Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 7.056

IJCSMC, Vol. 15, Issue. 5, May 2026, pg.154 – 162

Attribute-Guided Multi-Image Product Category Identification Using CLIP Representations

Gaon Cheon[1]; Taek Lee[2]

[1, 2]Division of Computer Science and Engineering, Sun Moon University, Republic of Korea
[1] gaon0427@sunmoon.ac.kr; [2] comtaek76@sunmoon.ac.kr
(Corresponding Author: Taek Lee)
DOI: https://doi.org/10.47760/ijcsmc.2026.v15i05.017

Abstract: In online shopping and livestream commerce environments, fashion products are often presented through multiple sales images or video frames rather than a single clean product image. Since several clothing items may appear together, identifying the target product requires considering not only the image content but also the attribute condition that describes the product of interest. In this paper, we define an attribute-conditioned multi-image product identification task and propose a lightweight CLIP-based framework for this setting. The proposed method uses a frozen CLIP image encoder and text encoder, extracts patch-level visual representations, and computes patch-text similarity with an attribute query such as color and clothing type. The selected patch evidence is aggregated within each frame and then combined across multiple images to predict the target product category. Experimental results show that using the penultimate representation of CLIP provides better performance than the last-layer representation and other CLIP-based adaptation baselines. Further patch distribution and patch removal analyses indicate that the penultimate representation forms more compact and stable evidence for the target product. These results suggest that attribute-guided patch-level evidence selection is effective for identifying fashion products in realistic multi-image sales scenarios.
Keywords: CLIP, Fashion Product Identification, Attribute Query, Patch-Level Evidence, Lightweight Adaptation, Livestream Commerce

I. INTRODUCTION

In recent online shopping and livestream commerce environments, products are no longer presented only as a single frontal image. A product can be shown in various forms, such as product pages, short videos, and live streams. In addition, real sales images or video frames often include not only the target product but also other clothing items and background information. Previous studies have reflected this environment in different ways. Some studies have addressed cross-domain product recognition across product page, short video, and live streaming domains [1]. Other studies have used image-title pairs and frame-description pairs collected from Taobao Mall and Taobao Live to study text-query-based image-level product retrieval and object-level product grounding [2]. These studies show that product identification in real e-commerce and livestream commerce is not simply a problem of classifying a whole image. Rather, it is also related to identifying which product inside an image or frame is referred to by the given information.

However, in real sales situations, the target product is not always given as a full product name or a detailed description. A seller may refer to a product using a simple attribute-based expression, such as “gray top,” which includes color and upper/lower clothing information. In this case, multiple clothing items may appear together in the image. In addition, previous fashion dataset studies have shown that clothing understanding involves various detailed tasks beyond simple category classification, such as attribute prediction, retrieval, detection, segmentation, and attribute localization [3][4][5][6].

Based on this background, this study defines product identification in real sales environments as an attribute-conditioned multi-image product category identification problem. That is, given a small number of sales images or frames and a simple attribute condition, the model should find the target product indicated by the condition among multiple visual elements and predict its category. To address this problem, we use CLIP as the base model. Since CLIP is trained to compare images and text in a shared embedding space, it is suitable for measuring the relationship between an attribute query, such as “gray top,” and visual representations in an image [7]. However, what is needed in this setting is not simply matching the query with the whole image. It is important to select local visual evidence related to the target product referred to by the query in an image where multiple clothing items appear together. Therefore, instead of using an image-level representation based on the CLS token of CLIP, we use patch representations and emphasize query-related patches based on the similarity between the text representation and each patch feature.

Based on this idea, we propose a lightweight product category identification framework using frozen CLIP. The input consists of up to four sales images or frames and one attribute query. The framework computes the similarity between each patch feature and the query representation to generate patch-level weights, and uses these weights to select visual evidence from regions related to the query. Then, it combines information from multiple frames and predicts one of 14 clothing categories.

