# FACap: A Large-scale Fashion Dataset for Fine-grained Composed Image Retrieval

François Gardères[*†]
francois.garderes@louisvuitton.com

Shizhe Chen[†]

Camille-Sovanneary Gauthier[*]

Jean Ponce[†‡]

## Abstract

The composed image retrieval (CIR) task is to retrieve target images given a reference image and a modification text. Recent methods for CIR leverage large pretrained vision-language models (VLMs) and achieve good performance on general-domain concepts like color and texture. However, they still struggle with application domains like fashion, because the rich and diverse vocabulary used in fashion requires specific fine-grained vision and language understanding. An additional difficulty is the lack of large-scale fashion datasets with detailed and relevant annotations, due to the expensive cost of manual annotation by specialists. To address these challenges, we introduce **FACap**, a large-scale, automatically constructed fashion-domain CIR dataset. It leverages web-sourced fashion images and a two-stage annotation pipeline powered by a VLM and a large language model (LLM) to generate accurate and detailed modification texts. Then, we propose a new CIR model **FashionBLIP-2**, which fine-tunes the general-domain BLIP-2 model on FACap with lightweight adapters and multi-head query-candidate matching to better account for fine-grained fashion-specific information. FashionBLIP-2 is evaluated with and without additional fine-tuning on the Fashion IQ benchmark and the enhanced evaluation dataset enhFashionIQ, leveraging our pipeline to obtain higher-quality annotations. Experimental results show that the combination of FashionBLIP-2 and pretraining with FACap significantly improves the model's performance in fashion CIR especially for retrieval with fine-grained modification texts, demonstrating the value of our dataset and approach in a highly demanding environment such as e-commerce websites. Code is available at https://fgxaos.github.io/facap-paper-website/.

## CCS Concepts

• **Information systems** → **Image search**.

[*]Louis Vuitton
[†]Inria, École normale supérieure, CNRS, PSL Research University
[‡]Courant Institute of Mathematical Sciences and Center for Data Science, New York University


*GENNEXT@SIGIR'25, July 17, 2025, Padova, Italy*

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

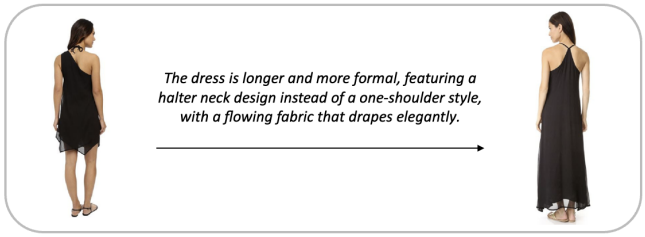
## Keywords

Composed Image Retrieval, Multimodal Fusion, Fashion Domain

**ACM Reference Format:**
François Gardères, Shizhe Chen, Camille-Sovanneary Gauthier, and Jean Ponce. 2025. FACap: A Large-scale Fashion Dataset for Fine-grained Composed Image Retrieval . In *Proceedings of GENNEXT@SIGIR'25*. ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 Introduction


**(a) Examples from the FashionIQ dataset [52].** Left: incorrect annotation (the target dress is not pink). Right: vague annotation lacking sufficient details to accurately retrieve the target image, like color or shape.




**(b) Example from our FACap dataset.**

**Figure 1: Our automatically constructed FACap dataset offers more detailed and accurate annotations** than existing datasets for the fashion CIR task.

Efficiently retrieving fashion images based on user preferences is crucial for enhancing e-commerce experience, from online shopping to inspiration and brand discovery. The preferences relate both to search interaction preferences —querying with images for instance —and taste and vocabulary preference to really adapt to user needs. Traditional image-to-image [42] or text-to-image [38] retrieval methods primarily support single-modality queries and fall short in handling more complex, real-world scenarios. For instance, a user may want to find a product similar to the one they have seen before, but with specific changes, like a different color, style, or feature. To address this, recent works have increasingly focused on composed image retrieval (CIR) [48, 52], which aims to

retrieve relevant fashion images by leveraging a reference image along with a modification text that describes specific alterations.

Most existing methods for Fashion CIR [5, 10, 34, 35, 55, 56] fine-tune pretrained vision-language models (VLMs) like CLIP [37] or BLIP-2 [25] to map images and texts into a shared multimodal space. The embeddings of the reference image and modification text are then fused and compared with the embeddings of candidate images to identify the most relevant match. However, these approaches are constrained by the limitations of current Fashion CIR datasets. For example, FashionIQ [52], a widely used dataset for this task, is limited in scale, containing only 18k *<reference image, modification text, target image>* triplets across just three fashion categories: dresses, shirts, and tops. A larger scale dataset would be able to better represent the concept diversity of fashion-domain knowledge. Furthermore, the crowdsourced captions in FashionIQ are short, noisy and lack details, as shown in Figure 1a and confirmed by our quality evaluation in Table 2. A better CIR experience is expected from three features of the modification text: faithfulness, levels of detail and discriminative power. Reaching a high level of quality is time-consuming and expensive, as it requires to manually annotate a large number of CIR triplets. The noisy and limited data available today hinder the existing models' ability to understand fine-grained fashion-related features crucial for fashion search tasks, such as specific collar types or textures.

To tackle the data scarcity challenge, some approaches have attempted to increase the data size, for example by generating reverse descriptions [34] for reference and target images; but generating difference descriptions for two images [36] is itself a challenging task. Other approaches attempt to eliminate the need for training data by performing zero-shot CIR [4, 21, 28, 40, 45, 54] with the help of pretrained VLMs, but their performance suffers from the absence of domain-specific representation learning. More recent efforts pretrain VLMs using web-crawled fashion images to learn more accurate multimodal representations for fashion [55, 56], but raw web data are often noisy and lack the necessary comparisons between pairs of images for effective CIR training.

In this work, we introduce Fashion Automatic Caption (FACap), a large-scale fashion-domain CIR dataset with fine-grained annotations. FACap automatically pairs web-sourced fashion images and employs a two-stage annotation pipeline to generate modification texts, hence creating CIR triplets. The first stage refines original noisy web captions using a VLM to produce long, faithful, and detailed descriptions for each image. The second stage utilizes a large language model (LLM) to analyze the differences between reference and target image captions, generating concise and accurate modification texts. With over 227k CIR triplets, FACap offers a high-quality dataset addressing the challenges of scale, accuracy, and detail in fashion CIR, as evidenced in our quality evaluation. We also introduce the FashionBLIP-2 model for Fashion CIR task using BLIP-2 [25] as backbone, and lightweight adapter modules to specialize it for fashion retrieval needs. Additionally, instead of relying on global features to match the multimodal query and candidate image, we design a multi-head query-candidate matching method that uses multiple feature representations to capture more fine-grained details. We evaluate the performance of our model in two settings: with and without fine-tuning on downstream fashion datasets. Experimental results demonstrate that pretraining on FACap significantly improves model performance for Fashion CIR and showcase the effectiveness of our FashionBLIP-2 model.

