공학석사학위논문

Image Captioning and Item Retrieval with Vision-Language Models and LLM-Generated Data in Fashion Domain

비전-언어 모델과 LLM 생성 데이터를 활용한 패션 도메인에서의 이미지 캡셔닝 및 아이템 검색

2025 년 2 월

서울대학교 대학원

산업공학과

정 철 환

Image Captioning and Item Retrieval with Vision-Language Models and LLM-Generated Data in Fashion Domain

비전-언어 모델과 LLM 생성 데이터를 활용한 패션 도메인에서의 이미지 캡셔닝 및 아이템 검색

지도교수 조 성 준

이 논문을 공학석사 학위논문으로 제출함

2024 년 11 월

서울대학교 대학원

산업공학과

정 철 환

정철환의 공학석사 학위논문을 인준함

2024 년 12 월

위 원 장 ______ 장 우 진 ______(인)

부위원장 ______ 조 성 준 ______(인)

위 원 ______ 이 성 주 ______(인)

Abstract

Image Captioning and Item Retrieval with Vision-Language Models and LLM-Generated Data in Fashion Domain

Cholhwan Jung

Department of Industrial Engineering

The Graduate School

Seoul National University

The recent advancement of vision-language models (VLMs) aligning vision and language capabilities of deep learning models has created new opportunities for domain-specific applications, such as image captioning and item retrieval, especially in fashion domain. However, general-purpose VLMs often struggle to handle fine-grained attributes in fashion domain, such as fabric textures, patterns, colors of clothes. This study addresses these problems by fine-tuning BLIP-2, recent VLM, with domain specific dataset.

We leverage BLIP-2 to jointly train image captioning and item retrieval ability to handle both task with same model and dataset at once. Also, we explore the impact of an auxiliary dataset generated by LLM, which has richer and detailed fashion attributes compared to publicly available datasets. The experimental results showed that public fashion captioning dataset with LLM-generated data, the model can achieve good performance in image captioning and item retrieval at the same time.

Our custom dataset significantly improved item retrieval tasks by providing fine-grained fashion attributes, outperforming retrieval benchmarks. Through single VLM architecture and creating complementary dataset without human annotation, we could successfully minimize cost and resources to attain high quality data while maintaining and surpassing performance of existing approaches.

Keywords: Vision-Language Models, Image Captioning, Item Retrieval, Fashion Domain Adaptation.
Student Number: 2023-26507

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The fashion industry is one of the largest and most dynamic sectors in the global economy, encompassing a wide array of activities ranging from design and production to marketing and retail. As the need of digital transformation and automation in fashion industry increases, the ability to process and analyze fashion-related visual data has become more important.

General vision-language models (VLMs) [21, 13, 15] have shown their strong multimodal capabilities, including image-text alignment, text generation. However, they often fail to maintain their performance in special areas like fashion domain due to the domain-specific features of clothing images. To bridge this gap, domain adaptation must be employed to make models to effectively understand and generate insights specific to fashion.

Understanding fashion images and their features differs significantly from understanding general image at some points. Fashion images often require fine-grained understanding of details such as fabric textures, patterns, colors, and styles. Also, model should understand context of fashion items, like how these items are worn, to consistently identify them regardless of posture of a model. In other words, understanding representation, or similarity between general images is a different task with understanding those in clothing images. This complexity makes domain adaptation to fashion difficult for vision-language models, which are pretrained on general datasets and lack the specificity required to perform well in the fashion domain.

This study explores the adaptation of recent VLM, BLIP-2 [13] to the fashion domain. By fine-tuning BLIP-2 using fashion domain-specific, this research aims to enhance the model's ability to generate detailed captions and perform accurate item retrieval.

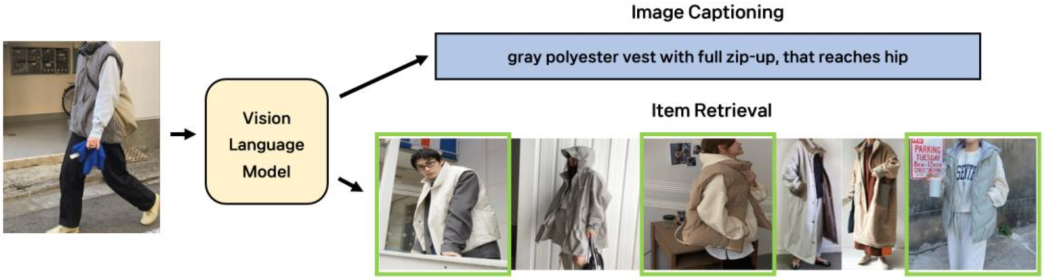


Figure 1.1: Image captioning and item retrieval with single VLM architecture.

1.1 Problem Description

The ability to effectively generate descriptive captions and retrieve relevant items is vital for advancing digital applications in the fashion industry, such as e-commerce and social media platforms. Historically, these tasks were tackled separately, often requiring large amounts of annotated data and relying on task-specific models. For instance, image captioning typically utilized encoder-decoder methods [2, 18], while item retrieval depended heavily on CNN structure that learns from annotated data such as bounding box, landmarks and segmentations of images [6, 16]. These approaches faced significant challenges, especially when applied to fashion images.

Also, fashion images are inherently complex, often requiring fine-grained recognition of attributes such as fabric texture, color, patterns, and stylistic details. Unlike general images, where objects and relationships are more straightforward, fashion imagery demands a deeper understanding of nuanced visual elements. Traditional models struggled with these challenges, as they lacked the multimodal capability to bridge the gap between visual features and textual descriptions effectively. Image captioning models require large amount of text data which is annotated to images, while item retrieval model require another large amount and human annotated data which are time-consuming and expensive to obtain.

The emergence of VLMs offers new opportunities to address these challenges by unifying image captioning and item retrieval tasks. Through single VLM architecture, we minimize cost and resources to attain high quality data to train image captioning and item retrieval ability. However, we cannot expect model to achieve same performance when applying general-purpose VLMs directly to the fashion domain due to the lack of domain-specific fine-tuning. This limitation underscores the need for tailored approaches that adapt VLMs to the unique demands of fashion, enabling better alignment between visual and textual representations.

1.2 Research Motivation and Contribution

This research aims to address the need for improved performance in fashion image captioning and item retrieval by adapting a recent vision-language model BLIP-2, specifically for the fashion domain. We also point out existing fashion image-text datasets often face inadequate images for real-world scenario and have lack of ability to accomplish successful image captioning and item retrieval task. By fine-tuning the model with both publicly available datasets and a custom-built auxiliary dataset that includes rich fashion-specific annotations, this work seeks to bridge the gap between general vision-language understanding and the specialized requirements of the fashion industry. Shown as Figure 1.1, through single VLM model and auxiliary dataset, we could handle image captioning and item retrieval at the same time.

The key motivations and contributions of this thesis are summarized as follows:

(a) The adaptation and fine-tuning of vision-language model BLIP-2, to the fashion domain.

(b) The development of LLM-generated custom dataset with detailed fashion-specific annotations to enhance model performance.

(c) Jointly training image captioning and item retrieval ability, with single model architecture and same dataset.

1.3 Organization of the Thesis

This thesis is organized into five chapters. Chapter 2 provides a literature review covering recent developments in vision-language models, approaches of image captioning and item retrieval, and approaches of leveraging LLM to replace human annotations and create datasets. Chapter 3 outlines the proposed methodology, detailing the model architecture of BLIP-2, the training detail, datasets. In Chapter 4, we present experimental settings, results, analyze the findings, and discuss the industrial applications of our model. Finally, Chapter 5 concludes the thesis with a summary of key contributions and potential directions for future work.

Chapter 2

Literature Review

2.1 Vision-Language Models

VLMs have emerged as powerful tools for integrating visual and textual data, enabling a wide range of applications such as image captioning, item retrieval, and multimodal understanding. One of early models, CLIP [21] had an approach to align image and text data by training large amount of image-text pair with contrastive learning. By leveraging millions of image captions from the internet, CLIP successfully learned high quality image-text representations and showed good performance in image-text matching task and zero-shot classification. However, the model is only trained for aligning image and text, it has lack of ability for generating text, task like image captioning. Also, model showed relatively low performance in zero-shot classification in specific domain such as fashion, as it was trained by general domain training data.

VLMs after CLIP have tried to utilize ability of both vision transformer (ViT) and LLMs, to perform further multimodal tasks. This approach made recent models to achieve capability of image captioning, as well as matching image and text. BLIP-2 [13] takes multimodal learning a step further by efficiently connecting ViT with LLMs such as OPT, FlanT5 using an intermediary module called querying transformer (Q-Former). This lightweight transformer-based module bridges the gap between visual and textual modalities, enabling effective alignment between ViT features and the input space of LLMs. With leveraging text generation ability of LLM, the model can perform image captioning and simple visual question answering. Q-former is known as efficient module to connect ViT and LLM while reducing number of training parameters, this concept is widely adapted to models after BLIP-2 [25]. Also, the model is trained with 2-stage training phase, which focus on image-text contrastive learning and text generation of its language model, therefore it is adequate architecture to handle both image captioning and item retrieval at once.