We also focus on the fact that the last layer of CLIP is not always the best representation for this problem. The last layer is strongly optimized for image-text alignment, so it is useful for matching the overall meaning between an image and text. However, it may have limitations in stably preserving local cues, such as color or clothing location [8]. In contrast, the penultimate layer may better retain local visual cues. Therefore, we compare the last layer and the penultimate layer, and analyze which representation is more suitable for attribute-conditioned product identification through patch distribution analysis and patch removal experiments.

The main contributions of this study are as follows. First, we define an attribute-conditioned multi-image product identification problem that reflects real sales images and livestream commerce environments. Second, we propose a lightweight framework that selects patch-level evidence related to the target product without additional localization annotations by using the similarity between frozen CLIP patch representations and an attribute query. Third, we compare the penultimate and last representations of CLIP, and experimentally show that the penultimate representation provides more compact and stable evidence. The overall framework of the proposed method is shown in Fig. 1.

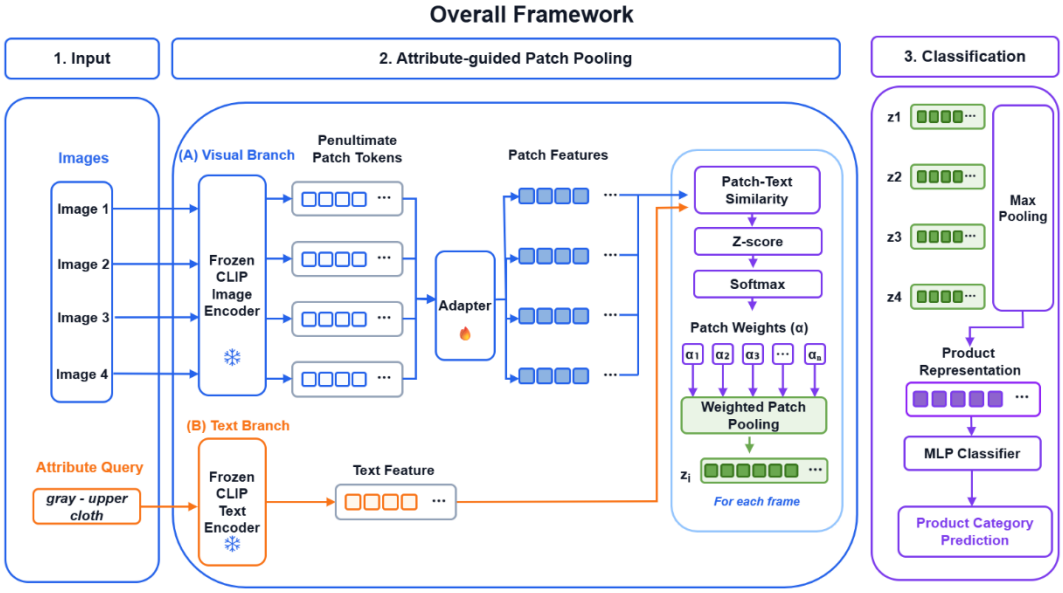

Fig. 1 Overall framework of the proposed method.

II. RELATED WORK

A. Attribute-Aware Fashion Understanding

In fashion product understanding, not only product categories but also fine-grained visual and textual information, such as style and attributes, have been used as important cues. FashionBERT addresses text-image matching and cross-modal retrieval in the fashion industry[9]. It explains that, unlike general-domain matching, fashion matching requires more attention to fine-grained information in both images and text. In particular, it focuses on the fact that fashion text often contains detailed information such as style and attributes, and learns high-level text and image representations by using patch image features. FashionCLIP trains a CLIP-like model adapted to the fashion domain based on contrastive learning, and shows its usefulness for fashion retrieval[10], classification, and grounding. FashionSAP proposes fashion symbols and attribute prompts for fine-grained fashion vision-language pre-training, and presents the learning of attribute-level fashion knowledge as its main contribution[11].

These studies show that fine-grained attribute information is important for fashion product understanding. However, they mainly focus on fashion image-text matching and cross-modal retrieval. In contrast, our study addresses a different setting: identifying the category of a target product referred to by a simple attribute query when multiple sales images or frames are given.

