Journal of Design Intelligence
https://www.ffspub.com/index.php/jdi
ISSN 3104-9001 (print)
E-mail: jdi.office@ffspub.com

Article

Comprehensive Fashion Understanding from Images through Multimodal LLMs

Yanghong Zhou[1,2], Hao Tian[1], Yang Chen[3], P. Y. Mok[3,4,†]
[1] *School of Fashion and Textiles, The Hong Kong Polytechnic University, Hong Kong, China*
[2] *Research Centre of Textiles for Future Fashion, The Hong Kong Polytechnic University, Hong Kong, China*
[3] *Division of Integrative Systems and Design, The Hong Kong University of Science and Technology, Hong Kong, China*
[4] *Research Institute for Intelligent Wearable Systems, The Hong Kong Polytechnic University, Hong Kong, China*
[†] *E-mail: tracy.mok@ust.hk*

Received: November 25, 2025 / Revised: February 15, 2026 / Accepted: March 25, 2026 / Published online: April 24, 2026

Abstract: With the popularity of social media platforms in the information era, people are sharing a large volume of digital content online, including photos and other media data. The availability of media data over the Internet makes it very attractive to mine useful information from these data. However, for fashion image recognition, most existing datasets and methods are limited to coarse-grained categories or a small subset of attributes, lacking a comprehensive dataset and a holistic understanding of clothing. In this paper, we construct a fashion dataset with structured and comprehensive clothing descriptions using an MLLM (Multimodal Large Language Model) grounded in fashion knowledge. To ensure accurate understanding by the MLLM, we define a fashion schema and propose a schema-guided prompting strategy. Leveraging this dataset, we train a model based on BLIP (Bootstrapping Language-Image Pre-training) to recognize category information and fine-grained attributes of clothing images, and to learn text–image aligned representations for clothing image retrieval. Experimental results demonstrate that our approach significantly improves both attribute recognition and cross-modal retrieval performance compared to existing baselines.

Keywords: Clothing Recognition; Fine-Grained Classification; Multi-Modal Large language Model; Clothing Retrieval
https://doi.org/10.64509/jdi.12.54

1 Introduction

Fashion has always been an important part of how people define themselves and others. From what people wear, we can find out their dressing style, clothing preferences, and social status. This type of content has many different uses in social media, dating, recommender systems, and surveillance. In recent years, with the development of deep learning (DL) technology, interesting works were reported where new DL models were proposed for fashion image understanding, for instance, the retrieval of similar products [1] and making clothing suggestions based on a street shot [2], and clothing recognition [3–7] . However, these fashion information from images are limited to satisfy the requirement of fashion understanding for fashion tag labeling and recommendation in practical scenarios. (1) Given that a picture is worth a thousand words in E-commerce, sellers need to understand fashion contents from images as much as possible. (2) Existing fashion datasets are insufficient for a comprehensive understanding of fashion content. Although many studies have proposed various methods to improve the performance of clothing recognition though different ways, such as incorporating clothing landmarks or addressing cross-domain challenges, these methods still rely on datasets with limited annotations and lack a holistic representation of fashion knowledge.

On many e-commerce platforms (e.g. ASOS, UNIQUE, etc), several images are displayed to showcase different characteristics of the clothing and its on-body appearance. Figure 1 shows different types of images on the clothing websites, including detail image, product-only image, part-body image and full-body image. In addition to clothing category and attributes information, which are studied in many works [8, 9], the image type should also be extracted to help sellers provide more complete visual representation of their products. In terms of the clothing category definition, there

† Corresponding Author: P. Y. Mok
*Academic Editor: Zhaohui Wang

is currently no standardized definition for clothing categorization. DARN [4] defines 9 attribute categories of upper clothing, which the clothes category is further divided to 20 fine-grained clothing categories. Deepfashion [5] dataset crawled millions of images, but the annotation is coarse and exists label noises. FashionGen [10] contains a wide range of clothing categories, covering diverse fashion items. Each image is paired with a corresponding textual description, typically detailing the garment's style, color, material, and design. However, these labels do not fully cover all possible fashion attributes. In addition, most of the images in the dataset are generated, which introduces a gap compared to real-world product photos. Deepfashion2 [7] provides the segmentation and landmarks, but only define 13 clothing categories. FashionAI defines a hierarchical structure of women clothing from the perspective of fashion design and annotated carefully by fashion experts. But this dataset only contains women images. There is currently no dataset that covers both men's and women's clothing images with comprehensive and fine-grained fashion annotations, and creating such a dataset requires significant time and labor costs.

To address these issues, this paper proposes an MLLM-based method for learning comprehensive fashion content from fashion images. This method includes recognizing the types of clothing presented, as well as the categories and fine-grained attributes of these fashion items. With the recent development of multimodal large language models (MLLMs) such as `ChatGPT-4` and `DeepSeek-2`, which demonstrate powerful capabilities in multimodal understanding, these models have shown promise in tasks such as question answering, dialogue generation, and information retrieval. Motivated by the success of these models in multimodal tasks, we leverage MLLMs to assist in the construction of our dataset. To achieve this, we propose a fashion-schema based and reverse validation across multiple views Chain of Thought (CoT) approach for data annotation. The fashion-schema defines a hierarchical tree of fashion categories and attributes, enabling the MLLM to generate structured and comprehensive descriptions for clothing items. To ensure accurate understanding, we instruct the MLLM to analyze the attributes of different views of a product item—such as the product-only image, detail image, and local-view—according to the fashion schema. For the image type classification, we will train a classifier to classify the image type. The MLLM will then cross-check the understanding across these different views and retain only the consistent inferences, eliminating any conflicting or inaccurate interpretations. Considering that different views highlight different features of the clothing item, we instruct the MLLM to assign higher weights to views that provide more accurate and relevant information for analyzing specific attributes. For example, the detail image may provide more precise information about the fabric texture or stitching, while the product-only image offers a broader view of the item's overall shape and design. Finally, the structured text description is generated for each product item by the MLLM. Then, with the constructed data, we fine-tune a model based on BLIP (Bootstrapping Language-Image Pre-training) for automatic labeling of fashion images and cross-modal retrieval applications. Figure 1 shows the overall framework for understanding fashion content from images. The major contributions of this work include:

- A new dataset of fashion images with comprehensive and structured clothing description are prepared using knowledge of the fashion industry
- A fashion-schema based and reverse validation across multiple views Chain of Thought (CoT) approach for automatic data annotation, and a BLIP-based model is fine-tune for automatic labeling of fashion images and cross-modal retrieval applications;
- Systematic experimental verification for the effectiveness of the proposed framework is presented.

2 Related Work

2.1 Clothing Recognition

In recent years, due to the popularity of online shopping, clothing classification and retrieval has been an increasingly popular research topic. As clothes often have large variations in style, texture and cutting, and are easily deformed in shape or are occluded, clothing classification and retrieval are very difficult. Therefore, clothing classification often requires the exact localisation of the clothing item. Because of the difficulty with localisation, most of the early work on attribute learning either assumed that the bounding box that contains the object of interest was pre-detected or used the image as a whole for feature extraction [11]. Later, with the improvement of the accuracy of face detection and pose estimation, most works on clothing localisation were based on face and body detection [12], as well as pose estimation [13] to detect the clothing region. But these methods can only handle the condition where face or person is shown in the clothing images. Otherwise, image segmentation methods are often used to localise the clothing region for these clothing images [14–18]. After the pre-detection of clothing, the image feature with regard to the location will then be extracted. A classifier will learn the extracted image feature by machine learning methods. The common features used in clothing attribute classification are colour histogram, SIFT [19], HoG [20] etc, while the common machine learning methods used in clothing classification are Support Vector Machines (SVMs) and Random Forest [21]. However, traditional feature extraction descriptors are primarily hand-crafted, and are not suitable for many kinds of clothing images. Due to the deformation, occlusion and large variation of clothing, it is difficult to design a feature descriptor to solve these problems. Moreover, these clothing classification methods depend greatly on the clothing detector, as well as the limitation of feature extraction. For example, the top5 classification accuracy on DeepFashion dataset [5] using the traditional method [22] is less than 40% [5].

