Multimodal AI for Clothing Assistive Solutions for the Visually Impaired

By

Kathure, Bessy Mukaria

169111

Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Data Science and Analytics at Strathmore University

Institute of Mathematical Sciences and iLab Africa

Strathmore University

Nairobi, Kenya

June, 2025

Declaration and Approval

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the dissertation contains no material previously published or written by another person except where due reference is made in the dissertation itself.

Student's Name: **Kathure Bessy Mukaria**

Sign: 

Date: 21st May 2025

Approval

The dissertation of **Kathure Bessy Mukaria** was reviewed and approved by the following:

Dr. Kennedy Senagi,
Supervisor, Institute of Mathematical Sciences,
Strathmore University.

Dr. Godfrey Madigu,
Dean, Institute of Mathematical Sciences,
Strathmore University.

Prof. Bernard Shibwabo,
Director of Graduate Studies,
Strathmore University.

Abstract

This study presents an Artificial Intelligence (AI) powered Image-to-Text-to-Speech (ITTS) system to enhance accessibility for visually impaired individuals in the clothing domain. Using the DeepFashion2 dataset, the Bootstrapped Language Image Pretraining (BLIP) model generated enriched captions, integrating metadata such as clothing scale, viewpoint, and category. These enriched captions were synthesized into audio using Google Text-to-Speech (gTTS), offering an accessible and descriptive experience. The system's performance was evaluated under zero-shot and fine-tuned settings, demonstrating substantial improvements in Bilingual Evaluation Understudy (BLEU)-1 (from 0.09 to 0.19), BLEU-2 (from 0.04 to 0.07), BLEU-3 (from 0.02 to 0.04), Recall-Oriented Understudy for Gisting Evaluation (ROUGE-L) remained stable at 0.16. At the same time, Metric for Evaluation of Translation with Explicit Ordering (METEOR) improved from 0.09 to 0.13. Although Consensus-based Image Description Evaluation (CIDEr) scores remained at 0.0, the fine-tuned model excelled in generating contextually rich and descriptive captions due to metadata integration. This study highlights the potential of multimodal AI systems,whose performance was evaluated using BLEU and other standard metrics, to address accessibility challenges, providing a solution to empower visually impaired users and laying the groundwork for future innovations in inclusive design.

Keywords: Multimodal AI, Assistive Reading, Digital Accessibility, Fashion Content, Image-to-Speech, Inclusive Design.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AI Artificial Intelligence

API Application Programming Interface

BLIP Bootstrapped Language Image Pretraining

BLEU Bilingual Evaluation Understudy

CIDEr Consensus-based Image Description Evaluation

CNN Convolutional Neural Network

CRISP-DM Cross Industry Standard Process for Data Mining

CSS Cascading Style Sheets

HTTPS HyperText Transfer Protocol Secure

gTTS Google Text-to-Speech

GPT-2 Generative Pretrained Transformer 2

HTML HyperText Markup Language

ICT Information and Communication Technology

ITTS Image-to-Text-to-Speech

JSON JavaScript Object Notation

JPG Joint Photographic Experts Group

METEOR Metric for Evaluation of Translation with Explicit Ordering

ML Machine Learning

PNG Portable Network Graphics

RNN Recurrent Neural Network

RGB Red, Green, Blue

ROUGE-L Recall-Oriented Understudy for Gisting Evaluation

S3 Simple Storage Service

TTS Text to Speech

SDG Sustainable Development Goal

URL Uniform Resource Locator

UN United Nations

USF Universal Service Fund

NLP Natural Language Processing

KISE Kenya Institute of Special Needs Education

LCS Longest Common Subsequence

ViT Vision Transformer

WCAG Web Content Accessibility Guidelines

WEBP Web Picture format

WHO World Health Organization

YOLO You Only Look Once

Acknowledgments

I am deeply grateful to the Almighty God for His unwavering grace, provision, and for sustaining my health throughout this academic journey.

I sincerely thank my parents, especially my father, for their constant encouragement, love, and support. Your belief in me has been a pillar throughout this endeavor.

My heartfelt appreciation goes to my supervisor, Dr. Kennedy Senagi, for his invaluable guidance, mentorship, and patience. Your insights and encouragement were instrumental in shaping this research.

I would also like to acknowledge the Chinese University of Hong Kong (CUHK) for providing access to the DeepFashion2 dataset, which played a critical role in developing and evaluating this project.

Lastly, I extend my thanks to all faculty, colleagues, and friends who provided support, inspiration, and motivation throughout my studies.

Chapter 1: Introduction

1.1 Background

The digital revolution has significantly transformed how people access, exchange and interact with information. However, the benefits of this digital transformation have not been equitably distributed, particularly for marginalized groups such as individuals with visual impairments. Approximately 2.2 billion people have some form of vision impairment, and at least 1 billion have vision impairment that could have been prevented or remains unaddressed (World Health Organization, 2020). In Kenya, visual impairment is a significant public health challenge, with an estimated 3.9 million people experiencing some form of vision loss, and around 290,000 being completely blind (The International Agency for the Prevention of Blindness, 2020). These figures highlight a glaring gap in digital accessibility, disproportionately affecting people in rural areas who face barriers to accessing essential digital content and services.

The Nakuru Eye Disease Cohort Study by (Bastawrous et al., 2016) contributes to this body of knowledge by presenting a 6-year cumulative incidence of visual impairment and blindness, providing a longitudinal perspective on these conditions in the adult population. Further research conducted as part of the Kenya Rural Blindness Prevention Project by (Whitfield et al., 1990) reveals that in rural Kenya, the prevalence of blindness reaches 0.7% according to World Health Organization (WHO) standards, with untreated cataracts being the leading cause. This rural-urban divide exacerbates the challenges of digital inclusion, as rural areas tend to have limited access to healthcare, education, and technological infrastructure.

In today's information-driven world, access to the internet and digital platforms is crucial for participation in everyday life, education, employment, and social engagement. Yet, visually impaired individuals face significant disadvantages due to digital platforms lacking necessary accessibility features (Suomi and Sachdeva, 2016; Joisten et al., 2015). This digital divide leaves many visually impaired Kenyans unable to fully engage with society, denying them opportunities for self-development, social inclusion, and economic empowerment (Takshara and Bhuvaneswari, 2025; Eziamaka et al., 2024).

The clothing industry presents a more complex challenge for people with visual impair-

ments. Clothing, a highly visual field, relies on detailed descriptions of colors, patterns, textures, and styles that current assistive technologies like screen readers often do not convey adequately (Joisten et al., 2015; Eziamaka et al., 2024). This limits the ability of visually impaired users to independently explore clothing products, styles, and trends, contributing to their exclusion from the growing digital clothing space. Addressing this gap is essential, as clothing is a form of self-expression and an important aspect of personal identity and social participation (Takshara and Bhuvaneswari, 2025).

This research advocates using an ITTS system to address this issue. Using advances in AI and ML, ITTS systems are designed to convert visual information into an auditory format, making digital content accessible to vision-loss users. This study particularly emphasizes the use of transformer models, which, unlike traditional models such as Recurrent Neural Network (RNN) like seq2seq (Sutskever et al., 2014) or Convolutional Neural Network (CNN) (Kalchbrenner et al., 2017), can process sequential data efficiently and effectively—their architecture, which can capture dependencies through attention mechanisms, suits transformers for ITTS systems.

This technological solution aligns closely with larger goals to promote inclusion and minimize inequalities, as outlined by the Sustainable Development Goal (SDG)s of the United Nations (UN). In particular, this study supports Goal 10: Reduced Inequalities by promoting inclusive and sustainable innovation (Tebbutt et al., 2016; Harb and Sidani, 2021).

Collaborative efforts from stakeholders, including inABLE and Kenya Institute of Special Needs Education (KISE), under the Ministry of Education, exemplify a commitment to inclusion and demonstrate the potential of technology to empower rather than exclude individuals. This research argues that access to digital media should be recognized as a universal right, not a privilege (World Wide Web Consortium (W3C), 2018). Creating a digital ecosystem characterized by integration and inclusion ensures that visual impairments do not prevent people from fully participating in the digital world. The goal is to make the benefits of the Internet accessible to everyone, regardless of physical limitations.

1.2 Problem Statement

In Kenya, visually impaired individuals, particularly those in rural areas, face significant barriers to digital accessibility despite government efforts through initiatives such as the

National Information and Communication Technology (ICT) Policy. Although the digital divide has been acknowledged, it remains wide, especially for individuals with visual impairments. It leaves many marginalized from accessing essential digital services for education, employment, and social participation (Nyamweya et al., 2024), including in visually driven sectors such as clothing. Assistive technologies, such as speech-generating devices and braille displays, can improve digital access. However, their implementation in the clothing industry is limited, with many clothing platforms failing to comply with Web Content Accessibility Guidelines (WCAG) (Himayah and Hasan, 2022). This results in usability challenges, where visually impaired individuals struggle to access detailed clothing information such as colors, textures, and patterns, crucial elements in clothing exploration and shopping experiences.

Additionally, there is a significant lack of targeted ICT capacity-building programs tailored to the specific needs of people with disabilities in sectors like clothing, further limiting their ability to engage effectively with digital platforms and tools. Although initiatives like the Universal Service Fund (USF) aim to bridge this gap, issues of transparency and allocation reduce their effectiveness, leaving the visually impaired, especially in rural regions, excluded from the opportunities provided by digital clothing platforms and other inclusive digital efforts (Wambua, 2021). These systemic challenges underscore the need for more comprehensive strategies to ensure that visually impaired individuals are not excluded from participating fully in the clothing industry's digital transformation.

1.3 Research Aim

The study's main objective is to develop and implement an AI-powered ITTS system to enhance digital accessibility for visually impaired individuals in Kenya.

1.4 Research Objectives

This research aimed to address the following objectives:

(a) To review the literature to identify existing Machine Learning (ML)-based approaches for tackling digital content accessibility for the visually impaired in Kenya.

(b) To develop and test the effectiveness of ViT and BLIP models tailored to image classification and caption generation, respectively.

(c) Integrate machine learning-based text-to-speech technology to generate natural context-aware speech from captions.

(d) Deploy the solution on a web browser.

1.5 Research Questions

The research questions addressed in this study were as follows:

(a) What are the existing ML-based approaches for improving the accessibility of digital content for visually impaired people?

(b) How can machine learning be optimized for extracting meaningful features from images to generate accurate and contextually relevant captions?

(c) How can ML-based text-to-speech technology be integrated with image captioning models to provide natural and context-aware speech for visually impaired users?

(d) How can deploying ITTS systems on a web browser contribute to digital accessibility efforts for the visually impaired?

1.6 Scope and Limitations of the Study

1.6.1 Scope

The study focuses on developing an ITTS system using AI and ML to improve digital accessibility for visually impaired individuals in rural and urban Kenya, enabling them to independently engage with clothing content during online and in-store shopping experiences. The accuracy of AI and ML models may limit the system's performance, particularly in complex environments where real-time processing of diverse clothing styles and textures may lead to less accurate descriptions.

1.6.2 Limitations

Here are some of the limitations of this research work:

(a) Real-Time Processing Challenges: In dynamic or fast-paced environments, real-time image processing may be less reliable, leading to delayed or inaccurate descriptions.

(b) Model Accuracy Constraints: The system's effectiveness depends on the accuracy

of the AI and ML models, which may struggle in cases involving complex clothing textures, patterns, and colors.

(c) Diversity of Clothing Styles: The dataset used for training may not cover the full range of cultural and regional clothing variations, potentially reducing performance for specific styles.

1.7 Research Justification

This study is justified by its contribution to improving digital accessibility for visually impaired individuals in Kenya, aligning with the *UN SDG*s. By developing an AI-powered *ITTS* system, this research directly supports *SDG 10* (Reducing Inequality) by enabling visually impaired individuals to access clothing-related content independently, bridging the accessibility gap in a highly visual industry. It also advances *SDG 4* (Quality Education) by providing visually impaired students with AI-driven tools for engaging with clothing education, ensuring equitable learning opportunities. Furthermore, this study contributes to *SDG 9* (Industry, Innovation, and Infrastructure) by leveraging AI and machine learning to create accessible and inclusive digital retail experiences. Beyond theoretical contributions, this research provides a practical solution for digital inclusion, ensuring visually impaired individuals can navigate clothing-related content with ease (United Nations, 2015; Government of Kenya, 2007).

Chapter 2: Literature Review

2.1 Introduction to Multimodal AI

Multimodal AI represents a significant advancement in creating systems that integrate visual, textual, and auditory data into a unified framework. By processing these different types of information, multimodal AI allows for a more holistic understanding of complex content. This is particularly valuable in developing an Image-to-Text-to-Speech (ITTS) system to improve accessibility in the clothing industry.

