BLIP-2 model (Li, Li, et al., 2023). The findings outlined in their qualitative analysis show that LLMs enable the generation of more accurate captions as well as captions adapted to the user's preferences. Zhu, Chen, Haydarov, et al. (2023) exploit LLMs' extensive intelligence to ask essential questions about the query image and use the Visual Language Model (VLM) to provide reasoning and answer LLMs' visual questions, mainly using ChatGPT and BLIP-2. To evaluate the ChatCaptioner in (Zhu, Chen, Haydarov, et al., 2023), human evaluation was undertaken to assess the informativeness of captions generated from their model in comparison to captions generated by the BLIP-2 captioning model and ground truth (GT) captions. Items (i.e., images and their GT captions) are obtained from three datasets: COCO (Lin et al., 2014), WikiArt (Saleh & Elgammal, 2015) and CC (Sharma et al., 2018). The results show that ChatCaptioner recieved three to four times more votes than the BLIP-2 and GT methods.

2.5.2 Clothing Image Captioning Methods

Clothing image captioning presents more challenges than general captioning because it requires fine-grained features to extract useful representations of clothing images rather than coarse representations, which answer only the questions of what and where (Mirchandani et al., 2022). Another challenge is that it is not optimal to adopt general pre-trained models in fashion tasks (Goenka et al., 2022; Han, Zhang, et al., 2023; Zhuge et al., 2021).

In light of the above-mentioned challenges in clothing image captioning, the encoder–decoder framework employed in general image captioning is unsuitable for efficient generation of useful descriptions from clothing images (Nguyen et al., 2021). The encoder–decoder is generally trained with maximum likelihood estimation, which generates the most likely next word based on the previous word without paying attention to attribute words, such as colour, size and material. Unlike image captioning for general images, presenting attribute descriptions, such as wool, sleeve and cotton, is essential to producing useful fashion descriptions (Yang et al., 2020).

Researchers propose several supportive techniques to overcome this challenge. Yan (2023) use an encoder–decoder with two classifiers to generate attractive captions. One classifier produces information about matching colours in clothing items, and the other determines the garment type (e.g., dress, skirt or shirt). This division is important in determining the choice of the most relevant attributes, as clothing items of the same type share the same attributes (Han, Zhang, et al., 2023; Yan, 2023; Yang et al., 2020). Their model was evaluated using their proposed clothing dataset and received scores of 0.205, 0.206 and 0.931 scores in BLEU-4, METEOR and CIDEr, respectively. Another researcher Yang et al. (2020) proposes the adoption of an encoder–decoder with hard attention (Vinyals et al., 2015) and rewarding reinforcement techniques (i.e., attribute-level and sentence-level semantics). Attribute-level semantics measure the consistency of the attributes between generated and GT captions, whereas sentence-level semantics focus on the sentence's semantic meaning. The generated captions were evaluated using their proposed Fashion Captioning Dataset (FACAD) and received scores of 24.2, 23.2, 47.2, 114.8 and 21.9 in BLEU-4, METEOR, ROUGEL, CIDEr and SPICE, respectively. Moreover, Moratelli et al. (2023b) propose enriching captioning by extracting additional information from external description sources. The proposed model was evaluated using the FACAD and received scores of 27.3, 0.6, 11.5, 22.3 and 84.5 in BLEU-1, BLEU-4, METEOR, ROUGE and CIDEr, respectively.

Following the greater success of the ROI-based approach in general image captioning (Anderson et al., 2018; Cornia et al., 2020; Kinghorn et al., 2018; Long et al., 2023; Luo et al., 2021; Shi et al., 2020), some researchers explored the further adoption of this approach in the clothing domain. Gao et al. (2020)compared results obtained from the existing ROI-based models with results generated from their grid-based proposed model called FashionBERT. The FashionBERT model cuts images into grids and generates features from each grid. Their assessment using the FashionGen dataset shows that the grid-based approach is superior to the ROI-based approach in the clothing domain. The KaleidoBERT model (Zhuge et al., 2021) similarly adopts the grid method but employs image–text pre-alignment to overcome the problem of semantic representation. A comparison of image captioning using the KaleidoBert and FashionBert models reveals that the former outperforms the latter using the FashionGen dataset. The Fashionbert model received scores of 3.30, 9.80, 29.7 and 30.1 in BLEU-4, METEOR, ROUGE-L and CIDEr, respectively, while KaleidoBert model scored 5.70, 12.8, 32.9 and 30.7, respectively. Both the FashionBERT and KaleidoBERT models outperform ROI-based models OSCAR (Li et al., 2020) in image captioning (Zhuge et al., 2021). However, these models are assessed based on the existing clothing dataset (i.e., FashionGen), which lacks sufficient clothing attributes (Gao et al., 2020; Nguyen et al., 2021), making it an inadequate gold standard dataset. Considering this limitation and the findings revealed by Li, Ye, et al. (2021), which show early defining of salient clothing elements,

and then applying the detection technique results in the generation of detailed descriptions with more focus. This suggests that the adoption of the ROI-based approach also generates captions with better focus, thereby enabling the generation of detailed descriptions of fine-grained clothing elements, such as graphics.

2.5.3 Clothing Image–Text Datasets

Recently, researchers have focused on the significance of exploring problems related to clothing image captioning. Several researchers have presented novel clothing datasets to train and assess models for generating descriptive captions tailored to clothing images. Jiang et al. (2022) introduced the DeepFashion-MultiModal dataset, which was built upon the DeepFashion clothing category prediction dataset (Liu et al., 2016). It includes more than 11,000 clothing items and covers 48 categories of clothing. Another example is the FACAD introduced by Yang et al. (2020), which contains over 900,000 images and detailed image information obtained from the Internet. The dataset not only provides a diverse range of clothing images - over 70 clothing categories and 900 attributes - but also offers fine-grained descriptions of clothing items. Another clothing image–text dataset was introduced by Sadeh et al. (2019). It contains 293,008 photographic images of clothing and descriptions annotated by fashion professionals. Forty-eight clothing categories were included in the dataset. In addition to this widely used clothing image–text dataset, other datasets, such as InFashAIv1 (Hacheme & Sayouti, 2021) and FACAD170K (Cai et al., 2022), have been developed to serve clothing image captioning applications.

The availability of clothing image captioning datasets has facilitated the development of clothing image captioning models, but researchers in the field of clothing image captioning have highlighted the lack of sufficient clothing attributes in these datasets (Gao et al., 2020; Nguyen et al., 2021). These limitations may mean that the datasets do not necessarily yield gold standard captions, so a comprehensive analysis of these datasets' limitations is essential. Identifying the limitations of clothing image captioning datasets guides the development of more advanced captioning models that are more efficient at capturing the fine-grained elements of clothing images. This will also shed light on the need for new datasets that mitigate current limitations.

2.6 Clothing Image Representation For VI Users

Several studies have been conducted to extract clothing data from images to assist people with visual impairment in understanding clothing images. For instance, by adopting a classification technique, Rocha et al. (2019) propose a solution to assist blind people in multiple tasks, including pattern and stain identification, as well as information about the state of the clothing (i.e. dirty, wrinkly or in good condition) by adopting a classification technique. They then extend their work (Rocha, Soares, et al., 2023) by adopting object detection methods (i.e. Mask R-CNN) to detect stains (i.e. coffee or wine). Further work (Rocha, Pinto, et al., 2023) adopts the YoloV5 detection model (Jocher, 2020) to identify defects, including holes. In addition to providing VI people with only raw clothing data, some work aims to suggest clothing combinations (Rocha et al., 2021; Silva et al., 2023; Yang, Yuan, et al., 2011). The application of classification, detection, and clothing combination techniques

provides limited insights into clothing elements. This limitation restricts VI users' ability to effectively visualise the clothing items based on the provided data.

Unlike classification and detection models, image captioning provides useful and significantly more information, making it a high-demand technology among VI individuals (Stangl et al., 2020b). The current work, however, which adopts image captioning to assist VI people, is designed for general captioning (Arystanbekov et al., 2023; Manay et al., 2022; Tiwary & Mahapatra, 2023). Tateno et al. (2020) were the first to take the initiative in addressing the needs of VI individuals in clothing image captioning, however, their system has several limitations. First, it adopts a standard encoder–decoder image captioning framework, which suffers from limitations in general-image (Ming et al., 2022; Stefanini et al., 2021) and fashion-image (Goenka et al., 2022; Han, Zhang, et al., 2023; Zhuge et al., 2021) captioning. Second, this model fails to provide information about essential fashion elements, such as the presence or number of pockets and is unable to describe multiple colours and graphics on items of clothing.

2.7 LLMs in Image Representation

LLMs such as Vicuna (Chiang et al., 2023), Alpaca (Taori et al., 2023), LLaMA (Touvron et al., 2023), GPT-3 (Brown et al., 2020) and GPT-4 (OpenAI, 2023) are trained on massive amounts of data with a large number of parameters, leading to impressive successes in recent years (Chowdhery et al., 2022; Thoppilan et al., 2022). These powerful models have been leveraged to generate useful applications, such as fine-tuning the LLM to follow human instructions, as with InstructGPT (Ouyang et al., 2022), and, in the case of ChatGPT, handling human-like conversations.

LLMs have demonstrated various pioneering abilities (Naveed et al., 2023; Yin et al., 2023), including comprehensive reasoning, chain-of-thought ability, linguistic knowledge and fluency and the generation of customised text based on different needs. These skills have been exploited in visual-text applications using multiple methods.

One method of employing an LLM in visual-text applications is to have it serve as a planner and controller. In HuggingGPT (Shen et al., 2023), the LLM (GPT-3.5 Turbo and text-Davinci-003) plays the role of core controller, while expert models on the Hugging Face platform are the executors. When a user inputs a query, the LLM breaks it down into several tasks and determines the sequence in which they should be executed. It then allocates these tasks to the relevant expert models, combines their results and creates useful responses. One notable advantage of HuggingGPT is its ability to establish connections with diverse expert models, enhancing its capability to process queries of varying modalities effectively. However, HuggingGPT is limited solely to the Hugging Face domain, which cannot be generalised to more complex tasks. Additionally, it cannot answer semantic questions or make scientific inferences to answer queries relying on scientific knowledge. To overcome the limitations of HuggingGPT, Chameleon (Lu, Peng, et al., 2023) employs an LLM (gpt-3.5-turbo and gpt-4) to act as a controller and to communicate with various domains, including Hugging Face, Web searches, code and GitHub. Involving a wider range of domains enables Chameleon to address more complex queries, such as semantic understanding, science and mathematical queries. GPT4Tools (Yang, Song, et al., 2023), on the other hand, fine-tuned ChatGPT (GPT-3.5 Turbo) using an instruction dataset that contains instances of user questions and

replies built based on chain of thought prompting. Instead of invoking expert models from certain domains, the GPT4Tools LLM invokes expert models from a tool pocket that includes certain expert models, such as image generation, image understanding (e.g., BLIP) and face recognition models. In visual-text dialogue scenarios, the LLM is employed to control and invoke required models in a multi-round manner. Visual ChatGPT (Wu et al., 2023) uses (text-davinci-003) LM and MM-REACT (Yang, Li, et al., 2023) uses the (gpt-3.5-turbo) model to control the dialogue. They consider the need for multi-rounds to extract dialogue history and collaborate with multiple expert models.

Another method implemented to leverage the instruction-following and resonating abilities of the LLM is to establish a conversation between the LLM and VLM. In this case, the LLM acts as a reasoner (questioner), while the VLM acts as an observer (answerer). Implementing prompt engineering techniques enables the LLM to continue asking questions to generate informative responses. This method has been adopted in the ChatCaptioner (Zhu, Chen, Haydarov, et al., 2023) image captioning model and in the IdealGPT (You et al., 2023) image questioning model. Both models employ ChatGPT(gpt-3.5-turbo") for question generation, while the BLIP-2 (Li, Li, et al., 2023) model is used to answer questions. ChatCaptioner creates a fixed question, "Describe this image in detail", to establish a conversation, while IdealGPT uses the user's inquiry to guide ChatGPT to generate sub-questions.

Unlike ChatCaptioner, IdealGPT implements a multiple-round technique to generate reliable responses to answer visual commonsense questions. Another similar example is the IMMO framework (Yang, Chen, et al., 2023), which adopted a multiple-round conversation between LLM and VLM using Vicuna-7b (Chiang et al., 2023) and BLIP-2. While ChatCaptioner and IdealGPT adopted zero-shot training using prompt engineering, IMMO applied fine-tuning training to enhance understanding between the observer and the reasoner. Specifically, this improved the ability of an observer to be optimised to the correct visual information for more accurate answers and improved the reasoner to ask questions with consideration for the reasoner's ability.

Text refinement and semantic understanding are also used to leverage the ability of the LLM. The MiniGPT model, for example, uses the Vicuna LLM to generate detailed descriptions using visual features generated from visual understanding models (such as ViT-G and Q-Former) and aligned with the LLM through linear layers. Similarly, Wang et al. (2023) uses text refinement and semantic understanding to generate object descriptions, asking the LLM to refine the descriptions based on user-required style, language or text length. In the same way, Lu, Lu, et al. (2023) borrows the knowledge and refinement ability of the LLM to generate recipes of users' demand dishes. In the first phase, the LLM proposes a recipe, which is then used to generate captions involving text-image and image captioning models. These captions are given to the LLM to consider when refining the proposed recipe and producing the overall output.

In addition to taking advantage of the above-mentioned skills of the LLMs to improve image-text generation models, using the simulation ability (Argyle et al., 2023; Simmons, 2022) is another approach worth exploring. Specifically, researchers might leverage the ability of LLMs to simulate human behaviour (Aher et al., 2023) and adopt it to develop models to generate text customised to the unique user's needs, for example, developing a model to generate textual data considering VI people's needs.

2.8 Image Captioning Evaluation

In the image captioning domain, models are evaluated using either automatic or human evaluation metrics. Existing automatic image captioning evaluation metrics, including BLEU (Papineni et al., 2002), ROUGE (Lin & Hovy, 2003), METEOR (Banerjee & Lavie, 2005), CIDEr (Vedantam et al., 2015) and SPICE (Anderson et al., 2016), have commonly been used in assessing image captioning models (Celikyilmaz et al., 2021; Lu et al., 2017; Moratelli et al., 2023b; SS et al., 2023; Tang et al., 2023; Vinyals et al., 2015; Zhang et al., 2021). These metrics use GT data, which are human-written captions, to assess automatically generated text by calculating similarity score. We discuss the automatic captioning metrics in subsection 4.5.3, where we review research that recruited humans to assess image captioning models.

Researchers in image captioning have conducted human evaluation with sighted evaluators to assess automatic captions using several criteria. The most widely assessed criterion when conducting human evaluation is judging the accuracy of such captions (Chen et al., 2022; Du et al., 2023; Fei, 2022; Hessel et al., 2021; Vinyals et al., 2015; Yi et al., 2020). This criterion is used to judge whether the generated caption contains any errors such as incorrectly describing the associated image. The informativeness or richness of information is also assessed when conducting human evaluation (Fei, 2022; Zhu, Chen, Haydarov, et al., 2023). This criterion measures the extent to which the automatically generated caption can produce sufficient information to represent the associated image. Human evaluation is also conducted to assess the naturalness of the generated text, which refers to its grammaticality and fluency (Fei, 2022; Lan et al., 2017). Human evaluators assess the generated captions in comparison to state of the art models. Another criterion assessed by conducting human evaluation is the object hallucination problem (Chen et al., 2022), which judges whether the automatic caption describes objects that do not exist. Human evaluation has also been conducted to judge how attractive the automatic clothing captions are to be used In marketing applications (Yang et al., 2020).

Among various research efforts on human image captioning evaluation, the comprehensive study by Kreiss et al. (2022) is notable for its inclusion of both VI and sighted evaluators. In their image captioning evaluation study, automatically generated captions were judged based on the ability to visualise the images relying on the captions alone. These automatically generated captions were also judged according to their relevance to the associated image.

In general, automatic and human evaluations have several strengths and weaknesses (Celikyilmaz et al., 2021). Implementing automatic evaluations is done by simply applying generated text to the evaluation metrics, which requires less time and effort compared to conducting human evaluations. In human evaluation, on the other hand, a number of tasks need to be considered. These tasks include designing a suitable evaluation study, considering ethical requirements, and obtaining ethical clearance. Furthermore, conducting human evaluation necessitates communicating and recruiting potential evaluators. In addition, unlike automatic evaluation, human evaluation can be implemented on a limited scale of data, as a large scale necessitates more resources such as time, effort, and a higher amount of remuneration to compensate the evaluators for their effort. On the other hand, results of human evaluation are more reliable compared to automatic evaluation (van der Lee et al., 2021). This is because, in human evaluations, end users assess the quality of the generated text in achieving its goal in several areas, including marketing and informativeness. While, in

automatic evaluations, the generated text is assessed based on fixed examples, GT captions, which make the quality of the evaluation solely rely on the quality of the used GT captions.

2.9 Gap in the Literature

Prior research has looked into the challenges VI individuals face when purchasing clothing, while other studies have put forward models for extracting meaningful textual information from images of clothing. We found significant limitations in the existing literature through our synthesis. Here, we address these gaps.

- Previous studies that have explored the challenges of clothes shopping for the VI people mainly relied on conducting interviews. However, observing VI people conducting shopping tasks provides objective and in-context evidence and demonstrates the challenges they face in a way that interviews cannot. Conducting an observational study also reminds the participants of challenges they might not have remembered during the interviews, thereby allowing us to mitigate recall bias and imperfect memory. Observing users also produces rich data that can be used to generate design implications for creating new technology that meets the intended users' needs (Rapp, 2021).

- To date, there is also a scarcity of research on how automated systems can be utilised to improve VI people's experiences. Interesting factors to consider include VI people's trust in automated systems, the kinds of automated services they would wish to have, and their fashion influences when purchasing or coordinating outfits. Research could also explore the offline environment to determine what services help VI people coordinate outfits in offline stores that are not offered so far in online stores.

- Existing clothing image–text datasets have many limitations that have not been explored in previous work. Specifically considering VI people's needs to mentally visualise clothing in the image-captioning domain (Alluqmani et al., 2023). From there, exploring clothing image–text dataset limitations from VI people's point of view is essential.

- Despite the greater success of the ROI-based approach in generating dense descriptions (Li et al., 2019; Long et al., 2023; Yin et al., 2019) and rich information on objects and their attributes (Anderson et al., 2018; Cornia et al., 2020; Kinghorn et al., 2018; Luo et al., 2021; Shi et al., 2020) in the context of general-image captioning, research in clothing domain has solely relied on the grid-based approach. However, previous research mainly focused on generating abstract descriptions of clothing elements (Cai et al., 2022; Hacheme & Sayouti, 2021; Han, Zhu, et al., 2023; Moratelli et al., 2023a; Yang et al., 2020), such as clothing categories (e.g., tops, skirts, and coats) and attributes (e.g., the existence of collars, patterns and graphics). No prior work has tackled the extraction of detailed descriptions of fine-grained clothing elements such as graphics. The unexplored area worth addressing is the adoption of the ROI-based approach to generate detailed descriptions of clothing graphic elements.

- Although Tateno et al. (2020) made an important initial contribution by addressing clothing image captioning for VI users, their research adopted a standard en-

coder–decoder image-captioning framework, which fails to generate sufficient descriptions (Ming et al., 2022; Stefanini et al., 2021). Additionally, this model fails to provide detailed information on fine-grained clothing elements such as graphics. Thus, developing a VI-friendly clothing-captioning model adopting recent technology is essential. Specifically, the advanced knowledge of LLM and its ability to understand humans could be leveraged to generate VI-friendly captions.

2.10 Conclusion

With the aim of providing a comprehensive understanding of the challenges faced by VI people when attempting to purchase clothing online, we conducted two complementary studies using both observational and interview approaches (Chapter 3). In Chapter 4, we propose a new VI-friendly clothing captioning model.

Chapter 3

The Barriers to Online Clothing Websites for Visually Impaired People

3.1 Introduction

VI people cite shopping for clothes as one of the most challenging everyday tasks they need to perform (Lee et al., 2020). Advances in technology have removed many of the physical challenges inherent in offline shopping (Williams, 2021). Thus, most VI people now engage in online shopping (Stangl et al., 2020a).

However, shopping for clothes online introduces barriers for VI people, including accessibility issues, an overreliance on images, and scarcity of accurate clothing information (Stangl et al., 2020a; Wang et al., 2021). As such, many VI online shoppers have to rely on sighted assistance from their friends and family. Unfortunately, most retailers rely heavily on product images to provide detailed information, but VI people need alternative text (alt text) to understand image content. In most cases, however, alt text is empty, lacks details or simply redirects to image paths (Wang et al., 2021). Many online stores also do not identify graphics (e.g., cartoon characters) on clothing products (Williams, 2021) or simply identify the characters, without describing what they are doing, assuming that this is already obvious in the images.

Previous studies that addressed VI people's challenges in purchasing clothing online (reviewed in Section 2.3.3) relied primarily on interviews, which do not offer a holistic understanding of the problem. Conducting observations in addition to interviews allowed us to obtain objective information about the specific issues encountered and the participants' subjective, holistic online shopping experience. Observing the participants conduct specific online shopping tasks provided us with in-context evidence of the limitations of retailers' websites and the opportunity to recall other challenges. Moreover, this approach produced rich data that could be used to generate design implications for creating new technology that meets the intended users' needs (Rapp, 2021).

Furthermore, previous studies have failed to explore VI people's challenges when attempting to coordinate outfits online and the services that can help them do so, which are currently not offered in online stores. In particular, no study has investigated how automated systems

can help VI people select and coordinate outfits online and understand the services they are likely to trust automated systems to perform. The factors that influence VI people when selecting clothing, such as celebrity or fashion style, which will provide insight into how to help provide better support for VI people, have also not been explored by previous studies.

To fill these research gaps, both observational studies and interviews were conducted to investigate the clothing selection and coordination challenges that VI people face when shopping for clothes online. Additionally, their wishes regarding how automated systems could be designed to improve their experiences were also explored. Specifically, this chapter serves to answer three of our research questions: *RQ1: What challenges do VI people face when independently selecting and coordinating outfits while online shopping? RQ2: To what extent do VI people consider a specific celebrity or fashion style when selecting outfits?*and *RQ3: What are the current VI-assistance systems, and what types of assistance do VI people like to have when purchasing clothing online?*

We break the third research question down into the following sub-questions:

- RQ 3.1: What services are offered to assist VI people in coordinating outfits in offline stores are not yet offered in online stores?
- RQ 3.2: What types of assistance would VI people like to have when coordinating outfits while shopping online?
- RQ 3.3: What type of assistance do VI people regularly make use of when shopping for clothes online?

We present the methodology we designed to answer our research questions and the procedure of our investigations in the 3.2 section, along with our insights into designing a study with VI individuals. Next, in the 3.3 section, we discuss the methods used to analyse our data, and following that, we report the outcomes in the 3.4 section. Finally, in the 3.5, we end this chapter by discussing those findings and suggesting potential design implications for online clothing stores.

3.2 Method

This chapter aims to understand VI people's behaviours and challenges while independently shopping for clothes online, including coordinating items. It also seeks to explore their influence on clothing preferences, the type of automated assistance they desire, and the level of trust they might place in such a system.

To achieve these aims and address our research questions (i.e., RQ1, RQ2 and RQ3), we conducted two complementary studies. We employed task observations and standalone interviews to leverage the advantages both approaches and overcome some of the limitations inherent in employing each on its own. Task observation is effective and widely used in the Human Computer Interaction (HCI) field for exploring potential design implications related to the development of a novel technology that meets the needs of its intended users (Rapp, 2021), while standalone semi-structured interviews are suitable for gaining a deep understanding of the intended users' views. Furthermore, the open-ended nature of the interview questions permits flexibility in the answers, allows unexpected responses to be elicited and

recorded, and permits interviewees to discuss their perceptions of the task(s) (Alshenqeeti, 2014).

In Study 1, task observation was conducted to understand the participants' in-context behaviours and identify barriers when performing specific shopping tasks. This approach allows the participants to recall and share examples of important details. The tasks were framed in the context of a pre-defined simulated situation (similar to Borlund (2000)'s simulated work tasks) to improve both internal and external validity by minimising confounding variables.

Participants were asked to imagine that they are going to a party and need to buy a new outfit to wear to it, and to select one top and one bottom item of clothing, along with a third compatible item. During the task observations, participants completed multiple tasks, including sharing their screens, accessing the Amazon website (logged out of their accounts), and selecting and coordinating three clothing items. This retailer was chosen due to its popularity, worldwide ubiquity and inclusion of various product-related information, such as product descriptions, images and reviews. Moreover, the implementation of the pre-planned tasks was restricted to a single retailer to minimise confounding or extraneous variables, thus simplifying the analysis. After completing the tasks, we conducted short follow-up interviews to ask questions about their experiences purchasing clothes online and the tasks they completed during the observation.

While Study 1 focused on obtaining an understanding of VI people's behaviours when purchasing clothes online, Study 2 gathered narrative data about their perspectives and needs. In Study 2, the participants were asked questions via standalone semi-structured interviews about their outfit selections, coordination processes, requirements and needs.

3.2.1 Participant Recruitment

A total of 17 participants took part in this research (see Table 3.1 for demographic information), including seven who participated in both studies. Overall, 10 participants took part in both parts of Study 1 (i.e., task observation and follow-up interview), and 14 were interviewed for Study 2.

Participant recruitment was performed by our contacts at several UK charities, including Sight Support West, Birmingham Royal Institution for the Blind and Sheffield Royal Society for the Blind, who advertised the studies on their social media accounts and in their online groups. Interested people contacted us via email, and we confirmed whether they met the criteria (explained later in this section). If the potential participant confirmed that he/she met the criteria, then we accommodated his/her preference to take part in one or both studies. Those who were willing to participate in both studies were able to choose which study to complete first based on their availability and preferences (some participants found it challenging to perform online shopping). We continued the recruitment process until we reached data saturation.

We applied the same three recruitment criteria for both studies: aged over 18, the absence of other disabilities aside from visual impairment and blindness (to exclude other challenges that may have affected purchasing behaviour, such as hearing or physical impairment), and prior experience with shopping online (at least once). Before conducting the study, all potential participants were emailed an information sheet and a consent form (Appendix A). To avoid any accessibility issues when providing consent, the participants had the option of verbally providing their consent before the interview (an audio recording was made to main-

tain integrity) or filling in the consent forms and returning them by email. All participants received a £10 Amazon voucher for participating in each study, resulting in remuneration at or above the UK living wage [1].

Table 3.1: *Participant demographic information.*

Perception level	Visual Experience	ID	Age Group	Gender	Type of assistance system	Stud[y/ies]
Blind	Acquired	1C	25-34	Male	Voice assistance(Siri) & screen readers	1 & 2
Blind	Acquired	1N	25-34	Male	Screen readers &Voice assistance	2
Blind	Acquired	2D	35-44	Female	Screen readers (Voiceover)	1 & 2
Blind	Acquired	2F	45-54	Male	Voice assistance	2
Blind	Congenital	1F	18-24	Female	Screen readers	2
Blind	Congenital	1D	35-44	Female	Screen readers	1 & 2
Blind	Congenital	1L	45-54	Female	Screen readers	2
Blind	Congenital	1S	35-44	Female	Magnification tool	1 & 2
Blind	Congenital	2E	45-54	Male	Screen readers	2
Blind	Congenital	2Q	45-54	Female	Screen readers	2
Blind	Congenital	2H	45-54	Male	Screen readers (Voiceover)	2
Blind	Congenital	1E	35-44	Female	Screen readers (Voice-over)	1
Blind	Congenital	1A	55-64	Male	magnification tool	1
Blind	Congenital	1B	45-54	Female	magnification tool	1
Low Vision	Congenital	1G	55-64	Female	Magnification tool	1 & 2
Low Vision	Congenital	1Q	45-54	Female	colour contrast settings	1 & 2
Low Vision	Congenital	1M	75 or more	Female	Large cursor setting	1 & 2

3.2.2 Research Procedures

As Figure 3.1 shows, we conducted two complementary studies. Both studies were co-designed in collaboration with representatives from the Sheffield Royal Society for the Blind (SRSB), a local society for VI people. We started our collaboration with the SRSB by discussing VI people's challenges and their experiences of communicating and working with VI people. Then, we discussed the initial proposals for study designs and explored whether they were suitable. For instance, were the tasks achievable by individuals with visual impairments, and would they require any support to complete the tasks? Alongside communicating with the charity, Our team (i.e., three in total) also validated the study designs in relation to our aims and research questions. This process was implemented iteratively. The study was approved by the University of Sheffield Research Ethics Committee (see the ethics document in Appendix B).

The interview design was inspired by the interview protocol refinement framework (Castillo-Montoya, 2016), whereby the interview is implemented in four steps. In the first step, the validity of the interview questions is tested by aligning them with the research questions. The researcher then, in the second step, set protocols for conducting the interviews as an inquiry-based conversation. As part of this, the questions must adhere to spoken language conventions and everyday practices so they are simple to understand and respond to. It is also essential to follow the rules of social conversation when designing these protocols. Examples of such social rules are asking one question at a time and using introductory questions, transitions between questions, and closing questions. Preparing a draft interview script can assist

[1] See http://livingwage.org.uk for more information

with developing such protocols and thereby smoothing the interview conversation. An interview script provides written information that directs the interviewee during the interview; it includes information the participant should know and guides them on where the conversation should be heading. In study 1, the interview questions are designed to address RQs 1 and 2 as well as questions related to the observation process, while the interview questions in study 2 address RQs 1, 2 and 3(see interview scripts in the Appendix C). After building the interview questions and protocols in these first two steps, in the third step, the interview protocols are then reviewed, with the team members providing feedback and suggestions on how to enhance reliability. In the final step, the interview is piloted on a small sample of participants (two participants, 1A and 1B) to check the functionality of the developed interview protocols, and appropriate amendments are made to resolve any issues identified.

The testing criteria in this pilot study were inspired by the criteria highlighted by Dikko (2016) and Hennink et al. (2020). These criteria are described as follows:

- Identifying any difficulty in implementing the proposed tasks in the task observation method
- Identifying any ambiguity in the interview questions
- Determining the expected time for completing each of the observations, subsequent interviews, and the separate standalone interviews.
- Checking whether the proposed tasks and questions cover all aspects of the research questions
- Testing the task observation and the interview guides.

After applying this framework and thereby confirming the quality of both study designs, we then conducted the studies without a strict order (e.g., we did not wait to complete study 1 before beginning study 2) to accommodate the preferences of the participants. Considering participant preferences is crucial for ensuring flexibility and comfort (Carter et al., 2021). Each participant was randomly assigned an ID to protect their identity. The task observations and the follow-up interviews in Study 1 lasted around 50 minutes in total - approximately 30 minutes for observation and 20 minutes for the follow-up interview - and the interviews in Study 2 lasted around 30 minutes. Due to Covid-19, both studies were conducted remotely using video conferencing applications. The task observations and the follow-up interviews in Study 1 lasted around 50 minutes in total - approximately 30 minutes for observation and 20 minutes for the follow-up interview - and the interviews in Study 2 lasted around 30 minutes. We audio-recorded the interviews in both studies, and video recorded (by capturing only the screen) the observation tasks to ensure that all relevant data was gathered (Jamshed, 2014). Throughout the task observations, we made notes in relation to our research questions and later used the recordings to make further notes.

The interview data were analysed based on Braun and Clarke (2006)'s recommendations, as discussed in 3.3. To become familiar with the interview data, we transcribed the recordings and read the transcripts multiple times. We then reviewed the transcripts and the recordings to ensure accuracy of the data. When transcribing the data, we took care to remove all identifiable data (i.e., the names of the participants and other names mentioned during the interviews). We also assigned IDs to each participant to ensure protection of their privacy

and to prepare the data for the coding stage. When analysing the data, we adopted a thematic approach combining inductive (codes derived from the content) and deductive (codes derived from our research questions) methods. In the early phase of coding the data, several interesting codes were identified, including how VI people communicate and conduct online shopping with sighted assistance, how COVID-19 impacts their shopping behaviour, and how they communicate and discuss clothing descriptions with sighted assistance. In this study, we focused on four main topics and their impacts on shopping among people with visual impairment: the use of assistance technology, the influence of cloth purchasing, the limitations in online shopping and the shopping mall environments

We reviewed and discussed the codes with the team members multiple times. After discussing the initial codes (more than 30) a number of times, we clustered related codes into a more manageable number and combined similar ones into one theme. In addition, we reviewed the codes considering the research questions of this study.

Analysis of the data from completed studies was done iteratively, with the new data from each additional completed study being incorporated into the analysis, allowing us to assess when saturation had been reached (DiCicco-Bloom & Crabtree, 2006). The coding and analysis of data from the two studies were performed separately, and then the findings of both studies were reported together as they were complementary.

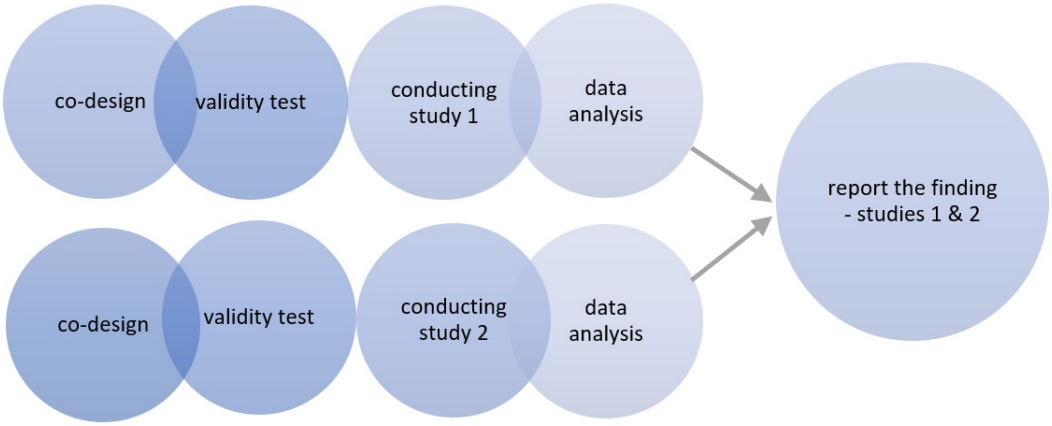


Figure 3.1: *Overview flow diagram of the research procedures: co-design with the Sheffield Royal Society for the Blind (SRSB), validity test, pilot studies, main studies, data analysis, and reporting the results. The research steps were implemented independently, but in the last step, the findings of both studies were merged as they were complementary.*

3.2.3 Study Involving Participants with Visual Impairments

It is important to point out that conducting studies with people who have visual impairments presents a number of additional challenges and requires the researcher to be more attentive to participants' well-being during the studies. We faced additional obstacles as we conducted our interviews and observational tasks using online video conferencing applications including Zoom and Google Meet. Some applications suffered from poor accessibility, and some VI people expressed discomfort because they were unfamiliar with them. Thus, providing familiar applications or alternative platforms, such as phone calls for the interviews, was essential.

Conducting the observational tasks online with VI people was challenging. They not only had to deal with the accessibility limitations of being online but also had to perform other tasks, such as sharing the screen. This required them to interact with several dialogs, which were mostly inaccessible to their screen readers. Some potential participants wanted to take part but were reluctant to perform the online shopping tasks, not only because of accessibility issues but also because they did not want to perform them in front of others to avoid being judged for their behaviour. All VI individuals involved in this research's observational study have successfully completed the pre-planned tasks.

It was also vitally important to ensure the pre-planned tasks were realistic for and achievable by the population. We did this by co-designing the study with people who have direct contact with the research population (i.e., representatives from the SRSB).

3.3 Data Analysis

Our analysis process was inspired by that used by Braun and Clarke (2006). However, we did not follow the suggested order of the analysis process exactly. In our analysis, there was some overlap between the steps of the iterative process (i.e., 'Generate initial codes', 'Search for themes', 'Review themes' and 'Define themes'). We began data familiarisation by transcribing the interviews using a partially-automated technique (i.e., Trint), repeatedly checking the transcripts and field notes against the audio and video recordings. We read and revised the collected data for clarity, and we annotated and revised the field notes while watching the video recordings of the participants' shopping tasks.

The early steps in the task observations followed Eriksson and Kovalainen (2015)'s suggested analysis method, including asking ourselves what the field notes indicate and what interesting and unique things are related to our research questions. However, while conducting subsequent observational studies and reading more into the data, we began to notice interesting behavioural patterns, including the process followed when searching for desired items and the sources of the descriptions relied upon (e.g., product titles, descriptions, customer reviews). We also identified the behaviours that occurred when participants encountered challenges that hindered their ability to make purchasing choices, and we noted the issues experienced with the tools they relied upon when conducting shopping activities, such as the limitations of the Amazon search engine from the perspective of VI users.

Then, with the qualitative data analysis software NVivo, we used an open coding techniques to develop and modify the codes as we proceeded through the coding process. During the analysis process, we derived codes from the data mainly by following an inductive approach. However, we also shifted to a deductive approach by considering the focus of this project on topics related to limitations in online shopping, VI people fashion influence, shopping mall environments, and the use of assistance technology in online shopping. To ensure reliability, the three team members discussed the codes and themes several times, iteratively revising the codes and checking them with each new transcript to see if new ones should be generated. We also reviewed the codes to ensure they were related to the research questions (Braun & Clarke, 2012).

The standalone interviews, Study 2, were analysed separately from the task observations, but the deeper the analysis went, the more we found strong connections between the codes generated from each. We therefore clustered the codes in each study and ended up with

the same four themes. We report the final themes and the findings of both studies as a complementary story in the following section.



Figure 3.2: *Diagram of themes identified from Studies 1 and 2.*

3.4 Findings

Four key themes emerged from our analysis of both studies (see figure 3.2); a total of 59 individual codes resulted from Study 2, and 34 codes from Study 1. Although the semi-structured interviews in Study 2 generated in-depth knowledge, the observational studies uncovered several aspects that did not emerge from the interviews. For instance, it revealed the inconsistency between how the participants represented their process of purchasing clothing online and their actual online shopping behaviour. Examples of observed behaviour include the prominent use of the search tool in the purchasing process and the trust in customer reviews over retailer descriptions when making a purchase choice. In addition, conducting purchasing tasks in Study 1 helped the participants bring evidence of the challenges they mentioned in the interviews and recall new ones, such as the restrictions of the search tool and the usage of conflicting terms when describing clothing items.

In this section, we discuss each theme, cite illustrative participant quotes, and refer to incidents noted during the task observations. It is worth noting that in discussing our findings we did not report the results according to the different categories of perception (such as low vision or blind) or visual experience (such as congenital or acquired). Because our discussions with participants revealed that being blind (registered blind) does not necessarily mean that one is completely without sight, we reported the findings of both categories to represent VI people in general. Some participants, such as 1A and 2D, were registered blind but had some sight, while participant 1B had no vision in the right eye and only low vision in the left eye. This indicates that not all blind people use screen readers to access the internet; some participants (including 1S, 1A and 1B) used the magnification tool instead. These observations, together with the results of our observations of the participants' behaviours in

selecting and coordinating clothing online, reveal only one notable difference between the two categories: those with low vision examined the product image and verified the content of the image by reading the descriptions and review data, while those with complete loss of sight gleaned their knowledge of the product solely from the textual data.

3.4.1 Limitations on Clothing Websites

Barriers to the Selection and Coordination of Outfits Online

In response to RQ1 (*What challenges do VI people face when independently selecting or coordinating outfits while online shopping?*), participants outlined many challenges that render their ability to select and coordinate clothing items online more difficult, including the restrictions of clothing descriptions and the unreliability of the current descriptions and tools.

Several limitations of clothing descriptions were brought up in relation to online shopping. Participants identified the lack of style descriptions as one of the most frequent constraints. They sought details of clothing items, including details about sleeves, pockets, buttons and zips. In the case of an item having buttons, for instance, VI people wanted to know,"How are the buttons? Are they press stud? Are they crest buttons? [1G, Study2]". Such information is almost never provided in the descriptions of items online.

The lack of adequate graphic descriptions was also identified in Study 2 and noted during the observational tasks. Participants indicated that retailers rarely provide descriptions of graphics and designs printed on clothes that would enable customers with VI to mentally imagine how it looks. Participant 1C referenced a T-shirt item he chose during the observation tasks, which was described as 'Take this with you on your summer holiday', without any other details describing the graphic on the shirt of a camper van. Commenting on this, he stated that this description was insufficient for a person with VI to predict the graphic. Instead, he suggested a better description, such as "a picture of the camper van in a green field with whatever cloudless sky [1C, Study2]". To obtain such a description, a VI shopper would need to rely on a sighted friend or relative.

How a piece of clothing with a graphic looks is about more than just what is represented by the graphic. For example, as reported by participant 1C, how the graphic on a clothing item feels (e.g., soft print) and the position of the graphic is also important. Participant 1S, who is congenitally blind, provided an example of how difficult it is to figure out the position of a graphic: "It might say a heart-printed t-shirt, but sometimes the heart can be quite abstract because you have to hunt to find out where it is on the top and especially if it's a similar colour as the rest of the t-shirt [1S, Study2]". Participant 1L also provided the case of an item with a tie-dye design. She simply wanted to know whether the design started on the right or left side of the item.

In addition to the lack of style and graphic information, colour descriptions were also highlighted by the participants as being either missing or described using ambiguous terms. For example, the use of terms like 'deep blue', 'sapphire', 'champagne' or 'charcoal', or uniformly described, such as grey, blue or pink, which do not provide a clear indication of the colour shade. Sighted people familiar with such terms find the uniformly described colours adequate, as they can obtain information on colour shade, for example, from images. However, VI people find it challenging to identify colour shades. Therefore, they require a new VI-friendly technique to assist them in better understanding colour shades. VI people

are interested in not only the colours or colour shades of an item, but also the position and extent of each color. Talking about this issue, 1L said: "Quite often, it will just say blue and white, but it wouldn't tell you [if] it's blue and white stripes or blue with white flowers or, you know, blue and white spots [...] I don't know which one is most... which one has the most colour [2Q, Study2]".

Unclear sizing information was another issue identified with clothing descriptions. This includes information such as the item's overall length, the length of its sleeves, or how loose or tight it is. Interviewees wanted to know the size of the other elements on the clothing items, such as pockets and buttons. One participant, 1L, described how challenging it is to identify size when retailers use the term 'one size', as this does not provide any information about which size group the item is for.

Another reported problem was the unreliability of the meta-information about the clothing and, subsequently, of the Amazon search tool, which uses this data to filter items. When discussing existing online descriptions, the participants used terms such as *deceptive* and *misleading*. Participant 1L shared her experience of encountering inaccuracy in the descriptions of pockets, which are her essential clothing item elements. She prefers pockets with fastenings because they give her more security when she takes her phone, money and keys outside. She reported her experience of ordering yoga fleece bottoms: "It said zip fastenings for the pockets. When I got it, it didn't have a zip fastening! It just had standard pockets that were on the side! [1L, Study2]". 1L also shared details about a time she bought a dress online. Although she received a dress with the same basic design as the one described on the website, it was in the wrong colours: "It was a black dress with red flowers on it, and the description was a white dress with black flowers [1L, Study2]".

In another related issue, participants also noted misclassification of items. For example, sometimes clothing was labelled as casual when it should be formal, and vice versa. The observational task revealed such a case of contradictory information. Participant 1D selected a blouse described as 'Chiffon Blouse Tops Elegant Casual', thinking it was more like party wear. However, when she checked the description, it said 'office wear' (see figure 3.3), so she decided to try a different item. She commented on this experience, stating that "if I went into every single [item's descriptions], I'd be here all day and all night [1D, Study1]". This is another clear example of an issue that is considerably less problematic for sighted individuals, who can either identify the misclassification error from the images, or can quickly skim the description. This is considerably more frustrating and time-consuming for shoppers who have to rely on a screen reader.