To summarize, our contribution is three-fold:

- We propose an automatic data construction method to scale up Fashion CIR datasets with web-sourced images and foundation models, resulting in a large-scale and high-quality dataset FACap.
- We propose the FashionBLIP-2 model, incorporating BLIP-2 with lightweight adapters for fashion domain adaptation and multi-head matching to cover fine-grained details.
- Experimental results on two benchmarks with and without downstream fine-tuning demonstrate the value of FACap and our model.

## 2 Related Work

### 2.1 Composed Image Retrieval

Existing approaches [5, 10, 34, 35, 55, 56] to composed image retrieval (CIR) mainly focus on learning a joint representation of the reference image and the modification text. The CLIP4CIR [5] model leverages CLIP [37] to encode images and texts and then uses MLPs to aggregate embeddings of the two modalities. To further enhance the modality representation, recent works [3, 34] have employed more powerful pretrained multimodal models such as BLIP [26] and BLIP-2 [25], yielding significant performance improvements. However, these methods rely on a single global vector for representation, which limits their ability to capture fine-grained details. To improve fine-grained CIR, TG-CIR [51] introduces both global and local attribute features with orthogonal regularization to learn more independent attribute features. ARTEMIS [8] and CaLa [19] propose two auxiliary methods leveraging image-text interactions in the CIR triplet to enhance query-target matching. Liu et al. [35] employ a two-stage approach, where the first stage uses a single global feature to filter out easy negatives, and the second stage leverages a dual-encoder architecture for fine-grained re-ranking. SPRC [3] proposes an additional sentence-level prompt for image-text fusion and a text prompt alignment loss to improve prompt learning.

One key challenge in CIR is the lack of high-quality supervised data. Existing CIR datasets, such as Fashion IQ [52], CIRR [33] and CIRCO [4], are significantly smaller than broader vision-language datasets like COCO [29] and LAION-5B [41]. To address this limitation, a line of work focuses on zero-shot CIR (ZS-CIR) [4, 21, 28, 40, 45, 54], aiming to develop generalized CIR models without the need for annotated data. ZS-CIR methods typically translate an image into text with a captioning model or textual inversion [11]. Yet, the performance gap between zero-shot methods and fully domain-adapted ones remains significant. Another type of approaches explores data augmentation [34] and synthetic data generation [10, 13, 22, 47]. BLIP4CIR+Bi [34] extends CIR datasets by adding reverse triplets, but results in less specific and less accurate modification texts. CompoDiff [13] uses an LLM to create new modification texts and generates the corresponding target images with a diffusion model[18, 39]. However, limitations in image generation models compromise the faithfulness and quality of the generated images. SPN [10], LaSCo [22], and CoVR-2 [47] instead only leverage VLMs and LLMs to generate modification texts for paired real images or videos. But these datasets focus more on general-domain images and fail to capture the fine-grained,

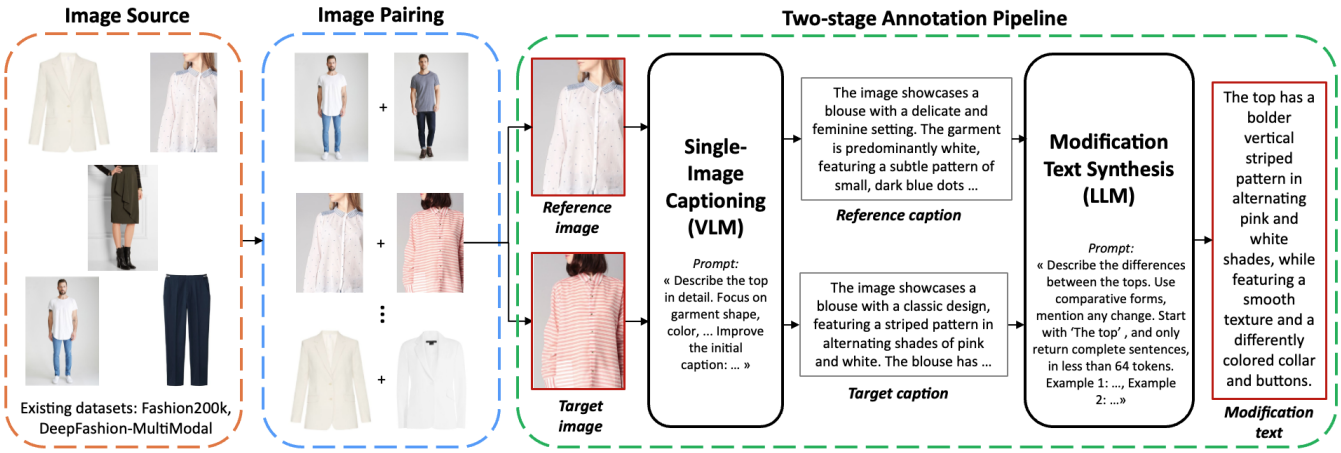



**Figure 2: The proposed data construction pipeline to automatically generate CIR triplets.** The images are extracted from large existing fashion datasets, then paired based on their visual similarity with images from the same product category. Then, our two-stage annotation process captions the images with a VLM, and an LLM generates a synthetic description of the changes applied on the reference image to obtain the target image.

fashion-specific vocabulary and visual details critical for fashion CIR tasks.

To improve fashion-domain CIR, recent efforts have focused on improving the pretraining of large multimodal models on fashion images. FashionViL [15] proposes a multi-view contrastive learning approach and pseudo-attributes classification to improve representation learning with fashion image-text pairs. Zhao et al. [56] proposes a progressive learning strategy, transitioning from general-domain pretraining to fashion domain pretraining. FAME-ViL [16] uses multi-task learning on heterogeneous fashion tasks, while UniFashion [55] further extends pretraining fashion datasets and tasks to include a broader range of multimodal generation and retrieval tasks, achieving state-of-the-art results in fashion CIR benchmarks. Nevertheless, existing fashion-focused pretraining mainly relies on image-text pairs rather than CIR triplets due to the scarcity of annotated triplets, limiting the model's ability to learn comparisons between two images. In this work, we address this gap by designing an automatic method to generate CIR triplets from fashion-domain images, and enhance pretraining efficiency for fashion CIR.