With the development of deep learning technology, some researchers began to use deep neural networks to learn the clothing attribute features from huge amounts of clothing attribute data. For example, Abdulnabi et al. [12] used a multi-task convolutional neural network to learn different attributes on a small dataset [11] with only 1856 images and 23 binary attributes. As the dataset was too small, they combined

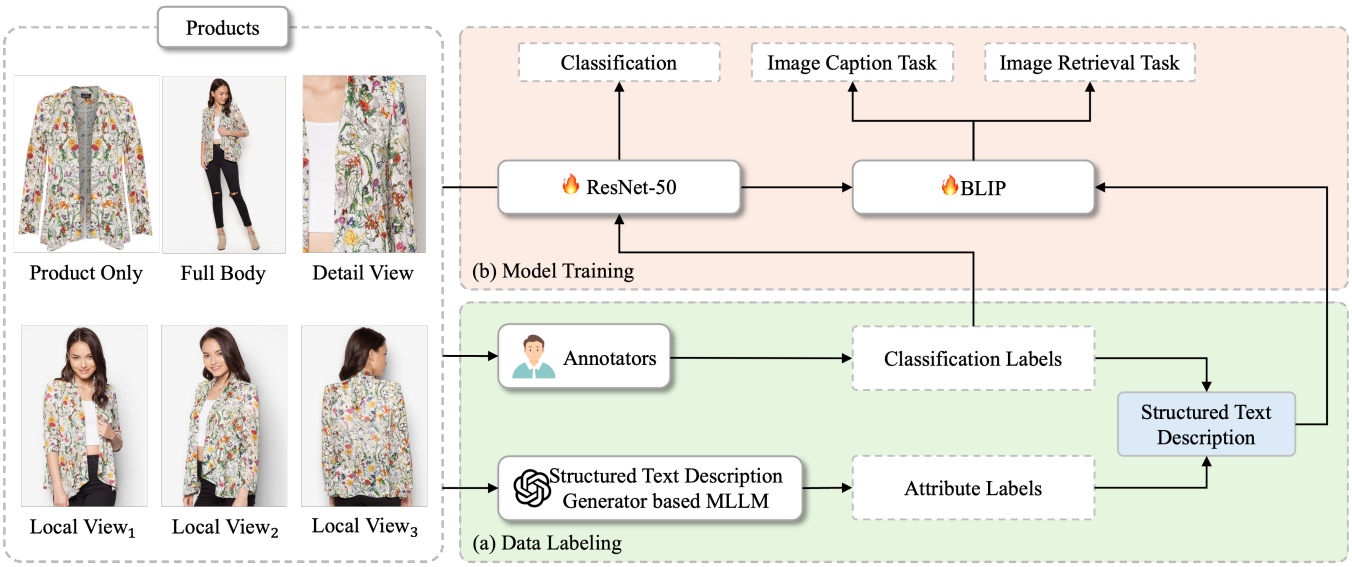

Figure 1: The overall framework for understanding fashion content from images.

the attributes into four categories. Their experimental results demonstrated the effectiveness of deep learning technology in clothing attribute classification. However, one of the major challenges of attribute learning using deep learning is the lack of training data with annotated clothing attributes, which is labour-intensive and time-consuming. To solve this problem, Liu et al. [5] proposed a large fashion dataset with clothing categories, attributes, locations and correspondence of images taken under different scenarios, including stores, street snapshots, and consumers. Based on the dataset, they designed FashionNet to train the clothing category and attribute classification and location prediction simultaneously, which was beneficial to the mutual learning between tasks. As the clothing localisation prediction and clothing category and attribute classification are all in a network for end-to-end training, the clothing category and attribute classification is very efficient. But their method requires the training data to be annotated with clothing locations. Dong et al. [6] proposed a multi-task curriculum transfer (MTCT) deep learning method to jointly learn all attributes for capturing the underlying correlations between different attributes, and also learn the feature similarity between cross-domain images. Their method did not require the training data to be annotated with clothing locations. They trained a clothing detector using Faster-RCNN [13] on an assembled clothing dataset with bounding box annotation to pre-detect the clothing. But the accuracy of the clothing detection was not high, around 70% for in-the-wild images when the threshold of correct detection Intersection over Union (IoU) is set to 0.6 [13].

2.2 Fine-Grained Attribute Classification

Unlike general classification tasks, fine-grained image classification focuses on classifying among categories that are both visually and semantically similar (e.g., bird species, car models, aircraft types). The task is challenging because the same categories contain significant differences and there are obvious similarities between different categories. A wide range of approaches has been proposed to tackle this problem. The main approach is discriminative localization, which can be divided into strongly or weekly supervised localization methods. The strongly supervised methods need not only the category annotations, but also the bounding box or part annotations. For example, Part R-CNN [18] learned whole-object and part detectors, and then localized the parts under geometric constraints and predicted a fine-grained category from a pose-normalization representation. Lin et al. [19] proposed a Deep LAC framework, which consists of localization, alignment and classification sub-networks. They also designed a value linkage function (VLF) to connect the localization and classification sub-networks. Part-Stacked CNN [20] used a fully convolutional network to locate multiple object parts and then a two-stream classification network was designed to encode object-level and part-level cues simultaneously. Instead of learning object detectors, Mask-CNN [21] learned a part segmentation model by using a fully convolutional network to localize part regions. Based on the part segmentation, a three-stream Mask-CNN model was proposed for fine-grained recognition. However, these methods require annotation of the bounding box and part location, which involve heavy manual labelling efforts, so some researchers proposed weakly supervised methods to locate regions. Xiao et al. [23] proposed two-level attention models for fine-grained image classification using the bottom-up attention to generate candidate patches, and then trained an object-level FilterNet to decide whether the patch was related to the basic category. The selected patches were used to train a DomainNet for fine-grained classification. As clustering patterns in the internal hidden representations are found inside the DomainNet, DomainNet is used to provide the part detectors for part-level attention. The patches selected by part detectors are then input into DomainNet to generate the activation score. They add the activation scores of different parts and the original images are then used to train the SVM part-based classifier. Lin et al. [24] proposed bilinear models, which consist of two feature extractors whose outputs are multiplied using the outer product at each location of the image and features are pooled to obtain the bilinear vector. Fu et al. [25] proposed a novel recurrent attention convolutional neural network (RA-CNN) which

recursively learns discriminative region attention and region-based feature representation at multiple scales in a mutually reinforced way. Zheng et al. [26] proposed a multi-attention convolutional neural network (MA-CNN), which enables part generation and feature learning to reinforce each other. MA-CNN proposed channel grouping loss to generate multiple parts by clustering.

Recently, some works have leveraged BERT [27] and CLIP [28] model for cross-modal image retrieval in fashion filed. For example, Gao et al. [29] proposed the FashionBERT framework, which segments each image into patches and feeds them, together with text tokens, into a BERT model for joint representation learning. To enhance fine-grained alignment between text and image at the patch level, Zhuge et al. [30] proposed to integrating BERT with local features extracted using Region of Interest (RoI) based methods. However, the popular RoI-based methods detect unsatisfactory region proposals, resulting in suboptimal performance. To address this challenge, EI-CLIP [31], which is based on the CLIP model, is proposed for e-commerce product retrieval. EI-CLIP incorporates an Entity-Aware Learning Module (EA-learner) and a Confounding Entity Selection Module (CE-selector). Similarly, another approach based on CLIP, FAME-ViL [32], introduces a Cross-Attention Adapter (XAA) and a Task-Specific Adapter (TSA) to adapt the pre-trained CLIP model for various fashion tasks within a unified framework. However, these methods mainly focus on category , and are limited in their ability to generate comprehensive and detailed descriptions. Therefore, we leverage BLIP to fine-tune the model for caption generation, ensuring that it produces comprehensive clothing descriptions.