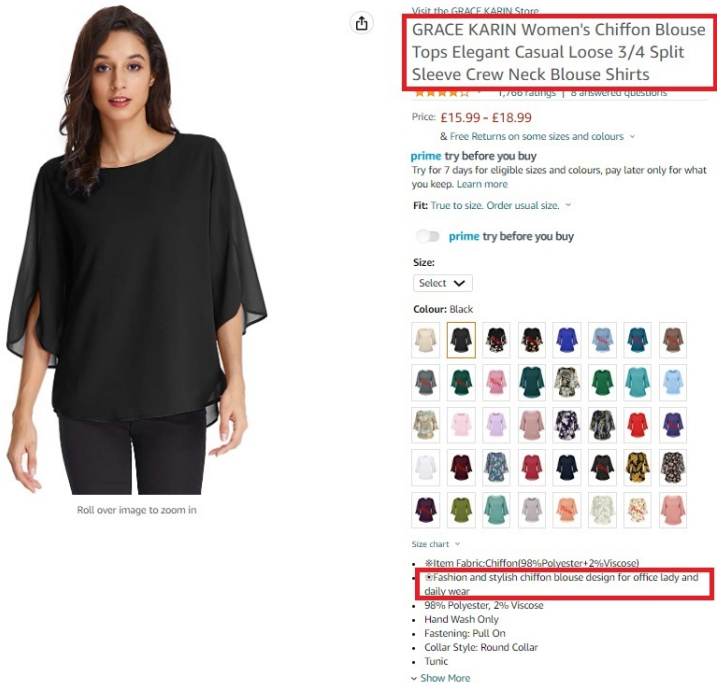


Figure 3.3: *An example of using contradictory terms (i.e., casual and office wear) when describing clothing items online.*

The Amazon search tool was also reported as being frequently unreliable. We observed that VI people rely mainly on the search feature on Amazon to look for their desired items. Although this search capability saves time and effort when filtering multiple items, the participants found it challenging to review the resulting items because they were very likely to receive extensive results, including entirely unrelated items. For example, participant 1M typed "ladies' smart shoes" to look for a third item compatible with her previous selections. She noted that the search engine suggested items related to the word 'shoes' and not the entire search query. This is another instance where the time cost of such imprecision for sighted users would be minimal but for those who have to rely on screen readers it is considerable.

Ways to Cope with the Online Restrictions

Participants highlighted several approaches they take to cope with the current limitations. One such technique is developing familiarity with particular websites to reduce the possibility of making mistakes. One of the interviewees commented on this technique: "If I go to a new shop I'm not familiar with, then it can become sort of really finding my way around and not knowing what to get [1S, Study2]". The obvious downsides to this is a reluctance to branch out from familiar sites and brands, and a need to refamiliarise oneself if the website changes its design.

Another technique that appeared in both the interviews and observational tasks was that the participants preferred to stick to dark colours, mainly black, which they called "*safe colours*". The term arose from the subjects' belief that dark colours match all other colours. In addition, to minimise the problems of matching outfits, most participants preferred to

select plain items (i.e., with no patterns or graphics). One of the respondents commented: "I tend to always wear the same colour: top and bottom. So it's safe, I suppose. I would say I'm a quite safe shopper [2D, Study2]".

The observational tasks also revealed that most of the subjects confirmed their choices by reading the reviews. Although this might also be true for sighted people, we noted that the VI people often did this to fill in the missing information about the selected product and validate the 'official' descriptions. When the participants were asked about the type of information they searched for when examining review data, they mainly mentioned sizing, length measurements and quality, and seeking other customers' general opinions. They trusted customer comments more than the descriptions provided by retailers. This was observed when participant 2D relied on customer descriptions, despite the fact that their descriptions directly contradicted those of the retailer. She selected a top that was described as a casual blouse and a cardigan described as a beach cardigan. However, in the reviews, she found that the items were considered more formal. Another participant, 1S, who uses magnification tools when browsing the internet, finds reading review data overwhelming; she described her method of scanning the review data: "I tend to [...] look at the latest ones [reviews] first and try and get a feel. Yeah. Because from that, you can generally normally get a feel for the item. And then I would move this across to see the different star ratings. And then quite often I'd look at the very best and the very worst might I get a good idea of what the issues are [1S, Study1]".

3.4.2 Fashion Influences

In response to RQ2 (*To what extent do VI people consider a specific celebrity or fashion style when selecting outfits?*), most of our participants stated that fashionistas, influencers or fashion styles do not typically influence their clothing choices. Instead, participants had other concerns, including how comfortable the item is or whether it includes certain practical elements, such as pockets. Some participants mentioned following specific brands, colour trends, or refer to their body shapes or to other people around them and what they wear.

The results also revealed that the lack of visual ability and fashion resources most likely affect the participants' fashion influences. This was exemplified by participant 1N, who has an acquired sight problem. He expressed how following fashion styles became less attractive when he became blind: "When I was sighted, I was wondering quite a lot, but not now, [...] Now it's kind of a regular thing [1N, Study2]". Another participant ID, argued that fashion sources such as magazines and TV shows do not really accommodate VI users. She commented: "A lot of the fashion information is about photos, and there's just not enough description for me to want to take an interest [1D, Study2]".

On the other hand, some participants have their own influences when it comes to fashion and clothing selection. Participant 1M likes to follow trends in colours rather than styles:"I probably follow the colours that [...] may be prevalent this year [1M, Study2]". In another example, participant 1Q, who is partially sighted, loves fashion and follows fashionistas. However, she prefers to shop based on her body shape, rather than following specific fashion trends.

Other participants stated that the main factors when selecting clothing items are comfort and ease of access. For example, participant 1F, a young blind participant, stated that for her outfits are functional rather than aesthetic. When asked about following fashionistas or

fashion styles, she laughed and said: "I don't tend to look for things to try to look better or to try to impress other people with what I look like or try to copy a [...] celebrity style [1F, Study2]".

3.4.3 Shopping Malls

With regard to RQ3.1 (*What services are offered to assist VI people in coordinating outfits in offline stores that are not yet offered in online stores?*) and RQ3.2 (*What types of assistance would VI people like to have when coordinating outfits while shopping online?*), the participants stated how offline shopping services impacted their ability to coordinate clothing items when compared to online shopping. They also discussed how automated assistance systems could improve the experience for VI shoppers.

Our findings indicate that the tactile and physical properties of offline stores, including the opportunity to feel and try on items, provide the ability for VI people to better coordinate clothing items when compared to purchasing items online. In online shopping, however, it was noted that coordinating items is done mainly "by coordinating in multiple windows [1S, Study2]", which can be hard on the eyes for people with limited vision as it requires focusing in and out multiple times. Moreover, this process is time-consuming. To avoid this, it was observed that participants mainly chose simple styles with dark colours or picked items based on colour coordination.

The participants also discussed how automated systems may be used to improve their experience in online shopping; below, we discuss these findings.

Detailed descriptive information The participants asked about the necessary requirements to improve online shopping and mentioned the need for better quality descriptive information, suggesting that clothing items should be described in greater detail. Participant 1D expressed her expectation for the availability of descriptions online: "if you pick a dress [...], it will say, What colour is that? [...] what's the pattern like? [...] Is it polka dots? or is it just squares, zigzags? [1D, Study2]".

While offering detailed textual information could be beneficial for avid users of screen readers, it could be tiring for others. To combat this, participant 2F suggested offering an audio button for verbal descriptions. Participant 2D felt that having clothing measurements in textual form, rather than in tables would help them to shop.

Another issue raised by the participants was enhanced colour descriptions. They suggested standardising colour palettes and associating them with codes or putting them on a consistent numerical scale. Participant 1M described how colour information can be presented: "If there are 45 blues for example and it would have a code that would tell you that it's blue which is this 'X241' and that's your colour code. [1M, Study2]".

Clothing and outfit recommendation systems Another important service requested by the participants was recommender systems. The majority of the participants sought a recommendation system that offered item coordination, rather than just suggesting similar ones. Participant 2F said: "If there was a way of being able to almost put two images together and the computer's telling you whether they matched or not, [...] Are they the same colour or are they matching colours or matching styles. If there was a way that the computer could tell you about that [it] would be fantastic [2F, Study2]".

Another participant, 1Q, transitioned from partially sighted to blind. When she coordinates outfits online, she tries to balance the style, such as a loose top with a tight bottom or vice versa. She also coordinates items based on colour agreement. However, she finds this process tiring, so she suggested a system that might take three of her selected items and put them together so she could see whether they are well coordinated.

Others suggested a system that alerts users to poor matches, rather than helping them to choose coordinating items. Such a system should alert customers when selected items do not match or are not appropriate for a specific occasion. Participant 2F expressed how these recommendations might be helpful: "It could tell [the automated system] that light green and bright orange don't really go together [... and that] having stripes one way and stripes the other way [is] not a good idea [2F, Study2]". This recommendation approach was generally preferred by the participants over recommending items that go together: "if it was telling you what to put together, it's a bit more restrictive [2D, Study2]".

In addition to these suggested recommendation systems, a system that narrows down search results based on previous purchases was also suggested: "Perhaps if it keeps track of what you've already purchased and then picks a similar style clothes that cuts out things, [...] you might not be interested in. Um, because, you know, there are several things I never wear.[...], if that would be removed from the search engine [1S, Study2]". An important consideration here would be whether to consider short-term items (e.g., those currently in the user's basket) or longer-term items, like those that the user had bought in recent weeks or months.

What should recommender systems know? The participants shared particular suggestions on how recommendation systems could be made more effective for their needs. They stated that it would be useful if the system recommended outfits based on customers' colour preferences and body shapes. Participant 1M, for example, suggested that the system should recommend items based on the user's colour season, such as whether they are an autumn, summer, winter or spring colour. For instance, the system could clarify that the colour white is not suitable for autumn colour. Another participant, 1Q, suggested a system that recommends items based on skin colour, as some colours are more complimentary toward certain skin tones than others. Participant 2D, on the other hand, suggested a system that recommends items based on the user's body shape, such as whether they have a pear or apple-shaped body. This is another good example of something that a sighted person might be able to determine from product images but which would be essentially hidden from an IV user.

3.4.4 Assistance Technology

With regard to RQ3.3 (*What type of assistance do VI people regularly make use of when shopping for clothes online?*), participants reported that they use screen readers, automated voice assistants (e.g., Alexa and Siri), and colour contrast and magnification tools to access the Internet. When asked about their trust in automated assistance, they stated that they are hesitant to trust such systems, instead placing more trust in sighted people (e.g., friends or relatives) when purchasing clothing. They believe that trust needs to be established based on experience. They are happy to use new assistance systems, but they do not trust them when selecting expensive clothing or clothing meant for special occasions. Participant 1C,

who is an early adopter of new technology and a technophile, commented: "I would certainly buy an average day-to-day outfit, summer clothes holiday clothes. [...] I would spend average prices [...], £23 maximum each for an item [...] using the automated system to begin with [1C, Study2]". This implies an expectation that using the automated system would lead to poor choices the user might regret.

The participants explained their reasons for not trusting the automatic systems to help in their clothes purchasing process; the main issue raised was that most automatic systems do not know their customers' specific needs or style preferences. Participant 1C, for example, discussed the importance of building a system that knows its customers and provides services based on customers' preferences. 1C shared her thoughts about automatic systems: "They [automated systems] don't know your needs. They don't know your preferred type of clothes. So they're not really going to help me in terms of telling me [...] if this meets [...] my specific needs. If this is going to fit me [1F, Study2]".

Another related point that explains the participants' mistrust is that automated systems do not have the ability to determine the suitability of clothing items for different occasions, for example a beach party or formal dinner. Automated systems also do not have the ability to predict how a piece of clothing might fit a user. Participant 1F commented: "a visual person, especially like a family member that I live with every day knows how big I am. [...] they will know what size is likely to fit me [1F, Study2]".

Participants were also more confident relying on human assistance because sighted humans can describe garments by comparing them to other items already known to or purchased by the VI shopper. Participant 1D commented: "what helps for me is they're able to say: 'Oh, it looks like that top that you've got', so they can reference it something that I actually have at home. So it makes it more real [1D, Study2]". This might be possible in a recommender system by considering a user's purchase history and would be a form of explainable recommendation.

3.5 Discussion

Our findings demonstrate the challenges VI people face when shopping on clothing websites, what influences their clothing preferences, what types of automated assistance they seek and what level of trust they might place in such a system. In this section, we discuss our findings and their contributions to the existing literature.

The initial objective of this chapter was to identify the barriers to using online clothing websites for VI people. Our results indicate that the scarcity of clothing descriptions is a key challenge, which is consistent with the findings in the literature (Liu, Ding, et al., 2019; Stangl et al., 2018; Wang et al., 2021; Williams, 2021; Yang et al., 2015). Our findings reveal that applying the observational approach provides a holistic understanding of the issues under investigation. This approach allowed the participants to provide prompt examples of the issues they encountered when shopping for clothes online, leading to better discussions. Our findings also indicate that the observational tasks disclose challenges that were not reported during the interviews. For example, our observational study reveals several behavioural patterns, including validating the retailer's description with review data and using the search function as a primary tool for findings items online, thereby suggesting ideas on how to offer better solutions. It supports the idea that using task observation in concert with interviews

produces rich data and insights that may not be identified from one of these methods alone. Such insights can be used to inform design implications for creating new technology (Rapp, 2021).

Our observational studies uncovered a previously unreported issue — deceptive or incorrect product descriptions and the knock-on effect on the reliability of the Amazon search tool — because in most cases, it led to unrelated items. However, there is a risk of bias, as this study was conducted exclusively using the Amazon website. One interpretation of the limitations outlined in this observational study is that the Amazon website consists of many retailers from various countries, which might explain the use of unfamiliar terms. For example, the word 'blouse' is used to describe items that are referred to as 'shirts' in some countries. However, retailers may also attempt to purposely mislabel items to gain sales that they would otherwise miss.

To cope with the limitations of online clothing websites, the results show that VI people achieve confidence in their selection by confirming their choices with sighted, trusted individuals or by reading customer reviews. These findings are in line with previous research (Liu, Ding, et al., 2019; Stangl et al., 2018; Wang et al., 2021). However, in contrast with previous studies (Wang et al., 2021; Williams, 2021), the results from this research did not report the use of chat services as assistance tools (e.g., Aira or Seeing AI). The reason for the lack of uptake of these support options among the study population is unclear, but may relate to the fact that most of the participants were blind (only three participants had low vision). This research, however, revealed that, potentially as a result of missing out on such support options, the participants restricted their shopping strategies to minimise risk (e.g., by selecting only black or plain-coloured items or by only visiting familiar websites). One issue that emerged from these findings was the participants' impressions of restriction and inflexibility when selecting clothing items online.

Our findings indicate that our participants were not interested in following specific designs, fashion models or influencers; instead, they follow specific brands, colour trends or refer to their individual body shapes. This might help us understand the most essential aspects of designing a personalised system for VI people.

Our results confirm that VI people are generally hesitant to trust automated systems, which is in line with the literature (Wang et al., 2021). One significant finding explaining their trust in humans is that humans can reference items already owned or known to the VI shopper, which would be helpful in visualising an image of how the new item looks.

Despite this scepticism, the investigation of possible systems to assist in coordinating and selecting clothing online revealed that VI people would be interested in improved assistance systems, including image captioning and recommendation systems. Our results have significant implications for designing such an assistance system for VI people (details of the implications for design will be discussed in the following subsection).

The contradiction between the necessity for automated systems and the reluctance to trust them leads to an essential discussion of the types of assistance systems that VI people are most likely to trust. According to these results, together with the findings of previous work (Stangl et al., 2018; Wang et al., 2021), it seems that VI people would welcome new assistive technology. However, they put a high degree of trust in systems that emphasise the probability of error (MacLeod et al., 2017) or systems that are more personalised (Lee et al., 2022). It is unclear whether VI people would trust a system to recommend clothing items

based on objective features, such as colour, size, and pattern (Gharaei et al., 2021) or more aesthetic qualities, such as judging style (Yu et al., 2018).

3.5.1 Design Implications

Our findings have a number of important design implications for future practice.

Developing Systems to Automatically Produce More VI People–Friendly Captions for Clothing Images

The main constraint VI people face in enjoying independent online shopping is the absence of clothing descriptions. Providing detailed clothing descriptions will enhance knowledge of the available products, thereby strengthening the ability to build a more reliable assistance system. Although the lack of clothing descriptions has been reported in previous literature (Liu, Ding, et al., 2019; Stangl et al., 2018; Wang et al., 2021; Williams, 2021; Yang et al., 2015) and many studies have attempted to offer solutions, such as developing an automatic fashion captioning system (Tateno et al., 2020), our findings indicate that the scarcity of clothing descriptions remains valid.

Our findings also suggest that briefly describing products does not assist VI individuals in imagining how they look. It is therefore important to develop systems that can produce more VI-friendly captions tailored to the needs of VI people instead of extant approaches, which are tailored to sighted users. One possible design that could help develop more suitable captions for VI people is to specify crucial clothing characteristics for such users. This includes mentioning the presence of graphics, buttons, zips and pockets on clothing items as well as providing descriptive information about these features, such as their size and kind (e.g., an invisible zip, press-stud buttons or a patch pocket). Methods have been proposed that can control the length of generated captions (Deng et al., 2020), forcing the model to produce more detailed descriptions. Other work has emphasised the importance of providing users with captions that realistically present the level of uncertainty/inaccuracy in the automated descriptions (MacLeod et al., 2017).

Validate Clothing Product Descriptions

Our results show that contradictory or deceptive descriptions on online clothing websites make it challenging for the participants to select their desired items. The contradictions appeared in the descriptions themselves as well as between how the retailers described the product and how customers described the same product in their reviews. One solution may be to filter product information by validating the information in the product descriptions with the review data. This can be done by providing additional information regarding the accuracy level of the product information and giving the VI shoppers the chance to make a more informed decision.

Colour Identification and Scaling System

Our participants reported restricting their choices when purchasing clothing on their own to predominantly dark colours (mainly black) to reduce the chance that they will receive undesirable or incompatible colours. This is not necessarily the colour they would want to

choose but is simply a safe choice. Future research might develop a colour identification and scaling system (e.g., assigning 01 as a metric number to the darkest blue colour, 02 to the lighter blue colour and 10 to the lightest blue colour) to provide a more flexible shopping environment for VI people and broaden their shopping choices. Such a system could analyse product images to identify colours and then clearly describe the item's colours using a recognised colour scheme/palette (e.g., blue, red, and green), each associated with specific scaling numbers. Another potential solution might be for the system to outline the percentage of each colour in a piece of clothing. These suggested systems would permit VI people to imagine how dark or light the selected item would be, helping them to match items better. It would also allow them to determine how similar or different the selected item is from the items they already own and the extent of each colour in a clothing item.

Cut Out Unlikely-to-Purchase Items from Search Engine Results

As shown in our observational study, the wide range of item results from the Amazon search engine introduces further challenges for VI people. One design idea is to develop a system that tracks users' shopping behaviours to understand their preferences (similar to existing recommender systems). The user preference information can then be used to curate the search engine results by excluding items unlikely to be purchased and presenting only user-preferred items. The system could also personalise the results based on the user's fashion influences. This design suggestion may also be of benefit to sighted users.

Fashion Warning System

Existing limitations of clothes shopping websites make it challenging for VI people to select their desired items successfully. Our results show that VI people mainly stick to familiar websites and restrict their selections to plain, dark clothing to avoid purchasing incompatible or undesirable items. As an implication of these results, it would be a good idea to develop a system that works as a monitor, warning its users when they make undesirable selections. For example, the system might warn users if they match casual items with formal ones or attempt to coordinate tops with busy patterns with bottoms with a different and clashing pattern. Such a system could be controlled by understanding the users themselves better. For example, the system could obtain information about the users' skin colour tone, their favoured colour season (e.g., spring or autumn colour ranges) or their body shape (e.g., apple or pear shape). Users may choose to alert the system if they plan to buy more than one outfit. With this information, the system would be able to generate personalised warnings.

3.6 Conclusions

In this chapter, a complementary approach (observations and interviews) was used to obtain insights into the selection and coordination challenges of VI people shopping for clothes online. Their ideas for design changes to automated systems in order to improve their experiences were also examined, contributing to research on the design of VI-friendly online assistance systems.

The study design enabled us to directly observe behaviours and the challenges faced in context, rather than relying on potentially unreliable or biased recall, providing useful con-

text that prompted the participants to recall and share examples not raised during interviews alone. For instance, when using Amazon's search tool during the observational task, they encountered problems such as misleading titles and the results including unrelated items, which made it harder to efficiently find the desired item. Further challenges included contradictory descriptions or customer comments that were inconsistent with the descriptions (e.g., using both 'casual' and 'formal' to describe clothing items). Observing the participants as they conducted online shopping tasks demonstrated the difficulties that VI users face when searching for and assessing desired items; avoiding deceptive and unclear information; and coordinating outfits. Observing these challenges helped us identify VI users' actual needs, which is essential for developing useful design recommendations, and allowed the participants to suggest useful ideas for how an automated system might help.

The standalone interviews revealed several interesting fashion-related factors that may influence VI people's clothing selections, such as brands, body shapes and colour trends, instead of fashionistas or style trends. Another important finding presented in this chapter is the improved understanding of the types of automated systems that VI people seek in clothing shopping websites and are likely to trust.

The combination of interviews and observational studies generated novel findings that led to several design suggestions, including systems to automatically produce clothing image captions that are more friendly to VI people, validated clothing product descriptions, colour scaling systems and fashion warning systems. Our results show the necessity for further research to understand the challenges VI people face and the need for a better online shopping experience. Researchers can benefit from the challenges discussed in the present study to run future studies and develop more accessible and usable technological solutions.

In an initial attempt to mitigate some online clothes shopping limitations faced by VI people, in the following Chapter 4, we propose a new VI-friendly clothing graphic captioning model. Among other limitations, we chose this one to be our focus due to the high impact of missing descriptions on VI shoppers' satisfaction, and thus worked on generating detailed descriptions for clothing graphics.

The findings outlined in the current chapter influenced the design of our image captioning model in Chapter 4. The fact that VI people seek detailed descriptions of graphic elements led us to explore approaches to understanding images (i.e., grid and ROI) and guided our decision to adopt the approach that meets their needs (i.e., generates focused, detailed descriptions). In addition, the finding that VI people seek certain information about graphics, such as colours, positions and the size of the graphic, influenced our assessments of the proposed model. Specifically, this finding was used to assess whether the generated captions meet VI people's needs. The other findings from the current chapter, such as the factors that affect VI people's selection of clothing and the factors affecting their trust in automated systems, will inform the further enhancement of the proposed image captioning model.

The proposed VI-friendly clothing graphic captioning model will lift one aspect of the current restrictions on VI people wishing to shop for clothes online. For instance, if provided with detailed descriptions, VI people will no longer need to choose only plain items in order to avoid purchasing incompatible or undesirable items. With this and other positive outcomes to be realised, overcoming the barrier presented by missing clothing graphic descriptions—though it has gone unaddressed until now— seems a worthy goal to achieve.

Chapter 4

Building a VI-friendly Captioning Model

4.1 Introduction

Clothing graphic elements contain several visual aspects that depend on sight, including object, colour and scale details that VI people report as challenging to visualise when relying on the existing captions. These captions either miss the graphic elements or provide abstract descriptions (Cai et al., 2022; Hacheme & Sayouti, 2021; Han, Zhu, et al., 2023; Moratelli et al., 2023a; Yang et al., 2020), negatively impacting VI people's behaviour when purchasing clothing online. Our findings outlined in Chapter 3 indicated that these limitations either restrict VI people's options to choosing plain items or lead them to rely on sighted assistance to provide useful descriptions of graphics. In addition, our finding shows that developing a VI-friendly captioning model is essential to ensure a welcoming online shopping environment for VI people and reduce their reliance on sighted assistance.

Several studies have been performed on clothing image descriptions, resulting in proposals of clothing image–text datasets that may help facilitate the development of clothing captioning systems, including FashionGen, DeepFashion-MultiModal and FACAD. These datasets include images of clothing items within several categories (e.g., tops, skirts, and coats) and with a range of attributes (e.g., the existence of collars, patterns and graphics), along with their associated captions. These datasets have been used for building and evaluating image captioning and other related tasks (Han, Zhang, et al., 2023; Jiang et al., 2022; Yang et al., 2020). However, previous research (Gao et al., 2020; Nguyen et al., 2021) has highlighted a lack of sufficient clothing attributes in these datasets, which may indicate that these datasets are not necessarily gold-standard resources. Additionally, these datasets may not contain the level of detail required by VI people, as crowdworkers face challenges in determining what VI people will look for in such descriptions (Simons et al., 2020). More work in exploring these datasets and understanding these aspects is essential.

To date, however, there has been little research on clothing image captioning (Li, Ye, et al., 2021; Tateno et al., 2020; Yang et al., 2020), and it is known that clothing items often contain fine elements that are difficult for automated systems to detect and describe. Previous research has generated captions of clothing images that focus purely on abstract and high-level characteristics, including the recognition of clothing categories such as tops,

skirts, and coats, as well as clothing attributes such as collars, cuffs, and patterns, and the further identification of fabric types, colours, and outfit styles. Such work has adopted a grid-based image captioning approach, which extracts image content by partitioning images into grids of equal size and mining visual information across all grids. This approach has been considered the best fit for images of clothing.

However, when it comes to generating detailed descriptions of clothing images, the commonly used approaches (i.e., grid-based) do not suffice to generate focused, informative descriptions. For example, it does not generate detailed descriptions of components of clothing graphics. Looking ahead, given the great success of taking the ROI approach to generating detailed descriptions of general images (Li et al., 2019; Long et al., 2023; Yin et al., 2019), we believe that adopting the ROI-based approach in the field of clothes will guide the development of more detailed captions.

In regard to developing a VI-friendly captioning model, the power of LLM and the ability to understand different users' needs can be leveraged. Recently, LLMs have shown great success in offering flexible content tailored to different users. For example, they have been successfully used to simulate humans (Aher et al., 2023) in psycholinguistic, and social psychology experiments and generate biased content based on the user's beliefs (Simmons, 2022). Additionally, they have successfully been used to generate relevant questions about images (You et al., 2023; Zhu, Chen, Haydarov, et al., 2023).

Caption generation models are typically evaluated using either automatic or human evaluation. In automatic evaluation, generated captions are assessed on the basis of GT. In human evaluation, generated captions are commonly assessed on the basis of accuracy (Chen et al., 2022; Du et al., 2023; Fei, 2022; Hessel et al., 2021; Vinyals et al., 2015; Yi et al., 2020) and informativeness (Fei, 2022; Zhu, Chen, Haydarov, et al., 2023). Accuracy evaluation involves assessing whether generated captions accurately depict image content, while informativeness evaluates the amount of information conveyed. For generated captions aimed at aiding VI users, informativeness of the generated captions is assessed based on their ability to visualise the image (Kreiss et al., 2022).

When we single out VI users as the target audience, this is a novel research stream that warrants our pursuit. VI people are reliant on textual descriptions, and generating detailed descriptions of clothing image elements such as graphics will provide them with important additional information. Such descriptions significantly improve their understanding of the items that they can purchase. This area of retail development, however, has not been tackled before. Generally, the existing research does not sufficiently address the particular needs of VI individuals with regards to clothing descriptions.

This chapter aims to answer our fourth research question *RQ4: How can we build an effective clothing graphic description model to generate useful descriptions for VI users?*

We address this research question by exploring the following sub-questions:

- RQ 4.1: What are the limitations of the available ground truth captions from existing image–text datasets, considering VI people's needs?
- RQ 4.2: How *informative* are the captions generated by our model compared to those generated by existing models and ground truth data for VI users?
- RQ 4.3: How *accurate* are the captions extracted by our model compared to those extracted by existing models and ground truth data?

Here, we explore benchmark clothing text-image datasets and assess these datasets considering VI people's needs. This is essential to obtain insights into the suitability of these datasets to be used to train or evaluate clothing captioning models. Furthermore, exploring these datasets is crucial to create a new clothing graphics dataset. Additionally, we propose a new VI-friendly clothing captioning model. We discuss our approach to adapting to VI people's needs without the need for GT data. Additionally, we discuss our approach to overcoming the challenge of generating detailed descriptions of fine-grain clothing elements (i.e., clothing graphics). To ensure a comprehensive evaluation of our approach, we conducted several different evaluations: automatic and human evaluations, as well as manual analysis. We believe that the results produced from this work will establish a new direction for generating VI-friendly clothing descriptions.

Section 4.2 provides details of our model architecture. In section 4.3, we explore the benchmark clothing datasets and propose our new dataset. Our experiments and evaluation methods are discussed in section 4.5. In section 4.4; we then illustrate our procedure for conducting this research. Our findings are presented in section 4.6. We end this chapter by discussing our findings in section 4.7.

4.2 Model Architecture

We propose a new model to generate useful descriptions of graphics in clothing images. The aim of this model is to aid VI shoppers by helping them visualise the content of clothing graphics when purchasing garments online. Our image captioning approach was informed by the findings presented in Chapter 3. For example, the preferences of VI individuals to limit their clothing purchases to plain items, in order to avoid unintentionally selecting clothing items with undesired graphics, highlighting the crucial need of generating graphic descriptions. Furthermore, the finding that VI individuals seek detailed descriptions of graphic elements guided our decision to implement the ROI approach in the design of our image captioning model. The model proceeds in two phases. First, it identifies graphic elements from clothing (i.e., graphic recognition) and then it generates graphic descriptions (i.e., description generation). Our model architecture is depicted in Figure 4.1.

4.2.1 Graphic Recognition

In this phase, we adopt the ROI-based approach to focus on generating descriptions on a predefined salient element (i.e., graphics). Unlike grid-based, ROI-based approach has shown success in generating focused fine-grained descriptions (Li, Ye, et al., 2021). In the first step, we classified clothing images as either having graphics or not having graphics. This step is essential to exclude images that do not contain graphic elements. If a clothing item that contained graphics is identified, then it was passed to a segmentation model. In this step, graphic elements are segmented from other clothing elements.

Clothing Image Classification We used the FashionCLIP classification model (Chia et al., 2022). The FashionCLIP model is the first model to leverage the strength of the CLIP model (Radford et al., 2021) and to fine-tune it on fashion products. Similar to the CLIP model, the FashionCLIP model is a multi-modal classification model that can be instructed

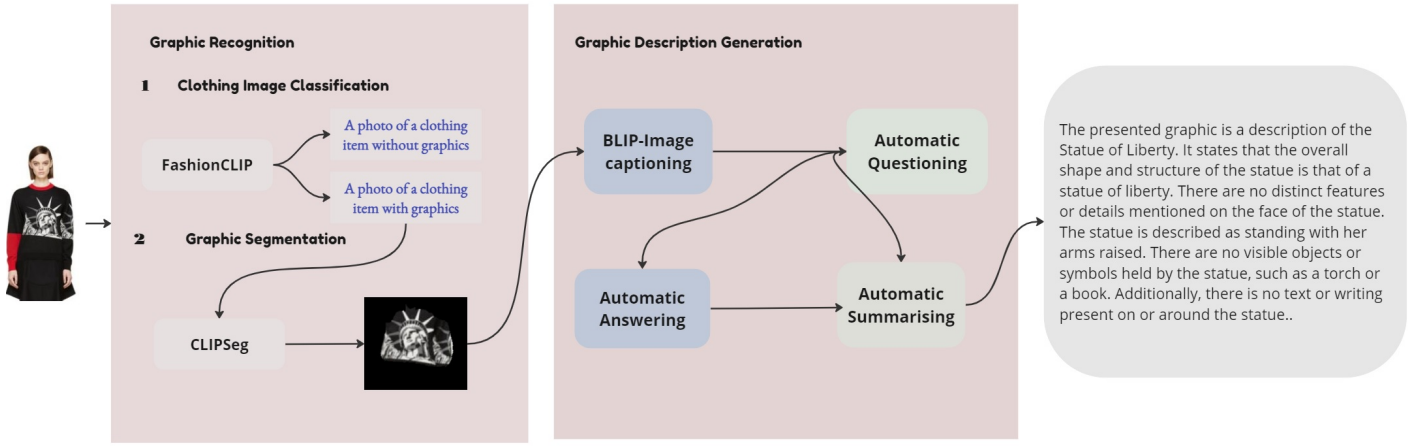


Figure 4.1: *The clothing graphic description model contains two phases: graphic recognition and graphic description generation.*

using prompt templates. In our model, we exploited the FashionCLIP model's knowledge of the clothing domain to classify whether the clothing image contained graphics. As prompts in our model, we utilised the template suggested by Radford et al. (2021), as follows:

"A photo of a clothing item {label}'"

Specifically, we used two prompts: *"a photo of a clothing item with graphics"* (prompt 1) and *"a photo of a clothing item without graphics"* (prompt 2). An example of the results generated by the implementation of the classification task is depicted in Figure 4.2. This figure shows three examples of clothing items: a plain item (left), an item with a graphic (middle) and an item with a pattern (right). Furthermore, the figure shows the predicted score for each prompt per item. For the plain item, the model predicted that the item most likely did not contain graphic elements (94% probability). For the items with the graphic and the pattern, the probability of identifying these elements was 74% and 96%, respectively. A detailed evaluation of the FashionCLIP model and the results are presented in 4.6.2.

When an image was classified as 'a clothing item without graphics', the following message was generated: 'This clothing item has no graphics'. If the model detected graphics on the clothing, the image was processed using the graphic segmentation step.

Graphic Segmentation Once a clothing item was identified as containing graphics, we applied the graphic segmentation task to crop the area of the image containing the graphics. This was essential to exclude other fine-grained fashion details such as buttons and pockets and force the graphic description generation (discussed below) to provide detailed information about only the graphics.

To implement the segmentation, we used the CLIPSeg model (Lüddecke & Ecker, 2022). A number of reasons underlie the rationale for using this model rather than traditional segmentation models, such as Mask R-CNN (He et al., 2018). First, employing traditional models requires a clothing graphic segmentation dataset to train the model on a specific task. However, existing clothing segmentation datasets do not include graphic segmentation. Instead, existing datasets mainly rely on segmenting clothing based on classes, such as skirts, shirts and scarves (Ge et al., 2019; Yang et al., 2014). Unlike these traditional

models, CLIPSeg, on the other hand, is a zero-shot learning model that does not require any pretraining process. This solves the problem of the lack of a clothing graphic segmentation dataset. Another reason for using the CLIPSeg model rather than traditional models is that the latter require a training process to recognise graphics in clothing images. CLIP model has the ability to understand images and text. CLIPSeg leverages this ability to perform the segmentation by simply using textual prompts to segment certain elements.

We employed the CLIPSeg model (Lüddecke & Ecker, 2022) using only one prompt: 'graphic'. We extracted the graphics and saved them for use in the following phase. Figure 4.3 shows examples of employing the CLIPSeg model in two examples of items. We then saved the segmented images to generate graphic descriptions in the second phase (i.e., graphic description generation).

			
Prompt 1: A photo of a clothing item with graphics	0.0593	0.7372	0.0403
Prompt 2: A photo of a clothing item without graphics	0.9407	0.2628	0.9597

Figure 4.2: *Examples of the results generated by FashionCLIP (Chia et al., 2022).*

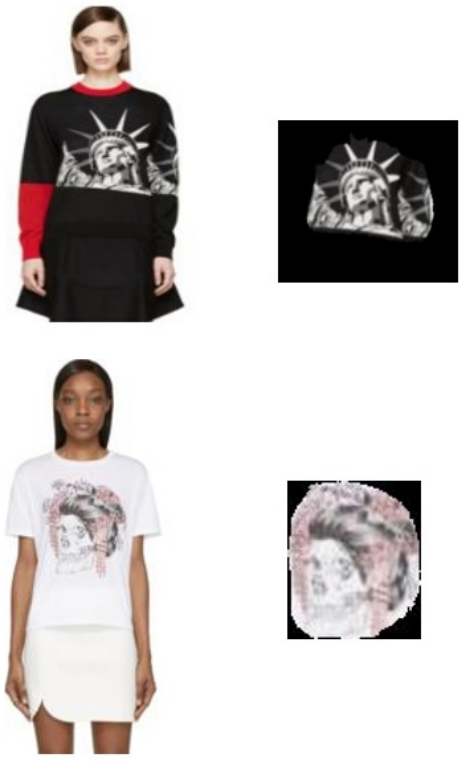


Figure 4.3: *Examples of results generated by CLIPSeg (Lüddecke & Ecker, 2022).*

4.2.2 Graphic Description Generation

The graphic segmentation obtained from the previous phase is used in this phase to generate focus descriptions. We aim to generate graphic descriptions tailored to VI people's needs. Instead of training our model on fixed examples of VI-friendly captions, which are most likely annotated by crowd workers who are unfamiliar with VI people's needs (Simons et al., 2020), we leverage the wide knowledge of LMs to build a zero-shot model (the models have not been trained). We employed ChatGPT, BLIP-2 captioning and BLIP-2 question-answering models. ChatGPT has shown wide knowledge, including linguistic and semantic reasoning ability (Naveed et al., 2023; Yin et al., 2023). BLIP-2 is an advanced model that is designed to generate textual data from images in several tasks, such as image captioning and question answering. Its ability to generate fewer incorrect answers to visual questions and its superiority to other powerful visual models have been demonstrated (You et al., 2023).

We have designed this phase to mimic image description conversations between VI individuals and sighted assistants. Specifically, we employ OpenAI's ChatGPT model [1] (gpt-3.5-turbo) to ask questions a blind person might ask in order to imagine graphic images, and we use the BLIP-2 model (Li, Li, et al., 2023) to answer the questions. Finally, ChatGPT is used again to summarise the responses generated by BLIP-2. This step is essential to generate well-structured, fluent captions.

We address the need to generate questions that are similar to questions a VI person might ask by aligning the LM with human behaviours. This alignment is done by designing prompt

[1] https://openai.com/

engineering to identify the targeted users (i.e., people with visual impairment). Aligning LMs with human behaviour has shown great success in generating personalised text (Argyle et al., 2023; Simmons, 2022). Furthermore, prompt engineering LLMs are capable of generating text simulating human behaviour (Aher et al., 2023).

The automated conversation was set up using the BLIP-2 image captioning model to generate contextual information about the graphic image. The automated conversations consist of three steps: automatic questioning, automatic answering, and automatic summarising. An example of prompts and results for each component is illustrated in Figure 4.4.

Automatic Questioning We designed our ChatGPT prompt, similar to the one used by Zhu, Chen, Haydarov, et al. (2023), to consist of three parts: a task request *Preq*, graphic information *Pqinfo*, and a task instruction *Pins*.

In the task request, we designed our prompt using the findings outlined in Chapter 3, which indicate that VI people seek information on graphic elements that help them mentally imagine graphics, such as colour, position, and the size of the graphic. Specifically, we request ChatGPT to produce questions that help VI people mentally imagine the graphic, which is to be based on the contextual information (graphic information) obtained from BLIP-2. In the task instructions, we guide ChatGPT to avoid any questions that BLIP-2 is not able to answer (i.e., sizing questions), as BLIP-2 would create answers that do not exist if these questions were not avoided. This was noted in our experiment. To prevent ChatGPT from producing too many questions, we limit the list to five using the following prompt:

"Print five questions a blind person would ask to mentally imagine the graphic of a {Pqinfo}. Avoid sizing questions."

Automatic Answering We used BLIP-2 to respond to the questions generated by ChatGPT, in the Automatic Questioning step, through prompting with the following template (this prompt template is the same for all five questions; "num" represents the question number): *Question: {ChatGPT question [num]} Answer:*

Automatic Summarising The dialogue produced from the Automatic Questioning and Automatic Answering process is then used as input to the automatic summarising task to generate captions for clothing images. Similar to the automatic questioning step, our prompt mechanism in this step consists of three components: the task request *Preq*, the context *Pc*, and task instructions *Pins*. The generation of each answer is prompted by *Preq* + *Pc* + *Pins*.

In the task request *Preq*, we ask ChatGPT to generate a summary of the dialogue obtained from the ChatGPT and BLIP-2 questioning answering system. As it has been observed that ChatGPT tends to use words similar to those we used in the prompt, we avoid using terms that we do not want to appear in the summary. For instance, using a prompt such as *"Generate a summary of the dialogue"* would result in the word "dialogue" appearing in the generated summary with high probability.

In the context *Pc*, the dialogue is structured based on the following template (this prompt template is the same for all five questions; "num" represents the question number):

"Question [num]: {question} Answer: {answer}".

In the task instructions *Preq*, we instruct ChatGPT to avoid including additional content and to disregard any questions for which no answer was provided by BLIP-2. A setup

phrase is also provided to structure the summary and prevent ChatGPT from generating an undesirable response. For example, ChatGPT might start the generated text with a phrase such as "this dialogue" or "this text", which is unsuitable to be included in the final caption.

"Generate a summary of the graphic descriptions. {Pd} *Don't add content. Exclude questions that lack answers. Start the summary with "The presented graphic is"".*

BLIP-2 Captioning:
a blue and white striped pillow with a bunny on it

Automatic Questioning:

Prompt:
Print five questions a blind person would ask to mentally imagine the graphic of a blue and white striped pillow with a bunny on it. Avoid sizing questions.
Result:
1. What shade of blue are the stripes on the pillow?
2. How many stripes are there on the pillow?
3. Is the bunny facing a certain direction on the pillow?
4. Are the stripes evenly spaced on the pillow?
5. Can you describe the expression on the bunny's face?

Automatic Answering:

Prompt: 1. What shade of blue are the stripes on the pillow? Answer:
Result: blue

Prompt: 2. How many stripes are there on the pillow? Answer:
Result: 2 stripes

Prompt: 3. Is the bunny facing a certain direction on the pillow? Answer:
Result: Yes

Prompt: 4. Are the stripes evenly spaced on the pillow? Answer:
Result : Yes

Prompt : 5. Can you describe the expression on the bunny's face?
Result :

Automatic Summarising:

Prompt : Generate a summary of the graphic descriptions. Question: 1. What shade of blue are the stripes on the pillow?
Answer: blue
Question: 2. How many stripes are there on the pillow?
Answer: 2 stripes
Question: 3. Is the bunny facing a certain direction on the pillow?
Answer: Yes
Question: 4. Are the stripes evenly spaced on the pillow?
Answer: Yes
Question: 5. Can you describe the expression on the bunny's face?
Answer:

Don't add content. Exclude questions that lack answers. Start the summary with "The presented graphic is"

Result :
The presented graphic is of a pillow with two evenly spaced blue stripes. The bunny on the pillow is facing a certain direction, and its expression is not described.

Figure 4.4: *Example of prompts and results for each component of the graphic description generation.*

4.3 Benchmark Datasets

This section explores the existing clothing datasets and introduces a novel clothing graphic dataset.

4.3.1 Clothing Datasets

We examined the available image-text datasets (i.e., clothing datasets that include images and their corresponding captions) to investigate their suitability for training or evaluating VI-friendly graphic captioning models. We chose to conduct our analysis on three large publicly available fashion image-text datasets: FashionGen (Sadeh et al., 2019), DeepFashion-MultiModal (Jiang et al., 2022) and FACAD (Yang et al., 2020). The rationale for choosing these datasets is that they are the most commonly used in image-text reserch (Gao et al., 2020; Han et al., 2022; Yang et al., 2020). In addition, the other clothing image-text datasets are either publicly unavailable (Cai et al., 2022; Sadeh et al., 2019) or designed for certain people, such as African style (Hacheme & Sayouti, 2021). Details of the chosen datasets are shown in Table 4.1 and discussed below.

We downloaded the datasets (i.e. FashionGen [2], Deepfashion-MultiModal [3] and FACAD [4]). directly from the respective websites. From each dataset, we randomly selected items to utilise in our analysis.

Dataset	Year	**No. of items (images per item)**	**Img size**	**Length of caption (no. of words per caption)**	**No. of Cat and Att**	**Sources & how were they annotated?**
FashionGen dataset (Rostamzadeh et al., 2018)	2018	293,008 (1–6 images)	1360X1360	1 paragraph Cap (/)	48 Cat	Images are photographed, and captions are annotated by a professional
Deepfashion-MultiModal (Jiang et al., 2022)	2022	11,484 images	1024 X 512	/	50 Cat & 1000 Att	DeepFashion(Liu et al., 2016)
FACAD (Yang et al., 2020)	2020	993K (6 $\sim$ 7*images*)	1560X2392	130K (approximately 21 words)	78 Cat & 990 Att	Google Chrome

Table 4.1: *Comparison of the three commonly used clothing image-text datasets, showing details of the year of publication, the number of items, image size and number of captions (Cap) included, the number of categories (Cat) such as dress and T-shirt and attributes (Att) such as colour and sleeves and the source. Details are obtained from the original sources. "/" denotes unavailable information.*

FashionGen is designed for image-text retrieval tasks. It contains 293,008 items, with

[2](n.d.). *Fashion-Gen Dataset.* Retrieved from https://fashion-gen.com/. Accessed 2022.

[3](n.d.). *DeepFashion-MultiModal Dataset.* Retrieved from https://github.com/yumingj/DeepFashion-MultiModal. Accessed 2022.

