Composed Image Retrieval in Fashion Using CLIP-based Multimodal Fusion

Chen Jing Wong

Master of Science
School of Informatics
University of Edinburgh
2024

Abstract

This research addresses the challenges of Composed Image Retrieval (CIR) in the fashion domain, where the rapid turnover of products and fragmented online inventories present unique difficulties for developing effective AI/ML solutions. Specifically, these challenges require systems that can accurately integrate and process multimodal queries—combining visual and textual information—to retrieve relevant images. The existing literature on CIR presents a variety of methods, but the state of the art remains unclear due to the fragmented nature of contributions and the uncertain value of different approaches. The central research question of this study is to identify the most effective approach for CIR by dissecting and evaluating the key components of existing methodologies.

To address this question, the study conducts an extensive ablation analysis of state-of-the-art CIR techniques, focusing on the evaluation of linear and non-linear multimodal fusion strategies and the implementation of a two-stage training approach. The experiments, performed on benchmark fashion datasets such as FashionIQ and Shoes, aim to clarify the impact of raw-data level multimodal fusion—particularly through visual and textual query unification—on retrieval performance. The findings indicate that a two-stage training scheme, when applied to these fusion techniques, significantly enhances retrieval accuracy. This research not only formulates a clearer understanding of the current state of the art in CIR but also offers actionable insights into best practices for developing more robust and efficient CIR systems. The study's conclusions contribute to advancing the field and provide a solid foundation for future research and practical applications in the fashion industry.

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Chen Jing Wong*)

Acknowledgements

I would like to express my deepest gratitude to my academic supervisor, Dr. Pavlos Andreadis, and my industrial supervisor, Dr. David Wood, for their invaluable guidance, time, and expertise throughout the course of this research. Their commitment to this project and weekly meetings have been instrumental in its development and success. I am particularly thankful to Moonsift for generously funding the computational resources necessary for this study. The collaboration with Moonsift has been pivotal in facilitating the advanced computational experiments required for this research. I am sincerely appreciative of the support, both academic and industrial, that I have received, and I am grateful for the opportunity to work with such knowledgeable and dedicated individuals. This research would not have been possible without their commitment and insightful contributions.

Table of Contents

Chapter 1

Introduction

1.1 Introduction

Composed Image Retrieval (CIR) is an emerging approach designed to retrieve images that closely match a source image combined with textual modifications provided by users. CIR holds potential for significantly advancing search systems within the fashion domain, where rapid product turnover caused by ”Fast Fashion” [7] and dispersed online inventories across different online retailers pose substantial challenges [44]. The need for a search system that can reduce consumers' substantial amounts of time researching products alongside changing inventories is critical. CIR's unique combination of visual and textual input offers a promising solution for improving search efficiency and user satisfaction in the fashion industry. CIR shows promise based on existing studies, and current research in this area primarily focuses on developing and refining both linear and non-linear multimodal fusion techniques and training strategies to improve retrieval accuracy.

However, the current research landscape for CIR reveals a gap in identifying the most effective components and methodologies due to the fragmented presentation of solutions across various studies, remaining a need for comprehensive evaluations that critically assess the limitations and practical effectiveness of these methods within the fashion domain. This research aims to critique existing works and synthesize their findings to identify the true state of the art in CIR. By conducting a comprehensive ablation study, this research seeks to explore effective approaches for CIR, transcending the invention of new models from various papers. This approach can not only identify best practices but also push the boundaries of what is considered state-of-the-art in CIR. Through thorough scientific investigation and critique, this study explores and validates

potential components to contribute to the development of more effective and robust CIR systems that enhance the online shopping experience.

The hypothesis guiding this research is that the integration of individual components from recent CIR methodologies, particularly those emphasizing raw-data-level multimodal fusion and the use of Visual Language Pretrained (VLP) models, like Contrastive Language-Image Pre-Training (CLIP) [49], can significantly enhance CIR performance in the fashion domain. Specifically, this research proposes that shifting the multimodal fusion process from the traditional feature level to the raw data level, coupled with a two-stage training approach, will result in improved retrieval accuracy and robustness. The hypothesis states that the use of CLIP additivity fine-tuning, combined with linear multimodal query fusion, will better align textual and visual features within the same embedding space, thereby optimizing the effectiveness of CIR systems. This approach is expected to leverage the inherent strengths of VLP models, particularly their cross-modal alignment capabilities, to create a more cohesive and precise retrieval process. Given this hypothesis, the research is further divided into specific investigative approaches to systematically study and validate the most effective methodologies for CIR in the fashion domain:

1.1.1 Research Question

- What constitutes the most effective approach for Composed Image Retrieval in the fashion domain?

1.1.2 Investigative Approach

- **Critique and Synthesize Existing CIR Methods**: To test the hypothesis, this approach involves performing experiments on benchmark fashion datasets, such as FashionIQ [64] and Shoes [20], to validate the proposed methods. The experiments will utilize metrics such as recall to compare the performance of the proposed CIR methodologies with state-of-the-art models, thereby identifying the strengths and limitations of these methods.

- **Evaluate the Effectiveness of Raw-Data Level Multimodal Fusion and Image Preprocessing Techniques**: Building on the hypothesis that raw-data level fusion and preprocessing can enhance retrieval accuracy, this approach will investigate the impact of techniques such as visual and textual query unification, alongside

various image preprocessing methods (e.g., standard CLIP preprocessing, target pad), on CIR performance.

- **Analyze Non-Linear vs. Linear Multimodal Fusion Models**: As part of validating the hypothesis, this approach will focus on assessing the impact of non-linear multimodal fusion models compared to linear models, specifically examining how these models influence retrieval performance.

- **Develop a Two-Stage Training Approach**:

  - **Stage 1**: Implement CLIP additivity fine-tuning to adapt the image and text encoders for better alignment with CIR tasks, thereby addressing the mismatch between large-scale pre-training and downstream task requirements as hypothesized.

  - **Stage 2**: Develop a linear multimodal query fusion technique using a Multi-Layer Perceptron (MLP) to combine the features of unified textual and visual queries, ensuring that the fused multimodal features remain within the original embedding space, as suggested by the hypothesis.

- **Identify Best Practices and Guidelines for CIR**: The final approach aims to synthesize findings from the experimental results to propose effective strategies for developing robust CIR systems. This will advance the field of CIR in the fashion domain by evaluating the effectiveness of innovative CIR methodologies and image preprocessing techniques, ultimately validating or refining the initial hypothesis.

1.1.3 Contributions

This research will contribute to the state-of-the-art in CIR by providing a thorough critique of existing methodologies and offering new insights into best practices for developing effective CIR systems. A key innovation in this study is the introduction of GradCAM as a qualitative metric to enhance AI explainability. GradCAM will be used to visualize and interpret how CLIP models process and output information, enabling a deeper analysis of whether the model is extracting useful features from the images. Additionally, this research will be the first to employ the t-SNE algorithm as a quantitative metric to analyze how embeddings are learned by the model and how they navigate toward the desired embedding in the embedding space. These visualizations

will provide critical insights into the effectiveness of different fusion strategies and training approaches. The findings from this study will serve as a foundation for future research and practical applications in the fashion industry, offering robust methodologies for developing more intuitive and efficient CIR systems.

1.1.4 Outline

The Introduction Section 1 sets the stage by defining the problem of CIR and articulating the research questions and objectives. This is followed by the Background Section 2, which reviews existing literature on image retrieval, focusing particularly on how CIR advances traditional approaches. The Methodology Section 3 then details the experimental design, explaining the choice of datasets, models, and evaluation metrics. This is directly linked to the Experimental Results Section 4, where the outcomes of the experiments are presented and analyzed, showcasing the effectiveness of the proposed approaches. The Conclusion Section 5 summarizes the key findings and contributions of this study, while the Impacts Section 6 outlines the practical applications and broader implications of the research in both academic and industrial contexts. Finally, Section 7 addressed the limitations, and future research directions are proposed.

Chapter 2

Background

2.1 Introduction to Image Retrieval (IR)

2.1.1 Overview of Image Retrieval Techniques

Image Retrieval (IR) refers to the process of searching and retrieving images from large databases based on visual content. It is a critical function in various applications, including search engines, digital libraries, and e-commerce. Traditional Content-Based Image Retrieval (CBIR) systems primarily rely on visual features like color, texture, and shape extracted from images to perform searches.

2.1.2 Traditional Content-Based Image Retrieval (CBIR)

While CBIR techniques have evolved significantly—transitioning from the use of engineered features such as Scale-Invariant Feature Transform (SIFT) [73] to the adoption of Convolutional Neural Networks (CNNs) [35]. However, CBIR is limited by its reliance on a single modality — visual content, leading to challenges when users need to express more complex search queries that involve textual elements.

2.1.3 Composed Image Retrieval (CIR) as an Advancement Over CBIR

This dual-modality approach, illustrated in Figure 2.1, allows for more nuanced and specific searches by integrating visual and textual information to retrieve images that closely match the user's desired modifications [39, 47, 48]. In the CIR process, the source image and the modification text are encoded separately through CLIP image and

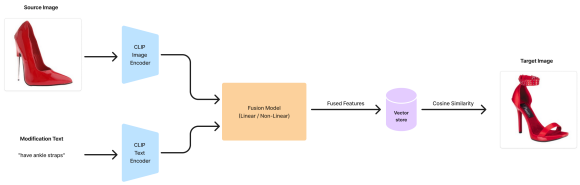

Figure 2.1: Example of Composed Image Retrieval (CIR) process

text encoders. The encoded features are then fused using a fusion model, which can be either linear or non-linear. The target image is retrieved by calculating the cosine similarity between the fused features and the stored vectors, ensuring that the retrieved image aligns with the modifications specified by the user.

The task of CIR distinguishes itself from traditional CBIR by allowing users to express their search intentions through a combination of both a source image and a textual modification, rather than being restricted to using either a pure text query or a visual query alone [39, 47, 48]. The limitations of CBIR, particularly its inability to handle complex queries that combine visual and textual elements, underscore the need for CIR. CIR's ability to process and fuse information from both modalities offers a significant advancement in retrieval accuracy and user satisfaction, particularly in domains like fashion, where nuanced modifications to an image are often necessary to meet user expectations.

2.2 Approaches for Feature Extraction

2.2.1 Traditional Model-Based Approaches

The feature extraction process involves mapping important characteristics of the data into separate embedding spaces, tailored to each modality (e.g., one space for visual features and another for textual features). Traditional model-based methods in feature extraction typically involve models like ResNet [22] and Long Short-Term Memory (LSTM) networks [24]. These models are trained on single-modality datasets to extract features from images or text. Each type of data (modality) — whether image or text — is usually mapped into its own separate space.