Other approaches to align ViT and LLM also exist. LLaVA [15] represents an alternative approach to vision-language integration. Unlike intermediary module like BLIP-2's Q-Former, LLaVA uses a simple linear projection layer to map vision transformer features directly to the LLM input space. The model is trained on a diverse instruction-following datasets, enabling the model to perform well in tasks that require an understanding of multimodal instructions. While LLaVA takes advantage to understand user instruction and relatively simple architecture that BLIP-2, its reliance on extensive instruction-tuning datasets which hard to attain in specific domain and its image-text matching ability is not yet proven.

While these recent VLMs aim general-purpose tasks, its performance in domain-specific applications is limited. They use train data in general domain, often show relatively poor performance in specific domains. Fashion domain is one of them, FashionCLIP [4] tried domain adaptation of CLIP by fine-tuning the model with clothing image-text pair from the internet. However, CLIP based model works only in multimodal matching tasks, and moreover FashionCLIP performs low performance in out-of-distribution real-world data.

2.2 Fashion Image Captioning and Item Retrieval

Image captioning and item retrieval are two critical tasks in the fashion domain, each addressing distinct challenges but they have been developed as separate streams of research. These tasks have relied on specialized approaches and architectures designed to excel at one while often being inadequate for the other.

Fashion image captioning focuses on generating descriptive text that encapsulates the key visual and contextual features of fashion items. These captions typically highlight details such as color, material, style, and occasion, aiming to enhance searchability and user experience in e-commerce platforms. General approaches for image captioning with domain specific dataset are used in this task including CNN-based encoder, RNN-based decoder, and models using RL [2, 8, 18, 23].

Item retrieval involves identifying and matching specific fashion items from a database based on a given input, which could be an image or a textual description. This task has traditionally relied on CNN architectures to extract features from images and match them to embeddings of textual queries or other images [6, 9, 12, 16]. While this approach has proven effective for this purpose, they come with notable challenges. These methods require substantial annotated datasets to learn robust embeddings and often fail to generalize well to unseen data.

The separate development of retrieval systems also results in inefficiencies when addressing tasks that require both descriptive and retrieval capabilities simultaneously. Different learning objectives and architectures require different annotated data for fashion item images. Image captioning models require description of images to generate caption, and item retrieval models require annotated data such as bounding box, landmarks, segmentations and detail attributes of images. Both formats of annotated data are often expensive and hard to attain which makes development of models inefficient.

VLMs offer a promising solution to bridge the gap between image captioning and item retrieval. By leveraging shared multimodal embeddings, these models can handle both tasks within a single architecture, eliminating the need for separate systems. Advanced architectures like BLIP-2 demonstrate the potential to generate high-quality textual descriptions while also producing embeddings suitable for retrieval tasks. The dual capabilities of these models enable a more integrated approach to fashion applications, where image captioning and item retrieval often need to work in single architecture.

We compared performance of our proposed method with baseline models of each task, image captioning and item retrieval. In image captioning, SCNST, SRFC [23] are our baseline models. These models utilize CNN as image encoder and integrate training process with maximum likelihood estimation (MLE), and reinforcement learning. MRCNN [6] and FashionCLIP [4] are our baseline models for item retrieval. MRCNN is CNN-based architecture trained by annotated image data of bounding box, landmarks, segmentations and detail attributes. FashionCLIP is CLIP-based vision-language model, but it cannot perform image captioning. Its multimodal alignment ability was used to measure baseline performance of CLIP-based model in fashion item retrieval task.

2.3 LLM Generated Data

Though we leverage VLMs to minimize annotated data, it still takes cost to attain large amount, high quality training data. There are some approaches to handle this problem by using LLMs to generate or augment training data to reduce cost.

The advent of LLMs has impacted the landscape of data augmentation across various domains and tasks, for instance text classification in NLP. By leveraging LLMs, researchers can generate diverse and high-quality training data, reducing the dependency on expensive and time-consuming human annotations [3, 7].

Recent study [19] has shown that fine-tuning supervised classifiers with LLM-generated data yields comparable performance to models trained on human annotated data. For example, in computational social science, commercial LLMs like GPT-4 preserved classification performance across various dataset, compared to models that trained on human annotated data. This suggests that replacing or augmenting human labels with LLM-generated data can be a viable and efficient alternative.

While the use of LLM-generated data has been extensively explored in NLP, its application in vision-language domains is relatively unexplored. VLMs require data that effectively bridges the gap between visual and textual modalities. LLMs with vision capabilities, such as GPT models with image input, offer new opportunities to generate rich, fine-grained captions and annotations. For instance, LLMs can classify visual attributes from images and transform these classifications into detailed captions using predefined prompts.

In the context of these related works to our research, we utilize BLIP-2, which can perform image captioning and item retrieval in single architecture to minimize annotation cost to generate training data. We also test LLM generated data augmentation in vision-language tasks to complement existing annotated dataset and enhance the model performance.

Chapter 3

Method

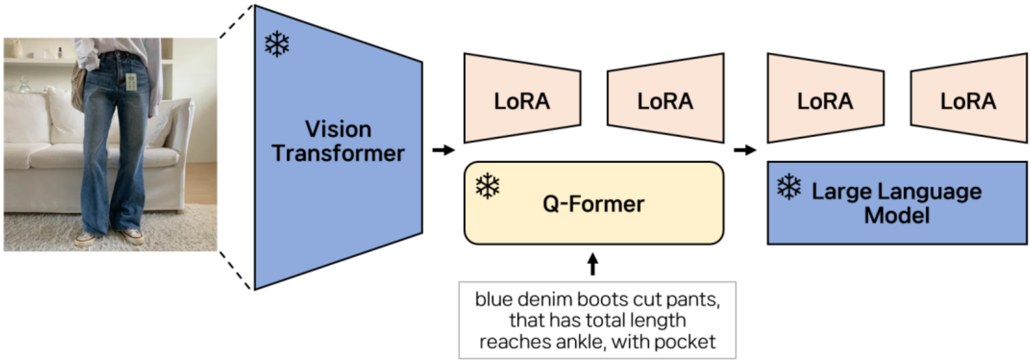

Vision Transformer
LoRA
LoRA
LoRA
LoRA
Q-Former
Large Language Model
blue denim boots cut pants, that has total length reaches ankle, with pocket

Figure 3.1: Fine-tuning using BLIP-2 with LoRA layer.

3.1 Model Architecture

BLIP-2 is designed to handle image-to-text generation tasks such as image captioning and simple visual question answering. The model integrates a frozen LLM as text decoder and ViT as image encoder. A key component of BLIP-2 is its intermediary called Q-Former, which acts as a bridge between ViT and LLM. The Q-Former employs learnable queries to represents visual features into a format that can be understood and processed by the LLM, addressing the modality gap between images and text.

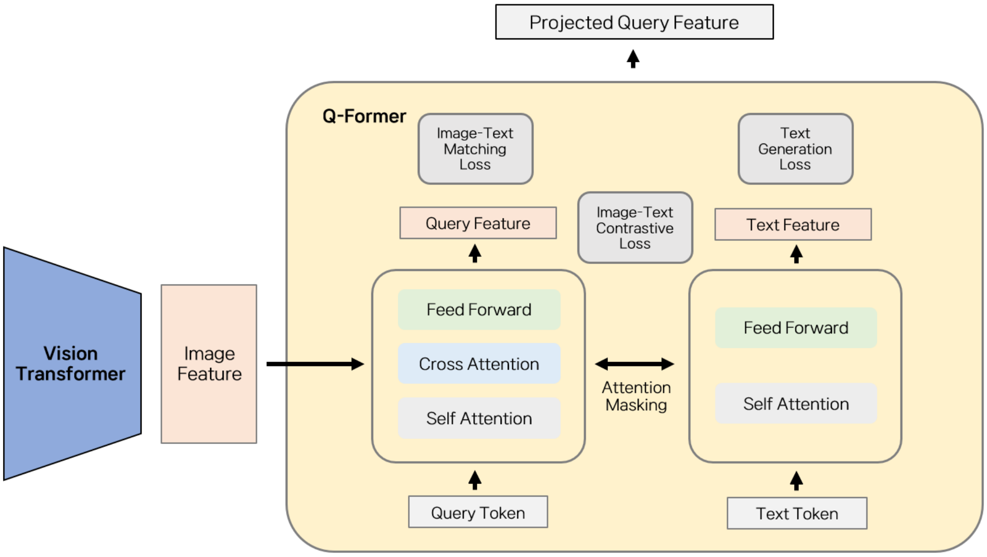


Figure 3.2: Architecture of Q-Former.

To adapt BLIP-2 to fashion image captioning and item retrieval, this work finetunes the model, leveraging the its abilities learned in pretraining stages. By optimizing the Q-Former and enhancing its alignment with the LLM, this fine-tuning approach ensures that the model can simultaneously learn both tasks. We aim to preserve the core strengths of BLIP-2 to align image-text data and generate robust text, also enhance its domain-specific applicability as well.

