# Scaling Prompt Instructed Zero Shot Composed Image Retrieval with Image-Only Data

Yiqun Duan<sup>1,2</sup> Sameera Ramasinghe<sup>1</sup> Stephen Gould<sup>2,3</sup> Ajanthan Thalaiyasingam<sup>1</sup> <sup>1</sup> Amazon <sup>2</sup> University of Techonlogy Sydney <sup>3</sup> The Australian National University

Abstract-Composed Image Retrieval (CIR) is the task of retrieving images matching a reference image augmented with a text, where the text describes changes to the reference image in natural language. Traditionally, models designed for CIR have relied on triplet data containing a reference image, reformulation text, and a target image. However, curating such triplet data often necessitates human intervention, leading to prohibitive costs. This challenge has hindered the scalability of CIR model training even with the availability of abundant unlabeled data. With the recent advances in foundational models, we advocate a shift in the CIR training paradigm where human annotations can be efficiently replaced by large language models (LLMs). Specifically, we demonstrate the capability of large captioning and language models in efficiently generating data for CIR only relying on unannotated image collections. Additionally, we introduce an embedding reformulation architecture that effectively combines image and text modalities. Our model, named InstructCIR, outperforms state-of-the-art methods in zero-shot composed image retrieval on CIRR and FashionIQ datasets. Furthermore, we demonstrate that by increasing the amount of generated data, our zero-shot model gets closer to the performance of supervised baselines.

Index Terms—Composed Image Retrieval, Multimodality retrieval

## I. INTRODUCTION

The objective of composed image retrieval [13], [24] is to search for an image that aligns with both a reference image and a textual input detailing the desired alterations to that reference. This allows users to modify an image-based search query with natural language, facilitating a clear articulation of intent. Such capabilities have wide-ranging applications, including e-commerce, recommendation systems, and search engines.

The effectiveness of recent CIR methods [6], [10], [13], [35], [47] largely depends on the pre-trained vision and language models such as CLIP [39] and BLIP [28], which utilize contrastive semantic matching. Nonetheless, these models need to be finetuned specifically for CIR using triplet data containing reference images, reformulation texts, and target images. This reliance on human-annotated triplet data hinders the scalability of CIR models. Additionally, the scarcity of triplet data can impede fine-grained reformulations across the interrelated visual and text modalities.

In contrast, two parallel studies, Pic2Word [41] and ZS-CIR [3], introduced the zero-shot composed image retrieval task by leveraging image inversion techniques [17]. Both



Fig. 1. Performance curve versus current zero shot composed image retrieval benchmarks, where grey dashlines - - indicates supervised baselines as intuitive references. Our zero-shot model (shown in purple) closes the gap with supervised baselines with increasing amount of generated data.

methods convert the reference image into a single text token and apply reformulation in the text domain. However, this transformation of images to singular text tokens may result in substantial loss of intricate visual details, leading to subpar zero-shot performance.

A concurrent work, TransAgg [34], suggests the use of an LLM (ChatGPT) to produce training data from existing image caption datasets. Yet, this approach remains reliant on preexisting image-caption paired data and uses generically trained image and text encoders while only optimizing an aggregation layer. Moreover, TransAgg has been explored only on a limited data scale.

Our approach further relaxes the data requirement by solely relying on unannotated image collections. We hypothesize that given an arbitrary image pair, one can generate the natural language description of the difference by utilizing large vision and language models, effectively replacing human annotations. To this end, we first leverage the LLaVa model [33] to generate captions for a randomly sampled image pair. Subsequently, we utilize an LLM to delineate the differences between these captions. This technique enables the formation of triplets in a zero-shot manner, relying exclusively on unannotated image collections.

Additionally, we introduce a simple, yet effective joint embedding reformulation architecture that fuses the image and text modalities using cross-attention at multiple levels. Such a latent fusion design enables fine-grained image manipulations using text and has been used in text-guided image generation [9], [40], [49]. Nevertheless, this embedding reformula-

This work was completed by Yiqun during his internship at Amazon, Australia.



Fig. 2. InstructCIR workflow. Given an image pair  $\{\mathbf{x}_a, \mathbf{x}_b\}$ , captions are generated by the LLaVA model. Next, the LLM generates the natural language description of the differences between these captions. Then, the reference image and the generated difference text are fused at multiple levels and the model is trained to minimize the embedding distance with respect to the target image.

tion approach has not been used in the CIR domain to our knowledge and outperforms late fusion techniques as well as direct pixel-space manipulations as shown in our experiments. Our contributions are three-fold:

- We introduce a scalable framework for training CIR models that relies solely on unannotated image collections, replacing the need for paired image-caption data and human annotations.
- We introduce a simple, yet effective latent fusion architecture to effectively combine image and text modalities for CIR.
- Our model sets the state-of-the-art on zero-shot CIR benchmarks and closes the gap between zero-shot and supervised settings aided by efficient data scaling; See Fig. 1.

## II. INSTRUCTCIR

This section presents a detailed overview of our zeroshot triplet data generation pipeline (Sec. II-A) and model architecture (Sec. II-B). Section II-C elaborates on the training process as well as objective functions for composed image retrieval.

