Project Report - NLP Spring'25
Enhancing Fashion Search

Kashyap J

May 7, 2025

1 Introduction

Online fashion platforms depend on rich textual descriptions for effective search, filtering and personalised recommendations. However, manual copywriting is time-consuming, prone to errors, and often inconsistent in style across catalogues. This project aims to automate description generation by combining image-based attribute extraction with large vision-language models, ensuring accurate and detailed descriptions. I have used two open-source datasets: FashionPedia (Western clothing) and IndoFashion (Indian clothing), encompassing over 30,000 images each. Using random sampling, I used a subset of four thousand images.

2 Methodology

The pipeline consists of three main stages:

2.1 Data Preparation

- Randomly sample 2,000 images from each dataset to create a balanced sandbox.
- Extract metadata attributes and image features.
- Clean metadata (remove brand noise, non-English tokens) for consistency in prompts.

2.2 Attribute Prediction

Then fine-tuned a ResNet-50 to predict multi-hot attribute vectors indicating visual features (colour, pattern, silhouette, etc).

- Inputs: raw images, metadata-encoded attributes.
- Output: probability scores per attribute, thresholded to binary predictions.
- Evaluated using macro- and micro-average metrics:
  - Macro Precision: 0.2605
  - Macro Recall: 0.1511
  - Macro F1: 0.1731
  - Micro F1: 0.6484

2.3 Description Generation

Using predicted attributes, metadata and raw images, I then constructed prompts for two vision-language pipelines:

1. **BLIP-2 + Flan-T5**: BLIP-2 for visual feature extraction, Flan-T5 for description generation.
2. **LLaVA 13B**: Unified vision-language model for feature extraction and generation of descriptions.

I iteratively refined prompts to avoid unwanted tokens (mentions of models or backgrounds) and to emphasise garment-focused details. LLaVA outperformed BLIP-2 + Flan-T5 in completeness and adherence to prompt constraints.

2.4 Embedding and Retrieval

Generated descriptions were embedded via MiniLM (sentence-transformer) and indexed in FAISS for similarity-based image retrieval:

- Embedded descriptions normalised to unit length.
- FAISS index built, queries embedded and matched by dot-product similarity.

3 Results

3.1 Description Generation

LLaVA-generated descriptions were more structured, complete and adhered to prompt constraints compared to BLIP-2 + Flan-T5.

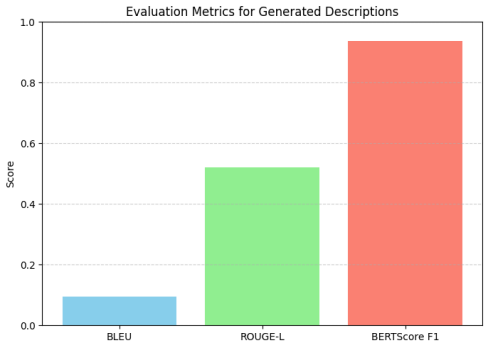
3.2 Retrieval Performance

- P@1: 0.998, R@1: 0.998
- P@5: 0.200, R@5: 1.000
- P@10: 0.100, R@10: 1.000

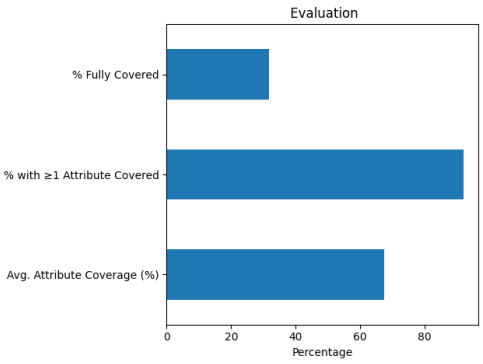
While top-1 retrieval was almost perfect, expanding to larger cutoffs revealed one relevant match per five candidates on average.

4 Evaluation

Qualitative evaluation involved comparing 100 human-written descriptions to model-generated outputs. Feedback indicated that LLaVA descriptions matched human level in detail and accuracy for key visual features, though occasional omissions occurred due to prompt length limits.



(a) Evaluation of Generated Descriptions



(b) Retrieval Performance Metrics

5 Limitations and Challenges

- **Rare Attribute Learning**: Insufficient examples leading to low recall for uncommon features.
- **Prompt Sensitivity**: Small variations altered the style or introduced unwanted tokens.
- **Max Token Constraint**: Truncated outputs required post-processing to ensure completeness.
- **Metadata Noise**: Particularly in IndoFashion, brand names and non-English labels necessitated extensive cleaning.
- **Intent and Event Context**: Descriptions lacked event-based qualifiers (e.g., festive, office, cocktail party) without additional prompt conditioning.