In the clothing industry, ITTS systems powered by multimodal AI offer several key benefits. They enable accurate descriptions of visual details, which can then be converted into natural speech. This allows visually impaired users to gain a richer understanding of clothing items. Traditional systems that rely on text or image data struggle to provide the same level of detail. By combining visual and textual inputs, multimodal AI ensures that descriptions are precise and contextually relevant, matching the clothing industry's fast-paced and visually rich environment. For instance, Srinivasan and San Miguel González (2022) highlights how empathy-inspired multimodal AI systems leverage visual and textual data to provide a deeper, context-aware understanding, demonstrating the potential for similar advancements in accessibility-focused applications like ITTS systems.

2.2 Related Works

Assistive technologies for visually impaired individuals have advanced significantly through multimodal AI, integrating visual, auditory, and textual data to enhance accessibility. Despite these advancements, current technologies still face limitations, particularly in dynamic environments such as shopping or clothing selection, where real-time assistance is crucial.

Virtual AI assistants capable of recognizing objects, text, and people for individuals with partial vision impairment are one such advancement (Raghavan, Rohith et al., 2021). These systems leverage CNNs to perform real-time recognition on Android devices, enabling accurate recognition in varied scenarios. However, reliance on mobile devices and a lack of physical assistance for complex tasks are limiting factors.

The VQAsk application, a multimodal platform utilizing Natural Language Processing

(NLP) and computer vision, effectively integrates NLP and visual analysis for real-time interactions (De Marsico et al., 2024). However, it struggles in complex visual scenes, limiting its use in scenarios such as shopping that require detailed contextual feedback.

Image captioning technology aims to generate descriptive captions for visual content, improving accessibility for visually impaired individuals. A study incorporating detected text into the AoANet model led to a 35% improvement in caption accuracy (Ahsan et al., 2021). However, the system remains limited to static images, which presents challenges in dynamic environments requiring real-time updates.

Multimodal AI systems combining CNNs for visual analysis and text-to-speech (TTS) for auditory output have shown promise, especially in educational contexts (A et al., 2023). However, adapting these systems for real-time assistive technologies remains challenging.

Large multimodal models, such as Be My AI, provide real-time visual descriptions using Vision Transformer (ViT) architecture (Xie et al., 2024). However, their lack of goal-oriented guidance reduces their effectiveness in specific tasks like shopping.

Wearable technologies, like a deep learning-powered wearable device using You Only Look Once (YOLO)v4 for real-time object detection and guidance, improve user safety and autonomy (Mohanraj et al., 2024). However, the high cost of hardware remains a significant barrier to accessibility.

In fashion, smart glasses equipped with YOLOv3 object detection and speech synthesis provide real-time auditory feedback on clothing items (Kumar et al., 2024). While effective, the cost limits their accessibility.

These studies highlight substantial progress in multimodal AI for visually impaired individuals, particularly in object detection, image captioning, and wearable technology. However, gaps remain in providing real-time, goal-oriented assistance in dynamic environments. Addressing these challenges is key to advancing the next generation of assistive technologies in clothing selection, navigation, and interactive settings.

2.3 Limitations of Current Assistive Technologies

Current assistive technologies, such as screen readers, struggle to convey essential visual information like color, texture, or style, which are crucial for understanding clothing.

Alt-text descriptions often fail to capture critical details such as fabric sheen, texture, or fit (Michele A. Williams and Hurst, 2013). Although sound-based feedback systems and interactive touchless technologies are being explored (Armstrong, 2015), a need remains for AI-driven solutions that can interpret and translate complex visual data into accessible formats for the clothing industry.

2.4 Advances in AI for Clothing Description Generation

Traditional methods like alt-text have been used to describe clothing items, but these descriptions are often too simplistic (Michele A. Williams and Hurst, 2013). Modern AI models generate more detailed, context-aware descriptions, analyzing garments to extract comprehensive details such as fabric quality, patterns, and styling (Jagadish et al., 2024).

2.4.1 Transformer-Based Image Captioning

The transition from rule-based systems to AI-driven models, especially transformers, has revolutionized clothing description generation. Transformer models like Vision Transformer (ViT) process images by dividing them into patches, capturing fine-grained details such as texture and design (Khalid and Gong, 2022). These models generate accurate, context-rich descriptions, enabling real-time applications for visually impaired users (Lamchoudi et al., 2024).

2.4.2 Bridging Visual Data and Language

BLIP integrates visual data with language, bridging the gap between modalities. BLIP excels in clothing applications, producing highly descriptive and accurate captions, enabling TTS systems to provide detailed, real-time information about clothing items (Li et al., 2022). The feature extraction capabilities of models like BLIP simplify generating captions from images, making them highly accessible for testing and practical applications. By leveraging the strengths of multimodal learning, these models can generate more accurate and contextually relevant captions, significantly improving digital accessibility for visually impaired individuals (Vaswani et al., 2023; Xu et al., 2021; Mokady et al., 2021).

2.4.3 Text-to-Speech Integration

Multimodal AI systems that integrate visual and language processing capabilities provide the foundation for advanced ITTS systems. Whisper, developed by OpenAI, offers an ideal Text to Speech (TTS) solution for clothing descriptions, delivering nuanced, context-aware speech tailored to visually impaired users (Radford et al., 2022). However, due to resource constraints, this study did not implement Whisper and instead used gTTS for its ease of integration and multilingual support.

2.5 Future Directions for ITTS in clothing

Future research should focus on improving TTS systems' expressiveness and naturalness to convey better the emotional and symbolic meanings embedded in clothing. Developing real-time captioning systems for fast-moving clothing events and expanding ITTS systems to include tactile descriptions will enhance their utility. By addressing these gaps, visually impaired users can engage more deeply with clothing, creating a more inclusive digital clothing landscape.

Applying transformer-based models like ViT, BLIP, and gTTS in ITTS systems holds immense potential to improve digital inclusion in the clothing industry. As clothing companies adopt these AI-powered solutions, the industry will be better positioned to close the accessibility gap, making clothing more inclusive for all.

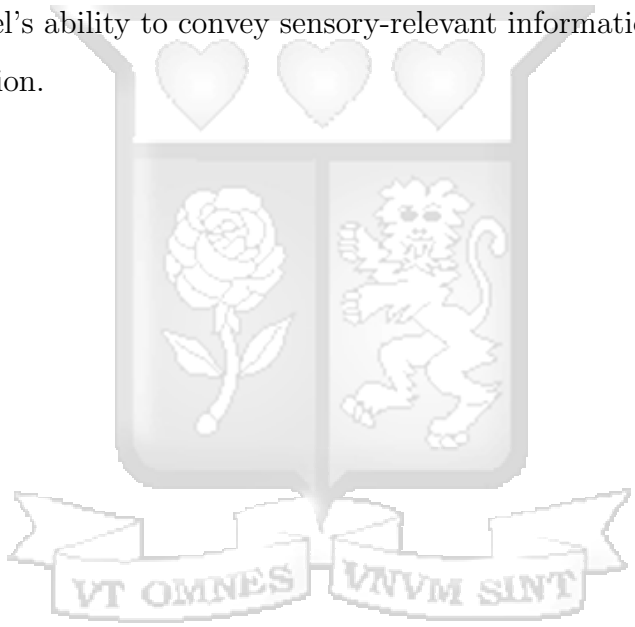
2.6 Gaps in clothing Accessibility for the Visually Impaired

The literature identifies several gaps in existing assistive technologies for visually impaired individuals in clothing. Burton (2011) emphasizes that while clothing is central to personal expression, current technologies fail to convey important aspects of clothing, such as the drape and texture of fabrics (Burton, 2011). Furthermore, Withana (2023) highlights the potential for personalized assistive technologies, such as wearable devices that mimic mainstream clothing accessories, to enhance accessibility and social inclusion. These findings suggest a clear need for more advanced assistive technologies to offer visually impaired users a richer and more nuanced understanding of clothing.

The clothing industry faces significant accessibility challenges due to the visual complexity of its content. Clothing items often feature intricate details that are difficult for current AI

systems to capture and describe in real time. These systems also struggle to convey tactile qualities, such as fabric texture and drape, which are crucial to understanding clothing. Addressing these gaps is essential for fostering greater inclusivity in the clothing industry.

This study responds to these gaps by leveraging structured data from the DeepFashion2 dataset to enhance the descriptive richness of visual outputs. While current assistive systems struggle to simulate tactile qualities such as texture, or weight, this research approximates them through metadata variables like clothing size, visibility, zoom level, and fit. These attributes, when converted into enriched textual descriptions (e.g., fully visible oversized coat or 'tight-fitted sleeveless top'), offer visually impaired users a bridge between visual details and inferred tactile understanding. This metadata-driven approach strengthens the model's ability to convey sensory-relevant information without requiring direct tactile simulation.



Chapter 3: Methodology

This research followed the Cross-Industry Standard Process for Data Mining Cross Industry Standard Process for Data Mining (CRISP-DM), a widely used framework for structuring machine learning and data-driven projects. The methodology (RuchaReads, 2021) comprised six key phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Each phase was crucial in ensuring a structured approach to developing an ITTS system that provided detailed auditory descriptions of clothing items for visually impaired individuals. CRISP-DM comprises several essential phases organized in an iterative process.

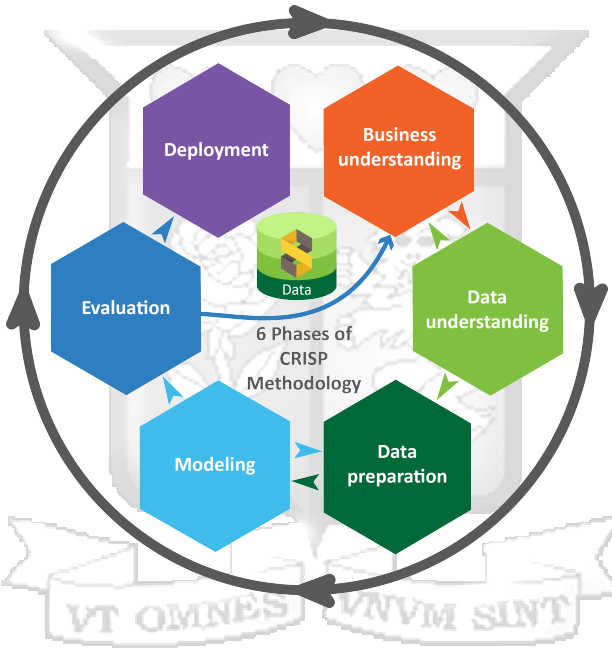


Figure 3.1: CRISP-DM framework outlining the standard process for data mining projects (source: (RuchaReads, 2021)).

3.1 Business Understanding

In Kenya and worldwide, visually impaired individuals faced challenges making independent clothing choices due to the limited descriptive capabilities of existing assistive technologies (Fernando et al., 2025). Statistically, over 2.2 billion (World Health Organization, 2020) people globally live with vision impairments. While some tools provided essential object recognition, they often failed to capture detailed clothing attributes and contexts in an audio way.

To address this gap, this study explored the use of multimodal AI, specifically vision

transformers and NLP, to develop an ITTS system. The system was designed to analyze images, generate text descriptions of clothing, and convert the descriptions into speech, allowing visually impaired individuals to receive meaningful auditory feedback.

3.2 Data Understanding

The DeepFashion2 dataset (Ge et al., 2019) was utilized as the primary data source for training and evaluating the ITTS system. DeepFashion2 provides a realistic alternative to datasets like Microsoft COCO (Lin et al., 2015), which features professionally captured and annotated images. This provides inclusive user-contributed data and real-world scenarios that enable the development of models capable of addressing practical challenges that visually impaired users face.

Although DeepFashion2 was collected for clothing recognition in commercial and academic settings, this research applies it to a different use case: enhancing clothing accessibility for blind users in Kenya. The dataset serves as a proxy for local clothing data, enabling the development of an assistive prototype in a low-resource context.

The dataset included various attributes that were essential for training the AI model. The key variables in the dataset are summarized in Table 3.1, detailing their descriptions, data types, and example values.

Table 3.1: Variables in the DeepFashion2 dataset.