**(a) The dress is sleeker and more form-fitting, with a round neckline and a sleeveless design, featuring a denser floral pattern in black and white on a white base.**

**(b) The skirt is shorter and more form-fitting with a mini length, featuring a classic tartan plaid pattern in black, white, and red, and a smooth woven texture instead of the layered effect.**

**Figure 3: Examples from the FACap dataset.** The caption of each image pair corresponds to their modification text.

## 2.2 Large vision and language models

Recently, large language models (LLMs) like GPT [6] and LLaMA [9] have achieved remarkable success on various textual tasks like text generation and reasoning. Building on this foundation, numerous models have been developed to extend LLMs with visual perception capabilities by encoding images as inputs to the LLMs, resulting in powerful large vision-language models (VLMs) like BLIP-2 [25], LLaVA [31], GPT-4V [1], InternVL [7] and many more [12, 49]. These VLMs effectively combine textual and visual information and have set new benchmarks across diverse tasks such as image captioning [29], visual question answering [2] and so on.

While most VLMs are designed to process single-image inputs, recent advancements [1, 23, 24, 30] have aimed to improve multi-image capabilities. However, this progress introduces two key challenges. First, on the model side, handling multiple images significantly increases the token count, leading to issues with context length. To address this, various image token compression techniques [27, 43] have been proposed for VLMs. Second, on the data side, multi-view image datasets [30] remain limited, restricting the ability of current VLMs to excel in multi-image reasoning tasks. In this work, instead of directly using VLMs to generate modification texts for image pairs, we propose a two-stage pipeline that leverages the strengths of VLMs for detailed single-image captioning and LLMs for advanced text reasoning. This approach ensures high-quality annotations that precisely capture the fine-grained details essential for fashion CIR.

**Table 1: Comparison of different datasets.** We exclude certain web image sources to avoid licensing constraints, resulting in fewer unique images than the FACAD dataset [53]. FACAD also includes noisy web descriptions with unstandardized language, leading to a larger vocabulary size. Instead, the captions in our FACap are automatically cleaned and contain more details.

| | #Uniq imgs | Ann. type | Pair type | #Pairs | Vocab size | Avg. length |
|---|---|---|---|---|---|---|
| MSCOCO [29] | 328,000 | Manual | <img, caption> | 1,640,000 | 26,848 | 10.5 |
| FACAD [53] | 993,000 | Web | <img, caption> | 130,000 | 15,807 | 21 |
| FashionIQ [52] | 25,136 | Manual | <ref img, mod txt, tgt img> | 18,000 | 4,401 | 6.36 |
| **FACap (Ours)** | 227,680 | Auto | <ref img, mod txt, tgt img> | 227,680 | 9,273 | 23.38 |
| | | | <img, caption> | 227,680 | 18,689 | 82.90 |

## 3 The Fashion Automatic Caption Dataset

To tackle the data scarcity challenge in fashion-domain CIR, we introduce a large-scale Fashion Automatic Caption dataset (FACap), generated automatically using existing fashion image datasets and foundation models. It provides detailed image captions and CIR triplets with both global and fashion-specific vocabulary, so that CIR methods can leverage precise fine-grained textual and visual concepts.

**Table 2: Quality evaluation of FashionIQ and our FACap dataset.** We randomly sample 216 triplets across categories for each dataset and ask three annotators to measure data quality from three aspects with scale from 1 (worst) to 5 (best).

| | Faithfulness | Details | Saliency |
|---|---|---|---|
| FashionIQ [52] | **4.48 ± 0.64** | 3.03 ± 0.67 | 3.60 ± 0.69 |
| FACap | 4.40 ± 0.60 | **4.09 ± 0.64** | **4.29 ± 0.60** |

### 3.1 Dataset Construction

Our goal is to generate triplets of the form *<reference image, modification text, target image>* for Fashion CIR. Figure 2 illustrates the automatic data construction pipeline, including image source collection, image pairing, and our two-stage annotation using single-image captioning and modification text generation.

**Image sources.** We use two publicly available fashion datasets: Fashion200k [14] and DeepFashion-MultiModal [20], both originally crawled from online shopping websites. The Fashion200k dataset comprises approximately 200k images across five categories: dresses, jackets, pants, skirts, and tops. The images are accompanied by product descriptions, which, while useful, tend to be noisy. DeepFashion-MultiModal [20] is a refined version of the DeepFashion [32] dataset. It consists of 44,096 high-resolution model-worn images of clothing, each annotated with automatically-parsed attributes from product descriptions and manually-labeled shape and texture information. Note that images in both datasets are distinct from those used in the downstream datasets, ensuring there is no information leakage.

**Image pairing.** From this large image pool, we extract pairs of images to create a list of reference and target image pairs for the CIR task. We constrain the visual similarity of the image pairs: if two images are too different, the modification text will focus on describing the target image, ignoring the reference image. On the other hand, if two images are overly similar, it can be challenging for automatic systems to describe their subtle differences. To address this similarity range, we first filter out images according to the initial datasets' file structure, to exclude pairings of different views of the same item, thus enhancing the diversity of the CIR triplets. Next, we encode each image using the CLIP image encoder [37], and compute its cosine similarity with all other images in the same image source and category. Inspired by [33], we randomly select one image among the top-20 most similar images to form the image pair. This randomized selection enhances dataset diversity, preventing consistent pairing with the most similar images.

**Two-stage annotation** We aim to utilize large vision and language foundation models [7, 25, 26] to automatically annotate image pairs. However, currently, only a few VLMs are capable of accurately comparing two images in detail, and they often struggle to directly generate accurate modification texts from two images due to the scarcity of multi-image training data, as observed in our initial experiments. Therefore, we propose a two-stage approach to generate more accurate and detailed image pair annotations.

In the first stage, we use the open-source VLM model InternVL [7], due to its good performance and modest computational requirements. The maximum token length for generation is set to 128, allowing the creation of long captions that capture as many fine-grained details as possible. These detailed captions are key in enhancing the precision of the modification texts generated in the second stage. To improve captioning accuracy and mitigate hallucinations which could introduce wrong elements in the caption, we prompt InternVL with the image category and available product descriptions and attributes from the image source. Although these additional inputs may be noisy, they often provide valuable context. Processing the entire dataset takes about 41 GPU hours on Nvidia A100 GPUs. In the second stage, we use the proprietary LLM GPT-4o mini [1] [1] to synthesize modification texts, benefitting from GPT4's strong capabilities in text reasoning. To guide the model in generating short and concise modification texts, we use clear instructions along with two in-context examples. This ensures that the LLM focuses on the most significant changes between the reference and target images. Figure 3 shows examples from FACap across different fashion categories.

