

# Caption Generation for Clothing Image Pair Comparison Using Attribute Prediction and Prompt-based Visual Language Model

Soichiro Yokoyama  
Faculty of Information  
Science and Technology  
Hokkaido University  
Sapporo, Japan  
email: yokoyama@ist.hokudai.ac.jp

Kohei Abe  
Graduate School of Information  
Science and Technology  
Hokkaido University  
Sapporo, Japan  
email: ko.abe@ist.hokudai.ac.jp

Tomohisa Yamashita  
Faculty of Information  
Science and Technology  
Hokkaido University,  
Sapporo, Japan  
email: tomohisa@ist.hokudai.ac.jp

Hidenori Kawamura  
Faculty of Information  
Science and Technology  
Hokkaido University  
Sapporo, Japan  
email: kawamura@ist.hokudai.ac.jp

**Abstract**—Detailed information for product comparisons is necessary for consumers' purchasing process, especially during the information search and choice evaluation phases. However, conventional product descriptions, which are the primary source of information, tend to focus only on the product in question and thus do not adequately express the differences between products. Garments are treated as target products, and the content required to compare items is assessed from clothing comparison articles in lifestyle magazines. Two generation methods are proposed for comparison of a pair of garment items. The first method separately generates captions for each item and selects a caption pair that expresses differences. The other utilizes a Visual Language Model with a prompt designed based on the assessment. Subject experiments confirmed that the proposed Visual Language Model method accurately represented the feature differences between garments and provided helpful information for consumers to compare garments.

**Keywords**-consumer support; information provision; clothing caption generation; clothing attribute estimation, visual language model.

## I. INTRODUCTION

This paper is based on the study presented initially at INTELLI 2024, The Thirteenth International Conference on Intelligent Systems and Applications [1]. An assessment of lifestyle magazine articles for clothing item comparison was added to organize the contents required in the captions. To generate captions that satisfy the requirements, a new method utilizing a Visual Language Model (VLM) was proposed, and its effectiveness was evaluated by comparison with the algorithm presented at the conference.

In the field of consumer behavior, the sequence of processes involved in the purchase of a product is widely recognized as the purchase decision-making process [2]. This process comprises five stages: problem recognition, information search, alternative evaluation, purchase decisions, and post-purchase evaluation. In the problem recognition phase, consumers identify their needs and problems, and collect information to

satisfy them in the information search phase. In the evaluation of alternatives, the consumer compares and evaluates products based on the collected information, and selects and purchases a specific product in the purchase decision stage. In the post-purchase evaluation, the degree of satisfaction was determined based on the results of the product use. During the information search and evaluation of alternatives phase, consumers need detailed information to understand the characteristics and differences of products and make the right choices. This information can originate from a variety of sources, such as user reviews, expert opinions, and comparison websites; however, product descriptions are one of the most important sources of information that consumers interact with in the early stages of their purchasing decisions. Product descriptions can successfully convey the basic features of a product; however, they tend to focus only on the product in question and do not adequately describe the differences between products. This lack of information may affect consumers' final purchasing decisions and post-purchase evaluations.

Image-caption generation is a research area for generating descriptive text from images; however, it primarily generates a single sentence for a single input image. It is impossible to generate a caption for each image by considering the relationships between multiple images. Some studies have aimed to generate distinctive image captions by comparing input images with similar images in a database; however, they cannot specify the images to be compared, as was the aim of this study. Recently, VLMs that receive prompt texts and images to generate text responses for general tasks have significantly improved and applied to the fashion domain. However, these models have yet to be utilized to generate such a caption pair.

This study aimed to provide adequate information to consumers when comparing products. As a concrete initial effort towards this goal, a method for generating captions that

highlight the differences between two products is proposed and evaluated. Clothing is selected as the target product. Clothing is an everyday purchase for consumers and has various features, such as pattern, material, length, and collar shape. Therefore, consumers need to compare product features during product selection. Articles from lifestyle magazine websites are gathered and analyzed to identify the key characteristics that the captions should include. Based on the assessment, this study considers captions for a pair of clothing items, generating one caption for each item that contains differences from the other item.

Two methods are proposed to achieve such a generation without requiring a large-scale dataset: caption pair selection and prompt-based VLM. The overview of each method is shown in Figure 1. In the caption pair selection method, two different garment images are independently input into an image-caption generation model to generate multiple captions. Next, the prominence of each attribute in each image is calculated using the garment attribute estimation model and the frequency of occurrence in the caption. This is compared between images, and the caption containing more salient attributes than one image is selected from the multiple captions generated for each image and output. The caption pair selection method yields captions that contain more salient features than one garment, with one sentence for each image and an average of approximately 14 words. Examples of the captions obtained are shown in Figure 1a. In the prompt-based VLM method, two garment images and carefully constructed prompts are given to a VLM, and the results are parsed to extract a caption pair. To fully cover the clothing attributes that should be included in the captions, chain-of-thought reasoning, where clothing attributes are first inferred, and captions are generated based on the attributes, is adopted.

In the subject experiment, it was evaluated whether the captions obtained using the proposed methods contained obvious errors, how well they described features that were only present in one garment, and whether they were useful for comparing garments. This experiment confirmed that the captions generated by prompt-based VLM adequately described the differences between products and provided useful information for product comparison.

The remainder of this paper is organized as follows. Section II describes work related to this study. Section III describes the proposed method. Section IV describes in detail the models and datasets used in the experiments. Section V describes the experiments on the comparative validation of the proposed method by employing different scoring methods. Section VI describes the experiments that qualitatively evaluate the captions generated by the proposed method. Finally, Section VII discusses the conclusion of this study and future perspectives.

## II. RELATED WORK

This section describes the main areas relevant to this study, namely image caption generation, caption generation for multiple images, garment attribute estimation, and garment image caption generation.

TABLE I  
COMPARISON OF IMAGE CAPTION GENERATION MODELS.

Model	BLEU4	METEOR
NIC [6]	27.7	23.7
NICA [9]	25.0	13.9
SCST [10]	31.9	25.5
ClipCap [11]	33.5	27.5
OFA [14]	44.9	32.5

### A. Image Caption Generation

Image-caption generation is the task of generating an appropriate description of a single-input image. A comparison of the main image-caption generation models for the benchmark dataset Microsoft Common Objects in Context (MS COCO) [3] is presented in Table I. Bilingual Evaluation Understudy (BLEU) [4] and Metric for Evaluation of Translation with Explicit Ordering (METEOR) [5] are automatic metrics that measure the similarity between the generated and correct captions, with higher values indicating better model performance. Vinyals et al. [6] proposed a model based on a deep recurrent architecture that combines a Convolutional Neural Network (CNN) [7] and Long Short Term Memory (LSTM) [8]. Subsequently, Xu et al. [9] introduced an attention mechanism that focused on specific regions in an image when generating different words. Furthermore, Rennie et al. [10] proposed a model that incorporates reinforcement learning. Recently, image-language pre-training models that learn using large amounts of image-text pair data have achieved higher accuracy than conventional models. Mokady et al. [11] proposed a model that combines the image language pre-training model Contrastive Language-Image Pre-training (CLIP) [12] and the language model Generative Pre-trained Transformer 2 (GPT-2) [13], which reduces training time and achieves highly accurate caption generation. Wang et al. [14] also proposed a pre-training model using 20 million image-text pair data. All these models generate a single-sentence caption for a single input image. In this study, one-sentence captions are generated for each of the two input images. A one-input, one-output image caption generation model is used independently to generate multiple captions for each input image. Each caption is then scored, and the highest caption is generated one sentence at a time to generate a one-sentence caption for each of the two images.

Vision Language Models (VLMs) that receive arbitral text prompts and images and generate appropriate response text regarding the task specified in the prompt have recently achieved remarkable performance improvement. Inspired by the success of Large Language Models (LLMs) in conversation tasks realized with a large amount of training corpus and model parameters represented by GPT series [15], training generic VLMs that are capable of solving a variety of multi-modal tasks of vision and language have been developed. GPT-4 [16] was trained on image input in addition to text corpus, resulting in its capability in text generation through image recognition. Wang et al. [17] showed that a controllable

