ProVLA: Compositional Image Search with Progressive Vision-Language Alignment and Multimodal Fusion

Zhizhang Hu
University of California, Merced
zh42@ucmerced.edu

Xinliang Zhu
Amazon Visual Search & AR
xlzhu@amazon.com

Son Tran
Amazon M5
sontran@amazon.com

René Vidal
Amazon Visual Search & AR
rvidal@amazon.com

Arnab Dhua
Amazon Visual Search & AR
aduha@amazon.com

Abstract

Traditional image-to-image and text-to-image search struggle with comprehending complex user intentions, particularly in fashion e-commerce, where users search for similar products with text modifications to a reference image. This paper introduces Progressive Vision-Language Alignment and Multimodal Fusion (ProVLA), a novel approach which utilizes a transformer-based vision and language model to generate multimodal embeddings. Our method involves a two-step learning process and a cross-attention-based fusion encoder to facilitate robust information fusion, and a momentum queue-based hard negative mining mechanism to handle noisy training data. Extensive evaluations on the Fashion 200k and Shoes benchmark datasets demonstrate that our model outperforms state-of-the-art methods.

1. Introduction

Image search, a cornerstone of computer vision, encodes the semantics of an input query to retrieve the most relevant images [9, 24, 31, 12, 10]. However, these traditional image-to-image and text-to-image modalities often struggle to accurately understand complex user intentions. Particularly in the fashion e-commerce domain, where users intend to search for a similar product given a reference picture, but they also have certain changes over the reference, such as changing the color and style. In such cases, the traditional image-to-image and text-to-image methods are limited in fulfilling the users’ search requirements.

Compositional image search [29, 1, 3] aims to address these challenges by using both both image and text inputs. Figure 1 depicts the process of compositional image search. The image serves as a reference, and the text is used as a modification guidance to describe how the reference image should be changed to approximate users’ intentions and obtain the target image. This approach enables users to express their ideas more precisely by leveraging both visual and textual information. However, one intrinsic challenge is the difficulty of compositing image-text representations that transform only the image features relevant to the text modification while preserving other features. Also, given the labor-intensive data collection and labeling processes, the training data is prone to be noisy, i.e., wrongly labeled color, feature, products, etc. Previous works [29, 1, 3] have made strides in this direction but still face several shortcomings.

This paper proposes ProVLA, a new compositional image search model that addresses these challenges. It employs transformer-based vision [21] and language [7] backbone models to generate embedding tokens for each modality. We introduce a two-step learning procedure and a cross-attention-based fusion encoder module to solve the first challenge, and a momentum queue-based hard negative mining mechanism to tackle the second challenge. We evaluate ProVLA on the Fashion 200k [15] and Shoes [14] benchmark datasets. Quantitative results indicate our ProVLA outperforms existing state-of-the-art (SOTA) methods, and qualitative visualization demonstrates effective interaction and exchange of information between modalities.
In summary, our main contributions are the following:

- The introduction of ProVLA, a multimodal compositional learning model for the combined representation of image and text queries.
- A two-step learning framework to progressively condition the model in the complex knowledge for downstream tasks.
- The development of a cross-attention-based multimodal fusion encoder and a momentum queue-based hard negative mining mechanism for robust and efficient information fusion.
- The validation of ProVLA on multiple image-text retrieval benchmarking datasets, demonstrating superior performance over recent state-of-the-art methods.

2. Related Works

Image Retrieval. Image retrieval, a longstanding problem in computer vision, relies on learning an embedding space that maps query and target retrieval candidates closely or far apart depending on relevance [5]. Traditional unimodal-based queries are effective for simple searches but struggle with more complex retrieval targets [12, 24, 22, 20, 32, 36; 11, 34, 35]. In response, Vo et al. proposed a compositional image retrieval model that combines image and text queries [29]. Subsequent models, like TIRG [29], ARTEMIS [6], VAL [3], DATIR [13], CoSMo [18], and MAAF [8], have further developed this approach. However, they tend to neglect contextualizing backbone models pre-trained on the open-domain data, which may limit the model to align domain-specific text terms with image contents.