[4](n.d.). *DeepFashion-MultiModal Dataset.* Retrieved from https://github.com/yumingj/DeepFashion-MultiModal. Accessed 2022.

an average of 1-6 images per item. The images in FashionGen are represented from different viewpoints (see Figure 4.5). Images of the same clothing item share the same description. In some cases, however, multiple images with different appearances share the same description. This ambiguity causes noise when training an ML model. Figure 4.6 shows an example of a description that describes only a T-shirt, even though an undescribed coat hides the T-shirt in the second image.



Figure 4.5: *Clothing images from the FashionGen dataset showing two different views of the same item.*



Figure 4.6: *Two images from the FashionGen dataset with different appearances sharing the same description.*

DeepFashion-MultiModal is designed for text-image retrieval tasks and comprises 11,484 items. Each item is represented in three positions - full, front and back - each of which has its own caption. Figure 4.7 depicts a clothing item displayed in images of the front and back views with different text descriptions. The GT text description suffers from two issues:

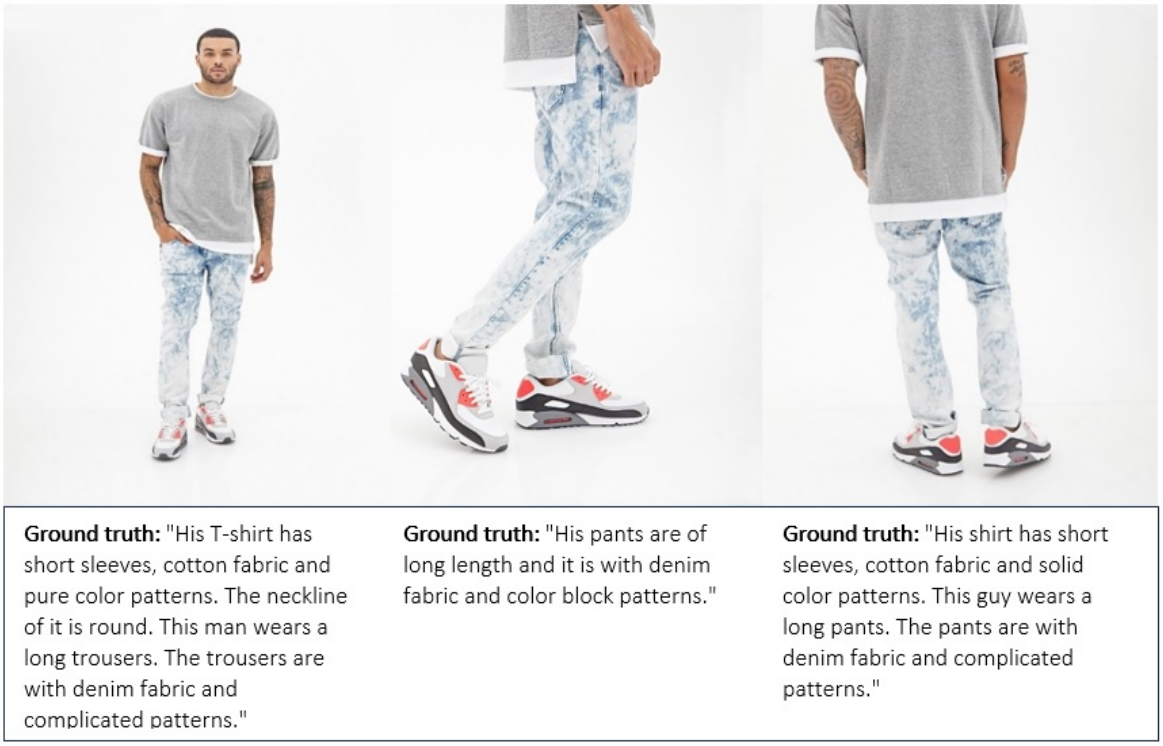


Figure 4.7: *Examples of clothing images from the DeepFashion multimodal dataset showing three different positions and the corresponding captions.*

- *Language misrepresentation* In the text descriptions, some inappropriate language forms are presented (see Figure 4.8). For instance, in some text descriptions, the same term (outlined in red) is repeated, or the same image is represented using female and male pronouns.

- *Conflicting description of the same image*

  Another issue is the use of different terms to describe the same clothing item. For example, in Figure 4.9, two distinct clothing types, particularly a shirt and a sweater, are used to describe the same clothing item.

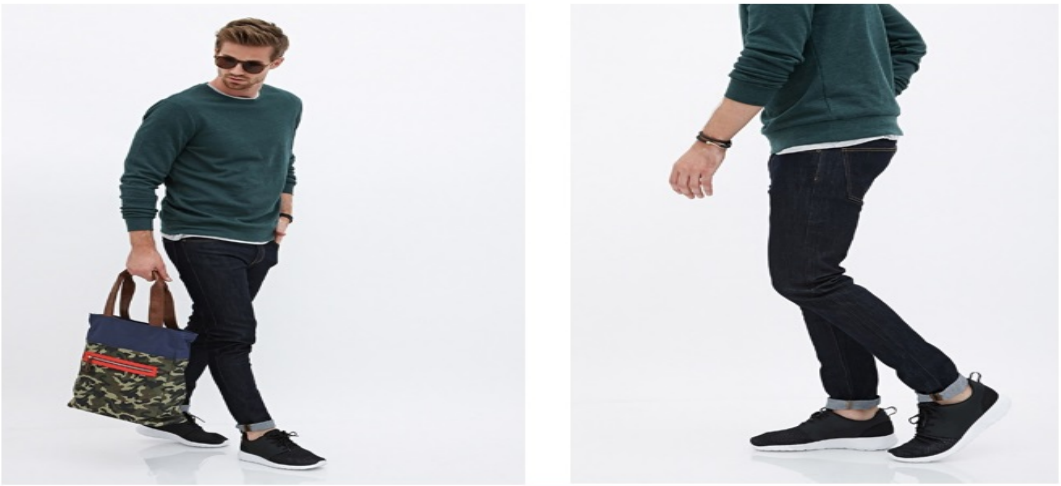


Ground truth: "The gentleman wears a tank tank top with pure color patterns. The tank top is with cotton fabric. It has a square neckline. The gentleman wears a long pants. The pants are with cotton fabric and solid color patterns."

Ground truth: "The female wears a medium-sleeve shirt with graphic patterns. The shirt is with knitting fabric and its neckline is round. There is an accessory in his her neck. There is an accessory on her wrist. There is a ring on her finger."

Ground truth: "The female is wearing a long-sleeve sweater with graphic patterns. The sweater is with cotton fabric. The pants the female wears is of three-point length. The pants are with cotton fabric and graphic patterns. The outer clothing this female wears is with knitting fabric and pure color patterns. This female wears a ring. There is an accessory on her wrist. There is an accessory in his her neck."

Figure 4.8: *Examples of description from the DeepFashion-MultiModal dataset with language misrepresentation problems.*



Ground truth: "The person is wearing a long-sleeve sweater with solid color patterns. The sweater is with cotton fabric and its neckline is crew. The person wears a long trousers. The trousers are with denim fabric and pure color patterns. The gentleman wears sunglasses. There is an accessory on his wrist."

Ground truth: "This person wears a long-sleeve shirt with solid color patterns. The shirt is with cotton fabric. The trousers this person wears is of long length. The trousers are with denim fabric and solid color patterns. There is an accessory on his wrist."

Figure 4.9: *Examples of different terms used to describe the same clothing item in DeepFashion multimodal text descriptions.*

FACAD is designed for clothing image description tasks and contains 993,000 items, with an average of 6–7 images per item. All the images of the same clothing item share the same text descriptions. Each image is represented in three separate channels (i.e., red, green and blue). This dataset, however, unlike the other two, is designed to attract customers and is not suitable to be employed to provide honest, objective descriptions to assist users with visual impairments (Yang et al., 2020). Figure 4.10 shows examples from the FACAD dataset, which exemplify the marketing-friendly nature of the captions.

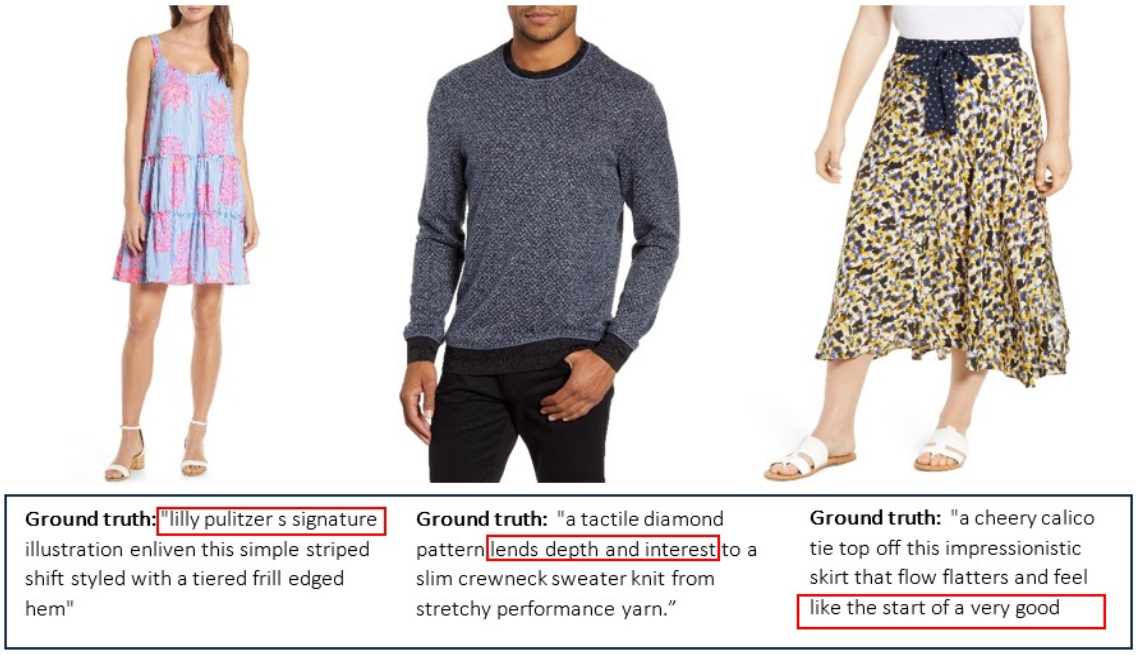


Figure 4.10: *Examples of descriptions from the FACAD; using marketing-friendly terms to entice customers.*

In our analysis we also investigated captions for clothing with graphics. It was noted that ground-truth captions in the three datasets attempted to provide characteristics/elements of the clothing items, however, they do not provide sufficient descriptions to permit VI users to visualise the graphics. Examples of images and their associated captions from the three datasets are depicted in Figure 4.11. All the examples lack graphic descriptions. In the case of the item from the FashionGen dataset (Figure 4.11, left), the caption only identifies what can be seen in the graphic (i.e., text and picture of a parrot). This provides clues for a VI person but does not allow a VI person to mentally visualise the graphics by relying on the captions alone. For example, a VI person would like to know what colours are presented in the graphic, what the text says and how they are positioned. In the case of the item from the DeepFashion-MultiModal dataset (Figure 4.11, middle), the caption only shows the presence of a graphic without identifying what it is and its colour. Finally, in the item from FACAD, the caption identifies the presence of a graphic and describes it as inspiring and eye catching without providing any information regarding the colour and the identifying characteristics of the object.

These limitations affect the quality of these datasets and thus their usefulness in training and evaluating image captioning models to generate detailed descriptions for applications designed for people with visual impairments.



'Linen short sleeve T-shirt in ivory. Crewneck collar. Text and parrot graphic printed at front in black, blue and yellow. Serge stitch at sleeve opening and hem. Tonal stitching.'

'This person is wearing a sleeveless tank top with graphic patterns. The tank top is with cotton fabric. It has a round neckline. The pants this person wears is of three-point length. The pants are with denim fabric and pure color patterns. There is a hat in her head. There is a ring on her finger. There is an accessory on her wrist.'

'[B]ring the good vibe in this easy cotton tee styled with an inspiring and eye-catching graphic.'

Figure 4.11: *Images and their ground-truth captions from FashionGen (left), DeepFashion-MultiModal(middle) and FACAD (right).*

4.3.2 A Novel Clothing Graphics Dataset

To date, no dataset has been dedicated to clothing graphics and descriptions. Therefore, to facilitate the evaluation of our model through automatic and human evaluations and to serve future research endeavours, we created a novel Clothing Graphic Dataset (CGD) dataset (discussion of how to utilise this dataset in future is provided in Chapter 5). The CGD includes items from all three datasets mentioned above (i.e., FashionGen, DeepFashion-MultiModal, and FACAD).

To streamline our selection process, when extracting items from the three datasets, we selected only items that were classified as 'graphic' in the original dataset. However, within our CGD, we only included items that our model classified as 'graphic'. We collected 135 items from the FashionGen dataset, 290 items from the DeepFashion-MultiModal dataset

and 143 items from the FACAD dataset.

The CGD contains graphic items (i.e., images, GT data and graphic descriptions generated from our model). We extracted items classified as 'graphic' from the FashionGen and DeepFashion-MultiModal datasets. Since the FACAD includes various types of items, such as shoes and accessories, that are outside the scope of our assessment, we limit the subset to include only items that are either 'tee', 'sweatshirt' or 'pullover'. Each item in our CGD includes an item ID, image names, GT captions and descriptions generated by our model. The image IDs and names remain consistent with those used in the original datasets. An excerpt of the CGD is shown in Figure 4.12.

This novel dataset will facilitate future applications in graphic clothing image-text applications, as it is the first dataset to accumulate a number of clothing items with graphics. Future work could use this novel dataset to generate a new clothing graphic segmentation dataset or employ it in a clothing image captioning model. Further work could use the GT and graphic descriptions (i.e., those generated by our model) to investigate and compare the performance of captions generated by LLM to human-written captions (i.e., GT data).


```
{
    "id": 86800,
    "images": [
        "000149.png",
        "000150.png",
        "000151.png",
        "000152.png"
    ],
    "description": "Long sleeve sweater in white. Crewneck collar. Raglan sleeves. Multicolored monkey graphic printed at front. Stretch sleeve cuffs and hem. Tonal stitching.",
    "graphic_description": "The presented graphic is a green wallet with one owl depicted on it. The owl is facing a random direction and there are no additional details or patterns on the wallet mentioned."
},
```

```
{
    "id": "00000019",
    "images": [
        "WOMEN-Graphic_Tees-id_00000019-01_1_front.jpg",
        "WOMEN-Graphic_Tees-id_00000019-01_2_side.jpg",
        "WOMEN-Graphic_Tees-id_00000019-01_4_full.jpg",
        "WOMEN-Graphic_Tees-id_00000019-01_7_additional.jpg"
    ],
    "description": "Her tank shirt has no sleeves, cotton fabric and graphic patterns. The neckline of it is round. This person wears a three-point shorts. The shorts are with denim fabric and pure color patterns. There is an accessory on her wrist. This female has neckwear. This lady wears a ring.",
    "graphic_description": "The presented graphic is a logo that uses the colors black, white, and grey. It is a combination of the letters i and o, creating a simple and clean design. The logo conveys a certain emotion or message, although the specific details are not provided. There are no specific elements in the logo that stand out or catch the eye, as it is designed to be simple and clean."
}
```

```
{
    "id": "22963",
    "images": [
        "22963_1.jpg",
        "22963_2.jpg",
        "22963_3.jpg",
        "22963_4.jpg",
        "22963_5.jpg",
        "22963_6.jpg"
    ],
    "description": "a cheeky upside down play graphic complete with signature heart logo splash across this soft cotton tee with signature whimsy",
    "graphic_description": "The presented graphic is a heart-shaped logo that incorporates distinct symbols and icons. The colors used in the logo are not specified. There are no specific patterns or textures within the logo. The logo conveys a certain mood or emotion."
}
```


Figure 4.12: *An excerpt from the CGD. It shows an item from the FashionGen (top), DeepFashion-MultiModal (middle) and FACAD (bottom). CGD: Clothing Graphic Description dataset, FACAD: Fashion Captioning Dataset.*

4.4 Research Procedures

In Chapter 3, we identified the end users (people with VI) and conducted a comprehensive study to explore their needs. We also discussed the limitations of online shopping that hinder their ability to independently purchase clothing online. Our findings and our literature synthesis revealed a need to produce better clothing representations that consider the needs of those with VI. Following our analysis, we proposed a new model to provide VI people with more informative detailed descriptions of clothing items that would better facilitate their mental visualisation of graphic elements. While developing a new model, we also explored the limitations of the GT captions used in the benchmark clothing image-text datasets, addressing the unique consideration of VI people's needs.

We conducted several evaluations to assess the proposed model, including preliminary analysis and automatic and human evaluations, as illustrated in Figure 4.13. Details of the evaluation methods and findings are presented in sections 4.5 and 4.6, respectively. The preliminary analysis assessment included recalling our previous knowledge of VI people's needs. When conducting the human assessment, we compared the captions generated from our model with those generated from the BLIP-2 captioning model and GT captions. Only the BLIP-2 captioning model was used in the automatic evaluation. In the early stage of our evaluation, we considered assessing our model against publicly available clothing image-text models, such as FashionViL (Han et al., 2022) and SRFC (Yang et al., 2020). However, these models were trained using fashion captioning datasets, such as FACAD and FashionGen, which have many limitations (see our discussion in section 4.3) and do not fulfil VI people's needs. Therefore, as we mentioned earlier, we chose to compare the performance of our model against BLIP-2, which has been recently widely employed in evaluating image captioning models (You et al., 2023; Zhu, Chen, Haydarov, et al., 2023). Furthermore, it was used in our evaluation to represent zero-shot models and the grid-based captioning approach. The GT captions were used in the human assessment to compare our model, which adopts the LLM to facilitate VI people's needs against captions written by humans (i.e. GT captions). Ethics approval was obtained to conduct human evaluations (see the ethical application and approval documents in Appendix E).

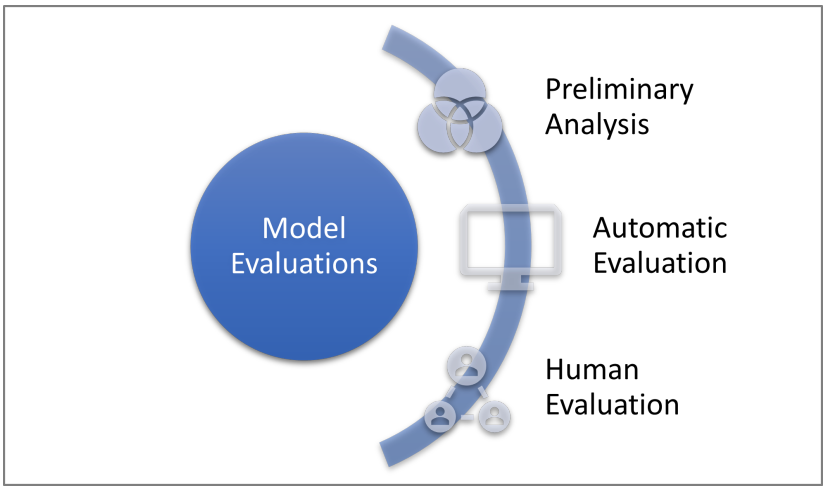


Figure 4.13: *Evaluation Methods.*

4.5 Experiments and Evaluation Methods

In this section, we present the experimental setup adopted to run our model and the evaluation methods implemented to examine the informativeness and accuracy of our model, aiming to answer our research questions RQs 4.1 - 4.4.

4.5.1 Experimental setup

In our experiments, we employed OpenAI's ChatGPT model (gpt-3.5-turbo) [5], recognised as the most powerful and accessible model at the time of this study's development. We set the decoding temperature to 0 to generate only results with higher certainty and to ensure stable results. We also employed BLIP-2 (blip2-opt-2.7b) (Li, Li, et al., 2023) in developing our model. When implementing our model, we used only the front view of the clothing items (ignoring other images, such as the right and left sides). Our graphical segmentation task may identify several segmentations for some images, especially when a clothing item consists of various distributed graphics. In this case, we only generated from the largest segment.

4.5.2 Preliminary Analysis

The proposed model was assessed to evaluate whether it generates informative descriptions that address VI people's needs. First, the Graphic Description Generation component (referring to the architecture of the proposed model in section 4.2) was analysed. This analysis focuses on assessing whether the questions generated by the automatic questioning task meet the essential features sought by VI people to visualise clothing graphic elements, considering the findings reported in chapter 3. Items were randomly selected from the FashionGen, DeepFashion-MultiModal and FACAD datasets. Second, the generated questions were analysed to assess whether the automatic questioning task produced questions relevant to the enquired clothing graphic image. This analysis was applied to a sample of 100 items randomly selected from the CGD. The informativeness of the clothing graphic descriptions generated from the proposed model was assessed in comparison to the GT captions.

4.5.3 Automatic Evaluation

Image Classification To establish whether the FashionCLIP model can determine if a clothing item contains graphics using the classification prompts applied in our model (refer to section 4.2), a classification assessment was adopted using 500 items randomly selected from the FashionGen dataset. In this assessment, the classification provided by the FashionGen dataset was used as reference data. The assessment adopted the standard classification evaluation metrics: accuracy, precision, recall and F1 scores. The accuracy metric calculates the number of correctly predicted cases by the FashionCLIP model about the total number of cases. The recall metric measures the number of correctly predicted positive cases compared to the actual number of positive cases. In contrast, the precision metric calculates the number of positive cases that are correctly classified. The F1 score, in contrast, provides a balance between precision and recall metrics by outputting their geometric mean.

[5] https://openai.com/

Image Captioning To assess how similar the captions generated by our model were to those generated by the BLIP-2 model, we compared our model's captions with those of BLIP-2. The BLIP-2 captioning model utilises a grid-based approach, whereas our model utilises an ROI-based approach. The descriptions generated by BLIP-2 were for the whole image, whereas those generated by our model were for the graphic element only.

Several automatic image-captioning models were used: BLEU, ROUGE, METEOR, CIDEr and SPICE. These metrics evaluate the performance of output text by computing the similarity of the generated text in comparison with human-annotated data (i.e., reference data). Here, we introduce these models and outline the rationale underlying their selection.

- BLEU The Bilingual Evaluation Understudy (BLEU) metric was originally introduced for machine translations (Papineni et al., 2002), and it was widely applied in image captioning (Celikyilmaz et al., 2021; Moratelli et al., 2023b; SS et al., 2023; Tang et al., 2023). BLEU is an n-gram metric used to measure the similarity between two sentences, a machine-generated candidate sentence and a reference date one. Specifically, it divides reference and candidate sentences into n-grams and compares the n-grams in both texts. For instance, in the BLEU-1 metric, the degree of similarity is assessed based on the matching of unigrams, but in the BLEU-2 metric one, the focus is on the matching of bigrams. Normally, up to 4 grams are used. As the n-gram value increases, the focus shifts towards the fluency of the text. BLEU-1, for instance, is used to assess effectiveness, whereas BLEU-4 is employed to evaluate fluency (Bai & An, 2018). In our evaluation, BLEU-1 and BLEU-4 are selected to provide information on the degree of similarity in terms of unigram and fluency levels. In addition, these metrics provide useful information on the n-gram precision results [6].

- ROUGE

  Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin & Hovy, 2003) is a set of multiple metrics designed to evaluate automatic summarisation but also used to assess image-captioning tasks (Sharma et al., 2023). While BLEU metrics rely on precision (i.e., how many n-grams in the output are correctly predicted), ROUGE metrics rely on recall (i.e., how many n-grams in the reference data are presented in the output text). Among the various versions of ROUGE metrics, ROUGE-L is the most applied (Luo et al., 2022). This variant measures the similarity between the output text and reference data based on the longest common sub-sequence (LCS) (i.e., words that are in sequential order but not necessarily consecutive). ROUGE metrics are a modification of those of BLEU based on further consideration of the structure of a sentence (Bai & An, 2018). Adopting the ROUGE-L metric in our evaluation enhances our comprehension of the structural similarity of sentences between our model and BLIP-2 descriptions compared to the reference data.

- METEOR

  The Metric for Evaluation of Translation with Explicit Ordering (METEOR) (Banerjee & Lavie, 2005) is another common evaluation metric in image captioning (Zhu, Chen, Haydarov, et al., 2023) that uses precision and recall. The significance of the METEOR

[6] aac-metrics. (n.d.). *Metrics for evaluating Automated Audio Captioning systems, designed for PyTorch.* Retrieved from https://pypi.org/project/aac-metrics/#bleu. Accessed February 2024.

is that it detects all similarities between text-generated and reference data using stemming and synonym matches. It is also designed to be applied at the sentence rather than the corpus level (Celikyilmaz et al., 2021).

- CIDEr Consensus-Based Image Description Evaluation (CIDEr) (Vedantam et al., 2015) was designed for image-captioning evaluation. This approach measures the similarity of the output text to a set of human-generated texts. The result is based on the calculation of the term frequency–inverse document frequency (TF–IDF), which assigns weights to each n-gram based on the frequency of each one appearing in the reference data. Low weight is given to an n-gram shown to be redundant throughout the reference data. The CIDEr correlates well with human judgement (Celikyilmaz et al., 2021). The CIDEr-D version is the most widely used in image-captioning evaluations (Sharma et al., 2023). The use of CIDEr-D indicates the similarity between captions generated by our model and BLIP-2 compared with the reference data, while excluding any redundant text that might have occurred by chance.

- SPICE Semantic Propositional Image Caption Evaluation (SPICE) (Anderson et al., 2016) was designed for image-captioning evaluation. This approach is based on the sense-graph generation of reference data that encodes attributes, objects and relationships. Rather than focussing on measuring the grammar and fluency of the output captions, SPICE measures semantic similarity. Compared with the CIDEr metric, SPICE correlates better with human judgement (Liu et al., 2017). Using the SPICE metric in our evaluation provides an understanding of the semantic similarity of our model and BLIP-2 captions compared with the reference data.

Our Method Our automatic evaluation was applied using a set of automatic image captioning metrics, including BELU, ROUGE-L, METEOR, CIDEr-D and SPICE. We conducted the evaluation on each subset of the FashionGen, DeepFashion-MultiModal and FACAD datasets collected in our CGD dataset. There were 135 items for FashionGen and 290 and 143 items for DeepFashion-MultiModal and FACAD, respectively. In our automatic evaluation, we compare the performance of our model against the BLIP-2 captioning model. BLIP-2 has recently been widely employed in evaluating image captioning models (You et al., 2023; Zhu, Chen, Haydarov, et al., 2023) and has been used to represent grid-based and zero-shot models (Li, Li, et al., 2023), that is, those not trained on a clothing captioning dataset.

4.5.4 Human Evaluation

Human evaluation is considered the gold standard for evaluating image captioning (Bernardi et al., 2016; Clark et al., 2021; van Miltenburg, 2023). One of its strengths is that captions are assessed by their end users, who judge whether they appreciate the output text (van der Lee et al., 2021). Another strength is that human evaluation can easily adapt to assessing several aspects of the generated captions, such as fluency, accuracy, informativeness and marketing (Yang et al., 2020). This evaluation method has recently gained more attention in the field of NLP (Schuff et al., 2023).

Evaluation Approach

When human evaluation is used on its own, two approaches are commonly used (Belz & Reiter, 2006): intrinsic and extrinsic.

The intrinsic approach is used to measure the effectiveness of a system in itself without considering other downstream tasks. Recent research in text generation systems relies on intrinsic evaluation (Celikyilmaz et al., 2021; van der Lee et al., 2021). This approach assesses the overall quality of the output on specific aspects such as *naturalness* (grammar and fluency) (Fei, 2022; Lan et al., 2017), *relevance* and *accuracy*, which involves how appropriate a caption is to the relevant image (Du et al., 2023; Fei, 2022; Hessel et al., 2021; Yi et al., 2020; Zhu, Chen, Haydarov, et al., 2023), and *richness*, which assesses how informative a caption is (Fei, 2022; Zhu, Chen, Haydarov, et al., 2023).

Unlike intrinsic evaluation, the extrinsic approach is a task-based measure of a system's usefulness in downstream tasks (van Miltenburg, 2023). The extrinsic approach measures a system's effectiveness in achieving its goal (van der Lee et al., 2019). Although the strength of the extrinsic approach lies in evaluating how a system actually performs, it is rarely employed to evaluate automatic text due to its complexity, cost, and time-consuming build (Gatt & Krahmer, 2018; van der Lee et al., 2019).

Evaluation Metrics

Two forms of evaluation are employed in automatic text evaluations: qualitative and quantitative. Qualitative evaluations aim to collect participants' assessments in words; data gathered in this approach is not numerical. This approach can involve several methods, including open-ended questionnaires, interviews, observations and qualitative text analysis. The qualitative approach is preferred if an evaluation aims to improve a system. The significant advantage of the qualitative approach is that it offers greater flexibility regarding participants' opinions compared with quantitative metrics (Schuff et al., 2023). Another important strength of employing a qualitative approach is that it typically requires a smaller sample size than quantitative methods (van der Lee et al., 2021). Although qualitative evaluation can provide a detailed assessment of a proposed system rather than comparing its quality with other models, it requires more analytical effort and suffers from reliability issues (van der Lee et al., 2021). Thus, qualitative evaluation methodologies are infrequently employed in automatic text evaluations (van der Lee et al., 2021).

By contrast, quantitative evaluation is commonly used in automatic text evaluations, mainly with questionnaires. This approach should be applied on a large scale (van der Lee et al., 2021). The most commonly used instruments for quantitative data collection are questionnaires with Likert scales, which can involve single or multiple items (Schuff et al., 2023).

In single-item evaluation, participants are asked to rate the quality of a text on its own, without judging the quality of the text based on other candidate data. For example, in image captioning evaluations, image caption pairs are presented on a Likert scale of several points, such as two (Zhu, Chen, Haydarov, et al., 2023), and five points (Sarto et al., 2023).

While single-item-based evaluation assesses the quality of output text without considering any candidate reference, multiple-item rating assesses a text using baseline reference data. The most common approach employed in image captioning is selecting the best or at least

most preferable caption among multiple candidates (Hessel et al., 2021; Sarto et al., 2023; Vedantam et al., 2015; Zhu, Chen, Haydarov, et al., 2023). For instance, Zhu, Chen, Haydarov, et al. (2023) asked participants to select the most informative caption among three candidates with the associated image present. Instead of choosing captions based on a reference image, Hessel et al. (2021) asked participants to select the better caption between two candidates compared with reference text. However, applying quantitative metrics alone does not allow participants to explain how they interpret what they have been asked or to provide clear explanation of their opinions (van der Lee et al., 2021).

Evaluation Design

Participants can be assigned to conditions (output text) in two ways (Schuff et al., 2023; van der Lee et al., 2021): within and between subjects.

In a between-subjects design, each participant is assigned to judge only one condition (i.e., output from a single algorithm/source). Thus, a larger number of participants is required to obtain reliable results than is needed for within-subjects designs. Between-subjects designs are not often used in automatic text evaluations; instead, within-subjects designs (also known as repeated measurements) are recommended and widely used (van der Lee et al., 2021). This approach exposes each participant to all conditions and has the strength of permitting participants to compare multiple conditions simultaneously. Another advantage is that it offers a large sample of evaluations. This approach, however, has several drawbacks. One negative aspect is that exposing participants to a large number of instances may lead to participant fatigue (Schuff et al., 2023). Another issue is that repeated evaluation may cause later conditions to suffer from a carry-over effect. To overcome these downsides, Schuff et al. (2023) suggests recruiting a large-scale sample and controlling the number of conditions presented to mitigate any possible fatigue. It is suggested that more than one participant judge each condition to achieve reliable and realistic evaluations (van der Lee et al., 2021; van Miltenburg et al., 2021). In image caption evaluations, conditions have been judged by five (Chen et al., 2022; Fang et al., 2015; Li & Yamaguchi, 2017; Vinyals et al., 2015; Yang et al., 2020), three (Hessel et al., 2021; Hodosh et al., 2013; Sadeh et al., 2019) or two (Lan et al., 2017; Yatskar et al., 2014) participants.

Sample Size

One important aspect to consider before conducting an evaluation study is sample size. That decision impacts the statistical power of the study, which in turn affects the level of reliability of the results. A small sample size leads to low levels of statistical power (Schuff et al., 2023; van der Lee et al., 2021). It is essential to ensure enough statistical power, as low power may result in reporting false negative results (Schuff et al., 2023). Generally, a power level of 0.80 is recommended (Lenth, 2004); this suggests that statistical differences will be identified 80% of the time.

In automatic text evaluations, two sample sizes are often used: large and small-scale samples. Large-scale samples normally consider recruiting between 10 and 670 annotators (van der Lee et al., 2021). They are anonymous crowdworkers, also referred to as non-experts, and are typically recruited through online platforms such as Amazon Mechanical Turk [7] and

[7] https://www.mturk.com/

Prolific [8]. This offers the potential to efficiently conduct a large-scale evaluation (Celikyilmaz et al., 2021). Recruiting participants from these platforms is commonly employed in evaluating automatic text-generation systems (Dhingra et al., 2019; Fang et al., 2015; Zhu, Chen, Haydarov, et al., 2023). However, the online crowdsourcing platform's inaccessibility makes it impossible for under-served communities, such as people with vision impairment, to be reachable (Vashistha et al., 2018).

Small-scale samples, on the other hand, generally involve one to five annotators (van der Lee et al., 2021). Recruiters in this sample size are typically referred to as *experts* and possess deep knowledge in a given domain (Celikyilmaz et al., 2021; Shimorina & Belz, 2021). Doctors, for instance, could evaluate a certain disease prediction system. It is worth noting that recruiting such participants is challenging. First, they are a very small portion of the general population and may be difficult to recruit (Kreiss et al., 2022). The difficulties encountered when recruiting such participants motivate researchers to complement results obtained from expert participants with recruiting large numbers of non-expert participants (Brulé et al., 2020; Hodosh et al., 2013; Kreiss et al., 2022) in order to achieve reliable results.

Our Method

The human evaluation conducted in this research was designed considering the best practice recommendations proposed by van der Lee et al. (2021). In this evaluation study, an intrinsic approach was adopted to assess our model by focusing on the quality of the online shopping descriptions it generated, without considering any other related factors (Shimorina & Belz, 2021) such as other textual data encountered on the shopping websites, including product descriptions and review data. This is the most common approach to evaluating NLG models (van der Lee et al., 2021). In applying this approach, we leveraged the advantages of both quantitative and qualitative methods, using statistical data to evaluate quality while also seeking to understand users' evaluations (Schuff et al., 2023; van der Lee et al., 2021).

We compared our model with the BLIP-2 captioning model (Li, Li, et al., 2023) and GT captions. We chose these reference data mainly because of the limitations of the existing clothing datasets, as demonstrated in section 4.3. The captions generated by BLIP-2 represented a zero-shot model (Li, Li, et al., 2023), that is, one not trained on a clothing caption dataset. The model was applied in recent image captioning research as reference data to assess text automation models (You et al., 2023; Zhu, Chen, Haydarov, et al., 2023). In addition, the BLIP-2 captioning model represents the adoption of the grid-based approach to extract visual descriptions from images. Using the same method, we examined the feasibility of adopting the ROI-based approach in our model compared with the grid-based approach in the BLIP-2 model. On the other hand, GT descriptions are human-generated and are used to train clothing language models to imitate such descriptions. They have been widely used in human evaluations to assess image-text generation models (Devlin et al., 2015; Fang et al., 2015; Zhu, Chen, Haydarov, et al., 2023). Applying GT descriptions (i.e., those used in our CGD dataset) in our evaluation enables the examination of the informativeness and accuracy of the descriptions generated by our model compared with those generated by humans. This allows us to determine whether implementing the LLM in our model facilitates the generation of descriptions better suited for individuals with visual impairment than those produced by

[8] https://www.prolific.com/

humans.

Our evaluation design used a graded scale rating system with a single-item rating to prevent any bias from the identification of the proposed method. The evaluation consisted of either three or four questions. The first question presents only a caption from our model, the BLIP-2 captioning model, or GT. The evaluator was asked to judge whether the caption was *informative* in helping them visualise the graphical elements of the clothing item without seeing the image. This assessment was performed to examine how successful the LLM in our model was in addressing the key components necessary for developing visual comprehension solely based on textual material. It also allowed for the examination of whether our model is feasible for creating detailed informative descriptions of clothing graphical elements. Responses were made using a 4-point Likert-type scale ('very informative', 'informative', 'less informative', or 'not informative') to prevent participants from choosing a neutral or middle value. In addition, the 4-point scale is one of the most commonly employed Likert scales for evaluating text generation systems (Hodosh et al., 2013; van der Lee et al., 2021; Vinyals et al., 2015).

The second question presented the same caption together with the associated image. The *accuracy* of the description was then rated using a 4-point Likert-type scale ('very good', 'good', 'bad', and 'very bad'). The third question asked participants to explain the reasons for their ratings for question two. Finally, the fourth question was asked only if the caption was generated by our model, and the participant rated the accuracy as either 'bad' or 'very bad'. In such instances, participants were asked to fix the caption to highlight the limitations and provide suggestions for enhancing the model. Before conducting the actual study, we carried out a pilot study on a small sample of participants to examine the functionality of our study and to mitigate any potential errors. Examples of the first, second, and third, and fourth questions are shown in Figure 4.14, Figure 4.15 and Figure 4.16.

A pool of 100 items was generated from our new dataset (i.e., the CGD dataset). The pool size had been previously suggested as appropriate for human evaluation of automatically generated text (van der Lee et al., 2019). Individuals with VI and sighted individuals performed the evaluation. One benefit of recruiting sighted evaluators alongside evaluators with VI is that sighted people usually provide useful insights into factors that make good image descriptions (Kreiss et al., 2022). Another benefit is that sighted evaluators can assess the accuracy of captions by referring to the associated images, while VI people may have difficulties doing so. In addition, recruiting sighted participants helps mitigate the challenges of recruiting large numbers of individuals from a low-incidence population (i.e., people with VI) (Brulé et al., 2020; Kreiss et al., 2022). In this study, all participants were presented with the required information and consent forms (see Appendix D). Their consents were collected via an opt-in checkbox included in the survey. Each participant evaluated 25 randomly-selected examples. The sighted evaluators answered all the questions, whereas the evaluators with VI were presented with the first question only, because the other questions required the ability to visually assess accuracy.

Demographic information was also collected. From the sighted evaluators, we gathered only information about their age groups. From the evaluators with VI, we collected information on age groups, levels of perception (i.e., blind or low vision), and types of VI (i.e., from birth or acquired), which was used to improve the replicability of the study and provide more meaningful results (van der Lee et al., 2019). Furthermore, to ensure reliability, we included

two attention-check questions (van der Lee et al., 2021) and removed evaluations where the participants failed to complete both questions correctly. Examples of the attention-check questions are presented in Figures 4.17 and 4.18.

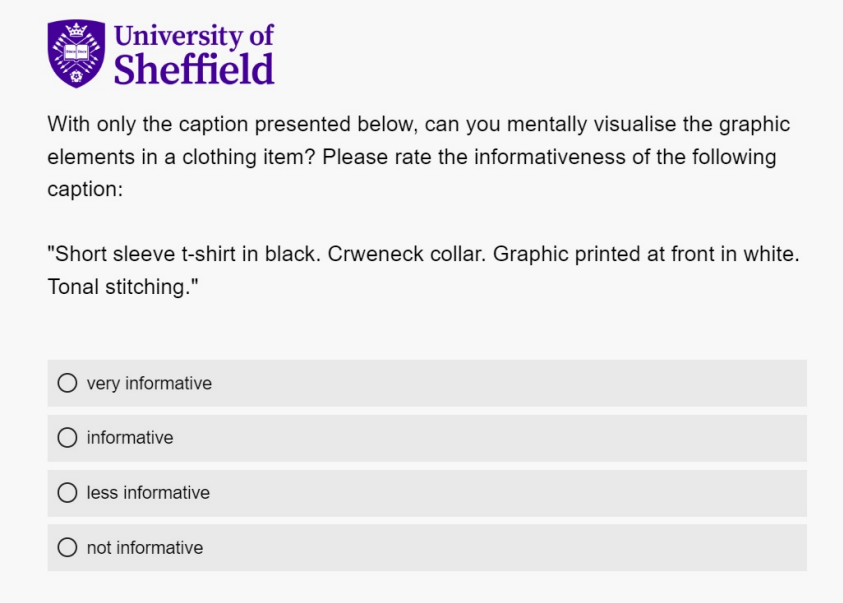


Figure 4.14: *Human Evaluation Interface of an example of Question 1.*

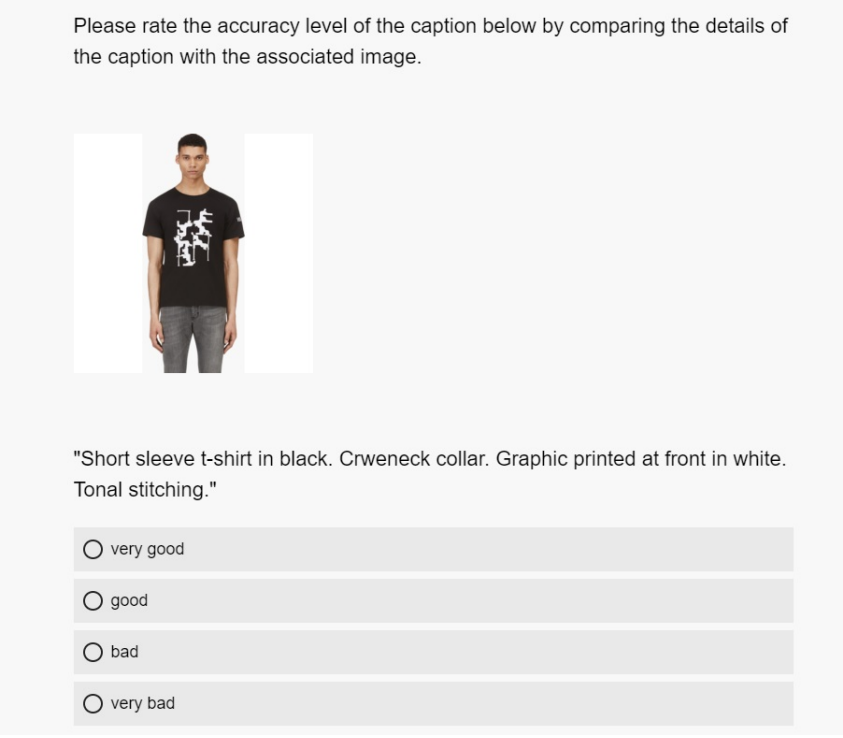


Figure 4.15: *Human Evaluation Interface of an example of Question 2.*

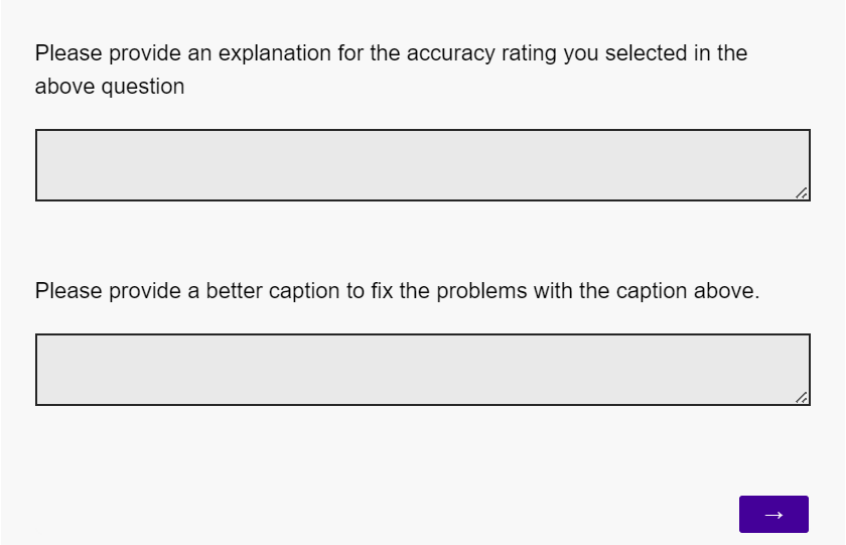


Figure 4.16: *Human Evaluation Interface of an example of Question 3 and 4.*

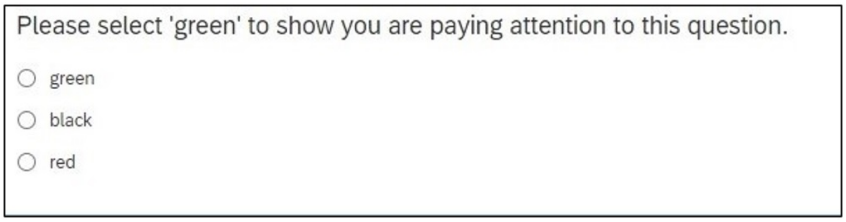


Figure 4.17: *Example of Attention Check Question 1.*

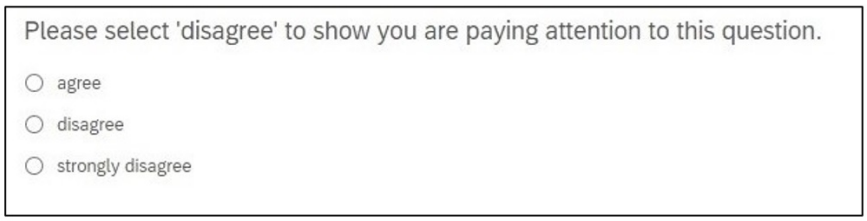


Figure 4.18: *Example of Attention Check Question 2.*

Participant Recruitment A total of 120 sighted evaluators were recruited (age range, 18–74 years). We also recruited 16 individuals with VI (age range, 18–54 years). Three evaluators with VI joined the assessment twice to complete 19 evaluations. This represents the maximum sample size we achieved within the challenges encountered in recruiting VI individuals (Brulé et al., 2020; Kreiss et al., 2022). Seven evaluators had low vision, and nine were blind. Eight evaluators had their VIs from birth, whereas the other eight acquired their visual conditions. Our sighted and VI sample sizes were within the average range (van der Lee et al., 2021). The former were recruited using the Prolific crowdsourcing platform [9].