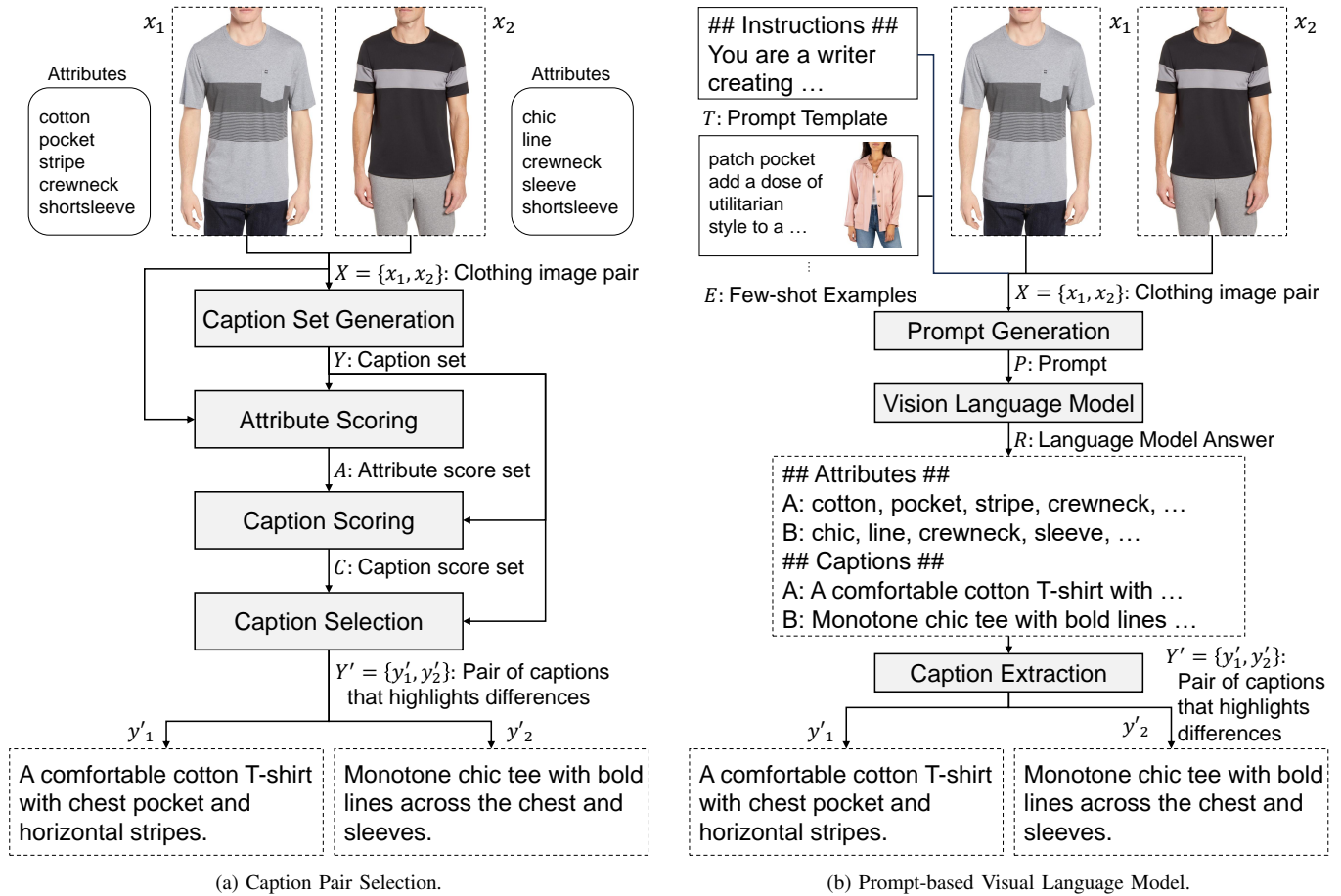


Figure 1. Overview of the proposed methods.

caption generation model can be obtained by training a VLM with a dataset containing multiple styles of captioning. In 2024, Ge et al. [18] presented a training-free pipeline that generates detailed captions by incorporating multiple VLMs and LLMs.

### B. Caption Generation for Multiple Images

Several efforts have been made to generate captions for multiple images as an application of conventional image-caption generation. One example is the change in the image-caption generation initiative. This method identifies changes between two input images and generates a one-sentence caption describing the change [19][20]. In this study, a caption is generated for each input image. In conventional image-caption generation, which tends to generate generic sentences, the distinctive parts of the input images are often ignored. To address this problem, an approach called feature-based image-caption generation is currently in progress [21][22]. In this approach, a single input image is compared to a set of similar images in a database to identify the distinctive aspects of the input image, which are then reflected in the caption. However, this approach does not specify similar images explicitly. In this study, two specified images are compared. The attribute estimates calculated for each image are compared, and a

relative score is calculated. The caption score is then calculated by summing the attribute estimates that appear in the caption and is used for caption selection.

### C. Clothing Attribute Estimation

Clothing attribute estimation is the task of estimating features, such as the material, pattern, collar shape, and sleeve length of clothing in an image. Examples of the estimated attributes include cotton, floral, sleeveless, and leather. This task has been applied to garment retrieval and recommendation. Chen et al. [23] proposed a model that combines a CNN [24] trained on a large image dataset, ImageNet [25] with a multilayer perceptron for a garment image retrieval task that matches images of garments worn by a person with those from a fashion e-commerce site. Similarly, Huang et al. [26] proposed a deep model that included two CNNs to handle street images and e-commerce site images in garment image retrieval. Both models were trained using bounding boxes to identify garment regions. In contrast, Liu et al. [27] proposed a model that learns garment landmark information, such as sleeve and collar positions, estimates the landmarks during inference, and uses this information as an aid for garment attribute estimation. A comparison of the garment-attribute estimation models on the benchmark dataset, DeepFashion

TABLE II  
COMPARISON OF CLOTHING ATTRIBUTE ESTIMATION MODELS.

Model	Top-3 Recall	Top-5 Recall
WBIT [23]	27.46	35.37
DARN [26]	40.35	50.55
FashionNet [27]	45.52	54.61

[27] is presented in Table II. The Top-k Recall [28] was used as an evaluation metric. This assigns the top-k attributes with the highest probability of estimation to each image and measures the number of correctly estimated attributes. By estimating landmark information, FashionNet can better recognize the shape and position of garments and perform better than models that use only bounding boxes. Here, consumer perceptions of attributes are subjective and depend on age and gender. Different consumers may consider different attributes important when comparing garments. However, as a first attempt in this study, the weighting of the attributes did not change. Only estimates objectively calculated using the model were used.

#### D. Clothing Image Caption Generation

Sonoda et al. [29] proposed a method for searching for similar input images from a set of garment images they collected and applied the obtained garment information and features of similar images to a template. Yang et al. [30] proposed a framework that supports the creation of product introductions on e-commerce websites. In their study, attribute- and sentence-level rewards were introduced to improve the quality of captions generated. They also adopted a method for integrating the training of the model using maximum likelihood estimation, attribute embedding, and reinforcement learning. In addition, a large dataset for garment image-caption generation containing approximately one million images was constructed. Cai et al. [31] removed noisy garment images and reconstructed a clean garment image dataset. These studies generated captions describing the salient features of a single-input garment image.

VLMs and LLMs are utilized in the fashion domain to provide more user-friendly interfaces for practical tasks such as retrieval and report generation. Chen et al. [32] integrate ChatGPT with a fashion retrieval system for understanding user queries. Ding et al. [33] propose a system that produces reports by analyzing catwalk images on fashion shows using a VLM. Maronikolakis et al. [34] evaluated the effectiveness of publicly available LLMs as a conversational agent in the fashion domain and showed promising results with GPT-4, a state-of-the-art LLM.

They are insufficient for the purpose of this research, that is, to provide information when comparing garments, in that they cannot express the detailed differences between different garments. In this study, a caption is generated that highlights the differences between two input garment images.

### III. PROPOSED METHOD

This section describes the clothing caption that highlights the differences between garment image pairs and the genera-

tion methods proposed in this study. First, real-world clothing captions that describe differences between garment pairs are assessed. Possible generation approaches are considered, and two promising methods implemented and evaluated in this paper are selected. The first method, caption pair selection, uses a conventional clothes attribute estimator and a clothes caption generator trained on an existing clothes dataset to sample caption candidates and select a pair that contains the most different attributes. The other utilizes a publicly available vision language model with a specifically designed prompt to generate differentiating caption pairs. Both proposed methods are trained on the existing clothing dataset or the general text corpus, requiring little to no additional dataset to generate differentiating caption pairs.

#### A. Generation of Clothing Captions that Highlight Differences

We collected articles from lifestyle magazine websites where multiple clothing items of the same category were compared for customers considering purchase and assessed clothing attributes discussed in the articles to determine the captions to be generated in the proposed method. An instance of clothing captions that highlight differences between a pair of clothing images is produced based on the articles. Finally, algorithms for generating the captions are discussed.

1) *Assessment of Magazine Articles on Garments Comparison*: Eight articles comparing multiple clothing items are collected from Japanese lifestyle magazine websites to assess their comparison approach and item attributes mentioned. Each article is published by lifestyle magazines and compares clothing items of the same category with similar price ranges, typically from different manufacturers, to provide customers with information about each item. Since the original magazine articles were written in Japanese, all the articles were translated into English for assessment. Table III summarizes translated titles, the number of items compared, the description approach for items, and discussed attributes in each article. The number of clothing items compared in one article ranges from 2 to 6, and four out of eight articles compare two items.

Approaches to describe differences among the clothing items can be divided into two categories. One comprises multiple paragraphs, each explaining the items' features on a specific attribute. The other consists of paragraphs explaining each item, enumerating notable attributes. For example, the former first states the design features of two items and their differences, such as the number and location of pockets. The difference of materials in a subsequent paragraph. In contrast, the latter describes the design and material features of one of the items. The design and coordination recommendation for the other item is discussed in successive paragraphs. The highlighting of differences is evident in the former as the differences are discussed per attributes contributing to understanding the breakdown of differences. In the latter, some articles contain sentences that explicitly state the difference from other items, such as "This one is smoother to touch." and "From the four tried-on items, this one was felt to fit my feminine style the best." for highlighting the differences.

TABLE III  
ASSESSMENT RESULT OF CLOTHING COMPARISON ARTICLES FROM LIFESTYLE MAGAZINE WEBSITES.