2.3 Multimodal Large Language Model

Multimodal large language models have demonstrated powerful abilities in multimodal understanding and have been successfully applied to a wide range of tasks, including visual question answering, image captioning, and open-domain multimodal dialogue. A common approach is to fine-tune large language models to enable them to handle image–text interactions by learning from multimodal instructions. Representative methods, such as LLaMA-Adapter [33] and LaVIN [34], employ parameter-efficient adapter modules or modality-mixing strategies to achieve this while minimizing training costs. In contrast to fine-tuning, prompting accomplishes specific tasks without updating model parameters by providing context, examples, or instructions. Due to its train-free nature, prompting has gained significant attention, particularly for addressing multimodal chain-of-thought (CoT) reasoning, where prompts guide models to generate both reasoning processes and answers. Chain-of-thought reasoning was first introduced by Wei et al. [35], enabling LLMs to improve accuracy on complex reasoning tasks while enhancing interpretability and transparency of decision-making. Building on this foundation, Zhang et al. [36] extended CoT reasoning to MLLMs to better capture intricate cross-modal relationships and achieve robust generalization across diverse multimodal benchmarks. Recent works further explore combining CoT reasoning with reranking strategies; for example, Wu et al. [37] proposed RankCoT, integrating reranking signals with CoT generation to improve retrieval-augmented generation (RAG) accuracy, while JudgeRank [38] employs explicit reasoning steps to mimic human judgment and enhance reranking in complex tasks. Conversely, Jedidi et al. [39] demonstrated that standard rerankers often outperform reasoning-based approaches, challenging the necessity of reasoning chains in reranking. However, few studies have explored the use of MLLMs for comprehensive and structured annotation in the fashion field.

3 Method

3.1 Data Collection

Only a few of the publicly available image classification datasets are focused on fashion, such as DeepFashion [5], a database with 800,000 fashion images in 50 categories and Deepfashion2 [7], a database with 491,000 images in 13 popular clothing categories. However, we find that these categories may not fit our use as the categorizations are not structured in a hierarchy of levels. We therefore develop our own dataset in this study. We crawled data from online shopping websites, including ZALORA[1], H&M[2], ASOS[3], and UNIQLO[4]. A total of 158,211 products were crawled from these websites, including 626,516 product images. Apart from product images, we also obtained different product information, including product category, brand, gender, price, and description. As the product description and category information from different websites are neither consistent nor complete, we define a new taxonomy structure to reclassify and unify the category and subcategories (also named fine-grained attributes hereafter) of the fashion products. We then performed extensive data cleaning and matched fashion images with correct labels. We excluded noisy images which are not related to clothing, like packing box images and size chart images, and also images that cannot be recognized by the human eyes. After cleaning, a total of 166,087 images for fashion items were used for model training in the experiment.

3.2 Data Labeling

We divide all clothing items into four sets, according to the image type, including product-only, local-view, total-look and details, as shown in Figure 1. These images show the clothes from different views. For example, product-only images are fashion items not being worn by people in the images, which are for analyzing clothing features. Local-view images show the effect of clothes with people wearing them, and we should remove other backgrounds before analysis; total-look images show the matching effect of different items on the person, which is useful for analyzing clothing coordination; details images show zoomed view of some clothing details, so they allow the study of detailed texture information. For each product item, we manually annotate the image type, clothing

[1] https://www.zalora.com
[2] https://www.hm.com
[3] https://www.asos.com
[4] https://www.uniqlo.com

category, and fine-grained category, ensuring that the dataset's ground truth is human-verified and accurate. The defined categories and their corresponding image counts are presented in Figure 2, while the fine-grained categories and image counts are shown in Figure 3. For detailed attribute annotation, we propose a fashion-schema based and reverse validation across m ultiple views Chain of Thought (CoT) approach, which will be detailed in this section.

Fashion schema We define a fashion schema organized as a hierarchical tree structure, starting from general clothing types (e.g., top, bottom, overall), which are further expanded into categories and fine-grained subcategories. Beyond categorical classification, the schema incorporates detailed attributes such as silhouette, color, style, pattern (including presence, size, and coverage), texture, fashion style, and usage scenario. More specifically, in the new taxonomy structure of fashion items, we defined 19 clothing categories, including 2 categories of full-body clothing, 5 categories of upper clothing, 4 categories of bottom clothing, 2 categories of intimates, and 6 categories of accessories. The number of images in each category under different image types are shown in Figure 2. As details images are not suitable for category recognition, we retained only the product-only, local-view and total-look images for the training of clothing category classifier. The dataset for category classification includes a total of 129,997 clothing images. For the 13 clothing categories, a total of 106 subclasses (also called fine-grained attributes) are defined. With the defined fine-grained attributes of each clothing category, all images in the category data set are labeled with associated fine-grained attribute class labels. For training attribute classifiers, we prepare datasets for upper clothing, bottom clothing, full-body clothing, and four separate datasets for each of the accessory categories. The attributes of each category and the number of images for each attribute classification model are shown in Figure 3.

MLLM annotation Since manual annotation is highly time-consuming and labor-intensive, we leverage an MLLM, Qwen2.5, to automatically annotate detailed attributes and generate structured fashion descriptions. To this end, design a Chain-of-Thought reasoning prompt that fully leverages all available image views of a product and incorporates a reverse validation mechanism to ensure accuracy. Specifically, for each clothing item, we employ Qwen2.5 to carefully analyze all images and answer the following questions step by step:

- Silhouette: What is the silhouette or shape? (e.g., A-line, H-shape, etc.)
- Color: What is the main color? (e.g., gray, red, blue, pink, etc.)
- Pattern: Is there any pattern? If so, what kind? (e.g., solid, small patterned, etc.)
- Texture: What is the texture or fabric? (e.g., silk, wool, cotton, etc.) Style: What is the overall style? (e.g., casual, formal, romantic, etc.) Occasion: What occasion is this clothing suitable for? (e.g., date, party, sport, etc.)
- Other details: Are there any other notable features? (e.g., V-neck, long sleeves, etc.)
- If additional obvious attributes or details are observed, please include them.

Because different types of images vary in their ability to reveal specific attributes, we instruct the MLLM to assign higher reference weights to views that provide more accurate information for analyzing corresponding attributes.

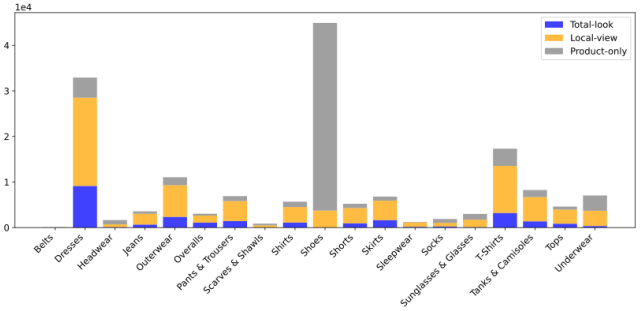

Figure 2: Number of images of each clothing category.

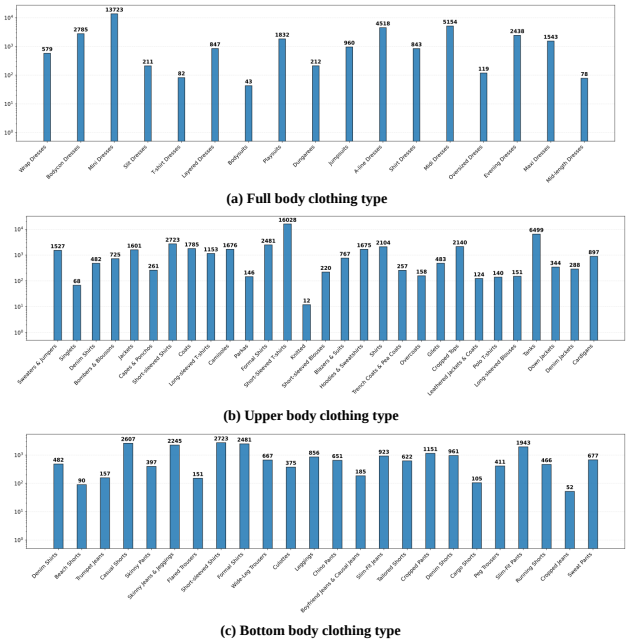

Figure 3: Number of images for each fine-grained clothing category.

Based on the above attributes, MLLM is prompted to construct a single sentence that contains all identified features and details of the clothing item. Since descriptions obtained from multiple image perspectives may introduce inconsistencies, we incorporate a reverse validation mechanism to check whether the summarized attributes and generated description fully and accurately correspond to every provided image. The reverse validation step is passed only when the MLLM verifies that the annotations do not conflict with any of the provided images. As a result, only attributes and descriptions that meet this strict consistency criterion are retained, ensuring high-quality data labeling. To further improve accuracy, we provide illustrative examples within the prompts. Finally, based on the attributes and final descriptions generated by the MLLM, we perform manual reviews to further ensure the accuracy of the annotations. Figure 4 illustrates the overall pipeline of our MLLM-based annotation process.