Variable	Description	Data Type
image_path	Path to the image file	String
category_name	Type of clothing item	String
category_id	Numerical ID assigned to clothing type	Integer
bounding_box	Coordinates of the bounding box for clothing item	List
segmentation	Pixel-wise segmentation mask for the item	List
landmarks	Key points on the clothing item (e.g., collar, sleeve)	List
occlusion	Level of occlusion (how much of the item is blocked)	Integer
scale	Relative size of the clothing item in the image	Integer
viewpoint	Perspective from which the clothing is viewed	Integer
zoom_in	Whether the image is zoomed in on the item	Integer
style	Clothing style information (if available)	Integer

3.2.1 Annotation Standards and Dataset Diversity

While DeepFashion2 provides high-resolution images and detailed attribute labels, it lacks transparency around annotation standards and contributor diversity. Sourced primarily

from online platforms, the dataset reflects commercial and Western-centric fashion norms. Annotations are largely focused on female clothing categories, with limited cultural or regional representation. These gaps challenge the development of inclusive assistive systems and may reduce caption relevance for users outside the dataset's primary demographic. Future work should explore datasets with more balanced category representation and culturally inclusive annotations.

3.3 Data Preparation

3.3.1 Dataset Curation and Cleaning

The project's initial phase involved curating and cleaning the dataset to ensure consistency and quality in preparation for model development. This process was conducted using the Python programming language (Van Rossum and Drake Jr, 1995), which offers a set of libraries appropriate for data wrangling and inspection.

The dataset comprised 32,153 annotation files and 21,919 corresponding images. To validate data integrity, the directory structures were traversed using the os and glob modules, enabling the identification of mismatches between image and annotation files. The annotation files, stored in JavaScript Object Notation (JSON) format, were parsed using the native JSON library. This made it possible to verify that each annotation file was adequately linked to a valid image and that key properties such as item categories and bounding boxes were intact.

A merge operation was carried out using the pandas (McKinney et al., 2010) library, wherein image file names were joined with corresponding annotations to address the observed discrepancy in file counts. Annotations lacking a matching image were excluded from the dataset to prevent model misalignment during training. This resulted in a reduced but cleaner set of valid image-annotation pairs.

Following cleaning, the dataset underwent a category review—an initial frequency distribution, performed using the collections.Counter utility revealed that some classes, specifically *sling* and *short sleeve outerwear*, were underrepresented. These categories were subsequently excluded from further use due to their insufficient sample sizes, which could introduce bias and hinder model generalization.

The remaining 11 categories were then balanced using stratified sampling. The StratifiedShuffleSplit function from the scikit-learn library (Pedregosa et al., 2011) was employed to ensure an even distribution of samples across classes. One hundred samples per category were selected, yielding a final, balanced dataset of 1,100 images. This curated dataset formed the foundation for all subsequent training and evaluation procedures.

3.3.2 Data Pre-processing for Model Compatibility

Following dataset curation, several pre-processing steps were applied to prepare the images for use with the BLIP and ViT models. These transformations were implemented using Python (Van Rossum and Drake Jr, 1995), with support from key libraries such as Pillow (Clark, 2015), torchvision (Marcel and Rodriguez, 2010), and PyTorch (Paszke et al., 2019a). These pre-processing techniques are widely adopted and recommended in deep learning pipelines for vision tasks and align with best practices outlined in literature (Heaton, 2018). The steps included:

- **Image Resizing:** All images were resized to 224×224 pixels using the Resize transformation from the torchvision.transforms module (Marcel and Rodriguez, 2010). This ensured consistency with the expected input dimensions of both models.
- **Color Normalization:** Each image was converted to Red, Green, Blue (RGB) format using PIL.Image.convert("RGB") from the Pillow library (Clark, 2015), ensuring a uniform color representation across the dataset.
- **Pixel Value Normalization:** The Normalize function from torchvision.transforms (Marcel and Rodriguez, 2010) was applied using ImageNet normalization statistics—mean $[0.485, 0.456, 0.406]$ and standard deviation $[0.229, 0.224, 0.225]$—to align the images with the pretraining conditions of the transformer models used.
- **Tensor Conversion:** Images were transformed into tensors using ToTensor() and adjusted to include a batch dimension for model compatibility. This was achieved using the PyTorch framework (Paszke et al., 2019a).
- **Sharpness Enhancement:** To improve image clarity and emphasize fine-grained details, a sharpening filter was applied using ImageFilter.SHARPEN from the Pillow library (Clark, 2015).

Image Enhancement for Captioning Accuracy Although visually impaired users do not see the images, enhancement improves visual input quality for the model. Clearer images help the captioning model extract more accurate features, leading to better audio descriptions. Thus, enhancement indirectly supports accessibility by improving output relevance.

3.3.3 Metadata Enrichment

To enhance the dataset with ground truth captions, metadata extracted from the DeepFashion2 (Ge et al., 2019) annotations was used to generate structured textual descriptions of clothing items for each image. This metadata included attributes such as clothing scale, occlusion level, viewpoint, zoom level, and category name, which were mapped to human-readable descriptions. For example, an image labeled as a large-sized short sleeve top, fully visible, not being worn, and shown in a close-up view was assigned the caption: "The clothing item is a large-sized short sleeve top, fully visible, not being worn, and shown in a close-up view." The metadata-enriched captions served as the ground truth for evaluating and comparing the outputs of the BLIP captioning model and the ViT classifier for each image. This enriched representation ensured an additional variable to serve as a ground truth to train the images. This is to overall improve model performance by adding a broader range of input variations as highlighted by (Yang et al., 2023).

3.3.4 Final Dataset Summary

The final dataset comprised 1,100 images, evenly distributed across 11 clothing categories. Each image was paired with a corresponding JSON annotation, containing category labels, bounding boxes, segmentation masks, and additional metadata. The dataset structure aligns with best practices observed in large-scale benchmarks such as ImageNet (Deng et al., 2009), where structured and annotated data significantly improve model training and evaluation. The dataset was preprocessed to meet the requirements of the BLIP captioning model, to enhance deep learning model performance (Heaton, 2018). After pre-processing, it was stored in a Simple Storage Service (S3) (Amazon Web Services, 2023) bucket to maintain a structured format optimized for model training and inference.

3.4 Machine Learning Model Development

The ITTS system was developed using two computer vision models: Bootstrapped Language Image Pretraining (BLIP) (Li et al., 2022) for image captioning and Vision Transformer (ViT) for image classification. This section details the modeling approach, architecture, evaluation strategy, and optimization processes for both models.

3.4.1 Vision Transformer (ViT): Image Classification

Model Architecture ViT-Generative Pretrained Transformer 2 (GPT-2) Transformer model (Dosovitskiy et al., 2021) combines the capabilities of a Vision Transformer (ViT) for image encoding and GPT-2 for language modeling. The Vision Transformer architecture comprises an encoder-decoder structure as introduced in the paper (Vaswani et al., 2023) "Attention Is All You Need".

The encoder consists of layers in Figure 3.2 that process the input iteratively, one layer after another. The decoder layers receive the output from the encoder and generate a decoded output. Simply put, the encoder maps the input sequence to a sequence fed into the decoder, generating an output sequence.

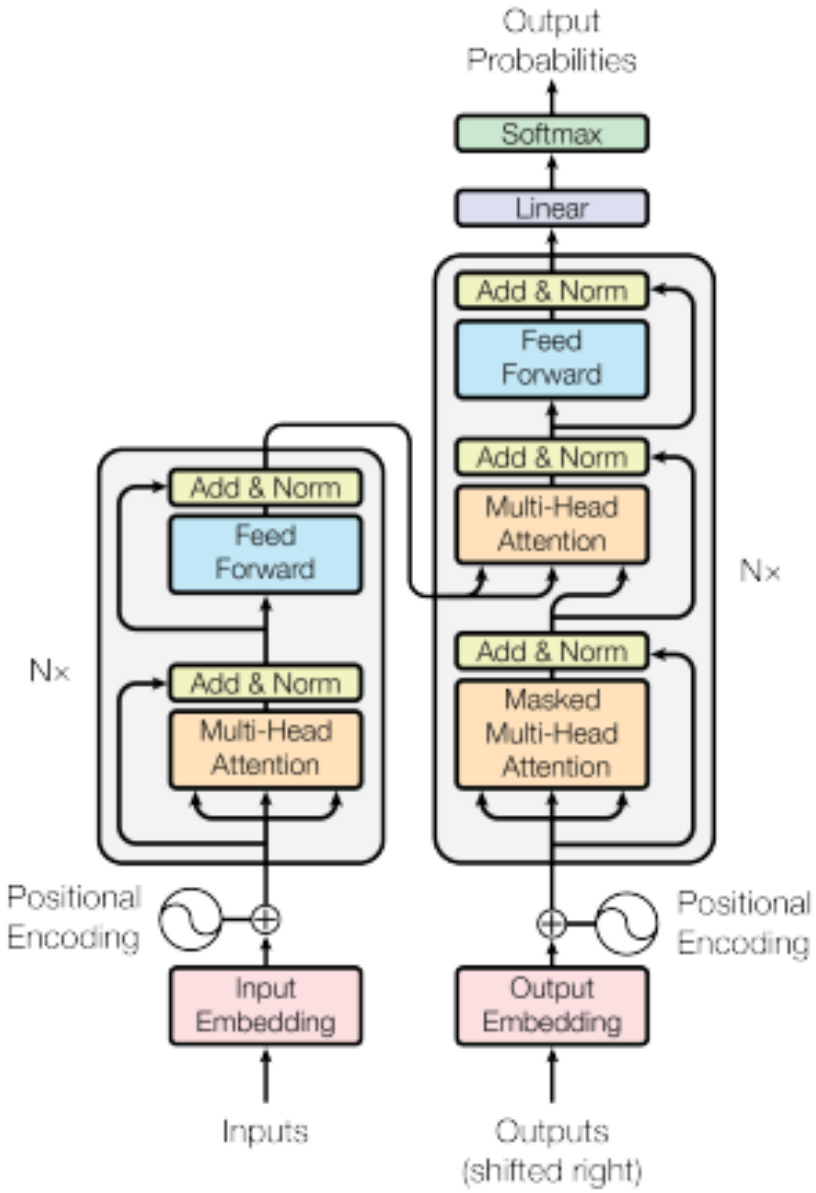


Figure 3.2: The ViT Transformer model architecture with encoder-decoder structure.

Zero-Shot Classification and Evaluation The ViT model was evaluated in a zero-shot setting, meaning it was used without any additional fine-tuning on the DeepFashion2 dataset. Images were resized to 224×224 pixels, converted into tensors, and classified directly. ViT predicted each image's most likely clothing category label. The generated descriptions were then compared to structured metadata-enriched ground truth captions using standard evaluation metrics such as BLEU, METEOR, ROUGE-L, and CIDEr, providing an initial assessment of the ViT (Dosovitskiy et al., 2021) model's performance in descriptively classifying clothing items.

3.4.2 Bootstrapped Language Image Pretraining (BLIP): Image Captioning

Model Architecture BLIP Li et al. (2022) integrates vision and language tasks within a single framework. It employs a multimodal encoder-decoder architecture to generate meaningful captions directly from image inputs. BLIP's architecture includes:

- *Unimodal Encoder*: Separately encodes image and text inputs using a vision transformer and a BERT-like text encoder.
- *Image-Grounded Text Encoder*: Enhances textual representations by incorporating cross-attention layers informed by visual features.
- *Image-Grounded Text Decoder*: Replaces bidirectional self-attention with causal self-attention layers to generate textual descriptions from images.

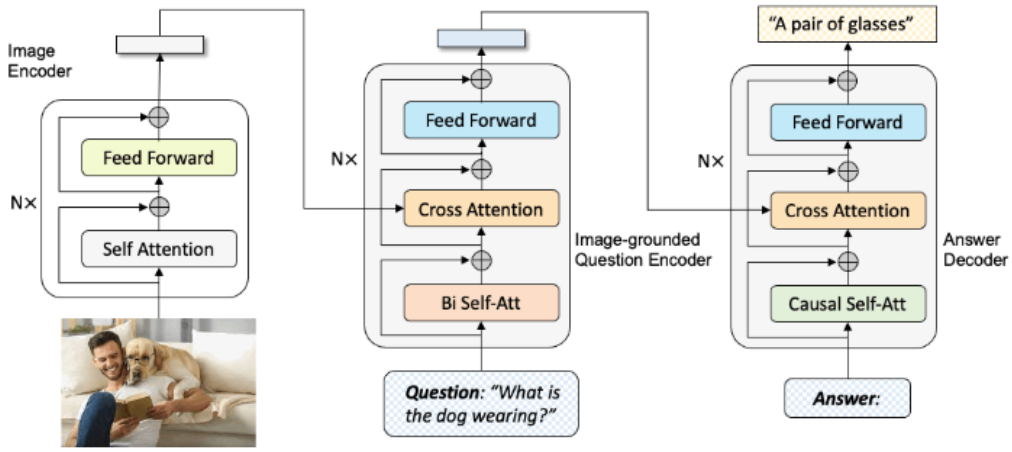


Figure 3.3: BLIP model architecture combining visual encoding and language decoding.