[1] https://platform.openai.com/docs/models/gpt-4o-mini#gpt-4o-mini

### 3.2 Quality evaluation

Given the large size of FACap, manually and exhaustively evaluating its quality is challenging. Therefore, we randomly sample 216 triplets from the FashionIQ and the same number from the FACap dataset to assess the quality of their modification texts. We evaluate the modification texts based on three key aspects:

- **Faithfulness**: Whether the modification text accurately describes the changes between the reference image and target images. Note that this criteria indirectly evaluates the presence of hallucinations generated by the VLM and LLM, as they lead to inaccurate differences.
- **Details**: Whether the modification text captures multiple elements present in the images.
- **Saliency**: Whether the modification text focuses on unique elements, reducing the number of possible false-positive target images. A vague text could have a high faithfulness value, but would score poorly on saliency.

Each criterion is manually scored by three annotators on a scale from 1 (worst) to 5 (best) for each randomly sampled triplet. The results are presented in Table 2. Notably, compared to the manually annotated FashionIQ dataset for CIR, our automatically constructed dataset exhibits even higher quality. The similar faithfulness values indicate that our pipeline's caption errors are comparable to the rate of mistakes made by human annotators, while improving the amount of details and the relevance of the texts for retrieval, as shown by the details and saliency values. This demonstrates the effectiveness of our annotation pipeline.

### 3.3 Dataset Statistics

Table 1 compares our FACap dataset with existing caption datasets in both the general and fashion domains. FACap offers two key advantages over existing CIR fashion datasets. First, it significantly expands the dataset's size in the fashion domain with minimal additional time and cost. The scale of FACap is closer to that of general-domain image-text datasets like MSCOCO [29], and we exclude other web sources to ensure our dataset can be publicly available. Second, FACap includes more accurate and detailed captions than existing datasets, as evidenced by our quality evaluation and average caption length. This can benefit fashion CIR tasks for fine-grained understanding of image-text alignment, particularly for fashion-related features and modifications, as illustrated in our qualitative results in figure 6.

## 4 The FashionBLIP-2 Model

### 4.1 Overall Framework

Given a reference image $I_r$ and modification text $T$, the objective of CIR is to retrieve the correct target image $I_t$ from an image database $\mathcal{D}$. The retrieved image $I_t$ should accurately reflect the specified modifications applied to $I_r$.

Figure 4 provides an overview of our FashionBLIP-2 model for the CIR task, which consists of three key modules: an image encoder for extracting image features, a light-weight Q-Former for compressing image features and performing multimodal fusion with text features, and a matching module for computing similarity between the query and the target image. The image encoder and Q-Former are adapted from the pretrained BLIP-2 model [25], to which we refer readers for a more detailed explanation.

Given $I_r$, the image encoder first extracts a feature map $f_r \in \mathbb{R}^{h\times w\times d_I}$, with $h, w$ the height and width of the encoded feature map and $d_I$ the feature dimensionality. Then, the Q-Former employs a set of learnable queries to distill $f_r$ into a compact set of embeddings $x_q \in \mathbb{R}^{n_q\times d_q}$ together with guidance from the modification text $T$, where $n_q \ll h \times w$. Similarly, each candidate image $I_c \in \mathcal{D}$ is sequentially processed by the image encoder and Q-Former but without any textual input, producing a corresponding set of embeddings $x_c \in \mathbb{R}^{n_q\times d_q}$ per image. Finally, the matching module takes the multimodal query embeddings $x_q$ and the candidate image embedding $x_c$ as inputs, computing a similarity score $s_{qc}$ between the query and the candidate image. During inference, the similarity between the multimodal query and all candidate images is computed. The candidate images are then ranked in descending order based on their similarity scores, resulting in the final retrieval list.

### 4.2 Adapter in Image Encoder

The BLIP-2 model [25] is initially trained on large-scale open-domain datasets, potentially reducing its effectiveness at capturing fine-grained visual details crucial to the fashion domain, such as features related to sleeve length or specific collar types. A straightforward approach to address this limitation is to fine-tune the BLIP-2 model alongside the CIR modules, to better adapt it to the fashion domain, but this may lead to high computational costs and catastrophic forgetting.

To overcome this challenge, we draw inspiration from [44] and introduce lightweight adapter modules into each transformer layer [50] of the image encoder. Instead of fine-tuning the entire BLIP-2 backbone, we freeze the pretrained weights in the image encoder and train only the newly introduced adapter modules along with the lightweight Q-Former. As illustrated in Figure 4, each adapter module comprises a downsampling layer, a non-linear operation, and an upsampling layer, with a residual link. Given an input $x \in \mathbb{R}^c$:

$$\text{Adapter}(x) = x + W_u\,(\sigma(W_d x)) \tag{1}$$

where $W_d \in \mathbb{R}^{c_b\times c}$, $W_u \in \mathbb{R}^{c\times c_b}$ are trainable parameters with $c_b \ll c$, and $\sigma$ denotes the GELU function [17].

This bottleneck architecture introduces a relatively small amount of additional parameters to the image encoder, ensuring a light-weight adaptation. Furthermore, the residual connection facilitates efficient gradient flow and helps preserve the original pretrained features, allowing the model to retain general-domain knowledge while learning fashion-specific details.

### 4.3 Multi-head Query-Candidate Matching

Previous works [3, 34] average the token embeddings $x_q$ and $x_c$ over the token dimension to obtain a single global vector for the query and each candidate image. However, this approach often suffers from the loss of fine-grained details crucial for retrieval.

