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

Wenliang Zhong[1], Rob Barton[2], Weizhi An[1], Feng Jiang[1], Hehuan Ma[1], Yuzhi Guo[1], Abhishek Dan[2], Shioulin Sam[2], Karim Bouyarmane[2], and Junzhou Huang[1]

[1]The University of Texas at Arlington, [2]Amazon

{wxz9204, weizhi.an, fxj8843, hehuan.ma, yuzhi.guo}@mavs.uta.edu

{rab, abhidan, shioulin, bouykari}@amazon.com

jzhuang@uta.edu

Abstract

Composed Image Retrieval (CIR) targets the retrieval of images conditioned on a reference image and a textual modification, but constructing labeled triplets (reference image, textual modification, target image) is inherently challenging. Existing Zero-Shot CIR (ZS-CIR) approaches often rely on well-aligned vision-language models (VLMs) to combine visual and textual inputs, or use large language models (LLMs) for richer modification understanding. While LLM-based methods excel in capturing textual details, they are computationally costly, slow to infer, and often restricted by proprietary constraints. In this paper, we argue that the superior performance of LLM-based ZS-CIR methods primarily stems from their capacity to follow instructions, an aspect largely missing in more efficient projection-based models built upon VLMs. To bridge this gap, we introduce DistillCIR, a dual-stream distillation framework that transfers LLMs' instruction-following capability into compact, projection-based architectures. By synthesizing triplet data with an LLM and incorporating a novel reasoning process, DistillCIR learns both composed retrieval and instruction awareness. In addition, we train an open-source multimodal LLM on the generated data, and further distill its instruction-aware embeddings into the projection-based model. Without any reliance on LLMs at inference, DistillCIR significantly surpasses state-of-the-art ZS-CIR methods in both performance and efficiency, offering a promising direction for instruction-aware, lightweight CIR. Project page is at: https://distillcir.github.io/.

1. Introduction

Composed Image Retrieval (CIR) is a retrieval task that refers to finding target images with a composition of reference images and a textual modification requirement [2, 12, 36]. Such a task is increasingly attractive in real-world scenarios, such as e-commerce platforms where users might want to find similar products with specific alterations [5, 63]. However, CIR often faces significant data acquisition challenges, as collecting training samples typically requires substantial human effort. In response, recent works explore Zero-Shot CIR (ZS-CIR) [4, 42], where models are pretrained on specific datasets and then applied directly to downstream tasks without additional tuning.

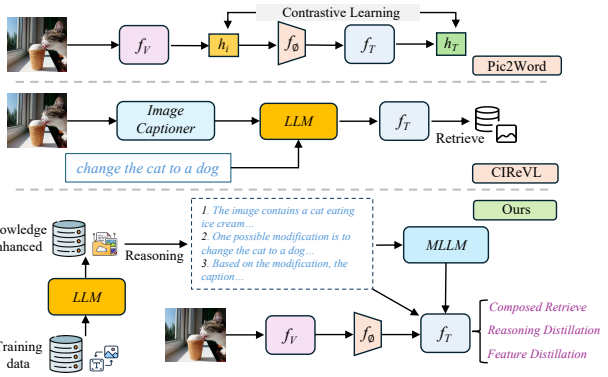

Figure 1. **Comparison with DistillCIR and Existing Methods**. Pic2Word relies solely on CLIP for both training and inference, offering efficiency but limited comprehension of textual modifications. In contrast, CIReVL employs an LLM to parse instructions during inference, making the process slower and less flexible. DistillCIR bridges this gap by distilling the LLM's instruction-following capabilities into a CLIP-based CIR model, achieving both superior performance and efficient inference.

Most ZS-CIR methods are projection-based models [2, 12, 36] founded on well-aligned Vision-Language Models (VLMs) such as CLIP [35]. Specifically, they generate

composed embeddings by projecting the image embedding from the VLM into the text embedding space and combining it with the textual modification for the output embedding. Despite their staged progress, concerns remain about whether they truly understand the textual modification or just simply combine multimodal information [48]. More recently, some methods [6, 19, 21, 22, 27] explore the involvement of Large Language Models (LLMs) [58] to enhance models' comprehension of modifications. These methods either process the textual modification and reference images with LLMs or directly leverage LLMs for embedding generation. Though qualitative experiments are impressive, these models are expensive to deploy and slow in inference due to the significantly larger number of parameters in LLMs. Worse still, some methods [21] inherently rely on proprietary LLMs in downstream tasks, which may be prohibited in commercial scenarios due to privacy concerns.

Witnessing the rise of LLM-based ZS-CIR models and their shortcomings, a question naturally arises: what internal capabilities make these LLM-based models more powerful than most projection-based models? In this paper, we argue that projection-based models based on pretrained VLMs excel at multimodal retrieval; however, they struggle to understand textual modifications. It is the instruction-following capability [47, 61] within LLMs that makes LLM-based models outperform projection-based models. Motivated by this argument, we propose to distill the instruction awareness of LLMs into projection-based models, making them small in size yet powerful in performance.

Nevertheless, distillation from LLMs to CIR models based on CLIP is challenging, primarily because of the inherent task difference: LLMs are trained to generate content, while CLIP is designed for multimodal representation alignment. To address this challenge, we propose a dual-stream distillation strategy. Our intuition is to leverage a powerful LLM [1, 6] to enhance existing data with its internal knowledge to train CIR models. Specifically, we prompt the LLM to extend an image-caption pair to a triplet similar to the CIR format, which contains an image, a textual modification, and a modified caption. While the triplet data can teach the projection-based model the basics of composed retrieval, it does not fully reflect the instruction-following capability of LLMs. To elicit instruction awareness, we add a reasoning process [49] during generation, where the LLM explicitly describes how it derives the modification and how the modification affects the final caption. A novel reasoning loss is then proposed to guide the projection-based model in learning from this detailed generation process. Besides learning from reasoning, it is also crucial to learn instruction-aware embeddings. To achieve this, we propose training an open-source Multimodal LLM (MLLM) [28, 29] using the triplet data and distilling its instruction-following capability into the projection-based model through another feature distillation loss. We term our dual-stream distilled model DistillCIR, which not only achieves superior performance compared to existing LLM-based ZS-CIR models but also remains efficient in model size. Furthermore, DistillCIR demonstrates extraordinary instruction-following capabilities without requiring any LLMs during inference.