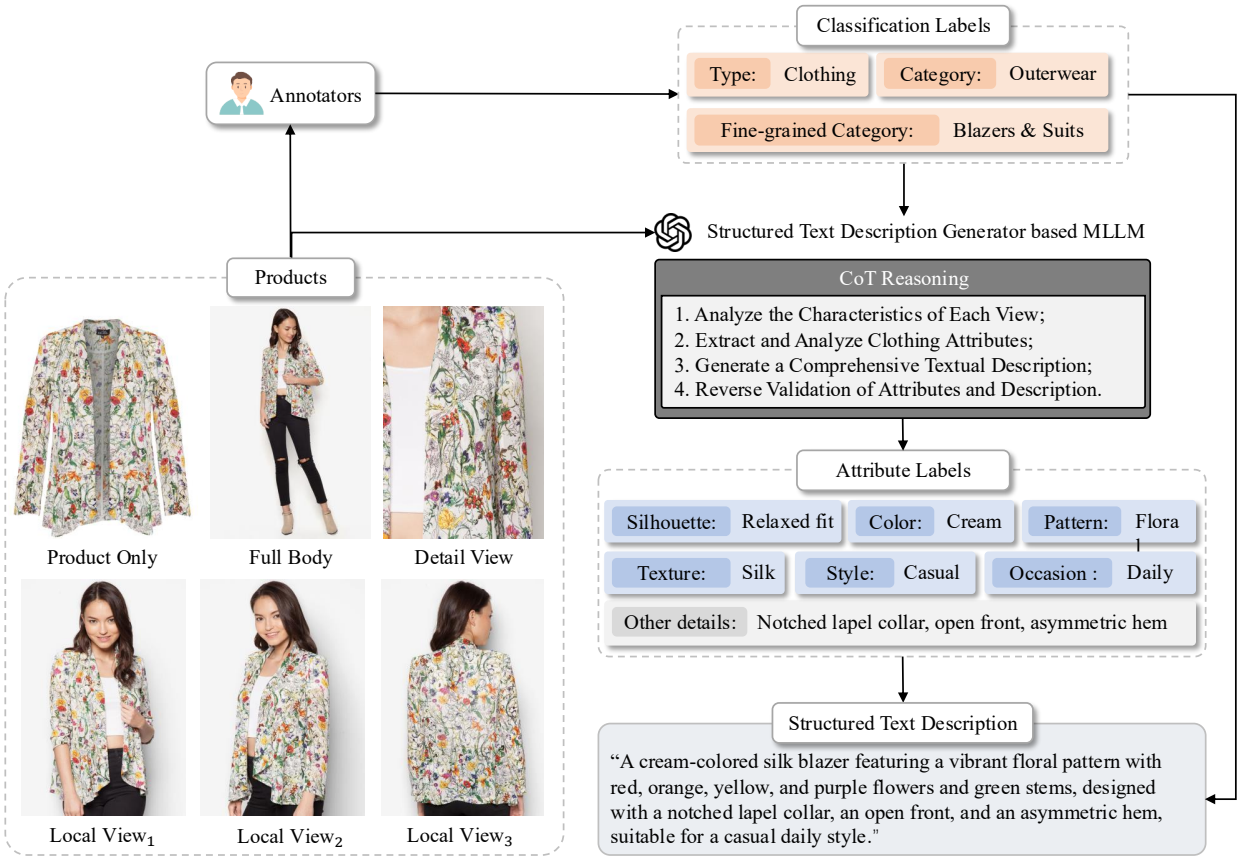

Figure 4: The pipeline for data annotation.

3.3 Fine-Tuning based on BLIP

Using the constructed clothing dataset with detailed item descriptions, we train a model based on the BLIP framework to perform image captioning and image retrieval tasks. Given an image I and its corresponding description T, we input them into an image encoder and a text encoder, respectively, to extract the image embedding f_I and the text embedding f_T. The image encoder consists of a set of self-attention layers followed by a feed-forward layer, and the text encoder consists of *12* bidirectional self-attention layers followed by a feed-forward layer. For the text encoder, a `[CLS]` token is appended to the text input to summarize the text description. Additionally, the image embedding, together with the text and an `[Encode]` token, is fed into an image-grounded text encoder to extract a multimodal representation f_m of the image-text pair. A cross-attention layer is incorporated between the self-attention and feed-forward network to capture the relationship between text and image embeddings. To generate captions for the given image, a grounded text decoder is employed, which outputs the prediction logits for the caption tokens. The structure of the grounded text decoder is similar to the text encoder, but replaces bidirectional self-attention with causal self-attention. Figure 5 shows the network structure of BLIP model.

The model is optimized using three objectives: the image-text contrastive (ITC) loss computed on f_I and f_T, the image-text matching (ITM) loss computed on f_I and f_m, and a language modeling (LM) loss computed on the predicted token probabilities p. The formulations are:

$$Loss_{\text{ITC}} = -\log\left(\frac{\exp(\text{sim}(f_I, f_T)/\tau)}{\sum_{i=1}^{N}\exp(\text{sim}(f_I, f_{T_i})/\tau)}\right), \quad (1)$$

where $\text{sim}(f_I, f_T) = \frac{f_I \cdot f_T}{\|f_I\|\|f_T\|}$ is the cosine similarity, τ is a temperature parameter, and N is the number of negative samples in the batch.

$$Loss_{\text{ITM}} = -\frac{1}{M}\sum_{j=1}^{M}\left[y_j \log p_j + (1-y_j)\log(1-p_j)\right], \quad (2)$$

where $y_j \in \{0,1\}$ indicates whether the image-text pair matches, and p_j is the predicted probability.

$$Loss_{\text{LM}} = -\sum_{t=1}^{T}\log P(w_t | w_{<t}, I), \quad (3)$$

where w_t is the token at position t, conditioned on previous tokens and image features. The overall loss is computed by:

$$\mathscr{L} = Loss_{\text{ITC}} + Loss_{\text{ITM}} + Loss_{\text{LM}}. \quad (4)$$

4 Experiments

We demonstrate the effectiveness of our framework by evaluating the performance of multiple tasks including image type classification, clothing category classification, image-to-text

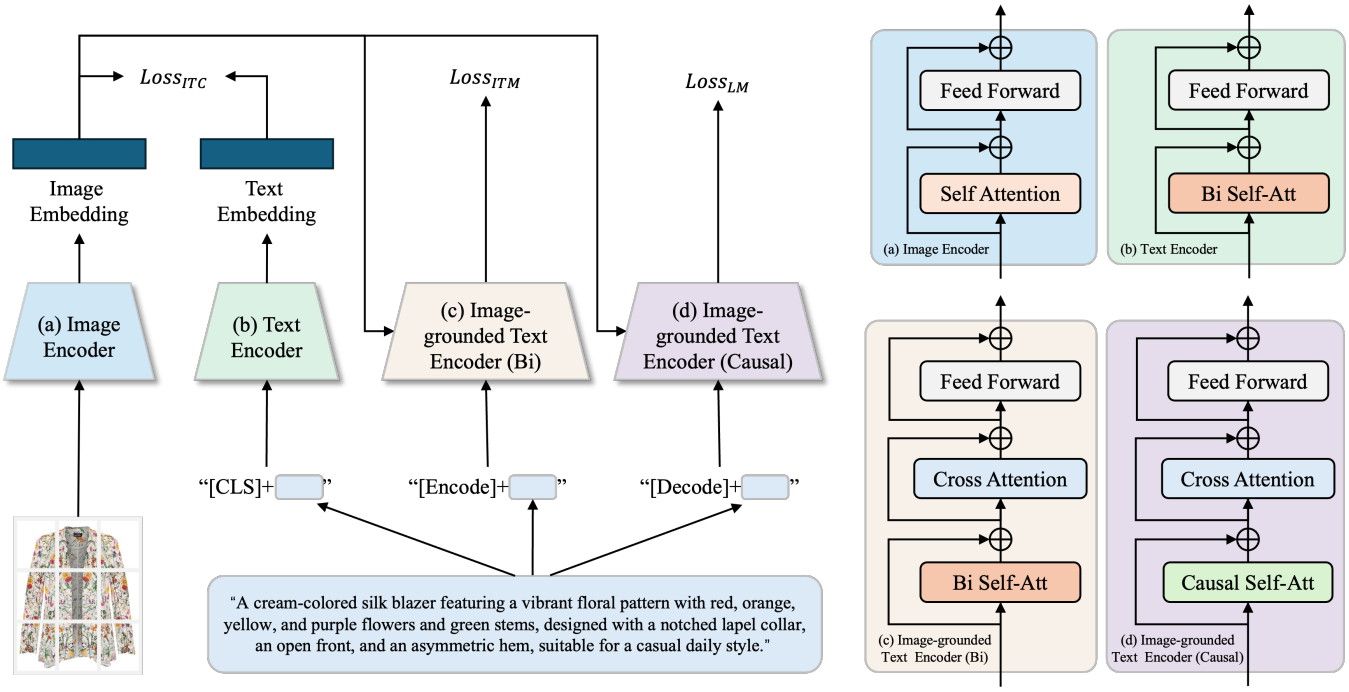

Figure 5: The network structure of BLIP model.

retrieval task, text-to-image retrieval task, and image caption task.