[9]https://www.prolific.com/

Meanwhile, the individuals with VI were recruited from our contact list of previous participants through snowball sampling. To ensure reliability, each item was evaluated more than once (van der Lee et al., 2021) by at least five sighted evaluators, while this could not be maintained with VI evaluators due to the challenges of recruiting from such a population (Brulé et al., 2020; Kreiss et al., 2022). All sighted evaluators were compensated £4.50 for their time and effort, and evaluators with VI were recruited for £10, equal or higher than the UK living wage [10]. VI individuals received higher compensation compared to sighted evaluators as a token of appreciation for their additional challenges in completing the evaluation task. The evaluators with VI and the sighted evaluators took approximately 30 and 20 minutes on average, respectively, to complete the assessment of 25 images.

Analysis We applied several analytical methods. In our analysis, we transformed the Likert scale to numerical scores, for example for the first question: '3'= 'very informative', '2' ='informative', '1' = 'less informative', and '0'= 'not informative'. Analysis of variance (ANOVA) was applied to analyse our results in the first and second questions. Furthermore, we conducted Tukey HSD tests to further explore the pairwise differences between the methods under comparison (i.e., GT, BLIP-2 and our model). To assess agreement within each group (i.e., sighted and VI individuals), we calculated Krippendorff's alpha. The selection of this metric was based on its flexibility and ability to be used with varying numbers of evaluators (Hayes & Krippendorff, 2007; Marzi et al., 2024). When calculating Krippendorff's alpha test on results obtained from our VI evaluators, raw scoring was used to compare the evaluators' agreement in each method. Items that obtained fewer than two evaluations were removed (the total number of included items is illustrated in subsection 4.6.3). When analysing the results obtained from sighted evaluators, no items were removed because all items received at least five evaluations from different participants. The Spearman correlation test was conducted to analyse the agreement between sighted and VI evaluators when judging informativeness in the first question. This correlation test was chosen because it relies only on a few assumptions. We also qualitatively analysed questions 3 and 4 using content analysis approaches (Vaismoradi et al., 2013). The identified themes included the weaknesses and strengths of each method.

4.6 Findings

This section presents the results obtained from evaluating the proposed model. Preliminary analyses of the results obtained from its components are reported in 4.6.1. This subsection also contains a qualitative comparison of the captions generated from the proposed model and the corresponding GT captions. The results of the automatic and human evaluations are illustrated in subsections 4.6.2 and 4.6.3, respectively. Failure analysis of the proposed model considering the results obtained from the human assessment, is also reported in 4.6.3.

4.6.1 Preliminary Analysis

The results generated by the proposed model were qualitatively analysed with a specific emphasis on determining whether these results effectively addressed VI people's needs, con-

[10] see https://www.gov.uk/national-minimum-wage-rates

sidering the findings presented in Chapter 3. Figures 4.19 and 4.20 show examples of the results obtained from its Graphic Description Generation component (refer to our model architecture in 4.2) using an image from each clothing dataset (i.e., FashionGen, DeepFashion-MultiModal and FACAD datasets). More examples are presented in Appendix F. Although the setup statement generated by the BLIP-2 captioning model provides only abstract descriptions of graphics, the adoption of Automatic Questioning to leverage the LLM's ability to imitate human behaviour permits the extraction of detailed information. When exploring the types of questions produced by the Automatic Questioning task, it was noted that the questions asked related mainly to the kinds of graphic element information that VI people want to know when purchasing clothing online (Alluqmani et al., 2023). An example of such an element is colour, which appears in all three examples (Q2 and Q1 on the left and right sides of Figure 4.19, respectively, and Q4 in Figure 4.20). Shape and position are the other essential elements that VI people require to visualise clothing graphics. These elements are also presented in Q1 and Q2, on the left and right sides of Figure 4.19, respectively. In addition to colour, shape and position, our Automatic Questioning considers asking questions regarding the identification of printed text in the graphic, as depicted in Q4 on the right side of Figure 4.19. These results demonstrate that adopting the LLM in our Automatic Questioning tasks generates questions similar to the queries that VI people, relying solely on textual descriptions, might ask so as to visualise clothing images.

Furthermore, the questions obtained from the automatic questioning task were manually analysed using 100 items randomly selected from the CGD dataset. The automatic questioning task successfully generated five questions all relevant to the enquired-upon clothing graphic for 97 items. The questions for the remaining three examples failed either because of hallucinations (see Figure 4.21) or inaccurate captions from BLIP-2. Figure 4.22 shows an example of an inaccurate description generated from BLIP-2 which fails to explicitly indicate that the described text is printed on the clothing item. This limitation affects the quality of the Automatic Questioning results. The quality of the information generated from the BLIP-2 captioning model affected the unsuccessful examples (more details are reported below). The automatic answering step responded to the questions by leaving a blank space (without providing an answer) in the case of an unknown answer, as shown in the example on the right side of Figure 4.19. As can be seen in Figures 4.19 and 4.20, the automatic summarising step provides a more detailed description of the clothing graphics.

Figure 4.23 compares the captions generated from the proposed model with the corresponding GT data using three items from the FashionGen, DeepFashion-MultiModal and FACAD datasets. The GT captions of all three datasets describe graphics using only a few words, providing only a superficial description that is unlikely to be sufficiently detailed or specific to be helpful to VI users. For example, the phrases 'graphic printed' and 'hand-drawn graphic', which do not provide detailed descriptions of graphic elements, are used to represent graphics. In contrast, while the captions generated by the proposed model provide more fine-grained and detailed information to describe the graphics, some aspects may not be fully accurate (see the description in Figure 4.20). These results demonstrate the lack of graphic descriptions in the GT data of the clothing datasets.

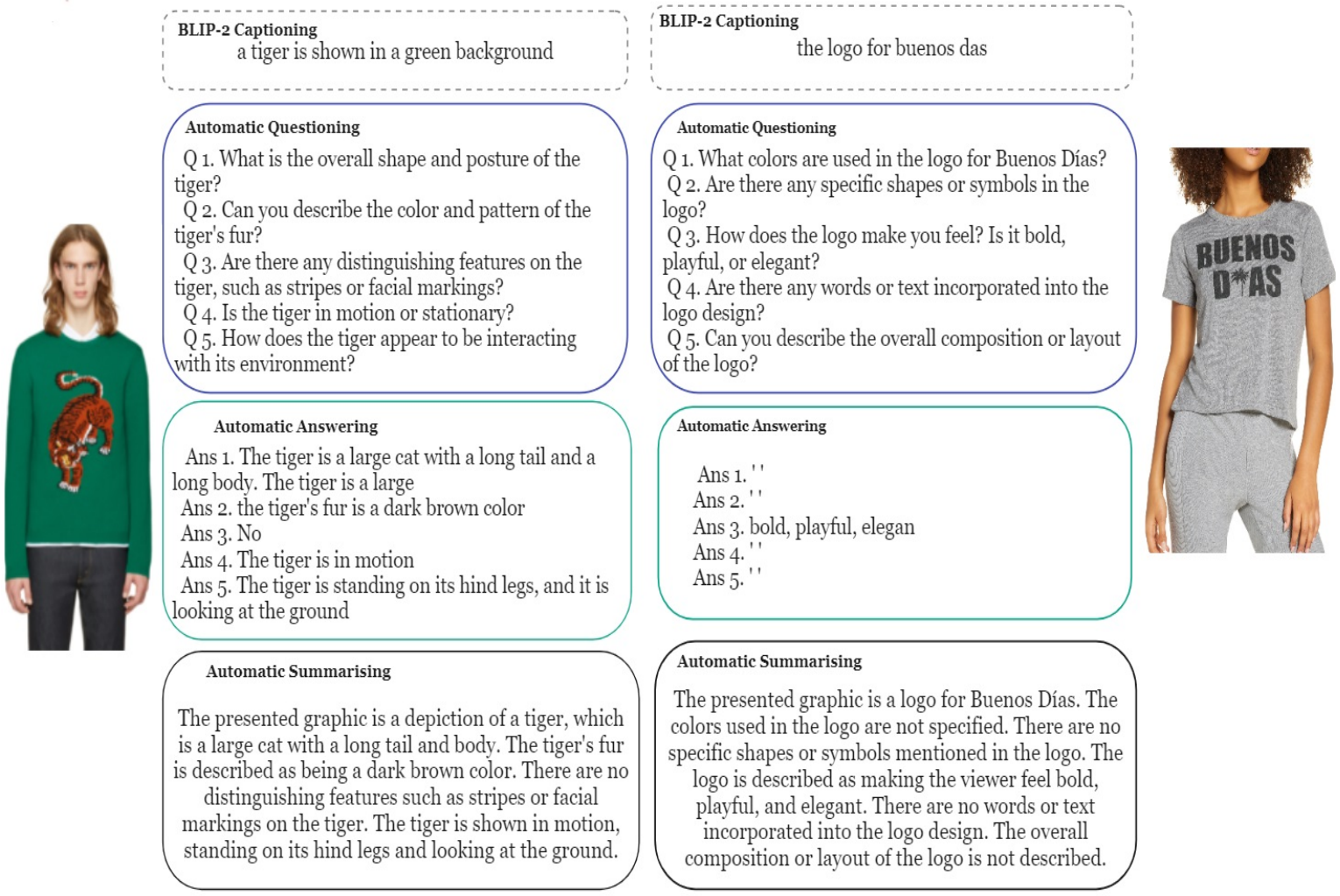


Figure 4.19: *Examples of results obtained from the graphic description generation component of the proposed model (including BLIP-2 captions, Automatic Questioning, Automatic Answering and Automatic Summarising) using images from the FashionGen (left) and FACAD (right) datasets.*

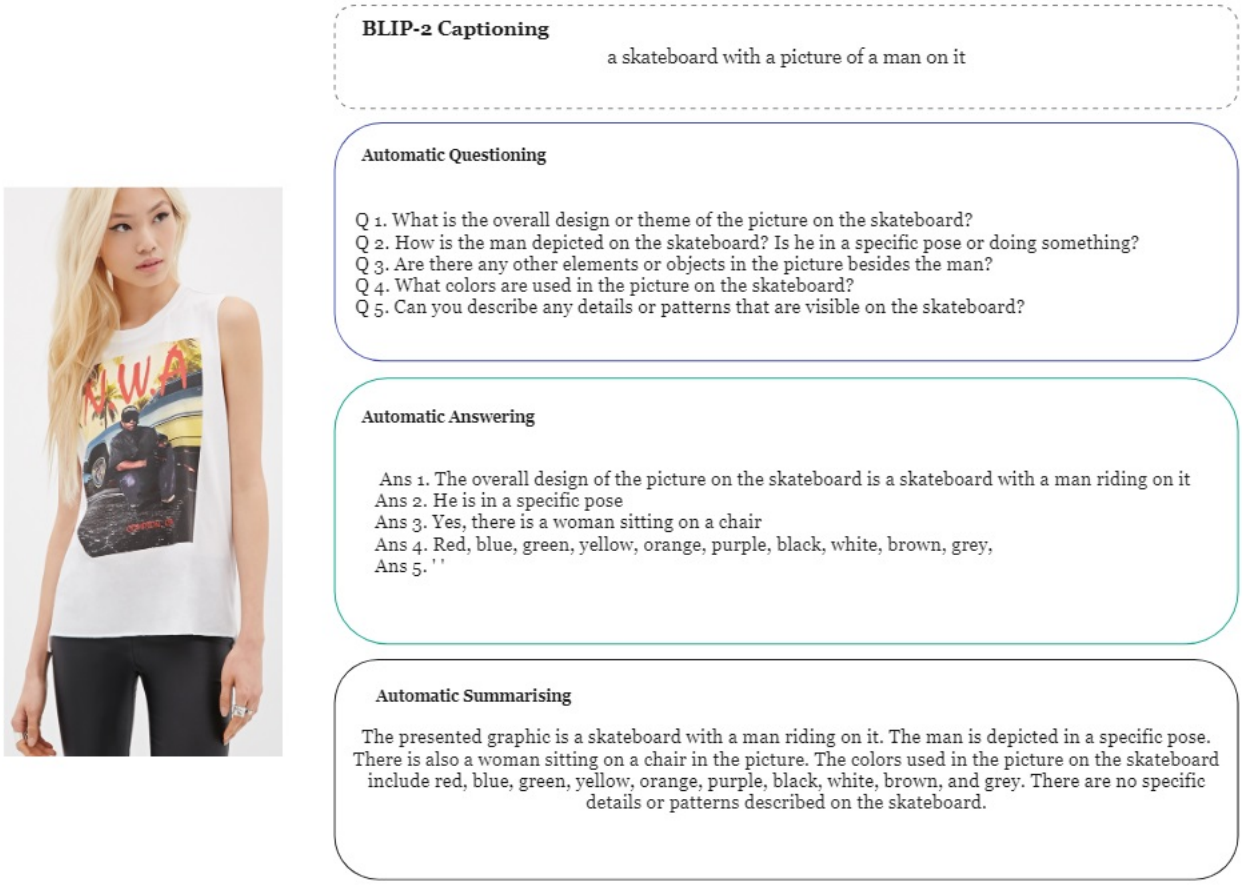


Figure 4.20: *Examples of results obtained from the graphic description generation component of the proposed model (including BLIP-2 captions, Automatic Questioning, Automatic Answering ,and Automatic Summarising) using an image from the DeepFashion-MultiModal dataset.*



BLIP-2 Captioning
A pink and white sign that says splash

Automatic Questioning
Q 1. What is the shape of the sign that says "splash"?
Q 2. How are the words "splash" and 'Avoid" arranged on the sign?
Q 3. Are the letters in the sign written in a specific font or style?
Q 4. What is the size of the sign? Is it small or large?
Q 5. Are there any additional elements or designs on the sign, apart from the words "splash" and 'Avoid"?

Automatic Answering
Ans 1. a
Ans 2.
Ans 3. Yes
Ans 4. small
Ans 5. Yes, there are additional elements on the sign, apart from the words "splash" and "

Automatic summarising
The presented graphic is a small sign in the shape of the letter "a" that says "splash." The words "splash" and "Avoid" are arranged on the sign, but the specific arrangement is not described. The letters in the sign are written in a specific font or style. There are additional elements on the sign, apart from the words "splash" and "Avoid," but these elements are not described .

Figure 4.21: *Example of results generated by the Automatic Questioning task fails to generate relevant questions about the inquired item due to the hallucination problem (highlighted in red boxes).*

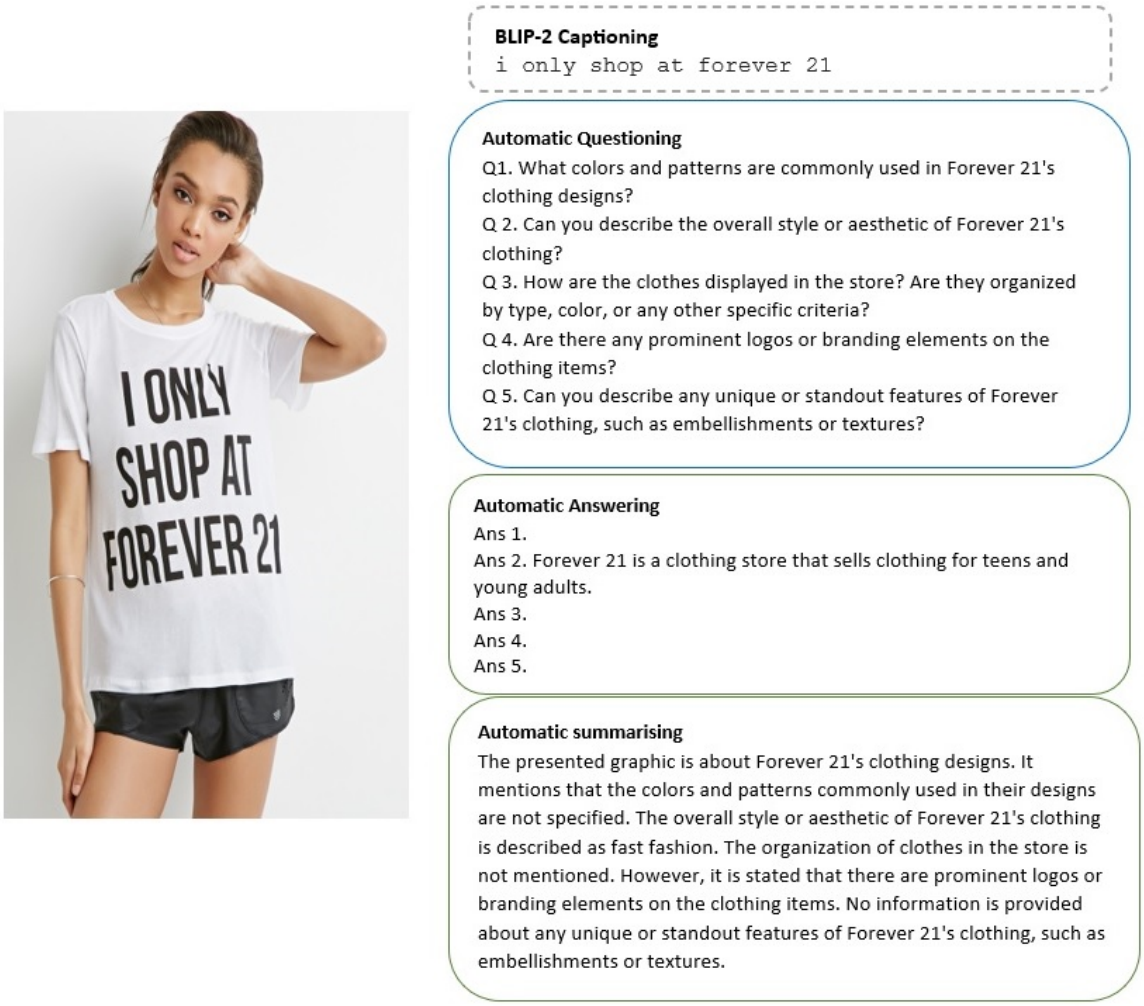


Figure 4.22: *Example of inaccurate descriptions generated from BLIP-2 affecting the quality of the Automatic Questioning results.*



"Tank top in black. Ribbed scoopneck collar and oversized armscye. Multicolored mohawk grpahic print at front. Tonal stitching."

"The presented graphic is of a cartoon character wearing a yellow hat. The hat is positioned on the character's head and does not have any specific design or pattern on it. It is plain and does not have any accessories like feathers or ribbons. The shape or style of the hat is not described."



"Long sleeve French terry sweatshirt in 'darkest' black. Rib knit crewneck collar, cuffs, and hem. Graphic printed in tones of pink at front. Tonal stitching."

"The presented graphic is of a woman standing and singing into a microphone. She is holding the microphone with both hands and appears to be fully engaged in her performance. The expression on her face suggests that she is passionate and immersed in her singing. Additionally, there are other elements or objects in the sticker that indicate the setting or atmosphere of the scene, although they are not described in detail. The microphone she is using has a distinct style or appearance, although no specific details are provided."



"bring a little taste of paradise to everyday life with this dreamy soft cotton crewneck sporting a throwback palm graphic"

"The presented graphic is a soft pink t-shirt with realistic, detailed illustrations of scattered palm trees. The sunset on the t-shirt is depicted in a more subtle manner, and there are no other elements or designs present apart from the palm trees and sunset."



"this cozy vintage inspired pullover feature a hand drawn graphic and extra soft fleece that make it a top you ll never want to take off"

"The presented graphic is a representation of Lake Tahoe, a large lake in the Sierra Nevada mountain range in California. The graphic typically features various shades of blue and green to depict the water of the lake. There are no specific descriptions of prominent landmarks or features that are commonly depicted in the graphic. The surrounding landscape of Lake Tahoe is often included in the graphic, showcasing the mountains and forests that surround the lake. The overall composition or layout of the graphic is not described in the given information."



"The lady is wearing a sleeveless tank shirt with graphic patterns. The tank shirt is with cotton fabric. The neckline of the tank shirt is round. The lady wears a ring."

"The presented graphic is the Miami Heat logo. The overall shape of the logo is circular, with a basketball in the center. The colors used in the logo are red, yellow, and black. There are no specific symbols or images within the logo other than the basketball. The different elements of the logo, such as the basketball and text, interact with each other by being placed within the circular shape. There is typography present in the logo, with the text \"Miami Heat\" written in bold, uppercase letters."



"The shirt this female wears has short sleeves, its fabric is cotton, and it has graphic patterns. The shirt has a round neckline. This female wears a three-point shorts, with denim fabric and pure color patterns. There is an accessory on her wrist. There is a ring on her finger."

"The presented graphic is the logo for MTV Music Television. It uses the colors black and white. There are no specific shapes or symbols incorporated into the logo. The logo visually represents music and television through the text \"Music Television.\" There are no distinct patterns or textures in the logo design. The overall composition or layout of the logo is not described."

Figure 4.23: *Examples of the lack of graphic descriptions in the ground-truth captions of two items each from three fashion captioning datasets: FashionGen (top), FACAD (middle), and DeepFashion-MultiModal datasets (bottom). The ground-truth captions are in the boxes with dotted borders, and the descriptions generated from the proposed model are in the blue boxes.*

4.6.2 Automatic Evaluation

Image Classification

Here, we present the results of the classification test using the FashionCLIP model on the FashionGen dataset (details on this dataset were discussed in subsection 4.3.1). As mentioned in subsection 4.5.3, the rationale for choosing the FashionCLIP model rather than other classification models, such as FashionNet (Liu et al., 2016), is because this is the most recent advanced classification model leveraging the power of the zero-shot CLIP model, and it has been fine-tuned with fashion examples. The details of the evaluation method are discussed in subsection 4.5.3.

Figure 4.24 shows the classification results of clothing items as either "graphic" or "non-graphic". The results demonstrate several failure cases, including inaccurately classifying 64 (13%) of the non-graphic clothing as having graphics and failing to recognise three instances (25%) of clothing with graphics. Notably, considering the very few instances of graphic images in the dataset, the accuracy of predicting non-graphic images is very high, while when predicting graphic images, the precision is extremely low (i.e., many false positives). To investigate the cases that were not categorised correctly, we conducted a failure analysis. When exploring the failure cases, we identified misclassification errors in the GT data. Figure 4.25 presents examples of items incorrectly identified as non-graphic clothing in the GT data, although graphics are clearly shown in the clothing images. Another misclassification example is presented in Figure 4.26, which presents an incorrectly classified graphic that appears to be a pattern embroidered on the fabric. This item shows patterns of colours and shapes rather than describable objects, such as characters, landmarks or printed texts. These results indicate that FashionGen's GT classification data are not appropriate for testing the graphic classification task.

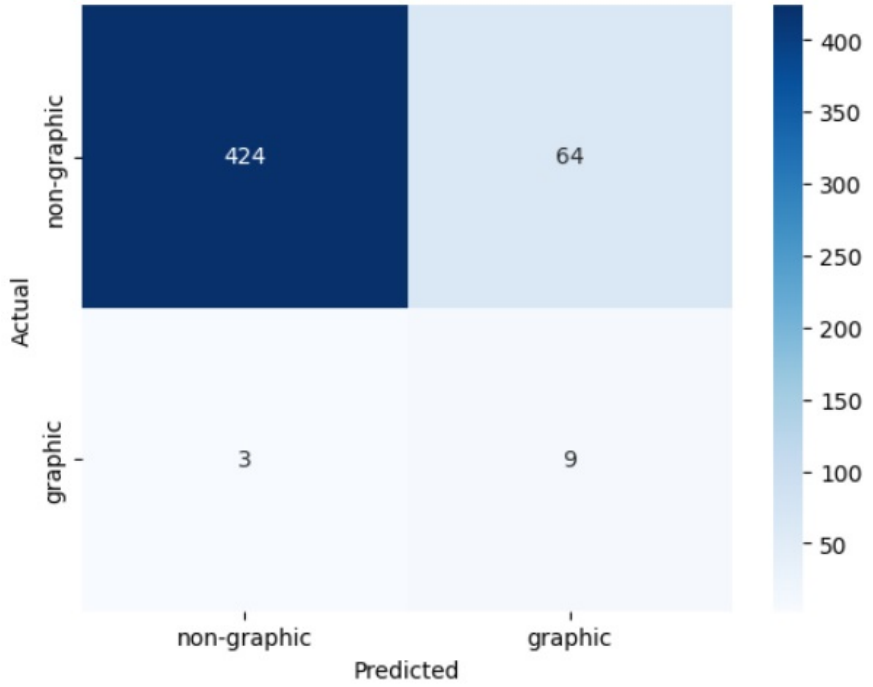


Figure 4.24: *Classification results of the FashionCLIP model using the Fashion-Gen dataset.*



Figure 4.25: *Examples of items incorrectly classified as non-graphic in the GT data (FashionGen).*



Figure 4.26: *Example of item incorrectly classified as graphic clothing in the GT data (FashionGen).*

To overcome this limitation, and to use a more reliable dataset as a baseline for our assessment, we created a new graphic clothing classification dataset. Considering VI people's needs to mentally visualise clothing items (Alluqmani et al., 2023), we based our classification on our definition of graphic elements (i.e., the inclusion of any drawings, images present or texts). Our classification dataset includes 162 graphic items and 300 non-graphic items collected from the FashionGen dataset (items may not be part of the sample used in the previous experiment). All items are randomly selected and manually classified. The FashionGen dataset was used as reference data to assess the FashionCLIP classification model.

Figure 4.27 indicates that, compared to the FashionGen dataset, our classification dataset shows notable improvement in the FashionCLIP model. While the results revealed an error of 42 items in accurately identifying graphical elements, we applied an error analysis. However, no clear results can be identified from these errors. Examples of unsuccessfully identified graphical elements are presented in Figure 4.28. Nonetheless, the overall results in Figure 4.27 demonstrate significant success in generating accurate predictions.

A summary of the FashionCLIP performance using the FashionGen dataset and our classification dataset is shown in Table 4.2. Note that redefining image classes in our classification dataset slightly increased the accuracy. When our dataset was used, the model showed significantly higher precision, indicating high accuracy in recognising graphic elements. As a result, our dataset demonstrates a far better F1 score. This reveals the remarkable effectiveness of

the FashionCLIP model in accurately identifying graphic elements when employing our definition of graphic elements on clothing items, and suggests that the FashionCLIP model is a good fit for the early stages of the proposed model.

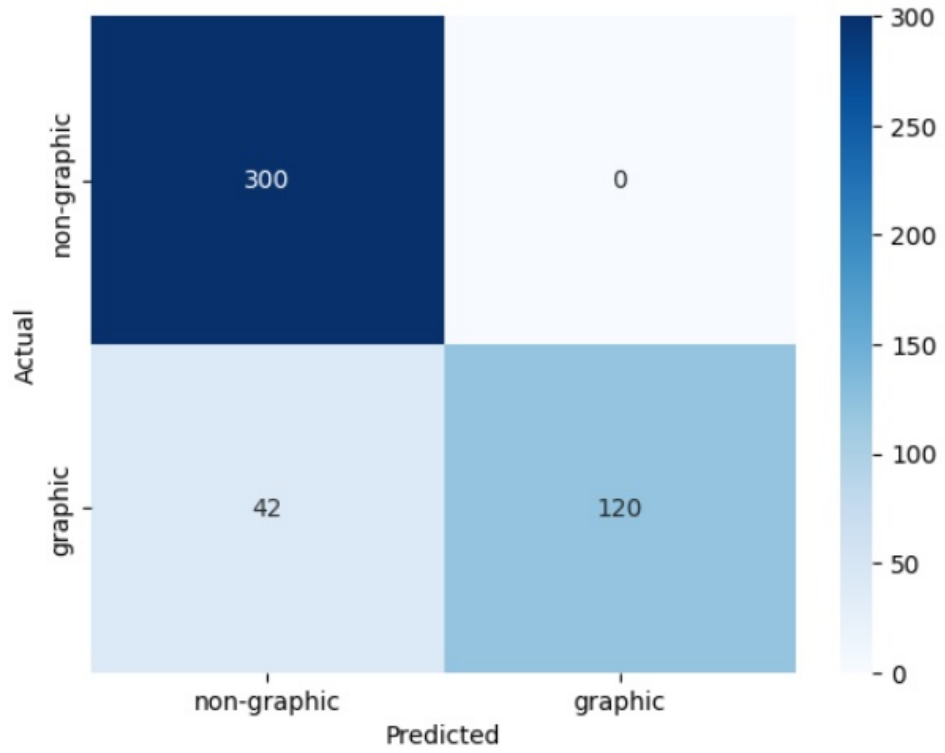


Figure 4.27: *Classification results of the FashionCLIP model using our classification dataset.*

Table 4.2: *Classification results of the FashionCLIP model using FashionGen and our classification datasets.*

Datasets	**accuracy**	**precision**	**Recall**	**F1**
FashionGen	0.86	0.12	0.75	0.21
Our graphic classification	0.90	1.0	0.74	0.85



Figure 4.28: *Example of items incorrectly classified as non-graphics by the FashionCLIP model.*

Image Captioning

We assessed the performance of the proposed model in comparison to the BLIP-2 captioning model. To ensure that our evaluation was comprehensive, we individually applied the assessment to each dataset (i.e., FashionGen, DeepFashion-MultiModal, and FACAD) encompassed within our CGD dataset. Details of the CGD dataset are discussed in 4.3. Additionally, we conducted a two-tailed t-test on the results from each metric to assess the effectiveness of the results. This will provide essential information regarding whether the difference between the captions generated by the proposed model and the BLIP-2 model is statistically significant. Further, we adopted a set of standard automatic evaluation metrics commonly used in image captioning: BLEU, ROUGE, METEOR, CIDEr and SPICE. Details of our evaluation method are given in subsection 4.5.3.

		BLEU-1	**BLEU-4**	**METEOR**	**ROUGE L**	**CIDEr-D**	**SPICE**
FashionGen	Our Model	0.070*	3.463e-07	0.052*	0.073*	0.002*	0.038*
	BLIP-2(baseline)	0.057	0	0.069	0.088	0.020	0.018
DeepFashion-MultiModal	Our Model	0.200*	0	0.085*	0.164*	0.011*	0.040*
	BLIP-2(baseline)	0.038	1.907	0.048	0.107	0.002	0.052
FACAD	Our Model	0.095	2.816e-10	0.068*	0.107*	0.001*	0.041
	BLIP-2(baseline)	0.105	2.292e-07	0.051	0.125	0.051	0.044

Table 4.3: *Comparison of the results obtained by our model and BLIP-2 in a set of automatic image captioning metrics using three fashion captioning datasets. An asterisk (*) indicates that the P-value is less than 0.05, denoting statistical significance.*

Results of our automatic evaluation are shown in Table 4.3. For the FashionGen dataset, our model scored higher in BLEU-1 (.070) and SPICE (.038) than did BLIP-2 (.057 and .018, respectively). Both metrics showed a statistically significant difference. This result indicates that our model is superior to the BLIP-2 model in terms of the BLEU-1 and SPICE metrics. and suggests that our model generates more unigrams and object-attribute relationships that are similar to FashionGen captions than does the BLIP-2 model. Figure 4.29 shows two examples of clothing images and captions generated from our model, BLIP-2 as well as their associated GT captions. It can be noted that captions generated from our model use more similar words to the GT data compared to BLIP-2's captions. For instance, in the right side item, the caption generated from our model uses similar words to the GT captions like "graphic", "pink", "shirt" and "front". While BLIP-2 uses less similar words like "black". The item on the left side shows captions of our model uses more similar words to the GT caption compared to the BLIP-2. The former uses "graphic", "of", "black", and "with" similar to the GT captions while the latter uses only the "t-shirt" word. These exemplified the higher score our model received compared to BLIP-2 in the BLEU-1. These results might be because our model generates captions with more words than BLIP-2 which leads to more chance of producing similar words and more word relationships.

In DeepFashion-MultiModal, our model outperforms the BLIP-2 model in the BLEU-1, METEOR, ROUGE L and CIDEr-D metrics, achieving scores of .200, .085, .164 and .011, respectively. BLIP-2 obtained scores of .038, .048, .107 and .002 in the same metrics, respectively. As depicted in Table 4.3, the results of BLEU-1, METEOR, ROUGE L and CIDEr-D metrics indicate statistically significant differences. Examples in Figure 4.30 exemplify the higher score of BLEU-1 received by our model compared to the BLEP-2. The examples show that captions generated from our model present more similar words to the GT captions compared to the BLIP-2 captions. For instance, the item on the right side uses more than four similar words (e.g., "graphic", "with", "t-shirt" and "with") while BLIP-2 shows one occurrence of similar words (i.e., "is"). These results show that our model generates captions that are more similar to DeepFashion-MultiModal captions in terms of word choice, overall caption structure and fluency than BLIP-2 does.

When using FACAD, our model achieved a slightly higher score in the METEOR metric (.068) compared with BLIP-2 (.051). Examples of items from the FACAD and captions from our model, BLIP-2 and GT data. The results show that these scores are statistically significant. The results show that our model generates semantic meanings that are slightly more similar to FACAD captions than BLIP-2 does.

In general, the automatic image captioning results demonstrate that our model performs similarly to DeepFashion-MultiModal data compared to FashionGen and FACAD data. When adopting BLEU-1, which assesses unigram similarity using precision evaluation, our model outperforms BLIP-2 using FashionGen and DeepFashion-MultiModal data. The METEOR metric assesses similarity by applying stemming and synonym matches using precision and recall evaluations. This shows that our model generates words and semantic meanings that are more similar to the DeepFashion-MultiModal and FACAD data than BLIP-2 does. ROUGE-L, on the other hand, measures similarity using the LCS technique to provide insight into the structural similarity between automatic and reference captions. This shows that our model generates captions with a structure that is closer to that of DeepFashion-MultiModal captions. CIDEr-D measures similarity in terms of generating diverse words and considering salient el-

ements. This shows that our model's captions are more similar to DeepFashion-MultiModal's captions than those of BLIP-2. SPICE shows the similarity of object identification and relationship in FashionGen.

Our model generally indicates more similar captions to the GT captions compared to the BLIP-2 model. This similarity does not necessarily mean that the captions generated from our model produce similar content or contain the same limitations as the GT captions (see our discussion in 4.3). Another point is that automatic image captioning evaluation does not assess the level of informativeness and accuracy of the generated captions. Thus, human evaluation was required to provide a better assessment of our model.

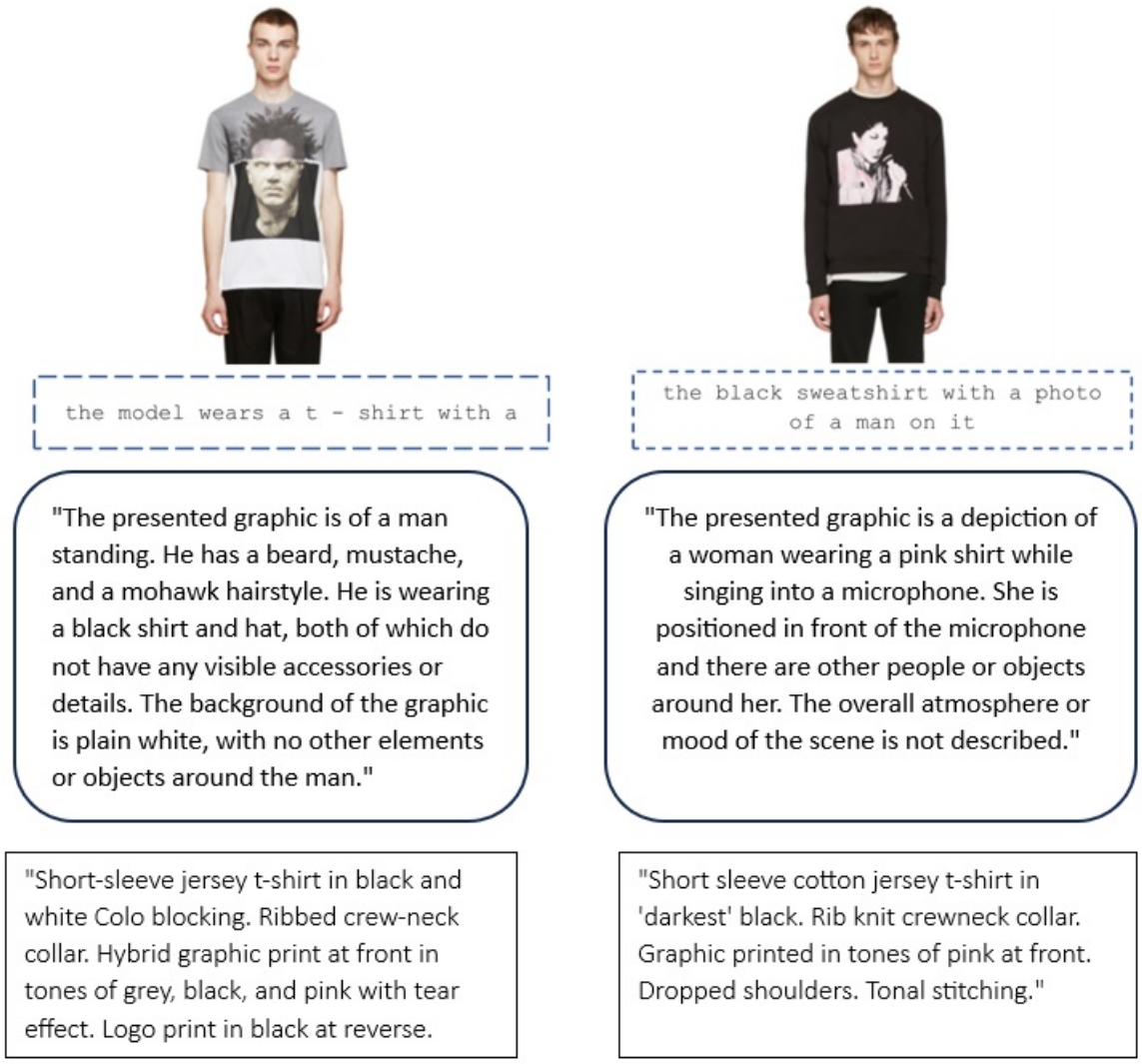


Figure 4.29: *Examples of two clothing items from the FashionGen dataset, comparing the rich, informative captions obtained from our model (middle) with captions generated from BLIP-2 (top) and GT captions (bottom).*



Figure 4.30: *Examples of two clothing items from the DeepFashion-MultiModal dataset, comparing the rich, informative captions obtained from our model (middle) with captions generated from BLIP-2 (top) and GT captions (bottom).*

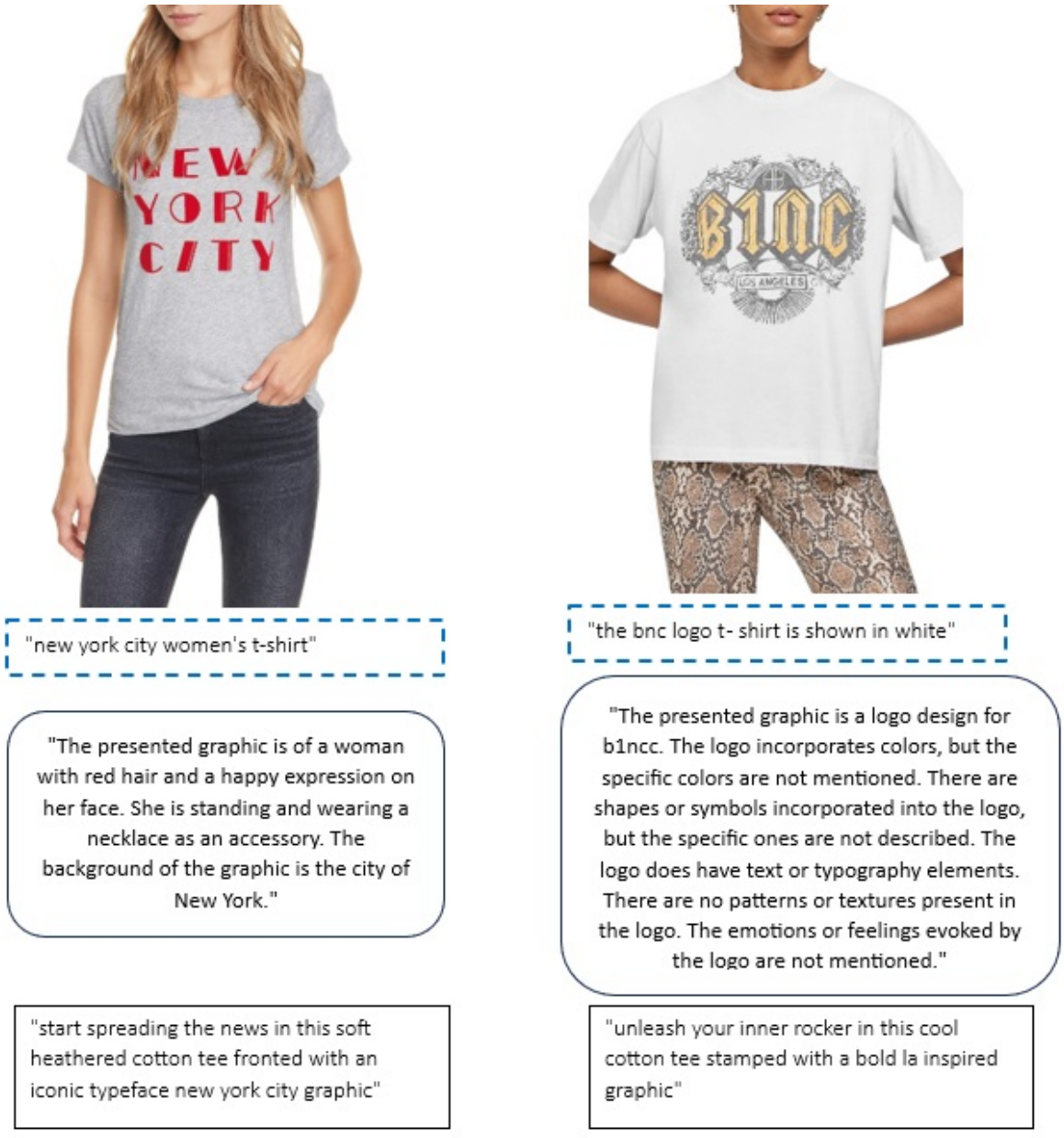


Figure 4.31: *Examples of two clothing items from the FACAD dataset, comparing the rich, informative captions obtained from our model (middle) with captions generated from BLIP-2 (top) and GT captions (bottom).*

4.6.3 Human Evaluation

In this evaluation, we chose a within-subjects study design that used 120 sighted and 16 VI evaluators to assess the informativeness and accuracy of the image captions generated from 100 items using three methods (GT data, BLIP-2 and our model).

Here, the evaluation results are reported in several areas. We discuss the results of the VI and sighted evaluators' judgments of the informativeness of the image captions. Next, we

discuss the results obtained from our sighted evaluators, who assessed the accuracy of the image captions. We interpret these quantitative results by discussing the qualitative results. In addition, we discuss the results of the failure analysis of the proposed model considering the limitations reported by the sighted evaluators. The details of our evaluation design are illustrated in 4.5.