To summarize, our contributions are threefold: (1) We propose a novel insight for enhancing projection-based ZS-CIR models through instruction awareness, which originally resides in LLMs but missing in retrieval VLMs. (2) We introduce DistillCIR, a novel method that learns the instruction-following capability from LLMs through dual-stream distillation. (3) Extensive experiments on four public datasets are conducted, and the results reflect the superiority of DistillCIR. Moreover, DistillCIR is instruction-aware without requiring any LLMs during inference.

2. Related Works

2.1. Composed Image Retrieval

Due to the scarcity of labeled data, recent research in Zero-Shot Composed Image Retrieval (ZS-CIR) has gained traction. Existing ZS-CIR methods generally fall into two categories: (1) **Projection-based Models** [2, 12, 36], which often build on established multimodal retrieval frameworks [25, 35]. These methods generate a composed embedding by projecting image embeddings into text space and combining them with modification embeddings [4, 36] or by jointly projecting both image and text embeddings into a shared interpolated space [18]. Although effective, they frequently struggle to interpret textual modifications [48]. (2) **LLM-based Models** [21, 27], which harness the capabilities of LLMs. Some leverage powerful LLMs at inference time to parse textual modifications [21], while others use LLMs primarily as text encoders [22, 27]. However, applying LLMs directly for downstream tasks may lead to slower, less efficient inference, and privacy issue in many commercial contexts. Although some approaches [19] sidestep this by using LLMs to synthesize training data, they generally rely on a specific data gallery, limiting broader ZS-CIR applicability. To address these challenges, we propose distilling the LLM's instruction-following ability into projection-based CIR models, thus enabling robust textual understanding without the computational overhead of LLM-based inference.

2.2. Vision-Language Model Distillation

Model distillation [11, 14, 34, 45] aims to transfer knowledge from a teacher model to a student model. Most existing approaches concentrate on feature-level alignment [8, 9, 33, 46, 51, 53], matching representations between teacher and student networks. With the rise of large language models (LLMs), recent work has also begun exploring how to

distill from a larger LLM into a smaller one [15, 32, 38, 47, 52]. One popular strategy prompts the teacher to synthesize data, which the student then uses for training to mimic the teacher's outputs. Beyond LLMs, some studies also investigate the distillation of retrieval models, but they primarily focus on feature alignment. Crucially, few endeavors address distilling instruction-awareness from LLMs into retrieval models. In this paper, we tackle this issue with a novel dual-stream distillation strategy.

3. Methodology

The overview of DistillCIR is shown in Figure 2. The subsequent sections provide details on the model architecture and the three training objectives: the composed loss, the reasoning distillation loss, and the feature distillation loss.

3.1. Model Architecture

DistillCIR uses a projection-based CIR model f based on the pretrained Pic2Word [35, 36], providing basic multimodal retrieval capability and efficiency. The image embedding h_i is obtained by passing the reference image i through the CLIP visual tower f_V (Equation 1). Subsequently, a projection module f_ϕ, consisting of a three-layer MLP with ReLU activations, maps the image embedding into the textual token embedding space (Equation 2). The textual modification t, represented by a sentence of instruction, is tokenized and combined with the projected image embedding. The composed input sequence is then fed to the CLIP text tower f_T to obtain the final composed embedding $h_{(i,t)}$ (Equation 3). D in all equations refers to the model embedding dimension and the subscript refers to the corresponding model. An overview of this composed embedding generation process is shown in Figure 3.

$$h_i = f_V(i), h_i \in \mathbb{R}^{D_V} \tag{1}$$

$$h'_i = f_\phi(h_i), h'_i \in \mathbb{R}^{D_T} \tag{2}$$

$$h_{(i,t)} = f_T(h'_i, t), h_{(i,t)} \in \mathbb{R}^{D_T} \tag{3}$$

In inference, all target images can be directly encoded by the visual tower f_V and cached in the database. Cosine similarity is then used to match the composed embedding with all target image embeddings.

3.2. Eliciting LLM's Knowledge via Data Extension

Obtaining large-scale triplet data (reference image, textual modification, target image) for CIR is challenging. In addition, the difference in tasks between LLMs and CIR models hinders the application of distillation techniques used for transferring larger language models to smaller ones in CIR. To address this challenge, we propose using publicly available image-text datasets and LLMs to derive triplet data with a format similar to CIR. This triplet data provides key benefits including knowledge enhancement from LLMs and training objectives akin to CIR.

Specifically, we leverage the CC3M [37] dataset, which is commonly used in existing ZS-CIR models for consistency. Given an image-caption pair (i, c) from the dataset, i and c represent the image and caption, respectively. We use a powerful frontier LLM in the class of Claude 3.5 Sonnet [3] or GPT-4o [1] to generate a triplet (i, t, m), where t is the textual modification, and m is the modified caption based on the image and modification. In detail, the caption c is fed to the LLM, which leverages its internal knowledge to brainstorm a suitable modification and generate the modified caption. By generating triplet data in a CIR-like format, we distill the LLM's internal knowledge for training CIR models. We then establish a composed training objective as in Equation 6, where we obtain the composed embedding h_{it} by forwarding (i, t) through the CIR model f and target embedding h_m by passing the modified caption through the text tower f_T.

$$h_{(i,t)} = f(i, t), h_{it} \in \mathbb{R}^{D_T} \tag{4}$$

$$h_m = f_T(m), h_m \in \mathbb{R}^{D_T} \tag{5}$$

$$\mathcal{L}_{com} = -\log \frac{\text{sim}(h_{(i,t)}, h_m)}{\text{sim}(h_{(i,t)}, h_m) + \sum_{m' \in \mathbb{N}} \text{sim}(h_{(i,t)}, h_{m'})} \tag{6}$$

$\text{sim}(\cdot, \cdot)$ represents the cosine similarity with temperature and $\mathbb{N}$ represents the negative set, which is from other samples in the same batch.