B. LIGHTWEIGHT ADAPTATION OF CLIP

CLIP is a vision-language model that connects images and text through natural language supervision. It was trained on 400 million image-text pairs and demonstrated strong zero-shot transfer ability to various downstream tasks by using natural language to refer to visual concepts. Since then, several methods have been proposed to adapt CLIP to downstream tasks with a small number of trainable components, without fine-tuning the entire CLIP model[7]. CLIP-Adapter proposes a feature adapter added to the visual or language branch of CLIP, which is different from prompt tuning[12]. CoOp models the context words of a prompt as learnable vectors and adapts pretrained CLIP to downstream image recognition tasks while keeping the CLIP parameters frozen[13]. CoCoOp points out that the learned context in CoOp has limited generalization ability to unseen classes, and proposes a lightweight neural network that generates input-conditional tokens for each image[14]. Tip-Adapter proposes a training-free adaptation method that builds a key-value cache model from a few-shot training set and complements CLIP’s prior knowledge through feature retrieval[15]. MaPLe proposes branch-aware multi-modal prompting, which jointly adjusts the vision and language branches of CLIP[16]. By applying prompts to both branches, MaPLe aims to improve vision-language representation alignment.

These studies show that CLIP can be efficiently adapted through adapters, learnable prompts, conditional prompts, cache models, and multi-modal prompts without full fine-tuning. However, they mainly focus on image-level recognition or few-shot classification and generalization. Their focus is different from our setting, where the model identifies the category of a target product in multiple sales images by using the similarity between an attribute query and patch-level visual evidence.

C. CLIP PATCH-LEVEL REPRESENTATION AND LAYER SELECTION

Recent studies have shown that CLIP representations can be used not only for image-level recognition, but also for tasks related to dense prediction and localization. DenseCLIP changes the image-text matching problem of CLIP into a pixel-text matching problem, and uses pixel-text score maps to guide the training of dense prediction models[17]. ResCLIP argues that CLIP performs well on image-level tasks but has limitations in dense prediction[18]. It uses the observation that cross-correlation self-attention in non-final layers has localization properties. Based on this, ResCLIP proposes a Residual Cross-correlation Self-attention module, which reconstructs the attention of the final block using attention from intermediate layers, and improves dense vision-language inference performance. TagCLIP also points out that CLIP’s multi-label classification performance strongly depends on discriminative local features, while CLIP’s global feature can be dominated by prominent classes. To address this issue, TagCLIP analyzes the preservation of patch-wise spatial information inside CLIP and proposes a local-to-global framework based only on frozen CLIP[19].

These studies show that CLIP contains information that can be used as patch-level or region-level visual evidence. Unlike dense prediction or open-vocabulary multi-label classification, our study uses patch-level evidence related to an attribute query for product category identification. In particular, we compare the last-layer and penultimate-layer representations, and analyze which representation forms more stable evidence for the target product through patch distribution analysis and patch removal experiments.

III. METHODOLOGY

This study focuses on identifying the category of a target product from multiple images or frames in real-world product sales and livestream commerce settings. Unlike clean product images, real sales images often contain several fashion items at the same time, such as tops, bottoms, and outerwear. As a result, the target product is not always isolated, and the model needs to understand which item is being referred to. We assume

that the seller's speech is first processed into a simple attribute query that contains color and clothing type. For example, an utterance such as "this gray top..." is converted into a query such as "gray top." Under this setting, the model receives multi-frame images and the attribute query as input, and predicts the category of the product indicated by the query.

A. *Overall Framework*

The proposed method is built on a pretrained CLIP model and selectively uses image regions based on the given attribute query. Each frame is passed through a frozen CLIP image encoder and converted into patch representations. The extracted patch features are then passed through a projection layer and a lightweight adapter to make them more suitable for the target task. The attribute query is encoded by the CLIP text encoder and converted into a text representation. After that, we compute the similarity between each patch feature and the text representation. Based on these similarity scores, the model generates patch-level weights. Patches that are more related to the query receive higher weights, and they are used to build a frame-level representation. Finally, the representations from multiple frames are combined, and the model predicts the final category.