Figure 3.1 shows the overall architecture of BLIP-2 with LoRA layers. In the architecture, output of ViT, visual encoder, and text query is fed to Q-Former and Q-Former output is projected to language model input to generate final text output. Figure 3.2 shows inside architecture of Q-Former. This module is pretrained with 3 different loss, image-text matching loss, image-text contrastive loss, and text generation loss in the first stage of pretraining, hence attaining high ability of aligning image and text. After finetuning pretrained weight of BLIP-2 with fashion image-text pairs, multimodal features for understanding fashion concepts can be extracted from query features, output embedding of Q-Former. We used this output embeddings to conduct item retrieval task in further evaluation.

3.2 Training Detail

Fine-tuning BLIP-2 for domain-specific tasks like fashion image captioning and item retrieval can be computationally expensive in limited resources, despite the Q-Former's design for training efficiency. We applied LoRA [10], a widely adopted parameter-efficient fine-tuning method that trains only low-rank perturbations of selected weight matrices, leaving the majority of the pre-trained model parameters frozen. This approach significantly reduces memory requirements while minimizing catastrophic forgetting of the model's general-purpose capabilities [1].
The large-scale pretraining on general-purpose datasets gives the model strong multimodal capabilities, but domain-specific finetuning requires adapting these capabilities without eroding its general text generation performance. Although LoRA tends to underperform compared to full fine-tuning on domain-specific tasks, it excels in maintaining the base model's performance across tasks outside the target domain. This makes LoRA particularly suitable for fine-tuning in scenarios where retaining general text generation abilities is also important.

We adapted LoRA layer with rank size 8 to key and query of attention layers in Q-Former and language model. BLIP-2 is pretrained with random cropping and resizing augmentation, but we did not apply image augmentation to our training image data. We use the AdamW optimizer with β1 = 0.9, β2 = 0.98 as same as original BLIP-2, and a weight decay of 0.01. Input image resolution is 224×224.

3.3 Training Dataset

Fine-tuning the BLIP-2 model for the fashion domain requires diverse and high-quality datasets that capture the nuances of fashion concepts. To achieve this, we utilized two public datasets, DeepFashion-MultiModal [11] and FACAD [23]. Also complemented by a newly created auxiliary dataset generated by LLM with vision capabilities. These datasets together provide a comprehensive resource for training and adapting the model to the unique requirements of the fashion domain.

Public Dataset

DeepFashion-MultiModal (DFMM) dataset is a collection of fashion-related images annotated with rich multimodal information. It was proposed by [11], manually annotating image captions to images from DeepFashion [17] dataset. It includes 44K images with short image caption.

The Fashion Captioning Dataset (FACAD) contains 993K images which has longer caption and more fine-grained attributes for fashion items than existing general domain and fashion comain datasets. FACAD focused on make fluent captions for fashion items compared with MS COCO caption, which mainly describes human rather than clothes. For our training dataset, we excluded FACAD captions for accessories, such as hat, shoes, and jewelry to focus on clothes. Also due to imbalance with amount of data between two datasets, we sampled only 70K of FACAD data.

While both public datasets contain relatively rich and fine-grained image captions for fashion items, they still have limitations. First, both datasets can contain noisy annotations, such as incorrect labels or incomplete descriptions. Second, many images in both datasets feature model are taken on a white and clean background, which significantly differs from real-world scenarios such as natural social media settings. Third, fashion is inherently detailed, requiring comprehensive annotations to capture elements such as fabric type, texture, pattern, fit, and style. Captions in these public datasets often lack the level of detail, and consistency of describing important details which is necessary to represent all relevant fashion attributes comprehensively.

LLM-Generated Dataset

To supplement the limitations of existing datasets, we created a new dataset with fine-grained and domain-specific information. This dataset consists of 100K fashion images curated from Korean e-commerce platforms and social media, capturing diverse clothing styles and contexts. Fashion images of this dataset contain either image of single clothes without model wearing it or image of model wearing clothes in real life. This differs to the public datasets, as most of fashion images have strict pose of model with white background. Also, it features more current trend of fashion and clothes as DFMM dataset is made before 2016, and FACAD dataset is made before 2020.

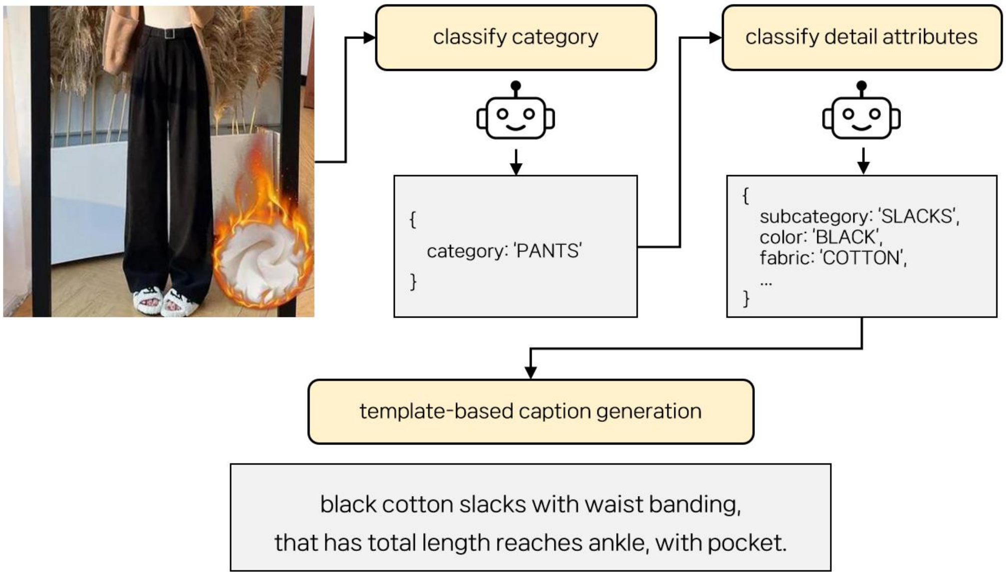

classify category
{
category: 'PANTS'
}
classify detail attributes
{
subcategory: 'SLACKS',
color: 'BLACK',
fabric: 'COTTON',
...
}
template-based caption generation
black cotton slacks with waist banding,
that has total length reaches ankle, with pocket.

Figure 3.3: Caption Generation Process of Custom Dataset.

Figure 3.3 shows the caption generation process of our new dataset. To generate fashion image caption for training data, we leveraged zero-shot image understanding ability of off-the-shelf LLM with vision capability, specifically gpt-4o-mini. Using a GPT model with vision capabilities, fashion attributes such as material, color, pattern, and fit were extracted with higher granularity compared to traditional benchmarks [6, 9, 17].

We first classified fashion attributes of given images by GPT zero-shot image classification. The process begins with categorizing the clothing item depicted in the image into one of the four primary categories: top, dress, pants or skirt. Initial classification establishes the foundation for applying category-specific prompts in subsequent classification for detailed attributes. Table 3.1 shows fashion attributes and possible candidates for classification results. Hierarchy between category and subcategory is show as Table 3.2. Classification result for fashion attributes including subcategory is used for caption generation.

Table 3.1: Structure of fashion attributes for classification.

Attribute	Candidates
category	top, dress, pants, skirt
color	red, orange, blue, navy, pink, brown, black, white, burgundy, purple, coral, mint, gray, green, khaki, yellow, beige
tone	normal, vivid, pastel/light, metallic/shiny
total length	belly, waist, hip, thigh, knee, shin, ankle, floor
sleeve length	sleeveless, short, long
fabric	cotton, cotton knit, cotton woven/tweed, seersucker, denim, corduroy/ribbed, wool knit, wool woven/tweed, polyester/nylon, linen, silk/chiffon, velvet/velveteen, leather, suede, fleece, fur, down
pattern	solid, dot, washed, check, cable, stripe, lettering, graphic, pattern/jacquard, argyle, camouflage, flower, checkerboard, animal, other
zip-up	none, half, full
button	none, half, full
neckline	round, square, v-line, sweetheart, slit, halter, turtle, boat
shoulder	normal, off-shoulder, one-shoulder
boolean attributes	pocket, hood, collar, see-through, waist banding, string, ribbon, ruffle, lace, pintuck

Table 3.2: Category-Subcategory hierarchy.

Category	Subcategory
top	t-shirt, cardigan, shirt, blouse, sweater, sweatshirt, sleeveless, vest, bomber jacket, varsity jacket, anorak jacket, blazer, down jacket, jacket, coat
dress	long dress, mini dress, shirt dress, overall dress, jumpsuit
pants	sweatpants, shorts, slacks, bermuda pants, cargo pants, overall pants, cigarette pants, boots cut pants, baggy pants, straight pants, harem pants
skirt	long skirt, flare skirt, peated skirt, mini skirt, balloon skirt, slit skirt

Figure 3.4 and Figure 3.5 show example prompts to classify category and detail attributes of given fashion items. Depends on first category classification result, we applied four different prompts to classify detail attributes. For example, if the given fashion item is classified to 'top', we should ask LLM to classify its subcategory within 't-shirt', 'cardigan', 'shirt', and so on, based on category-subcategory hierarchy in Table 3.2. This separated structure of prompts was developed to make understand our instructions more easily.