The caption generation process follows an autoregressive framework, predicting each subsequent word based on both the image content and previously generated words. This can be mathematically represented as:

$$P(C|I) = \prod_{t=1}^{T} P(c_t|c_{<t}, I) \tag{1}$$

where C is the full caption, I is the input image, T is the length of the caption, c_t denotes

the word predicted at time step t, and $c_{<t}$ represents the sequence of preceding words.

The model uses this probability distribution to sequentially generate the most likely next word in the caption, conditioned on the image and preceding words, following an autoregressive decoding strategy.

Zero-Shot Caption Generation BLIP was initially evaluated in a zero-shot scenario, meaning no additional fine-tuning was performed using the DeepFashion2 dataset. Pre-processed images were converted into PyTorch tensors, resized to 224 × 224 pixels, and passed through the pre-trained BLIP model to produce captions. These initial, zero-shot captions were generated directly from the model's existing knowledge without adaptation to the specific dataset, establishing a performance baseline.

Caption Comparison and Evaluation To evaluate BLIP's zero-shot descriptive capabilities, the generated captions were compared against structured metadata-enriched captions derived from the DeepFashion2 annotations. Standard caption evaluation metrics, BLEU, ROUGE-L, METEOR, and CIDEr—were employed to assess caption quality, coherence, and relevance quantitatively. This comparison provided insights into BLIP's initial ability to capture critical visual details and informed subsequent model fine-tuning strategies.

3.4.3 Model Selection

Following the initial zero-shot evaluations of ViT for image classification and BLIP for image captioning, a comparative assessment informed the model choice for subsequent fine-tuning. The evaluation focused on each model's ability to accurately represent clothing images, as measured by BLEU, ROUGE-L, METEOR, and CIDEr scores. The comparison provided insights into which model better captured essential visual details and generated more contextually relevant descriptions. Based on these evaluations, BLIP was selected for further fine-tuning due to its inherent capability to produce detailed and structured textual descriptions suitable for the ITTS system.

3.4.4 Fine-tuning BLIP's Image Captioning Model

The fine-tuning of the BLIP image captioning model (Li et al., 2022) was performed to produce informative, context-aware captions tailored for visually impaired users. Unified captions were generated by combining initial BLIP-generated descriptions with enriched metadata-driven descriptions, seamlessly integrating natural language context and structured clothing attributes.

A custom dataset class (`CaptionDataset`) was developed to facilitate data preprocessing. Images were retrieved from AWS S3 storage, resized uniformly to 256×256 pixels, and transformed into tensors suitable for model inputs. The initial BLIP-generated captions provided intuitive descriptions, while enriched captions included structured metadata such as clothing category, size, visibility, viewpoint, and zoom level. Both captions were tokenized using BLIP's tokenizer, applying padding and truncation to ensure uniform sequence lengths.

Fine-tuning was executed for three epochs with an AdamW optimizer (Loshchilov and Hutter, 2019), set at a learning rate of 5×10^{-5}. GPU acceleration was utilized to ensure efficient convergence (Kingma and Ba, 2017). Regularization techniques, including early stopping and weight decay, were employed to prevent overfitting. Batch processing was managed using PyTorch's DataLoader (Paszke et al., 2019b), and model performance was systematically monitored through validation phases, recording detailed loss metrics per epoch to inform training efficacy.

3.4.5 Audio Synthesis

To improve accessibility, enriched captions were converted into audio using Google Text-to-Speech (gTTS) (Taylor and Google, 2023)), enabling visually impaired users to receive spoken descriptions of clothing items. This integration provided an efficient, lightweight solution for real-time auditory feedback, enhancing the ITTS system's usability and accessibility. (gTTS) was selected for its ease of use, compatibility with Python-based web applications, and support for multiple languages. However, it has limitations in terms of voice naturalness and emotional expressiveness. More advanced models such as Whisper by OpenAI and Bark by Suno offer greater fidelity, multilingual robustness, and expressive capabilities. These models were not used in this prototype due to hardware

constraints, but are recommended for future iterations of the system where richer audio output is desired.

3.5 Performance Evaluation

The models were evaluated using four key metrics—**BLEU**, **ROUGE-L**, **METEOR**, **and CIDEr**—each designed to assess different aspects of caption quality, such as structural alignment, semantic relevance, and informativeness (Papineni et al., 2002; Lin, 2004; Banerjee and Lavie, 2005; Vedantam et al., 2015).

BLEU: This metric measures n-gram precision to evaluate lexical accuracy. BLEU-1, BLEU-2, and BLEU-3 focus on unigrams, bigrams, and trigrams. The BLEU score is calculated as follows:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right),$$

where $\text{BP} = e^{(1-r/c)}$ if $c \leq r$, else 1, accounts for brevity penalty, p_n represents n-gram precision, and w_n denotes uniform weights (Papineni et al., 2002).

ROUGE-L: This metric measures structural alignment by computing the Longest Common Subsequence (LCS) between generated and reference captions. It provides insights into the syntactic coherence of the captions (Lin, 2004).

METEOR: Designed to evaluate semantic relevance, METEOR incorporates word alignment, synonyms, and paraphrases. This approach goes beyond exact lexical matches, complementing BLEU (Banerjee and Lavie, 2005).

CIDEr: This metric assesses the informativeness of captions by comparing tf-idf-weighted n-grams with reference captions. CIDEr emphasizes the importance of generating contextually rich and specific descriptions (Vedantam et al., 2015).

Chapter 4: System Design and Architecture

This chapter describes the system design and architecture for the AI-powered ITTS system. The system is designed to assist visually impaired individuals in accessing clothing descriptions through Multimodal gTTS (Taylor and Google, 2023). The architecture comprises key components, including image processing, caption generation, and TTS synthesis using gTTS.

4.1 System Overview

The ITTS system is structured into three primary components:

1. **Image Processing and Captioning Module**: Extracts features from clothing images and generates descriptive captions using transformer-based models.
2. **TTS Module**: Converts the generated captions into natural-sounding speech using gTTS.
3. **Web-Based Deployment Interface**: Provides an accessible interface for users to upload images and receive auditory feedback.

4.2 System Modeling Framework

The system follows a modular modeling framework to ensure efficient design, implementation, and scalability. The ITTS system's modeling framework consists of:

1. **Functional Modeling**: Defines the functionalities of each component, including input processing, image-to-text transformation, and text-to-speech conversion.
2. **Data Flow Modeling**: Represents data flow from image input through processing and caption generation to the final speech synthesis output.
3. **Process Flow Diagrams**: Illustrate the interactions between system components and how data is processed through different modules.
4. **Use Case Modeling**: Captures user interactions with the system, such as uploading images, receiving captions, and listening to generated audio.

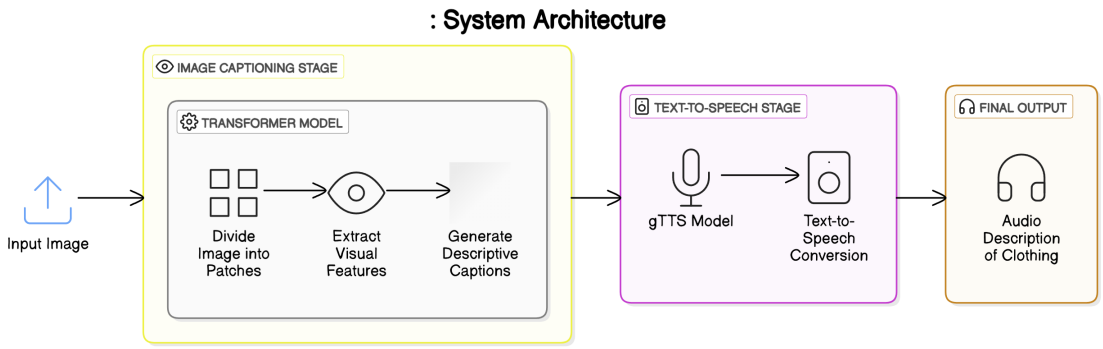


Figure 4.1: System architecture of the ITTS assistive solution, from image input to speech output.

Table 4.1: Description of system components in ITTS pipeline.

Component	**Description**
Input Image	The image of the clothing item uploaded by the user.
Transformer Model	Processes the image by dividing it into patches and extracting visual features.
BLIP Model	generates descriptive captions from extracted features.
gTTS	converts the generated caption into an audio description.
Final Output	The system delivers the generated audio description to the user.

4.3 System Components

4.3.1 Image Processing and Captioning Module

This module extracts features from clothing images and generates descriptive captions. It begins with a preprocessing pipeline that performs image resizing, normalization, and augmentation to ensure consistency and robustness in the input data. Once preprocessed, the images are passed through a transformer-based model, which encodes the visual data and generates context-aware textual descriptions. These captions capture details such as color, texture, and style, particularly relevant to clothing.

4.3.2 Text-to-Speech (TTS) Module

The TTS module converts the generated captions into spoken audio using the Google Text-to-Speech (gTTS) library (Taylor and Google, 2023). It takes the caption text produced by the image processing module and synthesizes it into an audio file, which is then played automatically for the user. The system supports multiple languages—English, Kiswahili, and French—allowing users to listen to the descriptions in a most comfortable language. This multilingual capability enhances accessibility, especially for users in diverse linguistic contexts. Using gTTS (Taylor and Google, 2023) ensures natural-sounding speech output, improving the overall user experience.

4.3.3 Web-Based Deployment Interface

The ITTS web application was deployed on Render, a modern cloud platform known for its ease of use and support for full-stack deployment. Render was chosen for its ability to integrate seamlessly with GitHub and support automatic builds and deployments. The backend, built using FastApplication Programming Interface (API) (Ramírez, 2023), handles image uploads and caption-to-speech processing, while the frontend uses HyperText Markup Language (HTML), Cascading Style Sheets (CSS), and JavaScript to provide an accessible, responsive user experience.

The interface features a design with a 5MB image upload limit, an icon-based upload section, and a loading spinner to keep users informed during processing. Users can listen to audio descriptions in English and optionally translate and play them in Kiswahili, enhancing accessibility for multilingual users. The application supports desktop and mobile use and is accessible via a secure HyperText Transfer Protocol Secure (HTTPS) Uniform Resource Locator (URL).

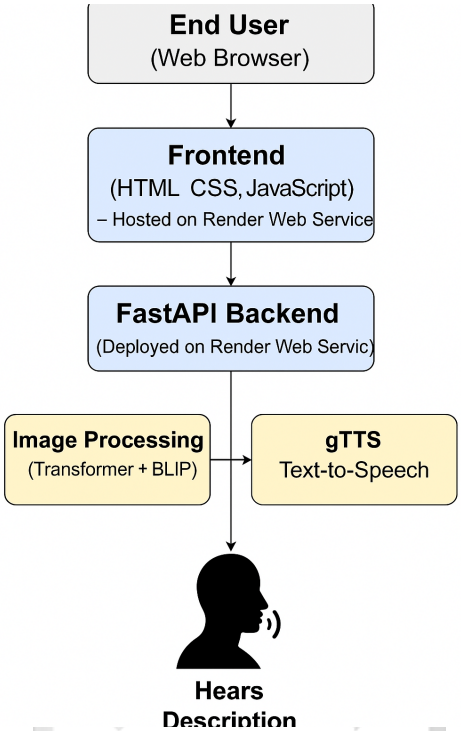


Figure 4.2: Deployment architecture of the ITTS system on Render cloud.

Chapter 5: System Implementation and Testing

5.1 Introduction

This section presents the implementation and testing of the AI-powered ITTS system. It outlines the core technologies used, key functionalities, and deployment considerations. Testing focused on evaluating usability, accessibility, and functional performance of the system.

5.2 System Implementation

The ITTS system was implemented as a browser-based prototype that integrates the BLIP model for caption generation,(gTTS) (Taylor and Google, 2023) for audio synthesis, and FastAPI (Ramírez, 2023) for backend services. Users can upload an image of clothing, receive a caption, and listen to an audio description, optionally translated into Kiswahili or French. The system is structured around three core modules. First, the caption generation module uses the BLIP model to extract visual features from the uploaded image and produce an accurate and context-aware English description. Second, the speech synthesis module converts this text into spoken audio using gTTS (Taylor and Google, 2023), while additional browser-based TTS engines handle multilingual translations. Finally, the web interface facilitates user interaction by providing a responsive, accessible, and intuitive environment. This front-facing layer ensures users can easily upload images, view the generated captions, and access audio feedback through a streamlined and minimalistic design.