2.2.2 Vision-Language Pretrained (VLP) Models

Vision-Language Pretrained (VLP) models have emerged as a powerful tool for CIR due to their ability to handle both visual and textual information simultaneously. VLP models like CLIP [49], ALIGN [26], and BLIP [33] represent a significant advancement over traditional approaches. By integrating these modalities into a shared embedding space, these models can perform complex cross-modal retrieval tasks with high accuracy, because they are pre-trained on vast datasets that include both images and their corresponding natural language descriptions, enabling them to encode both visual and textual information into a single, shared embedding space.

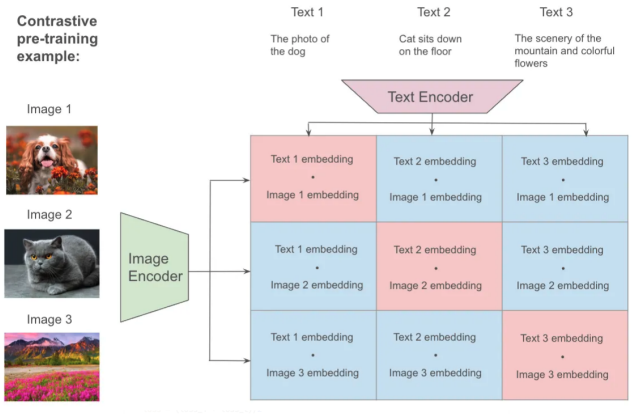

Figure 2.2: The illustration of the contrastive pre-training of CLIP models [67]

CLIP models are designed to align images and their corresponding text captions in a unified embedding space, making them particularly well-suited for cross-modal tasks like CIR. CLIP consists of two main components: an image encoder ψ_I and a text encoder ψ_T. Given an image I, the image encoder extracts a feature representation $\psi_I(I) \in \mathbb{R}^d$, where d is the embedding dimension, and similarly, given a text query T, the text encoder generates a corresponding feature representation $\psi_T(T) \in \mathbb{R}^d$. CLIP uses contrastive pre-training, where paired images and text descriptions (e.g., an image of a dog with the caption ”The photo of a dog”) are aligned in the embedding space if they are a correct pair and pushed apart if they are not. This training process shown in Figure 2.2 involves creating embeddings for both images and texts and ensuring that correct pairs (e.g., Image 1 with Text 1) are aligned closely in the embedding space, while incorrect pairs (e.g., Image 1 with Text 2) are not. This approach enables CLIP to learn robust cross-modal representations, essential for CIR tasks that require effective integration and retrieval of information across both visual and textual modalities.

2.2.3 Comparative Analysis of VLP Models vs. Traditional Models

Textual and visual features from traditional models are created in different spaces, and combining or comparing features across different types of data (e.g. matching text with images) can be challenging, often leading to less effective and suboptimal retrieval performance. While traditional models struggle with the alignment of features across different modalities, VLP models excel in this area due to their extensive training on large and diverse multimodal datasets. This extensive training allows VLP models to develop strong zero-shot cross-modal encoding capabilities, meaning they can accurately interpret and align images and text without requiring specific retraining for every new task. As a result, VLP models have demonstrated superior performance in CIR tasks [56, 15, 21], especially when handling complex, real-world queries that require the integration of both visual and textual information.

Given the success of native multimodal language models (Large Language Models (LLMs) trained with multimodal data) like Meta's Chameleon [2] and GPT-4o [45], which use the same VLP models to encode both text and image, there is a clear advantage to using a unified approach for the data encoding process. Therefore, this research employs CLIP as the feature extraction backbone to encode both text and image for target image retrieval.

2.3 Multimodal Fusion

Multimodal fusion in CIR involves the combination of visual and textual features to form a unified query representation for target image retrieval. Recent CIR methodologies have predominantly employed VLP models for feature extraction due to their ability to project images and text into a common embedding space. This shared embedding space naturally facilitates the fusion of multimodal queries, a critical process for accurately retrieving target images.

2.4 Evaluation of CIR Techniques

The primary metric used in this research is recall [23], which measures the proportion of relevant items retrieved out of all relevant items in the dataset. Recall is particularly important in CIR because it quantifies the system's ability to retrieve all the images that meet the user's specified criteria.

In mathematical terms, recall is defined as:

$$\text{Recall} = \frac{\text{Number of Relevant Items Retrieved}}{\text{Total Number of Relevant Items in the Dataset}} \tag{2.1}$$

This metric is crucial in the context of CIR, where the goal is often to ensure that as many relevant images as possible are retrieved, even at the expense of retrieving some non-relevant images. The recall metric is calculated at different thresholds (e.g., R@1, R@10, R@50), which represent the recall at the top 1, 10, or 50 retrieved results, respectively.

For instance, R@10 indicates the proportion of relevant images within the top 10 results returned by the CIR system. A higher recall at a given threshold suggests that the system is more effective at retrieving relevant images early in the ranking process, which is essential for user satisfaction in practical applications.

2.5 Related Work in CIR

In recent years, numerous approaches have been developed to address CIR tasks, leveraging advanced technologies such as transformer architectures [55], large VLP models, and generative models. These methods have diversified approaches to solving CIR tasks, utilizing attention mechanisms [61, 52, 34, 65, 68], conversational models [63, 6, 46], generative models [74, 18], Graph Neural Networks [71, 70], LLMs [16, 10, 38], multi-task learning [8], multi-expert models [21], prompt learning [3, 41, 19, 1, 30, 25, 28], re-ranking techniques [66, 9, 42], and similarity learning [59, 4, 27, 69, 32, 5]. Increasing computational power has led to notable improvements in CIR performance applied to the fashion domain. For instance, the authors in [18] propose a latent diffusion-based approach for composed image retrieval, focusing on generating high-quality fused multimodal features by diffusing latent representations of the source image and modification text. Similarly, [6] leverages LLMs to enhance multimodal search capabilities, improving CIR by utilizing advanced language understanding and processing capabilities. The work in [42] introduces a candidate set re-ranking approach using a dual multi-modal encoder to refine initial retrieval results, ensuring higher precision in CIR tasks. Additionally, [3] emphasizes using sentence-level prompts to enhance CIR processes, capitalizing on the ability of detailed textual descriptions to improve retrieval accuracy.

2.5.1 Multimodal Fusion Techniques in CIR

The development of effective multimodal fusion techniques has become a central focus in CIR research. The ability to effectively combine visual and textual features is crucial for achieving high retrieval accuracy in CIR systems.

2.5.1.1 DQU-CIR: Dual Query Unification

The DQU-CIR framework, introduced by Wen et al. [59], presents a linear and simple fusion model architecture that addresses the challenge of effectively combining multimodal queries in CIR. DQU-CIR [59] achieves absolute improvements of 5.51%, and 8.63% over the best baseline for the Avg. metric on FashionIQ VAL-Split [64], and Shoes [20] datasets respectively [59]. The key contribution of DQU-CIR [59] lies in its introduction of Dual Query Unification strategies—text-oriented unification and vision-oriented unification—that allow for a more effective combination of textual and visual information. Text-oriented unification involves concatenating the source image's caption with the modification text to form a unified textual query, while vision-oriented unification involves overlaying key modification words directly onto the source image to create a unified visual query. These unified queries are then encoded using the CLIP model, with the resulting features fused through a linear adaptive fusion mechanism. This approach significantly simplifies the multimodal fusion process, making it more efficient while still achieving State-of-the-art (SOTA) retrieval accuracy.

2.5.1.2 CLIP4CIR: Leveraging Non-Linear Fusion and Two-Stage Training

CLIP4CIR, presented by Baldrati et al. [4], advances CIR by introducing a non-linear fusion model architecture and a two-stage training process. The CLIP4CIR approach leverages the CLIP model's pre-trained capabilities to extract rich visual and textual features. The non-linear fusion model in CLIP4CIR [4] integrates these features through a more complex architecture involving additional layers, such as fully connected layers, which allow for more sophisticated interactions between modalities. This non-linear approach is designed to capture more intricate relationships between the visual and textual components, leading to enhanced retrieval performance. Furthermore, CLIP4CIR [4] employs a two-stage training process: the first stage fine-tunes the CLIP model to adapt its embeddings more effectively to the CIR task, while the second stage focuses on training the non-linear fusion model itself. This two-stage process allows for better alignment between the visual and textual embeddings, ultimately improving

the robustness and accuracy of the CIR system.

2.5.2 Importance of Ablation Studies in Multimodal Fusion

Given the varying approaches to multimodal fusion, an ablation study is crucial to dissect the impact of different strategies on CIR performance. DQU-CIR [59] and CLIP4CIR [4] represent two SOTA methodologies that approach fusion in fundamentally different ways—linear versus non-linear. However, these methods have not been thoroughly evaluated against one another in existing literature, leaving the question of which approach is more effective unanswered.

An ablation study focusing on these two methods will allow us to isolate the contributions of each approach and understand how they influence retrieval accuracy. This comprehensive analysis will not only clarify their roles in enhancing or hindering CIR performance but will also provide valuable insights into the effectiveness of different fusion strategies within the fashion domain. By thoroughly evaluating these strategies, this research aims to fill critical gaps in the current understanding of CIR, offering guidance for future studies and practical applications.

Chapter 3

Methodology

3.1 Problem Formulation

This research aims to address the task of Composed Image Retrieval (CIR), which involves retrieving a target image that satisfies the constraints imposed by a multimodal query consisting of a source image I_s and a modification text T_m. The CIR task is distinct from traditional image retrieval methods which typically rely on a single modality (either text or image) for querying. Instead, CIR requires the system to understand and integrate information from both the visual and textual domains to effectively retrieve a target image that aligns with the user's modification request.

Formally, the CIR problem can be defined as follows: Given a set of N training triplets $\mathcal{T} = \{(I_s, T_m, I_t)_i\}_{i=1}^{N}$, where I_s denotes the source image, T_m denotes the modification text, and I_t denotes the target image. Textual features f_{textual} and visual features f_{visual} are extracted by the CLIP model from the multimodal queries (I_s, T_m), then used as inputs to derive fused multimodal features output in a multimodal fusion process that encapsulates the user's essential search demand. The task is to align the fused multimodal features closely with the feature representation of the target image in the same embedding space.