## A. Zero-Shot Triplet Generation

A comprehensive workflow of our approach is depicted in Fig. 2. Initially, an image pair is randomly selected from the provided image collection. This pair is then transformed into textual descriptions via image captioning. Subsequently, the LLMs are prompted with these generated captions to produce the reformulation text that describes the difference between the captions.

*a) Image Captioning::* One key component in ensuring zero-shot CIR only depends on the given image collection is to convert images into a sufficiently detailed text description. We prompt a powerful visual instructed large language model

# TABLE I THE UPPER PART DENOTES PROMPTS USED TO CONVERT IMAGES INTO

TEXTS. THE LOWER PART INDICATES TEXT PROMPTS USED FOR REFORMULATING FROM IMAGE CAPTION A TO CAPTION B.

	LLM 🐝
Visual Encoder	Prompts <b><token></token></b> You are a helpful language and vision assistant, you are able to understand the visual content, and assist the
	user with a variety of tasks using natural language. Describe the image as detail as possible
Reformulation Pr	ompt:

You have two captions for two images, image A and image B, you are supposed to write a reformulation text describing changing from image A to image B. caption A: {Caption A} caption B: { Caption B } answer should be concise and within 12 words, only contain normal words, do not use special characters. Difference:

(LLaVA<sup>1</sup> [33]) to convert the sampled image pairs into a text pair. The prompt for the captioning model is provided in the top part of Table I, and example results are shown in Fig. 3. The image is tokenized and encoded through a ViT model and is fed into an LLM as context tokens. The LLM generates the detailed image caption given visual tokens and task prompts. The examples on the Fashion200k dataset show that, although LLM may generate common descriptions such as "posing for a photography shot", "a woman is wearing", it still accurately describes key elements of the given image, such as the object (*e.g.*., shirt, dress), style (*e.g.*., strapless, off-the-shoulder, split style), color, *etc.*.

b) Reformulation with LLMs:: Utilizing the high-quality image captions obtained as above, our approach leverages an LLM fine-tuned from vicuna 33B checkpoint<sup>2</sup> to generate reformulated text-prompts in a zero-shot manner. The language model is fine-tuned using low-rank adaptation finetuning (LORA) [25] with 10 epochs. The fine-tuning seed data is obtained by prompting ChatGPT-4 in zero-shot. Here, directly prompting ChatGPT-4 is also feasible to generate reformulation data. Here we fine-tuning our own model for the feasibility of the ablation study. We employ the prompt structure illustrated in Table I, which is crafted without prior training specific to this task. Illustrative examples of generated reformulation texts are presented in Fig. 3. These highlight the efficacy of our data generation pipeline. Rather than merely presenting experiment outcomes, we delve deeper to discern the influence of various language models on the performance metrics.

#### B. Model Structure

Our proposed architecture capitalizes on a streamlined yet potent framework; a layer-by-layer **text-guided embedding** 

<sup>&</sup>lt;sup>1</sup>https://huggingface.co/liuhaotian/llava-v1.5-13b

<sup>&</sup>lt;sup>2</sup>https://huggingface.co/lmsys/vicuna-33b-v1.3



Fig. 3. Caption results by prompting visual LLMs on Fashion200k and NLVR, where accurate information and unrelated information is highlighted in different color. Though unrelated information exists, it still suggests zero-shot image-text conversion is feasible with LLMs. Reform corresponds to the reformulation text generated for the images in the same column.



Fig. 4. Model structure of proposed embedding reformulation network. The text embedding is injected into the visual transformer through cross-attention layer by layer.

**reformulation network**, as illustrated in Fig. 4. This design is inspired by in the latest advancements in generative image manipulation [8], [40] where models adeptly utilize text instructions to guide image generation. Although the

generated images are visually plausible, they fail on recall metrics (as shown in Table II) due to overly strict pixel-topixel correlations. Instead of the pixel space, we harness the joint latent space [1] to perform the instructed embedding reformulation.

We modify each transformer block of the ViT model [14] by introducing cross attention from the text encoder. Our text encoder is the same as CLIP [39], which takes text instruction as input and outputs an embedding sequence. In each ViT block, one self-attention layer is first applied on visual embeddings and cross-attention is utilized to inject the text instruction on embeddings, where the text embeddings are treated as K and V and image embeddings as queries Q. Then, an MLP layer combined with layer norm [2] and residual connection is utilized to aggregate the features. Several such reformulation blocks are stacked to get the final output. On the top of the encoder, a predictor is used to obtain the final embedding for retrieval. For the target images, which do not have an instruct prompt, we use a null-text token as the text input. This enables us to embed both the source and target images within a shared model architecture. Additionally, the target encoder is updated using an exponential moving average similar to I-JEPA [1].

#### C. Composed Retrieval Training

We train the joint-embedding reformulation network by minimizing the distance between the reformulated query embeddings and the target embeddings. For a fair comparison, we maintain the model size consistent with preceding models which are similar to the CLIP model. For a given query with an image pair  $\mathbf{x}_a, \mathbf{x}_b$ , we obtain the associated captions  $C_a, C_b$  using the LLaVa model. Subsequently, the reformulation LLM produces the reformulation text  $t_{ab}$  in a zero-shot fashion, defined as  $t_{ab} = \text{LMM}_{\text{diff}}(C_a, C_b)$ .

