Like this one, but in blue: A new Dataset for Composed Image Retrieval in Fashion e-Commerce

Anonymous ACL submission

Abstract

Product discovery in online fashion retail requires a nuanced understanding of the needs and wants of users. One key challenge is to support users to express refinements for an item that is *almost right*. We cast this problem as a composed image retrieval (CIR) task, and introduce a new dataset, called *Textual Tailoring*, containing 870K examples, that reflects the in-store shopping experience for discovering new fashion items. We address this task by adapting Large Vision Language Models to represent a fashion item, together with refinement queries such as *show me more items like this, but in blue*, and retrieve relevant products.

1 Introduction

Online fashion platforms present vast catalogs, making it difficult for users to find items that match precisely their evolving preferences. While keyword search and collaborative filtering offer broad discovery, they often lack the granularity for specific iterative refinements and struggle to effectively address complex, multi-faceted user queries. For instance, as Figure 1 shows, a user might like a displayed dress but wants to see it in a different color, fabric, or with altered design elements such as longer or shorter sleeves, different or mixed styles. Or, they might want to find something similar to a query image or find a fitting item to an already existing piece of clothing. This task has been recently formulated as Composed Image Retrieval (CIR) (Saito et al., 2023)[1]. However, progress in CIR, particularly for the nuanced modifications common in fashion e-commerce, has been partly constrained by the scarcity of large-scale, specialized datasets designed for such interactions.

To address this lacunae, we introduce *Textual Tailoring* (TT), a novel CIR dataset for the fashion domain: given an anchor fashion item (image) and a natural language query specifying the desired modifications, the system should retrieve or present an item that incorporates these changes (Figure 1). This interaction paradigm allows for more intuitive and controlled exploration of fashion assortments. It also poses several challenges, such as adaptation to, or grounding in, the existing catalog of items. This requires building adapters that would project the representations of the understood tailoring from rich general Large Vision Language Models (LVLM) to the deployed representation spaces of a domain-specific e-commerce platform. Crucially, to facilitate the development and robust evaluation of systems for CIR, and to address the aforementioned scarcity, we introduce a new, large-scale dataset specifically curated for

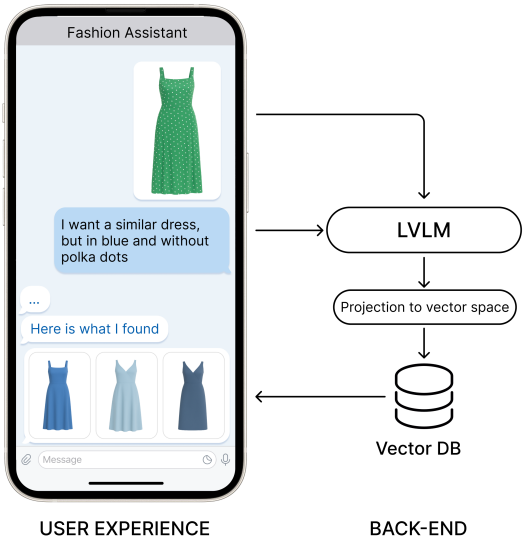

Figure 1: *Textual tailoring*: A user has found an item they like, but wants to make a change, s.a. the color or pattern of the item, and formulates the change as a text query. In the backend we use a LVLM that encodes the item the user found, together with their refinement query, and perform retrieval. Lastly, the user is presented with the items that match their query, while remaining faithful to the other aspects of the original item.

[1] Also known as Context-Based Image Retrieval and Text-Conditioned Image Retrieval (Baldrati et al., 2022a)