5.2.1 Overall UI Design

The UI was designed to support accessibility and demonstrate assistive functionality. It features segmented sections including How It Works, Assistive Feature Simulation, and Why It Matters. The layout prioritizes readability, icon-based controls, and real-time caption previews. TailwindCSS was used to streamline styling.

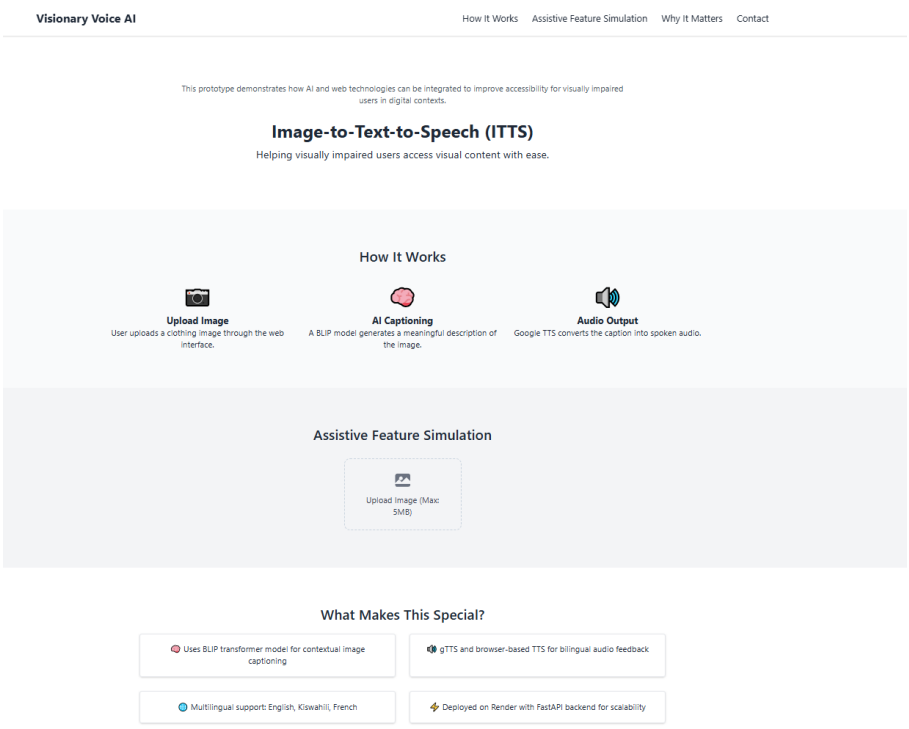


Figure 5.1: Overview of the ITTS interface layout, illustrating the structured arrangement of core sections designed for accessibility.

5.3 System Functionalities

5.3.1 Image Upload and Processing

Users upload images in Joint Photographic Experts Group (JPG), Portable Network Graphics (PNG), or Web Picture format (WEBP) format (max 5MB). Images are resized and normalized for input to the captioning model.

5.3.2 Caption Generation

The BLIP model generates captions in English that describe visual features such as clothing type, color, and design, given its context, as the model was selected for its contextual understanding and performance.

5.3.3 Text-to-Speech Conversion

Captions are converted to speech using gTTS (Taylor and Google, 2023), which autoplays in the browser. Users may translate the caption into Kiswahili or French via a dropdown and listen to the spoken translation using the browser's native TTS engine.

5.3.4 Web-Based User Interaction

The assistive feature simulation section enables real-time interaction, with visual feedback, audio playback, and language switching. Accessibility features like a loading spinner, dropdown language selection, and responsive layout are integrated.

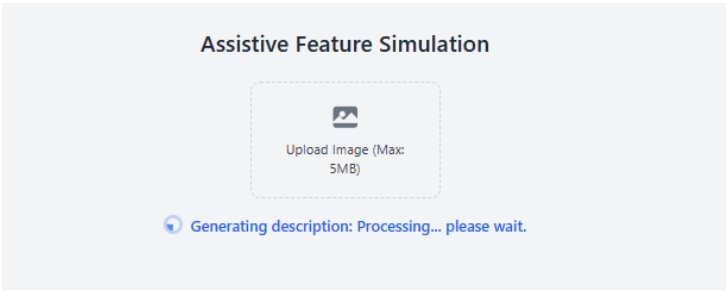


Figure 5.2: Loading indicator and user feedback during caption generation.

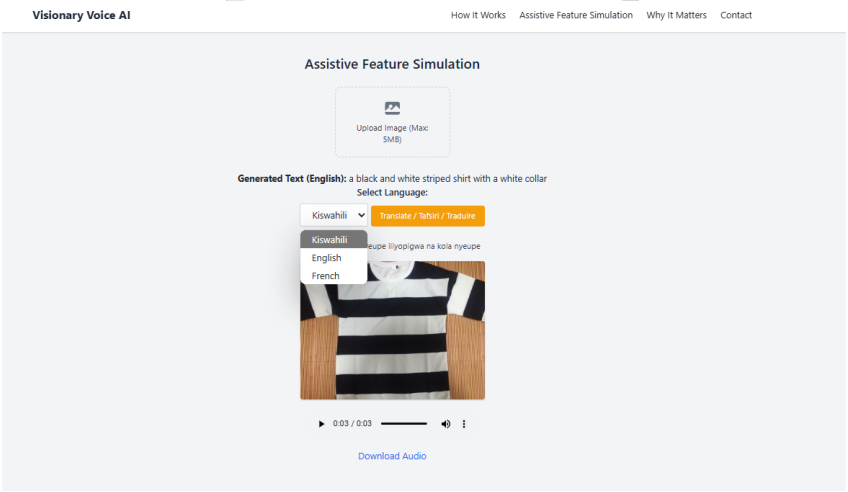


Figure 5.3: Assistive Feature Simulation section of the web interface showing image upload, caption preview, and audio playback.

5.4 System Testing

To ensure the reliability and practical usability of the ITTS system, a range of targeted tests were conducted. These focused on validating each functional module, assessing user experience, confirming cross-platform compatibility, and evaluating security protocols. While formal validation with visually impaired users is proposed as future work, the current prototype underwent structured testing to verify that the core components perform as expected.

5.4.1 Functional Testing

Functional testing focused on verifying that each integrated module performed according to specification. This included testing the image upload pipeline for supported formats and size constraints, validating that the BLIP model consistently generated captions relevant to the visual input, and ensuring that the generated audio files were correctly synthesized and auto-played in the browser. The language translation dropdown and playback of translated captions were also verified for responsiveness and completeness.

5.4.2 Usability Testing

Usability was assessed through informal walkthroughs and observation-based feedback sessions. These tests evaluated whether users could complete tasks such as uploading images, interpreting captions, and navigating between English and translated audio without confusion. Particular attention was paid to interaction clarity, error handling (e.g., oversized image uploads), and real-time feedback mechanisms like the loading spinner. The icon-based design and minimalistic layout were well-received for their intuitive guidance, particularly in simulating assistive technology use.

5.4.3 Compatibility Testing

The application was tested on major web browsers, including Google Chrome, Mozilla Firefox, and Safari, to ensure consistent performance across environments. Tests were conducted on desktop and mobile devices to evaluate responsiveness and layout adaptability. Elements such as audio playback, dropdown selection, and image rendering were observed to behave consistently across platforms, with no significant deviation in performance or appearance.

5.4.4 Security Testing

Security testing focused on input validation and safe data handling. Uploaded files were restricted to accepted image types and constrained to a 5MB limit to reduce attack surface. Additionally, backend logic ensured temporary storage of image and audio data without long-term retention. API endpoints were protected from malformed inputs, and error messages were sanitized to prevent leakage of implementation details. These

measures contributed to a secure runtime environment aligned with best practices for prototype-stage applications.

Chapter 6: Discussion of Results

This chapter reviews the outcomes achieved across the different stages of the CRISP-DM framework, from data understanding to model deployment. This segment presents the insights gathered from exploring and analyzing the dataset, highlighting the patterns and trends identified during data pre-processing and modeling. Moreover, it showcases the performance metrics and effectiveness of the developed models in achieving the research objectives highlighted in Chapter 1, Section 1.3.

6.1 Data Understanding

The initial Exploratory Data Analysis (EDA) of the DeepFashion2 dataset highlighted critical insights regarding the original distribution of clothing categories (see Figure 6.1). Before pre-processing, the dataset showed notable imbalances. Categories such as short sleeve dresses, vest dresses, and short sleeve tops dominated, with each exceeding 1000 occurrences. Conversely, slings and short sleeve outerwear categories were significantly underrepresented, each with fewer than 50 occurrences.

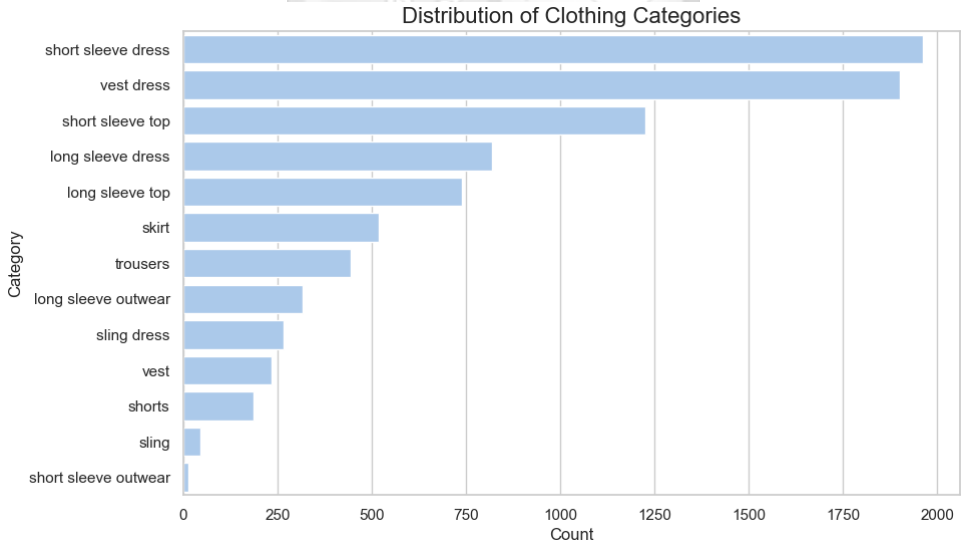


Figure 6.1: Clothing category distribution before stratified sampling.

To mitigate the risk of model bias towards these prevalent categories, stratified sampling was applied, resulting in a balanced dataset of 100 images per category for 11 selected categories, totalling 1,100 images.

Further EDA after stratification identified additional key characteristics. Source distri-

bution analysis (Figure 6.2) revealed a higher proportion of user-uploaded images than shop-sourced images, indicating potential variability in quality and styling.

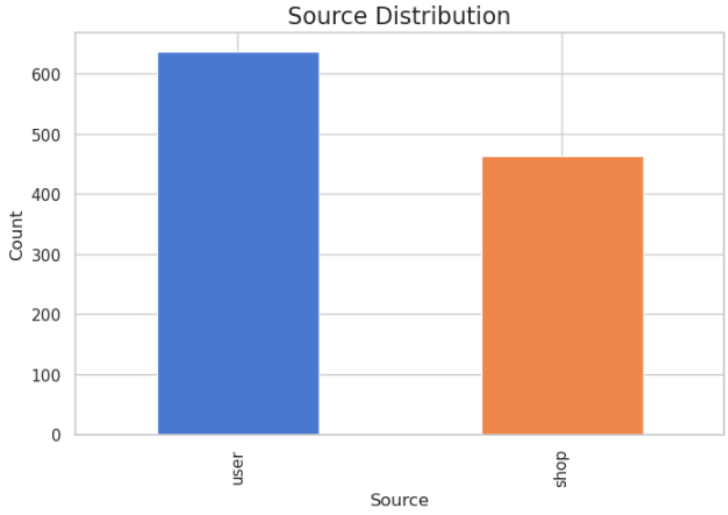


Figure 6.2: Distribution of image sources in the curated dataset.

The category-source distribution analysis further highlighted distinct differences between categories sourced from users and shops, with shop-sourced images demonstrating greater consistency (Figure 6.3).