4.1 Implementation Details

Configuration details Given the datasets, we split the images into three sets: 80% for training, and 20% for testing. For classification tasks, the training images are rescaled to 256×256, and augmented with a center crop to 224×224 and horizontal flip. We trained the classification networks by fine-tuning weights pre-trained on the ImageNet dataset with a mini-batch stochastic gradient descent with a momentum of 0.9, and fixed learning rate of 0.001. The batch size was set to 50. For image retrieval and image caption tasks, we fine-tuned the BLIP pre-trained on the COCO dataset with an initial learning rate of 0.00001, weight decay of 0.05, batch size of 5. All models used in these tasks were trained on an NVIDIA GTX 3090 GPU. During the inference phase, the input image size is 224×224. For the MLLM, we employed Qwen2.5 to perform complex visual reasoning for both image type and clothing category classification tasks.

Evaluation metrics The evaluation metrics used to assess the performance of the network are based on the confusion matrix, which show the difference between the predictions and ground truths. Table 1 is an example of a confusion matrix for a two-class classification problem, where the rows (positive/negative) represent the actual value and the columns (true/false) represent whether the prediction is true or false. The confusion matrix can provide detailed information on the classification results, and it also highlights the classes where the predictions are confused. Based on the confusion matrix, we can define the classification accuracy as follows:

$$\text{Acc} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}$$

Additionally, the recall measures the proportion of actual positive instances that are correctly identified by the model:

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}$$

To evaluate generated captions in image captioning, we use four standard metrics: BLEU-4 [40], METEOR [41], ROUGE [42], and CIDEr [43]. BLEU-4 assesses machine-human output correspondence, based on 'the closer a machine translation is to a professional human translation, the better it is,' using geometric mean of n-gram precisions (up to 4-grams) with brevity penalty. METEOR evaluates via unigram matching with exact, stem, and synonym alignments, computing harmonic mean of precision and recall (recall weighted higher). ROUGE, from Recall-Oriented Understudy for Gisting Evaluation, measures summary quality against human references via F-measure on longest common subsequence for structural similarity. CIDER provides consensus-based image description evaluation, better aligning with human judgments, through TF-IDF weighted n-gram cosine similarity ($n \leq 4$) emphasizing informative terms. These metrics collectively gauge lexical overlap, semantic fidelity, and human alignment.

Table 1: Confusion Matrix for Clothing Type Classification

		Predictions	
		True	False
Ground-truths	Positive	TP	FP
	Negative	TN	FN

4.2 Experimental results and analysis

To evaluate the performance of MLLMs in clothing image understanding, we first compare a ResNet-based model with an MLLM on image type classification and category classification tasks. Subsequently, we assess the fine-tuned BLIP model (denoted as BLIP-ft) on cross-modal retrieval and image captioning.

Evaluation on image type classification We evaluate the clothing type classification performance in this section. We compute the total accuracy of the model of each epoch on the validation dataset and choose the model with the highest accuracy as the best model. After training, we evaluate this best model on the test set. The average accuracy on the test data is 96.94%, and the per-class accuracy values are compared against MLLM in Table 2. As shown, the accuracy values for all image types on the test data are near or above 95% with ResNet. Specifically, the accuracy for the local-view and product-only images is nearly 98% and 97%, respectively, both of which are quite high. However, MLLM performs particularly poorly on local-view images. One possible reason for this discrepancy is the difference in the definitions of 'local-view' and 'total-look'. In our definition, 'total-look' refers to an image that must include a complete outfit, whereas 'local-view' may not necessarily adhere to this full clothing ensemble. Furthermore, this performance discrepancy highlights a divergence in feature learning mechanisms. The ResNet model, trained via supervised learning, effectively captures dataset-specific visual biases, such as cropping patterns and textural distributions intrinsic to the 'Local-view' class. Conversely, the MLLM relies on zero-shot semantic reasoning; without fine-tuning, it struggles to distinguish the arbitrary boundary between a 'Local-view' and a 'Product-only' image. This demonstrates that supervised learning remains essential for aligning models with precise, domain-specific taxonomic standards.

Table 2: Evaluation results on image type classification.

Image Type	ResNet50	MLLM
Total-look	94.68%	**98.96%**
Local-view	**97.77%**	33.15%
Product-only	**96.83%**	96.68%
Total	**96.94%**	67.66%

Evaluation on clothing category classification To study the ability of MLLM on clothing category classification, we also compare the performance of the model fined on ResNet-50 and MLLM. The comparison results are shown in Table 3. It can be observed that both models perform similarly in terms of mean accuracy. However, MLLM exhibits more extreme results: it performs exceptionally well in certain categories, such as Sunglasses & Glasses and Shirts, but performs poorly in others, such as Overalls and Tops, with accuracy dropping below 30%. This suggests that while MLLM demonstrates a strong understanding of common knowledge, it still lacks a deep understanding of more specialized fashion-related knowledge. Specifically, the performance difference in categories like 'Tops' and 'Overalls' highlights the gap in category definitions. In general training data, terms like 'Top' are often used as broad labels for various upper-body clothes. In contrast, our dataset uses a strict classification system where these categories do not overlap to ensure precision. The general MLLM is not aware of these specific definitions, whereas the supervised model effectively learns the clear class boundaries defined in our dataset. In future work, we plan to adopt a semi-supervised approach by manually labeling a subset of detailed attributes to train a deep learning model for attribute recognition. At the same time, we will guide the MLLM to understand these attributes, ensuring consistency between the two models' outputs. Any discrepancies will be manually reviewed and corrected.

Table 3: Evaluation results on clothing category classification.

Clothing Category	ResNet50	MLLM
Belts	77.78%	**100%**
Dresses	**89.25%**	89.12%
Headwear	**92.12%**	79.20%
Jeans	75.56%	**78.42%**
Outerwear	**73.53%**	50.52%
Overalls	**56.32%**	22.30%
Pants & Trousers	**79.78%**	78.20%
Scarves & Shawls	64.16%	**95.91%**
Shirts	93.58%	**97.89%**
Shoes	98.75%	**100%**
Shorts	74.55%	**78.97%**
Skirts	71.18%	**86.06%**
Sleepwear	46.56%	**64.08%**
Socks	94.16%	**98.14%**
Sunglasses & Glasses	99.33%	**100%**
T-Shirts	88.39%	**89.98%**
Tanks & Camisoles	**77.25%**	62.76%
Tops	**52.55%**	28.99%
Underwear	92.62%	**95.18%**
Total	**86.68%**	84.90%

Evaluation on image retrieval Table 4 and Table 5 show the results on the image-to-text retrieval and text-to-image retrieval tasks. It can be seen that BLIP-ft outperforms the original BLIP trained on general images by a large margin on both tasks. Specifically, compared to BLIP, BLIP-ft achieves higher recall at all ranks, with improvements of +3.4% at R@1, +14.5% at R@5, and +16.8% at R@10 for image-to-text retrieval, and +2.6% at R@1, +15.3% at R@5, and +18.9% at R@10 for text-to-image retrieval. This substantial improvement is directly driven by the model's ability to leverage the comprehensive fine-grained annotations provided in our dataset. Unlike standard pre-training which targets coarse-grained objects, our fine-tuning process aligns visual features with precise attribute descriptors (e.g., specific textures or collar types). This granular alignment optimizes the embedding space, enabling the model to effectively distinguish between visually similar items, thereby validating the effectiveness of our fine-grained annotation strategy.

