DualCIR: Enhancing Training-Free Composed Image Retrieval via Dual-directional Descriptions -Supplementary-

Jingjiao Zhao[1], Jiaju Li[2], Dongze Lian[3], Liguo Sun[1], Pin Lv[1]*
[1]Institute of Automation, Chinese Academy of Sciences (CASIA)
[2]Capital Engineering & Research Incorporation Limited (CERI), [3]ShanghaiTech University
{jingjiao.zhao, liguo.sun, pin.lv}@ia.ac.cn, lijiaju@ceri.com.cn, dzlianx@gmail.com

1. Additional Implementation Details

We refer to [3] and [4], following the Chain-of-Thought [5], to design the prompt for the LLM. We write the sample prompt to generate the descriptions: t_e, t_p, and t_n:

"I have an image. Given an instruction to edit the image, first generate a description of the edited image, and then detect the positive and negative parts in the generated description. Specifically, first, generate the description of the edited image. Avoid adding imaginary things. Then, divide the description into positive and negative attributes. Positive means the parts will appear in the edited image, and negative means the parts not appear in the edited image. Positive and Negative should be as simple as possible and combined by noun phrases in a few words."

2. Qualitative Results in the Field of Fashion

Due to space constraints in the main text, we only present qualitative results for the natural image dataset CIRCO [1]. However, composed image retrieval can also be applied to other domains, such as fashion. Therefore, we supplement our findings with results from the FashionIQ [2] validation set.

Figures 1-3 display some successful retrieval examples of the dress, shirt, and top tee subsets respectively. It can be observed from the figures that DualCIR can retrieve results that better match the semantics of the modified text compared to CIReVL [3]. For example, in the first row of Figure 1, where the modified text specifies "above the knees", CIReVL fails to deliver accurate results while DualCIR successfully obtains them. Furthermore, for more challenging queries such as the first row of Figure 3 requesting an "Ed Hardy shirt" in the target image, DualCIR successfully identifies a shirt with "Ed Hardy" printed on it.

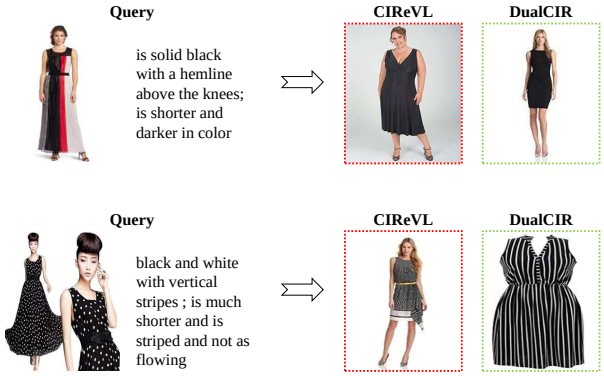

Figure 1. **Retrieval examples on the dress subset of FashionIQ.** The red boxes indicate incorrect retrieval results, while the green boxes represent correct results.

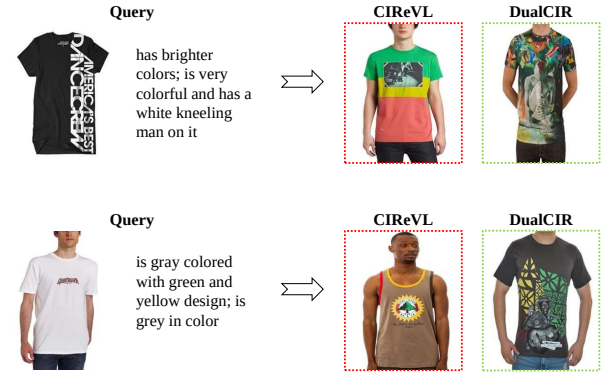

Figure 2. **Retrieval examples on the shirt subset of FashionIQ.** The red boxes indicate incorrect retrieval results, while the green boxes represent correct results.