Subsequently, the reformulated embedding  $\mathbf{e}_r$  is extracted by our proposed encoder and a predictor. Formally,

$$\mathbf{e}_r = g_\theta(\mathbf{z}_r, \phi_t(t_{ab})), \text{ where } \mathbf{z}_r = f_\theta(\mathbf{x}_a, t_{ab}),$$
 (1)

and  $\phi_t(\cdot)$  is the pre-trained text encoder. The predictor  $g_\theta(\cdot)$  is a MLP projection layer based on concatenated image and text embeddings  $\{\mathbf{z}_r, \phi_t(t_{ab})\}$ . Detailed description of the encoder  $f_\theta(\cdot, \cdot)$  is provided in the model structure section (Sec. II-B and Fig. 4). The target embedding is extracted using the same function with the null-text token embedding instead of input text,  $\mathbf{e}_t = g_\theta(\mathbf{z}_t, \phi_t(\emptyset))$ , where  $\mathbf{z}_t = f_\theta(\mathbf{x}_b, \emptyset)$ , Here,  $\mathbf{e}_r, \mathbf{e}_t \in \mathbb{R}^d$  where d is embedding dimensionality.

Our training objective is to minimize the distance between the query embedding  $\mathbf{e}_r$  and the target embedding  $\mathbf{e}_t$  for a given triplet  $\{\mathbf{x}_a, \mathbf{x}_b, t_{ab}\}$  from the zero-shot pipeline. Simultaneously, we maximize the distance between  $\mathbf{e}_r$  and embeddings of other target images within the batch. To accomplish this, we utilize a batch-centric contrastive loss:

$$\mathcal{L}_{\text{contr}} = \frac{1}{B} \sum_{i=1}^{B} -\log \frac{\exp\{\tau \cdot \kappa(\mathbf{e}_{r}^{i}, \mathbf{e}_{t}^{i})\}}{\sum_{j=1}^{B} \exp\{\tau \cdot \kappa(\mathbf{e}_{r}^{i}, \mathbf{e}_{t}^{j})\}}, \qquad (2)$$

Here,  $\mathbf{e}_r^i$ ,  $\mathbf{e}_t^i$  denote the embeddings of the reformulation encoder and the target encoder for the *i*-th triplet,  $\kappa(\cdot, \cdot)$ denotes the cosine similarity,  $\tau > 0$  is a temperature parameter that controls the range of the logits, and *B* is the number of triplets in a batch. Both text encoder and the reformulated image encoder are updated during the alignment training.

Given reformulated embeddings, we further boost the performance by fusing the pure text embedding with the reformulated embedding using the following settings from most of the previous papers [6], [36], [37]. Given the reformulated embedding  $\mathbf{e}_r$ , we calculate the final embedding as follows:

$$\mathbf{e}_f = \lambda \mathbf{e}_r + (1 - \lambda)\phi_t(t_{ab}),\tag{3}$$

where  $\lambda$  is the trained hyper-parameter weight to combine original features from two sides. After the whole network is trained, this combiner weight is fine-tuned using the same  $\mathcal{L}_{contr}$  by fixing backbone weights and replacing  $\mathbf{e}_r$  with  $\mathbf{e}_f$  in Equation 2.

## III. RELATED WORK

a) Composed Image Retrieval (CIR):: Composed image retrieval (CIR) retrieves images using a reference image-text pair [15], [16], [47], with applications in fashion [48] and scene composition [35]. Traditional methods merge latent embeddings [29]–[31] from both modalities to form retrieval queries. Techniques range from TIRG's gating and residual connections [47] to VAL's transformer-based hierarchical design [10]. Wu et al. [48] employ a custom transformer for early image-language fusion. In contrast, Goenka et al. [18] use BERT [12] for image-text-tag unified coding, while Han et al. [22] pre-train a model using a vast fashion dataset. Modern CIR approaches, like CLIP4CIR [7] and BLIP4CIR [36], leverage pre-trained visual-language models and apply late fusion. CASE [27] enhances this by adding external data. Candidate-R [37], aiming for peak performance, re-ranks retrieval candidates, albeit at a much higher computational cost. Notably, all these strategies require source-prompt-target triplets for training, and the high cost of obtaining such data constrains CIR's broader application.

b) Zero-Shot Compsed Image Retrieval:: The concept of zero-shot composed image retrieval has recently garnered significant attention. Two contemporaneous studies, Pic2Word [41] and ZS-CIR [3], utilize image-caption datasets to train networks that represent images as singular tokens, thus facilitating cross-modal retrieval in the text domain. CompoDiff [19] leverages a modified diffusion-denoising model to iteratively refine search queries and introduces a new dataset, SynthTriplet18M. This dataset comprises images synthesized through the prompt-to-prompt model [23], guided by corresponding captions.

Our concurrent work, TransAgg [34], harnesses ChatGPT combined with human-translated templates on selected image caption data from LAION [42], yielding impressive results. Distinctly, our approach aims to achieve zero-shot image retrieval relying solely on image distribution. By integrating image captioning models with large language models (LLMs) and capitalizing on scaling potential, InstructCIR sets new benchmarks in the realm of composed image retrieval.