3.3. Distillation from Reasoning Enhancement

Directly generating triplets from pairs using LLMs poses some potential issues. The generated triplet may be undermined by hallucinations [17, 54] within the LLM. More critically, the triplet can only teach the CIR model what targets to retrieve rather than how to retrieve the target, which we believe is the essence of instruction awareness. To learn the instruction-following capability, the reasoning process [43, 49] of how a triplet is generated by the LLM should be detailed and taught to the CIR model.

To achieve this goal, the triplet generation should not be free-form. Instead, a step-by-step reasoning process [49] should be incorporated into the generation. In the generation prompt, we first ask the LLM to identify objects or environments in the caption that are possibly changeable. Then, we instruct the LLM to brainstorm a plausible modification for the identified term. For example, given the caption "A student is playing basketball," one possibly changeable object is basketball, and the corresponding modification is "changing the sport to football." The precise generation process ensures that the modification is feasible, thereby mitigating issues caused by hallucinations within the LLM. The LLM is subsequently prompted to reason

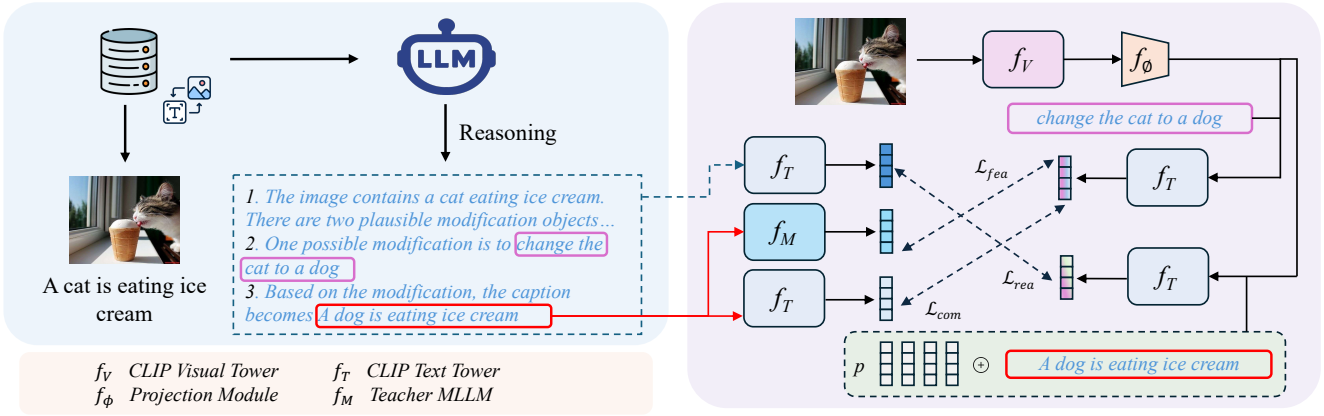

Figure 2. **Overview of DistillCIR**. We begin by augmenting the CC3M dataset via LLM-based reasoning, where the LLM produces plausible modifications and corresponding modified captions step by step, revealing how the modification is derived. We then train a projection-based CIR model on this LLM-enhanced data using three contrastive losses. The **composed loss** $\mathcal{L}_{com}$ forms the backbone of composed retrieval by mapping the reference image and its modification to the modified caption. The **reasoning distillation loss** $\mathcal{L}_{rea}$ contrasts the reference image and the modified caption against the LLM's reasoning process, employing learnable prompt tokens to differentiate it from the composed loss. Lastly, the **feature distillation loss** $\mathcal{L}_{fea}$ aligns the projection model's embeddings with those of a separately trained teacher MLLM. By combining these losses, our framework equips the projection model with strong instruction-following capabilities for composed image retrieval.

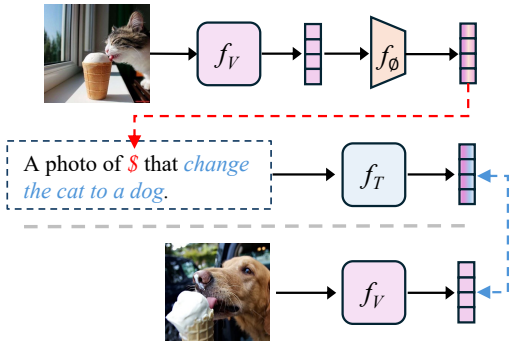

Figure 3. Model Architecture. The reference image is first processed by the visual tower f_V and then projected via f_ϕ. We then insert the projected image embedding into the text sequence by replacing the \$ token embedding. This combined input is fed into f_T, producing the composed embedding. At inference time, the target image is passed directly to f_V, and we measure the similarity between the composed embedding and the target embedding.

how the change will modify the original caption and ultimately derive a modified caption. Due to space limitations, the detailed prompt is shown in Appendix A. In a nutshell, the reasoning process consists of three parts: (1) identifying plausible terms to be changed; (2) brainstorming an applicable change for the identified terms; (3) analyzing how the modification affects the original caption and deriving the final modified caption. We denote the entire reasoning process as r, which will be used for our reasoning distillation to teach the CIR model instruction awareness.

To train the CIR model with the reasoning r, we propose a reasoning distillation training objective, which aims to align the image and modified caption (i, m) with its reasoning process. Specifically, the objective is a contrastive loss matching the consequential embedding $h_{(i,m)}$ and the reasoning embedding h_r as shown in Equation 9. However, in our experiments, we find that this loss can interfere with $\mathcal{L}_{com}$ because they share a similar format but differ in semantics. To distinguish it from $\mathcal{L}_{com}$, we introduce a set of learnable prompt tokens $p \in \mathbb{R}^{K \times D_T}$, where K is the number of tokens [23, 62]. These learnable tokens are prepended to m to capture specific reasoning semantics.

$$h_{(i,m)} = f(i, m, p), h_{(i,m)} \in \mathbb{R}^{D_T} \tag{7}$$

$$h_r = f_T(r), h_r \in \mathbb{R}^{D_T} \tag{8}$$

$$\mathcal{L}_{rea} = -\log \frac{\text{sim}(h_{(i,m)}, h_r)}{\text{sim}(h_{(i,m)}, h_r) + \sum_{r' \in \mathbb{N}} \text{sim}(h_{(i,m)}, h_{r'})} \tag{9}$$

3.4. Distillation from Feature Alignment

So far, $\mathcal{L}_{com}$ and $\mathcal{L}_{rea}$ learn the instruction-following capability at the semantic level. However, to generate instruction-aware embeddings, it is also important to learn at the feature level. Since most state-of-the-art LLMs are large or proprietary, we train an open-source Multimodal LLM (MLLM [26, 29]) f_M for this purpose.