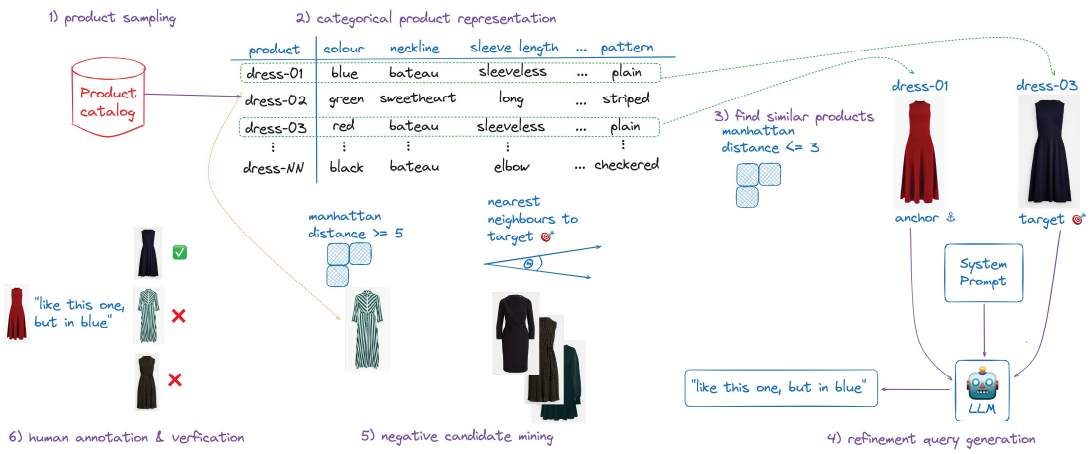

Figure 2: Dataset construction. After selecting a category and sampling products from it (1), we build a categorical representation based on product attributes (2), and select pairs – an anchor and a target – as positive candidates that have a manhattan distance of ≤ 3 (3). Subsequently, we pass the product images of a product pair to an LLM and task it to generate a refinement query for the given pair (4). We create negative candidates by (i) selecting products that have a manhattan distance of ≥ 5 to the anchor product and (ii) similar items to the target product via cosine similarity of product image embeddings (5). Lastly, we pass the positive and negative candidates, and the refinement query, to human annotators for verification and annotation.

this task. Fashion is a particularly good fit for the CIR task due to its visual nature and nuanced variety of styles, shapes, colours, patterns, etc. that a user might want to refine on any given product, while also being of high practical value.

This paper makes the following contributions:

1. We introduce the Textual Tailoring Dataset (TTD), containing over 870K (anchor image, modification query, target item) triplets[2].
2. We contribute a general methodology for CIR dataset construction from an existing product database that is applicable to other e-commerce domains.
3. We provide an experimental evaluation of the new dataset, leveraging no-train baselines and various architectures for fine-tuned LVLMs.

2 Related Work

Our work builds upon recent advancements in composed image retrieval (CIR) (Saito et al., 2023; Wang et al., 2025; Li et al., 2025; Huynh et al., 2025). LVLMs such as CLIP (Radford et al., 2021), LLaVA (Liu et al., 2023), and Qwen2-VL (Bai et al., 2023) have been extensively explored for the task of CIR. Incorporating commonsense knowledge, these models have shown strong capabilities in joint image-text understanding and generation. Conventional CIR models mainly rely on multimodal fusion approaches, which combine image and text features obtained from separate visual and textual encoders. In Baldrati et al. (2022b), the authors propose a Combiner model with fine-tuning of the CLIP text tower, which integrates embeddings of a reference image and a refinement query by optimizing a contrastive loss that aligns the composed representation with the target image in a shared embedding space. A limitation of these methods is the lack of large scale, reliable datasets of triplets—comprising an anchor image, a refinement query, and a target image. To address this, Saito et al. (2023) proposes training a mapper network that maps the output of the CLIP image tower to the input space of the text tower using unlabeled data with only the anchor image and the refinement query. Similarly, Liu et al. (2021) propose an early-fusion strategy in which the output of a ResNet model (He et al., 2015) applied to the reference image is integrated into the text input space and processed by a Transformer model. A key limitation of these methods is their inability to handle complex queries such as those containing negation. To facilitate research advancements in this area, several datasets have been introduced (Huynh et al., 2025; Gu et al., 2024; Wu et al., 2020). Notably, the FashionIQ dataset has been a primary resource for exploring the problem of Composed Image Retrieval within the fashion domain. Our dataset presents a substantial enhancement that addresses the most important limitations of the FashionIQ dataset, such as its small sam-

[2]Github link for dataset and code: LINK WILL BE ADDED IN CAMERA-READY VERSION

Figure 3: Examples from the Textual Tailoring dataset: green boxes indicate positives, red boxes indicate negatives.

ple size and restricted range of fashion categories, by providing a larger variety of nuance and depth of fashion products, while covering a wider range of product categories. Furthermore, we address this task by leveraging the language understanding capabilities of Large Vision-Language Models (LVLMs), by following Gunter et al. (2024), and training an adapter specifically for the task of article retrieval that matches the article embeddings in the production system.

3 Textual Tailoring for Fashion

Consider an anchor image I_a of a fashion product along with a natural language refinement query Q_r. In CIR, the task is to retrieve a target product image I_t from a product database $\mathcal{C}$. This results in I_t being visually similar to I_a and incorporating the refinement specified by Q_r (i.e., guidance). For example, if I_a is a red crew-neck t-shirt and Q_r is "in blue with a V-neck," I_t should be a blue V-neck t-shirt that otherwise resembles I_a. Here, the definition of similarity and guidance can be adjusted. In curating our dataset, we have implicitly used human judgment from fashion experts to implement both similarity and guidance.

3.1 The Textual Tailoring dataset

To develop and evaluate the adaptability of LVLMs for the Composed Image Retrieval task, we introduce and utilize a novel dataset — 'Textual Tailoring' (TT). This dataset consists of triplets $(I_a, Q_r, \{I_t^i\}_{i=1}^{10})$, where I_a is an anchor product image, Q_r is the textual refinement query, and $\{I_t^i\}_{i=1}^{10}$ is a set of 10 images for target products. If I_t^i satisfies the query relative to I_a it is annotated as $l_t^i = 1$, otherwise $l_t^i = 0$.