Table 4.4 shows a statistical summary of the total number of evaluators (including VI and sighted) who judged the informativeness and accuracy of the image captions. In the informativeness assessment, each item was judged by at least six evaluators, and a maximum of 17 participants assessed each item. As only sighted evaluators judged the accuracy of the captions, the number of assessors was lower than in the informativeness analysis (i.e. a minimum of five and a maximum of 15). The statistical summary shows that the informativeness of each image caption was assessed by approximately 11.66 evaluators, with moderate variability (SD = 2.59). The accuracy of each image caption was assessed by approximately 10.08, with moderate variability (SD = 2.76).

Statistic	Evaluators	
	Informativeness	Accuracy
Min	6	5
Max	17	15
Mean (SD)	11.66 (2.59)	10.08 (2.76)

Table 4.4: *A summary of the statistical results indicating the number of evaluators assessing the informativeness and accuracy of each image captions.*

Informativeness Evaluation

To assess the informativeness of the image captions, the VI and sighted evaluators were asked to assess their ability to visualise the graphics of the clothing items based on the captions alone. Further details of the evaluation method are included in 4.5.

Visually Impaired Evaluators The assessment results of our VI evaluators in terms of the *informativeness* of the captions are shown in Figure 4.32, which compares the scores, as percentages, obtained for all methods across four different evaluation categories: 'Very Informative', 'Informative', 'Less Informative' and 'Not Informative'). Our model achieved the highest (or among the highest) scores in the 'Very Informative' (30.60%) and 'Informative' (33.30%) categories. In contrast, the BLIP-2 and GT methods both obtained under 30% in the 'Very Informative' category and 33.30% and 24.70% respectively in the 'Informative' category. The boxplots in Figure 4.33 compare the average scores and the score distributions achieved through all three methods; the scores of all three varied widely, reaching the scale's upper and lower ends. Nonetheless, Table 4.5 shows a clear difference between our model's mean score and that of the GT and BLIP-2 methods. Our model was more successful than the other methods, scoring 1.737, with a standard deviation (SD) of 0.927, while the GT and BLIP-2 obtained similar average scores, at 1.440 (SD=0.875) and 1.460 (SD=0.764), respectively.

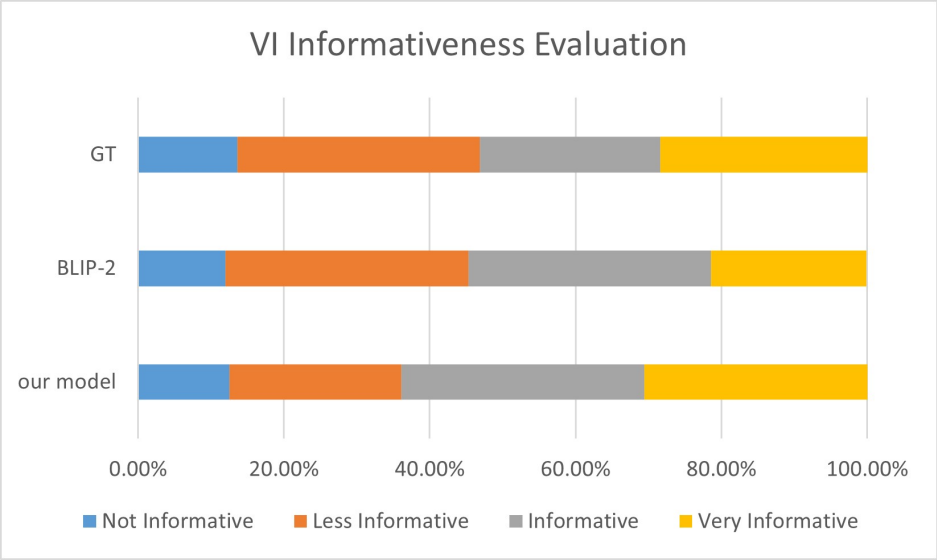


Figure 4.32: *The results of the VI evaluators' assessment of informativeness, comparing GT, BLIP-2 and our model. Our model achieved the highest score in the 'Very Informative' category and is among the top-rated in the 'Informative' category.*

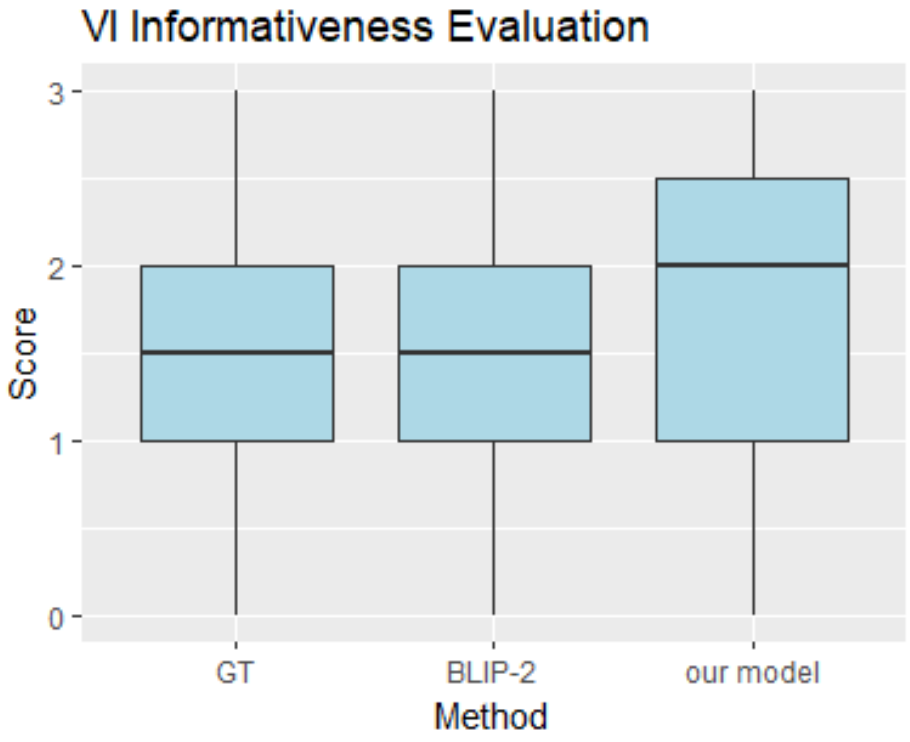


Figure 4.33: *Boxplot illustrating the average score given by VI evaluators' assessment of the informativeness of captions produced by GT, BLIP-2 and our model.*

Methods	Mean (SD)
GT	1.440 (0.875)
BLIP-2	1.460 (0.764)
Our Model	1.737 (0.927)

Table 4.5: *Means and standard deviation (SD) results of the VI informativeness assessment for the three methods (i.e., GT, BLIP-2 and our model).*

To investigate whether these differences were statistically significant, we applied an ANOVA test between our model and the BLIP-2 and GT methods among all items. Average scores for each item per method were computed. Results show that the difference between the means of all the methods was significant (p=0.0264; F=3.681). We also conducted a Tukey HSD test to explore the pairwise differences between the methods. The results indicated that, while our model's average score was significantly different from that of the GT (p=0.041), it was not significantly different from that of the BLIP-2 (p=0.05). These results mean that, compared to GT captions, VI evaluators found our model's captions of graphic elements to be more informative for visualising clothing graphics with no human assistance.

Agreement of VI Evaluators' Assessments We explore the agreements between the VI evaluators when judging items in each method (i.e., GT, BLIP-2 and our model) by calculating Krippendorff's alpha. The results are presented in Table 4.6. In this assessment, the number of items used varied because items that received fewer than two scores were removed. The number of items used for GT was 57, while our model used 55 and BLIP-2 used 61. It can be observed that the best agreement was obtained in the evaluation of our model, with a score of 0.311. Our VI evaluators agreed less when evaluating BLIP-2 and GT captions, which scored 0.188 and 0.226, respectively.

	# of Items	α VI informativeness
GT	57	0.226
BLIP-2	61	0.188
our model	55	0.311

Table 4.6: *Krippendorff's Alpha Coefficient for VI Evaluators' Agreement When Judging the Informativeness of Captions Extracted by Each Method (GT, BLIP-2 and our model).*

Sighted Evaluators The results of our sighted evaluators' assessments of the *informativeness* of our model and the BLIP-2 and GT methods are compared in Figure 4.34. Notably, our model was generally superior to the BLIP-2 and GT captions. Over half of the participants who were shown captions from our model found our model to be either 'Very Informative' (n=262; 25.8%) or 'Informative' (n=407; 40.1%). The GT and BLIP-2 captions, on the other hand, received lower scores (17.8% and 11.8%, respectively) in the 'Very Informative' category. The three caption methods had very similar results in the 'Informative' category. Figure 4.35 presents boxplots of the sighted evaluators' assessments of the informativeness of our model and the BLIP-2 and GT methods. As illustrated in Table 4.9, our model and GT achieved the highest average maximum ratings (2.667), while BLIP-2 obtained the lowest

(2.500). The minimum average score obtained by our model (0.545) is very similar to that obtained by the GT method (0.571). BLIP-2, on the other hand, obtained the lowest minimum average score (0.400). If we compare the average scores among the three methods, our model achieved the highest performance, with an average score of 1.774 (SD=0.462), compared to 1.671 (SD=0.453) and 1.565 (SD=0.449) obtained by GT and BLIP-2, respectively.

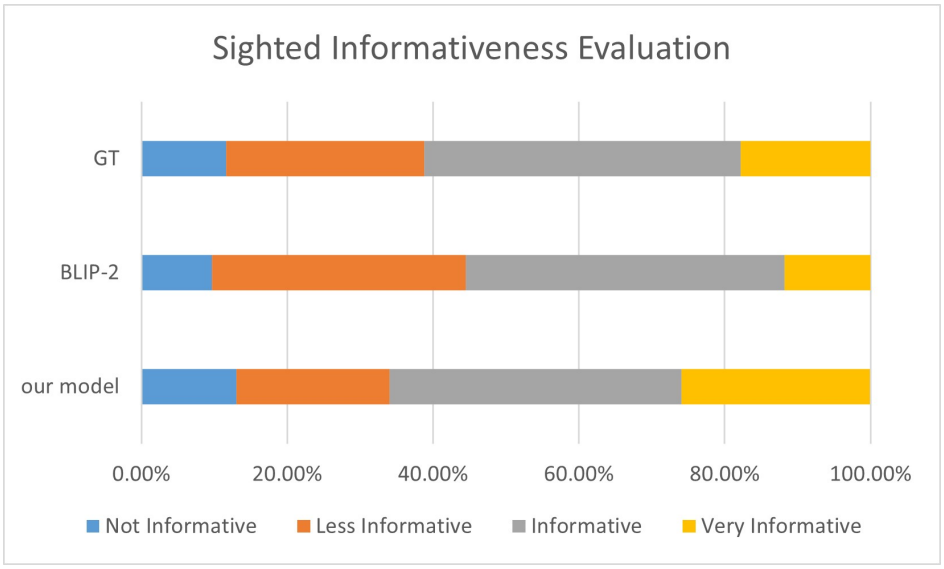


Figure 4.34: *Comparison of the results of our sighted evaluators' assessments of the informativeness of captions generated by GT, BLIP-2 and our model. Our model achieved the highest score in the 'Very Informative' category and was comparable to the other methods in the 'Informative' category.*

As in the VI evaluators' data analysis, we conducted an ANOVA test to explore whether the difference between the methods' means was statistically significant. The statistical results indicate that the mean difference between the three methods is statistically significant (p=0.0057, F=5.258). We also conducted a Tukey HSD test to explore the pairwise differences between each of the methods. The difference between our model and BLIP-2 is statistically significant (p=0.003), but the mean differences between GT and our model and between GT and BLIP-2 are not significant. These results suggest that the captions generated by our model are more informative than those generated by BLIP-2 in terms of creating visual pictures of the graphic elements of clothing without looking at the clothing image.



Figure 4.35: *Boxplot illustrating the average score given by sighted evaluators' assessing the informativeness of captions produced by GT, BLIP-2 and our model.*

Methods	Mean (SD)
GT	1.671 (0.453)
BLIP-2	1.565 (0.449)
Our Model	1.774 (0.462)

Table 4.7: *Mean and standard deviation (SD) results of the sighted informativeness assessment for the three methods (i.e., GT, BLIP-2 and our model).*

Agreement of Sighted Evaluators' Assessments To assess the agreement of our sighted evaluators' judgment of the informativeness of the image captions, we calculated Krippendorff's alpha. The results in Figure 4.8 show low agreement between the sighted evaluators. The agreement on judging the GT captions was 0.171, while the level of agreement was 0.19 and 0.126 when judging BLIP-2 and our model, respectively. These results show a significant inconsistency in how each evaluator perceived and rated the informativeness of the captions. To further understand these disagreements, we explored the comments and interpretations provided by the sighted evaluators.

	α sighted informativeness
GT	0.171
BLIP-2	0.19
our model	0.126

Table 4.8: *Krippendorff's Alpha Coefficient for sighted Evaluators' Agreement When Judging the Informativeness of Captions Extracted by Each Method (GT, BLIP-2 and our model).*

Agreement of VI and sighted assessments When assessing the three methods (i.e., GT, BLIP-2 and our model), Spearman's rank correlation was calculated between VI and sighted evaluators. The VI and sighted judgments obtained a moderate correlation of r(97) = .46 when assessing our model, but a weaker correlation was found when assessing BLIP-2 and GT, with r(95) =.38 and r(96)= .32, respectively. The results show significant positive correlations (p-value < 0.05) between the VI and sighted evaluators across all three methods.

To assess the informativeness of our model, we explore the disagreements between our VI and sighted evaluators. To begin with, we observed that, in contrast to sighted evaluators, those with visual impairments appreciated detailed descriptions of the clothing graphics. For instance, they remarked on the usefulness of details such as those recognising that some objects in the clothing graphic are unclear or those confirming that there are no other elements presented in the graphic apart from those already described. Figure 4.36 presents an example of a caption stating that there are no specific patterns or symbols in the graphic, which VI evaluators found to be extremely informative, while sighted evaluators disagreed.

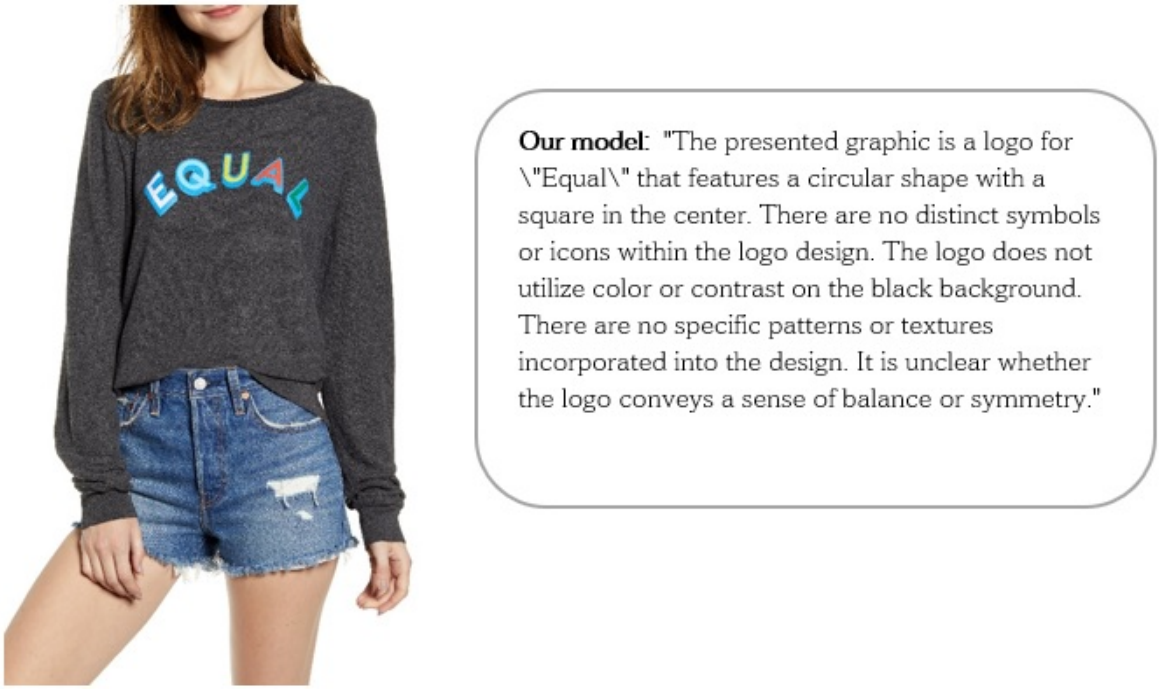


Figure 4.36: *An example of a caption for which VI and sighted evaluators disagreed. The former found it to be highly informative, while sighted evaluators did not.*

Accuracy Evaluation

We assessed the *accuracy* of the image captions by asking the sighted evaluators to compare the captions' content with the associated images. More details of our evaluation design are discussed in 4.5. Figure 4.37 presents the overall percentage of results produced by each of the three methods in four categories (i.e., 'Very Bad', 'Bad', 'Good' and 'Very Good'). Our model received only 11.1% of the 'Very Good' votes, while the GT and BLIP-2 captions received slightly higher scores (17.9% and 17.2%, respectively). Our model received the highest percentage of 'Very Bad' votes for a total of 312 (i.e., 30.8%). The boxplots in Figure 4.38 provide an overall comparison of the accuracy scores obtained for our model, BLIP-2 and GT. Our model and BLIP-2 received the same maximum and minimum scores of 2.667 and 0, respectively. In contrast, GT received the highest maximum (2.800) and minimum (0.3) scores. Our model had the lowest mean score at 1.199 (SD = 0.627), with BLIP-2 having a mean of 1.653 (SD = 0.569) and GT having a mean of 1.629 (SD = 0.486).

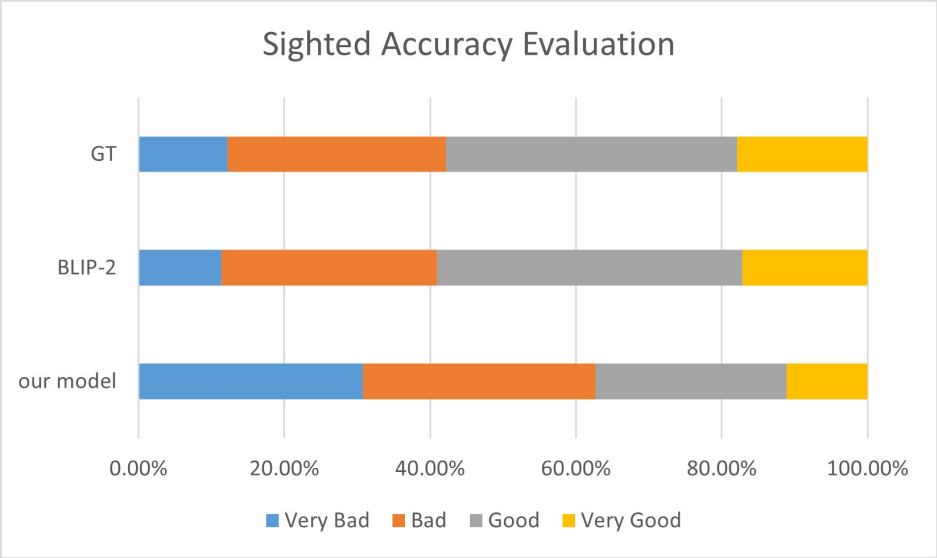


Figure 4.37: *Comparison of the results of our sighted evaluators' assessments of the accuracy of captions generated by GT, BLIP-2 and our model.*

We conducted ANOVA test to explore the significance of the difference in accuracy between the captions produced by our model, BLIP-2 and the GT. The test indicated a significant difference between the mean scores of the three methods ($p < 0.001$; $f = 20.6$). We also conducted a Tukey HSD test that found significant differences between the accuracy of our model and that of both GT and BLIP-2. No significant difference was found in the accuracy of captions produced using GT and BLIP-2. Although the proposed model has the lowest accuracy rating, it outperforms other methods in generating highly informative graphic descriptions (as discussed above). This indicates that, compared to the other methods, the proposed model succeeded more in addressing the essential aspects that VI (and sighted) individuals require to visualise clothing graphics when solely relying on textual descriptions. However, the low accuracy suggests a lack of accurate extraction of such aspects, highlighting the need for future improvements to ensure reliable descriptions.

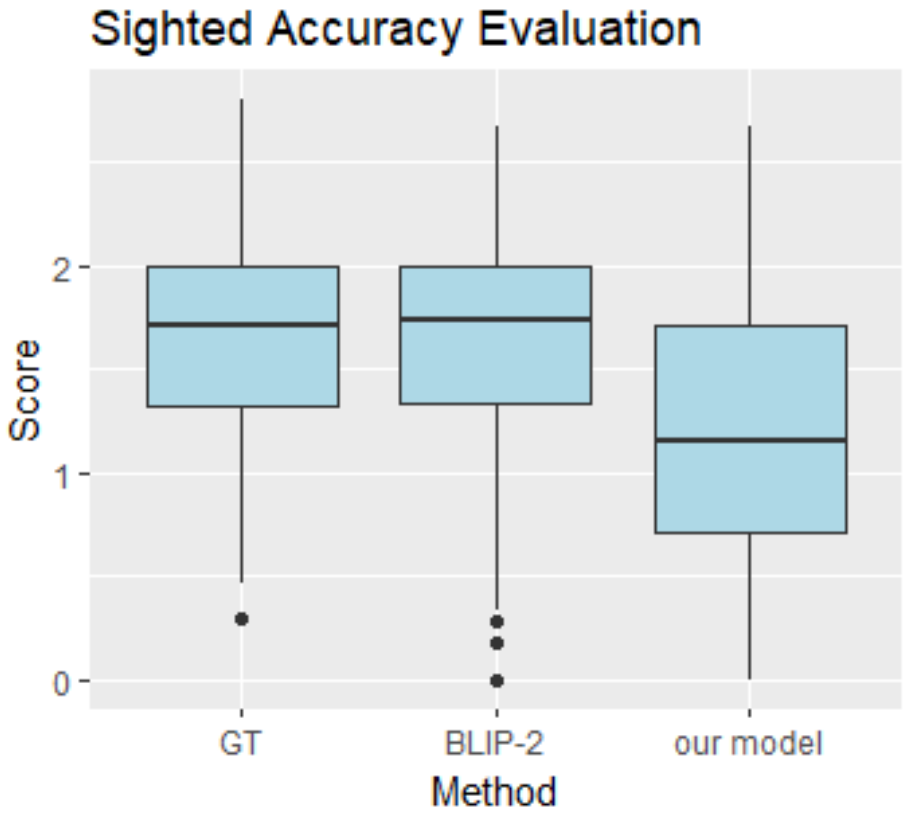


Figure 4.38: *Boxplot illustrating the average score given by sighted evaluators' assessment of the accuracy of captions produced by GT, BLIP-2 and our model.*

Methods	Mean (SD)
GT	1.629 (0.486)
BLIP-2	1.653 (0.569)
Our Model	1.199 (0.627)

Table 4.9: *Mean and standard deviation (SD) results of the sighted accuracy assessment for the three methods (i.e., GT, BLIP-2 and our model).*

Agreement of Sighted Evaluators' Assessments We used Krippendorff's alpha to explore the level of agreement between our sighted evaluators' assessments of the accuracy of the captions produced using the three methods (GT, BLIP-2 and our model). The results show that the agreement rate of the sighted evaluators in assessing the GT captions was 0.209, while the level of agreement was 0.302 and 0.344 when assessing BLIP-2 and our model, respectively. The evaluators were more consistent in their assessments of our model and less consistent in their assessments of the results produced using GT and BLIP-2. These results are further investigated by discussing the qualitative findings below.

Qualitative Results: characteristics and errors

In our evaluation study, we asked sighted evaluators to justify their judgment of the image captions and provide suggestions to enhance the descriptions (details are discussed in section 4.5). Here, we merge the results of the two open-ended questions asked in our study. We present characteristics and errors the participants reported following van Miltenburg et al. (2021) recommendations for evaluating NLG systems.

GT Captions were assessed as having middling accuracy and informativeness. Participants reported that GT captions "describe reasonably well with no glaring omissions [P104]", and the descriptions match the clothing in the image [P23, P46]. Regarding the graphic elements, which are the main focus of our research, the participants reported that these captions are "good at describing the vibe of the graphics" but "not very specific at all on what the graphic depicts"[P48]. The participants reported several factors that made it hard to mentally visualise graphics from the GT caption alone. One reported factor was that the caption "does not describe the graphic" and "just hint[s] at a theme [P98]". Another factor was that "there is no description of the graphic other than [that] it exists [P98]"; the only information describing the graphic is the term *'graphic patterns'*.

In addition to the lack of graphic descriptions, phrasing and the use of confusing language also affected participants' ability to obtain adequate information from the GT captions. They found that the GT captions did not follow the proper structure of the English language, which was confusing and hindered the participants' ability to picture clothing. For example, one participant [P89] explained that they rated the caption *'all thing[s] should be equal, even comfiness, so here's this graphic pullover in [a] buttery soft knit with a baggy fit'* as 'less informative' because it doesn't start with "this person is wearing clothes" or describe "the type of clothes this person is wearing".

Other participants reported that poor language affected their ability to obtain helpful information: (e.g., "Some grammar mistakes make the sentence unclear" [P46], and "All of it is just a mess of scrambled words that are very hard to imagine" [P76]). Additionally, our participants indicated that the word choice, such as the use of metaphor and fashion terms, which can be unfamiliar to people, also impacted the level of informativeness of the captions. For example, in response to the caption *"your yoga workout will feel like a mystical awakening when wearing this foiled graphic tee"*, one participant [P37] commented, "The foil graphic adds mystery".

BLIP-2 Captions were assessed to have the highest accuracy but the lowest informativeness; however, they are straightforward and helped participants easily visualise clothing items. In some cases, BLIP-2 captions accurately captured the graphics in the clothing. The participants appreciated the accuracy of these captions. The participants commented:

> "Overall, I would say it's good. It provides a simple account of the [sic] overal look of the design, including the colours. It would be hard to precisely describe the image on the shirt, so I think it's probably the best possible description. [P68]"
>
> "The description explained that the shirt was a white T-shirt with a floral print; this matches the image which is of a white T-shirt with a floral print on it. [P101]"
>
> "Just exactly how I imagined the image would look like, the description was exactly to the point. [P114]"

In other cases, however, our participants emphasised that the BLIP-2 caption only "gives a basic info [information] to visualise clothing [P78]", and generally "the caption is very broad [P79]". One limitation that affects the quality of the BLIP-2 captions, as reported by our participants, is the *lack of graphic descriptions*. Captions do not describe graphics or provide only the vibe of the graphic element and ignore essential details, such as describing the presented graphics and their features, including colour, scale and position. Participants commented on these limitations, highlighting that the description "[...] does not reveal much, only that there is a logo, but doesn't specify what logo. [P33]". The participants noted that identifying the existing objects alone does not help them visualise a complete picture of the clothing without looking at the image. Instead, they "[...] need more information about what the text says and what the image is. [P21]"

Participants explained why they provided low ratings when judging several BLIP-2 captions. For example, participant [P116] rated the caption *'a woman wearing a white t-shirt with the words.'* as 'not informative' because "it doesn't say what the words are. Nor does it describe the picture on the T-shirt or give any indication of the colours used." Another participant [P51] rated the same caption as 'less informative" and commented: "No description of the graphic and stops mid-sentence before describing the words on the garment." In a different example, participant [P119] rated the caption *'a man wearing a black t-shirt with a logo on the chest.'* as 'less informative' and commented: "I rated this low because it doesn't say what the logo is or what it says. It could be something rude, and a blind person wouldn't know."

Another limitation is that BLIP-2 often used graphics presented on clothing to describe the overall clothing item. For example, for clothing with jungle graphics printed on it, BLIP-2 outputted:: *'the white sweatshirt with a jungle scene on it'*. Participant [P29] noted that this description is "very ambiguous and hard to picture what is actually on the jumper".

Similarly, printed text is used to describe an overall item of clothing. Unfortunately, this misrepresentation causes misleading and confusing descriptions. For example, participant [P67], commented on the caption *'the splash graphic tee is white with pink and blue text'*: "You assume the design on the T-shirt is a splash graphic not the word splash." Another participant commented on the caption *'the purple t-shirt is made from cotton and has a logo on the front'*:"Makes you think the top is purple, not the brand – misleading. [P29]"

Additionally, our participants reported a case in which BLIP-2 misunderstood printed text and provided a very different description, which made them inaccurately visualise the item. One participant [P104], for instance, reported a case in which the brand text 'O'Neill' was described as 'I love it'. The participants reported a case of replacing the text used in the clothing with different wording:

> "The shirt is yellow but it states cake and walk not coffee and wine. [P39]" "The graphic is a brand name and not the words 'I love it', which is a very different proposition. [P98]"

Further limitations highlighted by our participants are that essential graphic features are not described, or some objects are missing. One participant [P48] reported a caption that identifies the existence of palm trees and boats on clothing, missing other objects such as flags, fireworks and waves. In another example, reported by participant [P88], the caption identifies only one piece of text printed on clothing and ignores other text: "It fails to describe that [the] top is covered with large city names – not just New York."

In some cases captions describe the object presented but miss features such as the position and colours, making it hard to visualise the graphics. For example, participant [P58] commented on a caption that states only the information that the object is a star: "Does not explain where the stars are on the clothing. I assumed on [the] chest area when imagining the graphic". Missing colours were also reported by our participants as hindering their ability to visualise the graphic accurately: "I imagined a very colourful vibrant eagle image. [P114]"

Generally, although BLIP-2 generates captions with some degree of accuracy, these captions have been shown to be unreliable because they do not contain sufficient descriptions of visual elements

Our Model captions were assessed to be the most informative but the least accurate, participants identified a number of strengths of our model. Firstly they were reported to be very informative when describing graphic elements: They "created expectations that very closely matched the image [P48]". Our participants reported several characteristics that make our model's captions helpful in visualising graphic clothing elements. An example of these characteristics is that our model caption "details what the picture shows [P43]", "uses a lot of descriptive words [P29]" and provides related information about graphics, "even what they symbolise [P34]".

Some graphic elements are complex and difficult to textually describe; however, our model provides descriptions that strongly match what is depicted in the clothing image. Participant [P54] explained that the captions generated from our model were very informative and accurate: "This is exactly how I would describe it after seeing the picture". Participant [P107] commented that one caption, *'The presented graphic is a depiction of a hand holding a red box with a heart on it. The hand is smooth in texture, as is the box. The heart on the box is raised, giving it a three-dimensional appearance. There are two other elements or details surrounding the hand and the box, although they are not described in further detail.'*, depicted in Figure 4.39, was "pretty good". They noted that although "this type of design is not easy to describe", the caption "conveys reasonably well what you are getting".

However, our participants also highlighted several errors in our model that hindered their ability to visualise clothing graphics. For example, our model generated graphic descriptions alone without providing information about the categories of clothing items (e.g., jumpsuit, dress or skirt). Participants reported facing challenges in mentally picturing graphics without also knowing the clothing category, especially when distinct scattered graphics were shown: "I feel like it gives lots of information, but it's hard to connect them all together to create a visual image [P78]". While it is essential to identify categories of clothing items when generating descriptions, it is also important to emphasise that the proposed model was mainly designed to generate descriptions focusing on clothing graphics alone. It appears that the evaluators expected the description to cover aspects of clothing other than graphics. This expectation arose as the other methods (i.e., BLIP-2 and GT) addressed other clothing aspects, such as clothing category and style. However, the assessment tasks specifically requested a focus on clothing graphics.

Another downside reported by our participants is graphic misrepresentation, including a failure to accurately describe all parts of a graphic and a tendency to provide inaccurate information about object features, such as count, type (text or picture), colour, scale and position.

Inaccurate descriptions of the objects presented in graphics impacted our participants' ability to accurately visualise graphics with the presence of captions alone. Our model successfully identifies the content types but the details are sometimes inaccurate. For example, Participant [P68] imaging a lone palm tree because the caption described there being a single tree (see Figure 4.40). However, when the image was presented, they found that there were actually multiple palm trees. Similar cases were reported by several participants. Participant [P120] commented on the same issue, noting that one image contained three owls in spite of being described as containing a single owl. In some cases, multiple different objects were presented in the clothing graphic but not all of them were identified in the caption, which made it hard for participants to establish a complete mental picture of the graphic. For instance, participant [P41] pointed out a case in which the caption described a graphic presented on a piece of clothing but ignored the text printed underneath. To overcome the issue of missing objects, one participant [P61] suggested that captions 'should add that the text appears several time[s]' or specify the count of the object presented and the location of each object. Participant [P48] suggested a different caption: "A photograph-style graphic with three flowers, one large [one] to the right and two smaller [ones] to the left [P48]".



"The presented graphic is a depiction of a hand holding a red box with a heart on it. The hand is smooth in texture, as is the box. The heart on the box is raised, giving it a three-dimensional appearance. There are two other elements or details surrounding the hand and the box, although they are not described in further detail."

Figure 4.39: *Example of a complex graphic that our model did a good job of describing.*



"The presented graphic is a picture of a palm tree against a striped background. The colors used in the picture are not specified. The palm tree appears in the center of the striped background. There are other elements or objects in the picture besides the palm tree and striped background, although they are not described. The texture or pattern of the palm tree leaves is described as being made of a thin layer of plastic. The background pattern of the stripes is horizontal."

Figure 4.40: *Example of a caption generated by our model, showing the failure to identify the type of the printed object (whether a printed text or picture).*

The lack of information regarding the type of graphics being described was another issue identified by our participants. For example, some objects (e.g., landmarks, plants or animals) were described without specifying whether they were represented in the form of text or a picture. Sometimes, the captions "included the lettering that appeared on the T-shirt but made it sound like this was an opinion or a descriptive phrase about the T-shirt overall [P49]". Our participants reported that this issue impacted how informative our model captions were and caused confusion when visualising clothing graphics. Participant [P116], for instance, visualised a t-shirt with an actual number printed with patterns inside, when reading the caption: *'The presented graphic is a fancy little number with distinct patterns and designs within it. The lines of the number appear to be of medium thickness. The number is printed on a white background with black lettering, which makes it stand out. The overall style or aesthetic of the fancy little number is a mix of the old and new.'*. However, when they looked at the image, they found that the t-shirt contained text saying 'fancy little number' instead. To improve the description, they suggested specifying that the statement was actually a printed caption. An example of a better caption would be: "The T-shirt is white and has short sleeves. Emblazoned on the front across the chest is the caption 'fancy little number.' The words are above one another, with 'fancy' at the top, 'little' just below it and 'number'

below that. The words are close together in a block. The font used is dark black, with bold capital letters."

Another participant commented on the same caption: "'Fancy little number' was not described as a text, [and this] makes it confusing [P116]". Another participant [P38] commented on the use of the statement 'representation of Los Angeles love' in a caption generated by our model (see Figure 4.41), as there was no way of knowing whether this referred to printed text or a pictorial representation. The participant suggested using a better caption: "A pair of greyscale human hands overlap to form the letters 'L' and 'A.' Above the hands, 'Los Angeles' is written in capitals in black font and the word 'love' is displayed in the same font below the image of the hands". Similarly, one participant [P40] reported a case where the clothing was described as including a graphic of New York City but actually included text reading 'New York'.



Figure 4.41: *Example of a caption generated from our model which shows lack of identification of object printed (text or picture printed).*

The lack of description of the scale and position of graphics on clothing items (e.g., centre, top, bottom, left or right side of the clothing) makes it hard to accurately visualise graphics. Our participants reported instances of missing information regarding the size and position of

the graphics or inaccurate descriptions of these features in our model's captions. Participant [P96], for example, was able to successfully visualise an eagle described in the caption, but the lack of information about the position of the eagle hindered their ability to formulate a complete picture: "It needs only explanation about the position of [the] eagle". Another aspect missing from our model captions is information about how graphics are formatted on clothing (e.g., text being presented forwards or backwards) and the relationship between different graphics when multiple graphics are depicted. In one example, our model described the presence of sun and moon pictures; however, "the positioning of the sun and moon in relation to each other [was] not described [P84]". Participant [P69] reported a case in which text was described without noting that the word was actually reversed.

Limitations surrounding colour identification (i.e., missing or incorrectly described colour information) was another issue reported by our participants. This limitation affects how informative and accurate our model's captions are. Our participants encountered difficulty in forming complete mental pictures of the graphics and ended up guessing how the clothing looked, and their guesses typically did not match the images. Participant [P38], for instance, commented on one item (see Figure 4.42: "Too much was left to the imagination, and therefore I felt the description was obscure and too open to interpretation". They suggested a better description: "Three light pink daisies (one large, one medium and one small) are positioned in a triangular formation on a greyish-brown square background".

Unhelpful detailed descriptions were another factor that impacted our participants' preferences for our model's captions. Participants reported that captions sometimes referred to things that were not presented in the image. For example, some captions stated that there were no objects depicted in the clothing image other than the ones described. Although these statements were accurate, our sighted participants did not like that these captions spent so much time explaining what was not contained in the image [P54] and reported that these kinds of captions would only cause confusion [P119].

Participant [P101] commented on one item (see Figure 4.43): "there are many things which are visible in the graphic that the description says are not, such as facial expressions, colours and visual elements". They suggested saying, "A white tank top which features the Boys 2 men [sic] logo, with bright red lettering on 'boys' and 'men' with a large black Roman numeral '2' in the middle. The main graphic below shows the band in a black-and-white image, all posing together with [their] faces and facial expressions clear to see".



"The presented graphic is of a heart-shaped pink flower positioned off to one side on a black background. There are two pink daisies also present in the graphic. The texture or pattern of the flower petals is not described."

Figure 4.42: *Example of a caption generated from our model shows incorrectly describing colours of clothing graphic.*



"The presented graphic is a boyz II men album cover that exudes a mood or atmosphere reminiscent of their music. The positioning or arrangement of the members of Boyz II Men is not described in the graphic. There are no distinct colors or patterns that stand out in the graphic. The facial expressions or body language of the members of Boyz II Men are not observable in the graphic. There are no visual elements or symbols that represent the music or style of Boyz II Men in the graphic."

Figure 4.43: *Example of a caption generated by our model includes unnecessary details.*

Unhelpful details were also provided via descriptions of items that were not depicted in the image at all (i.e., hallucinations). This affects the accuracy of our model's captions. For example, one caption identified that Michael Jackson was on the graphic but also described the presence of his arm and legs, which were not presented in the image. In response, participant [P49] commented, "The text describes a whole-body picture of Michael Jackson, but the T-shirt only contains a picture of his face." The participant suggested a better caption: "Ladies black T-shirt with red long [sic] sleeves. Contains a picture of Michael Jackson wearing sunglasses in the front centre. Blue rectangle in the background. Text 'Michael Jackson' above image (underlined) and 'Beat It' below image, both in cursive font". Participant [P41] reported a case in which our model identifies a person shown in the graphic while only the face is visible. However, our model provides inaccurate details that are invisible in the graphic (see Figure 4.44): "[There is] no way to know if the man is sitting on a bench or looking at the camera. There's also nothing to indicate the presence of weather."

Our participants also reported that our model sometimes identifies text printed on clothing, makes a random guess of what it is and then provides a description accordingly. The participants reported one incident in which the word 'doors' in the name of the popular music band *The Doors* was described as wooden doors. Based on the caption alone, they visualised a door used in buildings. Participant [P5] explained that this inaccuracy caused confusion: "The shirt was described to me as a door, so I was expecting to see a literal door, but the shirt was actually displaying a very famous band, The Doors." Another participant suggested the following alternative caption: "A blue marl sleeveless T-shirt, with a logo for the band The Doors featuring a black and white picture of singer Jim Morrison, overlaying a round red and yellow sunset and trees picture. 'Los Angeles, California' is printed under the picture." [P49]



"The presented graphic is a depiction of a man sitting on a bench. He is wearing a gray sweater and appears to be sitting upright. The man is looking at the camera, indicating a neutral facial expression. There are other objects or people near the man on the bench. Additionally, the graphic indicates the presence of weather or surroundings."

Figure 4.44: *Example of a clothing caption generated by our model shows a hallucination problem.*

Similarly, our participants highlighted a case in which the word 'cake', which was presented as text printed on clothing, was incorrectly described as a graphic (see Figure 4.45). One participant suggested providing a description of the word 'cake' so that they could correctly visualise the item: "The words 'Cake Walk' are printed on the chest area of a yellow T-shirt in cursive, italic font. 'Cake' is on the top line in white, and 'Walk' is on the bottom line in pink. Four pink cherries surround the word 'Cake'[P56]."



"The presented graphic is a circular logo for Cake Walk. The logo features a yellow cake with a cherry on top, symbolizing the sweet treat. The logo is positioned on a yellow background, although the specific positioning is not described. Additionally, the logo includes text or typography, although the exact words are not mentioned."

Figure 4.45: *Example of a clothing caption generated by our model shows inaccurately describing text printed.*

Failure Analysis of Our Model

To address the concerns that our sighted evaluators raised about our model's accuracy, we manually analysed the results to identify the specific components of our model (focusing on those used in the graphic description generation phase) that were causing errors. The details of our model architecture are discussed in 4.2 and illustrated in Figure 4.1. In our analysis, we explored the participants' comments on each clothing item (for all 100 items). For each clothing item, we identified each error raised by the evaluators and explored which component caused the error by analysing the input and output of each component. Each component's impact in the analysed error was assigned a score of 1.

The analysis results reveal that the BLIP-2 caption model caused 38 errors. When analysing these, we identified a number of issues. One was that the model mentions the image's content but does not provide any information to reveal whether the content is text or a graphic; for example, in Figure 4.46, the word 'purple' was used to represent the existence of the term 'purple' without mentioning if the word referred to the *text* purple or to *graphic* displayed in purple. This caused confusion and meant that the subsequent components in the proposed model might fail to represent the accurate meaning of the text. Another area of concern was that the BLIP-2 model added some descriptions that were not found in the image (i.e., a hallucination problem). For instance, in Figure 4.47, the model expected there

to be a picture of the moon represented in a graphic where there was actually no picture of the moon. This, again, impacted the subsequent components.



BLIP-2 Caption: purple - the lion

Automatic Questioning:

Question: 1. What shade of purple is the lion?
Question: 2. Is the lion standing or lying down?
Question: 3. Can you describe the lion's mane?
Question: 4. Are there any other animals or objects in the graphic with the lion?
Question: 5. How does the lion's expression appear?

Automatic Answering:

Answer 1: purple
Answer 2: standing
Answer 3: a purple plant with a lion's head
Answer 4: Yes, there are two other animals in the graphic with the lion
Answer 5: The lion's expression is a combination of the lion's head and the lion's tail.

Automatic Summarising:

The presented graphic is of a standing lion with a purple mane. The lion is the shade of purple and is accompanied by two other animals. The lion's expression is a combination of its head and tail. and tail

Figure 4.46: *Example of a clothing item and results obtained from the proposed model's components. The BLIP-2 caption model misrepresented objects in the graphic element without specifying if they are presented as text or a picture, as indicated in the red box.*

Automatic questioning, by contrast, led to two errors. These were caused either because the automatic questioning was not designed to ask follow-up questions or because of the hallucination problem (as depicted in figure 4.21). Figure 4.48 shows an example of a question asked about whether additional elements beyond those already identified were present in the graphic. If other elements were present, it did not offer a follow-up question that would have prompted the Automatic Answering task to generate descriptions for them.



BLIP-2 Caption: the tarot cards are shown with the sun and moon

Automatic Questioning:

Question: 1. Can you describe the colors and shapes on the card representing the sun?
Question: 2. How is the sun depicted in relation to other elements on the card?
Question: 3. What emotions or feelings does the image of the sun evoke?
Question: 4. Is there any symbolism or imagery associated with the sun on the card?
Question: 5. How does the image of the moon differ from the image of the sun on the card?

Automatic Answering:

Answer 1: the sun is a yellow card with a blue background
Answer 2: The sun is the most important element in the tarot
Answer 3: joy, happiness, love, peace, and love
Answer 4: Yes, the sun is the symbol of the sun god, the sun god is the symbol of the
Answer 5: the moon is a symbol of the feminine, while the sun is a symbol of the masculine

Automatic Summarising:

The presented graphic is a yellow card with a blue background representing the sun, which is depicted as the most important element in the tarot. The image of the sun evokes emotions of joy, happiness, love, peace, and positivity. Symbolism associated with the sun includes it being a symbol of the sun god. In contrast, the moon on the card symbolizes the feminine, while the sun symbolizes the masculine.