No.	Title	Items	Category	Description approach	Discussed attributes
1	GU's "Tucked Wide Pants" for a beautiful look. Comparison with the one from UNIQLO	2	Pants	Per attribute comparison	Design, Silhouette, Size, Material, Comfort, People recommended for
2	[A Thorough Comparison of the Most Cost-effective Products] Which is superior, UNIQLO or MUJI? Comparison of "linen shirt" priced at 3,990 yen	2	Tees	Description of each item and per attribute comparisons	Design, Material
3	[Workman VS North Face] Which cardigan to choose for autumn/winter? An enthusiast explains the recommendation!	2	Sweater	Description of each item with notable differencing points	Design, Material, Effect, Coordination, People recommended for
4	[Comparison Report] What are the differences between UNIQLO and GU's much talked-about "parachute cargo pants"?	2	Pants	Description of each item and per attribute comparisons	Design, Silhouette, Material, Impression, Effect, Coordination, Wearing scene
5	[Workman, UNIQLO, Muji] A thorough comparison of the "best men's T-shirts"! Which is the recommendation available under 2,000 yen?	3	Tee	Description per item with differences and similarities	Silhouette, Size, Material, Color, Impression, Effect, Wearing scene, People recommended for
6	Thorough Comparison of UNIQLO's 2024! "White T-shirts" that will be very useful this summer are here!	4	Tee	Description per item with unique feature of each item	Design, Silhouette, Size, Material, Impression, Coordination, People recommended for
7	A hot topic on SNS! Comparing 4 pairs of UNIQLO cargo pants. Which one is the best fit for you?	4	Pants	Description per item with unique feature or differences	Design, Silhouette, Material, Color, Impression, Wearing scene, Coordination, People recommended for
8	For those who can't choose from too many "UNIQLO White T's". Comparison of 6 models including the most popular No.1 and men's [Try-on review]	6	Tee	Description per item with differences	Design, Silhouette, Material, Impression, Effect, Wearing scene, Coordination, People recommended for

TABLE IV  
TEXT LENGTH OF MAGAZINE ARTICLES.

No.	Items	Average count for each item	
		Sentences	Words
3	2	4.5	78.0
5	3	5.3	110.3
6	4	4.0	79.5
7	4	5.3	90.8
8	6	5.8	110.8

This article considers caption generation for a pair of clothing images representing differences between items. A pair of captions are generated, each describing the corresponding item. This approach is commonly utilized in 5 out of 8 articles assessed. A comparison of only two items is considered to validate the basic feasibility and usefulness of caption generation that highlights differences. Note that all the articles assessed for more than three items employ this approach, which implies its extendability.

For the articles that explain each garment in different paragraphs, the text length for one item was approximately five sentences and 100 words in English. Table IV shows the average number of sentences and words. Each item was explained in about five sentences and 100 words for all the articles assessed. Therefore, the generated caption length for each item is also targeted at five sentences and 100 words.

We have organized clothing attributes commonly used in the articles. The following attributes were used to characterize garments, alone or in combination with others.

**Design, Silhouette, and Details** Shape of garments such as

V-neck and crewneck, clothing size outlines of the wearers like loose-fitting, tapered hem and smooth fit over the shoulders, and decoration or utility details such as the number and locations of pockets, ribbons, and straps.

**Material** Clothing materials and textures such as linen, hemp, glossy finish, and smooth texture.

**Color, Pattern, and Print** Available colors, patterns, and garment prints. e.g., red, solid color, bright color, and logos.

Additionally, the following derivative attributes were described in conjunction with the attributes above, often intended to provide solid evidence for more subjective derivative attributes with objective attributes.

**Impression** Impressions that others may receive from the wearer when wearing the garment. e.g., the material with a light sheen gives it a high and beautiful look.

**Effect** The effect of wearing the garment on the wearer's body shape and comfort. e.g., hip-hugging length for a slimming effect, soft against the skin with cotton blend material for a non-stress fit

**Wearing Scene** Situations where it is assumed that wearing clothing is appropriate and effective. e.g., the slightly longer sleeves are also perfect for the morning and evening temperature differences and the chilly rainy season.

**Coordination and Styling** Recommendation on other items that suit the garment. e.g., this shirt is not too long and can be easily matched with tapered silhouette pants

These derivative attributes were often written as recommendations for those with specific ideas, such as "The stretchy

cotton dobby material has a firm feel, making it ideal for those who want to enjoy an elegant look.” possibly because of the subjective nature of these attributes.

Reflecting on the assessment above, we produced an example of a pair of captions that describe the garment of two images as shown in Figure 2. Each caption contains the discussed attributes, explicitly comparing with the other garment and recommending specific people. The part of the text in the figure that refers to attributes is shown in bold. The part that compares with the other item is underlined. Each image of the example is obtained from the official website of UNIQRO and trimmed by the authors to align with other clothing images included in the dataset. The image for Garment A is taken from <https://www.uniqlo.com/jp/ja/products/E468503-000/00>, and <https://www.uniqlo.com/jp/ja/products/E472071-000/00> for Garment B. The caption refers to the other garment as “Garment A/B” for generality. Note that simple algorithms, such as substitution with product names, can easily alter this behavior.

2) *Possible Algorithms for Caption Generation*: We discuss possible algorithms for generating captions that highlight differences. Due to the lack of an existing dataset for such captions, this paper considers two algorithms: caption selection and a prompt-based VLM. First, existing caption generation models for garments are evaluated. Differentiation approaches for a pair of input garments are discussed. Finally, overviews of the two proposed methods are presented.

The caption selection method uses existing captioning models that generate captions from a single clothing image. Firstly, candidates of captions are separately generated for each clothing image of a pair using those models. Then, an additional algorithm selects the most appropriate pair. We propose a selection method based on an existing clothing attribute estimator. Another captioning model that accepts a clothing image and specific attributes to include in the caption could be used for greater efficiency. However, since the pre-trained weights of this model are not publicly available at the time of writing, such an approach is omitted from this paper.

VLMs are trained to generate response text based on text prompts and images for general tasks. With an appropriate prompt, these models can be used to generate captions that highlight differences. We propose a prompt-based VLM method for the caption generation. Most advanced VLMs can be accessed with API, allowing users to input custom text prompts and images and return a responding text. Preliminary experiments are conducted on three VLMs with publicly available APIs, and the most effective model is selected. With an abundant dataset of differentiating captions, VLMs could be fine-tuned for potentially more precise and context-aware caption generation.

### B. Caption Pair Selection Method

An overview of the method is presented in Figure 1a. The method considers a pair  $X = \{x_i \mid i = 1, 2\}$  of different garment images as input and outputs a caption pair  $Y' = \{y'_i \mid i = 1, 2\}$  corresponding to each image, where  $x_i$

is the  $i$ -th garment image, and  $y'_i$  is the output caption corresponding to  $x_i$ . In Figure 1a, the attribute set annotated to the image is displayed next to each image. This method comprises four modules: caption set generation, attribute scoring, caption scoring, and caption selection. The following sections describe these modules in detail.

1) *Caption Set Generation Module*: The caption set generation module considers a pair  $X$  of different garment images as input, inputs each image independently of the image caption-generation model, and outputs a caption set  $Y = \{y_{ij} \mid i = 1, 2; j = 1, 2, \dots, J\}$  corresponding to each image. Here,  $y_{ij}$  represents the  $j$ -th caption for image  $x_i$ . The image-caption generation model used in this study is described in detail in Section IV.

2) *Attribute Scoring Module*: The attribute scoring module considers a pair of different garment images  $X$  and a caption set  $Y$  as input and outputs a set of attribute scores  $A = \{a_{ik} \mid i = 1, 2; k \in K\}$  for each image. Here,  $K$  is the set of attributes to be evaluated and  $a_{ik}$  is the score of attribute  $k$  for image  $x_i$ . An attribute score is a numerical expression of the prominence of a particular attribute exhibited by a garment image; the higher the score, the stronger the garment image that exhibits that attribute. An example of an attribute score for the garment image  $x_1$  in Figure 1a is 0.20 for crewneck, 0.15 for pocket, and 0.01 for sleeveless, which were calculated to be higher when the image had the attribute prominently and lower when it did not. In this study, two methods of attribute scoring were considered: attribute scoring based on attribute estimation, and attribute scoring based on frequency of occurrence.

*Attribute scoring based on attribute estimation* uses a garment-attribute estimation model, whose output is the estimated probability of each attribute for an input-garment image. The estimated probability of an attribute for each image was calculated, and this value was used as the attribute score. This is illustrated in (1), where  $p_{ik}$  is the estimated probability of attribute  $k$  for image  $x_i$ . The clothing attribute estimation model used in this study is described in detail in Section IV.

$$a_{ik} = p_{ik} \quad (1)$$

*Attribute scoring based on frequency of occurrence* assumes that the caption generated for each garment image using the caption set generation module reflects the garment characteristics. If a particular attribute appears frequently in a caption set, it can be regarded as one of the main features of the garment. This method calculates the frequency of occurrence of each attribute in the caption set for each image and uses this value as the attribute score. This is illustrated in (2), where  $f_{ijk}$  is the number of occurrences of attribute  $k$  in the caption  $y_{ij}$ .

$$a_{ik} = \frac{1}{J} \sum_{j=1}^J f_{ijk} \quad (2)$$

3) *Caption Scoring Module*: The caption scoring module considers a caption set  $Y$  and an attribute score set  $A$  as inputs, and outputs a caption score set  $C = \{c_{ij} \mid i = 1, 2; j =$



This **dark brown, simple cotton T-shirt** features a **rounded neckline** that creates a more feminine look compared to Garment B. The **soft fabric** and **slightly slim silhouette** are distinctive, not only giving a **slimmer appearance** but also making it **perfect for creating a polished T-shirt style**, such as **layering it under a jacket**. With fewer decorations than Garment B, it has a more **refined appearance**, making it ideal for **daily outings and office casual wear**. Recommended for those who want to **project an air of elegance in a T-shirt outfit**.