Dataset construction methodology. We constructed the Textual Tailoring dataset on the basis of our internal product database. Figure 2 visualises the dataset construction process. We first select the target categories, and for each category we manually choose visually salient attributes (e.g. *colour* and *neckline* for *dresses*, or *toe shape* and *heel height* for *pumps*). We selected 15-25 attributes per category this way. We then randomly sample products for each category and create a categorical product representation for each product based on the selected attributes.

Next, we compute the pairwise Manhattan distances for all selected products, and sample anchor-target pairs (I_a, I_t) with a Manhattan distance of ≤ 3 as positive candidates. We pass each candidate pair to an LLM with the task to generate a suitable refinement query (Q_r).

The LLM is prompted with an in-store shopping scenario where a customer is presented the anchor item (I_a) by a shop assistant, but is actually looking for the product depicted in the target image (I_t). The LLM is given the product images, but not the attributes that differed, i.e. part of the task of the LLM is to identify the differences and formulate a refinement query on that basis. For the construction of our dataset we used GPT-4o to generate an English refinement query.[3]

We apply two methods for mining negative candidate products, (i) selecting products with a Manhattan distance ≥ 5 to the anchor product, and (ii) similar items to the target product, identified via cosine similarity of product image embeddings.

As the final step, we employ human annotators to verify the validity and coherence of the generated refinement query, and annotate the positive and negative product candidate pairs. All anchor, target, refinement query triplets, (I_a, Q_r, I_t), have been labeled by two human annotators, with a third annotator acting as tie breaker. All three human annotators are domain experts.

The final dataset contains 869,975 triplets, which are split into 611,175 triplets for training, 85,733 for validation, and 173,067 for evaluation, and is

[3]We note that our methodology is agnostic to the actual LLM used for generating refinement queries.

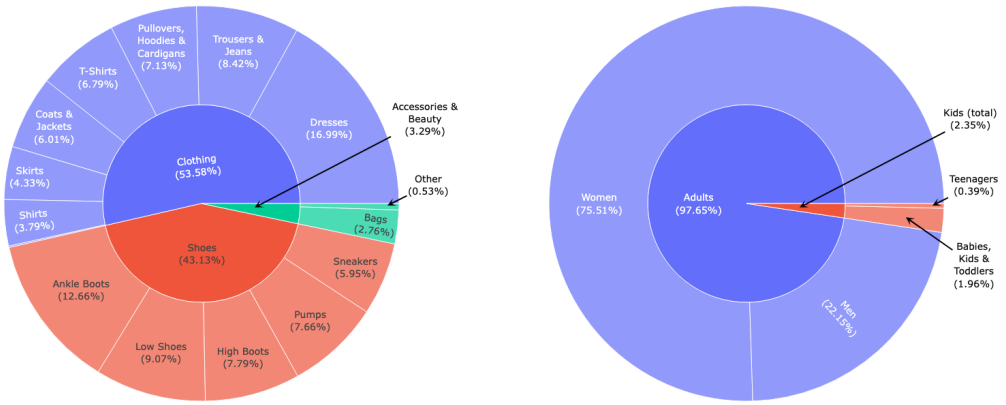

Figure 4: Category distribution (left) and target group distribution (right) of the textual tailoring dataset. The dataset covers products from 9 clothing categories, 5 shoe categories, 5 accessory categories, as well as beauty items. These categories are distributed across women's, men's and kids' items. All products are real fashion items that have been available for purchase.

available here. Figure 3 shows examples of anchor image, refinement query and positive and negative target image.

Product data statistics. Figure 4 shows the distribution of products in our dataset. The majority of items are various types of clothing and shoes, and are consistently distributed across 9 clothing categories and 5 shoe categories, with a smaller proportion of the data covering various accessories (e.g. bags, watches, etc.) and beauty items. Table 1 provides a complementary view on the category distribution, highlighting the datasets coverage of a substantial range and depth of products. Figure 4 furthermore highlights that the target group distribution is more uneven, with approx. 75% of products being targeted at women, with 22% for men, and the remaining 3% for kids' items, including for babies. Table 2 complements Figure 4 with a more granular view on individual groups.

Refinement queries. The utility of the dataset predominantly hinges on the quality of the refinement queries for any given anchor-target pair. For refinement query generation we used GPT-4o, and leveraged chain-of-thought prompting (Wei et al., 2022) with few-shot examples (Brown et al., 2020). We included good and bad examples of refinement queries — together with detailed reasoning — in our prompt.[4] Lastly, to ensure the automatically generated queries are of high-quality and suitable for the given pair of products, we employed domain experts for verifying the quality of the refinement queries and annotating the final (I_a, Q_r, I_t) triplets.