Table 4: Evaluation results on Image-to-Text retrieval.

Model	R@1	R@5	R@10	R@mean
BLIP	10.9%	48.4%	62.2%	40.5%
BLIP-ft	14.3%	62.9%	79.0%	52.1%

Table 5: Evaluation results on Text-to-Image retrieval.

Model	R@1	R@5	R@10	R@mean
BLIP	14.0%	49.0%	60.7%	41.2%
BLIP-ft	16.6%	64.3%	79.6%	53.5%

Evaluation on image caption Table 6 shows the comparison results for the image captioning task. As shown, the original BLIP performs poorly across all metrics, including BLEU-4, METEOR, ROUGE and CIDEr, indicating that the original BLIP struggles with clothing descriptions, particularly with providing comprehensive descriptions. After fine-tuning, the performance shows significant improvement: BLEU-4 increased from 0.9 to 7.8, METEOR improved from 6.2 to 17.3, and CIDEr rose from 2.7 to 28.9. This demonstrates that fine-tuning enhanced the model's understanding of the clothing domain, improving the grammatical structure, vocabulary choice, and semantic consistency of the generated text. Figure 6 presents a qualitative comparison of the generation results. While the original BLIP produces brief captions that often lack detail, BLIP-ft delivers more comprehensive and accurate descriptions with superior information density. This improvement stems from the incorporation of our proposed Fashion Schema during fine-tuning, which guides the model to prioritize a logical hierarchy of attributes—from global silhouettes to intricate design details. Consequently, the model effectively grounds the generation in visual evidence, ensuring the outputs are both structurally coherent and faithful to the specific product characteristics.

Table 6: Evaluation results on image caption.

Model	Bleu-4 [40]	METEOR [41]	ROUGE [42]	CIDEr [43]
BLIP	0.9	6.2	18.6	2.7
BLIP-ft	7.8	17.3	33.1	28.9

Figure 6: Qualitative comparison results on image caption.

4.3 Ablation Study

We conduct qualitative ablations by comparing the full pipeline with versions that omit one component at a time, allowing us to isolate the effects of Multi-View Fusion, Chain-of-Thought Prompting, Fashion Schema, and Reverse Validation.

Effect of Multi-view Fusion Integrating multi-view images is essential for reducing hallucinations in occluded regions and enhancing the faithfulness of garment-construction descriptions. As shown in Figure 7, the method without multi-view inputs exhibits severe hallucinations: it incorrectly predicts a front zipper closure and overlooks the navy-blue and white striped side panels, instead generating a generic floral description. In contrast, our proposed pipeline aggregates complementary information from different viewpoints, correctly identifying the back zipper closure and accurately capturing the side-panel details. These results demonstrate that multi-view inputs effectively prevent unsupported inference of invisible features and are thus indispensable for reliable attribute extraction.

Effect of Chain of Thought Prompting As shown in Figure 8, the method without CoT tends to hallucinate fine-grained subcategories, misclassifying the item as 'maternity peg trousers' and fabricating a 'black stretch panel' based solely on the presence of an elastic waistband. In contrast, our proposed method systematically methodically extract core features, correctly characterizing the item as standard 'Peg' silhouette trousers with a high, elasticated waist.This demonstrates that CoT effectively suppresses the model's impulse to over-associate generic visual features with niche categories, resulting in more precise and generalizable terminology.

Effect of Fashion Schema Figure 9 presents comparison examples between our proposed pipeline and the pipeline omitted the Fashion Schema. It can be observed that the pipeline without explicit schema constraints tends to conflate physical attributes with abstract stylistic classifications, leading to information redundancy and indexing failures. For instance, the pipeline lacking the Fashion Schema incorrectly populates the Style field with physical descriptors such as 'Sleeveless' and 'Fitted Bodice,' whereas the method equipped with the Fashion Schema, guided by the ontology, correctly assigns these physical characteristics to the Other details field while filling the Style field with an appropriate abstract label, 'Casual to semi-formal.' This demonstrates that the structured Fashion Schema plays a pivotal role in maintaining semantic consistency and ensuring adherence to field-specific definitions in the generated output.

Effect of Reverse Validation The Reverse Validation module functions as a critical filter for visual hallucinations, particularly those arising from biases in the training data distribution. As shown in Figure 10, the method without Reverse Validation misclassifies the silhouette as 'A-line,' likely relying on the high prior probability of this cut for sleeveless tops in the dataset rather than grounding its decision in concrete visual cues. With Reverse Validation enabled, the MLLM re-evaluates the visual evidence—such as the button-down front and chest pockets—and accordingly revises the prediction to the correct 'H-shape' (straight cut). This demonstrates that Reverse validation compels the model to align its outputs

Experiment Name	Type	Silhouette	Color	Pattern	Texture	Style	Occasion	Other details
Ours	Upper clothing	H-shape	White with pink, navy blue, and gray accents	Floral and geometric mixed print	Lightweight fabric, likely cotton or viscose blend	Casual, bohemian	Casual outings, summer wear, beach	V-neck, thin adjustable straps, **back zipper closure, side panels with navy blue and white stripes**, floral and geometric patterns
W/O Multi-view	Upper clothing	A-line	Multicolor (predominantly pink, white, navy blue, and gray)	Floral and geometric mixed print	Lightweight fabric, likely chiffon or crepe	Casual, romantic	Casual wear, summer outings, beach, date	Thin adjustable straps with metal rings, V-neckline, **front zipper closure** at neckline, **side panels with striped and polka dot patterns**

Figure 7: Qualitative analysis of the effect of Multi-view Fusion.

Experiment Name	Type	Silhouette	Color	Pattern	Texture	Style	Occasion	Other details
Ours	Bottom clothing	Peg	Beige	Solid	Cotton blend	Casual	Daily wear	High waist, elastic waistband, cuffed hems, side pockets
W/O CoT	Bottom clothing	Peg Trousers	Beige	Solid	Smooth, lightweight fabric with a slight sheen	**Maternity peg trousers with a high**, elasticated waistband designed to accommodate a growing belly, featuring cuffed ankles and a slim, tapered fit from knee to hem	Casual, everyday wear, maternity wear	The trousers feature a black, stretchy, high-waisted panel that sits over the bump, providing support and comfort. The main fabric is beige and appears to be a soft, flowing material. The waistband includes a white elasticated inner band for added comfort. The legs are tapered and end in neatly cuffed hems.

Figure 8: Qualitative analysis of the effect of CoT prompting.

Experiment Name	Type	Silhouette	Color	Pattern	Texture	Style	Occasion	Other details
Ours	**Full-body clothing**	A-line	Black, Red	Solid	Smooth, lightweight fabric (likely cotton or polyester blend)	**Casual to semi-formal**	Casual outings, semi-formal events, smart casual occasions	Sleeveless with thin spaghetti straps, scoop neckline, knee-length hem, pleated skirt, elastic waistband
W/O Fashion Schema	**Dress**	A-line	Black, Red	Solid	Smooth, Lightweight	**Sleeveless, Thin Spaghetti Straps, Fitted Bodice, Pleated Skirt**	Casual, Semi-Formal	**Round Neckline, Simple and Elegant Design**

Figure 9: Qualitative analysis of the effect of Fashion Schema.

with actual visual observations rather than probability-driven intuition, thereby correcting errors that may appear plausible but are ultimately unsupported.

4.4 Generalization Analysis

To assess the cross-domain robustness of the proposed framework and reduce potential domain bias introduced by standardized e-commerce datasets, we conduct a qualitative evaluation on unconstrained street-style scenarios. As shown in Figure 11, the model preserves high fidelity in attribute extraction when transitioning from studio-controlled imagery to complex real-world environments characterized by background clutter, environmental noise, and non-uniform lighting. In the pink jacket example on the left, the system successfully isolates the primary garment from substantial background distractions—including non-fashion objects such as toys and tableware—and accurately identifies fine-grained construction details such as the fur collar and front toggle

Experiment Name	Type	Silhouette	Color	Pattern	Texture	Style	Occasion	Other details
Ours	Upper clothing	**H-shape**	navy blue	solid	cotton blend	casual yet chic	casual outings, casual office wear	V-neckline, button-down front, two chest pockets, sleeveless, fitted waist
W/O Reverse Validation	Upper clothing	**A-line**	dark blue	solid	cotton blend	casual	casual outings, casual office wear	sleeveless, V-neckline, button-up front, pockets on either side of the chest

Figure 10: Qualitative analysis of the effect of the Reverse Validation.