To address this, we propose a multi-head query-candidate matching method based on a dual-level mixing operation like [46] to better capture fine-grained information. First, we use token mixing over the input tokens for $x_q \in \mathbb{R}^{n_q\times d_q}$, formulated as:

$$x_q^{tm} = W_{tm} \times x_q \tag{2}$$

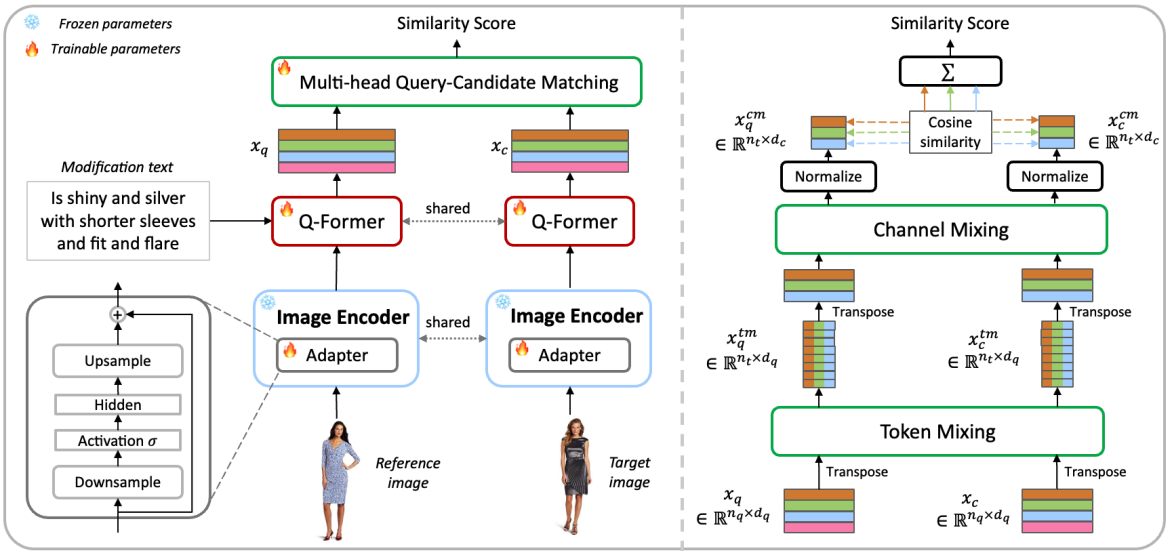



**Figure 4: Overview of the FashionBLIP-2 model.** Left: The input images are encoded using a pretrained image encoder with adapter modules, and further processed by a Q-Former module. The similarity between the two obtained representations is computed using multi-head query-candidate matching module. Right: Details of the matching module. The number of tokens and token dimensionality is reduced by token mixing and channel mixing respectively. The final similarity score is the sum of the cosine similarity for each paired vector.

where $W_{tm} \in \mathbb{R}^{n_t \times n_q}$ is a trainable parameter. Here, $n_t < n_q$ to reduce the redundancy across embeddings in the initial representation $x_q$, while retaining multiple vectors to capture multiple aspects of the inputs. We then perform channel mixing for each vector to project $x_q^{tm}$ into a lower-dimensional space:

$$x_q^{cm} = x_q \times W_{cm} \tag{3}$$

where $W_{cm} \in \mathbb{R}^{d_q \times d_c}$ with $d_c < d_q$. In order to encourage projecting $x_q$ and $x_c$ into a common low-dimensional embedding space, we use the same parameters $W_{tm} = W_{cm}$ to process $x_q$ and $x_c$.

Each row vector in $x_q^{cm}$ and $x_c^{cm}$ is viewed as one head for matching. The final similarity is the sum of the cosine similarities for each head as follows:

$$s_{qc} = \text{sim}\left(x_q^{cm}, x_c^{cm}\right) = \sum_{i=1}^{n_t} \frac{x_{q,i}^{cm} \cdot x_{c,i}^{cm}}{\|x_{q,i}^{cm}\|_2 \cdot \|x_{c,i}^{cm}\|_2} \tag{4}$$

## 4.4 Training

We train the FashionBLIP-2 model in two stages.

**Stage 1: Training on FACap.** The first stage aims to fine-tune a general-domain model to the fashion domain, to learn fine-grained visual and text representations. Since FACap contains both CIR triplets and image-caption pairs, we use two tasks in stage 1 training: the primary CIR task and an auxiliary Composed Text Retrieval (CTR) task.

For the CIR task, we employ the widely-used contrastive loss:

$$\mathcal{L}_{\text{CIR}} = -\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{\exp(x_i y_i)}{\exp(x_i y_i) + \sum_{\hat{y} \in \mathcal{N}_i} \exp(x_i \hat{y})}\right) \tag{5}$$

with $(x_i, y_i)$ a positive pair, and $\mathcal{N}_i$ the set of negative pairs. Here, the negative pairs correspond to the reference image $x_i$ and any target image other than $y_i$ in the batch.

The CTR task retrieves a target text corresponding to the target image from a text pool, rather than the target image as in CIR. This auxiliary task helps the model align the fused query embedding more effectively with the textual representation, complementing its alignment with the image representation in CIR task. The contrastive loss used for CTR task, $\mathcal{L}_{\text{CTR}}$, is defined as in Eq 5.

We fine-tune the whole FashionBLIP-2 except for the original image encoder, using the combined loss function $\mathcal{L}_{\text{CIR}} + \mathcal{L}_{\text{CTR}}$.

**Stage 2: Fine-tuning on downstream Fashion CIR dataset.** The second stage fine-tunes the FashionBLIP-2 model on the downstream dataset to further improve its performance. However, since FashionIQ does not contain image-caption pairs, we train the model exclusively with $\mathcal{L}_{\text{CIR}}$. Additionally, we freeze the image encoder and its adapter modules to preserve the fashion-domain knowledge learnt during the first stage of training.

# 5 Experiment

## 5.1 Experimental setup

**Datasets.** We use the FashionIQ dataset [52] for evaluation, which is the most widely used Fashion CIR dataset. With 18,000 training triplets and 6,016 validation triplets, it covers three categories: Dress, Shirt, and Toptee. Each reference and target image pair contains two manually annotated modification texts. Following previous works [5, 10, 34, 35, 55, 56], we concatenate the two annotated texts with an "and" word to form a single modification text, and evaluate the models on the validation split.