Instruction-tuned MLLMs provide an aligned vision-language space. To distill knowledge at the feature level, one remaining challenge is extracting instruction-aware embeddings from the MLLM. Similar to [20, 44], we propose

	Method	# Params.	CIRR						CIRCO			
			$R@1$	$R@5$	$R@10$	$R_s@1$	$R_s@2$	$R_s@3$	$mAP@5$	$mAP@10$	$mAP@25$	$mAP@50$
ViT-B	SEARLE	165M	24.00	53.42	66.82	54.89	76.60	88.19	9.35	9.94	11.13	11.84
	Slerp	88M	28.19	55.88	68.77	61.13	80.63	90.68	9.34	10.26	11.65	12.33
	CIReVL	12.3B*	23.94	52.51	66.00	60.17	80.05	90.19	14.94	15.42	17.00	17.82
	DistillCIR	153M	**29.88**	**58.75**	**70.26**	**62.33**	**82.46**	**90.75**	**16.38**	**18.55**	**19.61**	**20.84**
ViT-L	Pic2Word [36]	429M	23.90	51.70	65.30	53.76	74.46	87.08	8.72	9.51	10.64	11.29
	SEARLE [4]	442M	24.24	52.48	66.29	53.76	75.01	88.19	11.68	12.73	14.33	15.12
	KEDs [39]	429M	26.40	54.80	67.20	-	-	-	-	-	-	-
	Context-I2W [40]	496M	25.60	55.10	68.50	-	-	-	-	-	-	-
	LinCIR [12]	442M	25.04	53.25	66.68	57.11	77.37	88.89	12.59	13.58	15.00	15.85
	Slerp [18]	428M	30.94	59.40	70.94	64.70	82.92	**92.31**	18.46	19.41	**21.43**	22.41
	CIReVL [21]	12.5B*	24.55	52.31	64.92	59.54	79.88	89.69	18.57	19.01	20.89	21.80
	MCL [27]	7.4B	26.22	56.84	-	61.45	-	-	17.67	18.86	20.80	21.68
	DistillCIR	429M	**32.34**	**63.58**	**74.29**	**66.75**	**84.02**	91.83	**20.77**	**21.52**	21.41	**23.38**

Table 1. **Experiment Results on the CIRCO and CIRR Test Sets**. **Bold** values indicate the highest scores. *CIReVL includes ChatGPT as a part of components for retrieval. Only parameters of components with known sizes are reported. Our model achieves significant performance gains over the baselines while retaining a compact model size.

training a commonly used MLLM (LLaVA [29]) as an embedding model. Specifically, we use the triplet (i, t, m) and $\mathcal{L}_{com}$ to train the model. To acquire embeddings, the [EOS] token is appended to the textual input, and its corresponding output embedding is used as the representation. In training, the image and textual modification are fed to the MLLM for the composed embedding, while the modified caption is fed to the MLLM for the modified target embedding. The MLLM is then optimized by $\mathcal{L}_{com}$. Detailed training procedures are described in Appendix B. We then use the trained MLLM as the teacher model for feature distillation. Specifically, the CIR model generates the composed embedding for (i, t), which is projected and used to match the modified caption embedding generated by the MLLM in Equation 12.

$$h'_{(i,t)} = \text{Linear}(h_{(i,t)}), h'_{(i,t)} \in \mathbb{R}^{D_M} \tag{10}$$

$$h^M_m = f_M(m), h^M_m \in \mathbb{R}^{D_M} \tag{11}$$

$$\mathcal{L}_{fea} = -\log \frac{\text{sim}(h'_{(i,m)}, h^M_m)}{\text{sim}(h'_{(i,m)}, h^M_m) + \sum_{m' \in \mathbb{N}} \text{sim}(h'_{(i,m)}, h^M_{m'})} \tag{12}$$

Finally, we combine three losses to train the model for the composed retrieval, reasoning distillation, and feature distillation. The combined loss is presented in Equation 13.

$$\mathcal{L} = \mathcal{L}_{com} + \alpha \cdot \mathcal{L}_{rea} + \beta \cdot \mathcal{L}_{fea} \tag{13}$$

The final loss ensures that the CIR model not only learns to perform composed retrieval using the LLM's knowledge but also acquires instruction-following capabilities through the reasoning and feature distillation processes.

4. Experiment

4.1. Training Settings

For data processing, we use the frontier LLM to process the CC3M dataset and save all distilled tuples (i, t, m, r) as described in Section 3, which will be used for the model training.

We experiment with CLIP [35] ViT-B and ViT-L as the backbones of DistillCIR. In alignment with common baselines, all data processing and model training are conducted on the CC3M [37] dataset. For the teacher MLLM, we use xtuner/llava-phi-3-mini-hf [10]. After training the MLLM, we extract and cached modified caption embeddings from it. The training is conducted on a cluster of six H100 GPUs, with a learning rate of 2×10^{-5} and the batch size is set to 768. For parameter efficiency, we train the entire Pic2Word projection module while applying LoRA [16] on linear layers within the CLIP backbone. Appendix D includes further details and other hyperparameters.

4.2. Benchmarks and Baselines

We conduct our experiments on four prominent zero-shot CIR benchmarks: **FashionIQ** [50], **CIRR** [31], **CIRCO** [4], and **GeneCIS** [41]. As one of the earlier datasets, FashionIQ concentrates on fashion e-commerce imagery. By contrast, CIRR and CIRCO revolve around broader, real-world scenes. While CIRR pioneered CIR research on natural images, its design of assigning only a single target image per query may lead to overlooked positives. CIRCO addresses this concern by enabling multiple ground-truth targets per query, offering a more robust measure of retrieval performance. Meanwhile, GeneCIS specializes in conditional retrieval through four types of attribute and object modifications. For evaluation, we follow established practice by reporting Recall@k $(R@k)$ for FashionIQ, CIRR, and GeneCIS, with an additional subset metric $(R_s@k)$ for CIRR. Because CIRCO can include multiple correct targets for a single query, we employ mean Average Precision $(mAP@k)$ to balance both precision and recall across various ranking positions.