Figure 4.47: *Example of a clothing item and results obtained from the proposed model's components. The BLIP-2 caption model generated inaccurate information caused by the hallucination problem, as indicated in the red box.*



BLIP-2 Caption: a woman wearing a t - shirt with the words new york, paris and London.

Automatic Questioning:
Question: 1. What is the color of the woman's t-shirt?
Question: 2. Are the words "New York," "Paris," and "London" written in the same font?
Question: 3. Are the words "New York," "Paris," and "London" all in the same position on the t-shirt?
Question: 4. Is the woman wearing any other clothing or accessories?
Question: 5. Is there any additional imagery or design on the t-shirt, apart from the words?

Automatic Answering:
Answer1: white
Answer2: Yes
Answer3: Yes
Answer4: no
Answer5: Yes, there is a design on the t-shirt

Automatic Summarising: "The presented graphic is a white shirt with the words \"New York,\" \"Paris,\" and \"London\" written in the same font. The words are arranged horizontally on the shirt. There are additional graphics and designs on the shirt. The shirt is short-sleeved."

Figure 4.48: *Example of a clothing item and results obtained from the proposed model's components. It shows the limitations to providing sufficient descriptions of additional elements, as indicated in the red boxes.*

In contrast to the Automatic Question task, the Automatic Answering task caused 46 errors. The analysis showed that the Automatic Answering task failed to generate accurate information mainly because of two reasons. In some cases, it provided no answer, which caused essential information to be missed (e.g., a colour identification issue), as illustrated in Figure 4.49. In other cases, it provided an incorrect answer, which impacted the accuracy of the model (e.g., inaccurate descriptions of objects in graphics). Figure 4.50 shows an example of the Automatic Answering task generating inaccurate colour descriptions, such as the inclusion of red and orange colours in a graphic element. These inaccurate answers led to an inaccurate description.

Regarding the Automatic Summarising task, the analysis showed that no errors were originally caused by this task; however, the lack of accurate information provided by the prior components negatively impacted the accuracy of the description generated during the task. For example, the case of missing answers (caused by the Automatic Answering task) led the Automatic Summarising task to report 'colour is not specified' or simply to guess the answer, as shown in Figure 4.49, which our evaluators found confusing.

Another example was the failure by the Automatic Questioning task to generate a follow-up question, which led the Automatic Summarising task to mention only that other elements were present in the graphic, without providing descriptions, as shown in Figure 4.48. In some

cases, the Automatic Summarising task described some elements as though they were present in the graphic when, in fact, they were not.



BLIP-2 Caption: all you can eat happy vegan taco

Automatic Questioning:

Question: 1. What colors are used in the graphic of the all you can eat happy vegan taco?
Question: 2. Can you describe the different toppings and fillings on the taco in the graphic?
Question: 3. How is the taco presented in the graphic - is it on a plate, in a hand, or in a basket?
Question: 4. Are there any accompanying side dishes or drinks shown in the graphic?
Question: 5. Can you describe the overall mood or atmosphere conveyed by the graphic of the all you can eat happy vegan taco?

Automatic Answering:

Answer 1:
Answer 2:
Answer 3: it's on a plate
Answer 4: no
Answer 5: happy

Automatic Summarising:

The presented graphic is of an all you can eat happy vegan taco, featuring a variety of colorful toppings and fillings. The taco is shown on a plate, with no accompanying side dishes or drinks. The overall mood conveyed by the graphic is happy.

Figure 4.49: *Example of a clothing item and results obtained from the proposed model's components. It shows the limitations of the Automatic Answering task, which sometimes provided no answer or missed essential information, leading to inaccurate descriptions, as indicated in the red boxes.*



BLIP-2 Caption: kiss spirit of rock t-shirt

Automatic Questioning:

Question: 1. What colors are used in the graphic of the kiss spirit of rock t-shirt?
Question: 2. Can you describe the design of the kiss spirit of rock t-shirt?
Question: 3. Is the graphic of the kiss spirit of rock t-shirt detailed or more abstract?
Question: 4. Are there any specific symbols or images included in the graphic of the kiss spirit of rock t-shirt?
Question: 5. How does the graphic of the kiss spirit of rock t-shirt make you feel when you imagine it?

Automatic Answering:

Answer 1: black, white, red, blue, pink, purple, green, yellow, orange, and purple
Answer 2:
Answer: more abstract
Answer 3: Yes, there are a few symbols and images included in the graphic of the kiss spirit of rock t
Answer 4:

Automatic Summarising:

The presented graphic is a colorful design featuring black, white, red, blue, pink, purple, green, yellow, orange, and purple. The design of the kiss spirit of rock t-shirt is more abstract, with a few symbols and images included. The graphic evokes a variety of emotions when imagined.

Figure 4.50: *Example of a clothing item and results obtained from the proposed model's components. It shows the case of the Automatic Answering task failing to generate accurate information, as indicated in the red box.*

4.7 Discussion

The research in stages two and three (refer to the project procedure in section 1.5) demonstrates the development and testing of an effective clothing graphic captioning model for VI users. Here, we discuss our findings in relation to the three research questions (i.e., RQ 4.1, 4.2 and 4.3) synthesised from related work.

4.7.1 RQ 4.1: What are the limitations of ground truth captions of the image–text dataset considering VI people's needs?

Our findings indicate that GT captions in clothing image–text datasets (in this case, FashionGen, FACAD, and DeepFashion-MultiModal) suffer from several key limitations. First,

the GT captions do not provide detailed descriptions of graphics, which constitute essential information required by VI people in order to mentally visualise clothing (Alluqmani et al., 2023). Currently, the GT captions either identify merely the existence of graphics but provide no details, or they give a general sense of the graphics' nature but nothing more. Surprisingly, our analysis with the FashionCLIP model also revealed that graphics are sometimes inaccurately classified in the FashionGen dataset, as depicted in Figures 4.25 and 4.26. While the FashionGen dataset fails to uphold standard graphic identification, our results show that the FashionCLIP model is effective (i.e., achieving F1 score = 0.85) at identifying the graphics on the clothing in its images when applied to redefine the FashionGen dataset's graphical classification considering VI people's needs.

A second interesting finding was that the GT captions also fail to adhere to the standard English language structure and to use suitable words to provide informative captions. This finding is in accordance with those of recent studies indicating that clothing image–text datasets lack essential attribute descriptions (Gao et al., 2020; Nguyen et al., 2021). While previous research in clothing captioning models (Cai et al., 2022; Han, Zhu, et al., 2023; Nguyen et al., 2021; Xu et al., 2015) used GT captions to train the models developed, the results of our research demonstrate that this practice is flawed as the clothing image–text datasets fail to address essential fine-grained clothing elements. Additionally, the clothing image-text datasets disregard the necessity of addressing VI people's needs, making them unsuitable for training models to generate VI-friendly captions.

4.7.2 RQ 4.2: How informative are the captions generated by our model compared to those generated by existing models and ground truth data for VI users?

We discuss this research question from two perspectives. First, we discuss the impact of adopting an ROI-based approach to generate more detailed information compared to existing models, which adopt a grid-based approach. Next, we discuss the use of LLM in our model to address VI people's needs.

In terms of the adoption of the ROI-based approach, our findings illustrated the comparative performance of our ROI-based model in describing particular features (i.e., graphics) of clothing elements compared to BLIP-2, which uses a grid-based approach.

Our automatic evaluation revealed that the model we developed performed comparably to the BLIP-2 model in most metrics when tested on various clothing datasets (i.e., FashionGen, DeepFashion-MultiModal, and FACAD). Specifically, our model generated captions that matched the GT descriptions the most when applied to the DeepFashion-MultiModal dataset, potentially since this dataset includes captions for each image view (e.g., front, side, and back images), prompting the generation of detailed descriptions, while the FashionGen and FACAD datasets include just one caption that covers all image viewpoints.

However, this finding of comparative performance of our developed model must be interpreted with caution as automatic image-captioning evaluation metrics such as BLEU, METEOR, and ROUGE-L, which judge the generated text based on the similarity with the human-generated text (i.e., GT captions), may not serve to provide a reliable assessment of captioning models. They may fail to do so for several reasons. Firstly, the automatic captioning metrics judge the performance of the generated captions compared to the GT "gold standard" data, which we have already identified as being highly flawed in subsection 4.3.1.

Secondly, these automatic metrics assess generated captions based only on one variant of a caption, which may lead them to incorrectly judge the quality of captioning if there are other valid options that are not considered. That is often the case, because humans tend to provide varying descriptions of the same image (Ming et al., 2022). Furthermore, variant captioning of clothing imagery may arise if there is a need for varied descriptions that each have different goals, such as informational/utilitarian or marketing goals (Yang et al., 2020). Likewise, within those goals, there may be a number of approaches that are taken, such as top-down or bottom-up approaches (Ming et al., 2022). Beyond this, the intended recipient of the captions is another factor that influences how humans describe an image. For example, image descriptions may be produced that are intended for a VI audience, which communicate additional details so that the person can better visualise the image (Stangl et al., 2021), but those may not be the variants assessed by automatic metrics.

Nonetheless, our human evaluation results showed that the model we developed generated more informative captions when compared to BLIP-2, which applies a grid-based approach, and GT. This finding is contrary to previous studies, which suggested that an ROI-based approach is not effective when seeking to generate detailed descriptions in the clothing domain (Gao et al., 2020; Han et al., 2022; Zhuge et al., 2021). In contrast, we argue that applying the ROI-based approach to caption images with detailed features enables informative descriptions to be generated. The contradiction of our findings compared to previous studies could have arisen because the ROI-based approach was never applied to generate descriptions of detailed clothing elements such as graphics. We believe that such clothing elements need to be considered as separate components to the overall image and ones that require detailed descriptions. Graphical elements are unlike the other clothing elements, and as such, they require exploring from the first stage of captioning in order to permit further analysis if identified as necessary.

Regarding the use of LLM, our findings show that adopting an LLM in our image captioning model enables it to provide VI-friendly captions of clothing graphic elements, which VI (as well as sighted) evaluators considered to be more informative than BLIP-2 and GT captions. These results show that leveraging the LLM to consider VI people's needs could be a promising approach, which is in line with previous findings that LLMs are capable of simulating human behaviour (Aher et al., 2023). In our case, the LLM succeeds in generating questions that it has correctly inferred would come to mind for VI people seeking to visualise the clothing graphics. Nonetheless, although the LLM was shown here to outperform human-written captions (i.e., GT captions) in presenting essential information for VI people, these results must be taken with caution as none of the GT captions were aimed at serving VI people's needs. However, previous research (Simons et al., 2020) indicated that crowd workers, who are usually recruited to provide GT captions for image datasets, face challenges in determining what information needs to be included to provide VI-friendly captions. As such, it may be that our model would outperform GT captions even when those were written with the intent of being VI-friendly. In the future, it is worth building a dataset of VI-friendly human-written captions and comparing their informativeness with captions generated by our model.

Another finding of our study was a low correlation between VI and sighted people when judging the informativeness of captions. Our analysis of the relevant results revealed that the low agreement mainly concerned the LLM including all image details exhaustively, with

doubts raised around whether it provided complete information. For example, VI evaluators appreciated captions that recognised some objects in the clothing graphic were unclear or captions that confirmed there were no other elements presented in the clothing graphic apart from those already described, details that were not equally appreciated by the sighted evaluators. These findings are in line with previous work (MacLeod et al., 2017) that showed VI people seek honesty and transparency in descriptions; for example, they would rather there be a longer caption that informs them of the potential for errors rather than there be a lack of detail. These findings indicate that our model generates captions that better consider the needs of those with visual impairments.

4.7.3 RQ 4.3: How accurate are the captions extracted by our model compared to those extracted by existing models and ground truth data?

The results show that although the captions generated from the proposed model were assessed as being the most informative compared to the BLIP-2 and GT captions, they were rated the least accurate. Our analysis of these results shows that two main factors affect the proposed model's accuracy. The first factor is empty or inaccurate results produced by the BLIP-2 models (i.e. image captioning and question answering). In some cases, the BLIP-2 captioning model fails to generate accurate setup statements, which negatively impacts the quality of the questions produced by the automatic questioning task. BLIP-2 in the automatic answering task in some cases generates wrong or empty answers that negatively hinder the accuracy of the final captions. These results are in line with the findings of Zhu, Chen, Haydarov, et al. (2023), who showed that the inclusion of BLIP-2 in their model negatively affected the accuracy of the generated captions. The other factor that affects the accuracy of the captions generated from the proposed model is the hallucination problem. This is caused by adopting an LLM in the automatic questioning and automatic summarisation tasks, as well as adopting a VLM (i.e. BLIP-2) in the automatic answering task and the setup statement. The hallucination problem is widely identified in generative models and occurs when the model generates information that does not actually exist. This problem has been identified in VLMs, such as the BLIP-2 model, and in LLMs, such as ChatGPT (Li, Du, et al., 2023; Lovenia et al., 2023). Future work might tackle these limitations by adopting multiple rounds of automatic questioning and answering tasks to verify the generated descriptions. Another way of handling such limitations is to design precise prompts that check for uncertainty and request the generation of only information with less potential for error.

4.8 Conclusion

This chapter explored aspects of developing a VI-friendly clothing captioning model. It started by discussing our model framework in section 4.2. Then, section 4.3 investigated the existing GT captions (i.e. FashionGen, DeepFashion-MultiModal and FACAD) and proposed a new clothing graphic dataset (CGD). section 4.5 outlined the experimental setup adopted in this research, including the automatic and human evaluation methods, and illustrated the research procedure. This chapter concluded by reporting the findings in section 4.6 and further discussing these findings in section 4.7.

The results of assessing the existing clothing image-text GT captions were obtained from

the manual analysis in subsection 4.3.1 and the assessments of the VI and sighted evaluators in subsection 4.6.3. The overall results show that GT captions suffer from several limitations, such as omitting essential elements of clothing items like graphics. Such limitations affect the suitability of these captions to train VI-friendly models. Additionally, they fail to adhere to standard English-language syntax, which causes confusion in fully understanding descriptions. These results provide an answer to RQ4.1.

The proposed model was assessed by conducting automatic and human evaluations and manual analysis. The analysis was conducted in two stages. First, in subsection 4.6.1, the results obtained from the proposed model were analysed considering VI people's needs, as discussed in Chapter 3. Second, in subsection 4.6.3, the results of the proposed model were analysed to explore the limitations arising from the assessments obtained from the human evaluation. These assessments were applied to examine the informativeness of the proposed model, answering RQ4.2. Additionally, the assessments were considering two essential concepts: adoption of a ROI-based approach to produce detailed descriptions of fine-grained clothing elements (i.e., graphics) and adoption of an LLM to generate VI-friendly captions. The results demonstrate that the proposed model, which employs an ROI-based approach, generates more informative captions compared to the BLIP-2 model, which employs a grid-based approach. The proposed model received an average of 1.737 and 1.774 rating scores when the informativeness was assessed by the VI and sighted evaluators, respectively. BLIP-2, on the other hand, obtained an average rating of 1.460 and 1.565 by the VI and sighted evaluators, respectively.

Furthermore, the findings of this research indicate that adopting an LLM to simulate VI people's enquiries about clothing images to generate VI-friendly captions is a promising approach compared to the current practice of training new models using human-written captions (i.e. GT captions). These results demonstrated a low average rating of the informativeness of the GT captions (i.e. 1.440 and 1.671 given by the VI and sighted evaluators, respectively) compared to the average rating assigned to the proposed model (i.e., 1.737 and 1.774 scores by the VI and sighted evaluators, respectively).

The results of the human evaluation were reported in section 4.6.3, which also presented the accuracy assessment of the descriptions generated from our model to answer RQ4.3. These results show that although the proposed model produces informative captions considering VI people's needs, it struggles to maintain accuracy. However, these limitations can be tackled in future by conducting multiple rounds of the automatic questioning and answering tasks and designing prompts to avoid uncertain answers (Zhu, Chen, Haydarov, et al., 2023).

4.8.1 Limitations on the Human Evaluation Approach

The human evaluation approach adopted in this thesis was designed to assess the informativeness and accuracy of captions of three methods: GT, BLIP-2 and our model. The details of the evaluation approach have been discussed in 4.5.4. Here, we discuss several limitations that were observed when analysing the sighted evaluators' comments.

One drawback of our human evaluation approach was the lack of illustrations used to emphasise our focus on the descriptions of clothing graphics when assessing the informativeness of captions (i.e., we aimed to assess whether the descriptions of graphical elements were informative). From our analysis of comments from sighted participants, we found that participants interpreted informativeness differently. Some assessed the informativeness of

the captions based on whether they addressed all information about a clothing item. For example, this involved the inclusion of descriptions of clothing elements other than graphics, such as sleeves and clothing categories (e.g. shirt or t-shirt). Other evaluators rated the informativeness of captions based on whether they could build a visual picture of the graphical elements relying on textual data alone. This difference in the interpretation of caption informativeness resulted in low agreement among participants when judging the same item of clothing, as discussed in 4.6.3.

Another limitation was the lack of guidance on how accuracy should be assessed. The evaluators had varied views when judging the level of accuracy of clothing item descriptions. In our analysis of comments from sighted evaluators, some participants considered captions to be accurate if they included some correct details, while others only rated captions as highly accurate when they were as precise as possible. These differing approaches resulted in low agreement among the participants when judging the accuracy of a description of a given clothing item.

In addition to the lack of clarity regarding how the informativeness and accuracy of captions should be assessed, the sequence of the questions in our human evaluation design influenced the final evaluation results. We presented the first question (which asks evaluators to assess the informativeness of a caption based on whether they could build mental images of graphic elements by relying on the caption alone) and the second question (which asks evaluators to assess whether the caption in the previous question was accurate by checking it against the associated image) on the same page. Presenting both questions on the same page might have caused the evaluators to look at the image in the second question when assessing informativeness in the first question. Ultimately, this could have affected the quality of the evaluations. In particular, building a mental preconception of the clothing image based on the written description would have been disrupted by the presence of the image on the same page.

These limitations may have affected the evaluation results. However, they can be avoided in the future by providing examples showing that describing the graphic element is our main focus in our assessment of informativeness. Furthermore, to avoid varying evaluations of accuracy level, more descriptions could be provided about the number of incorrect elements, which could be considered very good, good, bad or very bad. In terms of the presentation of the evaluation questions, the questions should be presented on separate pages to avoid affecting the assessment.

Chapter 5

Conclusions and Future Work

This Chapter provides a conclusion to the thesis in section 5.1 including a thesis summary, key findings and overall thesis contributions. Section 5.2 provides a discussion of potential open areas and future directions in which this work can be expanded.

5.1 Conclusions

5.1.1 Thesis Summary

The thesis began with Chapter 1, which introduced the research motivation to explore VI people's online shopping challenges and to develop ways to assist them in this task. It briefly outlined the challenges typically encountered by VI people when performing such tasks and presents the rationale for choosing to investigate clothing products in particular among other shopping categories. It then discussed the current research challenges, thereby introducing the research questions, aim and objectives, and introduced the thesis' key research contributions.

Chapter 2 outlined the existing literature in several areas related to this research. This includes the broad area of VI people's challenges in daily activities. This is followed by an exploration of research specifically related to this project, namely VI people's shopping challenges in online and offline environments and in purchasing clothes in particular. Then the research that proposes VI people-friendly assistance systems for online clothes shopping is explored. Additionally, the research into image captioning in general, clothing image captioning, clothing image representation for VI users and clothing image-text datasets was discussed. Furthermore, research leveraging LLMs to generate image representations was examined. The chapter concludes by reviewing research on image captioning evaluation.

Chapter 3 employed a co-design approach with key stakeholders (i.e., VI people and their representatives at the SRSB) to design and develop two complementary studies. Task observation and standalone interviews were conducted with potential end users (i.e., those with visual impairments) to explore VI people's challenges in shopping for clothing online. In the task observation, the participants were asked to perform an online shopping activity to understand their shopping behaviour when shopping online. In the standalone interviews, the aim was to explore their outfit selections, coordination processes, requirements and needs.

Based on the key findings from Chapter 3, Chapter 4 described the development of a VI-friendly captioning model to enhance the online shopping experience for VI people. It dis-

cusses the architecture of the proposed model, examines existing clothing image-text datasets and introduces a novel clothing graphic dataset. Furthermore, it evaluates the proposed approach against baselines using both automatic and human image captioning evaluations. In addition, the chapter details the research procedure followed, and ends by reporting and discussing the chapter's findings.

5.1.2 Summary of Findings

The aim of this thesis is to examine the challenges that VI people experience when purchasing clothing online and to explore the creation of an approach for automated generation of effective VI-friendly clothing descriptions. This subsection provides a summary of how the adopted methods addressed the research questions and outlines the key findings.

Research Question One: What challenges do VI people face when independently selecting and coordinating outfits while online shopping?

Previous research has mainly used interviews to explore the challenges that VI people face when purchasing clothing online, so to obtain a more comprehensive understanding of those challenges, two complementary studies were conducted: task observations and standalone interviews. The task observations were used to understand the participants' in-context behaviours and identify barriers to performing specific shopping tasks and to allow the participants to consider, recall, and share important details. Standalone interviews were also conducted to gather narrative data about their perspectives and needs.

Conducting task observations alongside the standalone interviews provided immediate examples of the challenges that VI people face when online shopping. It also uncovered several challenges that were not mentioned during the interviews. For example, the observations revealed several behavioural patterns, such as validating retailers' descriptions with review data and using search functions as a primary tool for finding items online. These patterns led to discussions of the rationales for such behaviours, which in turn uncovered additional limitations, such as unreliable retailer descriptions and the frequent misclassification of clothing items. The studies revealed that VI people experience considerable limitations in product descriptions, including missing, insufficient, or unreliable descriptions of clothing styles, categories, graphics, and colours.

Research Question Two: To what extent do VI people consider a specific celebrity or fashion style when selecting outfits?

To address this research question, VI participants were asked, in the standalone interviews, about their fashion styles and the influence of celebrities. Our findings suggest that VI people do not show an interest in particular designs, fashion models, or influencers but prefer to follow brands or colour trends or to consider their own body shapes.

Research Question Three: What are the current VI-assistance systems, and what type of assistance do VI people like to have to purchase clothing online?

This research question was addressed by considering the task observations and through narrative discussions in the interviews. The results show that VI people use screen readers,

automated voice assistants (e.g., Alexa, Siri), and colour contrast and magnification tools when using the Internet. However, they are generally hesitant to place trust in automated systems, which is in line with the literature (Wang et al., 2021). Humans are capable of referencing items already owned or known to the VI shopper, which helps them to visualise how a new item looks, but automated systems are not. Nevertheless, VI people are interested in new assistance systems, including image captioning and recommendation systems. From these key findings, several design implications are identified and discussed in section 5.2. Examples of suggested implications for design include validating clothing product descriptions, removing unlikely-to-purchase items from search engine results, developing a VI-friendly captioning model, colour identification and scaling systems and fashion warning systems.

Research Question Four: How can we build an effective clothing graphic description model to generate useful descriptions for VI users?

The findings from the first three research questions suggest several design implications. Among them, generating detailed descriptions is one of the key challenges identified in Stage 1 of this project. Addressing such limitations is essential due to the large impact of missing, inaccurate, or vague descriptions on VI shoppers' satisfaction. Generating automatic detailed descriptions will enhance the shopping environment for VI people in several ways. First, automatic detailed descriptions help VI people to be more self-dependent and minimise the need to seek sighted assistance to verify visual information or provide detailed descriptions. For instance, providing detailed descriptions of objects presented in graphic elements and their attributes, such as colours, size and position, will reduce the need for sighted assistance to describe object attributes compared to providing only the word "graphic" to describe graphic elements.

Additionally, detailed automatic descriptions enhance the ability of people to mentally visualise clothing items by relying on textual data alone. Constructing mental visual representations of clothing items not only enables VI individuals to make selections when purchasing clothing online but also provides the opportunity to assemble more suitable matching outfits. Second, offering detailed descriptions will improve the confidence of VI people in making selections independently and eliminate the restriction on choosing only plain items to avoid purchasing incompatible or undesirable products. With this and other positive outcomes to be realised, overcoming the limitations resulting from missing descriptions of clothing graphics—a matter unaddressed until now—seems a worthwhile pursuit.

To answer this research question, we began by exploring image captioning research that might be applied to develop assistance systems for VI people when purchasing clothing online; however, we found a lack of such systems whose designs considered VI people's needs. We then analysed the GT captions commonly used in clothing image–text datasets (i.e., Fashion-Gen (Sadeh et al., 2019), DeepFashion-MultiModal (Jiang et al., 2022) and FACAD (Yang et al., 2020)) and found limitations within those datasets. For example, in FashionGen, multiple different images for the same clothing item can share the same caption, creating ambiguity that causes noise when training an ML model. DeepFashion, meanwhile, often uses conflicting terms in its captions, such as describing the same item as a 'sweater' and as a 'shirt', and FACAD uses a marketing-friendly style that aims to attract customers, rather than accurately describing the items, and which is thus unsuitable for VI people. This result is aligned with previous research, which indicates the clothing image-text datasets lack

essential attribute descriptions (Gao et al., 2020; Nguyen et al., 2021). We then conducted a human evaluation of the GT captions, which, together with the preliminary analysis, showed that the captions fail to consider VI people's needs and lack sufficient descriptions of essential elements, such as graphics.

To generate fine-grained descriptions, we adopted an ROI-based image captioning approach rather than a grid-based approach, following the former's success in generating fine-grained descriptions for general images (Anderson et al., 2018; Cornia et al., 2020; Kinghorn et al., 2018; Long et al., 2023; Luo et al., 2021; Shi et al., 2020). We compared the ROI-based approach in our model to the BLIP-2 captioning model, which uses a grid-based approach, with both automatic and human evaluations. The automatic evaluation showed that the captions were comparable to the BLIP-2 model; however, because the automatic evaluation assessed the generated captions based on the GT captions, which we had already identified have limitations, the automatic evaluation results should be considered with caution. The human evaluation, instead, showed that our model generates more informative captions than the BLIP-2 model, indicating that the ROI approach surpasses the grid-based approach in generating detailed descriptions of fine-grained clothing elements, such as graphics.

Considering the limitations of the GT data and the need to generate VI-friendly captions, it was essential to leverage the ability of LLMs to simulate human behaviour (Aher et al., 2023) and apply it to generate questions similar to VI people's enquiries regarding clothing images. As illustrated in section 4.2, these questions were then used to generate VI-friendly captions. To assess this approach, we conducted a human evaluation of the informativeness of captions generated by our model compared to GT captions and found that our model was rated as more informative by both sighted and VI participants. This indicates that the LLM effectively develop captions that it correctly predicts will be of interest to VI individuals trying to visualise clothing images. These results are in line with previous research on image description generation more generally, which indicates that LLMs are capable of generating relevant questions about images (You et al., 2023; Zhu, Chen, Haydarov, et al., 2023).

Next, the accuracy of the proposed model's results was evaluated compared to BLIP-2 and GT captions. The results show that its captions were less accurate than those produced by the other methods. To further explore these results, a failure analysis was conducted by manually exploring the results produced by each component (i.e., those used in the graphic description generation phase) of the proposed model in light of the sighted evaluators' comments. The analysis results indicate that two main aspects negatively affect the proposed model's accuracy. One aspect affecting overall accuracy is generating empty or inaccurate results produced by BLIP-2 models (i.e., image captioning and question answering). The BLIP-2 captioning model often fails to provide accurate set-up statements, which has a negative impact on the quality of the questions generated by the automated questioning task. BLIP-2 sometimes produces incorrect or empty answers in the automated answering task, which significantly hinders the correctness of the final captions. The results align with the findings of Zhu, Chen, Haydarov, et al. (2023), who indicate that including BLIP-2 into their model had a detrimental impact on the correctness of the generated captions.

The hallucination problem is another matter that negatively impacts the accuracy of the generated captions. It is caused mainly by adopting an LLM in the automatic questioning and automatic summarisation tasks and by adopting a VLM (i.e. BLIP-2) in the automatic answering task and the set-up statement. The issue of hallucination is often seen in generative

models when the model creates information that is not factually correct. This problem has been identified in VLMs, such as the BLIP-2 model, and in LLMs, such as ChatGPT (Li, Du, et al., 2023; Lovenia et al., 2023). Future work might tackle these limitations in several ways as discussed in section 5.2.

5.1.3 Summary of Contributions

In summary, our findings corroborate many conclusions from prior literature but also contribute several new, previously unidentified challenges and potential solutions to help VI people shop for clothes online. This is the first comprehensive study of VI people's experiences in selecting and coordinating clothing online, with the opportunity to share their thoughts on how an automated system could be used to improve their online experience. It also makes a major contribution to research on designing better online assistance systems for VI users. As well as reporting on our findings in detail, we translate these into design considerations, implications and recommendations to help make online shopping accessible and enjoyable for everyone.

Furthermore, this research is the first to tackle the problem of generating VI-friendly captions for clothing images. Although researchers have highlighted the limitations of the clothing datasets (Gao et al., 2020; Nguyen et al., 2021), this research is the first to explore these limitations comprehensively. Specifically, it shows that the most commonly-used clothing datasets (i.e., FashionGen, DeepFashion-MultiModal, and FACAD) are unsuitable for building VI-friendly clothing captions. Additionally, it shares the first clothing graphic dataset (i.e., CGD), which contains items from these three datasets. Additionally, this research assessed the effectiveness of the FashionClip model in identifying graphic clothing elements. Previous research in image captioning (Gao et al., 2020; Han et al., 2022; Yang et al., 2020) trains their models to simulate certain ground truth captions, which cause a lot of redundancy, limit the diversity of generated captions and restrict the generation of more human-like captions (Luo et al., 2022). Our model instead simulates VI-people's enquiries to mentally visualise clothing images by leveraging the advanced knowledge of the LLM. Another important contribution of this research is showing that predefining essential clothing elements and adopting an ROI-based approach are effective in generating detailed VI-friendly captions. In addition, this research sheds light on the importance of providing detailed descriptions of such missing elements (i.e., graphics) and establishing the foundation for creating detailed descriptions of clothing items with graphical elements.

5.1.4 Significance of This Thesis

This thesis is significant in several ways. One of its important contributions is that it addresses a critical gap in the literature by providing comprehensive knowledge of VI people's challenges and limitations in one of the most difficult activities (i.e. buying clothes online). The findings outlined from the first stage of this project (refer to the project stages in **??**) not only offer knowledge of VI people's challenges and online shopping limitations but also provide knowledge about how VI people cope with these limitations and how automated systems can be adopted to enhance their experience.

Another significant benefit of this thesis is that it provides the first comprehensive assessment of the widely used clothing image captioning datasets in train image captioning models.

In addition, this project contributes to the literature by creating the first clothing graphic image captioning dataset, which will facilitate future research on clothing graphics. Furthermore, this thesis proposes the first clothing graphic image captioning model considering VI people's needs.

In addition, this thesis lays the groundwork for future research in the clothing image captioning field. It demonstrates the importance of shifting the old-fashioned approach of developing image captioning models (i.e., training new image captioning models to mimic fixed ground-truth data) to a new advanced approach (i.e., using the powerful knowledge of LLM to generate more flexible user-based captions). Furthermore, this thesis highlights the essential of considering ROI approach to generate detailed descriptions of clothing fine-grained elements.

Ultimately, this thesis has the potential to inspire researchers interested in enhancing VI people's understanding of clothing images.

5.2 Future Work

As this work lays the groundwork for comprehending the essential needs of people with VI in clothing shopping online and for finding a suitable solution, it opens up several avenues for further research in clothing image captioning research. This section presents current limitations and potential areas for expansion of this work.

5.2.1 Further Investigation

This work excluded people with VI who had other disabilities or comorbidities from the participant sample to minimise confounding variables. However, to gather more widely applicable findings, future research should consider a broader sample of the population to understand whether and how various other disabilities interact with VI in clothing shopping online.

To permit practical data collection, observational tasks were conducted exclusively on the Amazon website. Most participants were already familiar with Amazon and considered it a more accessible website than many others. In addition, the use of a single site reduced the experimental complexity. However, in the future, the observational tasks should be repeated on a range of websites with varying levels of perceived accessibility and participant familiarity.

Future research should also examine how the online purchasing behaviour of people with VI differs in the presence of sighted assistance. An observational approach would provide a complete view of online issues. However, researchers should consider the possible difficulties that people with VI experience when completing the tasks and offer alternative tools if they express discomfort.

5.2.2 Possible Design Implications

This research has generated several design considerations, as detailed below:

Validating clothing product descriptions. Contradictory or deceptive descriptions were identified as one of the main limitations of online clothing websites. Contradictory descriptions appear in the descriptions themselves and become more apparent when comparing retailers' product descriptions and the descriptions for the same product provided by customers in reviews. One way to increase the confidence of VI users is to validate the information in

the product descriptions with the review data. This can be done by providing additional information regarding the accuracy level of the product descriptions, thereby enabling VI shoppers to make more informed decisions.

Colour identification and scaling systems. One hindrance to VI people purchasing items independently from online clothing websites is inaccurate and inconsistent colour descriptions. A possible solution, outlined from empirical studies, is to develop a standard to assign numbers to colours, such as 01 to the darkest blue, 02 to the next darkest blue, up to 10 for the lightest blue. This could help VI people determine different shades of colour and compare them with their existing clothing, allowing for the better coordination of colours. This would provide a more flexible shopping environment and mitigate the challenge of understanding unfamiliar or unstandardised names, such as 'sky blue' or 'royal blue', and their associated levels of darkness. Furthermore, descriptions could include the ratios of all of the colours present in a piece of clothing. These additional details could help VI users to construct a more precise representation of the clothing item.

Remove unlikely-to-purchase items from search engine results. One behaviour identified from observing VI people shopping for clothing online is that they relied mainly on the search tool to access their desired item. However, search tools typically generate lengthy lists of results and sometimes include unrelated items. Although this also affects sighted users, VI users require more time and effort to scan the resulting items, resulting in higher cognitive cost. One suggested solution is to track users' shopping behaviours to obtain insight into their preferences, similar to existing recommendation systems, with the aim of eliminating items that are unlikely to be purchased and displaying only user-preferred items. This would alleviate the difficulty of visually impaired individuals scanning items or blind users having to list each item using a screen reader before being able to choose the relevant item.

Fashion warning systems. Our results show that VI people rely on purchasing plain items, mainly in black, to prevent the possibility of mismatched clothing. Developing a fashion warning system is essential to mitigate this limitation. Such a system could alert VI users when they make potentially undesirable selections. For example, the system could provide a warning if the user attempts to combine incompatible items, such as matching a formal top with a casual bottom. This would increase the confidence of visually impaired individuals and reduce their current limitations, such as having to rely on familiar websites or choosing only plain items.

5.2.3 Enhancements of the Clothing Graphic Dataset

The clothing graphic dataset (CGD) developed for this thesis offers a set of clothing graphic images used in evaluating our model. However, some images have several limitations that affect the accuracy of our model. For example, some images have a low resolution, which hinders AI models from extracting useful information. Figure 5.1 shows an example of a low-resolution graphic. The object next to the deer is not well presented. This limitation provides the opportunity for VLMs and LLMs to guess the content, affecting the accuracy of our model.

Another limitation of the images in the CGD is that not all graphics are shown well. For example, in Figure 5.2, the text printed on the sleeve is hidden because the sleeves are rolled up, or none of the images show the printed text clearly.

The CGD can be further enhanced by filtering items with unclear graphics and enriching

the dataset with clearer graphic clothing images. Enhancing the dataset will facilitate the advancement of the research development of useful applications.



Figure 5.1: *A low-resolution image from the CGD.*



Figure 5.2: *A clothing image from the CGD, showing text printed on the left sleeve. Although the item is shown in various angles, none of the images present the text clearly.*

5.2.4 Improvements of the Performance of the Proposed Model

The proposed model will be enhanced in the future to generate comprehensive graphic descriptions of clothing items. Several areas could be considered to improve the proposed model.

First, our model was designed as a pipeline in which the quality of the proceeding output impacted the quality of the following task. In our model, the BLIP-2 captioning model was adopted to generate the setup statement, which was used in the following automatic conversation (see the details of our model in section 4.2). This shows many limitations

that impact the quality of the following tasks (see the findings discussed in section 4.6). In the future, we will use recent superior VLMs such as MiniGPT-4 (Zhu, Chen, Shen, et al., 2023), which generates more accurate and rich information from images than BLIP-2. This will eliminate the hallucination problem and enable the subsequent task (i.e., automatic conversation) to generate more informative and accurate captions.

Another way of enhancing the proposed model is to adopt the chain-of-thought method (Wei, Wang, et al., 2022), in which LLMs are prompted to solve a certain task in a series of steps. This can be adapted in our model by prompting ChatGPT in the automatic questioning tasks to ask one question at a time. Thereby, we leverage the multi-step reasoning ability of LLM (Wei, Wang, et al., 2022).

Addressing the confidence level is another essential element to be considered in enhancing our model in the future. To win the trust of people with VI in AI assistance, the confidence level should be addressed when generating captions (MacLeod et al., 2017). In our model, the VLM in the automatic answering task would be prompted to share the accuracy level and the reason if the confidence is low.

Furthermore, to generate comprehensive descriptions of graphic elements, the proposed model should consider clothing as an item rather than as an image, following Tang et al. (2024)'s suggestion. Thus, the model should consider all graphics in clothing items at different angles.

5.2.5 Flexible VI-friendly Captioning Models

The proposed model is mainly designed for people with VI (i.e., blind and low vision) in general. However, people who are blind and those with low vision differ in terms of understanding visual images. People with low vision might not use screen readers; instead, they might use magnification or colour contrast tools. Therefore, for them highly detailed descriptions might be boring or unnecessarily time-consuming; thus, they might only need descriptions adequate enough for them to verify the graphics they see on the screen. On the other hand, people who are blind would appreciate detailed descriptions that assist them in visualising the graphics.

Another factor that must be considered when building a flexible captioning model is the visual experience. People with acquired blindness have some knowledge of visual aspects such as colours and shade levels and other visual objects such as famous landmarks. Thus, our model must be enhanced to be flexible to address the various levels of perception and visual experience. More flexible personalised captions can be generated by modifying the prompt in the automatic summarising task to summarise the results, considering the user's unique needs. For example, for a user who has been blind since birth, the descriptions should include a detailed explanation of the object presented, including famous landmarks. Therefore, the model should ask the user if a detailed description is desired.

Bibliography

Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. *International Conference on Machine Learning*, 337–371.

Alluqmani, A., Harvey, M. A., & Zhang, Z. (2023). The barriers to online clothing websites for visually impaired people: An interview and observation approach to understanding needs. *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, 753–764. https://doi.org/10.1145/3563657.3595978

Alshenqeeti, H. (2014). Interviewing as a data collection method: A critical review. *English linguistics research*, *3*(1), 39–45.

Anderson, P., Fernando, B., Johnson, M., & Gould, S. (2016). Spice: Semantic propositional image caption evaluation. *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, 382–398.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6077–6086.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3), 337–351.

Arystanbekov, B., Kuzdeuov, A., Nurgaliyev, S., & Varol, H. A. (2023). Image captioning for the visually impaired and blind: A recipe for low-resource languages.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bai, S., & An, S. (2018). A survey on automatic image caption generation. *Neurocomputing*, *311*, 291–304.

Banerjee, S., & Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Belz, A., & Reiter, E. (2006). Comparing automatic and human evaluation of nlg systems. *11th conference of the european chapter of the association for computational linguistics*, 313–320.

Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., & Plank, B. (2016). Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, *55*, 409–442.

Bigham, J. P., Lin, I., & Savage, S. (2017). The effects of" not knowing what you don't know" on web accessibility for blind web users. *Proceedings of the 19th international ACM SIGACCESS conference on computers and accessibility*, 101–109.

Borges, M., Brasil, A., Oliveira, Z., & Silva, J. (2020). I don't see, but i'm seen -a study of garment consumption by visually impaired people. *19*, 186–202.

Borlund, P. (2000). Experimental components for the evaluation of interactive information retrieval systems. *Journal of documentation*, *56*(1), 71–90.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, *3*(2), 77–101.

Braun, V., & Clarke, V. (2012). *Thematic analysis.* American Psychological Association.

Brown, R. L., & Barrett, A. E. (2011). Visual impairment and quality of life among older adults: An examination of explanations for the relationship. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *66*(3), 364–373.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (pp. 1877–1901). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Brulé, E., Tomlinson, B. J., Metatla, O., Jouffrais, C., & Serrano, M. (2020). Review of quantitative empirical evaluations of technology for people with visual impairments. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14.

Bualar, T. (2018). Barriers to inclusive higher education in thailand: Voices of blind students. *Asia Pacific Education Review*, *19*(4), 469–477.

Burton, M. A. (2011). Fashion for the blind: A study of perspectives. *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*, 315–316.

Burton, M. A., Brady, E., Brewer, R., Neylan, C., Bigham, J. P., & Hurst, A. (2012). Crowdsourcing subjective fashion advice using vizwiz: Challenges and opportunities. *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility*, 135–142.

Cai, C., Yap, K.-H., & Wang, S. (2022). Attribute conditioned fashion image captioning. *2022 IEEE International Conference on Image Processing (ICIP)*, 1921–1925.

Carter, S. M., Shih, P., Williams, J., Degeling, C., & Mooney-Somers, J. (2021). Conducting qualitative research online: Challenges and solutions. *The Patient-Patient-Centered Outcomes Research*, *14*(6), 711–718.

Castillo-Montoya, M. (2016). Preparing for interview research: The interview protocol refinement framework. *Qualitative Report*, *21*(5).

Celikyilmaz, A., Clark, E., & Gao, J. (2021). Evaluation of text generation: A survey.

Chen, J., Guo, H., Yi, K., Li, B., & Elhoseiny, M. (2022). Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18030–18040.

Chia, P. J., Attanasio, G., Bianchi, F., Terragni, S., Magalhães, A. R., Goncalves, D., Greco, C., & Tagliabue, J. (2022). Fashionclip: Connecting language and images for product representations. *arXiv preprint arXiv:2204.03972*.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., & Xing, E. P. (2023). Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. https://lmsys.org/blog/2023-03-30-vicuna/

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., ... Fiedel, N. (2022). Palm: Scaling language modeling with pathways.

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., & Smith, N. A. (2021). All that's 'human' is not gold: Evaluating human evaluation of generated text.

Cornia, M., Stefanini, M., Baraldi, L., & Cucchiara, R. (2020). Meshed-memory transformer for image captioning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Deng, C., Ding, N., Tan, M., & Wu, Q. (2020). Length-controllable image captioning. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, 712–729.

Devlin, J., Cheng, H., Fang, H., Gupta, S., Deng, L., He, X., Zweig, G., & Mitchell, M. (2015). Language models for image captioning: The quirks and what works. *arXiv preprint arXiv:1505.01809*.

Dhingra, B., Faruqui, M., Parikh, A., Chang, M.-W., Das, D., & Cohen, W. W. (2019). Handling divergent reference texts when evaluating table-to-text generation.

DiCicco-Bloom, B., & Crabtree, B. F. (2006). The qualitative research interview. *Medical education*, *40*(4), 314–321.

Dikko, M. (2016). Establishing construct validity and reliability: Pilot testing of a qualitative interview for research in takaful (islamic insurance). *Qualitative Report*, *21*(3).

Du, S., Zhu, H., Lin, G., Wang, D., Shi, J., & Wang, J. (2023). Object semantic analysis for image captioning. *Multimedia Tools and Applications*, 1–28.

Eriksson, P., & Kovalainen, A. (2015). *Qualitative methods in business research: A practical guide to social research*. Sage.

Fang, H., Gupta, S., Iandola, F., Srivastava, R. K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J. C., et al. (2015). From captions to visual concepts and back. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1473–1482.

Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., & Forsyth, D. (2010). Every picture tells a story: Generating sentences from images. In K. Daniilidis, P. Maragos, & N. Paragios (Eds.), *Computer vision – eccv 2010* (pp. 15–29). Springer Berlin Heidelberg.

Farzana, J., Muhammad, A., Martinez-Enriquez, A., Afraz, Z., & Talha, W. (2014). Speech based shopping assistance for the blind. *Mexican Conference on Pattern Recognition*, 361–370.

Fei, Z. (2022). Attention-aligned transformer for image captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(1), 607–615. https://doi.org/10.1609/aaai.v36i1.19940

Fuchs, E., & Strauss, C. (2012). Online shopping involving consumers with visual impairments–a qualitative study. *International Conference on Computers for Handicapped Persons*, 378–385.