However, since the annotations in FashionIQ are noisy as shown in Figure 1 and Table 2, we enhance its quality by applying our

**Table 3: Results on the FashionIQ validation split for composed image retrieval, under two settings: with and without fine-tuning the model on FashionIQ.** Best and second-best results in each setting are highlighted in bold and underlined, respectively.

| Setting | Model | Dresses | | Shirts | | Tops&tees | | Averages | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | R@10 | R@50 | R@10 | R@50 | R@10 | R@50 | R@10 | R@50 | Avg. |
| Without fine-tuning | Pic2Word [40] | 20.00 | 40.20 | 26.20 | 43.60 | 27.90 | 47.40 | 24.70 | 43.70 | 34.22 |
| | SEARLE (ViT-L/14) [4] | 20.48 | 43.13 | 26.89 | 45.58 | 29.32 | 49.97 | 25.56 | 46.23 | 35.90 |
| | Context-I2W [45] | 23.1 | 45.3 | 29.7 | 48.6 | 30.6 | 52.9 | 27.8 | 48.9 | 39.37 |
| | FTI4CIR [28] | 24.39 | 47.84 | 31.35 | 50.59 | 32.43 | 54.21 | 20.39 | 50.88 | 40.14 |
| | LDRE (ViT-G/14) [54] | <u>26.11</u> | <u>51.12</u> | **35.94** | **58.58** | <u>35.42</u> | <u>56.67</u> | <u>32.49</u> | **55.46** | <u>43.97</u> |
| | FashionBLIP-2 (ours) | **32.52** | **53.25** | <u>34.79</u> | <u>52.40</u> | **36.66** | **58.13** | **34.66** | <u>54.59</u> | **44.63** |
| With fine-tuning | CLIP4CIR [5] | 33.81 | 59.40 | 39.99 | 60.45 | 41.41 | 65.37 | 38.32 | 61.74 | 50.03 |
| | BLIP4CIR+Bi [34] | 42.09 | 67.33 | 41.76 | 64.28 | 46.61 | 70.32 | 43.49 | 67.31 | 55.40 |
| | BLIP2-Cir [19] | 41.57 | 66.02 | 46.86 | 66.00 | 49.44 | 72.25 | 45.96 | 68.09 | 57.02 |
| | TG-CIR [51] | 45.22 | 69.66 | 52.60 | 72.52 | 56.14 | 77.10 | 51.32 | 73.09 | 58.05 |
| | FAME-ViL [16] | 42.19 | 67.38 | 47.64 | 68.79 | 50.69 | 73.07 | 46.84 | 69.75 | 58.29 |
| | Re-ranking [35] | 48.14 | 71.43 | 50.15 | 71.25 | 55.23 | 76.80 | 51.17 | 73.13 | 62.15 |
| | SPRC [3] | 49.18 | 72.43 | 55.64 | 73.89 | <u>59.35</u> | 78.58 | 54.92 | 74.97 | 64.85 |
| | UniFashion [55] | **53.72** | **73.66** | **61.25** | **76.67** | **61.84** | **80.46** | **58.93** | **76.93** | **67.93** |
| | FashionBLIP-2 (ours) | <u>51.41</u> | <u>73.53</u> | <u>57.02</u> | <u>75.32</u> | 58.95 | <u>79.60</u> | <u>55.79</u> | <u>76.15</u> | <u>65.97</u> |

automatic annotation process to the images from the Fashion IQ validation split. We create a triplet for each unique image in the validation split and generate a total of 15,536 CIR triplets, which we name enhFashionIQ, for fine-grained CIR evaluation.

**Evaluation metrics.** We use Recall@$k$ (with $k \in \{10; 50\}$ similarly to previous works) as the main metric. It computes the percentage of target images that appear in the top-$k$ retrieved images list. The recalls are computed for each category: dress, shirt, and toptee for FashionIQ and enhFashionIQ. We also report the average recall.

**Experiment settings.** We evaluate models under two settings:

- without fine-tuning setting: the model is only trained on our FACap dataset and then evaluated on downstream CIR datasets. This setting evaluates the generalization capacity of the model on a previously unseen fashion dataset, and its performance is compared to zero-shot methods [4, 21, 28, 40, 45, 54].
- fine-tuning setting: the model is fine-tuned on FashionIQ, and evaluated on FashionIQ and enhFashionIQ.

**Implementation details** We use the ViT-G version of the pre-trained BLIP-2 [25] model. For the adapter module in image encoder, we use a downsampling factor of 16. The Q-Former module is parametrized to take textual inputs of 128 tokens and $n_q = 32$ query tokens, and outputs 32 token embeddings with dimensionality of $d_q = 768$. The token mixing layer in our multi-head query-target matching reduces the 32 tokens to $n_t = 12$ tokens and channel dimension from 768 to $d_c = 256$. We run our experiments on NVIDIA H100 GPUs with batch size of 512 and AdamW optimizer.

## 5.2 Comparison with state-of-the-art methods

Table 3 presents the evaluation results on FashionIQ under the two settings - without and with fine-tuning. In the upper block, we compare the FashionBLIP-2 model, trained only on our FACap dataset, with zero-shot methods [4, 28, 40, 45, 54] to compare their generalization capacity on fashion data. Our model achieves an improvement over the state-of-the-art method LDRE [28], which uses pre-trained LLMs. The average gain is 0.66 absolute points with 2.17 in R@10, highlighting its ability to retrieve relevant images on a previously unseen fashion dataset. The most pronounced improvement is observed in the dress category, known for its high diversity in descriptions such as their length, pattern, neckline, and way of wearing, further demonstrating the effectiveness of our proposed dataset and approach.

The bottom section of Table 3 shows the comparison between our FashionBLIP-2 and existing methods [3, 5, 16, 19, 34, 35, 51, 55] fine-tuned on the FashionIQ dataset. Our FashionBLIP-2 achieves the second best results on average, only under-performing UniFashion [55] which utilizes more image-caption pairs —about 280k pairs —and generation tasks in training rather than CIR triplets. As our FACap dataset and method are complementary to UniFashion, we will leave it to future work.

## 5.3 Ablation study

**Pretraining on the FACap dataset.** In Table 4, we evaluate the contribution of the proposed FACap dataset using our FashionBLIP-2 model, SPRC [3] and UniFashion [55]. The evaluation is conducted on both FashionIQ and our enhFashionIQ, containing more fine-grained annotations. First, almost all models benefit from pre-training on FACap, improving performance on the two evaluation datasets. Second, our FACap dataset provides greater benefits to the FashionBLIP-2 model. This advantage is attributed to our model's multi-head matching mechanism, which effectively leverages the fine-grained details in FACap, whereas SPRC struggles to utilize such detailed information due to its reliance on global embeddings. Finally, FashionBLIP-2 achieves a better performance than SPRC on the FashionIQ split, and it demonstrates significantly higher

**Table 4: Impact of FACap pretraining on SPRC [3], UniFashion [55] (code reproduction) and our method.** The averaged Recall is reported for FashionIQ and enhFashionIQ validation splits.