#### **IV. EXPERIMENTS**

In this section, we provide comprehensive experiments to illustrate the state-of-the-art performance of InstructCIR on both zero-shot and supervised composed image retrieval.

#### A. Setup

The reformulation network is modified from CLIP pretrained ViT-L/14 and injects cross attention to each layer. The cross-attention layer is initialized with Xavier [26] initialization. The cross-attention layer has the same heads as the main backbone. The training data is sampled from Fashion200k [21] and NLVR [45] which respectively have 280k and 21.4k images. For the Fashion200k dataset, we randomly sample image pairs under the same meta class <sup>3</sup>, while for NLVR we sample the whole dataset. For the image caption model, we directly used LLaVA Vicuna 13B pre-train weights. We use our own language model with 33B as the text reformulator, also, we provide a comparison with different language models in supplementary.

Since the unique combination is scalable, we creating image pairs from 16k up to 500k to report the performance curve. The model is trained with AdamW [38] optimizer, with learning rate  $2 \times 10^{-6}$ , weight decay 0.1, batch size 32. The model is implemented using PyTorch and trained with eight A100 GPU instances.

<sup>&</sup>lt;sup>3</sup>Fashion200k has five meta classes: dresses, jackets, pants, skirts, tops.

## TABLE II

QUANTITATIVE COMPARISON WITH STATE-OF-THE-ART ZERO-SHOT METHODS (AND WITH SUPERVISED BASELINES). HERE, \* INDICATES DATA ARE SAMPLED FROM SUBSETS OF THE FASHION200K AND NLVR DATASETS. FOR ZERO-SHOT METHODS, EVALUATION METRICS ARE FROM THE SAME MODEL ACROSS TWO DATASETS. FOR SUPERVISED METHODS, EVALUATIONS ARE FROM THE BEST MODELS RESPECTIVELY ON EACH DATASET. <sup>†</sup> DENOTES THE TRAINING SCALE IS ADDITIONAL TO THE ORIGINAL CLIP PRE-TRAINED MODEL. CANDIDATE R<sub>50/100</sub> IS MARKED GREY AS IT REQUIRES A MUCH HIGHER COMPUTATIONAL COST. **RESULTS INDICATE THAT OUR INSTRUCTCIR REACHES STATE-OF-THE-ART PERFORMANCE ON BOTH ZERO-SHOT AND SUPERVISED BENCHMARKS. WE SELECT THE BEST PERFORMANCES IN EXISTING METHODS FOR FAIR COMPARISON. THE FIRST SECTION RESULTS (IMAGE OR TEXT ONLY) ARE WITH THE SAME CLIP-L14 ARCHITECTURE AS OUR METHOD.** 

			#Sample	CIRR		FashionIQ				
Methods	Zero-Shot	Data Source	Scale	R@1	R@5	R@50	$R_{Subset}@1$	R@10	R@50	Average
Image Only + CLIP-L14	$\checkmark$	CLIP [39]	-	8.42	23.81	61.03	22.98	6.33	15.21	10.78
Text Only + CLIP-L14	$\checkmark$	CLIP [39]	-	22.98	46.83	82.36	63.88	19.05	37.82	28.63
Image-Text Sum. +CLIP-L14	$\checkmark$	CLIP [39]	-	11.71	35.06	77.49	32.77	24.60	43.21	33.91
InstructPix2Pix +CLIP-L14	$\checkmark$	CLIP [39]/ LAION [43]	-	22.03	47.81	83.52	61.67	9.86	19.63	15.03
		Zero Sho	t Benchmarl	ks						
PALAVRA [11]	$\checkmark$	PerVL [11]	$\sim 1$ m.	16.62	43.49	83.95	41.61	19.76	37.25	28.51
Pic2Word [41]	$\checkmark$	CC3M [44]	3m.	23.90	51.70	87.80	-	24.70	43.70	34.20
SEARLE-XL-OTI [3]	$\checkmark$	COCO (CIRCO) [32]	118k.†	24.87	52.31	88.58	53.80	27.61	47.90	37.76
CompoDiff w/T5-XL [19]	$\checkmark$	SynthTriplets18m [44]	18m.	19.37	53.81	90.85	28.96	37.36	50.85	44.11
CASE Pre-LaSCo.Ca. [27]	$\checkmark$	LaSCo [27]	360k.†	35.40	65.78	94.63	64.29	-	-	-
TransAgg	$\checkmark$	LAION [43]	32k <sup>†</sup>	37.87	68.88	93.86	69.79	34.64	55.72	45.18
COVR [46]	$\checkmark$	WebVid- [46]	1.6m <sup>†</sup>	39.28	68.22	94.65	-	27.70	44.63	36.15
InstructCIR (Ours)	$\checkmark$	LAION [43]	300k <sup>†</sup>	38.56	69.21	95.21	68.22	36.56	56.33	46.89
InstructCIR (Ours)	$\checkmark$	Fashion200k [20]/ NLVR [45]*	$300k^{\dagger}$	39.28	69.62	95.88	69.87	<u>37.32</u>	56.84	47.08
		Compared to S	upervised L	earning.						
CLIP4CIR [4]	×	CIRR/FashionIQ	-	38.53	69.98	95.93	68.19	38.32	61.74	50.03
BLIP4CIR+Bi [36]	×	CIRR/FashionIQ	-	40.15	73.08	96.27	72.10	43.49	67.31	55.40
CASE Pre-LaSCo.Ca. <sup>†</sup> [27]	×	CIRR/FashionIQ	-	49.35	80.02	97.47	76.48	48.79	70.68	59.74
Candidate F [37]	×	CIRR/FashionIQ	-	44.70	76.59	97.18	75.02	46.15	69.15	57.65
Candidate $\mathbf{R}_{50/100}$ [37]	$\times$	CIRR/FashionIQ	-	50.55	81.75	97.18	80.04	51.17	73.13	62.15
COVR [46]	×	CIRR/FashionIQ	-	50.41	80.96	97.64	-	49.40	70.98	60.19
InstructCIR (Ours)	×	CIRR/FashionIQ	-	50.70	81.61	98.27	<u>76.10</u>	<u>49.03</u>	<u>70.96</u>	<u>60.00</u>