Vision-language Pre-training. Inspired by the success of transformer-based pre-training in natural language processing [7], vision-language pre-training leverages large-scale image-text pairs to learn a joint vision-language embedding space. Models like CLIP [26] and ALIGN [16] have shown promising results, yet they often ignore the interaction between modalities. Other studies such as UNITER [4], ViLT [17], ALBEF [19], and TCL [33] have proposed ways to learn joint embeddings of image content and natural language during pre-training, thereby increasing the interaction between modalities. This work builds upon these findings, proposing a vision-language training approach that tailors the backbone models to the joint embedding space with the semantics of the target retrieval task.

3. Proposed Method

The proposed framework, referred to as ProVLA, aims to create a fused representation of the reference image and modification text that is well aligned with the target image. It comprises a two-step progressive learning: Pre-fine-tuning and Image-Text Composition, and two main components: multimodal fusion encoder and momentum queue-based negative mining (Figure 2). The pre-fine-tuning step conditions the model into the domain-specific semantic spaces [37, 3]. During this phase, the model learns to align domain-specific text terms with image contents. Subsequently, the Image-Text Composition process executes multimodal fusion for the image search and retrieval task. The Pre-fine-tuning and Image-text composition networks utilize a vision encoder, a text encoder, and a multimodal fusion encoder, all based on transformer models [28]. The Pre-fine-tuning network inputs are the image and its caption, while the Image-text Composition network takes the reference image, target image, and corresponding modification text as inputs. In both steps, a Swin Transformer [21] is employed to extract visual representation from images. The compositional image search task requires both fine-grained visual information for the features mentioned in the modification text and coarse-grained visual information for the remaining features that should be preserved. As the Swin Transformer learns visual concepts at different levels of granularity in a hierarchical order, we concatenate output image tokens from the penultimate (Stage 3) layers and the final (Stage 4) as the image embedding. To encode the semantics of the modification text, we adopt a 6-layer transformer as the language backbone.

We view the image-text composition task as an image-to-image retrieval task, conditioned by the explicit description in the modification text. This process mainly relies on the cross-attention mechanism in the multimodal fusion encoder. The reference image embedding serves as the query in the cross-attention operation, acting as the initial search context for visual information. The modification text embedding, functioning as both keys and values, is the potential transformation repository, from which relevant modifications are identified and applied to the image, effectively bridging the image and text domains.

The primary objective of image-text composition training is to bring the embeddings of matched composed image-target image pairs closer while distancing the unmatched pairs. Therefore we employ a momentum queue-based negative mining strategy. For this, we leverage contrastive learning and a mechanism of momentum-based distillation, maintaining queues to store the most recent composed image, reference image, and target image embeddings [30, 19]. This results in optimized contrastive learning tasks and the calculation of softmax-normalized similarity.

More formally, let \( I^r \) denote the reference image, \( T \) the modifying text, \( C \) the composed image, and \( I^t \) the target image. Let \( Y^{c;2I^r}(C) \) and \( P^{c;2I^r}(C) \) denote, respectively, the one hot and soft-max normalized similarities between the composed image and the set of target images, as defined in [19]. Likewise, let \( Y^{I^t;c}(I^t) \) and \( P^{I^t;c}(I^t) \) denote the one-hot and soft-max normalized similarities between the target image and the set of composed images. Here, \( (I^t, C) \)
is a positive pair if the tuple \((I^t, T)\) used to compute \(C\) is paired with \(I^t\), in which case the one-hot similarity has a probability of 1. Otherwise, the pair is negative and the one-hot similarity has a probability of 0. The Composed-Target matching (CTM) loss is then mathematically defined as the cross-entropy \(H\) between softmax-normalized similarity \(P\) and the one-hot similarity label \(Y\) [19], i.e.:

\[
L_{CTM} = \frac{1}{2} E_{(C,I^t)}[H(C|Y^{c2I^t}(C), P^{c2I^t}(C)) + H(Y|I^{t,2c}(I^t), P^{t,2c}(I^t))].
\] (1)

In order to encourage the vision encoder to preserve visual attributes not modified by the text, we adopt the Reference-Target matching (RTM) loss. This loss employs the same design as the CTM loss, but the pair of inputs is the embedding of reference and target images. Also, to facilitate the training of the fusion encoder, we adopt the image-text matching (ITM) and masked language modeling (MLM) losses, which are widely used in previous vision-language fusion research [19, 33].