Given the results in CLIP4CIR [4] and DQU-CIR [59] papers, ablation studies in the following Section 3.2, 3.3, 3.4 are essential to systematically evaluate the innovating components utilized in this two papers. This study for multimodal fusion will also evaluate the effectiveness of the standalone components in improving retrieval performance. The scenarios covered include real-world use cases where customers search for products with desired adjusted features based on an input image and text modifications (e.g. "more blue and higher at the neckline" illustrated in Figure 3.1).

3.2 Raw-data level Multimodal Unification

The motivation behind this ablation study is grounded in the empirical results presented by DQU-CIR [59], which demonstrated the effectiveness of raw-data level multimodal (visual and textual) query unification. This approach has shown significant potential as a more efficient alternative to computationally intensive methods like image inpainting for image manipulation [18, 74]. By unifying queries at the raw-data level, DQU-CIR [59] effectively preserves the integrity of the original data while achieving a strong alignment between the visual and textual modalities. Given these findings, it becomes crucial to investigate whether this raw-data level fusion can serve as a superior approach to reduce computational complexity without sacrificing retrieval performance. This study aims to critically assess the potential of raw-data level query unification and whether can improve the retrieval performance in fashion products with different unification combinations, making it a crucial component of our ablation study.

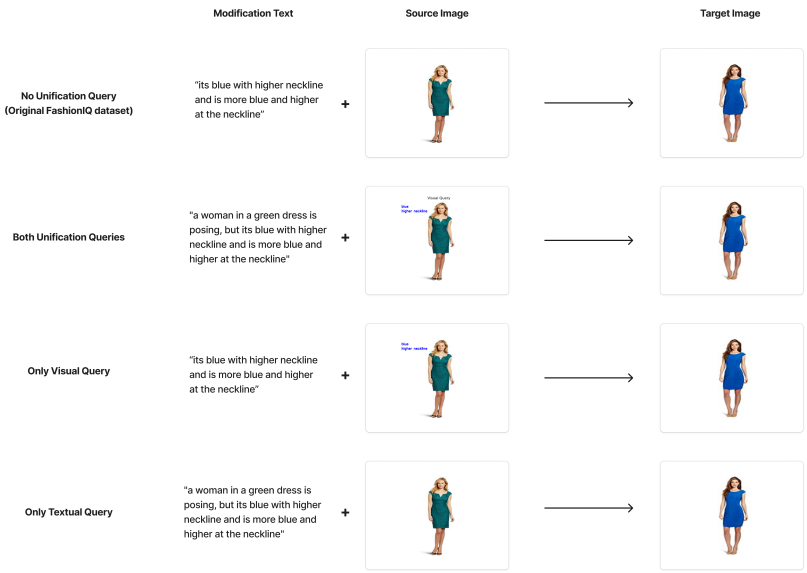

Figure 3.1: Different combinations for raw-data level multimodal (visual and textual) query unification

In this ablation study, the data preprocessing pipeline follows the methodology outlined in DQU-CIR [59], where the query unification occurs at the raw-data level by creating unified textual and visual queries as shown in Figure 3.1. A unified textual query is achieved by concatenating the modification text with a textual description of the source image, which is extracted using the BLIP-2 image captioning model [33].

This textual unification aims to retain the complete semantic content of both the source image and the modification text, making it particularly effective in CIR cases with complex modification requests, where the modification text is critical to expressing the user's search intent. A unified textual query is achieved by extracting key modification words from the modification text using a general-purposed Large Language Model [53], then directly inscribed onto the source image. This visual unification ensures that the visual integrity of the source image is preserved while embedding the crucial target attributes within the image. By utilizing CLIP's native transformer-based Optical Character Recognition (OCR) capability stated in [58, 60], the system can effectively reason about the user's search demand based on the unified visual query.

SOTA results in DQU-CIR [59] demonstrated the recall performance of text-only, visual-only, and combined query unification methods. However, the study did not compare these results with a baseline system that does not incorporate any query unification. To provide a fair and comprehensive evaluation of the effectiveness of raw-data level preprocessing, it is necessary to include a baseline comparison. Furthermore, while DQU-CIR [59] focused its query unification ablation study purely on the general image CIRR dataset [40], this study aims to explore the impact of these unification techniques in a domain-specific setting, specifically within the fashion industry. This study's findings will help clarify whether raw-data level fusion can indeed provide measurable improvements over the standard FashionIQ dataset. Additionally, the study will explore the applicability and limitations of these techniques in a fashion-specific context.

3.3 Linear and Non-Linear Multimodal Query Fusion

In composed image retrieval (CIR) tasks that utilize the CLIP model as the feature backbone, two primary variations of multimodal fusion are often considered: linear and non-linear fusion. These variations differ in how they combine the textual and visual features extracted by the CLIP model to form a unified query representation. The choice between linear and non-linear fusion approaches can significantly impact the effectiveness of the retrieval process. Linear fusion typically involves a straightforward weighted addition of the features from both modalities as implemented in DQU-CIR [59], while non-linear fusion incorporates additional transformations that may introduce complexity but also the potential for better alignment between the modalities as implemented in [12, 72, 61, 4].

3.3.1 Linear Fusion

DQU-CIR [59] presents a simple yet SOTA linear multimodal fusion approach where the fused multimodal features are computed as a weighted sum of the textual and visual query features. Formally, the combined feature q_f is calculated as:

$$q_f = \lambda \cdot f_{\text{textual}} + (1-\lambda) \cdot f_{\text{visual}} \tag{3.1}$$

where f_{textual} and f_{visual} represent the extracted textual and visual features, respectively, and λ is a learnable parameter with a range of 0-1 that adjusts the contribution of each modality. This linear combination ensures that the resulting feature resides within the same embedding space as the original features, thereby minimizing the risk of deviating from the common embedding space established by the VLP model [59]. The linear fusion approach is expected to work well in scenarios where the textual and visual information is complementary and straightforward to combine, which proves that latent text representations from CLIP exhibit geometric regularities and are highly structured [13].

3.3.2 Non-Linear Fusion

In contrast, CLIP4CIR [4] introduces a non-linear fusion approach that builds on the linear weighted addition by adding a second branch that outputs a non-linear mixture of the image and text features. Specifically, the combined features are obtained through the Combiner network $\mathcal{C}_\theta$ from CLIP4CIR [4], which learns to fuse the multimodal features via a more complex transformation:

$$\hat{q_f} = (1-\lambda) \cdot f_{\text{visual}} + \lambda \cdot f_{\text{textual}} + \nu \tag{3.2}$$

where ν represents the output of the non-linear branch with the textual and visual features as the inputs, the process involves linear transformations followed by ReLU activation and dropout layers, which introduce non-linearity into the fusion process. The purpose of this non-linear branch is to capture more complex relationships between the visual and textual features, potentially leading to better alignment in the embedding space. However, there is also a risk that this additional complexity could cause the fused multimodal features to deviate from the original embedding space, thus hindering retrieval performance.

3.3.3 Comparison and Evaluation

The effectiveness of linear versus non-linear fusion strategies remains an open question. This ablation study seeks to clarify their respective roles in enhancing CIR performance through a comprehensive ablation study and compares the retrieval performance of two SOTA approaches — linear fusion using the DQU-CIR model [59] and non-linear fusion using the CLIP4CIR model under identical preprocessing conditions and datasets. The model architecture in DQU-CIR [59] is used as the linear multimodal fusion model, and the model architecture in CLIP4CIR [4] is used as the non-linear multimodal fusion model for comparison. The reason for choosing these two specific models is that the CLIP4CIR model essentially includes a linear weighted addition branch, similar to the DQU-CIR model, but with an additional branch that outputs the non-linear mixture contribution of the image and text features. This setup allows for a fair contrast of the retrieval performance focusing on linear and non-linear fusion processes, instead of having performance comparison with different model setups as previous studies have done [72, 62, 58, 12, 59, 4].

Both models employ the same contrastive loss learning metric, ensuring that the comparison focuses solely on the differences between the fusion strategies. The contrastive loss function is formulated as:

$$\mathcal{L} = -\log \frac{\exp\left(\cos(f_q, f_{t_i})/\tau\right)}{\sum_{j=1}^{B} \exp\left(\cos(f_q, f_{t_j})/\tau\right)} \tag{3.3}$$

where $\cos(\cdot)$ represents the cosine similarity, f_q is the fused multimodal features query, f_{t_i} is the target image feature, and τ is the temperature parameter. This loss function encourages the fused multimodal features to be as close as possible to the corresponding target image feature while maximizing the distance from other target images within the batch.

The motivation for this investigation lies in the potential of non-linear fusion methods to capture complex relationships between visual and textual data that linear methods might miss. However, there is also a concern from DQU-CIR [59] that introducing non-linearity could disrupt the coherence of the fused multimodal features within the embedding space, leading to suboptimal retrieval performance. This ablation study aims to identify whether the additional complexity introduced by non-linear fusion offers a significant advantage over the simpler linear fusion approach.

3.4 Two-Stage Training

The two-stage training scheme originally proposed in CLIP4CIR [4] is designed to enhance the multimodal fusion process for CIR. This study applies this two-stage approach not just to the non-linear model as originally proposed but also to a linear multimodal fusion model. The goal is to evaluate whether this training strategy can similarly benefit the overall multimodal fusion task in a simpler, linear context, thus offering a broader understanding of its effectiveness.

3.4.1 CLIP Additivity Fine-Tuning (First Stage)

The first stage involves task-oriented fine-tuning of the CLIP model, where both the image and text encoders are adapted to better suit the CIR task. This process, as detailed in CLIP4CIR [4], aims to reduce the mismatch between the large-scale pre-training of CLIP and the downstream task by fine-tuning the encoders using an element-wise sum of image and textual features. The key objective here is to enhance the additive properties of the output features, allowing for better alignment between the source image and the modification text in the embedding space. For example, given an image of a black dress I_s and a corresponding text query T_m ("is blue"), the aim is for the fine-tuned model to generate embeddings where $\psi_I(I_s) + \psi_T(T_m) \approx \psi_I(I_t)$, with I_t representing the image of a blue dress. This fine-tuning process is expected to produce embeddings that exhibit strong "additivity properties," meaning the relative caption's embedding should correspond closely to the displacement vector between the source image and the target image features, i.e., $\psi_T(T_m) \approx \psi_I(I_t) - \psi_I(I_s)$.

3.4.2 Linear Multimodal Fusion with MLP (Second Stage)

The second stage of the fusion model training process directly addresses a known limitation of CLIP embeddings — while they support simple arithmetic operations, their effectiveness in fine-grained tasks or complex transformations can be limited, as highlighted in [14]. The need for more adaptive fusion mechanisms arises from the fact that simple arithmetic, such as the direct addition of embeddings, may not fully capture the nuanced relationships between visual and textual queries, especially in cases involving complex user modifications or specific retrieval tasks.