TABLE III Comparison on CIRCO dataset

Backbone	Method	K = 5	K = 10	K = 25	K = 50
B/32	SEARLE-OTI	7.14	7.83	8.99	9.60
	SEARLE	9.35	9.94	11.13	11.84
	InstrucCIR	<b>10.23</b>	<b>10.98</b>	<b>12.07</b>	<b>13.88</b>
L/14	Pic2Word	8.72	9.51	10.64	11.29
	SEARLE-XL-OTI	10.18	11.03	12.72	13.67
	SEARLE-XL	11.68	12.73	14.33	15.12
	InstructCIR	<b>12.94</b>	<b>13.84</b>	<b>15.62</b>	<b>16.34</b>

#### B. Zero-Shot Quantitative Evaluation

a) Baselines:: To illustrate the efficiency of our proposal, we provide a comparison with a wide range of zero-shot CIR baselines. CLIP [39] and PALAVRA [11] provide baselines that utilize frozen vision-language pretraining models. We respectively evaluate *text only, image only*, and direct embedding summation (*Image-Text Sum.*) to report the performance of each modality. Pic2Word [41] and SEARLE [3] represent realizing zero-shot CIR by converting an image into a single text token. Moreover, we provide an image editing baseline by using InstructPix2Pix [8] to edit reference images towards the target image with text prompt then applying pure image retrieval. CompoDiff [19] represents a diffusion-based generative model on latent space. CASE [27] and TransAGG [34] suggest using LLMs to generate reformulation data but rely on image-text pairs. *b) Performance::* The zero-shot evaluation is conducted across two datasets with the same model weights, FashionIQ [48] (fashion) and CIRR [35] (real-life scenarios) in Table II.

The first sector provides an intuitive understanding of zero-shot performance by using raw model analysis. Textonly and image-text summation using CLIP [39] exceed the performance of early baseline PALAVRA [11] and image only. For recent works, Pic2word [41] and SEARLE [3] introduce 2.09% to 7.09% performance improvement on recall metrics. CompoDiff [19] reaches highest 37.36% top 10 recall (R@10) on FashionIQ while failing to reach competitive performance on CIRR. TransAgg [34] and CASE further boost performance by introducing pretraining data.

While just given image distribution, InstructCIR reaches recall 38.18%, 69.62%, and 95.88% respectively on  $R@\{1,5,50\}$  on CIRR dataset. Using the same model, InstructCIR reaches recall 36.91% and 55.84% respectively on R@10 and R@50 on the FashionIQ dataset. Results demonstrate that our InstructCIR reaches new state-of-the-art performance across two major datasets.

c) Extensive Zero-Shot Performance on Benchmark CIRCO: We have now compared our method (InstructCIR) with two other zero-shot methods on an extensive dataset proposed recently called CIRCO [3]. The observation is similar to other datasets, that our approach clearly outperforms previous methods even on the CIRCO dataset. We will include these results in the revised manuscript.

## TABLE IV

Ablation study on different models, reformulation LLMs, and data distribution with zero-shot setting. 33B<sup>∓</sup> denotes our own Language model with 33 billion parameters. CIRR\* and FashionIQ\* denotes just utilizing images from GT dataset without annotation but creating zero-shot reformulation using our pipeline. 33B<sup>∓</sup>-GT denotes we use supervise data to fine-tune the Language model. InstructCIR outperforms TransAgg with respective to the quality of generated data as well as the MODEL ARCHITECTURE.