Sample Group	Type	Silhouette	Color	Pattern	Texture	Style	Occasion	Other details
Left	Upper clothing	A-line	pink	solid	satin	cute	casual	fur collar, black drawstring, front toggle closures, lace-trimmed pockets, ruffled lace hem, long sleeves with elastic cuffs
Right	Bottom clothing	A-line	off-white	solid	lightweight fabric with lace trim	romantic	casual, date	layered ruffles, lace hem, elastic waist

Figure 11: Generalization to unconstrained, less controlled real-world settings.

closures. Similarly, in the dynamic outdoor street-scene example on the right, the model exhibits strong adaptability to natural lighting variations and non-standard poses, correctly capturing the layered ruffles and A-line silhouette and assigning an appropriate 'sophisticated romantic' style label. These observations indicate that the integration of the structured Fashion Schema and Chain-of-Thought reasoning effectively prevents the model from overfitting to clean data distributions. By enforcing strict semantic boundaries and maintaining logical grounding, the framework mitigates domain bias and demonstrates practical utility for diverse downstream applications such as social media trend analysis and unconstrained street-style annotation.

4.5 Discussion

By leveraging the capabilities of MLLMs on multimodal data, we can automatically generate comprehensive clothing descriptions for apparel items, which improves the performance of image captioning and cross-domain retrieval. However, we observed that while MLLMs demonstrate strong understanding of general knowledge, they lack expertise in domain-specific knowledge, where supervised learning methods perform significantly better. This suggests that a hybrid approach combining the strengths of both paradigms could be highly effective. In future work, we plan to adopt a semi-supervised approach by manually labeling a subset of detailed attributes to train a deep learning model for detailed attribute recognition. At the same time, we will instruct the MLLM to understand these attributes, ensuring consistency between the two models' outputs. Any discrepancies will be manually reviewed and corrected.

5 Conclusion

This paper proposes a novel framework leveraging Multimodal Large Language Models (MLLMs) for comprehensive fashion understanding from images. We construct a fashion dataset with structured and detailed descriptions, manually annotating image type, category, and fine-grained category for each clothing item. Detailed descriptions are generated using a fashion-schema-based approach combined with reverse validation across multiple views through a Chain-of-Thought (CoT) process. Using this dataset, we fine-tune BLIP (denoted as BLIP-ft) and evaluate its performance on image captioning and cross-modal retrieval tasks. For future work, we aim to further enhance description quality by integrating deep learning models with MLLMs in a semi-supervised manner.

Funding

The work described in this paper was supported, in part, by the Innovation and Technology Fund (Project: ITP/004/24TP) and by the Research Institute for Intelligent Wearable Systems (Grant: CD95/ P0049355) and the Research Centre of Textiles for Future Fashion (Grants: BDVH/P0051330 & BBFL/P0052601) of The Hong Kong Polytechnic University.

Author Contributions

Conceptualization, Yanghong Zhou and Hao Tian; methodology, Yanghong Zhou and Hao Tian; validation, Yanghong Zhou, Hao Tian and Yang Chen; formal analysis, Yanghong Zhou; investigation, Yanghong Zhou; data curation, Yanghong Zhou and P. Y. Mok; writing—original draft preparation, Yanghong Zhou; writing—review and editing, P. Y. Mok; visualization, Yanghong Zhou and Yang Chen; supervision, P. Y. Mok; project administration, P. Y. Mok; funding acquisition, P. Y. Mok. All authors have read and agreed to the published version of the manuscript.

Conflict of Interest

All the authors declare that they have no conflict of interest.

Data Available

The dataset used in this study was constructed by the authors and is available upon reasonable request.

References

[1] Wang, X., Zhang, T., Tretter, D.R., Lin, Q.: Personal clothing retrieval on photo collections by color and attributes. IEEE transactions on multimedia **15**(8), 2035–2045 (2013). https://doi.org/10.1109/TMM.2013.2279658

[2] Liu, S., Song, Z., Wang, M., Xu, C., Lu, H., Yan, S.: Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In Proceedings of the 20th ACM International Conference on Multimedia, pp. 1335–1336 (2012). https://doi.org/10.1109/CVPR.2012.6248071

[3] Chen, Q., Huang, J., Feris, R., Brown, L.M., Dong, J., Yan, S.: Deep domain adaptation for describing people based on fine-grained clothing attributes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5315–5324 (2015). https://doi.org/10.1109/CVPR.2015.7299169

[4] Huang, J., Feris, R.S., Chen, Q., Yan, S.: Cross-domain image retrieval with a dual attribute-aware ranking network. In Proceedings of the IEEE International Conference on Computer Vision, pp. 1062–1070 (2015). https://doi.org/10.1109/ICCV.2015.127

[5] Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1096–1104 (2016). https://doi.org/10.1109/CVPR.2016.124

[6] Dong, Q., Gong, S., Zhu, X.: Multi-task curriculum transfer deep learning of clothing attributes. In 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 520–529 (2017). https://doi.org/10.1109/WACV.2017.64

[7] Ge, Y., Zhang, R., Wang, X., Tang, X., Luo, P.: Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5337–5345 (2019). https://doi.org/10.1109/CVPR.2019.00548

[8] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90

[9] Zou, X., Kong, X., Wong, W., Wang, C., Liu, Y., Cao, Y.: Fashionai: A hierarchical dataset for fashion understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 0–0 (2019). https://doi.org/10.1109/CVPRW.2019.00039

[10] Rostamzadeh, N., Hosseini, S., Boquet, T., Stokowiec, W., Zhang, Y., Jauvin, C., Pal, C.: Fashion-gen: The generative fashion dataset and challenge. arXiv preprint arXiv:1806.08317 (2018). https://doi.org/10.48550/arXiv.1806.08317

[11] Chen, Q., Li, J., Lu, G., Bi, X., Wang, B.: Clothing retrieval based on image bundled features. In 2012 IEEE 2nd International Conference on Cloud Computing and Intelligence Systems, pp. 980–984 (2012). https://doi.org/10.1109/CCIS.2012.6664323

[12] Abdulnabi, A.H., Wang, G., Lu, J., Jia, K.: Multi-task CNN model for attribute prediction. IEEE Transactions on Multimedia **17**(11), 1949–1959 (2015). https://doi.org/10.1109/TMM.2015.2477680

[13] Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE transactions on pattern analysis and machine intelligence **39**(6), 1137–1149 (2016). https://doi.org/10.1109/TPAMI.2016.2577031

[14] Sun, G.-L., Wu, X., Chen, H.-H., Peng, Q.: Clothing style recognition using fashion attribute detection. In Proceedings of the 8th International Conference on Mobile Multimedia Communications, pp. 145–148 (2015). https://doi.org/10.5555/2826112.2826142

[15] Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**(2), 91–110 (2004). https://doi.org/10.1023/B:VISI.0000029664.99615.94

[16] Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In 2005 IEEE Computer Society

Conference on Computer Vision and Pattern Recognition (CVPR'05), pp. 886–893 (2005). https://doi.org/10.1109/CVPR.2005.177

[17] Liaw, A., Wiener, M.: Classification and regression by randomForest. R news **2**(3), 18–22 (2002)

[18] Zhang, N., Donahue, J., Girshick, R., Darrell, T.: Part-based R-CNNs for fine-grained category detection. In European Conference on Computer Vision, pp. 834–849 (2014). https://doi.org/10.1007/978-3-319-10590-1_54

[19] Lin, D., Shen, X., Lu, C., Jia, J.: Deep lac: Deep localization, alignment and classification for fine-grained recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1666–1674 (2015). https://doi.org/10.1109/CVPR.2015.7298775

[20] Huang, S., Xu, Z., Tao, D., Zhang, Y.: Part-stacked CNN for fine-grained visual categorization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1173–1182 (2016). https://doi.org/10.1109/CVPR.2016.132