Given the triplet of the reference image, target image, and text, the task of image-text matching is to predict whether the text matches the difference between the two images. The masked language modeling task aims to predict the masked word in the given text by looking at the remaining text and two images. We adopt the same fusion encoder for the image-text composition, but the text embedding is the query to the cross-attention. For the ITM, the difference between target and reference image embedding is the input to the fusion encoder, and the concatenation of two image embeddings is used as the input for the MLM task.

In summary, the overall training objective of the Image-Text Composition is:

\[
L = L_{CTM} + L_{RTM} + L_{ITM} + L_{MLM}.
\] (2)

Please note that we omit the loss from the momentum model [19] for a clearer presentation.

### 4. Experiment

#### 4.1. Experimental Setup

**Datasets.** We test our method on two widely-used benchmarking datasets: Fashion 200k [15] and Shoes [14].

The Fashion 200k dataset, sourced from multiple websites, comprises roughly 200,000 fashion images with product information and product images. We adopt the approach introduced in TIRG [29], using only images and corresponding text captions, where the text style is an attribute list. We filter image pairs with a single word difference in the caption and generate modification text in a “replace A with B” format. This results in triplets of a reference image, target image, and modification text. The dataset comprises around 170,000 training and 30,000 testing triplets. The retrieval is performed by matching target attributes, not images.

The Shoes dataset was initially designed for attribute discovery [2], featuring image-text pairs of products and titles. It is then augmented with relative captions in natural language for dialog-based interactive image retrieval [14]. It consists of 10,000 training and 4,658 testing triplets. The modification text in this dataset is human-annotated, reflecting user-feedback in e-commerce.

**Baseline Models.** We compare ProVLA with a wide range of baselines including early works and recent SOTA models on this task, including TIRG [29], VAL [3], MAAF [8], CoSMo [18], ARTEMIS [6] and DATIR [13].
**Evaluation Metric.** For both datasets, we adopt the standard evaluation metric in retrieval, i.e., Recall@K, denoted as R@K for short.

**Implementation.** We use the PyTorch [25] deep learning framework to conduct all our experiments. The Swin Transformer is pre-trained on ImageNet-21k [27] at resolution 224x224. The text encoder and fusion encoder are initialized with a BERT$_{base}$ [7] model. We pre-fine-tune the model for 50 and 100 epochs for the Fashion 200K and Shoes datasets, respectively, using a batch size of 192 on 8 NVIDIA V100 GPUs. The optimizer is AdamW [23] with a weight decay rate of 0.02. We adopt a cosine-scheduled learning rate with a warmup set to $1e^{-5}$ in the first 30 epochs, and decayed to $1e^{-6}$. The momentum parameter is set to 0.995, and the size of the hard-negative mining queue is set to 38,400. The distillation weight $\alpha$ is linearly ramped up from 0 to 0.6 within the 1st epoch.

### 4.2. Compositional Search Performance

Table 1 and Table 2 illustrate the retrieval performance on the Fashion 200K and Shoes datasets. From those tables, we observe that our model achieves compelling results compared to other methods. On the Fashion 200k dataset, our model surpasses the SOTA method by a significant margin, with a 2.6 percent points increase in R@10 and a 1.9 percent points increase in R@50. However, we note a minor setback, as our model underperforms by 1.6 percent points on the R@1 metric. Similarly, in the Shoes dataset, our model exhibits a competitive edge, outperforming the SOTA by 0.5 percent points on R@1 and by a notable 3.1 percent points on R@50. A limitation surfaces on the R@50 metric, where our model falls behind the SOTA.

The improvement in both datasets can be attributed to our model’s efficacy in leveraging relevant image and text data, generating meaningful modifications for improved retrieval performance. However, the slight deficits in Recall@1 on the Fashion 200k dataset and Recall@50 on the Shoes dataset underscore room for further refinement. We conjecture that the shortcomings may stem from the model’s current handling of edge cases where minor attribute variations significantly impact the retrieval process.

