# CCIN: Compositional Conflict Identification and Neutralization for Composed Image Retrieval

Supplementary Material

## 1. More Experimental Details

In this section, we provide further details of the proposed CCIN in several aspects.

**Network Architecture.** The Adaptive Fusion Module is specifically designed to enable efficient multi-modal integration, which incorporates a Multi-Layer Perceptron (MLP) and a Sigmoid activation function to dynamically fuse diverse inputs. The detailed design of this module is presented in the left panel of Figure 8. In parallel, the Instruction-aware Q-Former extends the conventional Q-Former architecture by incorporating textual instructions as an additional input. This extension enables the module to focus on extracting task-specific features from images, which align with the semantic context provided by the instructions. The detailed architecture of this module is shown in the right panel of Figure 8.

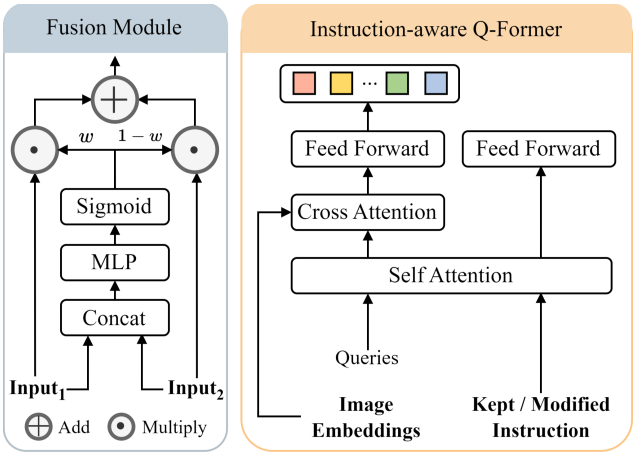


Figure 8. The architectural details of the Adaptive Fusion Module and the Instruction-aware Q-Former.

**Loss Function Details.** During training, we employ three distinct loss functions ($\mathcal{L}_{\text{ITC}}$, $\mathcal{L}_{\text{WRT}}$, and $\mathcal{L}_{\text{OPR}}$) to optimize the model and improve CIR performance.
*Contrastive Learning Loss* $\mathcal{L}_{\text{ITC}}$:

$$\mathcal{L}_{\text{ITC}} = -\frac{1}{B}\sum_{i}^{B}\log\frac{\exp\left(\boldsymbol{f}_{\text{query}}\boldsymbol{f}_{\text{tar}}^{\top}\right)}{\sum_{j\in\mathcal{B}}\exp\left(\boldsymbol{f}_{\text{query}}\boldsymbol{f}_{\text{tar}}^{\top}\right)}, \tag{9}$$

where $\mathcal{B}$ denotes the batch size, while $f_{\text{query}}$ and $f_{\text{tar}}$ represent the composed query representation and the target image representation, respectively.

*Weighted Regularization Triplet Loss* $\mathcal{L}_{\text{WRT}}$:

$$w_{ij}^{p} = \frac{\exp(d_{ij}^{p})}{\sum_{d_{ij}^{p}\in\mathcal{P}_i}\exp(d_{ij}^{p})}, w_{ik}^{n} = \frac{\exp(-d_{ik}^{n})}{\sum_{d_{ik}^{n}\in\mathcal{N}_i}\exp(-d_{ik}^{n})}, \tag{10}$$

$$\mathcal{L}_{\text{WRT}}(i) = \log(1+\exp(\sum_{j} w_{ij}^{p}d_{ij}^{p} - \sum_{k} w_{ik}^{n}d_{ik}^{n})), \tag{11}$$

where $(i, j, k)$ represents a hard triplet within each training batch. $p_i$ and $n_i$ denotes the positive and negative set for anchor $i$, respectively. $d_{ij}^{p}/d_{ik}^{n}$ denotes the distance between a positive/negative sample pair.
*Orthogonal Projection Regularization Loss* $\mathcal{L}_{\text{OPR}}$:

$$\tilde{\boldsymbol{f}}_{\text{con}} = W_1(\boldsymbol{f}_{\text{con}}), \tilde{\boldsymbol{f}}_{\text{tar}} = W_2(\boldsymbol{f}_{\text{tar}}), \tag{12}$$

$$\mathcal{L}_{\text{OPR}} = \alpha\cdot\frac{\|W_1W_2^{T}\|_F}{\|W_1\|_F\cdot\|W_2\|_F} + \beta\cdot\frac{\|\tilde{\boldsymbol{f}}_{\text{con}}\tilde{\boldsymbol{f}}_{\text{tar}}^{T}\|_F}{\|\tilde{\boldsymbol{f}}_{\text{con}}\|_F\cdot\|\tilde{\boldsymbol{f}}_{\text{tar}}\|_F}, \tag{13}$$

where $f_{\text{con}}$ and $f_{\text{tar}}$ donate the conflicting attribute feature and the target image feature, while $\|\cdot\|_F$ represents the Frobenius norm, respectively. Regularizing the projection matrix rather than the original features helps preserve feature information and prevent overfitting.