*Bold text indicates attributes and underlined text indicates comparison.*

(a) Garment A and corresponding caption.



This **taupe T-shirt** features ruffled sleeves, which Garment A lacks, adding not only **elegance** but also effectively **covering the upper arms**. The fabric has a **smooth texture**, making it **comfortable to wear even in summer**. The **slim design prevents it from bunching when tucked in**, which is a great advantage. It looks **stylish** on its own and **pairs well with skirts**, making it perfect for **dates and daily outings**. With a more girly impression than Garment A, it's recommended for those who find **plain T-shirts too simple**.

*Bold text indicates attributes and underlined text indicates comparison.*

(b) Garment B and corresponding caption.

Figure 2. Examples of captions for a pair of clothing images. (Images are taken from the UNIQLO official website and trimmed by the authors)

$1, 2, \dots, J\}$ . The caption score is a numerical expression of the extent that the caption reflects the salient attribute differences between the garment images and attributes specific to each image; a higher score is regarded as emphasizing the differences between one image and the other. Here,  $c_{ij}$  represents the score of the caption  $y_{ij}$ . In this study, two caption scoring methods were considered: caption scoring based on the comparison of top attributes and caption scoring based on the addition of relative scores. These methods are described in detail as follows.

*Caption scoring based on comparison of top attributes* first obtains an attribute set  $K_i^{top-n}$  with the top  $n$  attribute scores for each image. Next, the difference set  $D_i$  of  $K_i^{top-n}$  for each image is the difference attribute set, and the product set  $T$  is the common attribute set. These are presented in (3)~(5):

$$D_1 = K_1^{top-n} \setminus K_2^{top-n} \quad (3)$$

$$D_2 = K_2^{top-n} \setminus K_1^{top-n} \quad (4)$$

$$T = K_1^{top-n} \cap K_2^{top-n} \quad (5)$$

Finally, the difference between the number of attribute occurrences in the different attribute sets and the number of attribute occurrences in the common attribute set for each caption was calculated and used as a caption score. This process is illustrated in (6), where  $f_{ijk}$  is the number of occurrences of attribute  $k$  in caption  $y_{ij}$ .

$$c_{ij} = \sum_{k \in D_i} f_{ijk} - \sum_{k \in T} f_{ijk} \quad (6)$$

This method assigns higher scores to captions containing more differentiated and fewer common attributes.

*Caption Scoring Based on Relative Score Addition* first calculates the difference in attribute scores between images

to obtain the relative attribute scores  $\Delta a_{ik}$ . These are given by Equations (7) and (8), respectively.

$$\Delta a_{1k} = a_{1k} - a_{2k} \quad (7)$$

$$\Delta a_{2k} = a_{2k} - a_{1k} \quad (8)$$

Next, the relative attribute scores corresponding to the attributes in the caption are added and used as the caption score. The process is described in (9), where  $K_{y_{ij}}$  represents the set of attributes contained in the caption  $y_{ij}$ .

$$c_{ij} = \sum_{k \in K_{y_{ij}}} \Delta a_{ik} \quad (9)$$

Using this method, captions containing more attributes with relatively high attribute scores have higher scores.

*4) Caption Selection Module:* The caption selection module considers the caption sets  $Y$  and  $C$  as input, selects the caption with the highest caption score in the caption set corresponding to each image, and outputs a set of captions  $Y' = \{y'_i \mid i = 1, 2\}$  that highlights the differences. This process is represented by (10).

$$y'_i = \underset{y_{ij}}{\operatorname{argmax}} c_{ij} \quad (10)$$

### C. Prompt-based Visual Language Model

The method utilizes a VLM, which receives text prompts and images to generate responding text for general tasks. To acquire appropriate captions highlighting differences between clothing images, prompting techniques of few-shot examples and chain-of-thought reasoning are adopted.

An overview of this method is shown in Figure 1b, formulated as follows. First, a pair  $X = \{x_i \mid i = 1, 2\}$  of garment images is given. Combined with a text prompt template  $T$  and few-shot examples  $E = \{x_i \mid i = 1, 2\}$ , inputs for VLM

are formatted as  $P$  in the prompt generation module. With chain-of-thought reasoning, VLM first predicts the attributes of the garments to improve the coverage and then generates captions. Therefore, the VLM response  $R$  contains attributes and captions for both garments. The caption extraction module separates the desired caption pair  $Y' = \{y'_i \mid i = 1, 2\}$  corresponding to each image from the rest of  $R$ .

1) *Prompt Generation Module*: Prompts to the VLM significantly impact the quality of the generated caption and thus are carefully constructed with chain-of-thought reasoning and few-shot examples. Prompt template  $T$  specifies an instruction for caption generation, constraints on the output of each step of chain-of-thought reasoning should follow, followed by few-shot examples  $E$  that instantiate the desired generation contents and formats for specific inputs, and finally, a format that the response should follow and the input images. By concatenating the contents, the prompt generation module produces prompts  $P$ .

The contents of the prompt template  $T$  are as follows. Since VLM APIs utilized in this paper have a separate system input text that controls the role of the VLM in addition to user input texts, the template consists of system and user text parts.

The system input text specifies instructions for the VLM. The writer's role in an e-commerce site that compares two garments is given. Generation is structured in two steps. First, each garment image's previously discussed clothing attributes are inferred as a text of attributes enumeration. Then, a pair of captions is generated based on the input images and the results of the first step.

The user input text includes specific output constraints for each step. For the first step, garment features are categorized, and explanations and examples for each category are given. Constraints for the second step state the text length and the attributes that should be included in the captions. The example of input image pairs and corresponding output texts follows. Then, an output format is specified in JSON for both steps for easy extraction. Finally, clothing images of the pair  $X$  are appended to the prompt.

The few-shot examples  $E$  are produced by the authors regarding the assessment of the magazine articles. Two pairs of clothing items with images, corresponding attributes, and captions are presented, which consist of a pair of shirts, shown in Figure 2, and a pair of pants.

2) *Visual Language Model*: The constructed prompt  $P$  is given to the VLM, and the response  $R$  is returned. In principle, any VLM that accepts arbitrary images and text prompts can be used. For caption generation, however, the VLM has to be flexible enough to follow the instruction of  $P$  and produce a valid JSON formatted response that contains two corresponding captions for each input.

3) *Caption Extraction Module*: The response  $R$  from the VLM is formatted as text data, which can be parsed as valid JSON. The response should contain the estimated attributes for each image as the first step, followed by a corresponding caption for each item as the result of the second step. By accessing the relevant elements of the interpreted JSON object

from  $R$ , the caption extraction module extracts  $Y' = \{y'_i \mid i = 1, 2\}$  from  $R$ .

The extraction of captions is straightforward, provided that the response  $R$  adheres to the format specified in the prompt  $P$ . In our experiments, the VLM consistently produced responses in the expected format, thus eliminating the need for additional post-processing algorithms. Since most VLMs generate responses probabilistically in an auto-regressive manner, the generation process can be repeated if the initial response does not conform to a valid JSON format. This approach ensures that the final output meets the required structure.

#### IV. MODELS AND DATASETS

This section describes the image-caption generation models, garment attribute estimation models, garment image datasets, and VLMs used in the study.

##### A. Image Caption Model and Clothing Attribute Estimation Model

This study is looking at reflecting different national and regional fashion cultures in captions in the future. Therefore, image-caption generation models that can handle garment image data in various languages are desirable. Among the image-caption generation models compared in Section II, ClipCap [11] is a combination of CLIP and the language model GPT-2. It is easy to handle non-English data because CLIP exists for multiple languages [35], and the language model Generative Pre-trained Transformer 4 (GPT-4) [16], which is similar to GPT-2, supports multiple languages. Furthermore, as shown in Table I, the accuracy is sufficiently high among the major image-caption generation models. Therefore, in this study, ClipCap was used as the image-caption generation model in the caption set generation module. FashionNet [27] was used as the garment-attribute estimation model in the attribute-scoring module. This model estimates the landmarks of a garment and uses the obtained information for garment attribute estimation. This model can capture the fine-grained features of a garment image and is highly accurate.

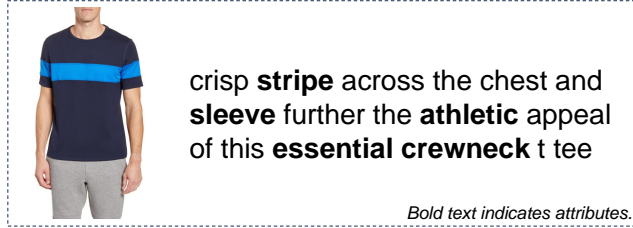
##### B. Clothing Image Dataset

A comparison of the main garment image datasets is shown in Table V. In this study, the FACAD170K garment image dataset [31] with both attributes and captions, which enables an attribute-based caption evaluation, was used to train the image-caption generation model. An example of the FACAD170K data is shown in Figure 3. Each garment image was crawled from a generic website, mainly Google Chrome, and was either an image of a person wearing the garment or an image of the garment alone, with a one-sentence caption from the web. The data collected using this method reflect the variety of styles and trends in clothing that real consumers interact with on a daily basis and are therefore highly suitable for simulation and analysis to mimic the context of consumers' clothing choices. The same caption is provided for garments of different colors. The bold text in the captions for Figure 3 represents multiple attributes assigned to a single garment image. FACAD170K



TABLE V  
COMPARISON OF CLOTHING IMAGE DATASETS.