[4]We will share the full prompt with the dataset, for revision the prompt is available here.

Group / Category	# of items (%)
Shoes / Ankle boots	44,411 (12.66%)
Shoes / High boots	27,327 (7.79%)
Shoes / Low shoes	31,825 (9.07%)
Shoes / Pumps	26,883 (7.66%)
Shoes / Sneakers	20,867 (5.95%)
Clothing / Coats & jackets	21,086 (6.01%)
Clothing / Dresses, one-piece combos	59,606 (16.99%)
Clothing / Pullovers, hoodies & cardigans	25,019 (7.13%)
Clothing / Shirts	13,280 (3.79%)
Clothing / Skirts	15,194 (4.33%)
Clothing / T-shirts	23,812 (6.79%)
Clothing / Trousers & jeans	29,552 (8.42%)
Clothing / Beachwear, underwear & nightwear	24 (0.01%)
Accessories & Beauty / Bags & wallets	9,688 (2.76%)
Accessories & Beauty / Headgear	123 (0.04%)
Accessories & Beauty / Lip cosmetics	318 (0.09%)
Accessories & Beauty / Scarves	177 (0.05%)
Accessories & Beauty / Watches	1,215 (0.35%)
Accessories & Beauty / Misc.	5 (0.00%)

Table 1: Fashion items per group and category.

Age group / Target group	# of items (%)
Adults / Women	264,884 (75.51%)
Adults / Men	77,702 (22.15%)
Kids / Babies & Toddlers	576 (0.16%)
Kids / Kids	6,295 (1.79%)
Kids / Teenagers	1,356 (0.39%)

Table 2: Fashion items per age and target group.

After generation and annotation, we analysed the style, length and quality of the refinement queries. Figure 5 and Table 3 show the distribution of refinement lengths, together with an example for each length bucket. The examples show that the in-store shopping scenario is well reflected in the queries.

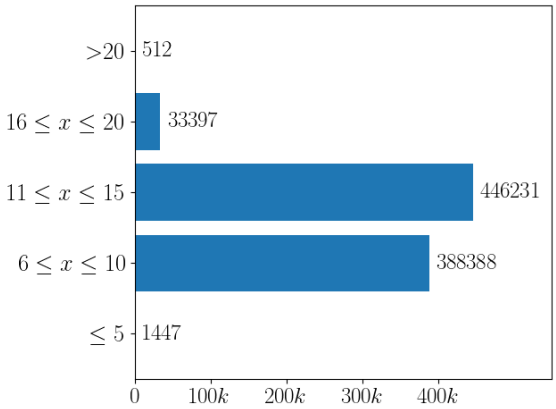

Figure 5: Length distribution (in words) of refinement queries.

Query length	Examples
< 20	I like the style but could you find it in a more vibrant red and with a more varied floral design?
$16 \leq x \leq 20$	Can I get a pair like this but in dark blue and with a skinny fit?
$11 \leq x \leq 15$	I am looking for a similar top but with a more breathable open-knit texture.
$6 \leq x \leq 10$	I would like it with a block heel instead.
≤ 5	Like this one but shorter.

Table 3: Refinement query examples by length bucket.

Our method for creating positive anchor-target pairs ensured sufficient variation for target product candidates that enabled the LLM to generate diverse refinement queries. We created 5 categories to asses the stylistic complexity of the queries:

- **simple**: A query changing only one attribute.
- **conjunction**: Refinement queries with at least one conjunction, i.e. where the anchor and target products differ in more than one feature.
- **negation**: Refinement queries where the target product *does not* have a particular feature that the anchor product has.
- **conjunction & negation**: A complex query containing a conjunction and a negation.
- **at least two conjunctions**: A complex query containing at least 2 conjunctions.

Figure 6 and Table 4 show the number of queries for each category, as well as an example for each bucket. More than half of the queries have at least a negation[5] or a conjunction, with a considerable amount containing both, or 2 or more conjunctions.

Human annotation. We focused on two classes:

- **relevant**: The retrieved product satisfies most of the specifications in the query and is similar to the anchor item.
- **not relevant**: The retrieved product does not satisfy the specifications in the refinement query, or is in any other way dissimilar to the anchor item.