| Model | Pretrain on FACap | Fine-tune on FIQ | FIQ | enhFIQ |
|---|---|---|---|---|
| SPRC [3] | ✗ | ✓ | 64.85 | 79.59 |
| | ✓ | ✓ | 64.87 | 80.29 |
| UniFashion [55]* | ✗ | ✓ | <u>65.34</u> | 81.97 |
| | ✓ | ✓ | 64.51 | <u>87.30</u> |
| FashionBLIP-2 (Ours) | ✗ | ✓ | 64.46 | 80.32 |
| | ✓ | ✓ | **65.97** | **87.93** |

* Results obtained using UniFashion's released code

**Table 5: Ablation study of different components in FashionBLIP-2 model.** The averaged Recall is reported for FashionIQ and enhFashionIQ validation splits under two settings. The acronym "MH" denotes the proposed multi-head matching.

| Training tasks | Adapter | Matching | no fine-tuning | | w/ fine-tuning | |
|---|---|---|---|---|---|---|
| | | | FIQ | enhFIQ | FIQ | enhFIQ |
| CIR | ✗ | Global | 41.04 | 87.81 | 64.36 | 86.42 |
| CIR | ✓ | Global | 42.62 | 88.20 | 64.38 | 86.75 |
| CIR | ✓ | MH | 43.31 | 89.14 | **65.97** | **87.93** |
| CIR+CTR | ✓ | MH | **44.63** | **89.45** | 65.62 | 86.93 |

improvements on enhFashionIQ under the same training configuration. This highlights the superior ability of our model to handle fine-grained fashion retrieval tasks. In addition, we observe that fine-tuning our model on FashionIQ does not degrade much its performance on enhFashionIQ compared to SPRC. This indicates that our method is more robust to noisy datasets, without losing its fine-grained performance.

**FashionBLIP-2 components.** Table 5 analyzes the individual contributions of each component in our FashionBLIP-2 model. The first row serves as the baseline, representing a model built on top of BLIP-2. In the second row, we specialize the image encoder with adapter modules, improving performance across both datasets and evaluation settings. The third row incorporates the proposed multi-head query-candidate matching mechanism. This boosts retrieval performance by enabling the model to capture and compare finer details between queries and candidates. Finally, the fourth row integrates the auxiliary CTR task during training on the FACap dataset. While it improves results in the setting without fine-tuning on FashionIQ, it decreases performance when fine-tuned on FashionIQ. We hypothesize that the CTR task may introduce a bias towards detailed textual descriptions, which might hinder adaptation to noisier datasets like FashionIQ.

**Size of the training data.** To investigate the impact of data quantity, we train FashionBLIP-2 on progressively larger subsets of FACap and evaluate the resulting models on FashionIQ and enhFashionIQ. As shown in Figure 5, the performance of our model improves with the increasing size of the training dataset across both

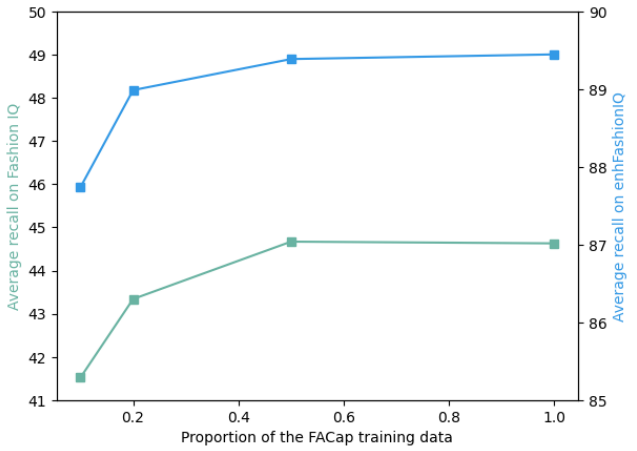



**Figure 5: Results on FashionIQ dataset without fine-tuning using different sizes of the FACap dataset.**

evaluation benchmarks. This highlights the critical role of having a large volume of diverse image-text pairs to effectively learn fine-grained multimodal representations. However, the performance gain from training on 50% to 100% of the dataset is relatively small: while data quantity is important, further improvements may require focusing on data quality and diversity rather than sheer volume.

## 5.4 Qualitative Results

Figure 6 presents qualitative results of FashionBLIP-2 on the Fashion IQ and enhFashionIQ validation data. The first two rows show that whether the modification text is precise (enhFashionIQ) or not (Fashion IQ), our model is able to combine it with characteristics of the reference images (for example clothing length and color). The third row presents a failure case of our model, revealing the difficulty of handling false negative examples: the correct target image is badly ranked, but all the top-3 retrieved images respect the given modification text and the information from the reference image (color and sleeves).

# 6 Conclusion

We have proposed two enhancements to tackle shortcomings of CIR in the fashion domain. Firstly, we designed an automatic pipeline to build a large-scale high-quality CIR dataset from a large list of images with noisy captions. Leveraging the strengths of a VLM and a LLM, this pairing and annotation method provides accurate modification texts, while adding relevant fashion details and focusing on salient changes. This method has allowed us to construct FACap, a higher quality dataset for fashion CIR. Secondly, we have introduced FashionBLIP-2, a method combining BLIP-2's general-domain comprehensive strength with an adapter module to adjust it to the fashion domain, and a new multi-head query-candidate matching mechanism to focus on fine-grained details and benefit from FACap high-quality captioning. Experiments show that FashionBLIP-2 trained on FACap outperforms state-of-the-art methods without fine-tuning on the downstream dataset. It also reaches competitive performance after fine-tuning on FashionIQ,


**Figure 6: Qualitative results of FashionBLIP-2 on Fashion IQ (rows 1 and 3) and enhFashionIQ (row 2).** The rank of the ground-truth image (framed in green) among the retrieved results is specified on the right.

making it well-suited for fast adaptation in the fashion domain, excelling in fine-grained retrieval tasks while remaining robust to vague modification texts.

## Acknowledgments

This project was granted access to the HPC resources of IDRIS under the allocation AD011015247R1 made by GENCI. It was funded in part by the French government under management of Agence Nationale de la Recherche as part of the "France 2030" program, reference ANR-23-IACL-0008 (PR[AI]RIE-PSAI project), and Paris Île-de-France Région in the frame of the DIM AI4IDF.

## References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.

[3] Yang Bai, Xinxing Xu, Yong Liu, Salman Khan, Fahad Khan, Wangmeng Zuo, Rick Siow Mong Goh, Chun-Mei Feng, et al. 2024. Sentence-level Prompts Benefit Composed Image Retrieval. In *The Twelfth International Conference on Learning Representations*.

