Zero-Shot Composed Image Retrieval via Dual-Stream Instruction-Aware Distillation

Supplementary Material

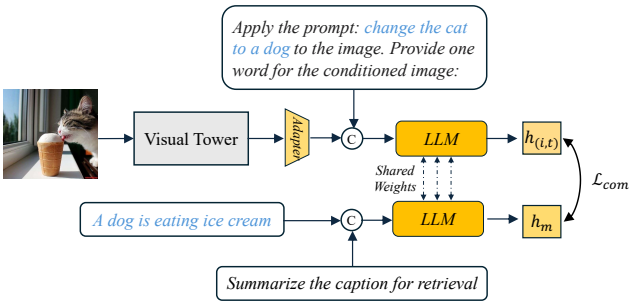

Figure 6. **Training Process of the Teacher MLLM.** Images are first processed by the visual tower and projected into the LLM embedding space. The textual modification, appended to a prompt template, is then combined with the projected image embeddings and fed into the LLM to generate the composed embedding $h_{(i,t)}$. For the modified caption, we add it to a separate prompt and again feed it into the LLM to produce h_m. The composed retrieval loss $\mathcal{L}_{com}$ is computed on $h_{(i,t)}$ and h_m to optimize the model.

A. Details of Data Generation

Existing retrieval models are largely trained on image–text pairs, limiting their direct applicability to CIR due to task discrepancies. In this paper, we extend existing paired datasets into triplets by leveraging knowledge from a powerful LLM. Given an image–text pair (i, c), where i is the image and c is the caption, we prompt the LLM with c using a specific template (see Figure 7). This prompt guides the LLM to derive a textual modification and a modified caption step by step. First, the LLM identifies changeable objects in the caption. Next, it proposes a modification for one of these objects. Finally, it explains how the modification affects the caption and produces a modified version. If no changeable object is detected, we instruct the LLM to add additional content to the caption instead. Dynamic examples, randomly selected from a curated collection, help the LLM better understand the task and promote more diverse generation outcomes.

B. Training of Teacher MLLMs

To further distill the instruction-following capability from LLMs, we propose the feature distillation, where we train a teacher MLLM to generate embeddings. Because MLLMs are inherently designed for text generation and cannot be used directly for retrieval, we adopt the "last token embedding" strategy similar to [20, 58] to transform the MLLM's function to usable retrieval representations.

The training procedure is illustrated in Figure 6. The teacher MLLM consists of three main components: the visual encoder f_V^M, the adapter f_A^M, and the LLM f_{LLM}^M. Given a triplet (i, t, m)—where i is the source image, t is the textual modification, and m is the modified caption—the visual encoder first converts the image into patch embeddings. These embeddings are then projected into the LLM's embedding space by the adapter (Equation 14). Next, the textual modification is tokenized within a prompt template, and those tokens are concatenated with the patch embeddings into a single input sequence. An [EOS] token is appended at the end, whose output embedding is taken as the overall representation $h_{(i,t)}^M$ (Equation 15). Finally, the modified caption is combined with a summary prompt and fed directly into the LLM (Equation 16) to produce its embedding h_m^M.

$$h_i^M = f_A^M(f_V^M(h_i)), h_i^M \in \mathbb{R}^{L^2 \times D_M} \tag{14}$$

$$h_{(i,t)}^M = f_{LLM}^M(\text{concat}(h_i^M, t, [\text{EOS}])), h_{(i,t)}^M \in \mathbb{R}^{D_M} \tag{15}$$

$$h_m^M = f_{LLM}^M(\text{concat}(m, [\text{EOS}])), h_m^M \in \mathbb{R}^{D_M} \tag{16}$$

Finally, we calculate the composed loss $\mathcal{L}_{com}$ using $h_{(i,t)}^M$ and h_m^M to optimize the teacher MLLM. After training, the modified caption embeddings h_m^M derived from the teacher MLLM are used to compute $\mathcal{L}_{fea}$.

C. Computation of Attention Map

Equations 17 to 21 show the detailed computation process of the attention map.

$$P = f_V(i), P \in \mathbb{R}^{L^2 \times D_V} \tag{17}$$

$$P' = f_T(f_\phi(P)), P' \in \mathbb{R}^{L^2 \times D_T} \tag{18}$$

$$A = P' \times h_{(i,t)}^T, A \in \mathbb{R}^{L^2} \tag{19}$$

$$A = \text{resize}(A), A \in \mathbb{R}^{L \times L} \tag{20}$$

$$A' = \text{interpolate}(A), A \in \mathbb{R}^{H \times W} \tag{21}$$

D. Hyperparameters of Training DistillCIR

Detailed training hyperparameters of the ViT-L version are shown in Table 7. For the ViT-B and other MLP-layer versions, we train corresponding Pic2Word f_ϕ as initialization.

You are helping to create a multi-modal dataset for Composed Image Retrieval (CIR). The dataset requires pairs of source and target image captions, along with a single, concise instruction describing how the source image is transformed into the target image.

Task
1. Input: A source image caption will be provided.
2. Brainstorming: Identify the key elements in the source caption (objects, actions, setting) and propose one significant, plausible modification.
3. Modification Instruction: Write a clear, succinct directive describing the intended change.
4. Modified Caption: Apply the change to the original caption, preserving all other details.

Output Requirements
Output exactly three items in this order:
1. Brainstorming – Briefly explain the original elements and the proposed change.
2. Modification Instruction – A short statement of the exact change to be made.
3. Modified Caption – The new caption reflecting the transformation.

Important
1. The modification instruction must focus on a single, significant change (e.g., changing an object's color, location, or action).
2. The modified caption must only incorporate this one change and remain otherwise consistent with the original caption.
3. The instruction and modified caption should be coherent and plausible.

Example:
…

Input: {original caption}

Figure 7. Prompt Template to process the image-caption data.

Training Config	**Value**
Visual Tower	CLIP ViT
Text Tower	CLIP Text Encoder
LoRA R	64
LoRA Alpha	16
Precision	FP16
Training Epochs	2
Batch Size	768
Learning Rate	2×10^{-5}
LR Scheduler Type	Cosine

Table 7. Hyperparameters of DistillCIR.