Model	Backbone	Images	Caption	#Reform LLM	#Scale	R@1	CIR R@5	<b>R</b> R <sub>Subset</sub> @1	Fashi R@10	ionIQ R@50
	CLIP L/14	TransAgg	LAION	Temp.	32k	33.04	64.39	63.37	32.63	53.65
TransAgg	CLIP L/14	TransAgg	LAION	ChatGPT	32k	32.67	64.05	62.98	32.45	53.15
	CLIP L/14	Fashion/NVLR	LLaVa	33B∓	32k	33.26	65.67	64.05	32.91	53.41
InstructCID	$\overline{\text{CLIP } L/14}$	TransAgg	LAION	-33B <sup>+</sup>	32k	35.58	67.85	66.87	33.52	54.07
liisuucicik	CLIP L/14	Fashion/NVLR	LLaVa	33B∓	32k	35.83	68.04	66.93	33.78	54.82
		Scali	ng-Up Exp	eriments with	Zero-Shot	t Pipeline				
	CLIP L/14	CIRR*	LLaVa	33B∓	3.6k	34.74	65.83	66.43	32.21	53.19
	CLIP L/14	FashionIQ*	LLaVa	33B∓	5.9k	33.08	65.98	66.38	33.64	54.62
	CLIP L/14	Fashion/NVLR	LLaVa	33B∓	32k	35.83	68.04	66.93	33.78	54.82
InstructCIR	CLIP L/14	Fashion/NVLR	LLaVa	33B∓	65k	36.72	68.49	67.03	34.85	55.16
	CLIP L/14	Fashion/NVLR	LLaVa	33B∓	95k	36.92	68.64	68.34	35.94	55.78
	CLIP L/14	Fashion/NVLR	LLaVa	33B∓	200k	38.04	69.56	69.45	36.64	56.18
	CLIP L/14	Fashion/NVLR	LLaVa	33B∓	300k	38.86	69.62	69.87	37.32	56.84
	CLIP L714	Fashion/NVLR	LLaVa	$3\overline{3}B^{\mp}-\overline{GT}$	<u>300</u> k	41.32	72.55	71.01	38.92	58.43

## C. Supervised Quantitative Evaluation

To underscore the efficiency of our proposed embedding reformulation network, we benchmarked it using standard supervised learning on two prominent CIR datasets: FashionIQ [48] (fashion-centric) and CIRR [35] (reflecting reallife scenarios). Unlike the zero-shot setting, which evaluates a single model across both datasets, the supervised benchmark trains optimized models for each dataset before evaluation.

It's worth noting that the Candidate Re-ranking process, which refines results from the top 100 retrieved candidates, incurs additional computational costs. Hence, we've highlighted Candidate  $\mathbf{R}_{100}$  in grey. In the CIRR evaluation, our InstructCIR outperforms most preceding benchmarks, even surpassing the Candidate  $\mathbf{R}_{100}$  in the Recall@K metrics. For the Recall<sub>Subset</sub>@K metric, while Candidate  $\mathbf{R}_{100}$  achieves the peak performance, both CASE and our proposal are closely competitive for the second-best score. In the FashionIQ evaluation, InstructCIR ranks slightly below CASE, with Candidate  $\mathbf{R}_{100}$  securing the third spot. These results suggest that our network architecture is on par with the state-of-the-art when compared against supervised baselines, underscoring the efficacy of our model design.

## D. Ablation Study

TransAgg [34] and our model significantly outperform previous models as these two models both utilize LLMs to create reformulated text prompts. The difference is that TransAgg utilizes existing image caption dataset, but InstructCIR creates captions given the image distribution. We further conducted detailed ablation studies in Table IV to help people understand the performance improvement.

*a)* Architectural Efficiency:: In this section, we provide an ablation study on architecture efficiency compared to concurrent work TransAgg [34]. The first sector of Table IV compares TransAgg [34] with our model by exchanging

TABLE V Ablation study on architectural design across CIRR and FashionIQ with zero-shot setting. Cross-attention-based latent fusion yields the most benefit.

Model	#Scale	CIRR R@1	R@5	Fashion R@10	IQ R@50
InstructCIR	32k	35.83	68.04	33.78	54.82
+w/o EMA	32k	35.04	67.85	33.16	54.23
+w/o CrossAtt.	32k	28.45	59.33	24.53	45.67

training data. By controlling data scale the same (32k), TransAgg R@{1, 5} of CIRR dataset respectively increases from 32.67%, 64.05% to 33.26%, 65.67%. This suggest the efficiency of zero-shot data creating pipeline of InstructCIR. Also, while using the TransAgg data training, our embedding reformulation network also could reach 35.58%, 67.85% on CIRR R@{1, 5}, which is still higher than previous model. This observation suggests that, while using the same data with the same scale of model, our methods outperforms TransAgg clearly. The improvement is compared smaller on the FashionIQ dataset, but the conclusion still stands.

Table V provides an ablation study on **architectual design**. The model is trained with InstructCIR data with a data scale of 32k samples. When removing EMA [1] update, the CIRR R@1 and FashionIQ R@10 drop slightly from 35.83% to 35.04% and from 33.78% to 33.16%. When further removing both EMA and cross attention to perform embedding reformulation, which is the original CLIP model, the average recall drops significantly by  $6.59\% \sim 8.52\%$  on both CIRR and Fashion IQ. This suggests the design efficiency of the proposed embedding reformulation network. We conducted a supervised CIR comparison in Section IV-C, which further illustrates the architectural efficiency by comparing with previous methods with the same training data.



