

Imagine and Seek: Improving Composed Image Retrieval with an Imagined Proxy

You Li Fan Ma Yi Yang[†]
ReLER, CCAI, Zhejiang University, Zhejiang, China

[†] Corresponding author

{uli2000, mafan, yangyics}@zju.edu.cn

Appendix

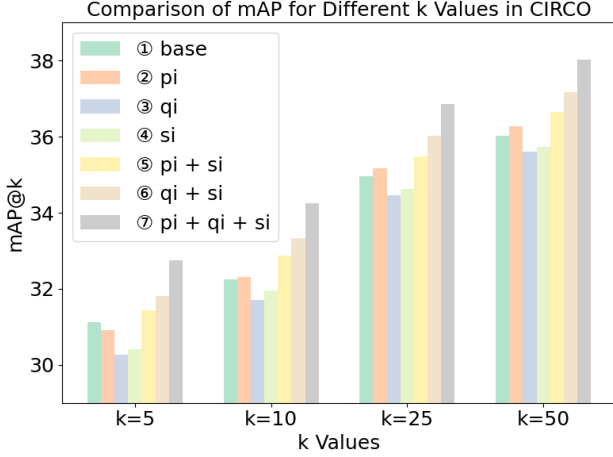


Figure 1. **Ablation results on the composition of robust proxy features** in the CIRCO dataset. **pi** indicates proxy features, **si** represents semantic perturbation, and **qi** indicates the query features.

A. More Analysis

In this section, we will conduct a deeper analysis of the experiments to better illustrate the details of our method.

A.1. The Construction of Robust Proxy

We analyze **the specific roles of each component in the robust features through Fig.1 and more visual examples in Fig.2**. The RP features include proxy image features, query image features, and semantic perturbation. Using the CIRCO dataset as an example, we construct RP features with different components for retrieval and obtain the top-1 retrieved image. This allows us to examine how varying the RP features influences the retrieved images.

We can draw the following conclusions: **1) The proxy image can provide richer information.** Since the proxy image contains semantic editing information as well as some information from the query image, using the proxy image can offer more comprehensive information. In Fig.1, compared to the baseline retrieval results ①, ② improves retrieval precision at $k = 10, 25, 50$, although it slightly reduces precision at $k = 5$. In contrast, directly using semantic perturbation or query features leads to an overall decline in retrieval precision. It can also be observed in Fig.2 that using only the query image (Qi) results in **an incorrect emphasis on background, even directly retrieving the original image itself**. While the second row shows that using only semantic perturbation (Si) **ignores details in the image such as angle**. In contrast, the proxy image achieves relatively better retrieval results because it provides background and angle information. **2) All the features are im-**

portant. In Fig.1, compared to ② and ③, when semantic perturbation is applied to enhance the text-driven editing information in proxy or query image features (as shown in ⑤ and ⑥), retrieval precision improves, highlighting the importance of semantic perturbation for robust features. Besides, the comparisons between ⑤ and ⑦, as well as ⑥ and ⑦, demonstrate that adding either proxy image features or query image features can further enhance the effectiveness of robust features. The second column in Fig.5 indicates that combining query image features can better preserve certain textual patterns and textures that are challenging for the proxy image to generate accurately.

A.2. The analysis of hyperparameters.

Since object content, associated text formats, and the alignment between text and target images vary across datasets, different datasets require different weighting parameters λ during retrieval. In experiments, we find that the suitable λ for CIRCO, CIRR, and FashionIQ are around 0.3, 0.0, and 0.8. Additionally, the quality of proxies constructed for different datasets may also vary. For example, we observed that in FashionIQ, **the constructed proxies struggle to fully describe corresponding text for logo patterns and face challenges in generating pure white backgrounds during the generation process**. As a result, the role of proxy images in FashionIQ is relatively weaker, necessitating a larger λ . Moreover, if different backbones are used, the extracted features may emphasize different attributes, requiring dataset-specific adjustments to the λ parameter during retrieval. **In practical applications, users have diverse retrieval goals. Therefore, by adjusting different parameters, they can control the retrieval process according to their preferences.** Users can refine robust features by adjusting feature weights to emphasize specific retrieval aspects. For example, increasing the weight of the query image enhances details like logos or highlights attributes such as angle or style. Conversely, emphasizing proxy images or semantic perturbations shifts focus toward text-based editing directions.

B. More Qualitative Results

In this section, we show more results on three datasets. We present the Query image, relative caption, one of our generated Proxy images, as well as the top-1 baseline retrieval result and top-1 retrieval result of our method.

B.1. More Qualitative results on CIRCO.

We show qualitative results on CIRCO in Fig.3. We can draw the following conclusions: 1) The generated proxy images demonstrate certain detailed features, such as the yel-



Figure 2. **Ablation result on the Robust Proxy.** We present the visualization result of using different compositions (**Qi** represents only using query image, **Si** represents only semantic perturbation, and **Pi** represents only using the proxy image) of features in the robust proxy.