We decided against a more granular annotation scale to reduce mental load on annotators and to decrease the influence of the inherent subjectivity of the task on the results. Annotation was done for all triplets, and annotators were instructed to flag any low-quality queries. For triplets where the initial annotators disagreed (15.8% of cases), we employed a third annotator to break the tie.[6]

Comparison with FashionIQ: Although the FashionIQ (FIQ) dataset (Wu et al., 2020) allows for a similar task, it differs from the Textual Tailoring dataset in many important ways. **(1)** FIQ contains about 30K data points with low resolution images. Conversely, our dataset contains over 870K data points with high resolution images. **(2)** the FIQ dataset is split into only three categories, namely dresses, shirts, and tops-tees. In contrast, the Textual Tailoring dataset is hierarchically split across: 3 demographics (women, men, kids), 3 main categories (clothing, shoes, accessories & beauty), and 20 subcategories, covering fashion in a wider, deeper and more diverse way. **(3)** the two datasets also strongly differ in the nature of the textual description between the reference and target image. While queries in FIQ are frequently describing the difference of two items (e.g. *The top is orange in color and more flowy.*), rather than expressing a refinement, our use of LLMs for the TT dataset creates more flexibility on instantiating a faithful in-store shopping scenario, which is reflected in the style of our refinement queries.

[5] See Appendix A.1.

[6] Although adding more annotators could potentially decrease annotator bias, we decided to focus more on consistency and the expertise of fashion experts.

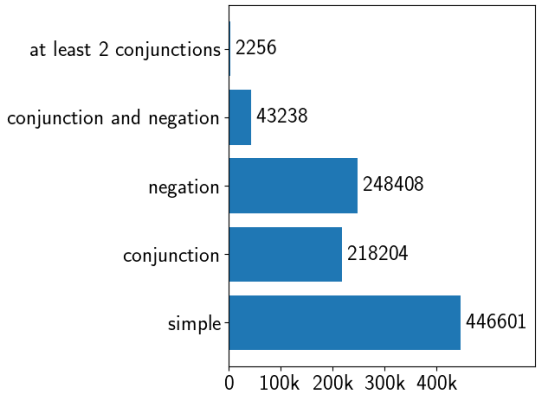

Figure 6: Complexity category distribution of refinement queries.

Category	Examples
simple	Can I get this skirt in a neutral tone?
conjunction	I would like this in a lighter grey tone and with shorter sleeves.
negation	I would prefer it without the pearl decorations.
conjunction & negation	I'm looking for a dress in white and without the criss-cross straps.
at least 2 conjunctions	Can I have it in plain navy and knee-length instead of with flowers and longer?

Table 4: Refinement query examples by complexity.

4 Models and Methods

4.1 Zero-shot Baselines

Image-Only. Retrieves target images based on image embedding similarity between I_a and I_t.
Text-Only. Retrieves target images based on embedding similarity between Q_r and I_t.
CLIP. Averages the CLIP embeddings for I_a and Q_r to retrieve I_t.
LVLM Textual Output with CLIP Retrieval. This baseline uses a pre-trained LVLM[7] in a zero-shot setting. For an anchor image I_a and query Q_m, the LVLM is tasked to create a descriptive search query that reflects the desired modified item, e.g.:

```
User: [Image of red sneaker]
      "Could you bring me a pair
       similar to this but from
       Onitsuka Tiger?"
Qwen: "White Onitsuka Tiger sneakers
        with colorful stripes..."
```

This generated query is then used with CLIP to find matching items.

4.2 Trained Baselines

CLIP Embedding Combiner. We use a *Combiner* to fuse CLIP anchor image embeddings $f(I_a)$ with refinement query embeddings $g(Q_m)$. The goal of this fusion is to align the multimodal representations and approximate the target image embedding $f(I_t)$. To achieve this, we define a fusion function $s(f(I_a), g(Q_m)) = W \cdot [f(I_a) \,\|\, g(Q_m)]$, where W is a learnable weight matrix and $\|$ denotes concatenation. The training objective combines an L2 loss with a contrastive loss to encourage better separation between relevant and irrelevant targets. During inference, we compute $E_{comb} = s(f(I_a), g(Q_m))$. Retrieval is carried out by identifying catalog items whose CLIP embeddings are most similar to E_{comb}.

[7] Qwen2-VL-72B-Instruct in our case

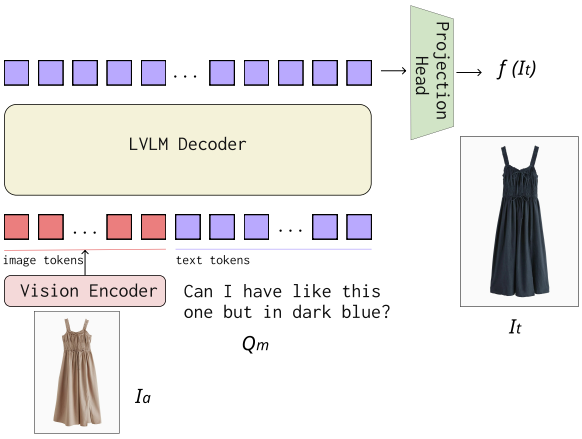

Figure 7: Architecture of the Qwen MLP Head model.