Dataset	Number of images	Attributes	Captions
FACAD170K [31]	178,862	yes	yes
DeepFashion [27]	289,222	yes	no
FashionGen [36]	325,536	no	yes
iFashion [37]	1,062,550	yes	no



(a) Clothing image A and corresponding caption.



(b) Clothing image B and corresponding caption.

Figure 3. Examples of data from the FACAD170K dataset.

TABLE VI  
HIGH-FREQUENCY ATTRIBUTES COMMON TO BOTH FACAD170K AND DEEPFASHION.

Attribute	Frequency (%)
cotton	4.53
cut	4.41
soft	3.76
sleeve	2.98
fit	2.81
leather	2.58
stretch	2.46
classic	2.45
knit	2.31
strap	2.25

has 990 attributes. In contrast, training the garment-attribute estimation model requires bounding boxes and landmark information to identify garment regions. However, FACAD170K did not contain these annotations. Because annotation is time-consuming, we used FashionNet's DeepFashion [27] pre-training model for garment attribute estimation. DeepFashion contains 1000 attributes, 292 of which match FACAD170K. The top ten attributes with the highest frequency of occurrence in FACAD170K and their frequencies are listed in Table VI. FACAD170K and DeepFashion data with these attributes were used to evaluate the proposed method.

### C. Visual Language Model

In the prompt-based VLM method, a VLM must follow a complex introduction and produce a result in a valid JSON format. We compared three state-of-the-art models that are publicly available through APIs and satisfy these requirements. They achieved competitive results with the vision-text benchmark, and it was determined that there was a need to compare performance on the tasks in this study.

**OpenAI GPT-4o [38]** A flagship model from OpenAI with multi-modal ability when the experiments are performed.

**Anthropic Claude 3.5 Sonnet [39]** The latest model from Anthropic at the time of the study.

**Google Gemini 1.5 Pro [40]** A VLM model from the Google Gemini Team reported the best performance on vision benchmarks across their models during the study.

Although models with publicly available weights, such as LLaVA [41], are attractive options since fine-tuning and integration with other models are practical, these models tend to perform inferior to the compared models in the fashion domain in zero-shot or few-shot settings. Therefore, they are excluded from comparison in this study.

### V. COMPARATIVE VERIFICATION OF MODULES IN THE PROPOSED METHODS

Each module in the proposed method is compared and validated in this experiment. Multiple algorithms or models are presented for some modules in the proposed methods. To identify the most suitable algorithm for each module, preliminary experiments were conducted before engaging in the more labor-intensive qualitative evaluation of the generated captions.

#### A. Caption Pair Selection Method

This experiment aimed to compare and validate attribute scoring based on attribute estimation and frequency of occurrence in the attribute scoring module and caption scoring based on the comparison of top attributes and the addition of relative scores in the caption scoring module to find the best combination of methods for generating captions that highlight differences.

1) *Methods*: In this experiment, the captions generated using the four proposed methods were automatically evaluated. In the caption set generation module, the image-caption generation model ClipCap was trained using 168,862 training data points from FACAD170K. The key parameters during training were set to a learning rate of  $2.0 \times 10^{-5}$ , a batch size of 40, and 10 epochs. These parameters were set based on the settings used in the original study [11].  $J = 100$  captions were generated for each image, based on the probability distribution of the language model. In the attribute scoring module, 292 attributes common to FACAD170K and DeepFashion were used as the attribute set  $K$  to be evaluated. Caption scoring based on top attribute comparisons in the caption scoring module uses the top  $n = 9$  attributes. The values were determined based on preliminary experiments that compared the estimated and correct attributes for different values of  $n$ .

The model was evaluated by comparing the inferred results of the model against FACAD170K and DeepFashion with correct labels. The evaluation metrics are as follows. The set of attributes annotated for a garment image  $x_i$  is the overall attribute set  $K_i^{GT}$ , and the set of attributes with only one garment image is the differential attribute set  $D_i^{GT}$ . This is expressed in (11) and (12).

$$D_1^{GT} = K_1^{GT} \setminus K_2^{GT} \quad (11)$$

$$D_2^{GT} = K_2^{GT} \setminus K_1^{GT} \quad (12)$$

Let  $K_{y'_i}$  be the attribute set contained in the caption  $y'_i$ . The precision, Recall, and F1 scores were calculated between  $K_{y'_i}$  and the differential attribute set  $D_i^{GT}$  to assess the degree of description of the attributes that differed between garments. Similar indices were calculated between  $K_{y'_i}$  and the overall attribute set  $K_i^{GT}$  as supplementary indices to assess the degree of description of the attributes in each garment image. Larger values of these indices are preferable. The evaluation was performed on 10,000 pairs, and the average value of each evaluation indicator was calculated.

2) *Results:* The evaluation results for the captions generated by the proposed method in FACAD170K and DeepFashion are listed in Tables VII and VIII. A comparison of the results across datasets shows that the evaluation values for FACAD170K are higher than those of DeepFashion for all indicators. This is because the image-caption generation model ClipCap was trained on the FACAD170K data; consequently, the attribute information of FACAD170K was more appropriately reflected in the captions. For attribute scoring methods, frequency-of-occurrence-based attribute scoring tends to perform better than attribute estimation-based attribute scoring on both datasets. In particular, FACAD170K outperformed the attribute scoring based on attribute estimation for all evaluation indicators. Regarding caption scoring methods, caption scoring based on relative score addition outperformed caption scoring based on top-attribute comparisons for all evaluation indices in both datasets. These results indicate that under the experimental conditions of this study, the combination of attribute scoring based on the frequency of occurrence and caption scoring based on relative score addition is the most effective.

### B. Prompt-based VLM Method

As stated in the previous section, three candidates exist for the VLM: OpenAI GPT-4o, Anthropic Claude 3.5 Sonnet, and Google Gemini Pro 1.5. Their effectiveness in generating a caption highlighting differences between clothing items is compared with the previously discussed prompts. In addition to that, the effectiveness and necessity of prompting techniques of few-shot examples and chain-of-thought reasoning are verified.

1) *Methods:* In this experiment, one of the authors annotated qualitative evaluations on generated captions for 15 pairs of clothing images. Since the annotations are time-intensive, experiments are systematically conducted on limited combinations of VLMs and prompting techniques.

Generated captions were evaluated by the following annotations and text length. A five-point Likert scale is utilized for concreteness and accuracy for the following clothing attributes. Concreteness is annotated based on whether the attribute is explained in the caption, while accuracy is given by whether the description of the attribute is correct.

- Design, Silhouette, and Details
- Material
- Color, Pattern, and Print
- Wearing Scene
- Comparison with another clothing item

for the derivative attributes, only concreteness is annotated because of their subjective nature.

- Impression
- Effect
- Wearing Scene
- Coordination and Styling
- People recommended for

In the preliminary experiment, the zero-shot setting does not produce captions containing all attributes specified in the prompt with any of the VLMs. Therefore, each VLM is combined with few-shot examples. Three VLMs were compared with few-shot examples without the chain-of-thought reasoning. The effectiveness of chain-of-thought reasoning is compared with the best-performed VLM.

2) *Result:* Of the captions generated by three VLMs with few-shot examples, the average text length did not significantly exceed the target of 100 words with all VLMs. Claude 3.5 Sonnet surpassed other models in all attributes for average concreteness and accuracy, describing design, color, impression, and effect for almost all pairs. The performance on material, coordination, and recommended people were worse than on other attributes. Gemini 1.5 Pro performed inferiorly, particularly in terms of descriptions of materials and coordination.

Introducing chain-of-thought reasoning improved the performance of Claude 3.5 Sonnet in terms of materials, coordination, and recommended people while preserving performance in other attributes and text length compliance. The prediction result of attributes as the first step was accurate for all pairs, possibly improving the captions.

From the experiment results, Claude 3.5 Sonnet with few-shot examples and chain-of-thought reasoning is adopted in the prompt-based VLM.

## VI. QUALITATIVE EVALUATION OF GENERATED CAPTIONS

### A. Objectives

This experiment evaluates the effectiveness of the proposed caption-generation methods by assessing the accuracy of attribute description, the clearness of explanation for differences between pairs, and the usefulness in clothing item comparison.

### B. Methods

In this experiment, the captions generated by the proposed methods were presented to a group of subjects for evaluation. The subject group comprised ten male and female subjects in

TABLE VII  
RESULTS IN FACAD170K.

Attribute Scoring	Caption Scoring	Differential Attributes			Overall Attributes		
		Precision	Recall	F1 Score	Precision	Recall	F1 Score
Attribute Estimation	Comparison of Top Attributes	0.145	0.171	0.144	0.198	0.174	0.173
	Relative Score Addition	0.157	0.223	0.172	0.212	0.225	0.206
Frequency of Occurrence	Comparison of Top Attributes	0.204	0.324	0.236	0.248	0.294	0.258
	Relative Score Addition	0.214	0.369	0.256	0.274	0.353	0.297

TABLE VIII  
RESULTS IN DEEPFASHION.