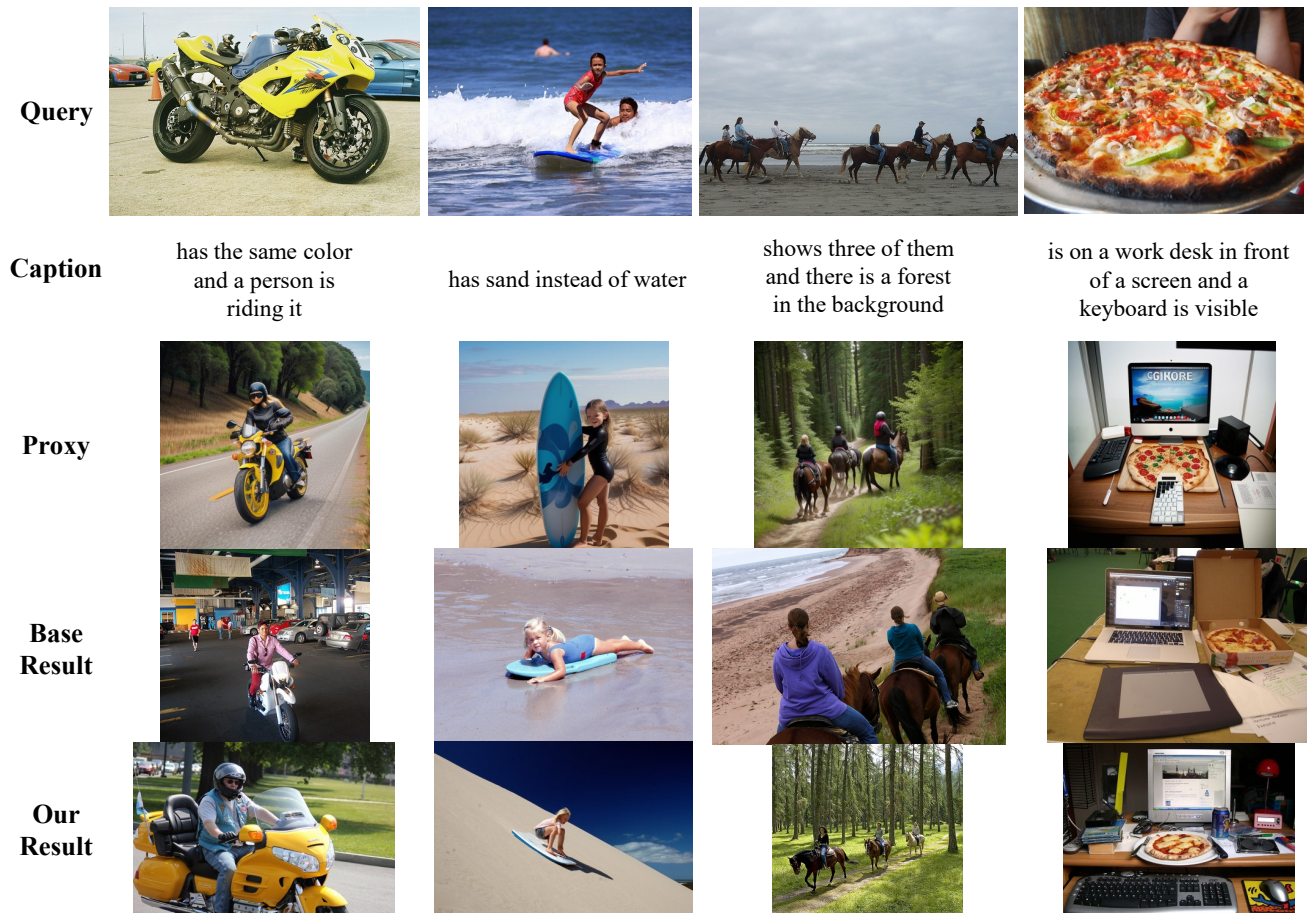


Figure 3. **Qualitative results on CIRCO dataset.** We show more improvement in top-1 retrieval results in the CIRCO dataset.