[21] Wei, X.-S., Xie, C.-W., Wu, J., Shen, C.: Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization. Pattern Recognition **76**, 704–714 (2018). https://doi.org/10.1016/j.patcog.2017.10.002

[22] Zhou, Y., Li, R., Zhou, Y., Mok, P.Y.: Describing clothing in human images: a parsing-pose integrated approach. In MCCSIS 2018 - Multi Conference on Computer Science and Information Systems; Proceedings of the International Conferences on Interfaces and Human Computer Interaction 2018, Game and Entertainment Technologies 2018 and Computer Graphics, Visualization, Computer Vision and Image Processing 2018, pp. 205–213 (2018)

[23] Xiao, T., Xu, Y., Yang, K., Zhang, J., Peng, Y., Zhang, Z.: The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 842–850 (2015). https://doi.org/10.1109/CVPR.2015.7298685

[24] Lin, T.-Y., RoyChowdhury, A., Maji, S.: Bilinear CNN models for fine-grained visual recognition. In Proceedings of the IEEE International Conference on Computer Vision, pp. 1449–1457 (2015). https://doi.org/10.1109/ICCV.2015.170

[25] Fu, J., Zheng, H., Mei, T.: Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4438–4446 (2017). https://doi.org/10.1109/CVPR.2017.476

[26] Zheng, H., Fu, J., Mei, T., Luo, J.: Learning multi-attention convolutional neural network for fine-grained image recognition. In Proceedings of the IEEE International Conference on Computer Vision, pp. 5209–5217 (2017). https://doi.org/10.1109/ICCV.2017.557

[27] Devlin, J.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018). https://doi.org/10.48550/arXiv.1810.04805

[28] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., *et al.*: Learning transferable visual models from natural language supervision. In International Conference on Machine Learning, pp. 8748–8763 (2021)

[29] Gao, D., Jin, L., Chen, B., Qiu, M., Li, P., Wei, Y., Hu, Y., Wang, H.: FashionBERT: Text and Image Matching with Adaptive Loss for Cross-modal Retrieval. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2251–2260 (2020). https://doi.org/10.1145/3397271.3401430

[30] Zhuge, M., Gao, D., Fan, D.-P., Jin, L., Chen, B., Zhou, H., Qiu, M., Shao, L.: Kaleido-bert: Vision-language pre-training on fashion domain. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12647–12657 (2021). https://doi.org/10.1109/CVPR46437.2021.01246

[31] Ma, H., Zhao, H., Lin, Z., Kale, A., Wang, Z., Yu, T., Gu, J., Choudhary, S., Xie, X.: Ei-clip: Entity-aware interventional contrastive learning for e-commerce cross-modal retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18051–18061 (2022). https://doi.org/10.1109/CVPR52688.2022.01752

[32] Han, X., Zhu, X., Yu, L., Zhang, L., Song, Y.-Z., Xiang, T.: Fame-vil: Multi-tasking vision-language model for heterogeneous fashion tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2669–2680 (2023). https://doi.org/10.1109/CVPR52729.2023.00262

[33] Zhang, R., Han, J., Liu, C., Zhou, A., Lu, P., Qiao, Y., Li, H., Gao, P.: LLaMA-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In The Twelfth International Conference on Learning Representations, pp. 1–30 (2024)

[34] Luo, G., Zhou, Y., Ren, T., Chen, S., Sun, X., Ji, R.: Cheap and quick: Efficient vision-language instruction tuning for large language models. In Proceedings of the 37th International Conference on Neural Information Processing Systems, pp. 29615–29627 (2023). https://doi.org/10.5555/3666122.3667410

[35] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., *et al.*: Chain-of-thought prompting elicits reasoning in large language models. In Proceedings of the 36th International Conference on Neural Information Processing Systems, pp. 24824–24837 (2022). https://doi.org/10.5555/3600270.3602070

[36] Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., Smola, A.: Multimodal chain-of-thought reasoning in language models. arXiv preprint arXiv:2302.00923 (2023). https://doi.org/10.48550/arXiv.2302.00923

[37] Wu, M., Liu, Z., Yan, Y., Li, X., Yu, S., Zeng, Z., Gu, Y., Yu, G.: RankCoT: Refining Knowledge for Retrieval-Augmented Generation through Ranking Chain-of-Thoughts. arXiv preprint arXiv:2502.17888 (2025). https://doi.org/10.48550/arXiv.2502.17888

[38] Niu, T., Joty, S., Liu, Y., Xiong, C., Zhou, Y., Yavuz, S.: Judgerank: Leveraging large language models for reasoning-intensive reranking. arXiv preprint arXiv:2411.00142 (2024). https://doi.org/10.48550/arXiv.2411.00142

[39] Jedidi, N., Chuang, Y.-S., Glass, J., Lin, J.: Don't" Overthink" Passage Reranking: Is Reasoning Truly Necessary? arXiv preprint arXiv:2505.16886 (2025). https://doi.org/10.48550/arXiv.2505.16886

[40] Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318 (2002). https://doi.org/10.3115/1073083.1073135

[41] Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the Acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation And/or Summarization, pp. 65–72 (2005). https://doi.org/10.5555/1626355.1626389

[42] Lin, C.-Y.: Rouge: A package for automatic evaluation of summaries. In Annual Meeting of the Association for Computational Linguistics, pp. 74–81 (2004)

[43] Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4566–4575 (2015). https://doi.org/10.1109/CVPR.2015.7299087

A CoT Reasoning Prompt

To ensure the MLLM generates consistent and domain-specific descriptions, we developed a hierarchical system prompt based on a predefined Fashion Schema. As detailed in Appendix A1, this prompt employs a Chain-of-Thought (CoT) strategy , directing MLLM, Qwen2.5, to perform structured attribute extraction before synthesizing a final caption. Furthermore, a reverse validation mechanism is integrated to cross-check inferences across diverse image views, ensuring the generated labels are accurate and visually grounded.

CoT Reasoning Prompt for Structured Text Description Generator based MLLM

You are given {image_count} images for single product, which include full body image, part body image, product only image, and details image (not all available). Its category is {category}, and its fine-grained category is {fine_grained}. Please follow the steps below:

1. Extract and Analyze Clothing Attributes
For the main clothing item only, carefully examine all images and answer the following questions step by step, ignoring any other clothing items, outfits, or styling elements present in the images. Focus exclusively on the attributes of the single main clothing item. Each attribute must have a value. Please fully utilize the characteristics of these four views (not all available) of full body image, part body image, product only image, and details image, and give higher reference weights to views that can more accurately analyze corresponding attributes.
The following constraints must be followed:

- **Type**: Full-body clothing, Upper clothing, Bottom clothing, or Accessories & Shoes.
- **Silhouette**: What is the silhouette or shape? (e.g., A-line, H-shape, etc.)
- **Color**: What is the main color? (e.g., gray, red, blue, pink, etc.)
- **Pattern**: Is there any pattern? If so, what kind? (e.g., solid, small patterned, etc.)
- **Texture**: What is the texture or fabric? (e.g., silk, wool, cotton, etc.)
- **Style**: What is the overall style? (e.g., casual, formal, romantic, etc.)
- **Occasion**: What occasion is this clothing suitable for? (e.g., date, party, sport, etc.)
- **Other details**: Are there any other notable features? (e.g., V-neck, long sleeves, etc.)

If you observe other obvious attributes or details, please add them.

2. Generate a Comprehensive Textual Description
Based on the above attributes, construct a single sentence that contains all identified features and details of the main clothing item only, ensuring the description covers attributes consistent across all images while excluding any references to other clothing items, outfits, or styling pairings.

3. Reverse Validation of Attributes and Description
Check whether the summarized attributes and the generated description perfectly cover each image provided, focusing solely on the main clothing item.

- If the attributes and description accurately represent the main item in every image, proceed.
- If there is any uncertainty or mismatch, include the word 'Uncertainty' in the Description field.

4. Return as JSON Format
Output all attributes, the lists of full-body images, and the final description in the following JSON format:

```
{
  "Type": "",
  "Silhouette": "",
  "Color": "",
  "Pattern": "",
  "Texture": "",
  "Style": "",
  "Occasion": "",
  "Other details": "",
  "Description": ""
}
```

Figure A1: The detailed CoT reasoning prompt designed for the MLLM to perform hierarchical fashion attribute extraction, structured text description and reverse validation.