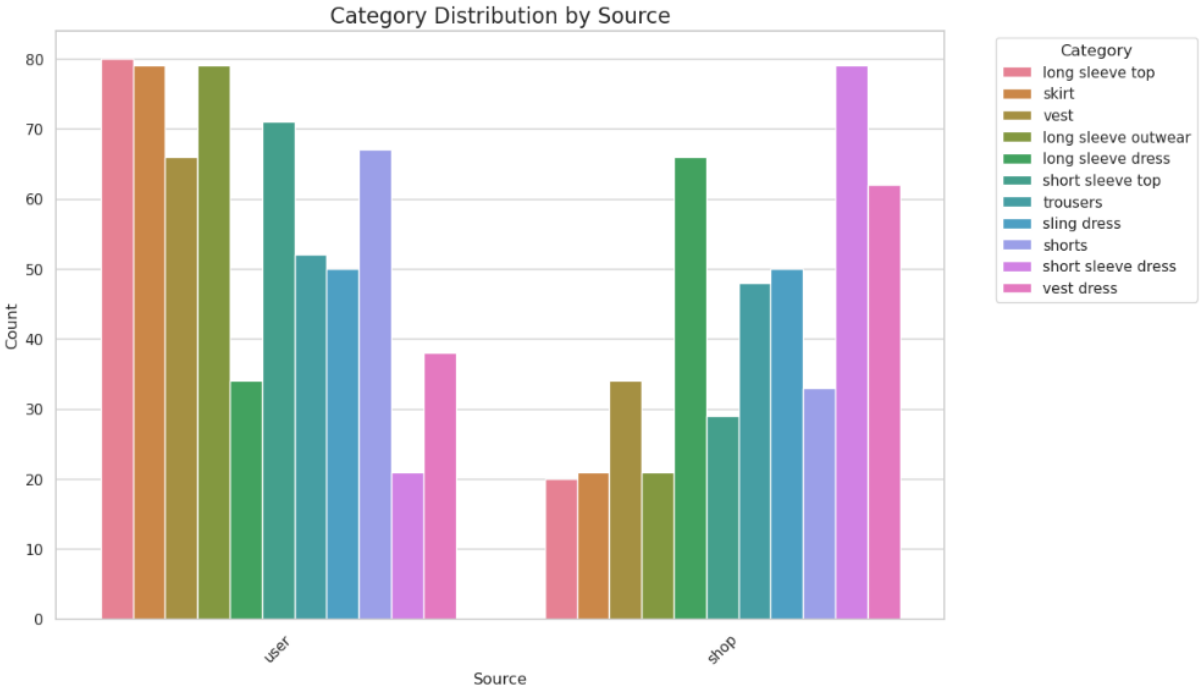


Figure 6.3: Clothing category distribution segmented by source type.

6.1.1 Bias in Clothing Category Distribution

The dataset analysis revealed a significant imbalance in the distribution of clothing categories. As shown in (Figure 6.1), categories such as short sleeve dresses, vests, and tops dominate the dataset, while items like trousers, shorts, and outerwear are notably underrepresented. This skew toward traditionally feminine clothing highlights a gender bias in the dataset, which may affect the model's ability to generalize across different user groups, particularly male or gender-neutral clothing scenarios. It also limits the captioning system's diversity in describing a wide range of garment types. Future development should consider more balanced or representative datasets to support equitable performance and broader applicability in real-world assistive contexts.

6.2 Data Preparation

The data preparation phase involved critical pre-processing steps to ensure compatibility with the machine learning models. Essential activities included uniformly resizing all images to 256×256 pixels, converting them to RGB format, and normalizing pixel values according to ImageNet standards. This standardization was crucial to maintaining consistency and enhancing model performance.

Metadata enrichment significantly enhanced caption quality. Attributes such as clothing scale, visibility, viewpoint, and zoom level were systematically extracted from annotations and integrated into the captioning framework. Examples of enriched metadata integration is provided below:

Figure 6.4: Metadata-enriched caption derived from structured attributes: example 1.

Figure 6.5: Metadata-enriched caption derived from structured attributes: example 2.

Figures 6.4 and 6.5 illustrate two distinct examples of captions generated through metadata enrichment, showcasing the integration of structured descriptors such as scale, viewpoint, and visibility.

6.3 Machine Learning Modeling

The ITTS system was developed using two machine learning models: ViT for image classification and BLIP for image captioning.

6.3.1 ViT Model Results

The ViT model (Dosovitskiy et al., 2021) was evaluated in a zero-shot setting, meaning it was used without additional fine-tuning on the DeepFashion2 dataset. The model performed well in classifying general clothing types but could not generate detailed textual descriptions. The results of its classification performance are illustrated in Figure 6.6.

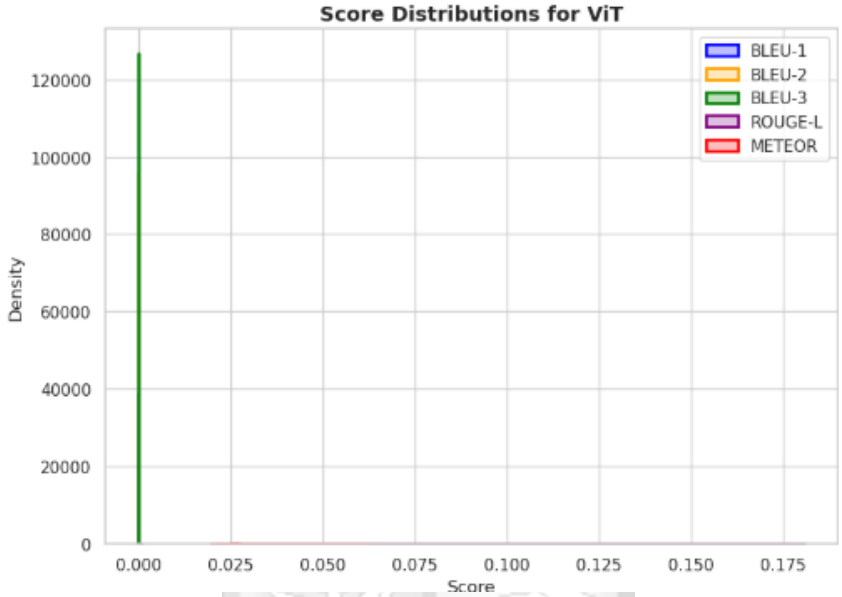


Figure 6.6: ViT model performance score distribution (zero-shot).

The distribution of ViT scores indicates a strong performance in basic classification tasks, as evident from the peak density in the graph. However, the near-zero values in BLEU and METEOR scores confirm that ViT lacks descriptive ability beyond categorical classification. This result aligns with the model's design, which excels at identifying object categories but does not generate contextual descriptions. The inability to produce textual descriptions limits its effectiveness for accessibility applications requiring rich, descriptive feedback.

Several factors may have contributed to ViT's underperformance in generating descriptive captions. Firstly, the model was evaluated in a zero-shot setting without fine-tuning on domain-specific clothing data, which likely limited its ability to generalize beyond generic classification. Secondly, ViT was originally designed for image classification tasks and lacks an inherent language modeling component. Unlike multimodal architectures such as BLIP, it does not align image features with textual outputs, which are essential for

rich captioning. This reinforces the importance of model-task alignment in accessibility-oriented applications.

6.3.2 BLIP Model Results

The BLIP model demonstrated superior performance in generating rich, detailed descriptions of clothing items. The model was tested in a zero-shot setting before being fine-tuned with metadata-enriched captions. After fine-tuning, the BLEU and METEOR scores substantially improved caption quality. Figure 6.7 illustrates the score distributions for BLIP's performance.

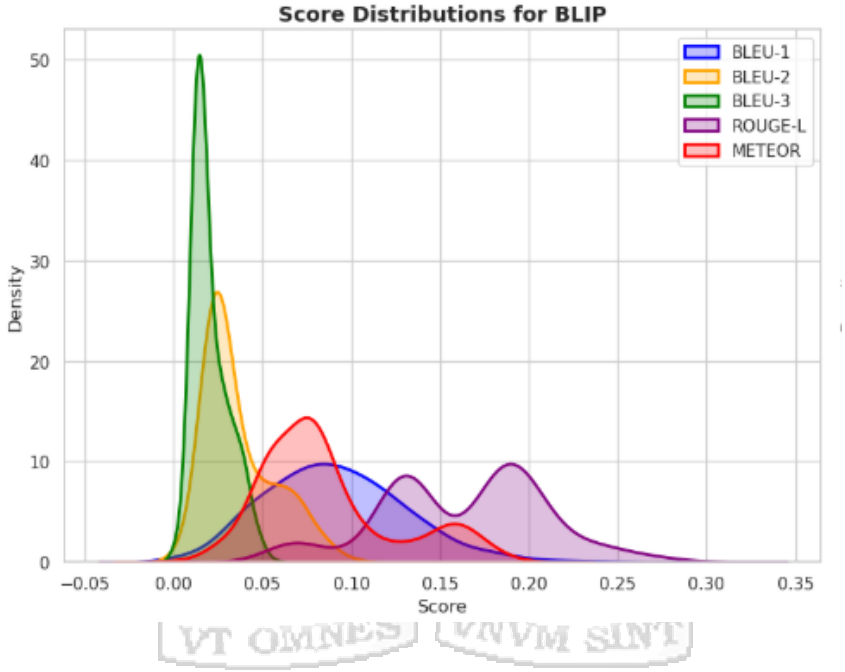


Figure 6.7: BLIP captioning score distribution (zero-shot setting).

The BLIP model's score distribution reveals a higher variance, indicating its ability to generate more nuanced and contextually relevant captions. Unlike ViT, BLIP could leverage metadata-enriched training to produce detailed descriptions of clothing attributes, styles, and conditions. The increased BLEU and ROUGE-L scores demonstrate the model's ability to align its output with reference captions effectively. This suggests that BLIP's multimodal learning approach enables it to interpret visual and textual data efficiently, making it highly suitable for accessibility-oriented applications.

6.3.3 Zero-Shot Model Comparison: ViT vs. BLIP

A direct comparison of ViT and BLIP in their zero-shot states was conducted using BLEU, ROUGE-L, METEOR, and CIDEr evaluation metrics. The results in Figure 6.8 clearly indicate that BLIP significantly outperformed ViT in all measured aspects.

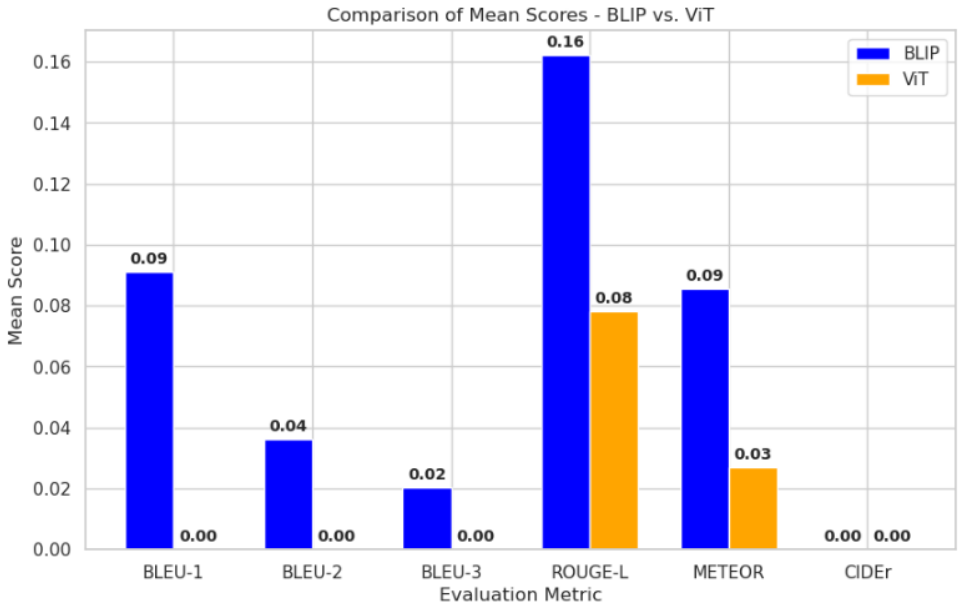


Figure 6.8: Mean evaluation scores: BLIP vs. ViT (zero-shot).

The BLEU-1, BLEU-2, and BLEU-3 scores highlighted BLIP's superior ability to generate captions that closely resembled reference descriptions, while ViT scores remained at near-zero levels, reflecting its categorical rather than descriptive nature. ROUGE-L, which measured textual overlap, confirmed BLIP's ability to construct semantically and syntactically coherent captions. The CIDEr score, which evaluated informativeness, remained low for both models, suggesting room for improvement in generating highly informative descriptions beyond basic attributes. These findings justified the selection of BLIP for fine-tuning to enhance its captioning performance.