B. *Attribute-guided Patch Pooling*

Instead of using the whole image as a single global representation, we compute the relevance between each image patch and the attribute query. This allows the model to focus on regions that are more related to the target product. For each frame, we extract patch features from the CLIP image encoder. We remove the CLS token and only use the patch tokens. The patch features of frame t are denoted as:

$$P_t = \{p_{t,1}, p_{t,2}, \cdots, p_{t,N}\}, \; p_{t,j} \in \mathbb{R}^d \tag{1}$$

where N is the number of patches and d is the feature dimension. The attribute query is encoded by the CLIP text encoder. Then, we compute the similarity between each patch feature and the query representation. Based on the similarity scores $s_{t,i}$, we generate patch-level weights using a softmax function:

$$\alpha_{t,i} = \frac{\exp(s_{t,i}/\tau)}{\sum_{j=1}^{N} \exp(s_{t,j}/\tau)} \tag{2}$$

where τ is a temperature parameter. The frame-level representation is obtained by the weighted sum of patch features:

$$z_t = \sum_{i=1}^{N} \alpha_{t,i} \; p_{t,i} \tag{3}$$

Through this process, the model can emphasize regions related to the target product without using explicit location annotations. This also helps reduce the influence of other clothing items or background regions in the image.

C. *Layer Selection and Multi-frame Aggregation*

Following the patch-level pooling step, we use patch representations from the penultimate layer of CLIP instead of the last layer. The last layer is mainly optimized for global image-text alignment, which may be less suitable for attribute-based identification using specific image regions. In contrast, the penultimate representation is relatively less specialized for image-text alignment and preserves more local visual information. Therefore, it is more suitable for identifying attributes such as color and clothing type. After obtaining the representation from each frame, we combine them into a single product representation. In this study, we use max pooling to preserve the strongest evidence among multiple frames. This allows the model to effectively use attribute information that may appear clearly only in certain frames.

IV. EXPERIMENTS & DISCUSSION

A. *Dataset Description*

In this study, we construct the experimental dataset based on two public fashion datasets to reflect real product sales environments. First, we use product images and model wearing images from the AIHub fashion product and wearing video dataset. We also use product-related frames from the Taobao Live Multimodal Video Product Retrieval dataset, which contains livestream commerce scenes. Since these datasets include clean product images, model wearing images, and live selling scenes, they are suitable for building the multi-frame product identification task addressed in this study[20]. To avoid identity leakage, product instances were separated at the product level between training and test sets. Attribute queries were automatically generated from metadata annotations including color and upper/lower clothing information.

The original data is reorganized according to our problem setting. Each sample consists of up to four images or frames of the same product, along with an attribute query that describes the product using color and upper/lower type information. The final dataset consists of 18,829 training samples and 4,706 test samples. The classification target includes 14 clothing categories: blazer, cardigan, casual pants, coat, dress, jacket, overall bottom, shirt/blouse, short, skirt, slacks, sweatshirt, t-shirt, and vest. For all experiments, we use the same 80%/20% train/test split to compare the performance of different models.

B. *Implementation Details*

The model uses CLIP ViT-B/16 as the backbone. During training, all parameters of the CLIP image encoder and text encoder are frozen. The trainable parts are limited to the lightweight adapter, the classifier after frame aggregation, and the task-specific modules added in each baseline. All experiments are repeated with 10 different random seeds, and the average performance is reported. We also use early stopping to prevent overfitting. The main hyperparameters are summarized in Table 1. For evaluation, we use Top-1 Accuracy, Top-3 Accuracy, and Balanced Accuracy. Balanced Accuracy is the average accuracy over all classes, and we use it to check whether the model is biased toward majority classes. We compare the proposed method with the original CLIP, the last layer adapter, the penultimate layer adapter, MaPLe, ResCLIP, and a gating model. Although MaPLe is a prompt learning method, we adapt it to our patch-level comparison setting by using patch representations from the image encoder instead of the CLS token for training and prediction. For ResCLIP, we follow the residual attention based structure from the original paper and adapt it to our setting in the form of RCS, which uses intermediate features from the 6th to 9th layers and the attention map from the last layer. The gating baseline is implemented as a late fusion method that uses both the penultimate and last layer representations of CLIP and adaptively controls their contributions through a learnable sigmoid gate.