In this stage, rather than relying solely on basic arithmetic operations, the model learns to fuse the fine-tuned image and text features using MLP. This approach aims to

enhance retrieval performance by learning an optimal weighted addition of the unified textual and visual query features in a linear context as stated in Equation 3.1. The MLP is tasked with refining the fused multimodal features to better align with the target image representation, thereby overcoming the limitations of straightforward embedding arithmetic. By employing triplets (I_s, T_m, I_t) during training, where (I_s, T_m) is the query and I_t is the target image, the model learns to minimize the distance between the fused multimodal features and the target image features. This learning process on top of the first stage of CLIP fine-tuning ensures that the model can handle more complex or fine-grained retrieval tasks effectively, beyond what simple embedding arithmetic can achieve.

3.4.3 Significance

The application of this two-stage training approach to both linear and non-linear models allows for a comprehensive evaluation of its potential to improve CIR tasks. By assessing whether this method enhances the effectiveness of linear multimodal fusion, this study contributes to the broader understanding of how best to align textual and visual features within the VLP model's embedding space. The results will inform future approaches to CIR, potentially offering a simpler yet effective alternative to more complex non-linear methods while addressing the inherent limitations of embedding arithmetic in handling intricate retrieval tasks.

Chapter 4

Experimental Results

To provide an intuitive demonstration of the effective components in CIR, this research illustrates the performance of different components from DQU-CIR [59] and CLIP4CIR [4] papers on two public datasets (i.e. FashionIQ [64] and Shoes [20]). The experiments in this section for each study mentioned in Section 3 are designed to facilitate a better understanding of the contribution of each key component in terms of quantitative and qualitative evaluation.

4.1 Datasets and Metrics

All the triplets in the training and test sets are structured as $< I_s, T_m, I_t >$, consisting of a source image (I_s), text modifications (T_m), and the corresponding target image (I_t). This research adhered to standard evaluation protocols for each dataset and utilized the Recall@k (R@k) metric for a fair comparison. Recall@k (R@k) [23] is particularly suitable for evaluating retrieval performance in this ablation study because it directly measures the model's ability to retrieve the correct target image within the top k results, which is critical in CIR tasks where the goal is to match a source image and modification text to a specific target image. This metric is highly relevant as it reflects how well the model handles large candidate sets by focusing on the top results, which is crucial for practical user scenarios where only the first few results are typically considered. Additionally, to provide deeper insights into how the CLIP models process visual information, this research proposed a novel evaluation of qualitative metrics towards CIR task - GradCAM heatmaps [51, 36], to visualize the regions of the image that receive the most attention during the retrieval process. The heatmaps generated by GradCAM highlight the attention distribution across the image corresponding to

the label without any additional training, where red heatmap regions indicate higher attention the model has given, and thus, greater relevance as perceived by the model. This visualization allows us to assess whether the model is focusing on the most relevant regions of the image in relation to the query to enhance the explainability of CLIP results, providing a qualitative evaluation of the model's interpretability and focus.

The FashionIQ dataset [64] comprises 46,609 training images, 15,536 validation images, and 15,538 test images, categorized into three fashion classes: Shirts, Dresses, and Toptees. It includes a total of 18,000 training triplets, 6,016 test queries, and 8,619 test targets [64]. Examples of modification text (T_m) in this dataset include descriptions like "has less sleeves," "has stripes," or "is more solid colored." The challenge's evaluation criteria defined in this dataset are followed, focusing on R@10 and R@50 across these categories. Notably, FashionIQ provides two evaluation protocols: the VAL-Split [11] and the Original-Split [64]. For consistency with existing literature, this study uses the VAL-Split which validates the recall metrics on the union of reference and target images in all validation set triplets.

The Shoes dataset [20] contains 8,990 training triplets, 1,761 test queries, and 4,658 test targets. It includes images of various women's footwear types, such as boots, sneakers, heels, and clogs. Examples of modification text include phrases like "has laces," "has a strap and buckle," or "is a shoe and contains no heel." In line with prior studies [11, 15, 62, 60], R@1, R@10, and R@50 are reported along with their calculated averages.

4.2 Experimental Setting

All experiments were conducted using the ViT-B/32 CLIP model [31] as the feature extraction backbone. This choice was made to ensure a fair and consistent experimental setup across all component variations, particularly because the DQU-CIR [59] used a considerably larger CLIP model (ViT-H/14). Although the ViT-H/14 CLIP model is doomed to offer superior performance due to its larger size, with approximately 5.6 times more text parameters and 7.2 times more image parameters than the ViT-B/32 model, its computational requirements make it less practical for extensive ablation studies. The ViT-B/32 CLIP model strikes a balance between performance and efficiency, making it ideal for in-depth experimental analysis in this study. Given the similarity between fashion product items, the recall empirical results vary to a small extent between different combinations of the ablation studies. Each experiment was run twice, and the

average results were reported to ensure reliability.

This study identified the effective components that improve CIR tasks, based on the methodologies from DQU-CIR [59] and CLIP4CIR [4]. The four distinctive components identified for preliminary experiments include linear and non-linear multimodal fusion models, image preprocessing for the CLIP model, model training scheme, and raw-data level multimodal query unification. For the image preprocessing pipeline, three methods were considered: CLIP standard, square pad, and target pad [4]. The standard CLIP preprocessing pipeline involves resizing the image so that the smaller side matches the input dimension, followed by a center crop to produce a square patch [43]. To address the loss of information caused by this operation, CLIP4CIR [4] proposed applying padding only when the aspect ratio exceeds a predefined target ratio. Preliminary results (see Table 4.1) show that while the target pad method performed slightly better, the differences were minimal, particularly given that most existing studies [12, 72, 61, 4] use the standard CLIP preprocessing method.

Preprocess	**Dresses**		**Shirts**		**Tops&Tees**		**Avg.**
	R@10	**R@50**	**R@10**	**R@50**	**R@10**	**R@50**	
Standard	52.11	75.71	56.28	76.59	60.79	82	67.25
Square pad	51.9	75.5	56.2	76.52	60.4	81.95	67.08
Target pad	51.95	75.9	56.27	76.7	60.88	82.05	67.29

Table 4.1: Retrieval performance of image preprocessing methods ablation study based on linear multimodal fusion with ViT-B-32 using FashionIQ VAL-split

All experiments were trained using the AdamW optimizer [17]. For both the FashionIQ and Shoes datasets, CLIP and fusion model parameters were optimized at different learning rates to ensure effective and gradual convergence, aligned with the three proposed ablation studies. Specific hyperparameters are detailed in their corresponding sections. 100 epochs are set for all experiment runs given the empirical observation that multiple test runs of each experiment showed an increase in validation loss after 100 epochs. For all experiments, the feature dimension of ViT-B/32 is 512, the batch size is fixed at 128, and the temperature factor in Equation 3.3 is set to 0.1 as DQU-CIR [59] proposed.

In conclusion, this section will conduct the three ablation studies proposed in Section 3, maintaining consistency with the CLIP standard image preprocessing method

and using the ViT-B/32 CLIP model as the feature extraction backbone to ensure computational efficiency and reliable comparisons.

4.3 Raw-data level Multimodal Unification Study

This ablation study investigate the effectiveness of raw-data level multimodal (visual and textual) query unification as proposed in DQU-CIR [59]. Each model was trained for 100 epochs, and experiments were conducted using a linear weighted addition model with end-to-end training by inputting different unification query combinations as training data, aiming to maintain a straightforward linear multimodal fusion process. The experiments utilized the FashionIQ and Shoes datasets, adhering to standard protocols and evaluation metrics, specifically focusing on Recall@k (R@k) to assess retrieval performance. For the FashionIQ dataset, the fusion model and CLIP model learning rates were 1e-5 and 1e-7 respectively. For the Shoes dataset, the fusion model and CLIP model learning rates were 5e-6 and 5e-7 respectively.

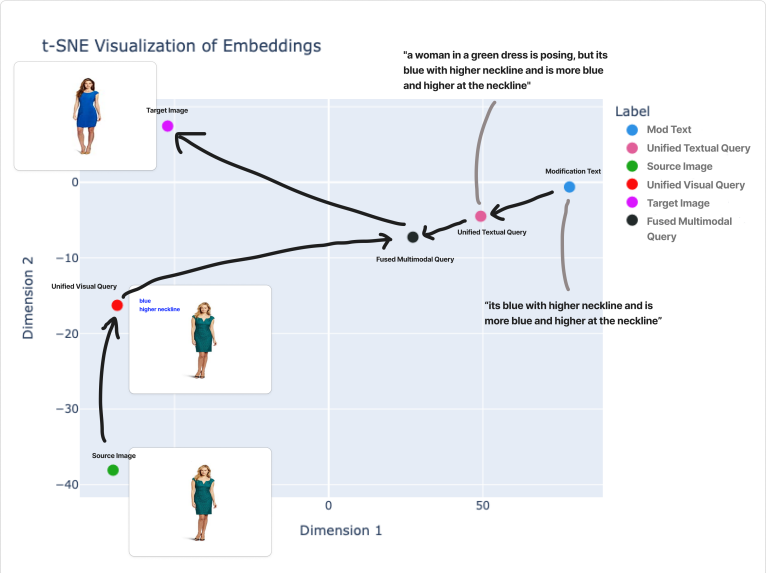

Figure 4.1: t-SNE Visualization of Embeddings: This figure illustrates the embedding space where different elements of the CIR process are projected onto a 2D space, showing how the system navigates from the source image and modification text to the target image.

To better understand the impact of query unification, t-SNE visualization is proposed

in this paper and used to project a triplet sample into a 2D embedding space. T-SNE algorithm aims to preserve the points that are close in the high-dimensional space to be close in the low-dimensional embedding as well [54]. Figure 4.1 illustrates that the embeddings of both the modification text (to unified textual query) and the source image (to unified visual query) became closer to the target image embedding after unification. This is significant because it demonstrates that raw-data level query unification effectively navigates the embedding space toward the desired target image. After the textual and visual query unification, the Euclidean distance between the text input and the target image decreased by 21.7% (130.61 to 102.27), and the Euclidean distance between the image input and the target image decreased by 43.7% (48.85 to 28.82), illustrated the method's effectiveness in improving retrieval accuracy, in which text on the image can harness the CLIP OCR property for the unified visual query, while the more detailed modification text including source image caption of the original object in the source image for the unified textual query. After conducting the t-SNE distance experiments on 6,016 test queries in the FashionIQ VAL-split dataset, the average difference in Euclidean distance from the modification text embedding to the unified textual query embedding is decreased by 17.67%, and the average difference in Euclidean distance from the source image embedding to the unified visual query embedding is decreased by 41.69%.

