Multi-Schema Proximity Network for Composed Image Retrieval

Supplementary Material

A. Overview

In this supplementary material, we first provide a detailed explanation of the Sinkhorn-Knopp algorithm used for solving the optimal transport plan. Next, we describe the datasets used in our experiments. Afterward, we analyze the effect of different hyperparameters values on performance. Following this, we present additional qualitative results to illustrate the effectiveness of our method. Lastly, we discuss the broader impacts of our work, emphasizing its practical applications and ethical considerations for responsible deployment.

A.1. Detailed Explanation of the Sinkhorn-Knopp Algorithm for Solving P*

To solve the optimal transport plan P^* in Eq. (I), we employ the iterative Sinkhorn-Knopp algorithm to efficiently minimize the objective while satisfying the constraints, as shown in Algorithm 1. Specifically, the optimization problem is defined as:

$$\begin{aligned} P^* = \min_P \langle P, -\log(C) \rangle + \tau KL\left(P \| \alpha\beta^T\right), \\ \text{s.t. } P\mathbb{1}_{N_B} = \mathbb{1}_{N_B} \cdot \frac{1}{N_B}, \\ P^T\mathbb{1}_{N_Q} = \mathbb{1}_{N_Q} \cdot \frac{1}{N_Q}. \end{aligned} \tag{I}$$

The algorithm initializes the kernel matrix as $K = \exp(-C/\sigma)$. The iterative updates are performed as:

$$u \leftarrow \frac{\alpha}{Kv}, \quad v \leftarrow \frac{\beta}{K^T u}, \tag{II}$$

where initialize u with $\mathbb{1}_{N_B}$ and v with $\mathbb{1}_{N_Q}$. After convergence, the transport plan is computed as:

$$P^* = \text{diag}(u)\, K\, \text{diag}(v), \tag{III}$$

The iterative Sinkhorn-Knopp algorithm ensures computational efficiency, satisfies the constraints, and incorporates entropy regularization, obtaining robust learnable queries. as shown in Algorithm 1

A.2. The Detailed Introductions of Datasets

CIRR is a real-life CIR dataset sampled from the dataset $NLVR^2$, which consists of 221, 552 real-life images with 36, 554 triplets. The dataset is grouped in multiple subsets of six semantically and visually similar images, randomly split into 80%, 10%, and 10% for training, validation, and test sets, respectively. **FashionIQ** is a dataset that focuses on the fashion domain, which contains three classes: Dress, Shirt, and Toptee. It is composed of 77, 684 fashion images with 30, 134 triplets, each triplet consists of a reference image, a pair of relative captions, and a target image. For the training set, there are 46, 609 fashion images forming 18, 000 training triplets. The validation and test sets consist of 15, 537 and 15, 538 images with 6, 017 and 6, 119 triplets, respectively. **LaSCo** is a large-scale challenging CIR dataset, labeled on COCO images based on VQA2.0, which contains an open and broad domain of natural images and rich text. It consists of approximately 121, 479 natural images with 389, 305 triplets, $\times 10$ larger than the CIRR dataset.

Algorithm 1 Multi-Schema Interaction

Input: Initialized model parameters W, query features F^q, target features F^t and transport cost C.
Output: Optimized model parameters W

```
for epoch=1:max_epoch do
    for batch=1:max_batch do
        Initialize α^0 with (1/N_B) · 𝟙 and β^0 with (1/N_Q) · 𝟙.
        while ||α^k − α^{k−1}||_1 < ϵ do
            ∀i : α_i^k ← [(C)β^{k−1}]_i^{−1},
            ∀j : β_j^k ← [α^{k−1^T}(C)]_j^{−1}.
        end while
        P* = diag(α)(C)diag(β).
        p_i = argmax(P_i^*),
        L_MSI = (1/N_B) Σ_{i=1}^{N_B} ((1/2) D(F_{p_i}^Q, sg(F_i^T))
        Optimize the model parameters W by L_MSI
        end for
end for
```

	Recall@K			
γ	**K=1**	**K=5**	**K=10**	**K=50**
0.3	56.26	85.08	91.39	98.11
0.4	<u>56.49</u>	84.96	**91.51**	98.11
0.5	**56.82**	84.93	91.46	**98.25**
0.6	56.35	<u>85.19</u>	<u>91.49</u>	<u>98.13</u>
0.7	56.27	**85.36**	91.28	98.09

Table I. **Detailed comparison** of relaxed proximity loss in our method in terms of recalls with regards to different values of γ on the **CIRR** validation dataset. The best and second-best results are marked in **bold** and <u>underlined</u>, respectively.