We compare DistillCIR against two categories of state-

Method		Focus Attribute			Change Attribute			Focus Object			Change Object			Average
		R@1	R@2	R@3	R@1	R@2	R@3	R@1	R@2	R@3	R@1	R@2	R@3	R@1
ViT-B	SEARLE	18.90	30.60	41.20	13.00	23.80	33.70	12.20	23.00	**33.30**	13.60	23.80	33.30	14.40
	CIReVL	17.90	29.40	40.40	14.80	25.80	35.80	**14.60**	**24.30**	**33.30**	16.10	27.80	**37.60**	15.90
	DistillCIR	**20.01**	**30.99**	**42.50**	**15.15**	**26.89**	**35.85**	13.17	23.77	32.96	**16.67**	**28.19**	37.10	**16.25**
ViT-L	Pic2Word	15.65	28.16	38.65	13.87	24.67	33.05	8.42	18.01	25.77	6.68	15.05	24.03	11.15
	SEARLE	17.10	29.60	40.70	**16.30**	25.20	34.20	12.00	22.20	30.90	12.00	24.10	33.90	14.35
	LinCIR	16.90	29.95	41.45	16.19	**27.98**	36.84	8.27	17.40	26.22	7.40	15.71	25.00	12.19
	CIReVL	19.50	31.80	42.00	14.40	26.00	35.20	12.30	21.80	30.50	17.20	**28.90**	37.60	15.85
	DistillCIR	**20.65**	**32.55**	**44.18**	16.06	27.58	**37.66**	**13.28**	**24.35**	**33.60**	**17.34**	28.84	**37.65**	**16.83**

Table 2. **Experiment Results on the GeneCIS Test Set**. **Bold** indicates the highest score.

of-the-art methods: (1) Projection-based models such as Pic2Word [36], SEARLE [4], Context-I2W [40], KEDs [39], Slerp [18], and LinCIR [12]; (2) LLM-based models such as CIReVL [21] and MCL [27]. To ensure fair comparisons, we mainly consider baselines trained on CC3M and exclude those [13, 19, 24, 57] relying on different training corpora.

4.3. Result Analysis

Table 1 presents our results on the hidden test sets for both CIRCO and CIRR, obtained from their respective submission servers. On the CIRCO benchmark, a dataset noted for its comprehensive annotations and allowance of multiple valid targets per query, our method achieves an $mAP@5$ of 20.77% surpassing LinCIR and Slerp by 8.18% and 2.31%, respectively. It also provides a 2.20% boost over CIReVL, which relies on two MLLMs, BLIP-2 [26] and ChatGPT. Turning to the CIRR dataset, which often features minimal overlap between reference and target images, our approach excels in handling text-driven retrieval. We obtain a 5.94% improvement over KEDs and a 6.74% improvement over Context-I2W in $R@1$. Additionally, we exceed the LLM-based baseline MCL by 6.12% in $R@1$ and 5.30% in $R_s@1$. These results underscore the strength and adaptability of our framework in managing the nuanced challenges posed by both CIRCO's multi-target scenarios and CIRR's instruction-dominated retrieval tasks. Noteably, besides the impressive performance, DistillCIR also has a compact size comparable to most projection-based baselines. Theses findings validate the efficiency and superiority of DistillCIR.

Table 2 compares our results on GeneCIS, a dataset distinguished by condition statements that may convey opposing meanings, which necessitates a model capable of nuanced interpretation. Our DistillCIR framework outstrips Pic2Word and CIReVL by 5.68% and 0.98% in average $R@1$, respectively. The superiority in this benchmark highlight the robustness of DistillCIR and its comprehension capability in following instruction.

Table 3 compares our approach against existing zero-shot baselines on FashionIQ revealing substantial gains in average $R@10$. Specifically, our model surpasses Pic2Word and KEDs by 5.98% and 3.98%, respectively. Although our training primarily utilizes natural imagery, FashionIQ is highly specialized in fashion e-commerce, making our strong performance there indicative of the model's robust generalization capabilities. Effectively transferring knowledge from general-purpose data to more domain-specific tasks underscores the versatility of our method across both fashion-centric and natural-image CIR settings.

Figure 4. **Qualitative Comparison between DistillCIR and Pic2Word on the CIRCO (Top) and FashionIQ (Bottom) Validation Set**. Green boxes indicate ground truths.

Figure 4 presents qualitative comparisons from both CIRCO and FashionIQ, demonstrating the adaptability of our method across diverse tasks. In particular, we observe Pic2Word correctly captures alterations in color within FashionIQ but struggles with shifts in style. By contrast, our approach successfully handles both color and style modifications, showcasing its superior ability to interpret and apply textual guidance.

	Method	Shirt		Dress		Toptee	
		$R@10$	$R@50$	$R@10$	$R@50$	$R@10$	$R@50$
ViT-B	SEARLE	24.44	41.61	18.54	39.51	25.70	46.46
	Slerp	23.06	41.95	19.24	42.14	26.57	47.78
	CIReVL	28.36	47.84	**25.29**	**46.36**	31.21	**53.85**
	DistillCIR	**29.34**	**49.05**	25.14	45.76	**33.87**	53.06
ViT-L	Pic2Word	26.20	43.60	20.00	40.20	27.90	47.40
	SEARLE	26.89	45.58	20.48	43.13	29.32	49.97
	KEDs	28.90	48.00	21.70	43.80	29.90	51.90
	Context-I2W	29.70	48.60	23.10	45.30	30.60	52.90
	LinCIR	29.10	46.81	20.92	42.44	28.81	50.18
	Slerp	29.64	46.47	23.35	45.12	31.97	51.20
	CIReVL	29.49	47.40	24.79	44.76	31.36	53.65
	DistillCIR	**30.73**	**50.12**	**25.51**	**46.64**	**33.88**	**53.74**

Table 3. **Experiment Results on the FashionIQ Validation Set**. **Bold** indicates the highest score.