1 You are an agent specialized in classifying clothings.
2 You will be provided with multiple images and for each image you should classify the clothing is which of TOP, ONEPIECE for top, which of PANTS, SKIRT for bottom.
3 If it is an wearing image by a model and there are both top and bottom, you might classify both clothings based on instruction.

4 You should classify the category of clothings to extract further features based on its category in the future.
5 For example, if a model in the image is classified to be wearing TOP and PANTS, you will be asked to tag detail attributes each TOP and PANTS later with different prompts respectively.

6 Constraints of each attributes are as followings:
7 - category: category of the clothing. sublist of [TOP, ONEPIECE, PANTS, SKIRT]
8 - category is TOP when the clothing is one of [CARDIGAN, JACKET, COAT, BOMBERJACKET, VARSITYJACKET, DOWN JUMPER, T-SHIRT, SHIRT, KNIT, SLEEVELESS, HOODIE, SWEATSHIRT, BLOUSE, VEST]
...

15 You will be given multiple images, return attributes of each image in the JSON format :

16 Output:

Figure 3.4: Example of Category Classification Prompt.

1 You are an agent specialized in tagging attributes to clothings.
2 You will be provided with multiple images and for each image if it is an wearing image by a model, you should depict one of outer, top, dress, pants, and skirt depends on instruction.
3 Your goal is to classify and extract attributes only for the item specified by key and value.

4 You should depict TOP.
5 Keys of attributes and values of each key are instructed below.
6 Definition and constraints of each attributes are as followings:
7 - subcategory: subcategory of the clothing. one of subcategory_values
...

29 Additional constraints of values of attributes are as followings:
30 - color should be a list of colors of main textiles, ignore small details like graphic, lettering.
...

35 You will be given multiple images, return attributes of each image in the JSON format of key1: value1, key2: value2, ... in a list.
36 Images will be provided in batch, the length of the list must be equal to the number of provided images. Tagging result should be extracted only one result per image.
37 Now you should depict clothings of the given images, output should be a JSON format.
38 Make sure attributes are aligned to single clothing.
39 Output:

Figure 3.5: Example of Attribute Classification Prompt.

Then we applied template-based caption generation for classified fashion attributes. By template-based generation, we could generate caption which is compact and contains highly intensive fashion attribute text. Number of words in caption in public dataset and LLM generated dataset is shown as Figure 3.6. Average number of words in public dataset is 26.7, while LLM generated dataset is 17.4. Relatively short and compact captions in this dataset work as complementary feature to public dataset because too long and noisy caption might make model hard to learn useful data from text. Figure 3.7 shows example image and caption of each dataset. Our new dataset contains many real-world fashion item images with natural background, and contains compact captions for fashion items.

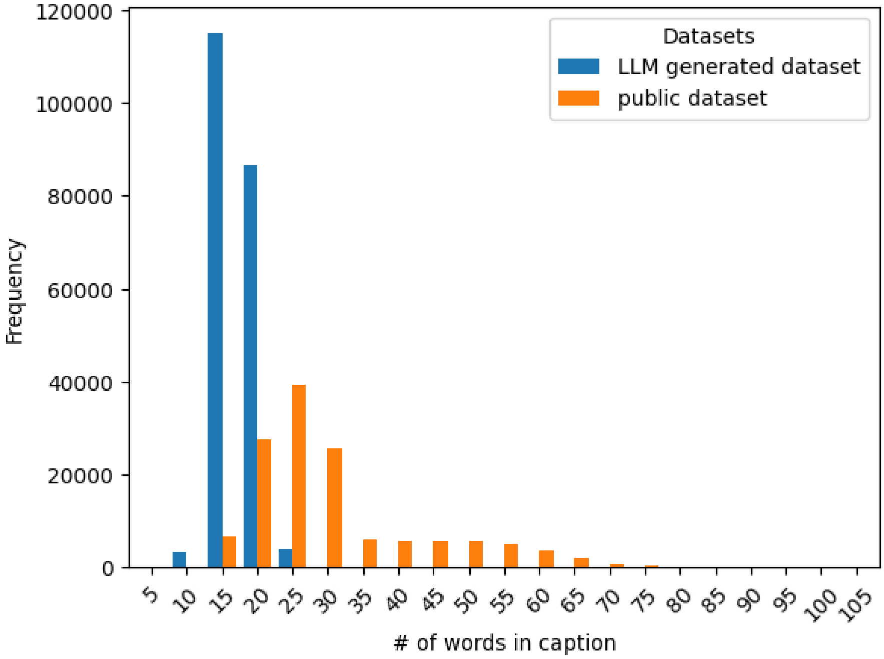
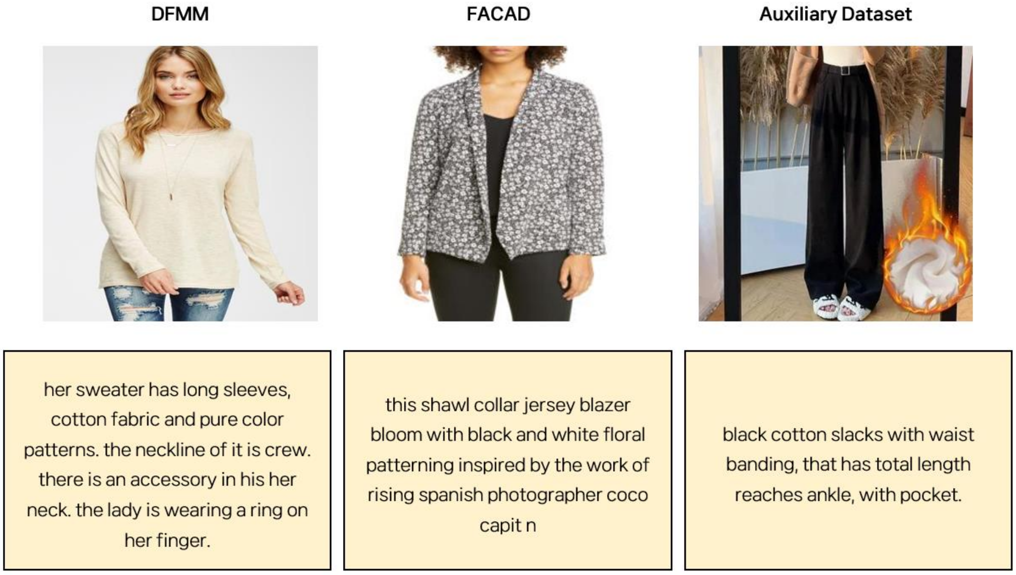


Figure 3.6: Number of words distribution in caption.

This dataset complements limitations of existing benchmarks by providing more granular and contemporary fashion attributes, particularly relevant for the dynamic nature of the fashion industry. In summary, by leveraging the public datasets, and the newly created LLM-generated dataset, we ensured that the training data provides both the diversity and specificity needed for fashion image captioning and item retrieval tasks. This combination of datasets helps the model to understand and generate fine-grained, engaging captions while being equipped to handle retrieval tasks effectively in the fashion domain.



her sweater has long sleeves, cotton fabric and pure color patterns. the neckline of it is crew. there is an accessory in his her neck. the lady is wearing a ring on her finger.

this shawl collar jersey blazer bloom with black and white floral patterning inspired by the work of rising spanish photographer coco capit n

black cotton slacks with waist banding, that has total length reaches ankle, with pocket.

Figure 3.7: Example image and caption in each dataset.

Chapter 4

Results and Discussion

4.1 Evaluation Metric and Experiment Setting

We evaluate the performance of the fine-tuned BLIP-2 in both image captioning and item retrieval. To evaluate the quality of generated captions, we use standard natural language generation metrics, BLEU [20], ROUGE [14], and METEOR [5]. These metrics are computed on the test set of the FACAD dataset.

BLEU score is an automatic metric for evaluating machine translation quality by comparing n-gram overlaps between the generated predictions and reference sentences. It is computed as Eq. 4.1, weighted sum of modified n-gram precision p_n, with brevity penalty BP. Modified precision is shown as Eq. 4.2, where C represent candidate sentences, P and R represent group of predicted sentences, and group of reference sentences respectively. In this study, we examine BLEU-4, where n is 4.