LVLM with Projection Head utilizes a pre-trained LVLM (i.e, Qwen2-VL-7B) and adapts it for downstream tasks through lightweight textual fine-tuning. The architecture (see Figure 7) consists of: *i) LVLM.* The anchor image I_a and the query Q_m are input into the LVLM to produce a joint multimodal representation. The outputs from the final transformer layer are passed to the projection head. *ii) Trainable projection head.* It uses a custom attention pooling layer with a learned latent query vector, applied via scaled dot-product attention with residual connections, with two projection layers. The model is trained to predict the internal CLIP embedding of the target image I_t. The loss function is cosine similarity loss, calculated between the output of the projection head and the

CLIP embedding of the target I_t. We evaluate the model using two training configurations:

(1) Frozen LVLM with trainable projection head. In this configuration, only the Projection Head is trained, while the LVLM remains frozen.

(2) LVLM + LoRa with trainable projection head. This configuration extends the previous approach by introducing lightweight fine-tuning of the LVLM using Low-Rank Adaptation (LoRA).

During inference, the projection head produces an embedding for the (I_a, Q_m) pair, used to retrieve the top-k most similar items from a pre-computed index of catalog CLIP embeddings.

5 Experiments and Results

Evaluation Metrics. We evaluate retrieval performance using Mean Reciprocal Rank (MRR), and Success@k – the proportion of queries for which a correct target is found within the top k results; k $\in \{10, 50, 100\}$.

Results. The results (Table 5) clearly indicate the superiority of the fine-tuned approaches. Our proposed Frozen LVLM with trainable projection head achieves an MRR of 0.108 and ranks the correct item in the top 10 for 20.5% of queries. This is a substantial improvement over over all zero-shot baselines. Among the zero-shot baselines, the Image-only model, which performs image similarity search, sets a reasonable lower bound (MRR 0.065, S@10 0.149). The Text-only baseline, which relies solely on textual matching without leveraging visual cues, performs worst overall (MRR 0.028, S@10 0.059), confirming the importance of multimodal alignment. The CLIP (Average Image+Text) baseline improves upon these results (MRR 0.076, S@10 0.161), showing that simple fusion of modalities can already capture complementary information, though it remains limited without fine-tuning. The LVLM Textual Output baseline (MRR 0.074, S@10 0.146) performs comparably to CLIP averaging, demonstrating that while pretrained LVLMs encode cross-modal representations, they are not fully optimized for retrieval tasks in a zero-shot setting. In contrast, all fine-tuned baselines deliver marked improvements. The CLIP Embedding Combiner achieves the strongest overall performance (MRR 0.114, S@10 0.237), indicating that jointly adapting image and text embeddings with a learned fusion module can effectively align modalities for retrieval. Our Frozen LVLM + Proj. Head closely follows (MRR 0.108, S@10 0.205), outperforming all zero-shot variants and confirming that even limited adaptation on top of frozen pretrained features can yield substantial gains. Finally, the LVLM + LoRa + Proj. Head variant (MRR 0.112, S@10 0.211) shows competitive results, suggesting that parameter-efficient fine-tuning can further enhance multimodal representations. Notably, the fine-tuned 7B parameter model outperforms the much larger 72B parameter model used in a zero-shot prompting setup. The CLIP embedding combiner also performs very competitively, suggesting that specialized fusion mechanisms can be highly effective.

Approach	**Success@k ↑**			**MRR ↑**
	@10	**@50**	**@100**	
Zero-shot Baselines				
Image-only	0.149	0.328	0.407	0.065
Text-only	0.059	0.132	0.173	0.028
CLIP (Avg. Image+Text)	0.161	0.350	0.446	0.076
LVLM Textual Output	0.146	0.346	0.463	0.074
Trained Baselines				
CLIP Embedding Combiner	**0.237**	**0.508**	**0.633**	**0.114**
Frozen LVLM + Proj. Head	0.205	0.469	0.594	0.108
LVLM + LoRa + Proj. Head	0.211	0.475	0.6	0.112

Table 5: Performance on the Textual Tailoring test set.

6 Conclusion

In this work, we introduced the Textual Tailoring dataset for Composed Image Retrieval. The dataset encapsulates a highly practical task for nuanced fashion product discovery. Furthermore, we proposed a general methodology for creating CIR datasets from existing product databases. We presented an experimental system to evaluate various approaches on our new dataset, demonstrating that fine-tuning a lightweight MLP head on features from a pre-trained VLM without and with LoRA fine-tuning outperforms existing baselines and zero-shot VLM prompting. Our findings showcase a successful possibility of general LVLM adaptation to increase impact in specialized domains, opening promising avenues for further research into instruction-driven multimodal systems for complex product discovery and user-centric online retail experiences.

7 Limitations

Our study, while demonstrating the potential of adapting LVLMs for fashion, has several limitations that offer suggestions for future work.