5. Ablation Study

In this section, we conduct an in-depth analysis of each part of DistillCIR and examine its instruction awareness learned from LLMs. Specifically, we aim to answer the following questions: (1) How does reasoning distillation affect the model performance, and what design choices should we make for the training objective? (2) What is the impact of feature distillation, and how do different MLLMs influence the results? (3) Does DistillCIR actually learn the instruction-following capability, and how should we evaluate it? We address these questions in each subsection. Unless otherwise specified, we use the ViT-L backbone and the validation sets in CIRR and CIRCO to conduct ablations.

5.1. Effect of Reasoning Distillation

While $\mathcal{L}_{com}$ focuses on composed retrieval, $\mathcal{L}_{rea}$ and $\mathcal{L}_{fea}$ further enhance instruction awareness. Unlike conventional distillation approaches, $\mathcal{L}_{rea}$ aims to enable CIR models to acquire instruction-following capabilities from LLMs at the semantic level by performing cross-task distillation from text generation to retrieval. This introduces a key challenge in designing the distillation loss, since the teacher and student models optimize fundamentally different objectives. We suggest two potential solutions: (1) aligning with the teacher (LLM) through text generation with a Negative Log Likelihood loss, or (2) aligning with the student (CIR) through retrieval via a contrastive loss. As shown in Cases 1-3 of Table 4, despite prior work [25, 26] suggesting that joint training for text generation and contrastive learning can be mutually beneficial, we observe that this approach actually underperforms compared to contrastive-only training. We attribute this discrepancy to differences in inputs and outputs: in contrast to the multitask setting described in [25] where both tasks share the same inputs and outputs, our $\mathcal{L}_{com}$ (composed retrieval) and $\mathcal{L}_{rea}$ (instruction reasoning) target distinct inputs and outputs, complicating training when the losses themselves also diverge.

Hence, we recommend using contrastive learning for $\mathcal{L}_{rea}$. However, Cases 3-4 shows that contrastive-only optimization does not adequately capture instruction reasoning. As discussed in Section 3, we attribute this to the semantic discrepancy between $\mathcal{L}_{com}$ and $\mathcal{L}_{rea}$; forcing them to share an identical loss complicates training. To address this, we propose a prompt-based approach. By introducing learnable prompt tokens into $\mathcal{L}_{rea}$, the model can better acquire instruction-awareness while maintaining a clear distinction from the basic composed retrieval task.

Cases	CIRR R@1	CIRCO mAP@10	GeneCIS R@1
1. $\mathcal{L}_{com}$ Only	30.38	19.75	16.04
2. + $\mathcal{L}_{rea}$ (NLL)	28.75	18.96	15.48
3. + $\mathcal{L}_{rea}$ (Contrast) w/o p	31.40	20.00	16.06
4. + $\mathcal{L}_{rea}$ (Contrast) w/ p	31.96	20.33	16.21

Table 4. **Ablations on the Reasoning Distillation**. $\mathcal{L}_{rea}$ (NLL) and $\mathcal{L}_{rea}$ (Contrast) refer to using the negative log likelihood for text generation and using the contrastive loss for retrieval as the reasoning distillation loss together with the composed loss, respectively. p are learnable prompt tokens.

Cases	CIRR R@1	CIRCO mAP@10	GeneCIS R@1
1. $\mathcal{L}_{com}$ Only	30.38	19.75	16.04
2. $\mathcal{L}_{com}$ + $\mathcal{L}_{fea}$ (TinyLLaVA)	30.62	19.78	16.02
3. $\mathcal{L}_{com}$ + $\mathcal{L}_{fea}$ (LLaVA-LLaMA3)	31.45	20.55	16.63
4. Teacher Model (TinyLLaVA)	31.93	21.42	16.95
5. Teacher Model (LLaVA-Phi3)	33.79	22.68	16.99
6. Teacher Model (LLaVA-LLaMA3)	34.17	23.20	17.15
7. $\mathcal{L}_{com}$ + $\mathcal{L}_{fea}$ (LLaVA-Phi3)	31.49	20.41	16.46

Table 5. **Ablations on the Feature Distillation**. Case 1-3 show results of feature distillation with different MLLMs together with the composed loss. Case 4-6 show result of directly using teacher models for CIR, which is large in parameters and slow in inference compared with DistillCIR.

5.2. Effect of Feature Distillation

While the reasoning distillation exploits LLM outputs at the semantic level, the feature distillation aligns representations from the CIR model with those of an open-source MLLM. As shown in Cases 1 and 7 of Table 5, incorporating the feature distillation loss $\mathcal{L}_{fea}$ improves performance beyond the base $\mathcal{L}_{com}$. Because open-source MLLMs vary in scale, we investigate different model sizes of them. In addition to xtuner/llava-phi-3-mini-hf (4.2B), we examine bczhou/tiny-llava-v1-hf (1.4B) [60] and xtuner/llava-llama-3-8b (8B) [10]. The results in Cases 2, 3, and 7 show that the smallest model yields the lowest performance, while the medium and larger models perform comparably. Given evaluations on public MLLM benchmarks [30, 56], this suggests that distillation performance is influenced more by the LLM's overall capability than its size. Since llava-phi-3 matches llava-llama-3 in accuracy but is more compact, we adopt llava-phi-3 to balance performance and efficiency.

5.3. Learning Instruction Awareness from LLMs

In this section, we examine how DistillCIR acquires instruction awareness. To assess its understanding of instructions during composed retrieval, we propose a specific criterion: whether the model can selectively focus on particular regions of an image according to textual modifications. Pretrained CLIP models often emphasize dominant objects in an image, but the composed embedding should instead shift attention to the details specified by the textual cues. Drawing inspiration from [40, 55], we visualize patch-level attention maps to investigate whether DistillCIR indeed adjusts its focus based on the modifications.

Specifically, let $i \in \mathbb{R}^{H \times W \times C}$ be an input image, where H, W, and C denote its height, width, and channel dimensions, respectively. The visual encoder f_V produces patch embeddings $P \in \mathbb{R}^{L^2 \times D_V}$, with L^2 representing the total number of patches. Next, we project these patch embeddings into the text embedding space using f_ϕ and f_T. Finally, we compute an attention map between the projected patch embeddings and the composed embedding, as illustrated in Equation 19. The attention map A is resized to a matrix and which is interpolated to the size of the image. The detailed computation process is shown in Appendix C and visualized examples are shown in Figure 5.