6.4 Model Optimization Results: Fine-Tuned BLIP

Following selecting BLIP as the optimal model for generating image descriptions, fine-tuning was conducted to enhance its captioning capabilities. The fine-tuning process incorporated structured metadata-enriched captions, refining the model's ability to generate more precise and contextually relevant descriptions.

6.4.1 Evaluation Metrics Performance

The performance of the fine-tuned BLIP model across key evaluation metrics is presented in Figure 6.9.

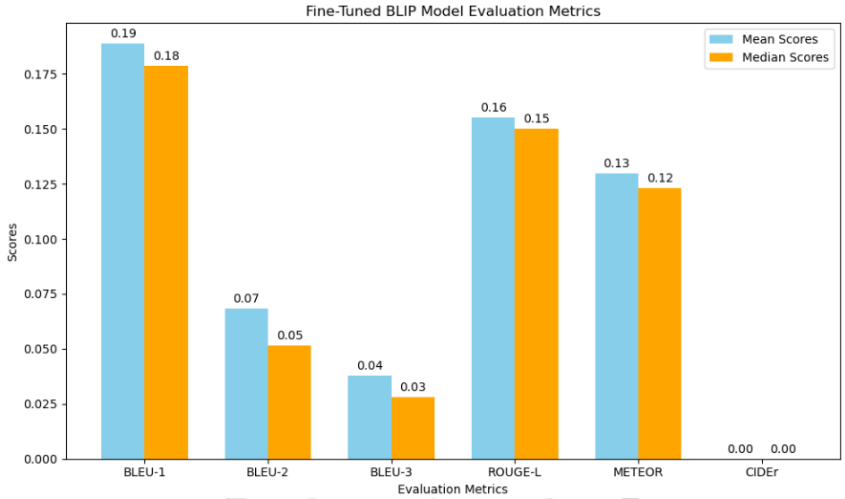


Figure 6.9: Evaluation metrics for the fine-tuned BLIP model.

The results indicate substantial improvements across BLEU, ROUGE-L, and METEOR scores compared to the zero-shot BLIP model. Table 6.1 summarizes the performance improvements.

Table 6.1: Performance comparison: zero-shot vs. fine-tuned BLIP model.

Evaluation Metric	**Zero-Shot BLIP**	**Fine-Tuned BLIP**
BLEU-1	0.09	0.19
BLEU-2	0.04	0.07
BLEU-3	0.02	0.04
ROUGE-L	0.16	0.16
METEOR	0.09	0.13
CIDEr	0.00	0.00

The following insights can be drawn from the results:

- **BLEU Scores:** The BLEU-1 score increased from 0.09 to 0.19, signifying that the fine-tuned model produced captions with better word alignment to reference

descriptions. Additionally, BLEU-2 and BLEU-3 scores showed improvement, indicating enhanced coherence in multi-word sequences.

- **ROUGE-L:** The stability of the ROUGE-L score at 0.16 suggests that the fine-tuned model maintained a high level of textual overlap with reference captions, ensuring consistency in generated descriptions.
- **METEOR:** The increase in METEOR from 0.09 to 0.13 highlights improvements in semantic accuracy, reflecting better word choice and sentence structure in generated captions.
- **CIDEr:** The CIDEr score remained at 0.00, indicating that while fine-tuning improved sentence-level metrics, the model still faced challenges producing highly informative and detailed captions beyond essential attributes.
- **CIDEr:** The CIDEr score remained at 0.00, indicating that while fine-tuning improved sentence-level metrics, the model still faced challenges producing highly informative and detailed captions beyond essential attributes. This outcome is largely attributed to the DeepFashion2 dataset providing only a single reference caption per image. Since CIDEr relies on computing tf-idf-based similarity across multiple reference captions, it cannot effectively evaluate semantic consensus in this setting. This limitation is acknowledged and suggests the need for either alternative metrics or datasets with richer reference annotations in future evaluations.

These findings show the impact of fine-tuning in enabling BLIP to leverage structured metadata effectively. The improvements in BLEU and METEOR suggest that the model became better at generating natural-sounding and relevant descriptions. However, the lack of progress in CIDEr highlights the need for further refinements, such as domain-specific training data or reinforcement learning strategies, to enhance the informativeness of captions.

6.4.2 Distribution of Evaluation Scores

The distribution of evaluation scores across the dataset is visualized in Figure 6.10 to analyze further the fine-tuned BLIP model's performance.

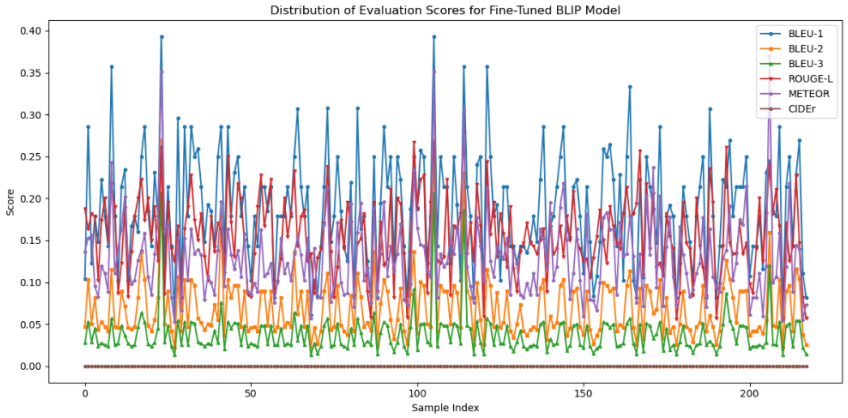


Figure 6.10: Score distribution across evaluation metrics (fine-tuned BLIP).

The distribution illustrates a clear upward trend in BLEU and METEOR scores, confirming that fine-tuning contributed to more consistent and accurate caption generation. Compared to the zero-shot model, there is an observable reduction in score variance, suggesting improved model stability.

The spikes in BLEU-1 and METEOR indicate that while most captions improved accuracy, some still showed variability. This could be attributed to the dataset's complex clothing styles, occlusions, or ambiguous viewpoints.

Overall, these results confirm that fine-tuning BLIP with enriched metadata significantly improved captioning performance by enhancing word alignment, semantic accuracy, and coherence, making it a more effective model for assisting visually impaired individuals in understanding clothing attributes.

6.5 Audio Synthesis

To enhance accessibility, the enriched captions were converted into audio descriptions using gTTS (Taylor and Google, 2023), a Python library for text-to-speech synthesis. For instance, an enriched caption such as *"The clothing item is a large-sized jacket, fully visible, viewed from the front, and shown in a close-up view"* was passed to the gTTS library (Taylor and Google, 2023). This process generated an audio file in MP3 format that could be played back to provide visually impaired users with descriptive information about the clothing items.

An example is shown in Figure 6.11, where a sling dress is described as *"The clothing item*

is a large-sized short sleeve top, partially visible, viewed from the front, and shown in a medium close-up". The audio file was generated successfully, demonstrating the seamless integration of image captioning and audio synthesis.

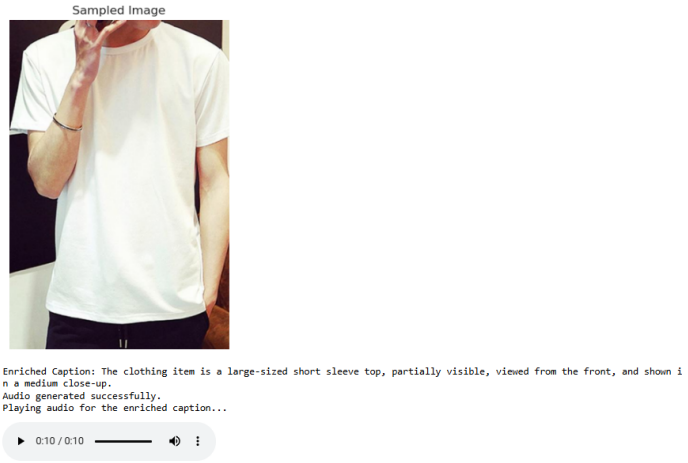


Figure 6.11: Example of enriched caption and audio synthesis for a sample image.

This integration highlights the system's potential to enhance accessibility by delivering rich, descriptive audio outputs for visually impaired users.

6.6 Summary of Results

The fine-tuned BLIP model demonstrated significant improvements over its zero-shot counterpart, particularly in BLEU-1 and METEOR scores, which increased from 0.09 to 0.19 and 0.09 to 0.13, respectively. These enhancements indicate improved lexical alignment and semantic accuracy in generated captions. ROUGE-L remained stable at 0.16, confirming strong textual coherence, while CIDEr remained at 0.00, suggesting limitations in generating highly informative captions.

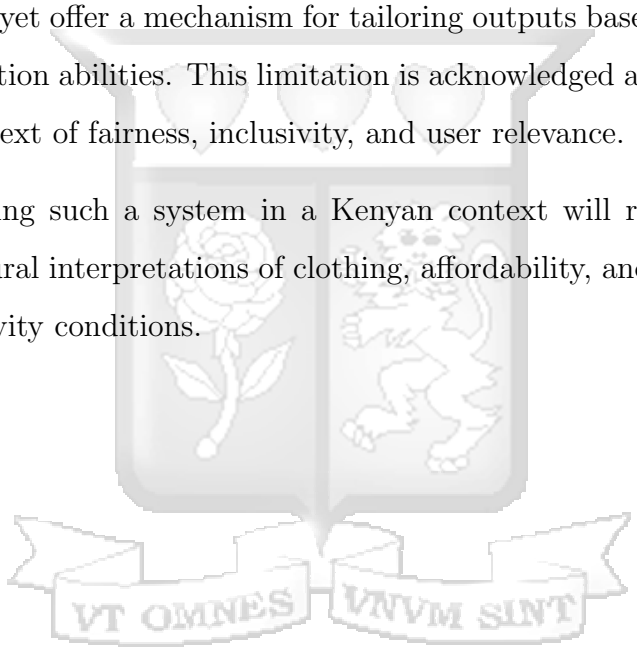
Comparative analysis reinforced BLIP's superiority over ViT for descriptive tasks, validating its selection for the ITTS system. Score distributions further highlighted increased consistency post-fine-tuning, demonstrating the effectiveness of structured metadata integration. These findings underscore the value of multimodal AI in accessibility solutions and provide a foundation for future enhancements in captioning accuracy and informativeness.

6.7 Ethical Considerations

This study acknowledges that visually impaired individuals represent a vulnerable population, and any system designed for their support must be grounded in ethical responsibility. The current prototype is a web-based tool that enables users to manually upload clothing images and receive audio descriptions. No personal data is collected, and there is no user profiling or customization at this stage.

However, the system's reliance on pretrained models means that all generated captions including color descriptors are presented without filtering. This presents a limitation for users who are congenitally blind and may have no conceptual understanding of color. The system does not yet offer a mechanism for tailoring outputs based on individual user preferences or perception abilities. This limitation is acknowledged as an ethical concern, especially in the context of fairness, inclusivity, and user relevance.

Furthermore, deploying such a system in a Kenyan context will require sensitivity to local languages, cultural interpretations of clothing, affordability, and accessibility across devices and connectivity conditions.



Chapter 7: Conclusions, Recommendations, and Future Work

7.1 Conclusions

This study successfully developed a Multimodal ITTS system tailored to improve clothing accessibility for visually impaired individuals. By leveraging the DeepFashion2 dataset, the BLIP model for image captioning, and Google Text-to-Speech (gTTS) for audio synthesis, the project demonstrated how ML can bridge the gap between visual content and auditory feedback. The system showed significant improvements in key evaluation metrics (BLEU-1 to BLEU-3, METEOR), particularly after fine-tuning the captioning model with metadata-enriched descriptions. Despite a flat CIDEr score, the enriched output offered contextually relevant captions, confirming the model's suitability for assistive applications. This research contributes to the growing field of inclusive AI solutions. It affirms the potential of multimodal systems in enhancing digital accessibility, especially in fashion, a domain traditionally inaccessible to the visually impaired. .

7.2 Recommendations

Future iterations of the system could benefit from captions that go beyond basic descriptions to include texture, functionality (e.g., casual vs formal), or recommended pairings. Increasing the diversity and specificity of the training dataset would help address underrepresented styles and garments. Real-world testing is essential to better align system outputs with the practical needs of visually impaired users. In particular, collecting direct user feedback could guide improvements to audio pacing, language tone, and interactivity. To support wider deployment, optimizing system efficiency for real-time use is recommended. Techniques such as model pruning, caching, or leveraging serverless architecture could reduce resource overhead while maintaining performance. Lastly, supporting user-selectable language or voice preferences would make the system more inclusive.