*Corresponding author.

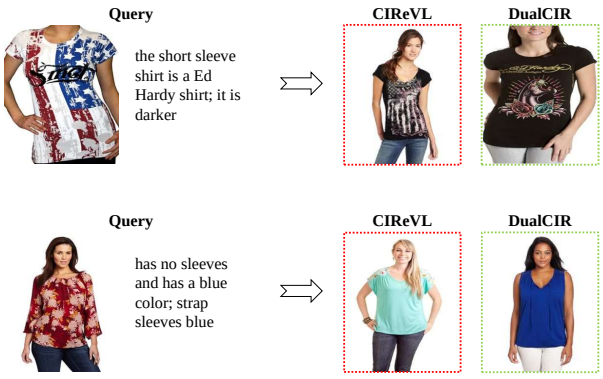

Figure 3. **Retrieval examples on the toptee subset of FashionIQ.** The red boxes indicate incorrect retrieval results, while the green boxes represent correct results.

We also present two typical failure cases in Figure 4 and Figure 5. Due to issues in the labeling and modified text design of FashionIQ, the recall score does not fully reflect the method's performance. However, it is evident from the examples that DualCIR's retrieval results are mostly reasonable based on the existing query requirements.

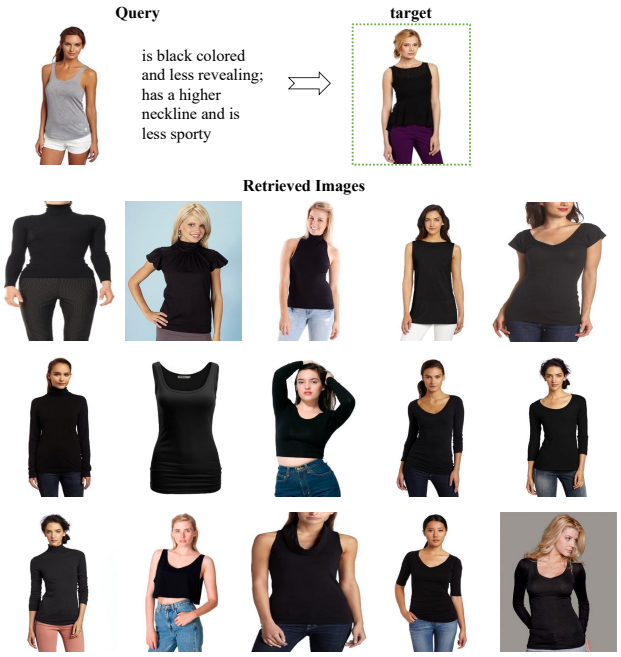

Figure 4. **A failure case on FashionIQ.** The initial 15 retrieved images do not match the target image. However, it is notable that almost all the retrieved images meet the query requirement ("is black colored and less revealing; has a higher neckline and is less sporty") and should be considered valid retrieval results.

Figure 5. **A failure case on FashionIQ.**The first 15 retrieved images did not match the target image. In this example (the target image needs to be "less attractive and more colorful"), "attractive" is subjective and difficult to assess as a retrieval criterion. At the same time, the shirt in the ground truth is red, which may not meet the requirement of being "colorful."

References

[1] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15338–15347, 2023. 1

[2] Xiaoxiao Guo, Hui Wu, Yupeng Gao, Steven Rennie, and Rogerio Feris. The fashion iq dataset: Retrieving images by combining side information and relative natural language feedback. *arXiv preprint arXiv:1905.12794*, 1(2):7, 2019. 1

[3] Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. Vision-by-language for training-free compositional image retrieval. *arXiv preprint arXiv:2310.09291*, 2023. 1

[4] Shitong Sun, Fanghua Ye, and Shaogang Gong. Training-free zero-shot composed image retrieval with local concept reranking. *arXiv preprint arXiv:2312.08924*, 2023. 1

[5] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 1