TABLE I
TRAINING CONFIGURATION

Hyperparameter	Value
Epoch	30
Early Stopping	5
Dropout	0.1
Learning Rate	2e-4
Pooling τ	0.5
Hidden Size	1024

C. *Baseline Comparison*

TABLE III
BASELINE COMPARISON OF CLIP-BASED METHODS

Baseline	Top-1 Accuracy	Top-3 Accuracy	Balanced Accuracy
Base CLIP	82.35±0.67	96.27±0.311	71.63±1.90
Penultimate layer + adapter	90.18±0.30	98.80±0.11	85.39±0.70
Last layer + adapter	87.48±0.46	98.30±0.13	79.42±1.122
Gating(Penultimate + Last)	88.81±0.53	98.50±0.25	81.13±1.20
MaPLe	86.80±0.58	98.10±0.20	79.01±0.80
ResCLIP	80.57±0.38	95.00±0.27	67.48±1.80

To evaluate the effectiveness of the proposed method, we compare it with several models based on CLIP. To isolate the effect of representation selection, all compared models were evaluated under the same multi-frame and patch pooling setting whenever applicable. The baselines include the original CLIP model, adapter models using either the last layer or the penultimate layer, MaPLe, ResCLIP, and a gating model that combines the penultimate and last layer representations. As shown in Table 2, the proposed method, which combines the penultimate representation with a lightweight adapter, achieves the best performance among all compared models. In particular, it improves Top-1 accuracy by more than 7 percentage points compared with the original CLIP model. This result shows that a large performance gain can be achieved through proper representation selection and lightweight adaptation, without full fine-tuning of the CLIP backbone. An interesting result is that the gating model, which combines the penultimate and last layer representations, also shows relatively strong performance, but it still performs worse than using the penultimate representation alone. This suggests that selecting a representation that fits the task can be more important than simply combining different representations.

In contrast, the adapter using the last layer shows lower performance, although it uses a representation that is strongly aligned with text. This suggests that attribute-based product identification requires local visual cues from regions related to the query, rather than relying only on the global alignment between the whole image and text. MaPLe and ResCLIP also show limited performance in this task. Although they use prompt learning and attention based features, respectively, they do not directly select patches related to the query in the same way as the proposed method. Overall, these results show that improving performance in this task depends more on using a suitable CLIP representation and patch features than on adding complex modules or combining multiple features. In particular, the penultimate layer provides a better balance between local visual information and semantic information, making it more suitable for attribute-based product identification.

D. Analysis of Penultimate and Last Representations

To analyze the difference between the penultimate and last representations, we examine how patch-level evidence is distributed in the image and how these patches affect the prediction. All experiments are conducted under the same lightweight adapter setting. First, we conduct a patch distribution analysis to examine which patches are used as evidence by the two representations. We select the top patches that have high similarity to the attribute query and measure how concentrated or scattered these patches are in the image.

TABLE IIIIIV
PATCH DISTRIBUTION ANALYSIS

Metric	Last layer	Penultimate layer
Distance among Top-20 Patches	0.20	0.12
Number of Separated Patch Groups	9.37	5.93

For this analysis, we use two metrics. Distance among Top-k Patches measures the average distance among the selected patches. A larger value means that the patches are more widely spread across the image. Number of Separated Patch Groups indicates how many separate regions the selected patches form. A larger value means that the evidence is distributed across more locations. As shown in Table 3, the last layer shows a larger average patch distance and a larger number of patch groups than the penultimate layer. This suggests that the last representation tends to respond to patches in multiple locations, rather than focusing on regions directly related to the target product. In contrast, the penultimate representation shows a more compact patch distribution, suggesting that it preserves local evidence related to the target product more stably.