### 4.3. Ablation Study

To better understand the impact of the two key components of our model, namely the two-step progressive learning and the queue-based hard negative mining, we conducted an ablation study. This analysis aims to isolate the contributions of each component and provide insights into their functionality and importance within the model. Table 3 depicts the result of the ablation study.

Initially, a simplified model without these components showed R@10 scores of 18.2 and 23.6 on the Fashion 200k and Shoes datasets, respectively. Adding queue-based hard negative mining significantly increased the scores to 30.3 and 53.7, demonstrating its vital role in improving retrieval precision by handling diverse negative samples. Incorporating both components further improved R@10 scores to 53.7 and 56.2 on the respective datasets, emphasizing the importance of progressive learning in refining the model’s understanding of the data.

### 5. Conclusion

In conclusion, our Progressive Vision-language Alignment and Multimodal Fusion (ProVLA) model introduces a more effective method for image search, especially in contexts where users aim to find items similar to a reference image with specific modifications. The model employs a two-step learning framework, a cross-attention-based fusion encoder, and a momentum queue-based hard negative mining mechanism to tackle inherent challenges. Evaluated on multiple image-text retrieval benchmark datasets, ProVLA outperforms recent state-of-the-art methods, illustrating its potential for robust and precise image search.

<table>
<thead>
<tr>
<th>Method</th>
<th>Fashion 200k</th>
<th>Shoes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@1</td>
<td>R@10</td>
</tr>
<tr>
<td>TIRG</td>
<td>14.1</td>
<td>42.5</td>
</tr>
<tr>
<td>VAL</td>
<td>22.9</td>
<td>50.8</td>
</tr>
<tr>
<td>MAAF</td>
<td>18.9</td>
<td>-</td>
</tr>
<tr>
<td>CosMo</td>
<td>23.3</td>
<td>50.4</td>
</tr>
<tr>
<td>ARTEMIS</td>
<td>21.5</td>
<td>51.1</td>
</tr>
<tr>
<td>DATIR</td>
<td>21.5</td>
<td>48.8</td>
</tr>
<tr>
<td>Ours</td>
<td>21.7</td>
<td>53.7</td>
</tr>
</tbody>
</table>

Table 1. Quantitative results for the Fashion 200k dataset. The highest value is colored green. "-" means the result is not reported in the original paper.

<table>
<thead>
<tr>
<th>Method</th>
<th>Fashion 200k</th>
<th>Shoes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@1</td>
<td>R@10</td>
</tr>
<tr>
<td>TIRG</td>
<td>12.6</td>
<td>45.5</td>
</tr>
<tr>
<td>VAL</td>
<td>16.5</td>
<td>49.1</td>
</tr>
<tr>
<td>MAAF</td>
<td>16.4</td>
<td>50.0</td>
</tr>
<tr>
<td>CosMo</td>
<td>16.7</td>
<td>48.4</td>
</tr>
<tr>
<td>ARTEMIS</td>
<td>18.7</td>
<td>53.1</td>
</tr>
<tr>
<td>DATIR</td>
<td>17.2</td>
<td>51.1</td>
</tr>
<tr>
<td>Ours</td>
<td>19.2</td>
<td>56.2</td>
</tr>
</tbody>
</table>

Table 2. Quantitative results for the Shoes dataset. The highest value is colored green.

<table>
<thead>
<tr>
<th>Method</th>
<th>Fashion 200k</th>
<th>Shoes</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Pre-fine-tuning and no queue-based negative mining</td>
<td>18.2</td>
<td>23.6</td>
</tr>
<tr>
<td>No Pre-fine-tuning</td>
<td>30.3</td>
<td>37.4</td>
</tr>
<tr>
<td>ProVLA</td>
<td>53.7</td>
<td>56.2</td>
</tr>
</tbody>
</table>

Table 3. Ablation study of removing two key components in the ProVLA.
References


transferable visual models from natural language supervision.
In International Conference on Machine Learning, pages 8748–8763. PMLR, 2021. 2


[29] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-an empirical odyssey. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 6439–6448, 2019. 1, 2, 3