The performance disparities across different product categories and query unification methods, as shown in the recall tables 4.2 and 4.3, further motivated a qualitative evaluation using GradCAM visualization. By computing gradients with respect to its final retrieval embedding output, it is observed how different fine-tuned CLIP models extracted features from the source and target images in a triplet example and then fused these features to retrieve the target image.

Query Unification	**Dresses**		**Shirts**		**Toptees**		**Avg.**
	R@10	**R@50**	**R@10**	**R@50**	**R@10**	**R@50**	**(R@10 + R@50)/2 (Rank)**
None	47.53%	72.56%	52.23%	71.2%	58.67%	80.89%	63.85% (4)
Both	52.11%	76.1%	56.13%	76.79%	61.04%	81.85%	67.34% (1)
Only Visual	48.56%	73.87%	53.11%	73.27%	61.24%	83.34%	65.57% (3)
Only Text	49.35%	74.54%	55.42%	75.83%	61.17%	82.14%	66.41% (2)

Table 4.2: Recall averages for Dresses, Shirts, Toptees of FashionIQ VAL-split dataset (6,016 test queries) at 10 and 50 thresholds with different Query Unifications

From both of the recall tables 4.2 and 4.3, it is apparent that the no unification query

Query Unification	**R@1**	**R@10**	**R@50**	**Avg. (R@1 + R@10 + R@50)/3 (Rank)**
None	25.34%	61.78%	82.87%	56.66% (4)
Both	27.81%	64.75%	84.97%	59.18% (2)
Only Visual	27.87%	65.1%	86.8%	59.92% (1)
Only Text	25.93%	64.14%	83.11%	57.72% (3)

Table 4.3: Recall averages for Shoes test dataset (1,761 test queries) at 1, 10, and 50 thresholds with different Query Unifications

consistently resulted in the worst retrieval performance across both datasets. This is reflected in the GradCAM heatmaps, where the CLIP model showed scattered focus for no unification query method, such as predominantly focusing on irrelevant areas like a woman's head in both source and target images in Figure 4.2, the same issue happened in the Shoes source and target images in Figure 4.3. In contrast, by writing the desired feature keywords in the source image, the unified visual query allowed the CLIP model to focus more effectively on the overall structure of the target object shown in both FashionIQ and Shoes images, leading to improved retrieval outcomes.

Interestingly, in the FashionIQ dataset, using only the unified textual query yielded better recall than using only the unified visual query in Table 4.2. This is likely because, while the unified visual query helps the CLIP model identify the correct object structure, fewer red heatmap areas are spread across the dress on the target image compared to the unified textual query in Figure 4.2. The unified textual query allows it to fully capture specific features like texture, shape, and length, which are crucial in fashion objects. This is because most dresses, shirts, and top tees have a similar shape in their own category, the unified textual query becomes important to add more fashion features on top of the structure and shape discovered by the unified visual query. Thus, in the FashionIQ dataset, the highest recall was achieved with both unification queries, combining the advantages of shape identification from the unified visual query and detailed feature extraction from the unified textual query. In the Shoes dataset, the unified textual query in the target image in Figure 4.3 guides the model to focus more on the fashion details on the shoes, rather than the background corner in the target image of no unification query method. However, the distinct differences between shoe types (e.g. sneakers, high heels, clogs) made the unified visual query more effective than using both unification queries or only the unified textual query method in terms of

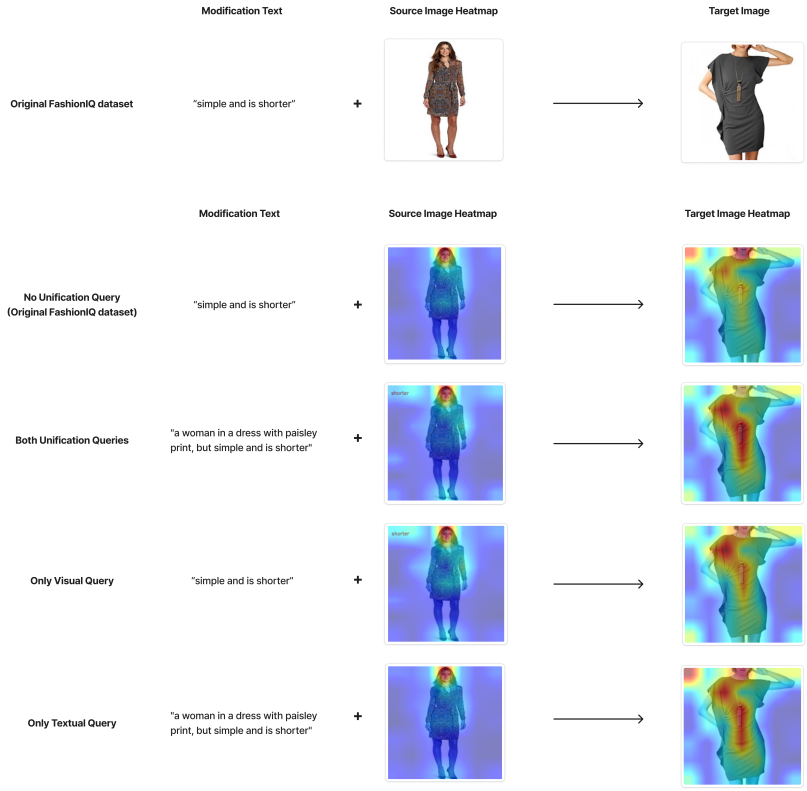

Figure 4.2: Triplet Example of GradCAM heatmap visualisation for different unification combinations on Dress source and target images of FashionIQ computed by the gradients of the CLIP model with respect to its final retrieval embedding output

recall metrics in Table 4.3.

The findings from this study provide compelling evidence that multimodal query unification plays a critical role in enhancing composed image retrieval (CIR) performance, particularly in the fashion domain. The empirical results demonstrate that the absence of query unification consistently leads to the poorest retrieval outcomes, as seen in both the recall metrics and GradCAM heatmaps, where the model's focus was often scattered and misaligned with relevant features. In contrast, integrating both unified visual and textual queries significantly improved the model's ability to capture and focus on the necessary features of the target image, as evidenced by the more targeted attention in the GradCAM visualizations and higher recall scores. Notably, the unified textual query alone was more effective than the unified visual query in the FashionIQ dataset, likely due to its ability to capture detailed fashion features that are crucial within similar-shaped categories. However, in the Shoes dataset, the unified

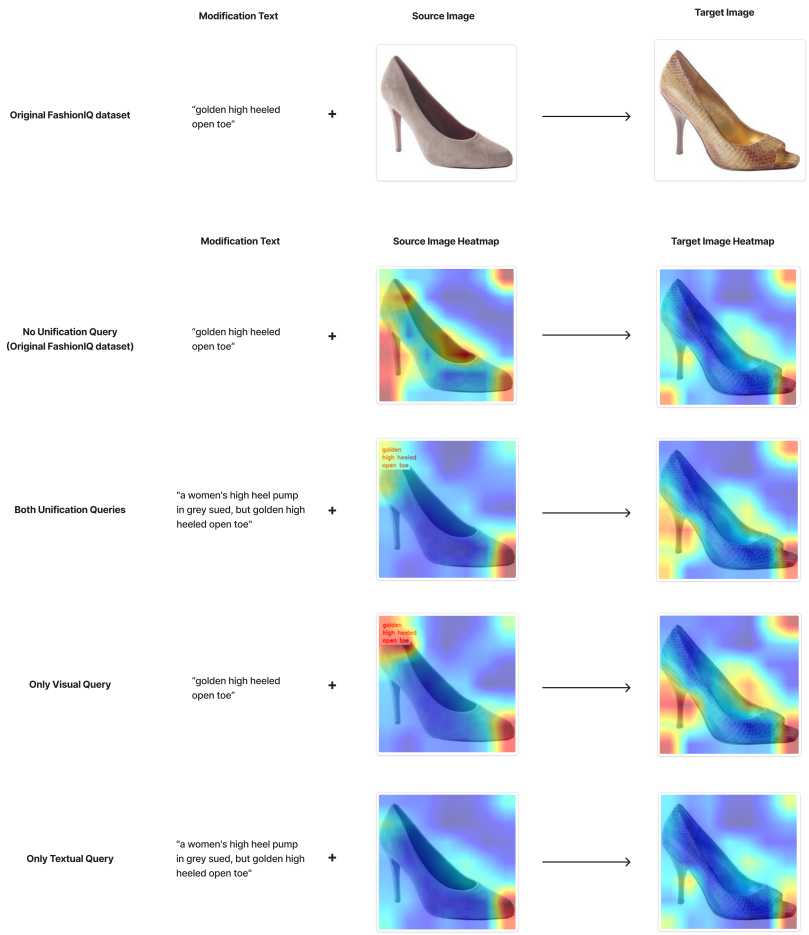

Figure 4.3: Triplet Example of GradCAM heatmap visualisation for different unification combinations on Shoes source and target images computed by the gradients of the CLIP model with respect to its final retrieval embedding output

visual query proved more effective, reflecting the importance of shape and structure in distinguishing between diverse footwear types. These results underscore the importance of raw-data level multimodal query unification, particularly when leveraging CLIP models, to optimize retrieval performance across varying product categories in the fashion industry. This study thus validates the effectiveness of this approach and highlights its potential to surpass traditional methods like image inpainting, offering a more efficient and robust solution for CIR tasks.

4.4 Linear and Non-Linear Multimodal Query Fusion Study

In this ablation study, the focus is on exploring the effectiveness of linear versus non-linear multimodal query fusion techniques in CIR tasks. The study aims to determine whether linear or non-linear fusion of features extracted by CLIP models enhances retrieval performance under the same experimental setup. The experiments were conducted using linear and non-linear multimodal fusion model architectures, as stated in Section 3.3, with end-to-end training over 100 epochs. For the FashionIQ dataset, the fusion model and CLIP model learning rates were set at 1e-5 and 1e-7 respectively. For the Shoes dataset, the fusion model and CLIP model learning rates were set at 5e-6 and 5e-7 respectively.