Fig. 5. Qualitative results on FashionIQ dataset. For each given reference image and text instructions, we visualize the top 10 retrieved candidates. Here, green boxes  $\Box$  denote reference images, and red boxes  $\Box$  denotes target images indicated by ground truth label.

b) Language Model Reformulation:: To illustrate how performance improvements related to pure image distribution we sampled from Fashion200k and NLVR. We directly take the images from the supervised dataset, FashionIQ, and CIRR and abandon the annotations. The images are fed through the zero-shot pipeline to train the model. Results in Table IV suggest that, although these two datasets show a slightly better performance on data (3.6k and 5.9k) closer to their own distribution, the differences are not significant and they still fall short of models with larger data scales. This suggests the good salable property of InstructCIR.

c) Scaling-Up:: The lower part of Table IV reports the scaling-up experiment of InstructCIR. The performance rises from 35.83% R@1 to 38.86% R@1 on the CIRR dataset when scaling from 32k to 300k. The findings indicate good scalability, demonstrating that as the data scale increases, so too does performance.

Furthermore, to obtain an approximate empirical performance upper bound, we augment our approach with semisupervised data. By utilizing triplet data from the CIRR [35] and the FashionIQ [48] datasets, we further boost the language model by leaking information from the ground truth. This exploration is critical, as it provides a trajectory of performance enhancement: starting from a zero-shot scenario and progressively approaching supervised benchmarks. With the help of the boosted language model, the performance could further rise to 41.32% R@1 on the CIRR dataset and 38.92% R@10 on the FashionIQ dataset, indicated as  $33B^{\mp}$ -GT in Table IV. This performance **is even comparable with advanced supervised baselines** CLIP4CIR [5] and BLIP4CIR [36].

*d) Test Domain Variance::* InstructCIR employs LLMs to generate text instructions in a zero-shot manner for training purposes. To showcase the generalization ability of Instruct-

TABLE VI Ablation study by replacing GT text instructions by generated text instructions. Small difference indicate the domain gap between generated and GT reformulation texts is small.

		CIRR		Fashion	IQ
Image	Test Text	R@1	R@5	R@10	R@50
GT	GT	35.83	68.04	33.78	54.82
GT	Generated	37.45	68.31	33.92	54.87
	$\Delta$ – – – – –	+1.62	+0.27	+0.14	+0.04

CIR in real-world test distributions, we have conducted an ablation study, presented in Table VI. In this study, while retaining the images from the ground truth (GT), we replaced the text GT with generated texts. Our hypothesis was that by using generated texts closer to the training distribution, the model would exhibit improved performance. Indeed, this change led to an increase in top-1 recall (R@1) as seen in Table VI. However, only marginal differences were observed for R@5, R@10, and R@50 metrics. Such results underscore the robust generalization capabilities of our proposed data creation pipeline.

#### E. Qualitative Analysis

To provide a clear understanding of the retrieval performance, we present visualizations of the retrieval results on the FashionIQ dataset (Figure 5) and CIRR dataset (Figure 6. For each reference image (highlighted with green boxes) paired with text prompts (displayed as the title of each line), we showcase the top 10 retrieved candidates. The target image's ground truth is marked with red boxes on each line.

It's widely recognized that metric-based evaluations can sometimes diverge from human judgments. This is because Giarda darker clothes and one of the mirrors overlapping the other.



Fig. 6. Qualitative results on CIRR dataset. green boxes  $\Box$  denote reference images and red boxes  $\Box$  denotes GT label.

there may be multiple reasonable results beyond just the ground truths. For instance, in the middle-right line of our visualization, all retrieved results adhere to the prompt "Has floral design and has flowers and is blue". However, the ground truth only indicates one of these valid candidates. Common failures occur when the original reference images still appear among the top 10 retrieved candidates. This is especially prevalent when the reference image captures comprehensive human attributes, like a face.

#### V. LIMITATIONS

Achieving optimal results requires sampling image pairs that have complementary features but maintain certain shared characteristics. When images with excessive dissimilarity are chosen, LLMs tend to falter, often generating descriptions that encompass both images rather than high-quality reformulation text prompts. However, such extreme scenarios are uncommon in CIR, where reference and target images usually differ only at a fine-grained level. Experiments show that these dissimilarities are tolerable as the training scale increases, without causing significant negative impacts on the model.

## VI. CONCLUSION

This work explores achieving zero-shot composed image retrieval based solely on an unannotated image collection. Our approach involves transforming images into detailed captions and then generating reformulated text prompts within the text domain. With the support of advanced large language models, this data-creating approach is not only efficient but also scalable, without requiring pre-existing caption data. Leveraging this scalable pipeline, InstructCIR reaches new state-of-the-art in the zero-shot domain. Furthermore, our proposed embedding reformulation network also attains stateof-the-art results in supervised benchmarks, underscoring the efficacy of our design.