Attribute Scoring	Caption Scoring	Differential Attributes			Overall Attributes		
		Precision	Recall	F1 Score	Precision	Recall	F1 Score
Attribute Estimation	Comparison of Top Attributes	0.051	0.089	0.058	0.077	0.091	0.077
	Relative Score Addition	0.070	0.136	0.084	0.123	0.164	0.131
Frequency of Occurrence	Comparison of Top Attributes	0.057	0.139	0.075	0.088	0.143	0.104
	Relative Score Addition	0.059	0.156	0.080	0.096	0.171	0.118

TABLE IX  
SET QUESTIONS AND OPTIONS.

No.	Question
Q1	Do you think the description of the attributes of Garment A/B is specific and accurate?
Q2	Do you think the description of the derivation based on the attributes of Garment A/B is specific and accurate?
Q3	Do you think that the attributes and derivatives unique to Garment A/B are described specifically and accurately?
Q4	Do you think a clear comparison is being made with Garment B/A in the caption to Garment A/B?
Q5	Do you think the two captions help you compare garments when you are choosing one to buy?

their 20s. Five pairs of clothing images were prepared. Two proposed algorithms are applied for each pair, and two pairs of captions are obtained. The clothing images and caption pairs were presented to the subjects without specifying the generation method and were evaluated.

Two examples of the presented pairs of clothing images and generated captions are shown in Figure 4. The other three pairs are provided in the supplementary. The pairs consist of highly similar clothing items based on preliminary experiments indicating that captions are most needed when distinguishing between highly similar clothing items. Each item of the pair is referred to as Garment A or B in the captions and questionnaire.

Table IX shows the questions and options set. To make the terms of attributes, derivatives, and these unique to one garment and clear comparison explicit to the subjects, the caption of Figure 2 and each corresponding part of the text was shown to the subjects in advance. Q1 and Q2 were designed to assess how accurately the caption represented garment attributes. Q3 and Q4 assessed how well the captions described the differences between items. Furthermore, Q5 was established to test the usefulness of the caption pairs provided for comparing garments. Q1 to Q4 were answered for each caption of pairs, whereas Q5 was asked for each caption pair. A five-point Likert scale was used to answer each question as

follows.

- Strongly Agree
- Agree
- Neutral
- Disagree
- Strongly Disagree

In addition, the subjects were asked to explain their answers to Q5 and any erroneous features and features not described in the caption.


Wilcoxon's signed-rank test was used as the test method. This test checked whether the answers to each question for the caption generated by prompt-based VLM were significantly biased from neutral and whether there were significant differences between the caption generation methods. The significance level was set at 5%. Furthermore, Bonferroni correction was applied to account for the effects of multiple tests.

### C. Results

The proportions of the responses to each question are shown in Figure 5. For the captions and pairs generated by the caption pair selection method, approximately half of the respondents positively answered 'strongly agree' or 'agree' in Q1. In the rest of the questions, the percentage of favorable responses was well below half. In particular, more than 95% of the answers were 'strongly disagree' or 'disagree' in Q4. For those generated by the prompt-based VLM method, more than 80% of the respondents answered 'strongly agree' or 'agree' in all questions, and more than half answered 'strongly agree.' The percentage of negative answers was relatively high in Q5, where about 10% of the respondents answered 'strongly disagree' or 'disagree'. The p-value of the Wilcoxon signed-rank test was below 0.5%, the significance level after adjusting with Bonferroni correction, for all comparisons between the caption pair selection and the prompt-based VLM method and bias from neutral for the prompt-based VLM method.

We obtained the following explanations from respondents for negative answers in Q5 towards the caption by prompt-based VLM.

1A




**Caption Pair Selection**

Breathable slubbed tech modal and piqu knit ensure performance comfort in a standby polo with easy appeal.

**Prompt-based VLM**

This light blue polo shirt offers a classic, versatile style perfect for various occasions. Its button-down collar and 3-button placket provide a refined touch, while the soft, lightweight knit fabric ensures all-day comfort. Unlike Garment B's bold stripes, this polo's solid heathered color allows for easy pairing with multiple bottoms. Ideal for those who prefer a timeless, understated look, this shirt transitions seamlessly from casual office settings to weekend outings, making it a wardrobe essential for the modern man seeking both style and versatility.

1B



**Caption Pair Selection**


Breathable slubbed tech modal and piqu knit ensure performance comfort in a standby polo with easy appeal.

**Prompt-based VLM**

Make a statement with this blue and white striped camp collar shirt. Its relaxed fit and full button-front closure create a breezy, laid-back silhouette perfect for summer events. Unlike Garment A's more conservative polo style, this shirt features a trendy camp collar and eye-catching vertical stripes, ideal for those who enjoy expressing their personality through fashion. The lightweight woven fabric keeps you cool during beach vacations or outdoor parties. Recommended for the fashion-forward individual who embraces bold patterns and a vacation-inspired aesthetic in their wardrobe.

(a) Pair 1.

2A




**Caption Pair Selection**

Outfitted with elasticized drawstring waist cargo pocket and utilitarian inspired patch pocket these crew pants are the ultimate in rugged utility.

**Prompt-based VLM**

These olive green straight-leg pants offer a perfect blend of casual comfort and smart style. Unlike Garment B's relaxed jogger design, these feature a more structured silhouette with a button closure and belt loops, making them versatile for both weekend outings and smart-casual work environments. The tapered ankle with elastic cuffs provides a modern touch, elongating the legs while maintaining comfort. Ideal for those seeking a polished casual look, these pants pair effortlessly with t-shirts or button-ups, offering greater versatility than the sportier alternative.

2B



**Caption Pair Selection**

Soft and stretchy cotton blend fleece mean easy comfort in comfort joggers with an elastic drawstring waistband.

**Prompt-based VLM**

Embrace comfort with these charcoal grey jogger-style pants, designed for ultimate relaxation and ease of movement. In contrast to Garment A's more structured design, these pants feature a soft, stretchy fabric with a drawstring waist, perfect for athletic activities or lounging. The heathered texture adds visual interest, while the tapered leg and cuffed ankles provide a trendy silhouette. Ideal for those prioritizing comfort and a sporty aesthetic, these pants are best paired with t-shirts and hoodies for a laid-back look that Garment A's more tailored style doesn't offer.

(b) Pair 2.

Figure 4. Examples of presented clothing images pair and captions shown to participants.

**Pair 1** It is difficult to read the intent compared to the reference captions, whether it is a comparison of clothing features or impressions.

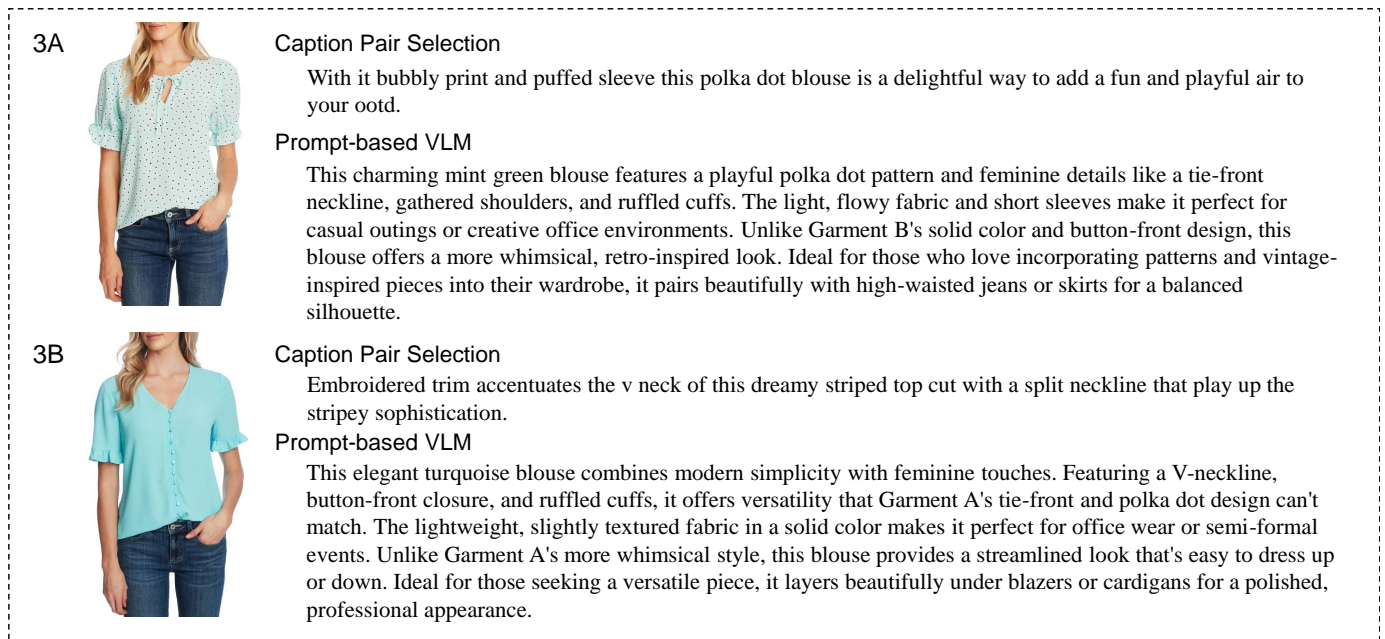
**Pair 4** Items are so different that I don't really feel the need to refer to the caption.