Fusion Model	**Dress**		**Shirt**		**Toptee**		**Recall (Avg)**
	R@10	**R@50**	**R@10**	**R@50**	**R@10**	**R@50**	
Linear	47.53%	72.56%	52.23%	71.2%	58.67%	80.89%	63.85%
Non-linear	53.94%	76.9%	56.08%	75.96%	62.88%	83.69%	68.24%

Table 4.4: Recall averages for Dresses, Shirts, and Toptees of FashionIQ at 10 and 50 thresholds with different Fusion Models.

Fusion Model	**R@1**	**R@10**	**R@50**	**Recall (Avg)**
Linear	25.34%	61.78%	82.87%	56.66%
Non-linear	28.39%	66.21%	86.77%	60.46%

Table 4.5: Recall averages for Shoes at 1, 10, and 50 thresholds with different Fusion Models.

From both the FashionIQ and Shoes recall tables 4.4 and 4.5, it is evident that the non-linear fusion models surpassed the linear fusion models in terms of retrieval performance, as indicated by higher recall metrics. By visualizing how the final retrieval embedding output relates to the source and target images in Figure 4.4, it is observed that the additional non-linear layer allows the model to intensify the attention on existing useful features mentioned in the modification text, such as the text ”short sleeve” made the corresponding red heatmap region in the source image more intense, and text ”darker

colored tee shirt" had the same effect in the target image, which indicates the non-linear layer has intensified the existing features in the final retrieval embedding. However, it should be noted that the non-linear layer simultaneously intensified the features learned from unrelated and redundant areas, such as the corner of the background in Figure 4.4. For the Shoes dataset, Figure 4.5 illustrates that more red heatmap areas are spread across the stiletto heel shoes in both the source and target image, while still providing attention to the existing textual features "ankle straps" in the modification text. At the same time, compared to the source and target images in the linear fusion model, the features learned by the non-linear fusion have been more spread out to the target object itself by sacrificing some intensities of the existing textual features, which potentially leads to better recall metrics.

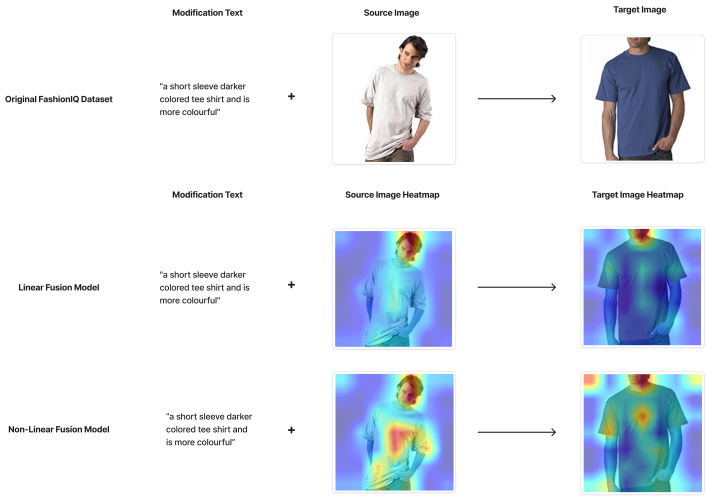

Figure 4.4: Examples of GradCAM heatmap visualisation for linear and non-linear fusion models on Shirt source and target images of FashionIQ dataset computed by the gradients of the CLIP model with respect to its final retrieval embedding output

This study has critically examined the effectiveness of linear and non-linear multi-modal fusion techniques in composed image retrieval (CIR) tasks within the fashion domain. The results, as evidenced by higher recall metrics in both the FashionIQ and Shoes datasets, suggest that non-linear fusion models can indeed capture more complex relationships between visual and textual data compared to their linear counterparts. The GradCAM visualizations further revealed that the non-linear layer intensifies attention on relevant features mentioned in the modification text, thereby enhancing the retrieval embedding's alignment with the target image. However, this intensification also occa-

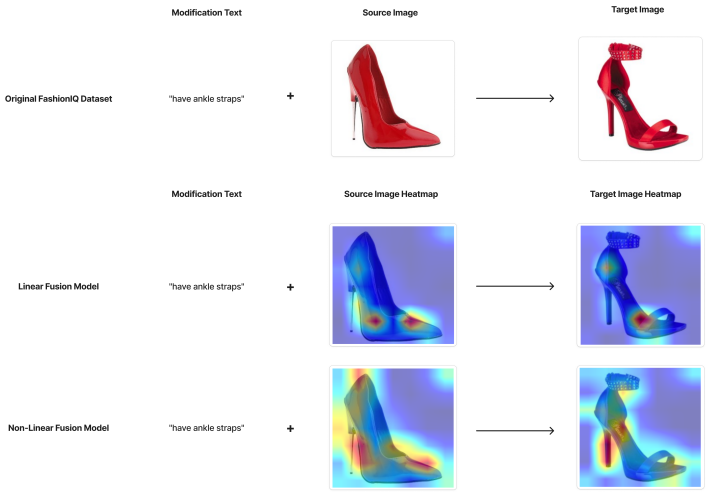

Figure 4.5: Examples of GradCAM heatmap visualisation for linear and non-linear fusion models on Shoes dataset source and target images computed by the gradients of the CLIP model with respect to its final retrieval embedding output

sionally extends to unrelated areas, indicating a trade-off between capturing essential features and introducing noise. Despite this, the overall retrieval performance suggests that the additional complexity introduced by non-linear fusion does provide a significant advantage, particularly in fashion CIR tasks where nuanced feature integration is crucial. Thus, contrary to concerns raised by DQU-CIR [59] about potential disruptions in the embedding space, this study demonstrates that non-linear fusion not only preserves but enhances the coherence of the fused multimodal features, thereby supporting its application for more effective image retrieval in the fashion domain.

4.5 End-to-End vs 2-Stage Training Study

The experiments in this study utilized a straightforward linear multimodal fusion process, employing a linear weighted addition model trained with two different schemes — end-to-end training and two-stage training. Both schemes were applied to the FashionIQ and Shoes datasets, incorporating textual and visual query unification. In the two-stage training method, the CLIP encoder and fusion model are trained separately, unlike the end-to-end approach where CLIP textual and image encoders are fine-tuned while simultaneously training the fusion model. The first stage of CLIP fine-tuning was conducted for 50 epochs, followed by a 100-epoch fusion model training stage. For the

FashionIQ dataset, the learning rates for the end-to-end training of the fusion model and CLIP model were set to 1e-5 and 1e-7, respectively, while for the two-stage training, these were adjusted to 1e-4 and 1e-6. In the Shoes dataset, the learning rates for the end-to-end training were 5e-6 and 5e-7, respectively, and for the two-stage training, they were 6e-5 and 1e-6. Preliminary experiments indicated that the two-stage training method could tolerate higher learning rates while maintaining effective and stable convergence, justifying the selection of higher learning rates for both datasets.

Training Scheme	**Dress**		**Shirt**		**Toptee**		**Recall (Avg)**
	R@10	**R@50**	**R@10**	**R@50**	**R@10**	**R@50**	
End-to-End	52.11%	76.1%	56.13%	76.79%	61.04%	81.85%	67.34%
2-Stage Training	53.21%	77.4%	56.18%	77.21%	63.22%	84.38%	68.6%

Table 4.6: Recall averages for Dresses, Shirts, and Toptees of FashionIQ VAL-split dataset (6,016 test queries) at 10 and 50 thresholds with different training schemes.

Training Scheme	**R@1**	**R@10**	**R@50**	**Recall (Avg)**
End-to-End	27.81%	64.75%	84.97%	59.18%
2-Stage Training	30.83%	68.3%	87.44%	62.19%

Table 4.7: Recall averages for Shoes test dataset (1,761 test queries) at 1, 10, and 50 thresholds with different training schemes.

Embedding Type	**Average Percentage Increase in Euclidean Distance**
Modification Text Embedding	27.16%
Textual Query Embedding	31.24%
Source Image Embedding	15.35%
Visual Query Embedding	13.66%
Retrieval Embedding	32.75%

Table 4.8: Average Percentage Increase in Euclidean Distance to Target Image Embedding for Shoes test dataset (1,761 test queries) by changing the training scheme from End-to-End to 2-Stage

The recall metrics in Tables 4.6 and 4.7 demonstrate that the 2-stage training scheme consistently outperforms the end-to-end approach for both the FashionIQ and

Embedding Type	**End-to-End Training (Rank)**	**2-Stage Training (Rank)**
Modification Text Embedding	171.92 (5)	88.27 (5)
Textual Query Embedding	142.93 (4)	66.41 (4)
Source Image Embedding	65.42 (2)	41.02 (2)
Visual Query Embedding	37.11 (1)	23.90 (1)
Retrieval Embedding	110.13 (3)	50.06 (3)

Table 4.9: Shoes triplet example on Euclidean Distance to Target Image Embedding: Comparison between End-to-End (Figure 4.6) and 2-Stage (Figure 4.7) Training Schemes (Ranked by the shortest distance to target image embedding)

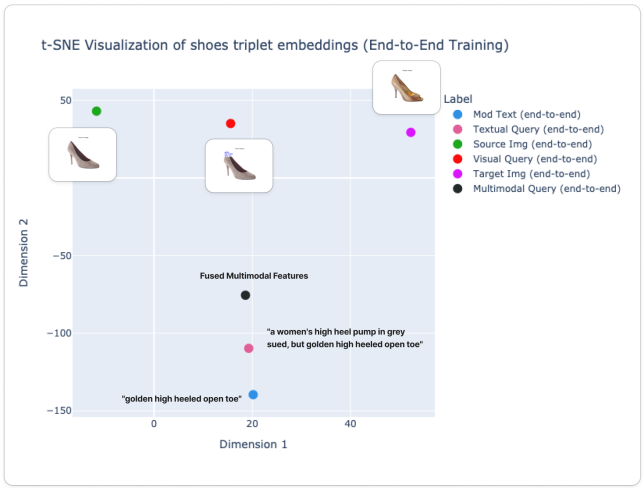

Figure 4.6: t-SNE Visualization of an end-to-end-trained Shoes embedding space

Shoes datasets, highlighting the effectiveness of the additional fine-tuning stage for increasing the additivity of CLIP embeddings. This improvement is further supported by projecting the query embeddings to a 2D space using t-SNE algorithm. In Table 4.8, after conducting the t-SNE distance experiments on 1,761 test queries in the Shoes dataset, the average Euclidean distance between each of the embedding types and the target image has a considerable decrease. Given the same Shoes triplet example, the ranking of shortest Euclidean distances to the target image embedding presented in Table 4.9 aligns with the ranking of retrieval performance in Table 4.3, showing that both embeddings of unified visual query and source image are closer to the target image

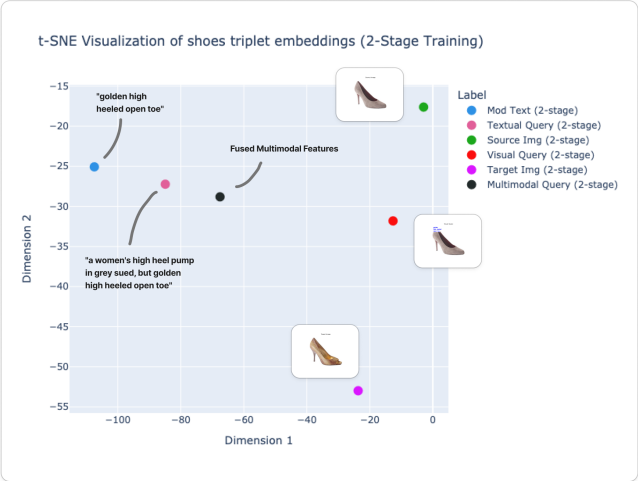

Figure 4.7: t-SNE Visualization of a 2-stage-trained Shoes embedding space

embedding in Figure 4.6 and 4.7 due to the fact of the importance of shape and structure in distinguishing between diverse footwear types. In Table 4.9, there is a significant and consistent reduction in the distance between the target image embedding and all other embeddings when using the 2-stage training approach. This indicates that the 2-stage training not only enhances recall metrics but also facilitates more effective navigation within the embedding space, making all embeddings closer to the target image and thereby improving retrieval performance.