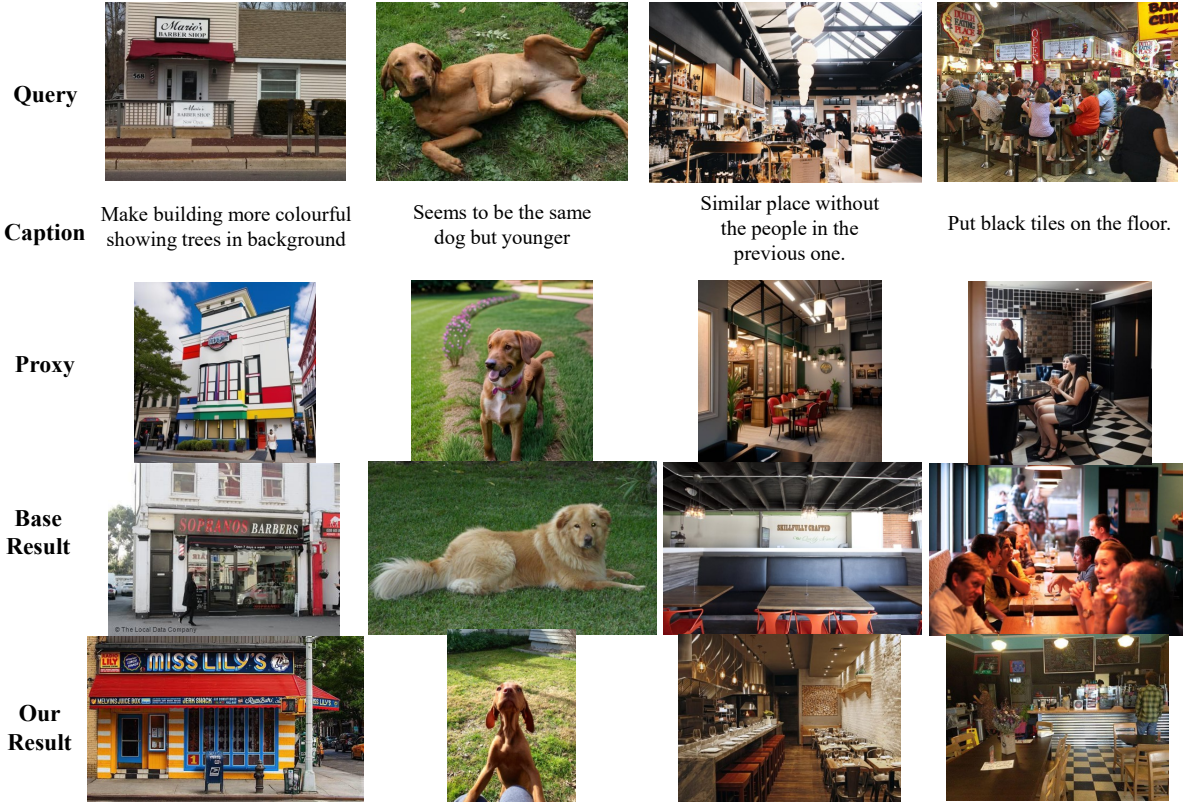


Figure 4. **Qualitative results on CIRR dataset.** We show more improvement in top-1 retrieval results in the CIRR dataset.

low car in the first column, the desert and blue skateboard in the second column, the forest background and three horses in the third column, and the screen and keyboard in the last column. 2) Our method improves the Top-1 retrieval results to better meet the requirements to some extent. For example, in the first column, the car we retrieved is yellow. In the second column, the background of our result has sand instead of water. In the third column, with a proxy to imagine the forest scene, we successfully retrieved three horses in a forest. In the last column, the retrieved result closely matches the features presented by the proxy and better fits the described spatial characteristics.

B.2. More Qualitative results on CIRR.

We show more qualitative results on CIRR in Fig.4 with the improvements in TOP-1 retrieval performance. As shown, our generated proxy images provide elements such as the target’s background ambiance and scene (e.g., more colorful background and black tiles), as well as the type and fine-grained attributes of the main objects (e.g., the same type of dog). These enhancements contribute to improved retrieval accuracy.

B.3. More Qualitative results on FashionIQ.

We show more qualitative results on FashionIQ in Fig.5 with the improvements in TOP-1 retrieval performance. FashionIQ provides text descriptions of attributes such as color and patterns, enabling our proxy features to achieve relatively better retrieval results. At the same time, we believe that query image features are also important for the FashionIQ dataset. For example, in the second column, although the baseline can retrieve clothes with pink color and black text, incorporating original image features allows the retrieval of images with logos that are more similar to the query while also aligning with the text description.

C. Limitation

Additional time overhead. While our method is plug-and-play in most scenarios and improves retrieval accuracy, it does introduce some time overhead. The process of layout generation and image generation adds certain time costs.

Sensitive to hyperparameters. The image-based retrieval enhancement is influenced by the performance of the constructed proxy images and the trade-off parameters used. Users need to carefully set reasonable weighting hyperparameters for the retrieval process. Thus, how to design a more reasonable method for combining weights or metrics





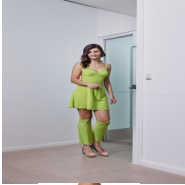

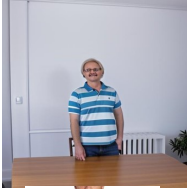









Query				
Caption	is a lime green	The shirt is pink in color with black writing.	is lighter blue with white stripes	its a greenish long dress
Proxy				
Base Result				
Our Result				

Figure 5. **Qualitative results on FashionIQ dataset.** We show more improvement in top-1 retrieval results in the FashionIQ dataset.

to reduce sensitivity to parameters is an important direction for future research.