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{4.1}$$

$$p_n = \frac{\sum_{gram_n \in C, C \in P} Count_{clip}(gram_n)}{\sum_{gram_n \in C, C \in R} Count(gram_n)} \tag{4.2}$$

ROUGE score assesses captioning quality by measuring overlapping units between predicted sentences and reference sentences. We specifically compute ROUGE-L, score based on the Longest Common Subsequence (LCS) between predictions and references. Eq. 4.3 and Eq. 4.4 computes precision and recall score between predicted sentences and reference sentences with LCS, and final ROUGE-L score is computed as Eq. 4.5, as F1-score of two metrics. P, R represent predicted sentence and reference sentence respectively

$$Precision = \frac{LCS(P,R)}{|R|} \tag{4.3}$$

$$Recall = \frac{LCS(P,R)}{|P|} \tag{4.4}$$

$$ROUGE - L = \frac{2Precision \cdot Recall}{Precision + Recall} \tag{4.5}$$

METEOR score evaluates quality by aligning predictions at the word and phrase level using external linguistic resources, offering improved performance over BLEU score. When computing $Precision_M$ and $Recall_M$, external word mapping is used to count unigrams considering synonyms, which make the metric more human-aligned than BLEU score. Harmonic mean between METEOR recall and precision is computed as Eq. 4.6. Final score is computed as Eq. 4.7, with chunk penalty to evaluate prediction which has longer sequence matched higher.

$$Fmean = \frac{10Recall_M Precision_M}{Recall_M + 9Precision_M} \tag{4.6}$$

$$METEOR = Fmean(1 - Penalty) \tag{4.7}$$

Also, qualitative evaluation was made on fashion images in Korean social media, which is out-of-distribution to FACAD images. For retrieval tasks, we measure the model's ability to identify matching items across datasets using top-k accuracy. We used consumer-to-shop benchmark test data in DeepFashion, and DeepFashion2. Both benchmark data consist of query image from consumer wearing certain fashion items, and image taken from shop and retailer which should be retrieved. DeepFashion benchmark has single ground truth retrieval result per query image, while DeepFahion2 has possible multiple images to be retrieved. As same evaluation setting in DeepFashion2, retrieval result is correct when at least one of the ground truth result is in top-k prediction. Both benchmarks provide bounding box of fashion item, we also evaluate item retrieval result with bounding box to crop image and without bounding box to retrieve item by full image.

We utilized BLIP-2 with decoder-only language model OPT and experimented with model variants of different language model size and training dataset. Two sizes of its integrated language model were used; OPT-2.7b and OPT-6.7b. To assess the impact of dataset composition, we conducted experiments using two different training dataset settings; training only public dataset without LLM-generated dataset and training with this dataset.

4.2 Experimental Results

Table 4.1: Overall result of model variants.

Model/Dataset	Image Captioning			Item Retrieval			
	BLEU-4	ROGUEL	METEOR	DF/c	DF/f	DF2/c	DF2/f
OPT-2.7b w/o Generated	1.8	18.4	17.1	0.158	0.101	0.359	0.207
OPT-2.7b w/ Generated	1.4	15.9	14.4	0.431	0.310	0.697	0.532
OPT-6.7b w/o Generated	2.0	18.7	17.2	0.081	0.053	0.177	0.097
OPT-6.7b w/ Generated	1.6	16.5	14.7	0.211	0.148	0.464	0.293

Overall results of model variants are shown as Table 4.1. DF/c, DF/f, DF2/c, DF2/f are respectively DeepFashion, DeepFashion2 benchmark with crop/full image settings. In image captioning both OPT-2.7b and OPT-6.7b models achieved better scores in BLEU-4 [20], ROGUE-L [14], and METEOR [5] metrics when trained only on public datasets. The increase of language model size of BLIP-2 made slight improvement of image captioning performance. The inclusion of the auxiliary dataset led to a slight decrease in performance for image captioning across both model sizes. This indicates through LLM-generated data, model could learn and understand more diverse attribute descriptions, generate captions closely matching exact captions of the benchmark dataset.

In item retrieval, models trained on public datasets with LLM-generated dataset showed significantly better performance. This indicates that the public datasets lacked sufficient fine-grained details to effectively capture meaningful fashion attributes. The addition of new custom data improved item retrieval performance substantially across all test sets, DeepFashion and DeepFashion2 with both crop and full image settings. The models benefited from the richer and more fine-grained fashion attributes provided by the LLM-generated dataset, enabling better alignment and representation in retrieval tasks. The improvements were particularly notable for the smaller OPT-2.7b model.

The smaller OPT-2.7b model consistently outperformed the larger OPT-6.7B model in item retrieval tasks, significantly when trained with LLM-generated data. This suggests that the smaller model may adapt better to the focused domain-specific dataset and there is a trade-off between image captioning and item retrieval performance.

The larger OPT-6.7b model showed slightly better performance in image captioning tasks when trained without the LLM-generated dataset, but considering this work aim to handle generation task and retrieval task at once, OPT-2.7b with the LLM-generated dataset would be the best model among the variants.

Table 4.2: Comparison with image captioning benchmark.

Model	Method	Training Dataset	BLEU	ROGUEL	METEOR
OPT-2.7b	Full-Finetune	DFMM 44K + FACAD 70K	5.9	19.9	14.4
	LoRA		1.8	18.4	17.1
	Pretrained	-	0.2	14.0	8.8
SCNST	-	FACAD 993K	6.1	22.5	12.3
SRFC			6.8	24.2	13.2

In comparison with image captioning benchmark, we also evaluated zero-shot performance of pretrained BLIP-2 without finetuning and full-finetuned BLIP-2 with public dataset, as using only public dataset performed best in image captioning task. Benchmark model SCNST, SRFC [23] was trained on whole FACAD data, hence scoring high BLEU and ROGUE metric. We finetuned BLIP-2 on training data that contains only 70K of FACAD, which is originally has 993K pair of data and made considerable improvements on METEOR. Pretrained base BLIP-2 scored low performance in BLEU and METEOR, and full-finetuned model scored 5.9 BLEU score, which indicates it succeed to make exact sequence of words as FACAD captions

Finetuned model with LoRA layer scored relatively high ROUGE, METEOR score than BLEU score, which indicates model could generate semantically right caption rather than just generating exactly matched words of training data. Considering we used only 70K image-text pairs of 993K total training data, BLIP-2 demonstrates powerful domain adaptation ability in image captioning task.

Table 4.3: Comparison in DeepFashion2 item retrieval benchmark.

Model	Training Data	Data Format	top-1	top-10	top-20
MRCNN	DeepFashion2	annotated images	0.268	0.574	0.665
FCLIP	Farfetch 800K	image-text pairs	0.162	0.364	0.443
Ours	Public/Generated 214K	image-text pairs	0.297	0.601	0.691

In item retrieval task, we compared our model, OPT-2.7b trained on public and LLM-generated dataset, with other fashion item retrieval benchmark model MRCNN and FashionCLIP in DeepFashion2 consumer-to-shop test data with full image setting. MRCNN was trained on DeepFashion2 390K annotated images, which provide bounding box, segmentation, landmarks and attributes of each item and image. Even though our method did not use heavily annotated images but only used image-text pair data, we outperformed baseline MRCNN model with 0.029, 0.027, 0.026 in top-1, top-10, and top-20 respectively.

FashionCLIP was trained on Farfetch 800K image-text pairs, it shows relatively lower performance, with a top-1 accuracy of 0.162, top-10 of 0.364, and top-20 of 0.443. The gap in performance could be derived by two aspects. As our method is based on BLIP-2 architecture and outperformed CLIP-based FashionCLIP, it indicates that multimodal alignment ability and image-text matching performance of BLIP-2 is better than simple CLIP architecture due to efficient intermediary module Q-Former. Also, the augmentation of our custom dataset enabled the model to achieve outstanding retrieval ability with only total 214K of image-text data, which is far less than 800K web-crawled data from Farfetch. We want to note that Farfetch web-crawled data has very similar image distribution with our public dataset, such as white background and clean item images, which might be hard to fit in consumer-to-shop retrieval scenario.

To summarize, smaller OPT-2.7b model is better suited for retrieval tasks, while the larger OPT-6.7b model shows marginal advantages in captioning. LLM-generated dataset improved item retrieval performance significantly, outperforming baseline model with heavily annotated data with relatively few training data. BLIP-

2 outperforms competing other VLM architectures like CLIP-based models and exhibits strong domain adaptability with limited data. However, there was a trade-off between image captioning and item retrieval performance in our model variants.