Therefore, the quantitative results from recall and Euclidean distance, along with the qualitative results from the t-SNE visualizations, validate the effectiveness of an embedding space with robust additivity properties. These properties facilitate the linear weighted addition of combining features, where the first-stage fine-tuning makes the model more task-oriented to the fashion domain, allowing for better learning of the final retrieval embedding in the subsequent multimodal fusion stage.

In conclusion, the 2-stage training scheme, initially proposed for non-linear fusion in CLIP4CIR [4], is also effective in a simpler linear fusion context. This study confirms that the additional complexity introduced by this training strategy benefits the overall multimodal fusion task, providing a broader understanding of its effectiveness in the context of composed image retrieval within the fashion domain.

Chapter 5

Conclusion

This research has systematically explored the key components that contribute to the effectiveness of composed image retrieval (CIR) systems within the fashion domain, focusing on the evaluation of raw-data-level multimodal fusion techniques and the implementation of a two-stage training approach. Through extensive experimentation on benchmark datasets such as FashionIQ and Shoes, the study has demonstrated that shifting multimodal fusion from the traditional feature level to the raw data level enhances retrieval accuracy. The integration of visual and textual query unification has been particularly effective, improving the alignment of multimodal features within the embedding space and leading to more accurate and robust retrieval results.

The two-stage training approach, originally designed for non-linear fusion, has proven to be effective in a linear multimodal fusion context. By fine-tuning the CLIP model to align more closely with CIR tasks and make the CLIP embeddings more additive, and subsequently applying a linear multimodal query fusion, the research has shown that this approach not only optimizes retrieval performance but also provides a broader understanding of its applicability and effectiveness.

Additionally, the comparative analysis between linear and non-linear fusion models revealed that the non-linear layer could capture more intense features at the target modification area specified by the user in the modification text and additional features across the whole object. However, this advantage comes with the potential introduction of noise, which could detract from retrieval accuracy. The proposed GradCAM qualitative metrics in this study provide a valuable tool for future research in visualizing whether the CIR model captures effective embeddings for retrieval, thereby enhancing AI explainability.

In conclusion, the findings from this research suggest that the most effective ap-

proach to achieve CIR in fashion involves a combination of raw-data level multimodal fusion with visual and textual query unification and a two-stage training process that fine-tunes both the image and text encoders.

Chapter 6

Impacts

Based on the insights gained from the ablation studies conducted in this research, the proposed approach in Section 5 demonstrates significant potential for application in both future research studies and industrial settings within the fashion domain.

In future research, this approach can be particularly valuable for investigating more complex image retrieval tasks that require a detailed understanding of subtle modifications to fashion items. The two-stage training process that fine-tunes both image and text encoders could serve as a foundation for further experimental work aimed at optimizing multimodal alignment and feature extraction in various CIR contexts. Future research could explore the scalability of this approach across larger and more diverse datasets, as well as its applicability to different machine learning paradigms, such as unsupervised or self-supervised learning. Moreover, the inclusion of advanced visualization techniques like GradCAM in future studies could provide deeper insights into how the models interpret and prioritize different features during the retrieval process. This would be crucial for refining model architectures and improving the interpretability and transparency of CIR systems. This comprehensive research strategy offers a robust framework for advancing CIR methodologies and expanding the boundaries of what is achievable in the field of CIR.

In industrial applications, particularly in online fashion retail, this methodology could be implemented to enhance search engines that cater to customers looking for specific modifications in products (e.g., adjusting the color or style of a dress). By adopting this approach, companies could significantly improve the customer shopping experience, making it more intuitive and tailored to individual preferences. Moreover, the two-stage training process that fine-tunes both image and text encoders ensures that the system remains adaptable and precise in rapidly changing fashion inventories,

making it a valuable tool for addressing the challenges posed by fast fashion dynamics.

In conclusion, the proposed approach not only advances the state-of-the-art in CIR research but also holds considerable promise for practical deployment in the fashion industry, offering a comprehensive strategy for developing sophisticated and user-centric image retrieval systems.

Chapter 7

Limitations and Future Work

7.1 Enhancing Visual Focus in CIR Systems

Limitation: One significant issue identified in Section 4.3 is the tendency of the fusion model to incorrectly focus on irrelevant areas, such as background corners or a person's head, rather than on the intended fashion object. This suggests that current methods may not fully capture the nuances of visual information that are critical for accurate retrieval. This limitation implies that retrieval accuracy may be compromised in scenarios where precise focus on the fashion item is crucial.

Future Work: To address this, future research could explore the integration of advanced automatic segmentation techniques, such as GroundedSAM [29, 50, 37], which could provide a more refined, feature-oriented segmentation. By accurately identifying and isolating the relevant fashion object within an image, such approaches could significantly enhance the quality of the visual query for feature extraction of CLIP models, thereby improving retrieval performance in extracting delicate fashion features.

7.2 Evaluating the Breadth of Non-Linear Models

Limitation: This study primarily focused on testing the essential MLP non-linear layer on top of the linear fusion models as a fundamental approach to non-linear model architectures in multimodal fusion. However, it is important to acknowledge that non-linear models encompass a wide range of architectures, each with the potential for significant variation in performance. This implies that by not exploring a broader range of non-linear models, the study may miss out on potentially more effective methods for enhancing CIR performance.

Future Work: Future work could include a more comprehensive evaluation of these non-linear models, particularly models such as transformers, which have demonstrated superior capabilities in capturing complex interactions between modalities. Both qualitative and quantitative assessments could be conducted to fully understand their potential to enhance multimodal fusion for composed image retrieval. Such experiments would provide a broader perspective on how different non-linear architectures can contribute to more accurate and efficient retrieval systems, and coupled with a 2-stage training scheme to evaluate how the retrieval performance is enhanced.

7.3 Addressing the Significance of Numerical Differences

Limitation: Another limitation of this study is the lack of a critical evaluation of the significance of differences observed in the numerical results across different sections. The differences in recall metrics reported in various tables may not always be statistically significant, and potential reasons for these differences need to be thoroughly explored. This implies that without statistical significance testing, the reliability of the results might be questioned, and the observed differences could be due to random variation rather than true performance improvements.

Future Work: For future work, conducting statistical significance tests or employing more robust metrics like Normalized Discounted Cumulative Gain (nDCG) [57] could provide a more accurate assessment of retrieval performance. Especially, nDCG considers ranking and similarity between the query and target images, which could be the potential new benchmark evaluation metrics for FashionIQ [64] and Shoes [20] datasets.

7.4 Reassessing Image Preprocessing Techniques

Limitation: One aspect that did not perform as expected was the application of the target pad image preprocessing method proposed by CLIP4CIR [4]. While it was expected to enhance retrieval performance, preliminary results in Table 4.1 showed similar recall performance to the CLIP standard image preprocessing method when applied to the linear multimodal fusion model. This finding implies that the target pad method may not offer the expected benefits in a linear model context, potentially

leading to inefficient preprocessing choices in future applications.

Future Work: This suggests that the target pad method may not be as effective in a linear context, and more studies are needed to explore how padding and cropping methods affect CLIP feature extraction across different model architectures. Investigating these preprocessing methods under different settings could provide insights into how to better optimize image preparation techniques for CIR tasks.

7.5 Quantifying GradCAM Visualization Effectiveness

Limitation: The qualitative nature of GradCAM visualization, while valuable for interpretability, could benefit from a more quantitative approach. This implies that the reliance on qualitative assessments alone may limit the ability to objectively compare model performance.

Future Work: Future research could develop a metric to quantify the effectiveness of GradCAM visualizations, offering a more objective way to assess whether the model's attention aligns with the most relevant features in the retrieval task. A quantitative approach would enhance the reliability of GradCAM as a tool for model interpretation in CIR.

7.6 Expanding Applications with Advanced Shopping Agents

Future Work: Regarding the future promising application of this study, I propose expanding the application of the CLIP-based multimodal fusion system with LLMs to incorporate a more advanced shopping agent, which has been a new research area proposed by recently published papers [6, 63, 10, 16]. This agent will be designed to better understand and adapt to the specific shopping stage the customer is in, such as product discovery, messy exploration, or a stage of specific fashion feature preference. The implication of developing such an agent is that it could revolutionize the shopping experience by providing highly personalized and dynamic assistance, which could lead to higher customer satisfaction and increased sales. By leveraging the capabilities of the LLM for sophisticated text processing and CLIP for precise visual-textual alignment, the agent can dynamically interpret and predict the customer's needs, thereby guiding them more effectively through their shopping journey. For instance, during the product

discovery stage, the agent could prioritize broader search results, offering a variety of styles and options. In contrast, in the specific feature preference stage, the agent would focus on narrowing down choices based on detailed textual inputs using a unified textual query studied in this paper, such as ”show me dresses with a V-neck and floral patterns”. This approach not only enhances the relevance of search results but also enriches the overall customer experience by providing a more intuitive and responsive shopping assistant. Future research could explore the integration of this advanced agent into the current system, aiming to further refine the user interaction model and improve the adaptability of the shopping assistant in real-time customer scenarios.