7.3 Future Work

While promising, this prototype remains an early-stage tool. Future work should prioritize piloting the system with a small group of visually impaired users to validate usability and gather actionable insights. Although expanding the language set is beneficial, focus-

ing first on refining support for English, Kiswahili, and French ensures that quality is not compromised for quantity. Real-time interactivity, such as asking for more details or clarifying outputs, could be explored, though this would require more sophisticated natural language processing capabilities. Finally, scaling the system for broader deployment would require addressing performance limitations, managing privacy in image handling, and ensuring robust language models suited for assistive applications. These steps, while ambitious, can be approached incrementally with clear focus on user-centered design and ethical deployment.

Bibliography

A, R., V, N., Jebadurai, I. J., Vedamanickam, A. M., and Kumar, P. U. (2023). Design of generative multimodal ai agents to enable persons with learning disability. In *Companion Publication of the 25th International Conference on Multimodal Interaction*, ICMI '23 Companion, page 259–271, New York, NY, USA. Association for Computing Machinery.

Ahsan, H., Bhatt, D., Shah, K., and Bhalla, N. (2021). Multi-modal image captioning for the visually impaired. In Durmus, E., Gupta, V., Liu, N., Peng, N., and Su, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 53–60, Online. Association for Computational Linguistics.

Amazon Web Services (2023). Amazon S3: Scalable Storage in the Cloud. Accessed: March 2025.

Armstrong, J. L. (2015). Seeing fashion through sound. Master's thesis, Ryerson University.

Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Goldstein, J., Lavie, A., Lin, C.-Y., and Voss, C., editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Bastawrous, A., Mathenge, W., Wing, K., Rono, H., Gichangi, M., Weiss, H. A., Macleod, D., Foster, A., Burton, M. J., and Kuper, H. (2016). Six-Year Incidence of Blindness and Visual Impairment in Kenya: The Nakuru Eye Disease Cohort Study. *Investigative Ophthalmology and Visual Science*, 57(14):5974–5983.

Burton, M. A. (2011). Fashion for the blind: a study of perspectives. In *The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '11, page 315–316, New York, NY, USA. Association for Computing Machinery.

Clark, A. (2015). Pillow (pil fork) documentation.

De Marsico, M., Giacanelli, C., Manganaro, C. G., Palma, A., and Santoro, D. (2024). Vqask: a multimodal android gpt-based application to help blind users visualize pictures. In *Proceedings of the 2024 International Conference on Advanced Visual Interfaces*, AVI '24, pages 2–5, New York, NY, USA. Association for Computing Machinery.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale.

Eziamaka, N. V., Odonkor, T. N., and Akinsulire, A. A. (2024). Ai-driven accessibility: Transformative software solutions for empowering individuals with disabilities. *International Journal of Applied Research in Social Sciences*.

Fernando, S., Ndukwe, C., Virdee, B., and Djemai, R. (2025). Image recognition tools for blind and visually impaired users: An emphasis on the design considerations. *ACM Trans. Access. Comput.*, 18(1).

Ge, Y., Zhang, R., Wu, L., Wang, X., Tang, X., and Luo, P. (2019). A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. *CVPR*.

Government of Kenya (2007). Kenya vision 2030.

Harb, B. and Sidani, D. (2021). Smart technologies challenges and issues in social inclusion – case of disabled youth in a developing country. *Journal of Asia Business Studies*.

Heaton, J. (2018). Ian goodfellow, yoshua bengio, and aaron courville: Deep learning. *Genetic Programming and Evolvable Machines*, 19(1):305–307.

Himayah, H. and Hasan, H. A. (2022). Digital inclusion for the faculty members: A case study. *Khizanah al-Hikmah : Jurnal Ilmu Perpustakaan, Informasi, dan Kearsipan*.

Jagadish, S., Lalitendra, M., Nikhita, K., Narsingarao, P., and Sreedhar, L. (2024). Artificial intelligence-powered mobile application to help visually impaired people. *International Journal of Innovative Science and Research Technology (IJISRT)*, 9(4):1100–1105.

Joisten, M., Zeng, L., Woletz, J. D., Brock, A. M., and Avila, M. (2015). Accessible interaction for visually impaired people. In *Universal Access in Human-Computer Interaction. Access to the Human Environment and Culture*, pages 379–381. De Gruyter.

Kalchbrenner, N., Espeholt, L., Simonyan, K., van den Oord, A., Graves, A., and Kavukcuoglu, K. (2017). Neural machine translation in linear time.

Khalid, L. and Gong, W. (2022). Vision4all — a deep learning fashion assistance solution for blinds. In *2022 5th International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pages 156–161.

Kingma, D. P. and Ba, J. (2017). Adam: A method for stochastic optimization.

Kumar, N., Verma, K., and Ahmad, E. (2024). Artificial intelligence based virtual assistant for vision impaired person. *International Journal for Research in Applied Science and Engineering Technology*, 12(2):1293–1298.

Lamchoudi, S., Miftah, R., and Mourhir, A. (2024). Clothing type and color classification using tinyml. In *2024 International Conference on Intelligent Systems and Computer Vision (ISCV)*, pages 1–6.

Li, J., Li, D., Xiong, C., and Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. (2015). Microsoft coco: Common objects in context.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization.

Marcel, S. and Rodriguez, Y. (2010). Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 1485–1488, New York, NY, USA. Association for Computing Machinery.

McKinney, W. et al. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX.

Michele A. Williams, C. N. and Hurst, A. (2013). Preliminary investigation of the limitations fashion presents to those with vision impairments. *Fashion Practice*, 5(1):81–105.

Mohanraj, P., Rajasekar, T., Sivaelango, N., Sri Karthickraja, V., and Vignesh, N. (2024). Wearable device for visually impaired using deep learning. In *2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, pages 348–352.

Mokady, R. et al. (2021). Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Nyamweya, W. K., Muchelule, Y., and Mwalili, T. (2024). Digital inclusion practices and technology accessibility performance for people with disabilities in tertiary education institutions in nairobi county. *International Academic Journal of Economics and Finance*, 4(1):473–523.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In Isabelle, P., Charniak, E., and Lin, D., editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019a). Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M.,

Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019b). Pytorch: An imperative style, high-performance deep learning library.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision.

Raghavan, Rohith, Krishnan, Vishodhan, Nishad, Hitesh, and Shaikh, Bushra (2021). Virtual ai assistant for person with partial vision impairment. *ITM Web Conf.*, 37:01019.

Ramírez, S. (2023). Fastapi. `https://fastapi.tiangolo.com/`. Web framework for building APIs with Python.

RuchaReads (2021). Crisp dm framework. `https://ruchareads.wordpress.com/2021/03/29/1-crisp-dm-framework/`. [Accessed 05-04-2024].

Srinivasan, R. and San Miguel González, B. (2022). The role of empathy for artificial intelligence accountability. *Journal of Responsible Technology*, 9:100021.

Suomi, R. and Sachdeva, N. (2016). Internet accessibility for visually impaired. In Mesquita, A., editor, *Handbook of Research on Human Social Interaction in the Age of Mobile Devices*, pages 250–259. IGI Global.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks.

Takshara, K. S. and Bhuvaneswari, G. (2025). Empowering visually impaired individuals: The transformative roles of education, technology, and social connections in fostering resilience and well-being. *British Journal of Visual Impairment*.

Taylor, P. and Google (2023). gtts: Google text-to-speech python library. `https://pypi.org/project/gTTS/`. Available at `https://pypi.org/project/gTTS/`.

Tebbutt, E., Brodmann, R., Borg, J., MacLachlan, M., Khasnabis, C., and Horvath, R. (2016). Assistive products and the sustainable development goals (sdgs). *Globalization and Health*, 12(1):79.

The International Agency for the Prevention of Blindness (2020). Vision loss in kenya 2020.

United Nations (2015). The 2030 agenda for sustainable development.

Van Rossum, G. and Drake Jr, F. L. (1995). *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2023). Attention is all you need.

Vedantam, R., Zitnick, C. L., and Parikh, D. (2015). Cider: Consensus-based image description evaluation.

Wambua, J. (2021). Inclusive ict and disability. `https://www.enableme.ke/en/article/inclusive-ict-and-disability-399`. Accessed: 2024-09-05.

Whitfield, R., Schwab, L., Ross-Degnan, D., Steinkuller, P., and Swartwood, J. (1990). Blindness and eye disease in kenya: ocular status survey results from the kenya rural blindness prevention project. *British Journal of Ophthalmology*, 74(6):333–340.

Withana, A. (2023). Co-designing personalized assistive devices using personal fabrication. *Commun. ACM*, 66(7):89–90.

World Health Organization (2020). *World Report on Vision*. World Health Organization, Geneva.

World Wide Web Consortium (W3C) (2018). *Web Content Accessibility Guidelines (WCAG) 2.1*.

Xie, J., Yu, R., Zhang, H., Lee, S., Billah, S. M., and Carroll, J. M. (2024). Emerging practices for large multimodal model (lmm) assistance for people with visual impairments: Implications for design. *ArXiv*, abs/2407.08882.

Xu, J. et al. (2021). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*.

Yang, S., Xiao, W., Zhang, M., Guo, S., Zhao, J., and Shen, F. (2023). Image data augmentation for deep learning: A survey.

Appendices

Appendix A: Similarity Report

BessyMukaria_169111_Dissertation_Final_V01.pdf

ORIGINALITY REPORT

12%	11%	10%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	de.overleaf.com Internet Source	1%
2	Submitted to Strathmore University Student Paper	1%
3	digital.lib.washington.edu Internet Source	1%
4	arxiv.org Internet Source	1%
5	proceedings.neurips.cc Internet Source	1%
6	dspace.cvut.cz Internet Source	<1%
7	su-plus.strathmore.edu Internet Source	<1%
8	www.iajournals.org Internet Source	<1%
9	Submitted to The University of Tokyo Student Paper	<1%

BessyMukaria_169111_Dissertation_Final_V01.pdf

ORIGINALITY REPORT

12%	11%	10%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	de.overleaf.com Internet Source	1%
2	Submitted to Strathmore University Student Paper	1%
3	digital.lib.washington.edu Internet Source	1%
4	arxiv.org Internet Source	1%
5	proceedings.neurips.cc Internet Source	1%
6	dspace.cvut.cz Internet Source	<1%
7	su-plus.strathmore.edu Internet Source	<1%
8	www.iajournals.org Internet Source	<1%
9	Submitted to The University of Tokyo Student Paper	<1%

10	mmidentity.fmk.sk Internet Source	<1%
11	erepository.uonbi.ac.ke:8080 Internet Source	<1%
12	ojs.aaai.org Internet Source	<1%
13	www.analyticsvidhya.com Internet Source	<1%
14	assets.amazon.science Internet Source	<1%
15	M. Raja, J. Deny, Nagaraj P, Muneeswaran V. "A Peculiar Reading System for Blind People using OCR Technology", 2023 Second International Conference on Electronics and Renewable Systems (ICEARS), 2023 Publication	<1%
16	Submitted to Intercollege Student Paper	<1%
17	Submitted to Massey University Student Paper	<1%
18	Submitted to RMIT University Student Paper	<1%
19	Submitted to Liverpool John Moores University Student Paper	<1%

20	cdn.openai.com Internet Source	<1%
21	"Advances in Information and Communication Technology", Springer Science and Business Media LLC, 2025 Publication	<1%
22	dokumen.pub Internet Source	<1%
23	ijerd.com Internet Source	<1%
24	macsphere.mcmaster.ca Internet Source	<1%
25	Submitted to De Montfort University Student Paper	<1%
26	Submitted to Adtalem Global Education Student Paper	<1%
27	Submitted to University of East London Student Paper	<1%
28	www.mdpi.com Internet Source	<1%
29	cs.rochester.edu Internet Source	<1%
30	journal3.uin-alauddin.ac.id Internet Source	<1%

Appendix B: Ethical Clearance Confirmation

Strathmore UNIVERSITY

19th December 2024

Ms Mukaria Bessy,
bessy.kathure@strathmore.edu

Dear Ms Mukaria,

RE: Leveraging Multimodal AI for Clothing Assistive Reading Solution for Visually Impaired Users

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** proposal. Your application reference number is **SU-ISERC2475/24.** The approval period is from **19th December 2024 to 18th December 2025.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used.
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 72 hours of notification.
iv. Any changes anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 72 hours.
v. Clearance for the export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days of completion of the study to SU-ISERC.

Before commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

Mr Ambrose Rachier,
Chairperson; SU-ISERC

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email admissions@strathmore.edu www.strathmore.edu