[4] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. 2023. Zero-shot composed image retrieval with textual inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15338–15347.

[5] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2022. Conditioned and composed image retrieval combining and partially fine-tuning clip-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4959–4968.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[7] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24185–24198.

[8] Ginger Delmas, Rafael Sampaio de Rezende, Gabriela Csurka, and Diane Larlus. 2022. Artemis: Attention-based retrieval with text-explicit matching and implicit similarity. *ICLR* (2022).

[9] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[10] Zhangchi Feng, Richong Zhang, and Zhijie Nie. 2024. Improving composed image retrieval via contrastive learning with scaling positives and negatives. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 1632–1641.

[11] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. [n. d.]. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *The Eleventh International Conference on Learning Representations*.

[12] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, Jianfeng Gao, et al. 2022. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision* 14, 3–4 (2022), 163–352.

[13] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, HeeJae Jun, Yoohoon Kang, and Sangdoo Yun. 2023. CompoDiff: Versatile Composed Image Retrieval With Latent Diffusion. *arXiv preprint arXiv:2303.11916* (2023).

[14] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. 2017. Automatic spatially-aware fashion concept discovery. In *Proceedings of the IEEE international conference on computer vision*. 1463–1471.

[15] Xiao Han, Licheng Yu, Xiatian Zhu, Li Zhang, Yi-Zhe Song, and Tao Xiang. 2022. Fashionvil: Fashion-focused vision-and-language representation learning. In *European conference on computer vision*. Springer, 634–651.

[16] Xiao Han, Xiatian Zhu, Licheng Yu, Li Zhang, Yi-Zhe Song, and Tao Xiang. 2023. Fame-vil: Multi-tasking vision-language model for heterogeneous fashion tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2669–2680.

[17] Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415* (2016).

[18] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-Prompt Image Editing with Cross Attention Control. arXiv:2208.01626 [cs.CV] https://arxiv.org/abs/2208.01626

[19] Xintong Jiang, Yaxiong Wang, Mengjian Li, Yujiao Wu, Bingwen Hu, and Xueming Qian. 2024. CaLa: Complementary Association Learning for Augmenting Comoposed Image Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2177–2187.

[20] Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. 2022. Text2Human: Text-Driven Controllable Human Image Generation. *ACM Transactions on Graphics (TOG)* 41, 4, Article 162 (2022), 11 pages. https://doi.org/10.1145/3528223.3530104

[21] Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. 2023. Vision-by-language for training-free compositional image retrieval. *arXiv preprint arXiv:2310.09291* (2023).

[22] Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. 2024. Data roaming and quality assessment for composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 2991–2999.

[23] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326* (2024).

[24] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895* (2024).

[25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.

[26] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*. PMLR, 12888–12900.

[27] Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. 2024. Tokenpacker: Efficient visual projector for multimodal llm. *arXiv preprint arXiv:2407.02392* (2024).

[28] Haoqiang Lin, Haokun Wen, Xuemeng Song, Meng Liu, Yupeng Hu, and Liqiang Nie. 2024. Fine-grained textual inversion network for zero-shot composed image retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 240–250.

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. 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, 740–755.

[30] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/

[31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems* 36 (2024).

[32] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1096–1104.

[33] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. 2021. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2125–2134.

[34] Zheyuan Liu, Weixuan Sun, Yicong Hong, Damien Teney, and Stephen Gould. 2024. Bi-directional training for composed image retrieval via text prompt learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 5753–5762.

[35] Zheyuan Liu, Weixuan Sun, Damien Teney, and Stephen Gould. 2024. Candidate Set Re-ranking for Composed Image Retrieval with Dual Multi-modal Encoder. *Transactions on Machine Learning Research* (2024). https://openreview.net/forum?id=fJAwemcvpL

[36] Dong Huk Park, Trevor Darrell, and Anna Rohrbach. 2019. Robust change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4624–4633.

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[38] Jun Rao, Fei Wang, Liang Ding, Shuhan Qi, Yibing Zhan, Weifeng Liu, and Dacheng Tao. 2022. Where Does the Performance Improvement Come From? -A Reproducibility Concern about Image-Text Retrieval. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*. 2727–2737.

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

[40] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. 2023. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19305–19314.

[41] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems* 35 (2022), 25278–25294.

[42] Shihao Shao, Kaifeng Chen, Arjun Karpur, Qinghua Cui, André Araujo, and Bingyi Cao. 2023. Global features are all you need for image retrieval and reranking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11036–11046.

[43] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. 2024. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434* (2024).

[44] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. 2022. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5227–5237.

[45] Yuanmin Tang, Jing Yu, Keke Gai, Jiamin Zhuang, Gang Xiong, Yue Hu, and Qi Wu. 2024. Context-I2W: Mapping Images to Context-dependent Words for Accurate Zero-Shot Composed Image Retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 5180–5188.

[46] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. 2021. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems* 34 (2021), 24261–24272.

[47] Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. 2024. CoVR-2: Automatic Data Construction for Composed Video Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

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

[49] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. GIT: A Generative Image-to-text Transformer for Vision and Language. *arXiv preprint arXiv:2205.14100* (2022).

[50] A Waswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, A Gomez, L Kaiser, and I Polosukhin. 2017. Attention is all you need. In *NIPS*.

[51] Haokun Wen, Xian Zhang, Xuemeng Song, Yinwei Wei, and Liqiang Nie. 2023. Target-guided composed image retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*. 915–923.

[52] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. 2021. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*. 11307–11317.

[53] Xuewen Yang, Heming Zhang, Di Jin, Yingru Liu, Chi-Hao Wu, Jianchao Tan, Dongliang Xie, Jue Wang, and Xin Wang. 2020. Fashion Captioning: Towards Generating Accurate Descriptions with Semantic Rewards. In *ECCV*.

[54] Zhenyu Yang, Dizhan Xue, Shengsheng Qian, Weiming Dong, and Changsheng Xu. 2024. LDRE: LLM-based Divergent Reasoning and Ensemble for Zero-Shot Composed Image Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 80–90.

[55] Xiangyu Zhao, Yuehan Zhang, Wenlong Zhang, and Xiao-Ming Wu. 2024. Unified Vision-Language Model for Multimodal Fashion Retrieval and Generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 1490–1507. https://doi.org/10.18653/v1/2024.emnlp-main.89

[56] Yida Zhao, Yuqing Song, and Qin Jin. 2022. Progressive learning for image retrieval with hybrid-modality queries. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1012–1021.