4.3 Qualitative Evaluation

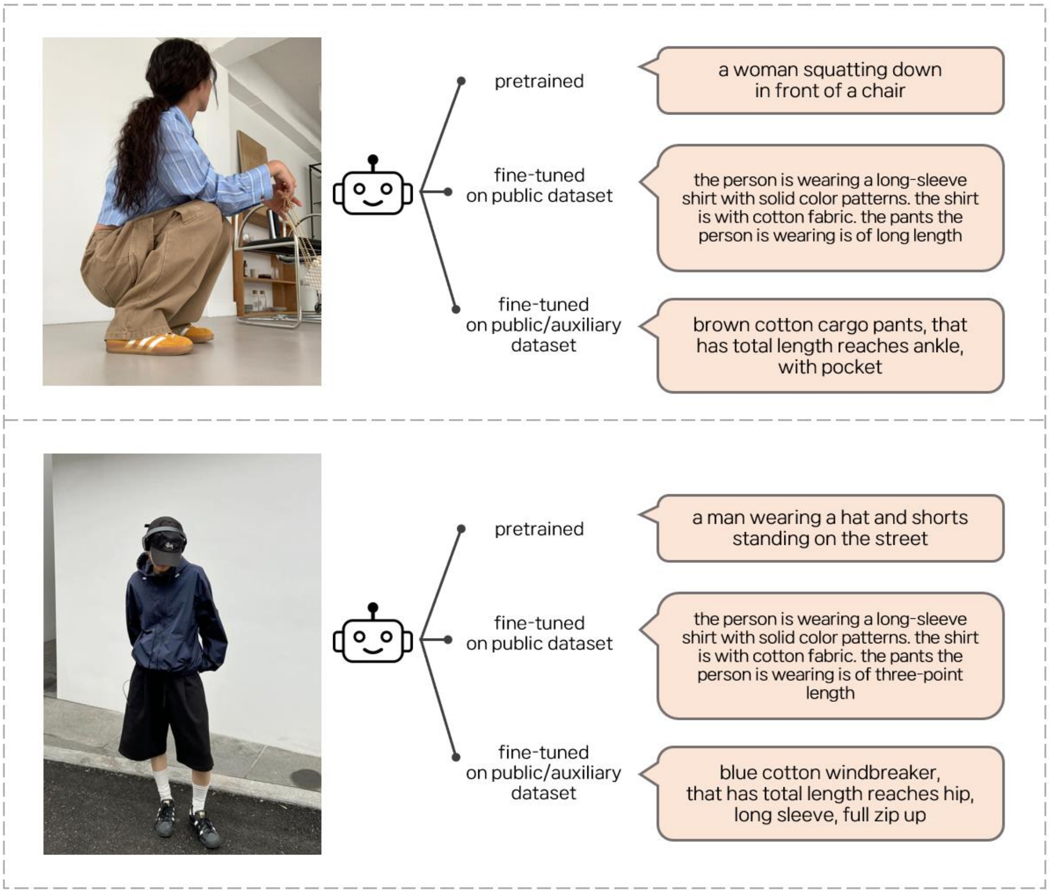

pretrained
a woman squatting down in front of a chair
fine-tuned on public dataset
the person is wearing a long-sleeve shirt with solid color patterns. the shirt is with cotton fabric. the pants the person is wearing is of long length
fine-tuned on public/auxiliary dataset
brown cotton cargo pants, that has total length reaches ankle, with pocket
pretrained
a man wearing a hat and shorts standing on the street
fine-tuned on public dataset
the person is wearing a long-sleeve shirt with solid color patterns. the shirt is with cotton fabric. the pants the person is wearing is of three-point length
fine-tuned on public/auxiliary dataset
blue cotton windbreaker, that has total length reaches hip, long sleeve, full zip up

Figure 4.1: Image captioning of BLIP-2 in different settings.

A qualitative evaluation was also conducted using random fashion images from Korean e-commerce platforms and social media, which is closer to real-world scenario. Figure 4.1 shows BLIP-2 captioning ability to out-of-distribution data in pretrained, finetuned on only public dataset, finetuned on public data with auxiliary dataset settings.

Pretrained BLIP-2 has robust ability to caption given images but it is trained on general caption data, generating captions focused on general facts about human in the given image. As shown in the figure, caption 'a woman squatting down in front of a chair', and 'a man wearing a hat and shorts standing on the street' do not give enough fashion related description from the images. BLIP-2 finetuned on only public dataset failed to generate proper caption, as generating almost same caption in different out-of-distribution images. Two captions from given images are fashion related, but as looking closely, phrase like 'the person is wearing a long-sleeve shirt' is shown in both image, which is actually incorrect information. Finetuned on public .data with LLM-generated dataset could generate reasonable caption in each image, but due to template-based training data, the output captions follow same template scheme.

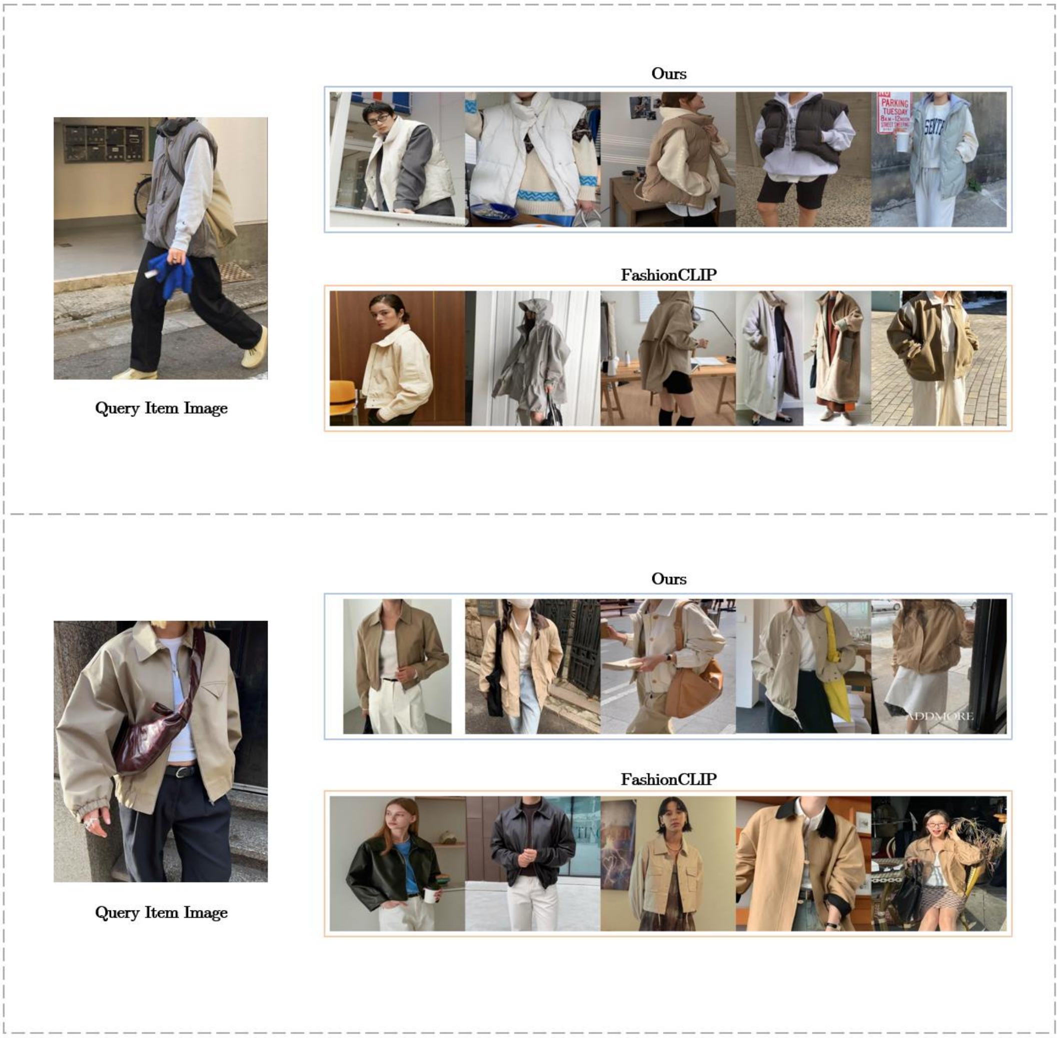


Figure 4.2: Example 1 of Retrieval Results comparing to FashionCLIP.

Figure 4.2, 4.3 and 4.4 shows item retrieval result of FashionCLIP and ours in custom consumer-to-shop retrieval settings. Query images are gathered from Korean social media and images to be retrieved are from fashion e-Commerce clothes images, which is about 50K in total. In retrieval result, our model has better understanding of general fashion concepts than FashionCLIP, while it confuses category or other detail of fashion item in given image when model is not in simple posture or style is more complicated than training data.

In the first case of Figure 4.2 our model could retrieve clothes category of vest consistently as LLM-generated data contain caption with vest category. In contrast, FashionCLIP failed to achieve this result due to its image distribution of training data and inconsistent fashion attributes in text data. Following cases also show the gap between ability to understand fashion concepts of two models. Our model could consistently retrieve similar items in terms of category, color, texture and other attributes.