TABLE VV
ACCURACY IMPROVEMENT AFTER PATCH REMOVAL

K	Last layer (Misleading Removal)	Last layer (Random Removal)	Penultimate layer (Misleading Removal)	Penultimate layer (Random Removal)
5%	33.10	5.14	4.10	6.23
10%	44.83	7.21	4.79	7.37
20%	51.38	11.28	5.39	10.42

However, the difference in patch distribution alone is not enough to determine whether this property actually causes misclassification. To analyze this more clearly, we conduct a patch removal experiment on the cases where each representation makes a wrong prediction. The detailed results of this experiment are reported in Table 4. Specifically, we separately analyze samples where the last representation predicts the wrong class while the penultimate representation predicts the correct class, and samples where the penultimate representation predicts the wrong class while the last representation predicts the correct class. For each case, we define patches that have higher similarity to the wrong class as misleading patches. We then remove these patches and measure how much the probability of the correct class increases. The results show that when the last representation makes a wrong prediction, removing only a small portion of top patches, such as 5% to 10%, greatly improves the prediction. In particular, removing misleading patches leads to a much larger improvement than removing random patches. This means that the last representation can be easily affected by a small number of misleading patches. In contrast, when the penultimate representation makes a wrong prediction, the performance change is relatively small under the same patch removal setting. This suggests that the penultimate representation does not rely too strongly on a few specific patches, but makes predictions based on more stable evidence.

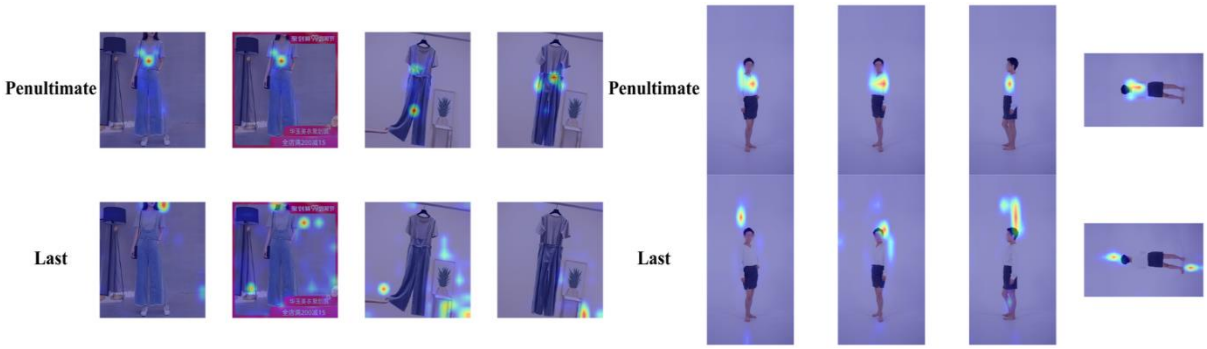

Fig. 2 Patch Evidence Visualization

The visualization results in Fig. 2 provide an intuitive example of this analysis. The penultimate representation mainly shows activation on the target product region, whereas the last representation often shows high activation on regions that are not directly related to the target product, such as the background or other clothing items. This can be interpreted as a result of the strong image-text alignment of the last representation, where some unrelated patches are overly associated with a certain class. Overall, the last representation has strong semantic alignment, but it is also more sensitive to misleading patches, which can lead to misclassification. In contrast, the penultimate representation better preserves local visual cues and uses patch-level evidence more stably. Therefore, for tasks where local visual cues are important, such as attribute-based product identification, the penultimate layer representation is a more suitable choice than the last layer representation.

E. Ablation Study

To analyze the difference between the penultimate and last representations, we examine how patch-level evidence is distributed in the image and how these patches affect the prediction. All experiments are conducted under the same lightweight adapter setting.