Dataset-Related Limitations:
(1) The introduced TT dataset is constructed based on a *predefined set of visually salient attributes*. This may not encompass the exhaustive range of fashion attributes or some of their subtle nuances. Consequently, the model's ability to understand and act upon modifications related to missing attributes or highly specific details not well-represented in the dataset might be constrained. The selection of anchor-target pairs based on a fixed number of differing attributes might also simplify complex sartorial relationships. One way to mitigate this could be use conditional generative image models to generate target images given the modifier query, in the manner proposed in the (Yin et al., 2025). Adding such generated images with fine-grained differences would make the problem much harder, although less realistic, given that in practice the assortment being searched over is always limited, so it's possible that such target products do not exist.
(2) The textual queries, though curated via GPT-4o and human annotation, may contain *ambiguities* or not fully represent the diversity of genuine user language. We did instruct the annotators to filter out unrealistic sounding, not human-like refinements. The queries have mean length 60$\pm$13 characters, which is longer than a typical search query, but closer in length to the queries where a customer has a specific idea in mind, and needs to be more verbose to express their intent. More significantly, fashion language is dynamic, with new trends and "cold fashion semantics" emerging constantly. The current static datasets (TT and FashionIQ) might not equip the model to effectively interpret or act on queries related to very recent trends or novel stylistic expressions not present in its training. While strategies like RAG could potentially mitigate this for new textual concepts, they are not part of the current evaluation framework.
(3) The primary language for queries in the evaluated datasets is English. This limits the direct applicability of the trained models in non-English speaking contexts or for *fashion terminologies and tailoring preferences unique to other cultures*.

All of these issues related to synthetic-real gap will be addressed by obtaining the reformulation data from the customers using the deployed system trained that can solve this task.

Methodological and Technical Limitations:
(1) The proposed Frozen LVLM + Proj. Head approach relies on a projection head to project features from the frozen Qwen2-VL model into the target internal CLIP embedding space. This projection is an approximation. The projection, being a relatively small network, may *not project LVLM embeddings with high fidelity* to the production model space for all types of complex modifications, potentially leading to a loss of nuanced information captured by the larger LVLM. (2) Utilizing a frozen LVLM backbone is computationally efficient for fine-tuning but means the core visual and textual representations are *static*, fixed by the LVLM's original pre-training. This can lead to *representation bias*, where inherent biases (e.g., regarding specific fashion subcultures, niche aesthetics, or newly emerged concepts underrepresented in pre-training data) or knowledge gaps of the pre-trained LVLM are propagated. The lightweight projection head might not be able to fully rectify these, in contrast to approaches involving dynamic fine-tuning of the LVLM itself.

Scope and Generalizability Limitations: (1) The textual tailoring task, as defined and evaluated, is framed as a *single-turn interaction*. Real-world shopping experiences often involve iterative, multi-turn dialogues for progressive refinement, which is outside the scope of the current work. (2) While this work focuses on fashion, adapting the proposed approaches to other e-commerce domains (e.g. home decor, electronics) or even to different fashion catalogs (e.g., with different attribute distributions or image styles) would require substantial dataset re-curation (though the proposed methodology is transferable) and retraining of the projection head, and potentially different choices for the base LVLM and target embedding space.

Ethical considerations include the risk of inadvertently amplifying subtle societal biases present in fashion data relating to socioeconomic status (e.g., by implicitly favoring characteristics of higher-priced items), body image (if successful tailoring is less effective or available for underrepresented body types), or age. Additionally, the energy consumption required for training and deploying LVLMs, even the adapter-based versions in this work, remains a pertinent environmental concern for AI systems. Finally, ensuring that the textual tailoring interface and its outputs are fully accessible to users with diverse abilities (e.g., visual, motor impairments) is crucial for equitable application and to prevent digital exclusion.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and more. *arXiv preprint arXiv:2308.12966*.

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2022a. Conditioned and composed image retrieval combining and partially fine-tuning clip-based features. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4955–4964.

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2022b. Conditioned and composed image retrieval combining and partially fine-tuning clip-based features. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4955–4964.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shay, Burak Ring, and Daniela Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Geonmo Gu, Sanghyuk Chun, Wonjae Kim, HeeJae Jun, Yoohoon Kang, and Sangdoo Yun. 2024. Compodiff: Versatile composed image retrieval with latent diffusion. *Transactions on Machine Learning Research*. Expert Certification.

Tom Gunter, Zirui Wang, Chong Wang, Ruoming Pang, Andy Narayanan, Aonan Zhang, Bowen Zhang, Chen Chen, Chung-Cheng Chiu, David Qiu, Deepak Gopinath, Dian Ang Yap, Dong Yin, Feng Nan, Floris Weers, Guoli Yin, Haoshuo Huang, Jianyu Wang, Jiarui Lu, and 136 others. 2024. Apple intelligence foundation language models. *Preprint*, arXiv:2407.21075.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Chuong Huynh, Jinyu Yang, Ashish Tawari, Mubarak Shah, Son Tran, Raffay Hamid, Trishul Chilimbi, and Abhinav Shrivastava. 2025. Collm: A large language model for composed image retrieval. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3994–4004.