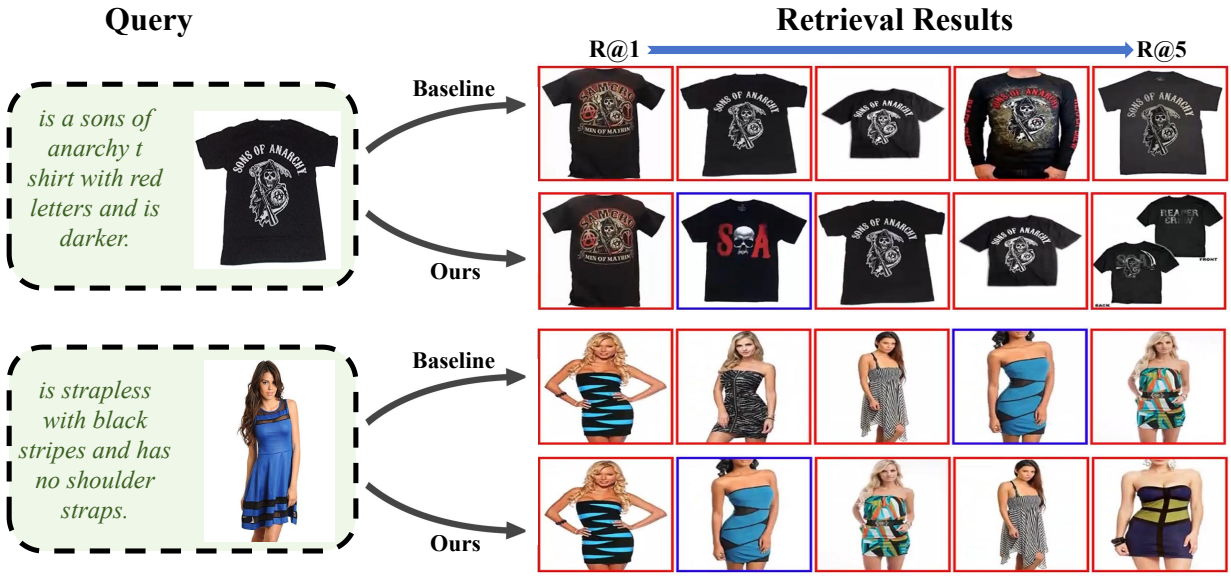

Figure I. Comparison of Recall@5 retrieval results between Baseline and our methods on the Fashion-IQ dataset. The blue border denotes the true retrieval result and the red border represents the false retrieval result.

	Dress		Shirt		Toptee		Average		
γ	**R@10**	**R@50**	**R@10**	**R@50**	**R@10**	**R@50**	**R@10**	**R@50**	**Avg.**
0.3	49.67	73.08	55.25	74.04	57.62	77.97	54.18	75.03	64.61
0.4	49.87	71.69	55.20	73.34	57.57	<u>78.83</u>	54.21	74.62	64.42
0.5	**51.17**	**74.12**	**56.37**	**75.17**	**59.56**	**79.30**	**55.70**	**76.20**	**65.95**
0.6	49.68	<u>73.57</u>	55.64	<u>74.88</u>	<u>57.62</u>	78.53	<u>54.31</u>	<u>75.66</u>	<u>64.99</u>
0.7	<u>49.92</u>	72.78	<u>55.74</u>	74.49	56.81	78.23	54.16	75.17	64.67

Table II. **Detailed comparison** of alignment loss in our method in terms of average recalls with regards to different values of γ on the **Fashion-IQ** dataset. The best and second-best results are marked in **bold** and <u>underlined</u>, respectively.

A.3. Detailed Analysis of Hyperparameters.

In Tables I and II, we analyze the impact of different γ values on the performance of our method across the **CIRR** and **Fashion-IQ** datasets. On the CIRR dataset, our method achieves the best **Recall@1** value of **56.82** when $\gamma = 0.5$, indicating that this setting effectively balances the attracting and repelling terms in RPLoss to optimize the alignment of positive pairs while mitigating the influence of noisy negatives. Similarly, on the Fashion-IQ dataset, our method achieves the highest average recall (**Avg. Recall = 65.95**) when $\gamma = 0.5$, outperforming other γ values. The method also secures the best results for multiple categories, such as **Shirt** (**R@10 = 56.37**) and **Toptee** (**R@10 = 59.56**), further highlighting the effectiveness of $\gamma = 0.5$ in balancing positive and negative pair alignments. Overall, the results suggest that $\gamma = 0.5$ is the optimal setting for both datasets, consistently ensuring robust and competitive recall performance by effectively addressing the challenges of noisy negatives while maintaining strong alignment between queries and targets. The impact of varying the number of learnable queries is not explored due to the need to load the BLIP2 pre-trained model.

A.4. Additional Qualitative Analysis.

We provide additional qualitative results on the **Fashion-IQ** dataset in Fig. I, illustrating the retrieval performance of our method compared to the baseline. Each row represents a query, which consists of a reference image and a relative caption describing the desired modification. The retrieval results are sorted by rank, with the blue border denoting the true match and the red borders representing false matches.

As shown in the first example, the query specifies a "sons of anarchy t-shirt with red letters and a darker appearance". While the baseline retrieves T-shirts that broadly align with the query's attributes, they lack the precise multi-schema interactions required for an accurate match, such as the specific darker color and red lettering. In contrast, our method successfully identifies the correct target (ranked at **R@2**)

and retrieves items that better adhere to the specified attributes, demonstrating the effectiveness of our method in understanding and addressing complex multi-schema interactions.

Overall, these examples reinforce the robustness of our method in handling complex queries and capturing multi-schema interactions, resulting in more accurate and relevant retrieval outcomes. The qualitative results align with the quantitative improvements reported in the main paper, further validating the effectiveness of our proposed approach.

A.5. Broader Impacts

This work was developed using publicly available datasets for Composed Image Retrieval (CIR) and aims to enhance the capabilities of multimodal retrieval. CIR has significant applications in domains such as e-commerce, digital content management, and information retrieval, enabling users to formulate more intuitive and effective queries by integrating textual and visual inputs. These advancements contribute to improving the user experience in various practical scenarios.

This work raises no ethical, safety, or environmental concerns. The datasets used are publicly available and widely adopted in academic research, ensuring compliance with ethical standards.