1) *Number of Frames:* In Table 5, we analyze the effect of multi-frame input by changing the number of input frames to 1, 2, and 4. The results show that performance tends to improve as the number of frames increases, and the best performance is achieved when using four frames. This can be interpreted as the effect of complementary attribute information from multiple frames, which may not be sufficiently observed in a single image.

TABLE V
PERFORMANCE BY NUMBER OF IMAGES

Images	Top-1 Accuracy	Top-3 Accuracy	Balanced Accuracy
1	87.34	98.62	83.55
2	88.89	98.92	84.63
4	90.18	98.80	85.39

TABLE VVI
PERFORMANCE BY FRAME AGGREGATION METHOD

Pooling	Top-1 Accuracy	Top-3 Accuracy	Balanced Accuracy
Max	90.18	98.80	85.39
Mean	89.88	98.91	85.45
Attention	90.13	98.88	85.49

2) *Frame Aggregation Method:* In Table 6, we compare max pooling, mean pooling, and attention pooling to analyze the effect of the frame aggregation method. The performance difference among the three methods is not large. Although attention pooling shows slightly higher balanced accuracy, the difference is very small. Therefore, we use max pooling as the default setting because it provides stable performance without increasing additional complexity.

TABLE VII
IMPACT OF PRODUCT NAME AUGMENTATION

Query Type	Top-1 Accuracy	Top-3 Accuracy	Balanced Accuracy
Query + Product Name	92.72	98.82	89.38
Attribute Query Only	90.18	98.80	85.39

3) *Effect of Additional Text Information:* In Table 7, we conduct an experiment where the product name is included in the query. The results show that adding the product name further improves performance compared with the original query. This suggests that richer text information helps the model identify the target product in the image.

TABLE VIII
QUERY COMPOSITION ABLATION

Query Type	Top-1 Accuracy	Top-3 Accuracy	Balanced Accuracy
Full Query	90.18	98.80	85.39
Type Only	89.52	98.62	84.70
Color Only	80.22	92.78	71.57
Wrong Type Query	42.10	62.37	32.28

4) *Query Composition:* In Table 8, we analyze the effect of each component in the attribute query. We compare the Full Query, which includes both color and upper/lower type information, with Type-only, Color-

only, and Wrong Type Query. The results show that the Full Query achieves the best performance, while Type-only maintains a similar level of performance. Color-only also shows a reasonable level of performance, but the performance drops significantly with the Wrong Type Query. This suggests that color information also works as a useful cue for finding the target product, but the model mainly uses upper/lower type information as a key condition for identifying the target product.

V. CONCLUSIONS

In this paper, we defined an attribute-conditioned multi-image product category identification task for realistic fashion sales images and livestream commerce environments. We proposed a lightweight CLIP-based framework that selects target-related patch evidence using the similarity between frozen CLIP patch representations and a simple attribute query. Experimental results showed that the proposed penultimate-layer adapter achieved the best performance among all compared baselines. It improved Top-1 accuracy by 7.83% over the original CLIP, 2.70% over the last-layer adapter, and 3.38% over MaPLe. It also improved Balanced Accuracy by 13.76% over the original CLIP. Further analysis showed that the penultimate representation produced more compact and stable patch evidence, while the last-layer representation was more affected by misleading patches. However, this study has several limitations. First, the query setting is limited to simple attributes such as color and clothing type, and does not fully cover more complex product descriptions such as material, pattern, style, or detailed product names. Second, the dataset was reorganized according to our problem setting, so further validation on larger and more diverse real-world commerce datasets is needed. Finally, the comparison is mainly conducted with CLIP-based adaptation baselines, and direct comparison with broader attribute-based fashion recognition models remains limited.

ACKNOWLEDGEMENT

This research was supported by the MSIT(Ministry of Science, ICT), Korea, under the National Program for Excellence in SW, supervised by the IITP(Institute of Information & communications Technology Planning & Evaluation) in 2026"(2024-0-00023)

REFERENCES

[1]. X. Bai, Y. Li, Y. Cheng, W. Yang, Q. Chen, and H. Li, "Cross-domain product representation learning for rich-content e-commerce," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2023, pp. 5697–5706.