Bibliography

[1] Lorenzo Agnolucci, Alberto Baldrati, Marco Bertini, and Alberto Del Bimbo. isearle: Improving textual inversion for zero-shot composed image retrieval, 2024.

[2] Meta AI. Generative ai text and images: cm3leon. https://ai.meta.com/blog/generative-ai-text-images-cm3leon/, 2023.

[3] Yang Bai, Xinxing Xu, Yong Liu, Salman Khan, Fahad Khan, Wangmeng Zuo, Rick Siow Mong Goh, and Chun-Mei Feng. Sentence-level prompts benefit composed image retrieval, 2023.

[4] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Composed image retrieval using contrastive learning and task-oriented clip-based features. *ACM Transactions on Multimedia Computing, Communications and Applications*.

[5] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21434–21442, 2022.

[6] Oriol Barbany, Michael Huang, Xinliang Zhu, and Arnab Dhua. Leveraging large language models for multimodal search, 2024.

[7] Rachel Bick, Erika Halsey, and Christine C. Ekenga. The global environmental injustice of fast fashion. *Environmental Health*, 17(1):92, 2018.

[8] Junyang Chen and Hanjiang Lai. Pretrain like your inference: Masked tuning improves zero-shot composed image retrieval, 2023.

[9] Junyang Chen and Hanjiang Lai. Ranking-aware uncertainty for text-guided image retrieval, 2023.

[10] Qianqian Chen, Tianyi Zhang, Maowen Nie, Zheng Wang, Shihao Xu, Wei Shi, and Zhao Cao. Fashion-gpt: Integrating llms with fashion retrieval system. pages 69–78, 10 2023.

[11] Yanbei Chen, Shaogang Gong, and Loris Bazzani. Image search with text feedback by visiolinguistic attention learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2998–3008. IEEE, 2020.

[12] Yanzhe Chen, Jiahuan Zhou, and Yuxin Peng. Spirit: Style-guided patch interaction for fashion image retrieval with text feedback, 2024. Accepted as a conference paper in IEEE ICIP 2024.

[13] Guillaume Couairon, Matthieu Cord, Matthijs Douze, and Holger Schwenk. Embedding arithmetic of multimodal queries for image retrieval. In *O-DRUM Workshop at CVPR*, 2022.

[14] Guillaume Couairon et al. Embedding arithmetic of multimodal queries for image retrieval. *arXiv preprint arXiv:2112.03162*, 2022.

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

[16] Chun-Mei Feng, Yang Bai, Tao Luo, Zhen Li, Salman Khan, Wangmeng Zuo, Xinxing Xu, Rick Siow Mong Goh, and Yong Liu. Vqa4cir: Boosting composed image retrieval with visual question answering, 2023.

[17] Xavier Gastaldi. Shake-shake regularization, 2017.

[18] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, HeeJae Jun, Yoohoon Kang, and Sangdoo Yun. Compodiff: Versatile composed image retrieval with latent diffusion, 2024.

[19] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, Yoohoon Kang, and Sangdoo Yun. Language-only efficient training of zero-shot composed image retrieval, 2024.

[20] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. Dialog-based interactive image retrieval. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 676–686. Curran Associates, Inc., 2018.

[21] Y. Zhao C. Zhang S. Huang H. Zhu, Y. Wei. Amc: Adaptive multi-expert collaborative network for text-guided image retrieval, 2024. arXiv preprint arXiv:2312.12273.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[23] Steven A. Hicks, Inga Strümke, Vajira Thambawita, Malek Hammou, Michael A. Riegler, Pål Halvorsen, and Sravanthi Parasa. On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports*, 12(1):5979, 2022.

[24] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997.

[25] Young Kyun Jang, Dat Huynh, Ashish Shah, Wen-Kai Chen, and Ser-Nam Lim. Spherical linear interpolation and text-anchoring for zero-shot composed image retrieval, 2024.

[26] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision, 2021.

[27] Xintong Jiang, Yaxiong Wang, Mengjian Li, Yujiao Wu, Bingwen Hu, and Xueming Qian. 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*, volume 19742 of *SIGIR 2024*, page 2177–2187. ACM, July 2024.

[28] Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. Vision-by-language for training-free compositional image retrieval, 2024.

[29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[30] Xiang Zhang Chun-Liang Li Chen-Yu Lee Kate Saenko Tomas Pfister Kuniaki Saito, Kihyuk Sohn. Pic2word: Mapping pictures to words for zero-shot composed image retrieval, 2024. arXiv preprint arXiv:2302.03084.

[31] LAION. Large scale openclip: L/14, h/14, and g/14 trained on laion-2b. *LAION Blog*, 2022.

[32] Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Data roaming and quality assessment for composed image retrieval, 2023.

[33] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[34] Shenshen Li, Xing Xu, Xun Jiang, Fumin Shen, Zhe Sun, and Andrzej Cichocki. Cross-modal attention preservation with self-contrastive learning for composed query-based image retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20:1 – 22, 2024.

[35] Xiaoqing Li, Jiansheng Yang, and Jinwen Ma. Recent developments of content-based image retrieval (cbir). *Neurocomputing*, 452:675–689, 2021.

[36] Yi Li et al. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv preprint arXiv:2304.05653*, 2023.

[37] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[38] Yikun Liu, Jiangchao Yao, Ya Zhang, Yanfeng Wang, and Weidi Xie. Zero-shot composed text-image retrieval, 2024.

[39] Yu Liu, Guihe Qin, Haipeng Chen, Zhiyong Cheng, and Xun Yang. Causality-inspired invariant representation learning for text-based person retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14052–14060. AAAI Press, 2024.

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

[41] Zheyuan Liu, Weixuan Sun, Yicong Hong, Damien Teney, and Stephen Gould. Bi-directional training for composed image retrieval via text prompt learning, 2023.

[42] Zheyuan Liu, Weixuan Sun, Damien Teney, and Stephen Gould. Candidate set re-ranking for composed image retrieval with dual multi-modal encoder, 2024.

[43] MLFoundations. Openclip: An open source implementation of clip. *GitHub Repository*, 2022.

[44] Moonsift. Mapping shopping journeys with the help of ai. https://www.moonsift.com/guides/mapping-shopping-journeys-with-the-help-of-ai, 2024. Accessed August 21, 2024.

[45] OpenAI. Gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024.

[46] Anwesan Pal, Sahil Wadhwa, Ayush Jaiswal, Xu Zhang, Yue Wu, Rakesh Chada, Pradeep Natarajan, and Henrik I. Christensen. Fashionntm: Multi-turn fashion image retrieval via cascaded memory, 2023.

[47] Leigang Qu, Meng Liu, Jianlong Wu, Zan Gao, and Liqiang Nie. Dynamic modality interaction modeling for image-text retrieval. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1104–1113. ACM, 2021.

[48] Leigang Qu, Wenjie Wang, Yongqi Li, Hanwang Zhang, Liqiang Nie, and Tat-Seng Chua. Discriminative probing and tuning for text-to-image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–11. IEEE, 2024.

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[50] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024.

[51] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021.

[52] Chull Hwan Song, Taebaek Hwang, Jooyoung Yoon, Shunghyun Choi, and Yeong Hyeon Gu. Syncmask: Synchronized attentional masking for fashion-centric vision-language pretraining, 2024.

[53] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805, 2023.

[54] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[56] 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 Conference on Computer Vision and Pattern Recognition*, pages 6439–6448. IEEE, 2019.

[57] Yining Wang, Liwei Wang, Yuanzhi Li, Di He, Tie-Yan Liu, and Wei Chen. A theoretical analysis of ndcg type ranking measures, 2013.

[58] Zixiao Wang, Hongtao Xie, Yuxin Wang, Jianjun Xu, Boqiang Zhang, and Yongdong Zhang. Symmetrical linguistic feature distillation with clip for scene text recognition. In *Proceedings of the ACM International Conference on Multimedia*, pages 509–518. ACM, 2023.

[59] Haokun Wen, Xuemeng Song, Xiaolin Chen, Yinwei Wei, Liqiang Nie, and Tat-Seng Chua. Simple but effective raw-data level multimodal fusion for composed image retrieval. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2024.

[60] Haokun Wen, Xuemeng Song, Xin Yang, Yibing Zhan, and Liqiang Nie. Comprehensive linguistic-visual composition network for image retrieval. In *Proceedings*

of the International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 1369–1378. ACM, 2021.

[61] Haokun Wen, Xuemeng Song, Jianhua Yin, Jianlong Wu, Weili Guan, and Liqiang Nie. Self-training boosted multi-factor matching network for composed image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3665–3678, 2024.

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

[63] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models, 2023.

[64] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. 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*, pages 11307–11317, 2021.

[65] Yahui Xu, Yi Bin, Jiwei Wei, Yang Yang, Guoqing Wang, and Heng Tao Shen. Multi-modal transformer with global-local alignment for composed query image retrieval. *IEEE Transactions on Multimedia*, 25:8346–8357, 2023.

[66] Yahui Xu, Jiwei Wei, Yi Bin, Yang Yang, Zeyu Ma, and Heng Tao Shen. Set of diverse queries with uncertainty regularization for composed image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2024.

[67] Aman Yadav. Explainable ai for clip: The architecture explanation and its application for segment anything, 2023.

[68] Qu Yang, Mang Ye, Zhaohui Cai, Kehua Su, and Bo Du. Composed image retrieval via cross relation network with hierarchical aggregation transformer. *IEEE Transactions on Image Processing*, 32:4543–4554, 2023.

[69] Yuchen Yang, Min Wang, Wen gang Zhou, and Houqiang Li. Cross-modal joint prediction and alignment for composed query image retrieval. *Proceedings of the 29th ACM International Conference on Multimedia*, 2021.

[70] Feifei Zhang, Mingliang Xu, Qirong Mao, and Changsheng Xu. Joint attribute manipulation and modality alignment learning for composing text and image to image retrieval. pages 3367–3376, 10 2020.

[71] Feifei Zhang, Mingliang Xu, and Changsheng Xu. Geometry sensitive cross-modal reasoning for composed query based image retrieval. *IEEE Transactions on Image Processing*, 31:1000–1011, 2022.

[72] Yida Zhao, Yuqing Song, and Qin Jin. Progressive learning for image retrieval with hybrid-modality queries. *arXiv preprint arXiv:2204.11212*, 2022.

[73] Liang Zheng, Yi Yang, and Qi Tian. Sift meets cnn: A decade survey of instance retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(5):1224–1244, 2017.

[74] Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. Vista: Visualized text embedding for universal multi-modal retrieval, 2024.