#### REFERENCES

- [1] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Selfsupervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023. 3, 6
- [2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. arXiv preprint arXiv:1607.06450, 2016. 3
- [3] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. arXiv preprint arXiv:2303.15247, 2023. 1, 4, 5
- [4] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Conditioned and composed image retrieval combining and partially fine-tuning clip-based features. In CVPR Workshops, 2022. 5
- [5] A. Baldrati, M. Bertini, T. Uricchio, and A. Del Bimbo. Conditioned and composed image retrieval combining and partially fine-tuning clip-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022. 7
- [6] A. Baldrati, M. Bertini, T. Uricchio, and A. Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022. 1, 4
- [7] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In CVPR, pages 21466–21474, 2022.
- [8] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. arXiv preprint arXiv:2211.09800, 2022. 3, 5
- [9] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv preprint arXiv:2303.12712, 2023. 1
- [10] Y. Chen, S. Gong, and L. Bazzani. Image search with text feedback by visiolinguistic attention learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1, 4
- [11] Niv Cohen, Rinon Gal, Eli A. Meirom, Gal Chechik, and Yuval Atzmon. "this is my unicorn, fluffy": Personalizing frozen visionlanguage representations. In ECCV, 2022. 5
- [12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018. 4
- [13] E. Dodds, J. Culpepper, S. Herdade, Y. Zhang, and K. Boakye. Modalityagnostic attention fusion for visual search with text feedback. arXiv preprint arXiv:2007.00145, 2020. 1
- [14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 3
- [15] Yiqun Duan, Zhen Wang, Yi Li, and Jingya Wang. Cross-domain multi-style merge for image captioning. *Computer Vision and Image Understanding*, 228:103617, 2023. 4
- [16] Yiqun Duan, Zhen Wang, Jingya Wang, Yu-Kai Wang, and Chin-Teng Lin. Position-aware image captioning with spatial relation. *Neurocomputing*, 497:28–38, 2022. 4
- [17] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618, 2022. 1
- [18] Sonam Goenka, Zhaoheng Zheng, Ayush Jaiswal, Rakesh Chada, Yue Wu, Varsha Hedau, and Pradeep Natarajan. Fashionvlp: Vision language transformer for fashion retrieval with feedback. In CVPR, pages 14105– 14115, 2022. 4
- [19] 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. 4, 5
- [20] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. Automatic spatially-aware

fashion concept discovery. In *Proceedings of the IEEE international conference on computer vision*, pages 1463–1471, 2017. 5
[21] X. Han, Z. Wu, P. X. Huang, X. Zhang, M. Zhu, Y. Li, Y. Zhao, and

- [21] X. Han, Z. Wu, P. X. Huang, X. Zhang, M. Zhu, Y. Li, Y. Zhao, and L. S. Davis. Automatic spatially-aware fashion concept discovery. In *IEEE International Conference on Computer Vision*, 2017. 4
- [22] Xiao Han, Licheng Yu, Xiatian Zhu, Li Zhang, Yi-Zhe Song, and Tao Xiang. Fashionvil: Fashion-focused vision-and-language representation learning. In ECCV, 2022. 4
- [23] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626, 2022. 4
- [24] Mehrdad Hosseinzadeh and Yang Wang. Composed query image retrieval using locally bounded features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3596– 3605, 2020. 1
- [25] 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. 2
- [26] Siddharth Krishna Kumar. On weight initialization in deep neural networks. arXiv preprint arXiv:1704.08863, 2017. 4
- [27] Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Data roaming and early fusion for composed image retrieval. arXiv preprint arXiv:2303.09429, 2023. 4, 5
- [28] 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. 1
- [29] Yi Li, Hualiang Wang, Yiqun Duan, and Xiaomeng Li. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv e-prints*, pages arXiv–2304, 2023. 4
- [30] Yi Li, Hualiang Wang, Yiqun Duan, Hang Xu, and Xiaomeng Li. Exploring visual interpretability for contrastive language-image pretraining. arXiv preprint arXiv:2209.07046, 2022. 4
- [31] Yi Li, Hualiang Wang, Yiqun Duan, Jiheng Zhang, and Xiaomeng Li. A closer look at the explainability of contrastive language-image pretraining. *Pattern Recognition*, page 111409, 2025. 4
- [32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 5
  [33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual
- [33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 1, 2
  [34] Yikun Liu, Jiangchao Yao, Ya Zhang, Yanfeng Wang, and Weidi
- [34] Yikun Liu, Jiangchao Yao, Ya Zhang, Yanfeng Wang, and Weidi Xie. Zero-shot composed text-image retrieval. arXiv preprint arXiv:2306.07272, 2023. 1, 4, 5, 6
- [35] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained visionand-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2125–2134, 2021. 1, 4, 5, 6, 7
- [36] Zheyuan Liu, Weixuan Sun, Yicong Hong, Damien Teney, and Stephen Gould. Bi-directional training for composed image retrieval via text prompt learning. arXiv preprint arXiv:2303.16604, 2023. 4, 5, 7
- prompt learning. arXiv preprint arXiv:2303.16604, 2023. 4, 5, 7
  [37] Zheyuan Liu, Weixuan Sun, Damien Teney, and Stephen Gould. Candidate set re-ranking for composed image retrieval with dual multi-modal encoder. arXiv preprint arXiv:2305.16304, 2023. 4, 5
- [38] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017. 4
- [39] 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, 3, 5
- [40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 10684–10695, 2022. 1, 3
- [41] 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, 4, 5
- [42] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale

dataset for training next generation image-text models. *arXiv preprint* arXiv:2210.08402, 2022. 4

- [43] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114, 2021. 5
- [44] 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, 2018. 5
- [45] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 217–223, 2017. 4, 5
- [46] 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, volume 38, pages 5270–5279, 2024. 5
- [47] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6439–6448, 2019. 1, 4
  [48] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie,
- [48] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11307–11317, 2021. 4, 5, 6, 7
- [49] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. arXiv preprint arXiv:2302.05543, 2023.