**Pair 5** The caption for Garment A describes it as "Unlike Garment B, this jacket is ideal for mild weather" but the caption for Garment B states that it is "While less suitable for rainy conditions than Garment A" giving the impression that the opposite is true

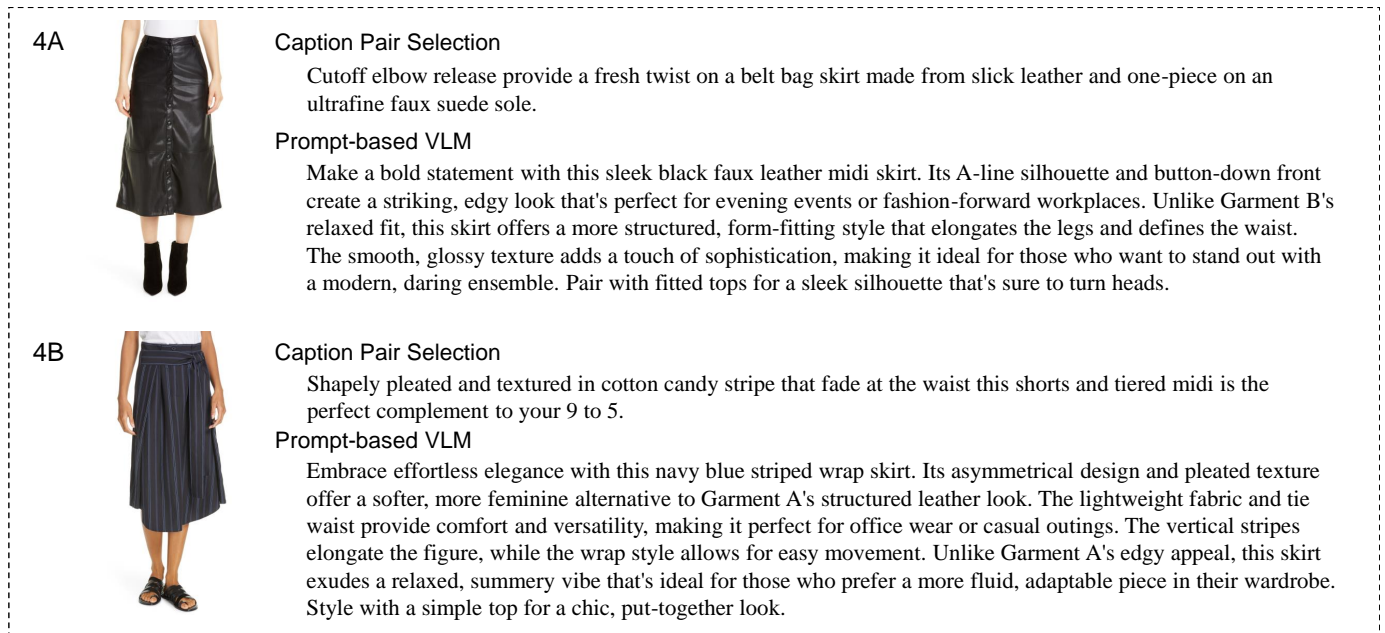
**Pair 5** There are some errors in the characteristics, and I feel a discrepancy with the impression I actually have.

**Pair 5** It is hard to tell what they are comparing. I felt like there was a mix of functional and visual comparisons compared to the reference captions.

In most explanations for negative answers to the caption pair selection method, it is pointed out that the captions are short, and described attributes are easily understandable only from the images without reading the caption.



(c) Pair 3.



(d) Pair 4.

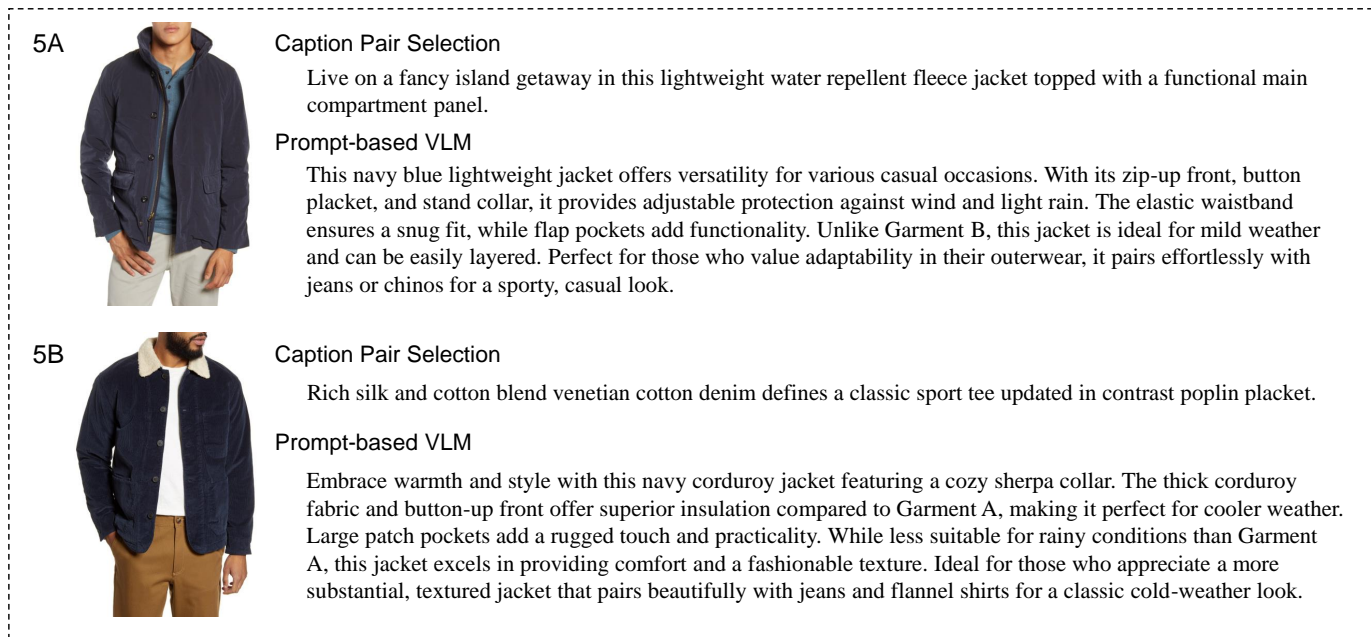
Figure 4. Examples of presented clothing images pair and captions shown to participants.

#### D. Discussions

The overall result indicates that the caption pairs generated by the prompt-based VLM method are preferred to those generated by the caption pair selection method. From the respondents' explanation, the utilized existing caption generator does not cover all the attributes required for the captions that highlight differences. The fact that the proportion of positive answers in Q2 is lower than in Q1 for caption pair selection implies that the caption generator only describes clothing attributes such as design and material and fails to explain

derivatives like impression and coordination. Additionally, a high proportion of negative answers in Q4 shows the limitation of caption pair selection in that the existing generator cannot generate captions that explicitly describe differences, and thus, such captions cannot be selected. These weaknesses can be covered by fine-tuning, although it requires a dataset of captions that contains derivatives. On the other hand, the flexibility of the state-of-the-art VLM can address such an issue by in-context learning with only a few examples.

The negative answers in Q5 for prompt-based VLM indicate



(c) Pair 5.

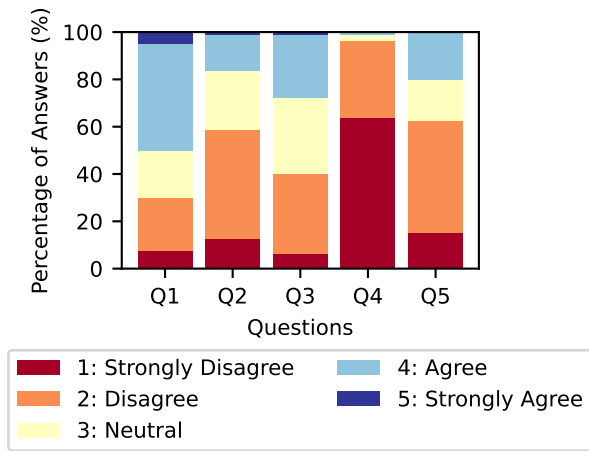
Figure 4. Examples of presented clothing images pair and captions shown to participants.

room for improvement in this approach. One of the reasons is the lack of logical consistency in the generated captions. For example, a further interview for a negative answer in Pair 5 revealed that the respondent felt uncomfortable with the phrase “The thick corduroy fabric and button-up front offer superior insulation compared to Garment A”. It is natural to think that insulation is introduced only by corduroy fabric, not by button-up front. Another example is that one of the respondents found that the part of the caption for 1A, “Unlike Garment B’s bold stripes, this polo’s solid heathered color allows for easy pairing with multiple bottoms.” describes a clothing feature, while the corresponding description for 1B, “Unlike Garment A’s more conservative polo style, this shirt features a trendy camp collar and eye-catching vertical stripes, ideal for those who enjoy expressing their personality through fashion.” shows an impression, giving a misleading feel. The caption for 1A implicitly includes a nuance that 1A is less eye-catching and thus can be combined with various bottom items, which can be explicitly stated for a clearer caption. These logical issues could potentially be resolved by extending the chain-of-thought reasoning so that relations between attributes are inferred. Another reason can be the subjective nature of derivative attributes since one respondent found a discrepancy between the description and her or his impression for pair 5 while others necessarily did not. This issue suggests that personalization can be required to describe such attributes.