You Li, Fan Ma, and Yi Yang. 2025. Imagine and seek: Improving composed image retrieval with an imagined proxy. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3984–3993.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. 2021. Image retrieval on real-life images with pre-trained vision-and-language models. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2105–2114.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

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

Lan Wang, Wei Ao, Vishnu Naresh Boddeti, and Ser-Nam Lim. 2025. Generative zero-shot composed image retrieval. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29690–29700.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Ni, Xiaohua Cao, Jonathan Ku, and Quoc V. Lu. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. 2020. Fashion iq: A new dataset towards retrieving images by natural language feedback. *Preprint*, arXiv:1905.12794.

Deqiang Yin, Junyi Guo, Huanda Lu, Fangyu Wu, and Dongming Lu. 2025. Editgarment: An instruction-based garment editing dataset constructed with automated mllm synthesis and semantic-aware evaluation. *Preprint*, arXiv:2508.03497.

A Appendix

A.1 Dataset

A.1.1 General Information

Figure 3 shows examples of anchor image, refinement query and positive and negative target image. The final dataset contains 611,175 triplets for training, 85,733 for validation, and 173,067 for testing. The test set of the TTD contains 14,025 unique (I_a, Q_m) pairs, with a total of 75,564 unique target SKUs. Each pair is associated with a list of positive and negative target SKUs for ranking evaluation. An example data point might be (SKU_dress1, "show me this in green", {SKU_dress2_green, SKU_dress3_green}). The primary language for queries in the evaluated dataset is English. The purpose of TDD is to support the development, training, and rigorous evaluation of models for the textual tailoring task in the fashion domain, where a system retrieves items that match an anchor image as modified by a natural language query.

A.1.2 Negation list

The following is a list of negating words that we included in our complexity analysis:

- without
- not
- doesnt, doesn't
- excludes, exclude, excluding
- except
- avoid
- never
- instead

A.2 Experimental setup for LVLM with trainable projection head baseline

In the Frozen LVLM with trainable projection head baseline, LVLM outputs are precalculated and cached to avoid redundant forward passes, significantly improving training efficiency. The projection head includes one attention layer with 8 heads, residual connections, two intermediate projection layers, and two final feed-forward layers (both of size 512) with GELU activations. The total number of trainable parameters is approximately 80M. Training is conducted for 25 epochs using the Adam optimizer with a learning rate of 0.001, a batch size of 1280, and in-batch negatives. It completes in around 4 hours on a single NVIDIA H200 (141 GB) GPU.

In the LVLM with LoRA fine-tuning baseline, LoRA adapters with rank 16 (chosen for optimal performance) are inserted into the LVLM's attention layers—specifically the query and value projections. The combined number of trainable parameters—including the projection head and LoRA adapters—totals approximately 85 million, whereas the full model encompasses 7.8 billion parameters. The Projection Head and adapters are fine-tuned jointly. A batch size of 15 is used, and for each positive example, a random negative sample is drawn from the dataset. Training is conducted on an NVIDIA H200 (141 GB) instance with 8 GPUs using data parallelism. The model is trained for 26 hours using the Adam optimizer with a learning rate of 2e-4 and a cosine learning rate scheduler.

A.3 Avg. CLIP (Image+Text) details

Computes separate CLIP embeddings for the anchor image $f(I_a)$ and the textual query $g(Q_m)$, where $f(.)$ and $g(.)$ are the CLIP image and text encoders, respectively. A combined query embedding $E_{comb} = \alpha f(I_a) + (1 - \alpha) g(Q_m)$ is formed (or using vector arithmetic like $f(I_a) + g(Q_m) - g(\text{generic reference text})$). Retrieval is performed by finding catalog items whose CLIP embeddings are closest to E_{comb}.

A.4 Artifact Licenses

We release Textual Tailoring dataset under MIT license[8]. The other artifacts that were used in this work are licensed as follows:

1. FashionIQ (FIQ) dataset is published under Community Data License Agreement (CDLA) License[9].
2. Qwen2-VL-7B-Instruct[10] is available under Apache 2.0 license[11].
3. Qwen/Qwen2-VL-72B-Instruct[12] is available under tongyi-qianwen licence[13]

All artifacts are used according to their intended use.

[8] https://opensource.org/license/mit
[9] https://cdla.dev/
[10] https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct
[11] https://choosealicense.com/licenses/apache-2.0/
[12] https://huggingface.co/Qwen/Qwen2-VL-72B-Instruct
[13] https://huggingface.co/Qwen/Qwen2-VL-72B-Instruct/blob/main/LICENSE