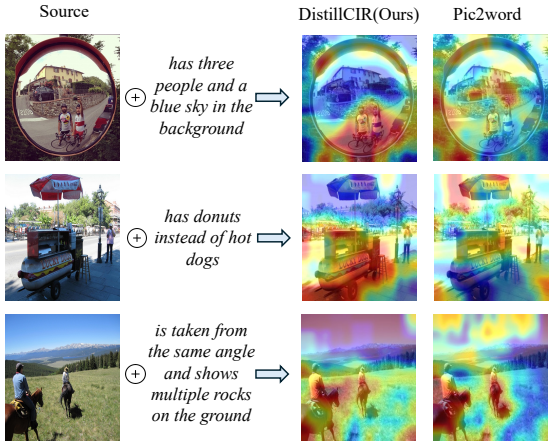

Figure 5. **Attention Map Comparison between DistillCIR and Pic2Word on the CIRCO Validation Set with Color Map JET**. Warmer colors indicate higher attention than cooler colors.

From these visualizations, it is evident that the earlier CIR model, Pic2Word, fails to highlight the specific regions indicated by textual modifications. We conjecture that previous projection-based models, trained solely on image-caption pairs, primarily merge image and text features without fully addressing the discrepancy between standard image-text retrieval and composed image retrieval [7]. In contrast, DistillCIR selectively attends to the relevant areas, reflecting its instruction awareness acquired via LLM-based distillation. While there has been limited research on instruction-following metrics for retrieval, the attention map serves as an initial quantitative metric. We leave the development of qualitative evaluation metrics for future work.

Cases	CIRR R@1	CIRCO mAP@10	GeneCIS R@1
1. LoRA on f_V	32.38	20.94	16.27
2. LoRA on f_T	31.40	21.36	16.05
3. Train f_ϕ from scratch	29.65	20.07	15.85
4. MLP layers in $f_\phi = 1$	30.38	20.19	16.44
5. MLP layers in $f_\phi = 2$	32.89	21.47	16.59
LoRA on (f_V, f_T), Full on f_ϕ (Ours)	33.45	21.52	16.83

Table 6. **Ablations on Other Training Settings**. Case 1-3 show results of training different components only and freeze others. Case 4-5 show results of training different MLP layers in f_ϕ under the same training and freezing strategy. Ours indicate train f_V and f_T with LoRA and full f_ϕ.

5.4. Analysis of Other Training Settings

In this section, we explore additional training setups. Table 6 reports the performance of DistillCIR under varying component configurations. As shown in Cases 1, 2, 3, and 6, applying LoRA to f_V and f_T while training full f_ϕ achieves the highest scores. This suggests that training more components generally enhances the model's learning capacity. Specifically, Applying LoRA on f_V yields better performance than f_T, indicating that the visual tower potentially acts a more important role. In addition, Cases 4-5 examines the effect of different MLP depths in f_ϕ. We observe continuous gains with increasing layers, reaching a maximum at three layers. Overall, we find that applying LoRA to both the visual and text towers, while restricting MLP layers in f_ϕ, achieves the best performance.

6. Conclusion

In this paper, we address instruction-aware Zero-Shot Composed Image Retrieval by distilling instruction-following capabilities from large language models into compact projection-based architectures. Our DistillCIR framework adopts a dual-stream strategy: it synthesizes pseudo triplet data with rationales to guide learning, and further leverages an open-source MLLM to embed instruction-aligned knowledge. This approach achieves strong performance while removing the need for LLMs during inference. DistillCIR highlights a practical path to instruction-aware retrieval, bridging projection-based and LLM-based models.

7. Acknowledgments

This work was partially supported by US National Science Foundation IIS-2412195, Microsoft Accelerate Foundation Models Research, 2024, CCF-2400785, the Cancer Prevention and Research Institute of Texas (CPRIT) award (RP230363) and the National Institutes of Health (NIH) R01 award (1R01AI190103-01).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 3

[2] Lorenzo Agnolucci, Alberto Baldrati, Marco Bertini, and Alberto Del Bimbo. isearle: Improving textual inversion for zero-shot composed image retrieval. *arXiv preprint arXiv:2405.02951*, 2024. 1, 2

[3] Anthropic. The claude 3 model family: Opus, sonnet, haiku. 3

[4] 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, 5, 6

[5] Oriol Barbany, Michael Huang, Xinliang Zhu, and Arnab Dhua. Leveraging large language models for multimodal search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1201–1210, 2024. 1

[6] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 2

[7] Jaeseok Byun, Seokhyeon Jeong, Wonjae Kim, Sanghyuk Chun, and Taesup Moon. Reducing task discrepancy of text encoders for zero-shot composed image retrieval. *arXiv preprint arXiv:2406.09188*, 2024. 8

[8] Defang Chen, Jian-Ping Mei, Yuan Zhang, Can Wang, Zhe Wang, Yan Feng, and Chun Chen. Cross-layer distillation with semantic calibration. In *Proceedings of the AAAI conference on artificial intelligence*, pages 7028–7036, 2021. 2

[9] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4794–4802, 2019. 2

[10] XTuner Contributors. Xtuner: A toolkit for efficiently fine-tuning llm. https://github.com/InternLM/xtuner, 2023. 5, 7

[11] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021. 2

[12] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, Yoohoon Kang, and Sangdoo Yun. Language-only efficient training of zero-shot composed image retrieval–appendix–. 1, 2, 5, 6

[13] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, HeeJae Jun, Yoohoon Kang, and Sangdoo Yun. Compodiff: Versatile composed image retrieval with latent diffusion. *arXiv preprint arXiv:2303.11916*, 2023. 6

[14] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2

[15] Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301*, 2023. 3

[16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 5

[17] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025. 3

[18] Young Kyun Jang, Dat Huynh, Ashish Shah, Wen-Kai Chen, and Ser-Nam Lim. Spherical linear interpolation and text-anchoring for zero-shot composed image retrieval. *arXiv preprint arXiv:2405.00571*, 2024. 2, 5, 6