[2]. H. Li, H. Jiang, T. Jin, M. Li, Y. Chen, Z. Lin, Y. Zhao, and Z. Zhao, "DATE: Domain adaptive product seeker for e-commerce," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 19315–19324.

[3]. Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "DeepFashion: Powering robust clothes recognition and retrieval with rich annotations," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 1096–1104.

[4]. Y. Ge, R. Zhang, X. Wang, X. Tang, and P. Luo, "DeepFashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019, pp. 5337–5345.

[5]. M. Jia, M. Shi, M. Sirotenko, Y. Cui, C. Cardie, B. Hariharan, H. Adam, and S. Belongie, "Fashionpedia: Ontology, segmentation, and an attribute localization dataset," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2020, pp. 316–332.

[6]. H. Wu, Y. Gao, X. Guo, Z. Al-Halah, S. Rennie, K. Grauman, and R. Feris, "Fashion IQ: A new dataset towards retrieving images by natural language feedback," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 11307–11317.

[7]. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in Proc. 38th Int. Conf. Mach. Learn. (ICML), Proc. Mach. Learn. Res., vol. 139, 2021, pp. 8748–8763.

[8]. F. Sammani and N. Deligiannis, "Interpreting and analysing CLIP's zero-shot image classification via mutual knowledge," in Adv. Neural Inf. Process. Syst. 37 (NeurIPS), 2024, pp. 39597–39631.

[9]. D. Gao, L. Jin, B. Chen, M. Qiu, P. Li, Y. Wei, Y. Hu, and H. Wang, "FashionBERT: Text and image matching with adaptive loss for cross-modal retrieval," in Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR), 2020, pp. 2251–2260.

[10]. P. J. Chia, G. Attanasio, F. Bianchi, S. Terragni, A. R. Magalhães, D. Goncalves, C. Greco, and J. Tagliabue, "Contrastive language and vision learning of general fashion concepts," Sci. Rep., vol. 12, no. 1, Art. no. 18958, Nov. 2022.

[11]. Y. Han, L. Zhang, Q. Chen, Z. Chen, Z. Li, J. Yang, and Z. Cao, "FashionSAP: Symbols and attributes prompt for fine-grained fashion vision-language pre-training," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 15028–15038.

[12]. P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, "CLIP-Adapter: Better vision-language models with feature adapters," Int. J. Comput. Vis., vol. 132, no. 2, pp. 581–595, Feb. 2024.

[13]. K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," Int. J. Comput. Vis., vol. 130, no. 9, pp. 2337–2348, Sep. 2022.

[14]. K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 16816–16825.

[15]. R. Zhang, W. Zhang, R. Fang, P. Gao, K. Li, J. Dai, Y. Qiao, and H. Li, "Tip-Adapter: Training-free adaptation of CLIP for few-shot classification," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2022, pp. 493–510.

[16]. M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan, "MaPLe: Multi-modal prompt learning," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 19113–19122.

[17]. Y. Rao, W. Zhao, G. Chen, Y. Tang, Z. Zhu, G. Huang, J. Zhou, and J. Lu, "DenseCLIP: Language-guided dense prediction with context-aware prompting," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 18082–18091.

[18]. Y. Yang, J. Deng, W. Li, and L. Duan, "ResCLIP: Residual attention for training-free dense vision-language inference," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2025, pp. 29968–29978.

[19]. Y. Lin, M. Chen, K. Zhang, H. Li, M. Li, Z. Yang, D. Lv, B. Lin, H. Liu, and D. Cai, "TagCLIP: A local-to-global framework to enhance open-vocabulary multi-label classification of CLIP without training," in Proc. AAAI Conf. Artif. Intell., vol. 38, no. 4, pp. 3513–3521, Mar. 2024.

[20]. Alibaba Cloud Tianchi, "Taobao Live Multimodal Video Product Retrieval Dataset," Tianchi Datasets. [Online]. Available: https://tianchi.aliyun.com/dataset/75730.