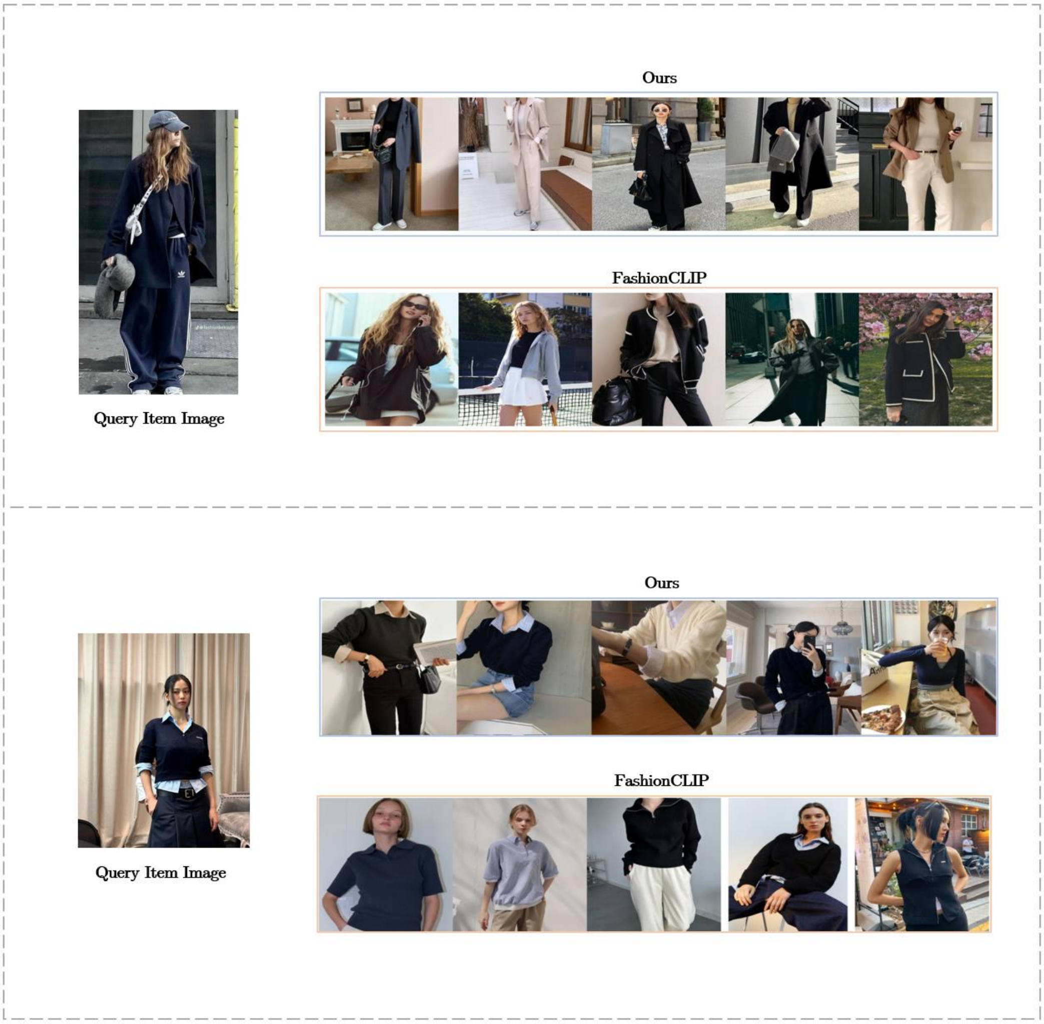


Figure 4.3: Example 2 of Retrieval Results comparing to FashionCLIP.

Figure 4.3 also shows different retrieval results between our model and FashionCLIP. For each image of black coat and black sweater with layered shirt inside, our model could retrieve more relevant and consistent category and details such as neckline and collar. Small mismatch of color appeared in retrieval results in both model.

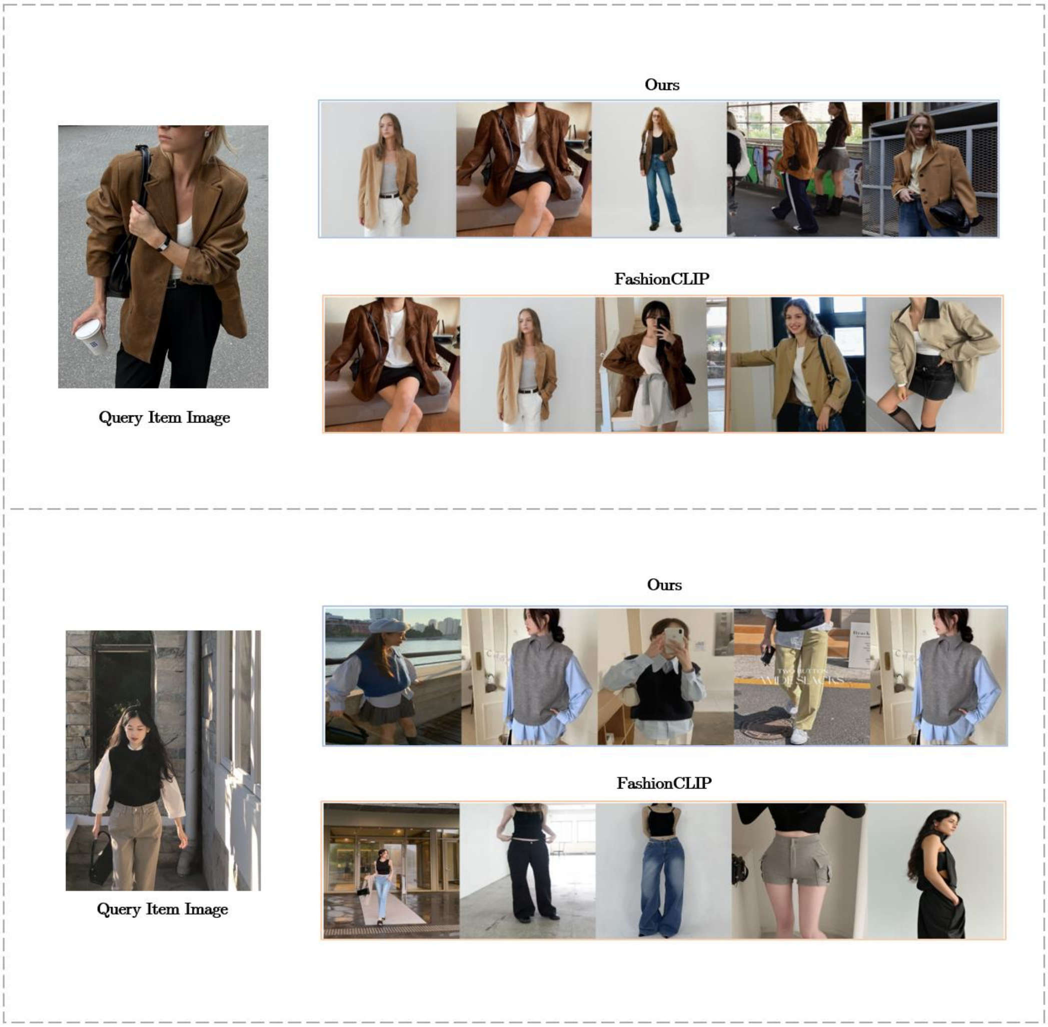


Figure 4.4: Example 3 of Retrieval Results comparing to FashionCLIP.

Figure 4.4 shows query images of brown jacket and black vest with inside white shirt. Our model made better retrieval result in color in the first case, searching jackets that have relatively similar level of brown colors. Also the model successfully retrieved fashion item with vest category, while FashionCLIP searched items among sleeveless tops.

We have evaluated qualitative results in image captioning and item retrieval out of training data distribution in this section. Our LLM-generated dataset gave model robust performance in both task by providing knowledge of images closer to real-world scenario and fine-grained fashion attribute text to understand complicated fashion concepts.

4.4 Industrial Applications

As our model demonstrates strong item retrieval ability in fashion items, one of possible applications is fashion-specialized image, item retrieval service. VLM-based item retrieval systems can serve as a cost-effective alternative to commercial off-the-shelf retrieval service, such as Google Lens by allowing businesses to create custom, domain-specific retrieval pools. Just like Figure 4.5, showing example user interface of Google Lens, retailers can curate their inventory into searchable datasets, enabling users to upload images and find visually or semantically similar products.

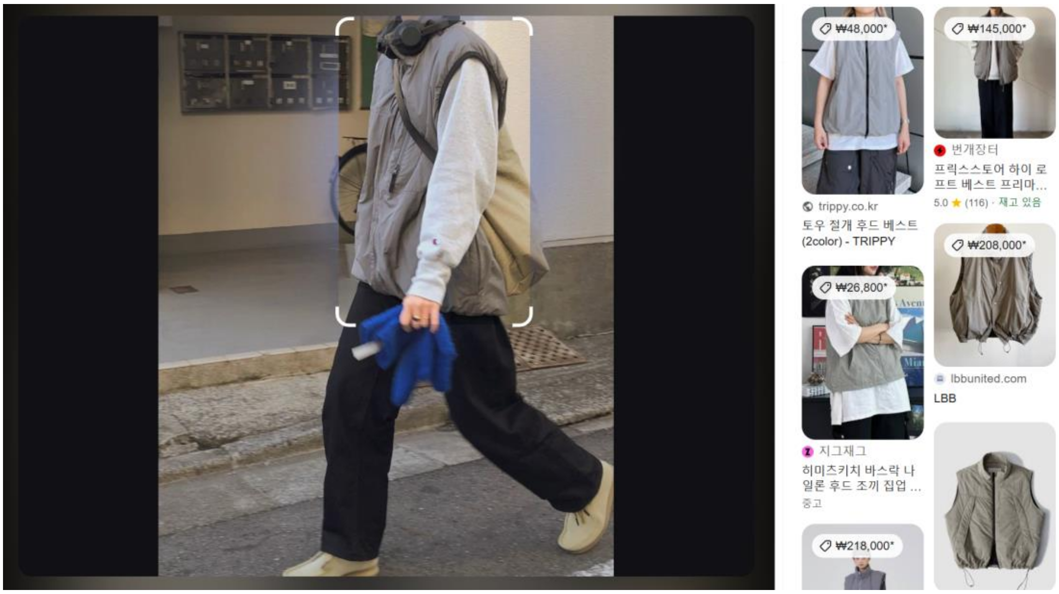

₩48,000*
trippy.co.kr
토우 절개 후드 베스트 (2color) - TRIPPY
₩145,000*
번개장터
프릭스스토어 하이 로프트 베스트 프리마...
5.0 ★ (116) · 재고 있음
₩26,800*
지그재그
히미츠키치 바스락 나일론 후드 조끼 집업 ...
중고
₩208,000*
lbbunited.com
LBB
₩218,000*

Figure 4.5: Image Retrieval Service, Google Lens.