[19] Young Kyun Jang, Donghyun Kim, Zihang Meng, Dat Huynh, and Ser-Nam Lim. Visual delta generator with large multi-modal models for semi-supervised composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16805–16814, 2024. 2, 6

[20] Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-v: Universal embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580*, 2024. 4, 1

[21] Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. Vision-by-language for training-free compositional image retrieval. *arXiv preprint arXiv:2310.09291*, 2023. 2, 5, 6

[22] Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal inputs and outputs. 2023. 2

[23] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021. 4

[24] Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Data roaming and quality assessment for composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2991–2999, 2024. 6

[25] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 2, 7

[26] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 4, 6, 7

[27] Wei Li, Hehe Fan, Yongkang Wong, Yi Yang, and Mohan Kankanhalli. Improving context understanding in multimodal large language models via multimodal composition learning. In *Forty-first International Conference on Machine Learning*. 2, 5, 6

[28] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 2

[29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2, 4, 5

[30] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 7

[31] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2125–2134, 2021. 5

[32] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023. 3

[33] Peyman Passban, Yimeng Wu, Mehdi Rezagholizadeh, and Qun Liu. Alp-kd: Attention-based layer projection for knowledge distillation. In *Proceedings of the AAAI Conference on artificial intelligence*, pages 13657–13665, 2021. 2

[34] Antonio Polino, Razvan Pascanu, and Dan Alistarh. Model compression via distillation and quantization. *arXiv preprint arXiv:1802.05668*, 2018. 2

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 5

[36] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19305–19314, 2023. 1, 2, 3, 5, 6

[37] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, 2018. Association for Computational Linguistics. 3, 5

[38] Krishna Srinivasan, Karthik Raman, Anupam Samanta, Lingrui Liao, Luca Bertelli, and Michael Bendersky. QUILL: Query intent with large language models using retrieval augmentation and multi-stage distillation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 492–501, Abu Dhabi, UAE, 2022. Association for Computational Linguistics. 3

[39] Yucheng Suo, Fan Ma, Linchao Zhu, and Yi Yang. Knowledge-enhanced dual-stream zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26951–26962, 2024. 5, 6

[40] Yuanmin Tang, Jing Yu, Keke Gai, Jiamin Zhuang, Gang Xiong, Yue Hu, and Qi Wu. Context-i2w: Mapping images to context-dependent words for accurate zero-shot composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5180–5188, 2024. 5, 6, 8

[41] Sagar Vaze, Nicolas Carion, and Ishan Misra. Genecis: A benchmark for general conditional image similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6862–6872, 2023. 5

[42] Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. Covr: Learning composed video retrieval from web video captions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5270–5279, 2024. 1

[43] Boshi Wang, Xiang Yue, and Huan Sun. Can chatgpt defend its belief in truth? evaluating llm reasoning via debate. *arXiv preprint arXiv:2305.13160*, 2023. 3

[44] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*, 2023. 4

[45] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018. 2

[46] Xiaobo Wang, Tianyu Fu, Shengcai Liao, Shuo Wang, Zhen Lei, and Tao Mei. Exclusivity-consistency regularized knowledge distillation for face recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 325–342. Springer, 2020. 2

[47] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*, 2022. 2, 3

[48] Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. Uniir: Training and benchmarking universal multimodal information retrievers, 2023. 2

[49] 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. 2, 3

[50] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. The fashion iq dataset: Retrieving images by combining side information and relative natural language feedback. *CVPR*, 2021. 5

[51] Kunran Xu, Lai Rui, Yishi Li, and Lin Gu. Feature normalized knowledge distillation for image classification. In *European conference on computer vision*, pages 664–680. Springer, 2020. 2

[52] Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*, 2024. 3

[53] Chuanguang Yang, Zhulin An, Libo Huang, Junyu Bi, Xinqiang Yu, Han Yang, Boyu Diao, and Yongjun Xu. Clip-kd:

An empirical study of clip model distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15952–15962, 2024. 2

[54] Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, Yu-Yang Liu, and Li Yuan. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469*, 2023. 3

[55] Runpeng Yu, Weihao Yu, and Xinchao Wang. Attention prompting on image for large vision-language models. *arXiv preprint arXiv:2409.17143*, 2024. 8

[56] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 7

[57] Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. Magiclens: Self-supervised image retrieval with open-ended instructions. *arXiv preprint arXiv:2403.19651*, 2024. 6

[58] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023. 2

[59] Wenliang Zhong, Weizhi An, Feng Jiang, Hehuan Ma, Yuzhi Guo, and Junzhou Huang. Compositional image retrieval via instruction-aware contrastive learning. *arXiv preprint arXiv:2412.05756*, 2024. 1

[60] Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. Tinyllava: A framework of small-scale large multimodal models. *arXiv preprint arXiv:2402.14289*, 2024. 7

[61] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023. 2

[62] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 4

[63] Xinliang Zhu, Sheng-Wei Huang, Han Ding, Jinyu Yang, Kelvin Chen, Tao Zhou, Tal Neiman, Ouye Xie, Son Tran, Benjamin Yao, et al. Bringing multimodality to amazon visual search system. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6390–6399, 2024. 1

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

Supplementary Material

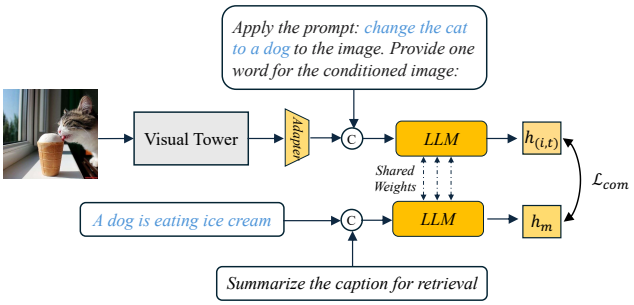

Figure 6. **Training Process of the Teacher MLLM.** Images are first processed by the visual tower and projected into the LLM embedding space. The textual modification, appended to a prompt template, is then combined with the projected image embeddings and fed into the LLM to generate the composed embedding $h_{(i,t)}$. 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, 59] 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.