**Dataset Details.** In the experimental setup, we utilize three benchmark datasets: FashionIQ [48], CIRR [27], and Shoes [16]. The statistics details of these datasets, including the number of triplets, images, and their respective domains, are summarized in Table 7.

| Dataset | Triplets | Images | Domain |
|---|---|---|---|
| FashionIQ [48] | 30,134 | 77,684 | Fashion |
| CIRR [27] | 36,554 | 21,552 | Natural |
| Shoes [16] | 10,751 | 14,764 | Fashion |

Table 7. Existing Composed Image Retrieval (CIR) datasets, FashionIQ [48], CIRR [27], and Shoes [16], respectively.

**Training Details.** Following [2], the input images are preprocessed to a fixed resolution of $224 \times 224$. And the batch size is consistently set to 64 for experiments on FashionIQ [48], CIRR [27], and Shoes [16] datasets.

## 2. More Visualization Results

In this section, we provide more visualization results of the proposed CCIN in several aspects.

**Retrieval Results.** To further evaluate the efficacy of CCIN, we perform additional comparative experiments against SPRC [2], which lacks mechanisms for addressing

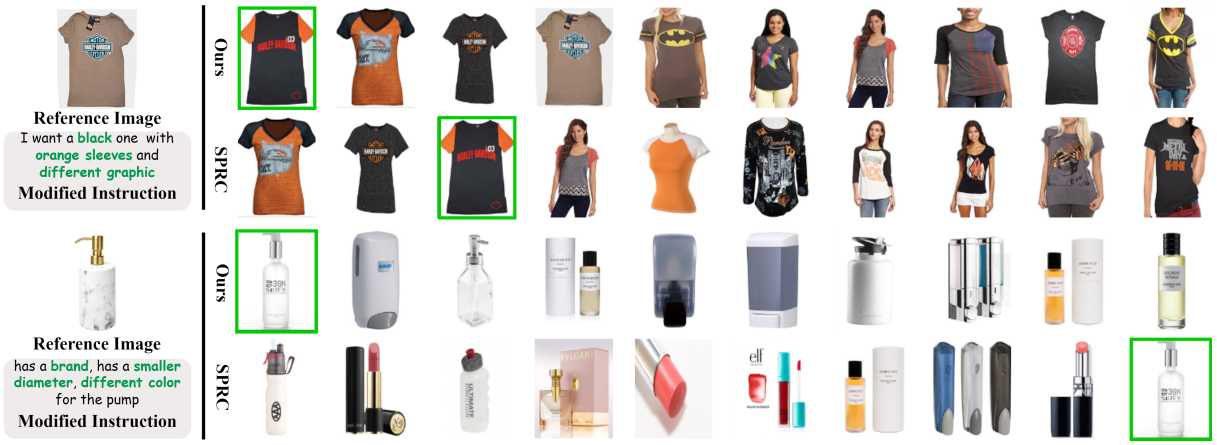


Figure 9. More qualitative results comparing SPRC [2] with the proposed method, with respect to Recall@1 metric on FashionIQ [48] and CIRR [27] datasets. The reference image and modified instruction are displayed on the left, while the retrieval rank is presented on the right. Correct matches are highlighted within green rectangles, respectively.


Figure 10. More qualitative results of CCI on three datasets [16, 27, 48]. Specifically, $T_{ref}$ and $T_{mod}$ denote the reference image caption and the modified instruction. The conflict identification results are highlighted in purple, respectively.

compositional conflicts. More visualization results on FashionIQ [48] and CIRR [27] are presented in Figure 9.

**Conflict Identification Results.** To demonstrate the effectiveness of CCI, we provide additional visualizations of conflict identification results from the FashionIQ [48], CIRR [27], and Shoes [16] datasets, as illustrated in Figure 10.

| Type | FashionIQ | CIRR | Shoes | Type | FashionIQ | CIRR | Shoes |
|---|---|---|---|---|---|---|---|
| | Mean | Mean | Mean | | Mean | Mean | Mean |
| Cos | 64.04 | 81.02 | 59.26 | Direct | 62.90 | 79.20 | 59.01 |
| Ours | **64.59** | **81.66** | **59.42** | Ours | **64.59** | **81.66** | **59.42** |

Table 8. Left: Comparison with using cosine similarity loss (Cos) to regularize original feature distances. Right: Comparison with directly (Direct) using LLMs to formulate target instruction (reference caption + modified instruction).