This enables item retrieval integrates into platforms like YouTube to identify clothing or accessories worn in videos in real-time. Assume we extract fashion item images and those features and store to database in advance. It would work as multimodal search engine when users make search query in either image or text format, that retrieves similar items to the search query and web links of retrieved search results, shown as Figure 4.6.

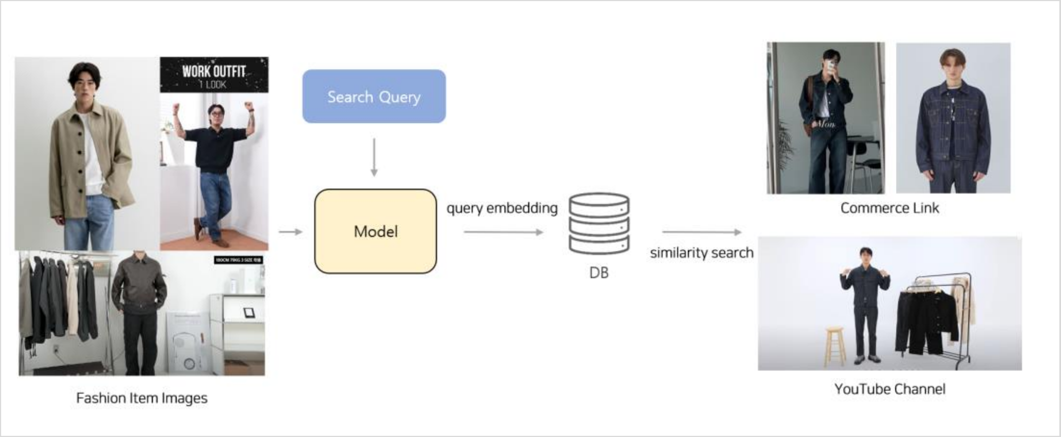


Figure 4.6: Multimodal Search Engine using VLM.

Chapter 5

Conclusion

5.1 Conclusion

This study explored the adaptation of the vision-language model BLIP-2 for the fashion domain, focusing on two primary tasks: fashion image captioning and item retrieval. By leveraging both publicly available datasets and a newly created LLM-generated dataset enriched with fine-grained fashion attributes, our approach demonstrates the potential for improving domain-specific model performance. By using single VLM architecture to handle both tasks at once, and creating complementary dataset without human annotation, we could successfully minimize cost and resources to attain high quality data while maintaining and surpassing performance of existing approaches. Our method is expected to be effectively applicable to image and item retrieval service, possibly substituting existing retrieval service to fashion domain specific search engine.

5.2 Limitations and Future Direction

We could observe compatible performance in image captioning and item retrieval with benchmark models, but still few vision-language tasks remain to explore. We could not cover visual question answering in fashion domain, future work can further optimize and create dataset to handle broad and general text generation tasks. Domain specific auxiliary dataset should be also considered.

Bibliography

[1] Biderman, Dan, et al. "Lora learns less and forgets less." arXiv preprint arXiv:2405.09673 (2024).

[2] Cai, Chen, Kim-Hui Yap, and Suchen Wang. “Attribute Conditioned Fashion Image Captioning.” 2022 IEEE International Conference on Image Processing (ICIP). NEW YORK: IEEE, 2022.

[3] Chen, Zhikai, et al. "Label-free node classification on graphs with large language models (llms)." arXiv preprint arXiv:2310.04668 (2023).

[4] Chia, Patrick John, et al. "Contrastive language and vision learning of general fashion concepts." Scientific Reports 12.1 (2022).

[5] Denkowski, Michael, and Alon Lavie. "Meteor universal: Language specific translation evaluation for any target language." Proceedings of the ninth workshop on statistical machine translation. 2014.

[6] Ge, Yuying, et al. "Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

[7] Golde, Jonas, et al. "Fabricator: An Open Source Toolkit for Generating Labeled Training Data with Teacher LLMs." arXiv preprint arXiv:2309.09582 (2023).

[8] Hacheme, Gilles, and Noureini Sayouti. "Neural fashion image captioning: Accounting for data diversity." arXiv preprint arXiv:2106.12154 (2021).

[9] Hadi Kiapour, M., et al. "Where to buy it: Matching street clothing photos in online shops." Proceedings of the IEEE international conference on computer vision. 2015.

[10] Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." arXiv preprint arXiv:2106.09685 (2021).

[11] Jiang, Yuming, et al. "Text2human: Text-driven controllable human image generation." ACM Transactions on Graphics (TOG) 41.4 (2022): 1-11.2

[12] Kucer, Michal, and Naila Murray. “A Detect-Then-Retrieve Model for Multi-Domain Fashion Item Retrieval.” 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). vol. 2019.

[13] Li, Junnan, et al. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." International conference on machine learning. PMLR, 2023.

[14] Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." Text summarization branches out. 2004.

[15] Liu, Haotian, et al. "Visual instruction tuning." Advances in neural information processing systems 36 (2024).

[16] Liu, Si, et al. "Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set." Proceedings of the 20th ACM international conference on Multimedia. 2012.

[17] Liu, Ziwei, et al. "Deepfashion: Powering robust clothes recognition and

retrieval with rich annotations." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[18] Moratelli, Nicholas, et al. "Fashion-oriented image captioning with external knowledge retrieval and fully attentive gates." Sensors 23.3 (2023): 1286.

[19] Pangakis, Nicholas, and Samuel Wolken. "Knowledge distillation in automated annotation: Supervised text classification with LLM-generated training labels." arXiv preprint arXiv:2406.17633 (2024).

[20] Papineni, Kishore, et al. "Bleu: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.

[21] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

[22] Wu, Hui, et al. "Fashion iq: A new dataset towards retrieving images by natural language feedback." Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. 2021.

[23] Yang, Xuewen, et al. "Fashion captioning: Towards generating accurate descriptions with semantic rewards." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16.

[24] Zhang, Beichen, et al. "Long-clip: Unlocking the long-text capability of clip." European Conference on Computer Vision. Springer, Cham, 2025.

[25] Zhu, Deyao, et al. "Minigpt-4: Enhancing vision-language understanding with

advanced large language models." arXiv preprint arXiv:2304.10592 (2023).

국문초록

비전-언어 모델의 급속한 발전은 이미지 캡셔닝 및 아이템 검색과 같은 도메인 특화 작업 및 응용 분야에서 새로운 가능성을 열어주었다. 그러나 일반 목적의 비전-언어 모델은 패션 도메인에 특화된 세부적인 속성을 이해하는 데 어려움을 겪는다. 본 연구는 최신 비전-언어 모델 중 하나인 BLIP-2를 패션 도메인에 finetuning 하여 이러한 문제를 해결한다.

본 연구는 BLIP-2를 활용하여 이미지 텍스트 생성과 아이템 검색 기능을 동시에 학습시켜 동일한 모델과 데이터셋으로 두 가지 작업을 처리할 수 있도록 설계했다. 이 과정에서 공개된 데이터셋과 세부적인 패션 속성이 풍부하게 포함된 LLM 기반 보조 데이터셋 결합하여 학습에 활용했다. 실험 결과, 보조 데이터를 포함한 공개 데이터셋을 통해 모델이 이미지 캡셔닝과 아이템 검색 작업에서 우수한 성능을 동시에 달성할 수 있었다. 단일 비전-언어 모델 구조와 LLM 기반 보조 데이터셋을 통해, 기존 접근법의 성능을 유지하거나 능가하면서도 고품질 데이터를 확보하기 위한 비용과 자원을 성공적으로 최소화할 수 있었다.

주요어: 비전-언어 모델, 텍스트 생성, 아이템 검색, 의류 및 패션 도메인

학번: 2023-26507

감사의 글

지도해주신 조성준 교수님께 감사드립니다. Big Data AI 연구실원들과 서울대학교 산업공학과 모든 식구들께 감사드립니다.