Gan, Z., Gan, C., He, X., Pu, Y., Tran, K., Gao, J., Carin, L., & Deng, L. (2017). Semantic compositional networks for visual captioning. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5630–5639.

Gao, D., Jin, L., Chen, B., Qiu, M., Li, P., Wei, Y., Hu, Y., & Wang, H. (2020). Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2251–2260.

Gatt, A., & Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, *61*, 65–170.

Ge, Y., Zhang, R., Wang, X., Tang, X., & Luo, P. (2019). Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5337–5345.

Gharaei, N. Y., Dadkhah, C., & Daryoush, L. (2021). Content-based clothing recommender system using deep neural network. *2021 26th International Computer Conference, Computer Society of Iran (CSICC)*, 1–6.

Goenka, S., Zheng, Z., Jaiswal, A., Chada, R., Wu, Y., Hedau, V., & Natarajan, P. (2022). Fashionvlp: Vision language transformer for fashion retrieval with feedback. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14105–14115.

Goferman, S., Zelnik-Manor, L., & Tal, A. (2011). Context-aware saliency detection. *IEEE transactions on pattern analysis and machine intelligence*, *34*(10), 1915–1926.

Gupta, A., Verma, Y., & Jawahar, C. (2012). Choosing linguistics over vision to describe images. *Proceedings of the AAAI conference on artificial intelligence*, *26*(1), 606–612.

Habib, F., & Irshad, E. (2018). Impact of visual impairment on quality of life among adolescents. *FWU Journal of Social Sciences*, *12*(1).

Hacheme, G., & Sayouti, N. (2021). Neural fashion image captioning: Accounting for data diversity. *arXiv preprint arXiv:2106.12154*.

Han, X., Yu, L., Zhu, X., Zhang, L., Song, Y.-Z., & Xiang, T. (2022). Fashionvil: Fashion-focused vision-and-language representation learning. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, & T. Hassner (Eds.), *Computer vision – eccv 2022* (pp. 634–651). Springer Nature Switzerland.

Han, X., Zhu, X., Yu, L., Zhang, L., Song, Y.-Z., & Xiang, T. (2023). Fame-vil: Multi-tasking vision-language model for heterogeneous fashion tasks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2669–2680.

Han, Y., Zhang, L., Chen, Q., Chen, Z., Li, Z., Yang, J., & Cao, Z. (2023). Fashionsap: Symbols and attributes prompt for fine-grained fashion vision-language pre-training. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15028–15038.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. communication methods and measures, 1 (1), 77-89.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2018). Mask r-cnn.

Hennink, M., Hutter, I., & Bailey, A. (2020). *Qualitative research methods*. Sage.

Hessel, J., Holtzman, A., Forbes, M., Bras, R. L., & Choi, Y. (2021). Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Hodosh, M., Young, P., & Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, *47*, 853–899.

Hossain, M. Z., Sohel, F., Shiratuddin, M. F., & Laga, H. (2019). A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, *51*(6), 1–36.

Jamshed, S. (2014). Qualitative research method-interviewing and observation. *Journal of basic and clinical pharmacy*, *5*(4), 87.

Jeamwatthanachai, W., Wald, M., & Wills, G. (2019). Indoor navigation by blind people: Behaviors and challenges in unfamiliar spaces and buildings. *British Journal of Visual Impairment*, *37*(2), 140–153.

Jiang, Y., Yang, S., Qju, H., Wu, W., Loy, C. C., & Liu, Z. (2022). Text2human: Text-driven controllable human image generation. *ACM Transactions on Graphics (TOG)*, *41*(4), 1–11.

Jocher, G. (2020). *YOLOv5 by Ultralytics* (Version 7.0). https://doi.org/10.5281/zenodo.3908559

Jones, N., Bartlett, H. E., & Cooke, R. (2019). An analysis of the impact of visual impairment on activities of daily living and vision-related quality of life in a visually impaired adult population. *British Journal of Visual Impairment*, *37*(1), 50–63.

Karim, K. A. A. (2020). The challenges faced by visual impairments in buying clothes and how to wear clothes. *Idealogy Journal of Arts and Social Science*, *5*(2), 71–76.

Karpathy, A., Joulin, A., & Fei-Fei, L. (2014). Deep fragment embeddings for bidirectional image sentence mapping. *arXiv preprint arXiv:1406.5679*.

Kaufman-Scarborough, C., & Childers, T. L. (2009). Understanding markets as online public places: Insights from consumers with visual impairments. *Journal of Public Policy & Marketing*, *28*(1), 16–28.

Kinghorn, P., Zhang, L., & Shao, L. (2018). A region-based image caption generator with refined descriptions. *Neurocomputing*, *272*, 416–424. https://doi.org/https://doi.org/10.1016/j.neucom.2017.07.014

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). Segment anything.

Kreiss, E., Bennett, C., Hooshmand, S., Zelikman, E., Morris, M. R., & Potts, C. (2022). Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics.

Krishnamoorthy, N., Malkarnenkar, G., Mooney, R., Saenko, K., & Guadarrama, S. (2013). Generating natural-language video descriptions using text-mined knowledge. *Proceedings of the AAAI conference on artificial intelligence*, *27*(1), 541–547.

Kuznetsova, P., Ordonez, V., Berg, T. L., & Choi, Y. (2014). Treetalk: Composition and compression of trees for image descriptions. *Transactions of the Association for Computational Linguistics*, *2*, 351–362.

Lan, W., Li, X., & Dong, J. (2017). Fluency-guided cross-lingual image captioning. *Proceedings of the 25th ACM International Conference on Multimedia*, 1549–1557. https://doi.org/10.1145/3123266.3123366

Lebret, R., Pinheiro, P., & Collobert, R. (2015). Phrase-based image captioning. *International Conference on Machine Learning*, 2085–2094.

Lee, J., Herskovitz, J., Peng, Y.-H., & Guo, A. (2022). Imageexplorer: Multi-layered touch exploration to encourage skepticism towards imperfect ai-generated image captions. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–15.

Lee, J., Kim, J., & Jung, H. (2020). Challenges and design opportunities for easy, economical, and accessible offline shoppers with visual impairments. *Proceedings of the 2020 Symposium on Emerging Research from Asia and on Asian Contexts and Cultures*, 69–72.

Lenth, R. V. (2004). Power analysis for experimental research: A practical guide for the biological, medical and social sciences and statistical power analysis: A simple and general model for traditional and modern hypothesis tests.

Li, F. M., Dorst, J., Cederberg, P., & Carrington, P. (2021). Non-visual cooking: Exploring practices and challenges of meal preparation by people with visual impairments. *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, 1–11.

Li, J., Li, D., Savarese, S., & Hoi, S. (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Li, S., & Yamaguchi, K. (2017). Attention to describe products with attributes. *2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA)*, 215–218. https://doi.org/10.23919/MVA.2017.7986839

Li, X., Jiang, S., & Han, J. (2019). Learning object context for dense captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 8650–8657. https://doi.org/10.1609/aaai.v33i01.33018650

Li, X., Ye, Z., Zhang, Z., & Zhao, M. (2021). Clothes image caption generation with attribute detection and visual attention model. *Pattern Recognition Letters*, *141*, 68–74.

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al. (2020). Oscar: Object-semantics aligned pre-training for vision-language tasks. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, 121–137.

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., & Wen, J.-R. (2023). Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Lin, C.-Y., & Hovy, E. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. *Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics*, 150–157.

Lin, H.-T., Lin, C.-J., & Weng, R. C. (2007). A note on platt's probabilistic outputs for support vector machines. *Machine learning*, *68*, 267–276.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755.

Liu, G., Ding, X., Yu, C., Gao, L., Chi, X., & Shi, Y. (2019). " i bought this for me to look more ordinary": A study of blind people doing online shopping. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–11.

Liu, J.-J., Hou, Q., Cheng, M.-M., Feng, J., & Jiang, J. (2019). A simple pooling-based design for real-time salient object detection. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3917–3926.

Liu, S., Zhu, Z., Ye, N., Guadarrama, S., & Murphy, K. (2017). Improved image captioning via policy gradient optimization of spider. *Proceedings of the IEEE international conference on computer vision*, 873–881.

Liu, Z., Luo, P., Qiu, S., Wang, X., & Tang, X. (2016). Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1096–1104.

Long, Y., Wen, Y., Han, J., Xu, H., Ren, P., Zhang, W., Zhao, S., & Liang, X. (2023). Capdet: Unifying dense captioning and open-world detection pretraining. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15233–15243.

Lovenia, H., Dai, W., Cahyawijaya, S., Ji, Z., & Fung, P. (2023). Negative object presence evaluation (nope) to measure object hallucination in vision-language models. *arXiv preprint arXiv:2310.05338*.

Lu, J., Xiong, C., Parikh, D., & Socher, R. (2017). Knowing when to look: Adaptive attention via a visual sentinel for image captioning. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 375–383.

Lu, P., Peng, B., Cheng, H., Galley, M., Chang, K.-W., Wu, Y. N., Zhu, S.-C., & Gao, J. (2023). Chameleon: Plug-and-play compositional reasoning with large language models.

Lu, Y., Lu, P., Chen, Z., Zhu, W., Wang, X. E., & Wang, W. Y. (2023). Multimodal procedural planning via dual text-image prompting.

Lüddecke, T., & Ecker, A. (2022). Image segmentation using text and image prompts. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7086–7096.

Luo, G., Cheng, L., Jing, C., Zhao, C., & Song, G. (2022). A thorough review of models, evaluation metrics, and datasets on image captioning. *IET Image Processing*, *16*(2), 311–332.

Luo, Y., Ji, J., Sun, X., Cao, L., Wu, Y., Huang, F., Lin, C.-W., & Ji, R. (2021). Dual-level collaborative transformer for image captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(3), 2286–2293. https://doi.org/10.1609/aaai.v35i3.16328

MacLeod, H., Bennett, C. L., Morris, M. R., & Cutrell, E. (2017). Understanding blind people's experiences with computer-generated captions of social media images. *proceedings of the 2017 CHI conference on human factors in computing systems*, 5988–5999.

Manay, S. P., Yaligar, S. A., Thathva Sri Sai Reddy, Y., & Saunshimath, N. J. (2022). Image captioning for the visually impaired. *Emerging Research in Computing, Information, Communication and Applications: ERCICA 2020, Volume 1*, 511–522.

Mao, J., Xu, W., Yang, Y., Wang, J., & Yuille, A. L. (2014). Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*.

Marzi, G., Balzano, M., & Marchiori, D. (2024). K-alpha calculator–krippendorff's alpha calculator: A user-friendly tool for computing krippendorff's alpha inter-rater reliability coefficient. *MethodsX*, *12*, 102545.

Mason, R., & Charniak, E. (2014). Nonparametric method for data-driven image captioning. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 592–598.

Mazurowski, M. A., Buda, M., Saha, A., & Bashir, M. R. (2019). Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on mri. *Journal of magnetic resonance imaging*, *49*(4), 939–954.

Ming, Y., Hu, N., Fan, C., Feng, F., Zhou, J., & Yu, H. (2022). Visuals to text: A comprehensive review on automatic image captioning. *IEEE/CAA Journal of Automatica Sinica*, *9*(8), 1339–1365.

Mirchandani, S., Yu, L., Wang, M., Sinha, A., Jiang, W., Xiang, T., & Zhang, N. (2022). Fad-vlp: Fashion vision-and-language pre-training towards unified retrieval and captioning.

Mitchell, M., Dodge, J., Goyal, A., Yamaguchi, K., Stratos, K., Han, X., Mensch, A., Berg, A., Berg, T., & Daumé III, H. (2012). Midge: Generating image descriptions from computer vision detections. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 747–756.

Moratelli, N., Barraco, M., Morelli, D., Cornia, M., Baraldi, L., & Cucchiara, R. (2023a). Fashion-oriented image captioning with external knowledge retrieval and fully attentive gates. *Sensors*, *23*(3), 1286.

Moratelli, N., Barraco, M., Morelli, D., Cornia, M., Baraldi, L., & Cucchiara, R. (2023b). Fashion-oriented image captioning with external knowledge retrieval and fully attentive gates. *Sensors*, *23*(3). https://doi.org/10.3390/s23031286

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2023). A comprehensive overview of large language models.

Nguyen, B. T., Prakash, O., & Vo, A. H. (2021). Attention mechanism for fashion image captioning. In Y.-P. Huang, W.-J. Wang, H. A. Quoc, L. H. Giang, & N.-L. Hung (Eds.), *Computational intelligence methods for green technology and sustainable development* (pp. 93–104). Springer International Publishing.

Oliver, M. (2013). The social model of disability: Thirty years on. *Disability & society*, *28*(7), 1024–1026.

OpenAI. (2023). Gpt-4 technical report.

Ordonez, V., Kulkarni, G., & Berg, T. (2011). Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, *24*.

Otyola, W. R., Kibanja, G. M., & Mugagga, A. M. (2017). Challenges faced by visually impaired students at makerere and kyambogo universities. *Makerere Journal of Higher Education*, *9*(1), 75–86.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in neural information processing systems* (pp. 27730–27744). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Park, J., Han, Y., & Oh, U. (2020). Understanding smartphone-based online shopping experiences and behaviors of blind users. *International journal of advanced smart convergence*, *9*(3), 260–271.

Pedersoli, M., Lucas, T., Schmid, C., & Verbeek, J. (2017). Areas of attention for image captioning. *Proceedings of the IEEE international conference on computer vision*, 1242–1250.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. *International conference on machine learning*, 8748–8763.

Rahman, I. U., Wang, Z., Liu, W., Ye, B., Zakarya, M., & Liu, X. (2020). An n-state markovian jumping particle swarm optimization algorithm. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *51*(11), 6626–6638.

Rapp, A. (2021). In search for design elements: A new perspective for employing ethnography in human-computer interaction design research. *International Journal of Human–Computer Interaction*, *37*(8), 783–802.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf

Rocha, D., Carvalho, V., Soares, F., & Oliveira, E. (2019). Extracting clothing features for blind people using image processing and machine learning techniques: First insights. In J. M. R. S. Tavares & R. M. Natal Jorge (Eds.), *Vipimage 2019* (pp. 411–418). Springer International Publishing.

Rocha, D., Carvalho, V., Soares, F., & Oliveira, E. (2021). A model approach for an automatic clothing combination system for blind people. In E. I. Brooks, A. Brooks, C. Sylla, & A. K. Møller (Eds.), *Design, learning, and innovation* (pp. 74–85). Springer International Publishing.

Rocha, D., Pinto, L., Machado, J., Soares, F., & Carvalho, V. (2023). Using object detection technology to identify defects in clothing for blind people. *Sensors*, *23*(9). https://doi.org/10.3390/s23094381

Rocha, D., Soares, F., Oliveira, E., & Carvalho, V. (2023). Blind people: Clothing category classification and stain detection using transfer learning. *Applied Sciences*, *13*(3), 1925.

Roh, M., Selivanova, A., Shin, H. J., Miller, J. W., & Jackson, M. L. (2018). Visual acuity and contrast sensitivity are two important factors affecting vision-related quality of life in advanced age-related macular degeneration. *PloS one*, *13*(5), e0196481.

Rostamzadeh, N., Hosseini, S., Boquet, T., Stokowiec, W., Zhang, Y., Jauvin, C., & Pal, C. (2018). Fashion-gen: The generative fashion dataset and challenge. *arXiv preprint arXiv:1806.08317*.

Sadeh, G., Fritz, L., Shalev, G., & Oks, E. (2019). Generating diverse and informative natural language fashion feedback. *arXiv preprint arXiv:1906.06619*.

Sahasrabudhe, S., & Palvia, P. (2013). Academic challenges of blind students and their mitigation strategies. *Proceedings of the Nineteenth Americas Conference on Information Systems, Chicago, Illinois*, 15–17.

Saleh, B., & Elgammal, A. (2015). Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *arXiv preprint arXiv:1505.00855*.

Sarto, S., Barraco, M., Cornia, M., Baraldi, L., & Cucchiara, R. (2023). Positive-augmented contrastive learning for image and video captioning evaluation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6914–6924.

Schuff, H., Vanderlyn, L., Adel, H., & Vu, N. T. (2023). How to do human evaluation: A brief introduction to user studies in nlp. *Natural Language Engineering*, *29*(5), 1199–1222. https://doi.org/10.1017/S1351324922000535

Sharma, D., Dhiman, C., & Kumar, D. (2023). Evolution of visual data captioning methods, datasets, and evaluation metrics: A comprehensive survey. *Expert Systems with Applications*, *221*, 119773. https://doi.org/https://doi.org/10.1016/j.eswa.2023.119773

Sharma, H., Agrahari, M., Singh, S. K., Firoj, M., & Mishra, R. K. (2020). Image captioning: A comprehensive survey. *2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC)*, 325–328.

Sharma, P., Ding, N., Goodman, S., & Soricut, R. (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2556–2565.

Shen, Y., Song, K., Tan, X., Li, D., Lu, W., & Zhuang, Y. (2023). Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face.

Shi, Z., Zhou, X., Qiu, X., & Zhu, X. (2020). Improving image captioning with better use of captions. *arXiv preprint arXiv:2006.11807*.

Shimorina, A., & Belz, A. (2021). The human evaluation datasheet 1.0: A template for recording details of human evaluation experiments in nlp.

Silva, L., Rocha, D., Soares, F., Esteves, J. S., & Carvalho, V. (2023). Automatic system to identify and manage garments for blind people. *Acta IMEKO*, *12*(3), 1–10.

Simmons, G. (2022). Moral mimicry: Large language models produce moral rationalizations tailored to political identity. *arXiv preprint arXiv:2209.12106*.

Simons, R. N., Gurari, D., & Fleischmann, K. R. (2020). " i hope this is helpful" understanding crowdworkers' challenges and motivations for an image description task. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), 1–26.

Socher, R., Karpathy, A., Le, Q. V., Manning, C. D., & Ng, A. Y. (2014). Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, *2*, 207–218.

Sohaib, O., & Kang, K. (2017). E-commerce web accessibility for people with disabilities. *Complexity in information systems development* (pp. 87–100). Springer.

SS, R. A., Priyadharshini, M., & Kalinathan, L. (2023). Image caption generation for blind users of social media websites.

Stangl, A., Morris, M. R., & Gurari, D. (2020a). " person, shoes, tree. is the person naked?" what people with vision impairments want in image descriptions. *Proceedings of the 2020 chi conference on human factors in computing systems*, 1–13.

Stangl, A., Morris, M. R., & Gurari, D. (2020b). "person, shoes, tree. is the person naked?" what people with vision impairments want in image descriptions. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3313831.3376404

Stangl, A., Verma, N., Fleischmann, K. R., Morris, M. R., & Gurari, D. (2021). Going beyond one-size-fits-all image descriptions to satisfy the information wants of people who are blind or have low vision. *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*. https://doi.org/10.1145/3441852.3471233

Stangl, A. J., Kothari, E., Jain, S. D., Yeh, T., Grauman, K., & Gurari, D. (2018). Browsewithme: An online clothes shopping assistant for people with visual impairments. *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, 107–118.

Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., & Cucchiara, R. (2021). From show to tell: A survey on image captioning. *arXiv preprint arXiv:2107.06912*.

Sun, C., Gan, C., & Nevatia, R. (2015). Automatic concept discovery from parallel text and visual corpora. *Proceedings of the IEEE international conference on computer vision*, 2596–2604.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, *27*.

Szpiro, S., Zhao, Y., & Azenkot, S. (2016). Finding a store, searching for a product: A study of daily challenges of low vision people. *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 61–72.

Tang, Y., Zhang, L., & Yuan, Y. (2023). Fashion item captioning via grid-relation self-attention and gated-enhanced decoder. *Multimedia Tools and Applications*, 1–25.

Tang, Y., Zhang, L., & Yuan, Y. (2024). Fashion item captioning via grid-relation self-attention and gated-enhanced decoder. *Multimedia Tools and Applications*, *83*(3), 7631–7655.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., & Hashimoto, T. B. (2023). Stanford alpaca: An instruction-following llama model.

Tateno, K., Takagi, N., Sawai, K., Masuta, H., & Motoyoshi, T. (2020). Method for generating captions for clothing images to support visually impaired people. *2020 Joint 11th International Conference on Soft Computing and Intelligent Systems and 21st International Symposium on Advanced Intelligent Systems (SCIS-ISIS)*, 1–5.

Thoppilan, R., Freitas, D. D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., Li, Y., Lee, H., Zheng, H. S., Ghafouri, A., Menegali, M., Huang, Y., Krikun, M., Lepikhin, D., Qin, J., ... Le, Q. (2022). Lamda: Language models for dialog applications.

Tiwary, T., & Mahapatra, R. P. (2023). An accurate generation of image captions for blind people using extended convolutional atom neural network. *Multimedia Tools and Applications*, *82*(3), 3801–3830.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). Llama: Open and efficient foundation language models.

Tu, G. (2020). Image captioning on general data and fashion data: An attribute-image-combined attention-based network for image captioning on mutli-object images and single-object images.

Ushiku, Y., Harada, T., & Kuniyoshi, Y. (2012). Efficient image annotation for automatic sentence generation. *Proceedings of the 20th ACM international conference on Multimedia*, 549–558.

Ushiku, Y., Yamaguchi, M., Mukuta, Y., & Harada, T. (2015). Common subspace for model and similarity: Phrase learning for caption generation from images. *Proceedings of the IEEE international conference on computer vision*, 2668–2676.

Vaismoradi, M., Turunen, H., & Bondas, T. (2013). Content analysis and thematic analysis: Implications for conducting a qualitative descriptive study. *Nursing & health sciences*, *15*(3), 398–405.

van der Lee, C., Gatt, A., van Miltenburg, E., & Krahmer, E. (2021). Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, *67*, 101151. https://doi.org/https://doi.org/10.1016/j.csl.2020.101151

van der Lee, C., Gatt, A., van Miltenburg, E., Wubben, S., & Krahmer, E. (2019). Best practices for the human evaluation of automatically generated text. *Proceedings of the 12th International Conference on Natural Language Generation*, 355–368. https://doi.org/10.18653/v1/W19-8643

van Miltenburg, E. (2023). Evaluating nlg systems: A brief introduction. *arXiv preprint arXiv:2303.16742*.

van Miltenburg, E., Clinciu, M.-A., Dušek, O., Gkatzia, D., Inglis, S., Leppänen, L., Mahamood, S., Manning, E., Schoch, S., Thomson, C., & Wen, L. (2021). Underreporting of errors in nlg output, and what to do about it.

Vashistha, A., Sethi, P., & Anderson, R. (2018). Bspeak: An accessible voice-based crowdsourcing marketplace for low-income blind people. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3173574.3173631

Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3156–3164.

Wang, R., Chen, Z., Zhang, M. R., Li, Z., Liu, Z., Dang, Z., Yu, C., & Chen, X. (2021). Revamp: Enhancing accessible information seeking experience of online shopping for blind or low vision users. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–14.

Wang, T., Zhang, J., Fei, J., Zheng, H., Tang, Y., Li, Z., Gao, M., & Zhao, S. (2023). Caption anything: Interactive image description with diverse multimodal controls.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent abilities of large language models.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837.

Whitney, M. (2020). Teaching accessible design: Integrating accessibility principles and practices into an introductory web design course. *Information Systems Education Journal*, *18*(1), 4–13.

Williams, J. B. (2021). *The online apparel shopping experience of blind consumers*. North Carolina State University.

Williams, M. A., Hurst, A., & Kane, S. K. (2013). " pray before you step out" describing personal and situational blind navigation behaviors. *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*, 1–8.

Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., & Duan, N. (2023). Visual chatgpt: Talking, drawing and editing with visual foundation models.

Wu, Q., Shen, C., Liu, L., Dick, A., & Van Den Hengel, A. (2016). What value do explicit high level concepts have in vision to language problems? *Proceedings of the IEEE conference on computer vision and pattern recognition*, 203–212.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In F. Bach & D. Blei (Eds.), *Proceedings of the 32nd international conference on machine learning* (pp. 2048–2057). PMLR. https://proceedings.mlr.press/v37/xuc15.html

Yan, B. (2023). A cnn-lstm-based model for fashion image aesthetic captioning. *Third International Conference on Computer Vision and Data Mining (ICCVDM 2022)*, *12511*, 309–315.

Yan, F., & Mikolajczyk, K. (2015). Deep correlation for matching images and text. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3441–3450.

Yang, D., Chen, K., Rao, J., Guo, X., Zhang, Y., Yang, J., & Zhang, Y. (2023). Tackling vision language tasks through learning inner monologues.

Yang, H., Peng, Q., Gao, Q., & Patrick Rau, P.-L. (2015). Design of a clothing shopping guide website for visually impaired people. In P. Rau (Ed.), *Cross-cultural design methods, practice and impact* (pp. 253–261). Springer International Publishing.

Yang, R., Song, L., Li, Y., Zhao, S., Ge, Y., Li, X., & Shan, Y. (2023). Gpt4tools: Teaching large language model to use tools via self-instruction.

Yang, W., Luo, P., & Lin, L. (2014). Clothing co-parsing by joint image segmentation and labeling. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3182–3189.

Yang, X., Yuan, S., & Tian, Y. (2011). Recognizing clothes patterns for blind people by confidence margin based feature combination. *Proceedings of the 19th ACM International Conference on Multimedia*, 1097–1100. https://doi.org/10.1145/2072298.2071947

Yang, X., Zhang, H., Jin, D., Liu, Y., Wu, C.-H., Tan, J., Xie, D., Wang, J., & Wang, X. (2020). Fashion captioning: Towards generating accurate descriptions with semantic rewards. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, 1–17.

Yang, Y., Teo, C., Daumé III, H., & Aloimonos, Y. (2011). Corpus-guided sentence generation of natural images. *Proceedings of the 2011 conference on empirical methods in natural language processing*, 444–454.

Yang, Z., Li, L., Wang, J., Lin, K., Azarnasab, E., Ahmed, F., Liu, Z., Liu, C., Zeng, M., & Wang, L. (2023). Mm-react: Prompting chatgpt for multimodal reasoning and action.

Yatskar, M., Galley, M., Vanderwende, L., & Zettlemoyer, L. (2014). See no evil, say no evil: Description generation from densely labeled images. *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (* SEM 2014)*, 110–120.

Yi, Y., Deng, H., & Hu, J. (2020). Improving image captioning evaluation by considering inter references variance. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 985–994.

Yin, G., Sheng, L., Liu, B., Yu, N., Wang, X., & Shao, J. (2019). Context and attribute grounded dense captioning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6241–6250.

Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., & Chen, E. (2023). A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

You, H., Sun, R., Wang, Z., Chen, L., Wang, G., Ayyubi, H. A., Chang, K.-W., & Chang, S.-F. (2023). Idealgpt: Iteratively decomposing vision and language reasoning via large language models.

You, Q., Jin, H., Wang, Z., Fang, C., & Luo, J. (2016). Image captioning with semantic attention. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4651–4659.

Yu, H., Tullio-Pow, S., & Akhtar, A. (2015). Retail design and the visually impaired: A needs assessment. *Journal of Retailing and Consumer Services*, *24*, 121–129.

Yu, W., Zhang, H., He, X., Chen, X., Xiong, L., & Qin, Z. (2018). Aesthetic-based clothing recommendation. *Proceedings of the 2018 world wide web conference*, 649–658.

Yuan, C. W., Hanrahan, B. V., Lee, S., Rosson, M. B., & Carroll, J. M. (2019). Constructing a holistic view of shopping with people with visual impairment: A participatory design approach. *Universal Access in the Information Society*, *18*(1), 127–140.

Zhang, X., Gregg, E. W., Cheng, Y. J., Thompson, T. J., Geiss, L. S., Duenas, M. R., & Saaddine, J. B. (2008). Diabetes Mellitus and Visual Impairment: National Health and Nutrition Examination Survey, 1999-2004. *Archives of Ophthalmology*, *126*(10), 1421–1427. https://doi.org/10.1001/archopht.126.10.1421

Zhang, X., Sun, X., Luo, Y., Ji, J., Zhou, Y., Wu, Y., Huang, F., & Ji, R. (2021). Rstnet: Captioning with adaptive attention on visual and non-visual words. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15465–15474.

Zhao, Y., Kupferstein, E., Tal, D., & Azenkot, S. (2018). " it looks beautiful but scary" how low vision people navigate stairs and other surface level changes. *Proceedings of the*

20th International ACM SIGACCESS Conference on Computers and Accessibility, 307–320.

Zhu, D., Chen, J., Haydarov, K., Shen, X., Zhang, W., & Elhoseiny, M. (2023). Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions. *arXiv preprint arXiv:2303.06594*.

Zhu, D., Chen, J., Shen, X., Li, X., & Elhoseiny, M. (2023). Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Zhuge, M., Gao, D., Fan, D.-P., Jin, L., Chen, B., Zhou, H., Qiu, M., & Shao, L. (2021). Kaleido-bert: Vision-language pre-training on fashion domain. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12647–12657.

Appendices

Appendix A

Empirical Study Documents

The University Of Sheffield.

Information School

Participant Information Sheet - Study 1

Project Title: Assisting the visually impaired and blind people in online shopping

You are being invited to take part in a research project. Before you decide whether or not to participate, it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask us if there is anything that is not clear or if you would like more information. Take time to decide whether or not you wish to take part.

---

1. What is the purpose of the study?

The purpose of this study is to explore the potential challenges visually impaired and blind people might experience when selecting and coordinating outfits online.

2. Why have I been invited?

You have been invited to this study because you are of age 18 or older, either blind or partially sighted and have some prior experience with online shopping.

3. Do I have to take part?

By participating in this study, you will receive a £10 Amazon voucher. It is up to you to decide whether or not to take part. All information related to this study along with the informed consent form will be emailed to you. It would be appreciated if you could email back to show that you are agree to take part in this study. Before starting the interview, the researcher will explain all information related to this study. Moreover, your consent will be taken verbally and recorded to maintain integrity. You are free to withdraw at any time within 2 weeks, without giving a reason. If you choose to withdraw your participation then any and all data collected about you and during any studies you have participated in will be deleted.

Please note that by choosing to participate in this research, this will not create a legally binding agreement, nor is it intended to create an employment relationship between you and the University of Sheffield.

4. What will happen to me if I take part?

If you agree to participate, the researcher will arrange a meeting to conduct the study.

This study includes task observation and a follow-up interview. In the task observation, you will be provided with an imaginary scenario and asked to select and coordinate new outfits online. Then, a follow-up interview will be conducted to ask questions related to the observation process. During the task observation, a video recording will capture only the screen to record your behaviour when performing the planned tasks. Once the interview started the video recording will be stopped and start an audio recording. To ensure a comfortable environment, you will be informed when an audio or video recording begins.

The purpose of the recording is to allow the researcher to capture all the information discussed during the task observation and interview, which is important to permit later analysis. The task observation will take about 30 minutes and followed by 20 minutes to perform the interview.

5. What will I have to do?

During the task observation, you will be asked to complete a number of tasks. This includes sharing your screen (Please let the researcher know ahead of time if you need help sharing your screen), accessing the Amazon website (please do not log in to your own account), as well as performing shopping tasks such as outfit selection and coordination. To better understand your shopping behaviour, you will be asked to think aloud while the researcher is quietly observing and taking notes. The observation concludes either by confirming that you have completed the tasks or by indicating that you wish to terminate the observation.

In the follow-up interview, you will be asked questions related to the selection tasks you have just completed. However, you may refuse to perform any tasks or answer any questions that make you feel uncomfortable, and you can stop the interview at any time without any negative consequences.

6. What are the possible disadvantages and risks of taking part?

There are no known disadvantages or risks associated with this research. However, if you feel uncomfortable at any point and no longer wish to continue then you only need to mention this to the researcher and the study will be terminated and any collected data will be deleted.

7. What are the possible benefits of taking part?

While there are no immediate benefits for those people participating in the project, it is hoped that this work will have a beneficial impact on improving the shopping services in online stores for visually impaired and blind people. Moreover, the participants will have the opportunity to try and test any automatic solution that will be built as a result of this study.

8. Is the purpose of this study educational?

Yes. the data in this research will be used for a Ph.D. study.

9. Will my taking part in this project be kept confidential?

All the information that we collect about you during the course of the research will be kept strictly confidential and will only be accessible to members of the research team. The collected data will be deposited in the secure and password-protected University of Sheffield repository, and other researchers may use the collected data only if they agree to preserve the confidentiality of the information as requested in this project.

10. What will happen to the data collected, and the results of the research project?

The collected data will be transcribed and analysed, then the result will be published. You will not be individually identified in any report, publication, or presentation without seeking your full consent. Direct quotes from the interviews may be used in reports and publications; however, these will be anonymised to ensure that you cannot be identified.

11. Will I be recorded, and how will the recorded media be used?

The audio and/or video recordings of your activities made during this research will be used only for analysis and for illustration in conference presentations and lectures. No other use will be made of them without your written permission, and no one outside the project will be allowed access to the original recordings. Your likeness will not be included in any publications resulting from this research.
Remember, you will not be identified in any report, publication or presentation without seeking your full consent.

12. Who is organising and funding the research?

The University of Sheffield is the sponsor of this study, and Umm Al Qura University is funding this research.

13. Who is the Data Controller?

The University of Sheffield will act as the Data Controller for this study. This means that the University of Sheffield is responsible for looking after your information and using it properly.

14. Who has ethically reviewed the project?

This project has been ethically approved via the University of Sheffield's Ethics Review Procedure, as administered by the Information school.

15. What if something goes wrong and I wish to complain about the research or report a concern or incident?

If you are dissatisfied with any aspect of the research and wish to make a complaint, please contact *Dr Morgan Harvey; m.harvey@sheffield.ac.uk* in the first instance. If you feel your complaint has not been handled in a satisfactory way you can contact the Head of the information school *Professor Val Gillet; v.gillet@sheffield.ac.uk* If the complaint relates to how your personal data has been handled, you can find information about how to raise a complaint in the University's Privacy Notice: https://www.sheffield.ac.uk/govern/data-protection/privacy/general.

If you wish to make a report of a concern or incident relating to potential exploitation, abuse or harm resulting from your involvement in this project, please contact the project's Designated Safeguarding Contact *Dr Morgan Harvey; m.harvey@sheffield.ac.uk*. If the concern or incident relates to the Designated Safeguarding Contact, or if you feel a report you have made to this Contact has not been handled in a satisfactory way, please contact the Head of the information school *Professor Val Gillet; v.gillet@sheffield.ac.uk* and/or the University's Research Ethics & Integrity Manager (Lindsay Unwin; l.v.unwin@sheffield.ac.uk).

16. Contact for further information

Amnah Alluqmani
PhD student
Information school
The University of Sheffield
Western Bank
Sheffield
S10 2TN
Email: asalluqmani1@sheffield.ac.uk

Dr. Morgan Harvey
Lecturer in Data Science
Information school
The University of Sheffield
Western Bank
Sheffield
S10 2TN
Email: m.harvey@sheffield.ac.uk

Dr. Ziqi Zhang
Senior Lecturer in Social Media
Information school
The University of Sheffield
Western Bank
Sheffield
S10 2TN
Email: ziqi.zhang@sheffield.ac.uk

The University Of Sheffield.

Information School

Participant Information Sheet – Study2

Project Title: Assisting the visually impaired and blind people in online shopping

You are being invited to take part in a research project. Before you decide whether or not to participate, it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask us if there is anything that is not clear or if you would like more information. Take time to decide whether or not you wish to take part.

---

1. What is the purpose of the study?

The purpose of this study is to explore the potential challenges visually impaired and blind people might experience when selecting and coordinating outfits online.

2. Why have I been invited?

You have been invited to this study because you are of age 18 or older, either blind or partially sighted and have prior experience with online shopping.

3. Do I have to take part?

By participating in this study, you will receive a £10 Amazon voucher. it is up to you to decide whether or not to take part. All information related to this study along with the informed consent form will be emailed to you. It would be appreciated if you could email back to show that you are agree to take part in this study. Before starting the interview, the researcher will explain all information related to this study. Moreover, your consent will be taken verbally and recorded to maintain integrity. You are free to withdraw at any time within 2 weeks, without giving a reason. If you choose to withdraw your participation then any and all data collected about you and during any studies you have participated in will be deleted.
Please note that by choosing to participate in this research, this will not create a legally binding agreement, nor is it intended to create an employment relationship between you and the University of Sheffield.

4. What will happen to me if I take part?

If you agree to participate, the researcher will arrange a meeting to conduct the interview.

During the interview, the researcher will ask general questions, questions related to selecting and coordinating clothes online and questions to allow the chance to hear your opinion and suggestions. The interview will be audio recorded. The purpose of the recording is to allow the researcher to capture all the information discussed during the interview, which is important to permit later analysis. The interview will last for about an hour.

5. What will I have to do?

The purpose of this user study is to evaluate our new system. The system is designed to generate useful descriptions of graphics on clothing items. Its aim is to benefit VI shoppers by helping them visualise the content of clothing graphics when purchasing garments online.

6. What are the possible disadvantages and risks of taking part?

There are no known disadvantages or risks associated with this research.
However, if you feel uncomfortable at any point and no longer wish to continue then you only need to mention this to the researcher and any collected data will be deleted.

7. What are the possible benefits of taking part?

While there are no immediate benefits for those people participating in the project, it is hoped that this work will have a beneficial impact on improving the shopping services in online stores for visually impaired and blind people. Moreover, the participants will have the opportunity to try and test any automatic solution that will be built as a result of this study.

8. Is the purpose of this study educational?

Yes. the data in this research will be used for a Ph.D. study.

9. Will my taking part in this project be kept confidential?

All the information that we collect about you during the course of the research will be kept strictly confidential and will only be accessible to members of the research team. The collected data will be deposited in the secure and password-protected University of Sheffield repository and other researchers may use the collected data only if they agree to preserve the confidentiality of the information as requested in this project.

10. What will happen to the data collected, and the results of the research project?

The collected data will be transcribed and analysed, then the result will be published. You will not be individually identified in any report, publication or presentation without seeking your full consent. Direct quotes from the interviews may be used in reports and publications; however, these will be anonymised to ensure that you cannot be identified.

11. Will I be recorded, and how will the recorded media be used?

The audio recordings of the interview will be used only for analysis and for illustration in conference presentations and lectures. No other use will be made of them without your written permission, and no one outside the project will be allowed access to the original recordings
Remember, you will not be identified in any report, publication or presentation without seeking your full consent.

12. Who is organising and funding the research?

The University of Sheffield is the sponsor of this study, and Umm Al Qura University is funding this research.

13. Who is the Data Controller?

The University of Sheffield will act as the Data Controller for this study. This means that the University of Sheffield is responsible for looking after your information and using it properly.

14. Who has ethically reviewed the project?

This project has been ethically approved via the University of Sheffield's Ethics Review Procedure, as administered by the Information school

15. What if something goes wrong and I wish to complain about the research or report a concern or incident?

If you are dissatisfied with any aspect of the research and wish to make a complaint, please contact *Dr Morgan Harvey; m.harvey@sheffield.ac.uk* in the first instance. If you feel your complaint has not been handled in a satisfactory way you can contact the Head of the information school *Professor Val Gillet; v.gillet@sheffield.ac.uk* If the complaint relates to how your personal data has been handled, you can find information about how to raise a complaint in the University's Privacy Notice: https://www.sheffield.ac.uk/govern/data-protection/privacy/general.

If you wish to make a report of a concern or incident relating to potential exploitation, abuse or harm resulting from your involvement in this project, please contact the project's Designated Safeguarding *Dr Morgan Harvey; m.harvey@sheffield.ac.uk*. If the concern or incident relates to the Designated Safeguarding Contact, or if you feel a report you have made to this Contact has not been handled in a satisfactory way, please contact the Head of the information school *Professor Val Gillet; v.gillet@sheffield.ac.uk* and/or the University's Research Ethics & Integrity Manager (Lindsay Unwin; l.v.unwin@sheffield.ac.uk).

16. Contact for further information

Amnah Alluqmani
PhD student
Information school
The University of Sheffield
Western Bank
Sheffield
S10 2TN
Email: asalluqmani1@sheffield.ac.uk

Dr Morgan Harvey
Lecturer in Data Science
Information school
The University of Sheffield
Western Bank
Sheffield
S10 2TN
Email: m.harvey@sheffield.ac.uk

Dr. Ziqi Zhang
Senior Lecturer in Social Media
Information school
The University of Sheffield
Western Bank
Sheffield
S10 2TN
Email: ziqi.zhang@sheffield.ac.uk

The University Of Sheffield.

Participant Consent Form

Assisting the visually impaired and blind people in online shopping

Consent Form Study 1

Please tick the appropriate boxes	**Yes**	**No**
Taking Part in the Project		
I have read and understood the project information sheet or the project has been fully explained to me. (If you will answer No to this question, please do not proceed with this consent form until you are fully aware of what your participation in the project will mean.)	☐	☐
I have been given the opportunity to ask questions about the project.	☐	☐
I agree to take part in the project.	☐	☐
I understand and agree that taking part in the project will include: participating in a task observation	☐	☐
I understand and agree that taking part in the project will include: participating in an interview	☐	☐
I agree that whilst I am participating in this task observation, a video recording will record only the screen, I agree to being video recorded and for these anonymised video recordings to be used in the research.	☐	☐
I agree that whilst I am participating in this interview, an audio recording will be made. I agree to being audio recorded and for these anonymised audio recordings to be used in the research.	☐	☐
I understand that by choosing to participate in this research, this does not create a legally binding agreement nor is it intended to create an employment relationship with the University of Sheffield.	☐	☐
I understand that I can withdraw from the study at any time within 2 weeks. I do not have to give any reasons for why I no longer want to take part and there will be no adverse consequences if I choose to withdraw.	☐	☐
How my information will be used during and after the project		
I understand my personal details such as name, phone number, address and email address etc. will not be revealed to people outside the project.	☐	☐
I understand and agree that my words may be quoted in publications, reports, web pages, and other research outputs. I understand that I will not be named in these outputs unless I specifically request this.	☐	☐
I understand and agree that other authorised researchers will have access to this data only if they agree to preserve the confidentiality of the information as requested in this form.	☐	☐
I understand and agree that other authorised researchers may use my data in publications, reports, web pages, and other research outputs, only if they agree to preserve the confidentiality of the information as requested in this form.	☐	☐
I give permission for the data that I provide to be deposited in the University of Sheffield's data repository so it can be used for future research and learning	☐	☐
So that the information you provide can be used legally by the researchers		
I agree to assign the copyright I hold in any materials generated as part of this project to The University of Sheffield.	☐	☐

Name of participant [printed]	Signature	Date

Name of Researcher [printed]	Signature	Date
Amnah Alluqmani		06/07/2022

Participant Consent Form

Project contact details for further information:

Amnah Alluqmani

PhD student
Information school
The University of Sheffield
Western Bank
Sheffield
S10 2TN
Email: asalluqmani1@sheffield.ac.uk

Dr Morgan Harvey

Lecturer in Data Science
Information school
The University of Sheffield
Western Bank
Sheffield
S10 2TN
Email: m.harvey@sheffield.ac.uk

Professor Val Gillet

Information School
The University of Sheffield
Western Bank
Sheffield
S10 2TN
Email: v.gillet@sheffield.ac.uk

The University Of Sheffield.