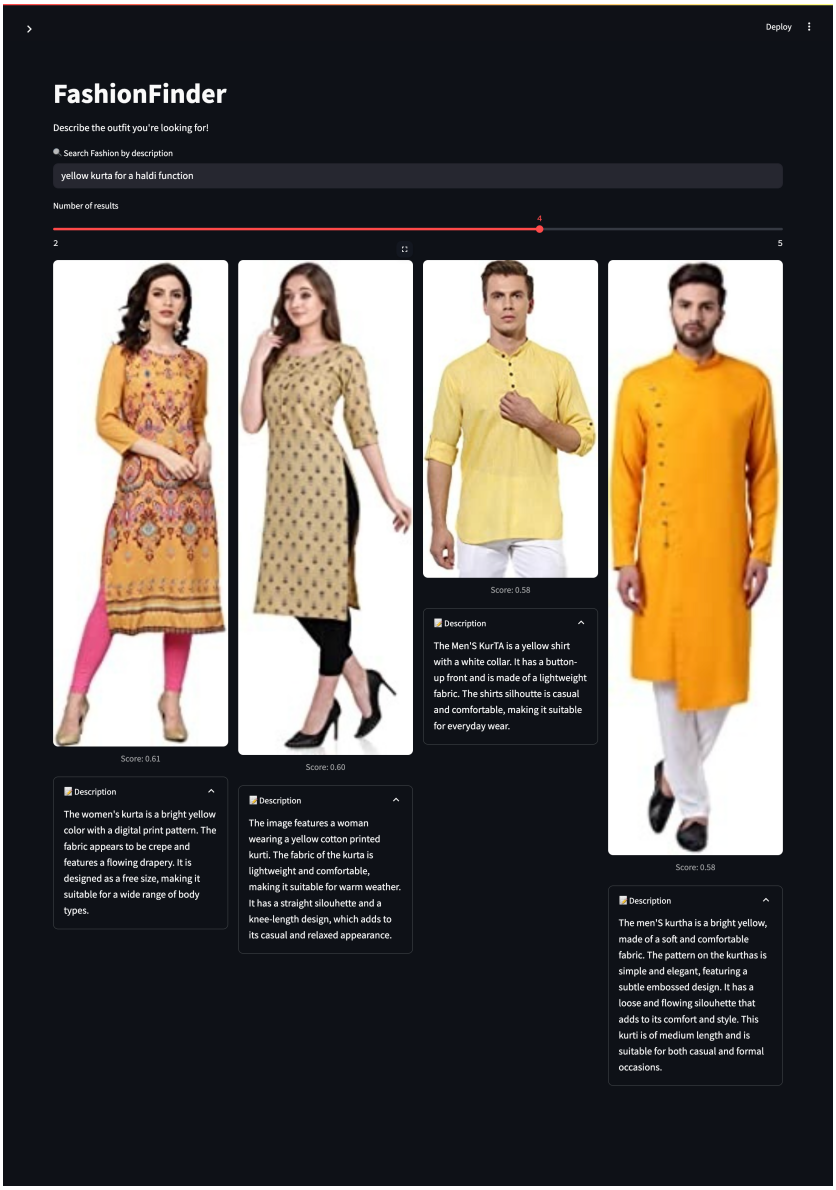
6 Novelty and Motivation

This project uniquely integrates fine-grained attribute prediction and SOTA vision-language models to automate fashion copywriting. By combining structured attribute vectors with open-domain LLMs, the project achieves detailed, garment-centric descriptions that enhance e-commerce search and recommendation systems. The dual-dataset approach validates robustness across Western and Indian fashion domains.

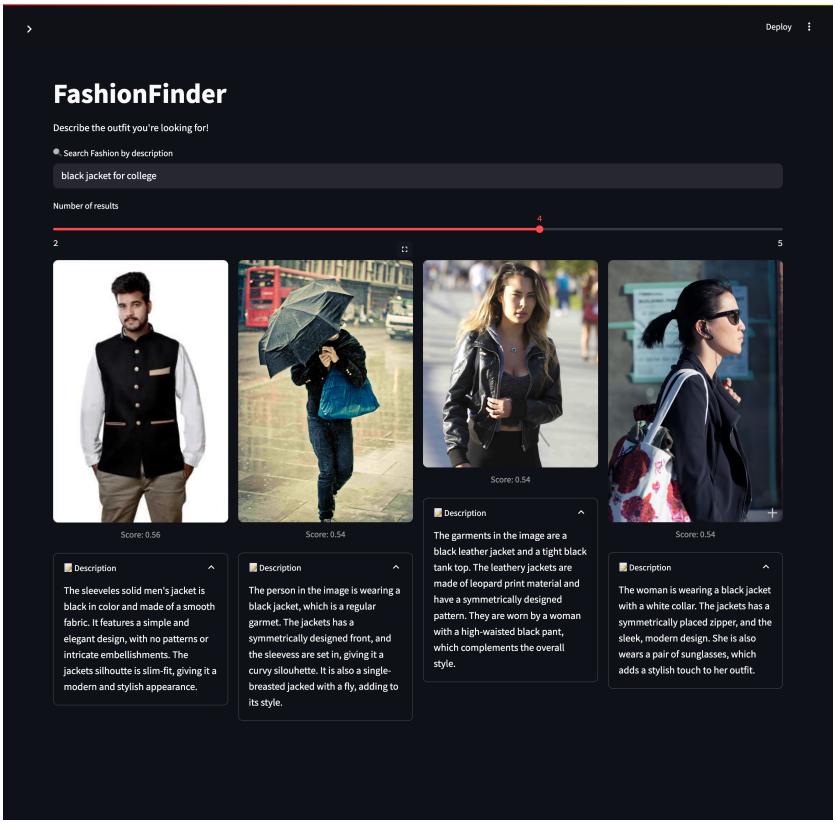
This system addresses the constraints of manual cataloguing, reduces copywriting costs and supports dynamic personalisation by embedding descriptions for real-time retrieval and recommendation. There are now startups that have started adopting something similar using image to image recognition - shoppin' and polopan.

7 Sample Outputs

Below are examples of the system's UI and output performance:



(a) UI Example - Yellow Kurta for a haldi function



(a) UI Example - Black Jacket for college