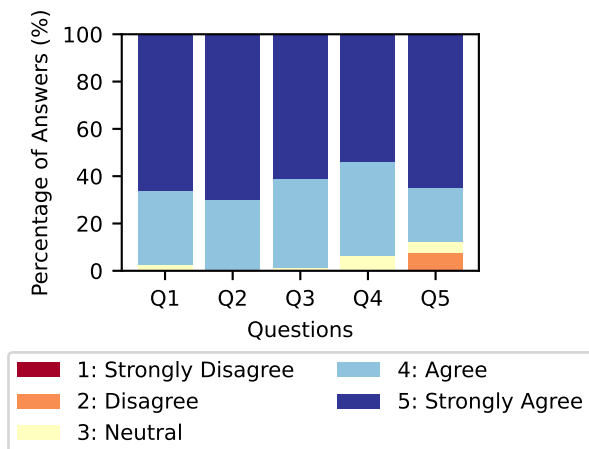
## VII. CONCLUSION AND FUTURE WORK

This study proposed and evaluated a caption-generation method that highlights the differences between pairs of garment images to provide helpful information for consumers when comparing products. The content that should be included

in the captions for the clothing item comparison is assessed from lifestyle magazine articles. The method to generate the captions without preparing a task-specific dataset is discussed, and two methods are proposed, namely caption pair selection and prompt-based VLM. In the caption pair selection method, two garment images are input independently into an image caption generator to generate multiple captions. Attribute scores are then calculated for each image. A caption score is then calculated for each caption in the multiple captions generated for each image using the attribute scores. Finally, the captions are selected and output based on caption scores. Automatic evaluation experiments were conducted on attribute scoring and caption scoring, focusing on accurately describing the features of a single garment and the differences between garments. In the prompt-based VLM method, two images are given to the VLM, and a pair of captions are generated simultaneously. The prompt is designed based on the assessment to describe clothing attributes and item comparisons. Additionally, prompting techniques for few-shot examples and chain-of-thought reasoning are utilized. Since there are multiple approaches and VLMs to implement these methods, preliminary experiments are conducted. Methods employing attribute scoring based on the frequency of occurrence and caption scoring based on relative score addition were rated highly. Attribute scoring based on frequency of occurrence uses the frequency of an attribute’s occurrence in the caption as the attribute score, whereas caption scoring based on relative score addition calculates the relative value of the attribute score and adds it to the number of attributes that appear. Furthermore, captions generated by a combination of methods that received high ratings in the automatic evaluation experiment were presented to the subjects, and a qualitative evaluation of their useful-



(a) Caption Pair Selection.



(b) Prompt-based VLM.

Figure 5. Percentage of answers to each question for captions generated by proposed methods.

ness was conducted. Multiple state-of-the-art VLMs publicly available through APIs are evaluated by annotating generated captions using several criteria determined by the assessment of the item comparison article. As a result, Claude 3.5 Sonnet is selected, and the effectiveness of chain-of-thought reasoning by estimating clothing attributes is verified. The quantitative evaluation with a questionnaire revealed that the prompt-based VLM generates captions containing the required content for comparison and provides helpful information for comparing two garments with the flexibility of the state-of-the-art VLM. Furthermore, it is confirmed that the proposed method has room for improvement due to the need for more logical consistency and subjectivity of the clothing attributes.

The proposed method can only specify two garment images as input images. We plan to extend this approach to handle more than three garment images to meet consumer garment comparison needs better. Based on the assessment of clothing item comparison articles, captions generated for each item can be concatenated to provide information for comparison. The prompt-based VLM method can be easily extended for

multiple clothing items, provided that the sequence length for inputs and outputs of VLM is sufficient.

## REFERENCES

- [1] A. Kohei, Y. Soichiro, Y. Tomohisa, and K. Hidenori, "Generation of Captions Highlighting the Differences between a Clothing Image Pair with Attribute Prediction," in *INTELLI 2024, The Thirteenth International Conference on Intelligent Systems and Applications*, 2024, pp. 7–16.
- [2] M. R. Solomon, *Consumer Behavior: Buying, Having, and Being*. Pearson, 2020.
- [3] T.-Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context," in *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, Springer, 2014, pp. 740–755.
- [4] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.
- [5] S. Banerjee and A. Lavie, "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*, 2005, pp. 65–72.
- [6] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and Tell: A Neural Image Caption Generator," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3156–3164.
- [7] S. Ioffe, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *arXiv preprint arXiv:1502.03167*, 2015.
- [8] A. Graves and A. Graves, "Long Short-Term Memory," *Supervised Sequence Labelling with Recurrent Neural Networks*, pp. 37–45, 2012.
- [9] K. Xu *et al.*, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," in *Proceedings of the 32nd International Conference on Machine Learning*, 2015, pp. 2048–2057.
- [10] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical Sequence Training for Image Captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7008–7024.
- [11] R. Mokady, A. Hertz, and A. H. Bermano, "ClipCap: CLIP Prefix for Image Captioning," *arXiv preprint arXiv:2111.09734*, 2021.
- [12] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the 38th International Conference on Machine Learning*, 2021, pp. 8748–8763.
- [13] A. Radford *et al.*, "Language Models are Unsupervised Multitask Learners," 2019.

- [14] P. Wang *et al.*, “OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework,” *CoRR*, vol. abs/2202.03052, 2022.
- [15] L. Ouyang *et al.*, “Training Language Models to Follow Instructions with Human Feedback,” in *Advances in Neural Information Processing Systems*, S. Koyejo *et al.*, Eds., vol. 35, Curran Associates, Inc., 2022, pp. 27 730–27 744.
- [16] J. Achiam *et al.*, “GPT-4 Technical Report,” *arXiv preprint arXiv:2303.08774*, 2023.
- [17] N. Wang, J. Xie, J. Wu, M. Jia, and L. Li, “Controllable Image Captioning via Prompting,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, 2023, pp. 2617–2625.
- [18] Y. Ge *et al.*, “Visual Fact Checker: Enabling High-Fidelity Detailed Caption Generation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 033–14 042.
- [19] H. Jhamtani and T. Berg-Kirkpatrick, “Learning to Describe Differences Between Pairs of Similar Images,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 4024–4034.
- [20] D. H. Park, T. Darrell, and A. Rohrbach, “Robust Change Captioning,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4624–4633.
- [21] J. Wang, W. Xu, Q. Wang, and A. B. Chan, “Group-based Distinctive Image Captioning with Memory Attention,” in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 5020–5028.
- [22] Y. Mao *et al.*, “Rethinking the Reference-based Distinctive Image Captioning,” in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 4374–4384.
- [23] M. Hadi Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg, “Where to Buy It: Matching Street Clothing Photos in Online Shops,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3343–3351.
- [24] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [25] J. Deng *et al.*, “ImageNet: A Large-Scale Hierarchical Image Database,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
- [26] J. Huang, R. S. Feris, Q. Chen, and S. Yan, “Cross-Domain Image Retrieval with a Dual Attribute-Aware Ranking Network,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1062–1070.
- [27] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, “Deep-Fashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1096–1104.
- [28] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe, “Deep Convolutional Ranking for Multilabel Image Annotation,” *arXiv preprint arXiv:1312.4894*, 2013.
- [29] A. Sonoda, “Apparel EC Saito Ni Okeru Setsumei Bun Jidou Seisei (Automatic Generation of Descriptions in Apparel E-Commerce Sites),” in *Proceedings of the Japan Society of Management Information National Conference, 2018 Autumn*, Japan Society of Management Information, 2018, pp. 125–127.
- [30] X. Yang *et al.*, “Fashion Captioning: Towards Generating Accurate Descriptions with Semantic Rewards,” in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, Springer, 2020, pp. 1–17.
- [31] C. Cai, K.-H. Yap, and S. Wang, “Attribute Conditioned Fashion Image Captioning,” in *2022 IEEE International Conference on Image Processing (ICIP)*, 2022, pp. 1921–1925.
- [32] Q. Chen *et al.*, “Fashion-GPT: Integrating LLMs with Fashion Retrieval System,” in *Proceedings of the 1st Workshop on Large Generative Models Meet Multimodal Applications*, 2023, pp. 69–78.
- [33] Y. Ding *et al.*, “FashionReGen: LLM-Empowered Fashion Report Generation,” in *Companion Proceedings of the ACM on Web Conference 2024*, 2024, pp. 991–994.
- [34] A. Maronikolakis, A. P. Ramallo, W. Cheng, and T. Kober, “What Should I Wear to a Party in a Greek taverna? Evaluation for Conversational Agents in the Fashion Domain,” *arXiv preprint arXiv:2408.08907*, 2024.
- [35] F. Carlsson, P. Eisen, F. Rekathati, and M. Sahlgren, “Cross-Lingual and Multilingual CLIP,” in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 6848–6854.
- [36] N. Rostamzadeh *et al.*, “Fashion-Gen: The Generative Fashion Dataset and Challenge,” *arXiv preprint arXiv:1806.08317*, 2018.
- [37] S. Guo *et al.*, “The iMaterialist Fashion Attribute Dataset,” in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 3113–3116. DOI: 10.1109/ICCVW.2019.00377.
- [38] OpenAI. “Hello GPT-4o,” Accessed: 2024-12-12. [Online]. Available: <https://openai.com/index/hello-gpt-4o/>.
- [39] Anthropic. “Introducing Claude 3.5 Sonnet,” Accessed: 2024-12-12. [Online]. Available: <https://www.anthropic.com/news/claude-3-5-sonnet>.
- [40] Gemini Team, Google *et al.*, “Gemini 1.5: Unlocking Multimodal Understanding Across Millions of Tokens of Context,” *arXiv preprint arXiv:2403.05530*, 2024.
- [41] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual Instruction Tuning,” in *NeurIPS*, 2023.