Participant Consent Form

Assisting the visually impaired and blind people in online shopping

Consent Form Study 1

Please tick the appropriate boxes	**Yes**	**No**
Taking Part in the Project		
I have read and understood the project information sheet or the project has been fully explained to me. (If you will answer No to this question, please do not proceed with this consent form until you are fully aware of what your participation in the project will mean.)	☐	☐
I have been given the opportunity to ask questions about the project.	☐	☐
I agree to take part in the project.	☐	☐
I understand and agree that taking part in the project will include: participating in a task observation	☐	☐
I understand and agree that taking part in the project will include: participating in an interview	☐	☐
I agree that whilst I am participating in this task observation, a video recording will record only the screen, I agree to being video recorded and for these anonymised video recordings to be used in the research.	☐	☐
I agree that whilst I am participating in this interview, an audio recording will be made. I agree to being audio recorded and for these anonymised audio recordings to be used in the research.	☐	☐
I understand that by choosing to participate in this research, this does not create a legally binding agreement nor is it intended to create an employment relationship with the University of Sheffield.	☐	☐
I understand that I can withdraw from the study at any time within 2 weeks. I do not have to give any reasons for why I no longer want to take part and there will be no adverse consequences if I choose to withdraw.	☐	☐
How my information will be used during and after the project		
I understand my personal details such as name, phone number, address and email address etc. will not be revealed to people outside the project.	☐	☐
I understand and agree that my words may be quoted in publications, reports, web pages, and other research outputs. I understand that I will not be named in these outputs unless I specifically request this.	☐	☐
I understand and agree that other authorised researchers will have access to this data only if they agree to preserve the confidentiality of the information as requested in this form.	☐	☐
I understand and agree that other authorised researchers may use my data in publications, reports, web pages, and other research outputs, only if they agree to preserve the confidentiality of the information as requested in this form.	☐	☐
I give permission for the data that I provide to be deposited in the University of Sheffield's data repository so it can be used for future research and learning	☐	☐
So that the information you provide can be used legally by the researchers		
I agree to assign the copyright I hold in any materials generated as part of this project to The University of Sheffield.	☐	☐

Name of participant [printed] Signature Date

Name of Researcher [printed] Signature Date

Amnah Alluqmani



Participant Consent Form

Project contact details for further information:

Amnah Alluqmani

PhD student
Information school
The University of Sheffield
Western Bank
Sheffield
S10 2TN
Email: asalluqmani1@sheffield.ac.uk

Dr Morgan Harvey

Lecturer in Data Science
Information school
The University of Sheffield
Western Bank
Sheffield
S10 2TN
Email: m.harvey@sheffield.ac.uk

Professor Val Gillet

Information School
The University of Sheffield
Western Bank
Sheffield
S10 2TN
Email: v.gillet@sheffield.ac.uk

Appendix B

Empirical Study Ethics

Application 043417

Amendment - Complete (Submitted on 22/02/2022)

Description of changes

I would like to add that when conducting the studies, I will collect participants' demographic information. The information will include their age, gender, frequency of internet use, degree of perception (blind or low vision), and if their visual impediment is congenital or acquired. Further details are provided in the attached document. This demographic data is required as it helps the reader learn more about the population under study, and to ensure the high quality of research.

Additional ethical considerations

Do the proposed changes pose any additional ethical considerations?
Yes

Description of how these ethical considerations are addressed
The participants may feel irritated and refuse to disclose their data. In this case, they have the option to select "I prefer not to mention" to prevent any inconvenience, as well as the ability to end the interview at any time. The collected data will thereafter be deleted. If any participant does not want their demographic data gathered, the researcher will simply start the study without collecting this information. The participant's response to the demographic questions will not be part of the audio recording; instead, the researcher will take notes.

Additional risks

Do any of the proposed amendments to the research potentially change the risk for any of the researchers?
Yes

Description of the risks and how they are addressed
I do not think there is any potential risk to the researcher as only general information will be gathered. However, if the participants feel irritated or frustrated, they can end the interview at any time. The demographic data will be collected to ensure the high quality of research by providing some information about the population under study. The participants will be informed about the purpose of collecting this data.

Supporting documentation revisions

Do the proposed amendments require revisions to any of the supporting documentation? Please note that when uploading new versions of documents which you have previously provided, you should give a description of the document which clearly indicates that this is a new version, e.g. by providing an appropriate version number. It is also helpful to the reviewers if you clearly mark the changes you have made in the document itself (e.g. by highlighting new text or using tracked changes in Word).
Yes

Uploaded documentation
- DemographicInfo.pdf

Other relevant information

Decision

Approved

Original application

Section A: Applicant details

Date application started:
Mon 13 September 2021 at 13:26

First name:

Amnah

Last name:
Alluqmani

Email:
asalluqmani1@sheffield.ac.uk

Programme name:
not applicable

Module name:
not applicable
Last updated:
24/02/2022

Department:
Information School [a.k.a iSchool]

Applying as:
Postgraduate research

Research project title:
Assisting visually impaired and blind people in online shopping

Has your research project undergone academic review, in accordance with the appropriate process?
Yes

Similar applications:
- not entered -

Section B: Basic information

Supervisor

Name	Email
Morgan Harvey	m.harvey@sheffield.ac.uk

Proposed project duration

Start date (of data collection):
Fri 1 October 2021

Anticipated end date (of project)
Wed 30 March 2022

3: Project code (where applicable)

Project externally funded?
No

Project code
- not entered -

Suitability

Takes place outside UK?
No

Involves NHS?
No

Health and/or social care human-interventional study?
No

ESRC funded?
No

Likely to lead to publication in a peer-reviewed journal?
Yes

Led by another UK institution?
No

Involves human tissue?
No

Clinical trial or a medical device study?
No

Involves social care services provided by a local authority?
No

Is social care research requiring review via the University Research Ethics Procedure
No

Involves adults who lack the capacity to consent?
No

Involves research on groups that are on the Home Office list of 'Proscribed terrorist groups or organisations?
No

Indicators of risk

Involves potentially vulnerable participants?
Yes
Involves potentially highly sensitive topics?
No

Section C: Summary of research

1. Aims & Objectives

Visually impaired (VI) and blind people experience multiple difficulties in performing daily activities. One of the main challenges is being able to perform online shopping independently. Generally, they use automated assistance tools, such as alternative text and voiceover from smartphones, to help them access a product description and reviews. However, they still cannot confirm their selection of clothes without the help of a sighted person who ensures that the selected products are appropriate for them or who ensures the coordination of the selected outfit. To avoid this dependency, VI and blind people often purchase plain clothes or only buy from stores that offer coordinated outfits. Most studies on assisting VI and blind people with online shopping have exclusively focused on the accessibility and information retrieval problem. The main aim of this research is to focus on understanding the challenges of VI and blind people in selecting and coordinating outfits during online shopping and build a suitable automated assistance tool. This project will be carried out by conducting three stages as the following:
Stage 1:
The aim of this stage is to investigate the potential challenges of VI and blind people while shopping for clothes online. This stage will be obtained by implementing the following studies
- Study 1, planned task observation will be conducted to gain in-depth knowledge of these obstacles. A follow-up interview will then be implemented to discuss any ambiguity that arises during the observation process.
- Study 2, a standalone interview will be conducted to ask questions related to the experiences of VI and blind people in selecting and coordinating outfits online. As such, this study will also provide the opportunity to listen to their opinions and suggestions about improving the online shopping process.

Stage 1 objectives :
- To observe VI and blind people's behavior when shopping online.
- To explore how they select and coordinate clothes online
- To identify the difficulties they face when shopping online.
- To inform how an automated support system should be built to provide suitable assistance in online shopping

Results obtained from this stage will be applied to build a suitable automated assistance tool (in stage 2).

Stage 2:
The aim of this stage is to build an automated assistance tool to limit the difficulties of shopping clothes online for VI and blind people.

Stage 3:

The aim of this stage is to evaluate the proposed system

2. Methodology

The data will be collected by performing two studies.

Study1: In this study, the participants will be asked to describe what kind of clothes they would like to buy for a party. They will be asked to select one top and one bottom item, along with a third compatible item. After describing their desired choices, they will be asked to share their screen and start shopping using the website of their preferred online retailer by performing the following tasks:
• Select a top item you would like to wear
• Select a bottom item you would like to wear
• Select any compatible third item you would like to wear

To gather sufficient information, the participants will be asked to think aloud when performing the above-mentioned tasks. The average time given for this task observation will be 30 minutes (this is an inductive duration; a more accurate time estimation will be given after conducting the pilot study).
After performing the task observation, the participants will be interviewed to share their experiences and clarify any ambiguity arising from the observation process. The questions are divided into two parts. The first part of the questions is designed to analyse how satisfied the participants were with their shopping experience during the task observation. In addition, these questions open a discussion about the challenges they faced when selecting and coordinating their outfits. This will allow them to share their suggestions. In the second part, the participants will be asked specific questions related to their task observation. This interview is expected to last for 30 minutes

Study 2: In this study, the interview questions are divided into three parts. In the first part, information on the participants' backgrounds and shopping experiences will be collected. In the second part, the questions focus on finding out the challenges that they face selecting and coordinating clothes when shopping online. In the last part, the participants are asked to share their opinions on specific aspects of the process of selecting and coordinating clothes when shopping online. This interview will last for approximately one hour.

An audio recording will be made while performing both studies. However, in Study 1, once participants start to share their screen and perform the tasks, a video recording will capture their behaviour while they select clothes online. This recording will capture only their screens and no likenesses of the participants or their homes will be recorded.

To ensure a comfortable environment, the participants will be informed when an audio or video recording begins. These recordings will be used for later analysis. If any participant would like sighted assistance, they either ask a friend or family member to assist them with the study or perform the study at the SRSB., which will offer the necessary assistance.

Before performing the actual study, the proposed studies will be piloted to ensure better time estimation and improve the proposed protocol. As the transcript is generated, it will be analysed, and the theme and pattern will be identified using NVivo and Microsoft Excel.

3. Personal Safety

Have you completed your departmental risk assessment procedures, if appropriate?

In progress

Raises personal safety issues?

No

I believe this research does not involve any issues for the researcher and participants, as shopping is not a sensitive topic, and both studies will be conducted remotely. If any participant is frustrated or irritated by the tasks (this is not expected, as the planned tasks are not complicated), they can terminate the interview at any time.
If any participant would like sighted assistance, they either ask a friend or family member to assist them with the study or perform the study at the SRSB
The project aims, objectives, and expected outcomes will be explained to all the participants before they take place in the study.

Section D: About the participants

1. Potential Participants

As the study aims to explore the challenges of visually impaired and blind people, the recruitment criteria will be as the following:
• Age of minimum 18 years
• Varied genders (to maintain research validity)
• Absence of other disabilities aside from visual impairment or blindness (to exclude other challenges that may affect purchasing behaviour, such as hearing or physical impairment)
• Experience in shopping online (at least once)

2. Recruiting Potential Participants

The participants will be recruited in collaboration with The Sheffield Royal Society for the Blind based on the aforementioned criteria. Before conducting the study, they will receive an email that includes the research information sheet, the informed consent form, and any information related to the research.

2.1. Advertising methods

Will the study be advertised using the volunteer lists for staff or students maintained by IT Services? No

- not entered -

3. Consent

Will informed consent be obtained from the participants? (i.e. the proposed process) Yes

Nominated participants will receive an email with the research information, including an information sheet and an informed consent form. Before conducting the interview and the observation, the information related to the research will be explained, and the participants will be informed that their consent will be taken verbally and recorded to maintain integrity.
Participants will be given the opportunity to ask any questions they might have before giving consent.

4. Payment

Will financial/in kind payments be offered to participants? No

5. Potential Harm to Participants

What is the potential for physical and/or psychological harm/distress to the participants?

There is a chance that the participants might feel frustrated or irritated while performing the tasks.

How will this be managed to ensure appropriate protection and well-being of the participants?

As stated in the information sheet, participants may refuse to perform any tasks or answer any questions which make them feel uncomfortable and they can stop the interview at any time without giving any reason.
So, if they feel frustrated or uncomfortable at any point and no longer wish to continue then they only need to mention this to the researcher and the study will be terminated and any collected data will be deleted.

6. Potential harm to others who may be affected by the research activities

Which other people, if any, may be affected by the research activities, beyond the participants and the research team?

Any person being available at the participant location while performing the study.

What is the potential for harm to these people?

As this study will be audio recorded, there is a chance that other people may speak in the background and be recorded.

How will this be managed to ensure appropriate safeguarding of these people?

If a voice of someone appears in the background, the recording will be stopped to make sure this background is removed. After completing the interview, the section including the background distraction will be deleted.

7. Reporting of safeguarding concerns or incidents

What arrangements will be in place for participants, and any other people external to the University who are involved in, or affected by, the research, to enable reporting of incidents or concerns?

If participants are dissatisfied with any aspect of the research and wish to make a complaint, They might contact Dr. Morgan Harvey; m.harvey@sheffield.ac.uk in the first instance. If they feel the complaint has not been handled in a satisfactory way They can contact the Head of the information school Professor Val Gillet; v.gillet@sheffield.ac.uk If
the complaint relates to how the participant's personal data has been handled, They can find information about how to raise a complaint in the University's Privacy Notice: https://www.sheffield.ac.uk/govern/data-protection/privacy/general

Who will be the Designated Safeguarding Contact(s)?

The primary supervisor Dr. Morgan Harvey; m.harvey@sheffield.ac.uk.

How will reported incidents or concerns be handled and escalated?

If participants wish to make reports of concerns or incidents relating to potential exploitation, abuse, or harm resulting from their involvement in this project, They might contact the project's Designated Safeguarding Dr. Morgan Harvey; m.harvey@sheffield.ac.uk. If the concern or incident relates to the Designated Safeguarding Contact, or if the participant feel the reports they have made to this Contact has not been handled in a satisfactory way, the participants might contact the Head of the information school Professor Val Gillet; v.gillet@sheffield.ac.uk and/or the University's Research Ethics &

Integrity Manager (Lindsay Unwin; l.v.unwin@sheffield.ac.uk).

Section E: About the data

1. Data Processing

Will you be processing (i.e. collecting, recording, storing, or otherwise using) personal data as part of this project? (Personal data is any information relating to an identified or identifiable living person).
No

Please outline how your data will be managed and stored securely, in line with good practice and relevant funder requirements

The video and audio files obtained from performing both studies will be stored in the University of Sheffield's Google Drive, as will all paper documents and written research diaries. Each participant will be assigned to a randomly generated ID number that will be used to refer to all of their recordings and transcripts. The ananomization will take place as soon as the recordings start, and the data will be stored in the University's Google Drive and shared only with the supervisors.

Section F: Supporting documentation

Information & Consent

Participant information sheets relevant to project?
Yes

Document 1097515 (Version 1)	All versions
Document 1097514 (Version 1)	All versions

Consent forms relevant to project?
Yes

Document 1097517 (Version 1)	All versions
Document 1097516 (Version 1)	All versions

Additional Documentation

Document 1097518 (Version 1) Data Management Plan	All versions

External Documentation

- not entered -

Section G: Declaration

Signed by:
Amnah Alluqmani
Date signed:
Sat 2 October 2021 at 16:29

Offical notes

- not entered -

Downloaded: 08/02/2024
Approved: 18/10/2021

Amnah Alluqmani
Registration number: 200296368
Information School [a.k.a iSchool]
Programme: not applicable

Dear Amnah

PROJECT TITLE: Assisting visually impaired and blind people in online shopping
APPLICATION: Reference Number 043417

On behalf of the University ethics reviewers who reviewed your project, I am pleased to inform you that on 18/10/2021 the above-named project was **approved** on ethics grounds, on the basis that you will adhere to the following documentation that you submitted for ethics review:

- University research ethics application form 043417 (form submission date: 02/10/2021); (expected project end date: 30/03/2022).
- Participant information sheet 1097515 version 1 (21/09/2021).
- Participant information sheet 1097514 version 1 (21/09/2021).
- Participant consent form 1097517 version 1 (21/09/2021).
- Participant consent form 1097516 version 1 (21/09/2021).

The following amendments to this application have been approved:

- Amendment approved: 22/02/2022

If during the course of the project you need to deviate significantly from the above-approved documentation please inform me since written approval will be required.

Your responsibilities in delivering this research project are set out at the end of this letter.

Yours sincerely

Claire Du Puget
Ethics Administrator
Information School [a.k.a iSchool]

Please note the following responsibilities of the researcher in delivering the research project:

- The project must abide by the University's Research Ethics Policy: https://www.sheffield.ac.uk/research-services/ethics-integrity/policy
- The project must abide by the University's Good Research & Innovation Practices Policy: https://www.sheffield.ac.uk/polopoly_fs/1.671066!/file/GRIPPolicy.pdf
- The researcher must inform their supervisor (in the case of a student) or Ethics Administrator (in the case of a member of staff) of any significant changes to the project or the approved documentation.
- The researcher must comply with the requirements of the law and relevant guidelines relating to security and confidentiality of personal data.
- The researcher is responsible for effectively managing the data collected both during and after the end of the project in line with best practice, and any relevant legislative, regulatory or contractual requirements.

Appendix C

Interview Scripts

Table C.1: *Observation and Follow-up Interview Protocol for Study 1. "IQ" denotes interview questions, "RQ" denotes research questions.*

Introduction script:
Thank you for agreeing to take part in this study.
I'll explain some information related to the study, and you're welcome to interrupt me at any time and ask questions.
The purpose of this study is to explore the potential challenges visually impaired and blind people might experience when selecting and coordinating outfits online.

This study comprises task observation which will last up to 30 minutes, followed by a 20-minute follow-up interview.
During the task observation, you will be asked to do a number of tasks.
This includes sharing your screen,
accessing the Amazon website (please do not log in to your own account), and then performing shopping tasks such as outfit selection and coordination. To better understand your shopping behaviour, you will be asked to think aloud
while the researcher is quietly observing and taking notes. The observation concludes either by confirming that you have completed the tasks or by indicating that you wish to terminate the observation.

In the follow-up interview, you will be asked questions related to the selection tasks
you have just completed. However, you may refuse to perform any tasks or answer any questions that make you feel uncomfortable, and you can stop the interview at any time without any negative consequences.

A video recording will capture only the screen to record your behaviour when performing the planned tasks, while an audio recording will be used in the follow-up interview.
To ensure a comfortable environment, you will be informed when an audio or video recording begins.
Do you have any questions before we start the observation?
If you did not fill out the consent form formally, your consent will be taken verbally and audio recorded.

Now I'll start the observation.

Imagine that you are going to a party and need to buy a new outfit to wear. Please share your screen and start shopping on Amazon by performing the following tasks:
Task 1: Select a top item you would like to wear
Task 2: Select a bottom item you would like to wear
Task 3: Select any compatible third item you would like to wear

Interview script:

Thank you for taking part in the observation study.
In the follow-up interview, I'll ask you about your online shopping experience and the task observation you have completed.
Do you have any questions before we start the interview?
You are free to ask questions you may have st any time during this interview.

	Opening Information	RQ1: What challenges do VI people face when independently selecting or coordinating outfits during online shopping?	RQ2 To what extent,do VI people consider a specific celebrity style or fashion style when selecting outfits?	Questions related to the observation process
Part1: In this part I will ask general questions				
IQ1.Tell me a little bit about yourself	X			
IQ2.How pleased were you with the choices you have made?	X			
IQ3. Describe any challenges you faced when selecting the products		X		
IQ4. Describe any challenges you faced when you were coordinating the items		X		
IQ5.Would you please talk a bit about your fashion inspiration? Do you follow a specific celebrity's style; Why?			X	
IQ7 Would you prefer to have an automatic personal shopper that can select and coordinate clothes for you based on your favourite celebrity style / fashion style?			X	
Part 2: Thank you for your responses, now I will ask questions related to the observation process				

Table C.2: *Standalone Interview protocol matrix for study 2. "I" denotes interview questions, "RQ" denotes research questions.*

Interview's introduction script:

Thank you for agreeing to take part in this interview.
The purpose of this study is to better understand the difficulties that visually impaired and blind people face when selecting and coordinating clothing online,as well as to solicit their feedback and suggestions for improving the online shopping experience.
In this interview, I will ask you about online shopping, your experience selecting and coordinating clothing while shopping online, and how to improve your online shopping experience. The interview will last up to half an hour.
It will be audio recorded in order to capture all of the information discussed during the interview, which is necessary for later analysis.
To ensure a comfortable environment, you will be informed when the audio recording begins.

Do you have any questions before we start the interview?
You are free to ask any questions you may have at any time during this interview.
I would be delighted to respond. If you did not fill out the consent form formally, your consent will be taken verbally and audio recorded.

	Background Information	RQ1.What challenges do VI people face when independently selecting and coordinating outfits while online shopping?	RQ3.3: What type of assistance do VI people regularly make use of when shopping for clothes online?	RQ2. To what extent do VI people consider a specific celebrity or fashion style when selecting outfits?	RQ3.1: What services are offered to assist VI people in coordinating outfits in offline stores are not yet offered in online stores?	RQ 3.2: What types of assistance would VI people like to have when co-ordinating outfits while shopping online?
Part1: Part1: I will begin this interview by asking some General questions						
IQ1.Tell me a little bit about yourself	X					
IQ2: Describe how you manage to shop for clothes online during Covid-19 lockdowns	X					
Part2: Thank you for your responses, now I will ask questions regarding the outfit selection and coordination process						
IQ3: when selecting or coordinating clothes online with a collaborative assistance, what types of questions do you usually ask?		X				
IQ4: Do you use any automated assistance tool when shopping for clothes online? Describe its strengths and weaknesses			X			
IQ5: Describe your experience when you are shopping for clothes online?		X				
IQ6:Tell me about your fashion interests? How do you find inspiration around fashion? Do you follow a specific celebrity's style; explain?				X		
IQ7: Please describe a time when you bought clothes online independently, if appropriate. How did you coordinate the outfits while shopping online?		X		X		

IQ8: Please describe how do you coordinate outfits when shopping in online stores versus when physical stores Which scenario do you prefer and why?					X	
Part 3:Thank you for your responses. The questions in part 3 were designed to allow a chance to listen to your opinion						
IQ9: Would you please talk about any potential challenges you experience when selecting or coordinating outfits online?		X				
IQ10: What is your suggestion for enhancing the process of shopping for clothes online?						X
IQ11: Do you think you would like being assisted by an automatic personal shopper? If yes/no Would you trust an automatic personal shopper to select or coordinate an outfit for you? Why? If yes/no Please describe what services you would trust your personal shopper to perform for you?						X
IQ12: Have you tried any of the current automated personal shoppers (e.g. Amazon personal shopper and Stitch Fix), why?						X
Before we conclude the interview, do you have and comments or experience you would like to add						

Appendix D

Human Evaluation Documents

The University Of Sheffield.

Participant Consent Form
Fashion Image Captioning Evaluation Study

Consent Form

Please tick the appropriate boxes	**Yes**	**No**
Taking Part in the Project		
I have read and understood the project information sheet or the project has been fully explained to me. (If you will answer No to this question, please do not proceed with this consent form until you are fully aware of what your participation in the project will mean.)	☐	☐
I have been given the opportunity to ask questions about the project.	☐	☐
I agree to take part in the project.	☐	☐
I understand and agree that taking part in the project will include: evaluating image captioning systems	☐	☐
I understand that by choosing to participate in this research, this does not create a legally binding agreement nor is it intended to create an employment relationship with the University of Sheffield.	☐	☐
I understand that I can withdraw from the study at any time within 2 weeks of completing the online survey. I do not have to give any reasons for why I no longer want to take part and there will be no adverse consequences if I choose to withdraw.	☐	☐
How my information will be used during and after the project		
I understand my personal details such as name, phone number, address and email address etc. will not be revealed to people outside the project.	☐	☐
I understand and agree that other authorised researchers may use my data in publications, reports, web pages, and other research outputs, only if they agree to preserve the confidentiality of the information as requested in this form.	☐	☐
I give permission for the data that I provide to be deposited in the University of Sheffield's data repository so it can be used for future research and learning	☐	☐
So that the information you provide can be used legally by the researchers		
I agree to assign the copyright I hold in any materials generated as part of this project to The University of Sheffield.	☐	☐

Name of participant [printed] Signature Date

Name of Researcher [printed] Signature Date:

Project contact details for further information:

Amnah Alluqmani

PhD student
Information school
The University of Sheffield
The Wave, 2 Whitham Rd, Broomhall
Sheffield
S10 2SJ
Email: asalluqmani1@sheffield.ac.uk

Participant Consent Form

Dr. Morgan Harvey

Lecturer in Data Science
Information school
The University of Sheffield
The Wave, 2 Whitham Rd, Broomhall
Sheffield
S10 2SJ
Email: m.harvey@sheffield.ac.uk

Professor Val Gillet

Information School
The University of Sheffield
Western Bank
Sheffield
S10 2TN
Email: v.gillet@sheffield.ac.uk

The University Of Sheffield.

Information School

Participant Information Sheet

Project Title: Fashion Image Captioning Evaluation Study

You are being invited to take part in a research project. Before you decide whether or not to participate, it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask us if there is anything that is not clear or if you would like more information. Take time to decide whether or not you wish to take part.

---

1. **What is the purpose of the study?**
   The purpose of this user study is to help us evaluate various systems, which are designed to generate useful descriptions of graphic elements on clothing items. Their aim is to benefit visually impaired shoppers by helping them visualise the content of clothing graphics when purchasing garments online.

2. **Why have I been invited?**
   You have been invited to this study because you are of age 18 or older.

3. **Do I have to take part?**
   It is up to you to decide whether or not to take part. You are free to withdraw at any time within 2 weeks of completing the online survey, without giving a reason. If you choose to withdraw your participation then any and all data collected about you and during any studies you have participated in will be deleted. Please note that by choosing to participate in this research, this will not create a legally binding agreement, nor is it intended to create an employment relationship between you and the University of Sheffield.

4. **What will happen to me if I take part?**
   If you are an individual with a visual impairment, the researcher will send you an email with instructions and a link to access the study. Once you access the link, further information will be provided, and you may provide your consent to continue.

   If you are a crowdsourced participant, once you access the study link, you will find information about the study and have the option to provide your consent before beginning the study.

5. **What will I have to do?**

   Your task will be to judge whether the caption is informative in helping you visualise the graphical elements of the clothing item without seeing the image. Responses are given using a four-point Likert-type scale (`very informative', `informative', `less informative' or `not informative').

   If you are a sighted individual, images will be presented and you will be asked to evaluate the correctness of the descriptions compared to the information depicted in

the images. To rate the accuracy of the captions, the participants will be asked to choose from a Likert scale (i.e. 'very good', 'good', 'bad', and 'very bad'). You will then be required to provide an explanation for your ratings as well as an alternative caption to address any issues with the one presented. Once you have finished evaluating the captions, click the "submit" button to save your participation.

6. What are the possible disadvantages and risks of taking part?

There are no known disadvantages or risks associated with this research. However, if you feel uncomfortable at any point and no longer wish to continue then you only need to mention this to the researcher and any collected data will be deleted.

7. What are the possible benefits of taking part?

If you are a visually impaired participant, you will receive a £10 Amazon voucher via email.
If you are a crowdsourced participant, you will receive a £3.63 payment via the Prolific platform.

8. Is the purpose of this study educational?

Yes. the data in this research will be used for a Ph.D. work.

9. Will my taking part in this project be kept confidential?

All the information that we collect about you during the course of the research will be kept strictly confidential and will only be accessible to members of the research team.

10. What will happen to the data collected, and the results of the research project?

The collected data will be analysed, then the result will be published. You will not be individually identified in any report, publication, or presentation without seeking your full consent.

11. Will I be recorded, and how will the recorded media be used?

No, you will not be recorded

12. Who is organising and funding the research?

The University of Sheffield is the sponsor of this study, and Umm Al Qura University is funding this research.

13. Who is the Data Controller?

The University of Sheffield will act as the Data Controller for this study. This means that the University of Sheffield is responsible for looking after your information and using it properly.

14. Who has ethically reviewed the project?

This project has been ethically approved via the University of Sheffield's Ethics Review Procedure, as administered by the Information school.

15. What if something goes wrong and I wish to complain about the research or report a concern or incident?

If you are dissatisfied with any aspect of the research and wish to make a complaint, please contact *Dr Morgan Harvey; m.harvey@sheffield.ac.uk* in the first instance. If you feel your complaint has not been handled in a satisfactory way you can contact the Head of the information school *Professor Val Gillet; v.gillet@sheffield.ac.uk* If the complaint relates to how your personal data has been handled, you can find information about how to raise a complaint in the University's Privacy Notice: https://www.sheffield.ac.uk/govern/data-protection/privacy/general.

If you wish to make a report of a concern or incident relating to potential exploitation, abuse or harm resulting from your involvement in this project, please contact the project's Designated Safeguarding Contact *Dr Morgan Harvey; m.harvey@sheffield.ac.uk*. If the concern or incident relates to the Designated Safeguarding Contact, or if you feel a report you have made to this Contact has not been handled in a satisfactory way, please contact the Head of the information school *Professor Val Gillet; v.gillet@sheffield.ac.uk* and/or the University's Research Ethics & Integrity Manager (Lindsay Unwin; l.v.unwin@sheffield.ac.uk).

16. Contact for further information

Amnah Alluqmani
PhD student
Information school
The University of Sheffield
The Wave, 2 Whitham Rd, Broomhall
Sheffield
S10 2SJ
Email: asalluqmani1@sheffield.ac.uk

Dr. Morgan Harvey
Senior Lecturer in Data Science
Information school
The University of Sheffield
The Wave, 2 Whitham Rd, Broomhall
Sheffield
S10 2SJ
Email: m.harvey@sheffield.ac.uk

Appendix E

Human Evaluation Ethics

Downloaded: 03/11/2023
Approved: 07/08/2023

Amnah Alluqmani
Registration number: 200296368
Information School [a.k.a iSchool]
Programme: not applicable

Dear Amnah

PROJECT TITLE: Fashion Image Captioning Evaluation Study
APPLICATION: Reference Number 054252

On behalf of the University ethics reviewers who reviewed your project, I am pleased to inform you that on 07/08/2023 the above-named project was **approved** on ethics grounds, on the basis that you will adhere to the following documentation that you submitted for ethics review:

- University research ethics application form 054252 (form submission date: 14/07/2023); (expected project end date: 06/08/2023).
- Participant information sheet 1125177 version 2 (27/06/2023).
- Participant consent form 1125178 version 2 (14/07/2023).
- Participant consent form 1126618 version 1 (14/07/2023).

The following amendments to this application have been approved:

- Amendment approved: 01/11/2023
- Amendment approved: 11/09/2023

If during the course of the project you need to deviate significantly from the above-approved documentation please inform me since written approval will be required.

Your responsibilities in delivering this research project are set out at the end of this letter.

Yours sincerely

Peter Bath
Ethics Administrator
Information School [a.k.a iSchool]

Please note the following responsibilities of the researcher in delivering the research project:

- The project must abide by the University's Research Ethics Policy: https://www.sheffield.ac.uk/research-services/ethics-integrity/policy
- The project must abide by the University's Good Research & Innovation Practices Policy: https://www.sheffield.ac.uk/polopoly_fs/1.671066!/file/GRIPPolicy.pdf
- The researcher must inform their supervisor (in the case of a student) or Ethics Administrator (in the case of a member of staff) of any significant changes to the project or the approved documentation.
- The researcher must comply with the requirements of the law and relevant guidelines relating to security and confidentiality of personal data.
- The researcher is responsible for effectively managing the data collected both during and after the end of the project in line with best practice, and any relevant legislative, regulatory or contractual requirements.

The University Of Sheffield.

Application 054252

Amendment - Complete (Submitted on 01/11/2023)

Description of changes

In this amendment, we are increasing the compensation amount of our expert users from £3.63 to £10, as this change comes after piloting our study.

Additional ethical considerations

Do the proposed changes pose any additional ethical considerations?
No

Additional risks

Do any of the proposed amendments to the research potentially change the risk for any of the researchers?
No

Supporting documentation revisions

Do the proposed amendments require revisions to any of the supporting documentation? Please note that when uploading new versions of documents which you have previously provided, you should give a description of the document which clearly indicates that this is a new version, e.g. by providing an appropriate version number. It is also helpful to the reviewers if you clearly mark the changes you have made in the document itself (e.g. by highlighting new text or using tracked changes in Word).
No

Other relevant information

Decision

Approved

Amendment - Complete (Submitted on 11/09/2023)

Description of changes

Instead of conducting a within-subject design, we have revised our method to conduct a between-subject approach to avoid collecting biased data. In general, we employ a scale rating approach so that different captions from the same test pool can be compared. The evaluation itself consists of either three or four questions. The first question presents only a caption from either our pipeline or from other models, and the evaluator is asked to judge whether the caption is informative in helping them visualise the graphical elements of the clothing item without seeing the image. Responses are given using a four-point Likert-type scale (`very informative', `informative', `less informative' or `not informative'). In the second question, the same caption is presented together with the associated image; the caption's accuracy is then rated using a four-point Likert-type scale (`very good', `good', `bad', and `very bad'). The third question asks participants to explain their reasons for the rating given. The fourth question is asked only if the caption was generated by our pipeline and the participant rated the accuracy as 'bad' or 'very bad'. The participants are asked to fix the caption, and the intent is to explore the limitations of the pipeline. In this amendment, we only revised the first question to be a Likert scale question and added a new question (the fourth question) to collect alternative captions, aiming to fix any issues with the captions generated from our pipeline.

Additional ethical considerations

Do the proposed changes pose any additional ethical considerations?
No

Additional risks

Do any of the proposed amendments to the research potentially change the risk for any of the researchers?
No

Supporting documentation revisions

Do the proposed amendments require revisions to any of the supporting documentation? Please note that when uploading new versions of documents which you have previously provided, you should give a description of the document which clearly indicates that this is a new version, e.g. by providing an appropriate version number. It is also helpful to the reviewers if you clearly mark the changes you have made in the document itself (e.g. by highlighting new text or using tracked changes in Word).
Yes

Uploaded documentation

- InformationSheet-Stage3.doc

Other relevant information

Decision

Approved

Original application

Section A: Applicant details

Date application started:
Thu 1 June 2023 at 12:23

First name:
Amnah

Last name:
Alluqmani

Email:
asalluqmani1@sheffield.ac.uk

Programme name:
not applicable

Module name:
not applicable
Last updated:
03/11/2023

Department:
Information School [a.k.a iSchool]

Applying as:
Postgraduate research

Research project title:
Fashion Image Captioning Evaluation Study

Has your research project undergone academic review, in accordance with the appropriate process?
Yes

Similar applications:
- not entered -

Section B: Basic information

Supervisor

Name	Email

Name	Email
Morgan Harvey	m.harvey@sheffield.ac.uk

Proposed project duration

Start date (of data collection):
Tue 6 June 2023

Anticipated end date (of project)
Sun 6 August 2023

3: Project code (where applicable)

Project externally funded?
No

Project code
- not entered -

Suitability

Takes place outside UK?
No

Involves NHS?
No

Health and/or social care human-interventional study?
No

ESRC funded?
No

Likely to lead to publication in a peer-reviewed journal?
Yes

Led by another UK institution?
No

Involves human tissue?
No

Clinical trial or a medical device study?
No

Involves social care services provided by a local authority?
No

Is social care research requiring review via the University Research Ethics Procedure
No

Involves adults who lack the capacity to consent?
No

Involves research on groups that are on the Home Office list of 'Proscribed terrorist groups or organisations?
No

Indicators of risk

Involves potentially vulnerable participants?
Yes
Involves potentially highly sensitive topics?
No

Section C: Summary of research

1. Aims & Objectives

Visually impaired (VI) people experience multiple difficulties in performing daily activities. One of the main challenges is being able to perform online shopping independently. Generally, they use automated assistance tools, such as alternative text and voiceover from smartphones, to help them access a product description and reviews. However, the alternative text and product descriptions provided by online stores is often incomplete, inaccurate, or entirely absent, making it difficult or even impossible for VI people to make infomred choices.
We have developed a new system that automatically generates detailed descriptions of clothing graphics to help VI people mentally imagine the details of the graphic information when purchasing clothing items online. We now need to evaluate our system to determine whether the generated descriptions provide a better understanding of the clothing items. Additionally, we will assess whether the proposed system meets the needs of the intended users (i.e. people with visual impairment).

The objectives of this work are:
- to evaluate the accuracy of the generated descriptions
- to compare the generated captions against those of existing image-captioning models
- to assess whether the generated descriptions are beneficial for VI shoppers

The results of this stage will be used to assess our system, address its weaknesses and inform the development of an improved version.

2. Methodology

Our evaluation study will consist of two parts:
The first part will evaluate our system against an existing image-captioning system and examine whether our system helps users better understand the associated images without looking at them. Only one pair of captions will be presented, each describing the same image of a clothing item and focusing on the graphic elements of the clothing. Participants will be asked to select the caption that best helps them mentally imagine the graphics. They will choose the best caption by checking the box next to their preferred selection.
In the second part, images will be presented and participants will be asked to evaluate the correctness of the descriptions compared to the information depicted in the images. To rate the accuracy of the captions, the participants will be asked to choose from a Likert scale (i.e. ‘very good’, ‘good’, ‘bad’, and ‘very bad’). Participants will then be asked to explain the reasons behind their ratings.

Each participant will evaluate approximately 50 items. We aim to recruit 40 VI people via our collaboration with the Sheffield Royal Society for the Blind. We also aim to recruit around 300 sighted people using the Prolific crowdsourcing platform. The second part of the study will only include sighted participants. Our selections will be informed by participants’ self-identification as sighted or VI at the beginning of the study.
It is estimated that the first and second parts of the evaluation will each take about 20 minutes to complete (this duration was calculated inductively; a more accurate time estimation will be provided after we conduct the pilot study).

Initially, we will conduct a pilot study to assess the study design and obtain a more accurate time estimation. Subsequently, the design of the actual study will be reviewed, and the duration will be adjusted as necessary. The study will be conducted using the Qualtrics tool.

3. Personal Safety

Have you completed your departmental risk assessment procedures, if appropriate?

Not applicable

Raises personal safety issues?

No

I do not believe this research causes any problems for the researcher or participants. The participants will participate in the study by clicking on links and offering their assessments without the necessity of disclosing any personal information, and and all submissions will be made anonymously.
The two parts of the study will be conducted remotely at the participants' convenience. Sighted participants will receive a link distributed through the Qualtrics tool, participants with visual impairments will receive the link via email. If sighted assistance is needed, potential participants will be asked to conduct the study at a local charity.

Section D: About the participants

1. Potential Participants

We are looking to recruit individuals who are 18 or older. A total of 40 visually impaired people and 300 sighted people.

2. Recruiting Potential Participants

VI participants will be recruited in collaboration with The Sheffield Royal Society for the Blind. Sighted participants will be recruited via the Prolific crowdsourcing platform.

Before conducting the study, VI people will receive an email that includes the research information sheets, the informed consent forms, and any other information related to the research. After giving their consent and completing the study, they will receive the voucher payment via email.

The participants who will be involved through crowdsourcing will receive information about the study and give their consent through the survey interface before beginning the study. Then, they will automatically receive the payments once they complete the study.

2.1. Advertising methods

Will the study be advertised using the volunteer lists for staff or students maintained by IT Services? No

- not entered -

3. Consent

Will informed consent be obtained from the participants? (i.e. the proposed process) Yes

Nominated VI participants will receive an email with the research information, including an information sheet and an informed consent form. The participants will be asked to complete the consent sheets and email them back. Once we receive their consent, we email them the study links and instructions on how to complete the study.

Crowdsourced participants will receive information about the study and give their consent through the survey interface before beginning the study. Once they give their consent, they can proceed to join the study.

4. Payment

Will financial/in kind payments be offered to participants? Yes

VI participants will be compensated with a £3.63 Amazon voucher via email.
Crowdsourced participants will be compensated with a £3.63 payment via the Prolific platform.

5. Potential Harm to Participants

What is the potential for physical and/or psychological harm/distress to the participants?

There is a chance that the participants might feel frustrated or irritated while participating in the study.

How will this be managed to ensure appropriate protection and well-being of the participants?

As stated in the information sheet, participants may refuse to complete the study.
So, if they feel frustrated or uncomfortable at any point and no longer wish to continue, then they only need to mention this to the researcher and the study will be terminated and any collected data will be deleted.

6. Potential harm to others who may be affected by the research activities

Which other people, if any, may be affected by the research activities, beyond the participants and the research team?

No other people are expected to be affected by the research activity.

What is the potential for harm to these people?

None

How will this be managed to ensure appropriate safeguarding of these people?

The study does not require any carers. However, if sighted assistance is needed, potential participants will be encouraged to conduct the study at a local charity.

7. Reporting of safeguarding concerns or incidents

What arrangements will be in place for participants, and any other people external to the University who are involved in, or affected by, the research, to enable reporting of incidents or concerns?

If participants are dissatisfied with any aspect of the research and wish to make a complaint, They might contact Dr. Morgan Harvey; m.harvey@sheffield.ac.uk in the first instance. If they feel the complaint has not been handled in a satisfactory way They can contact the Head of the information school Professor Val Gillet; v.gillet@sheffield.ac.uk If the complaint relates to how the participant's personal data has been handled, They can find information about how to raise a complaint in the University's Privacy Notice: https://www.sheffield.ac.uk/govern/data-protection/privacy/general

Who will be the Designated Safeguarding Contact(s)?

The primary supervisor Dr. Morgan Harvey; m.harvey@sheffield.ac.uk.

How will reported incidents or concerns be handled and escalated?

If participants wish to make reports of concerns or incidents relating to potential exploitation, abuse, or harm resulting from their involvement in this project, They might contact the project's Designated Safeguarding Dr. Morgan Harvey; m.harvey@sheffield.ac.uk. If the concern or incident relates to the Designated Safeguarding Contact, or if the participant feel the reports they have made to this Contact has not been handled in a satisfactory way, the participants might contact
the Head of the information school Professor Val Gillet; v.gillet@sheffield.ac.uk and/or the University's Research Ethics & Integrity Manager (Lindsay Unwin; l.v.unwin@sheffield.ac.uk).

Section E: About the data

1. Data Processing

Will you be processing (i.e. collecting, recording, storing, or otherwise using) personal data as part of this project? (Personal data is any information relating to an identified or identifiable living person).
No

Please outline how your data will be managed and stored securely, in line with good practice and relevant funder requirements

The obtained data will be stored on a Qualtrics server, which meets the university's requirements for data security and is GDPR-compliant. Then copies of the data will be downloaded for analysis and saved on the University of Sheffield's Google Drive. The data will be securely stored in the University's Google Drive and shared only with
the supervisors.

Section F: Supporting documentation

Information & Consent

Participant information sheets relevant to project?
Yes

Document 1125177 (Version 2) — All versions

Consent forms relevant to project?
Yes

Document 1125178 (Version 2) — All versions

Document 1126618 (Version 1) — All versions
Data management plan

Additional Documentation

External Documentation

- not entered -

Section G: Declaration

Signed by:
Amnah Alluqmani
Date signed:
Fri 14 July 2023 at 21:02

Offical notes

- not entered -

Appendix F

More Results of The Graphic Description Generation



BLIP-2 Caption : lady gaga hat tote bag.

Automatic Questioning:
Question: 1. What color is the hat on the tote bag?
Question: 2. Can you describe the shape of the hat on the tote bag?
Question: 3. Is the hat on the tote bag plain or does it have any embellishments?
Question: 4. Are there any other designs or patterns on the tote bag besides the hat?
Question: 5. Can you describe the overall style or aesthetic of the tote bag featuring Lady Gaga's hat?

Automatic Answering:
Answer 1: pink
Answer 2:
Answer 3: plain
Answer 4: yes, there are other designs and patterns on the tote bag besides the hat
Answer 5: Lady Gaga hat tote bag

Automatic Summarising: The presented graphic is a pink tote bag featuring Lady Gaga's hat. The hat is plain and has no embellishments. In addition to the hat, there are other designs and patterns on the tote bag.



BLIP-2 Caption: peace and love - peace and love - peace and love - peace and love - peace and love -

Automatic Questioning:
Question: 1. What colors are used in the graphic of peace and love?
Question: 2. Are there any symbols or images that represent peace and love in the graphic?
Question: 3. How are the words "peace and love" visually represented in the graphic?
Question: 4. Is there a specific layout or design that conveys the message of peace and love in the graphic?
Question: 5. Can you describe the overall feeling or mood that the graphic of peace and love evokes?

Automatic Answering:
Answer 1: red, black, yellow, green, blue, purple, orange, pink, white, and black
Answer 2: Yes, there are symbols and images that represent peace and love in the graphic
Answer 3:
Answer 4: Yes, the design conveys the message of peace and love
Answer 5: peace and love

Automatic Summarising:
The presented graphic is a colorful representation of peace and love, featuring a variety of colors such as red, yellow, green, blue, purple, orange, pink, white, and black. There are symbols and images that represent peace and love included in the graphic, and the words "peace and love" are visually represented as well. The layout and design of the graphic effectively convey the message of peace and love, evoking a feeling of peace and love overall



BLIP-2 Caption: a red shirt with the words burberry established in london

Automatic Questioning:
Question:1. Can you describe the shade of red on the shirt?

Question:2. How is the word "Burberry" written on the shirt? Is it in cursive or block letters?
Question:3. Where on the shirt is the word "Burberry" located?
Question:4. Are there any other designs or patterns on the shirt besides the word "Burberry"?
Question:5. Can you describe the overall style or cut of the shirt?

Automatic Answering:
Answer 1: red
Answer 2: Block letters
Answer 3: On the left side of the shirt
Answer 4: Yes, there are other designs on the shirt besides the word "Burberry"
Answer 5: The overall style of the shirt is a classic, with a classic collar and a classic button down front

Automatic Summarising:
The presented graphic is a red shirt with the word "Burberry" written in block letters on the left side. In addition to the word "Burberry," there are other designs on the shirt. The overall style of the shirt is classic, with a classic collar and a button-down front.