

Scale Up Composed Image Retrieval Learning via Modification Text Generation

Yinan Zhou, Yaxiong Wang*, Haokun Lin, Chen Ma, Li Zhu*, Zhen Dong Zheng

Abstract—Composed Image Retrieval (CIR) aims to search an image of interest using a combination of a reference image and modification text as the query. Despite recent advancements, this task remains challenging due to limited training data and laborious triplet annotation processes. To address this issue, this paper proposes to synthesize the training triplets to augment the training resource for the CIR problem. Specifically, we commence by training a modification text generator exploiting large-scale multimodal models and scale up the CIR learning throughout both the pretraining and fine-tuning stages. During pretraining, we leverage the trained generator to directly create Modification Text-oriented Synthetic Triplets (MTST) conditioned on pairs of images. For fine-tuning, we first synthesize reverse modification text to connect the target image back to the reference image. Subsequently, we devise a two-hop alignment strategy to incrementally close the semantic gap between the multimodal pair and the target image. We initially learn an implicit prototype utilizing both the original triplet and its reversed version in a cycle manner, followed by combining the implicit prototype feature with the modification text to facilitate accurate alignment with the target image. Extensive experiments validate the efficacy of the generated triplets and confirm that our proposed methodology attains competitive recall on both the CIRR and FashionIQ benchmarks. Codes and datasets will be made publicly accessible.

Index Terms—Composed image retrieval, Text generation, Metric learning, Information retrieval.

I. INTRODUCTION

IN the context of Composed Image Retrieval (CIR), a given reference image and the modification text (also known as modifier) are utilized to amalgamate information across both visual and textual modalities, to pinpoint the most congruous

* indicates corresponding authors.

Y. Zhou and L. Zhu are with the School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: zyn13572297710@stu.xjtu.edu.cn; zhuli@mail.xjtu.edu.cn).

Y. Wang is with the School of Electronics and Information Engineering, Hefei University of Technology, Hefei 230009, China (e-mail: wangyx15@stu.xjtu.edu.cn).

H. Lin is with the School of Artificial Intelligence, University of the Chinese Academy of Sciences, Beijing 101408, China (e-mail: haokun.lin@cripac.ia.ac.cn).

C. Ma, H. Lin and Y. Zhou are with the Department of Computer Science, City University of Hong Kong, Hong Kong 999077, China (e-mail: chenma@cityu.edu.hk; haokun.lin@cripac.ia.ac.cn; zyn13572297710@stu.xjtu.edu.cn).

Z. Zheng is with Faculty of Science and Technology, and Institute of Collaborative Innovation, University of Macau, Macau 999078, China (e-mail: zhedongzheng@um.edu.mo).

The paper is supported by the NSFC project under grant No. 62302140, the Fundamental Research Funds for the Central Universities (Academic Newcomer Support Program of Hefei University of Technology with project No. JZ2024HGTB0261), and University of Macau Start-up Research Grant SRG2024-00002-FST and Multi-Year Research Grant MYRG-GRG2024-00077-FST-UMDF.

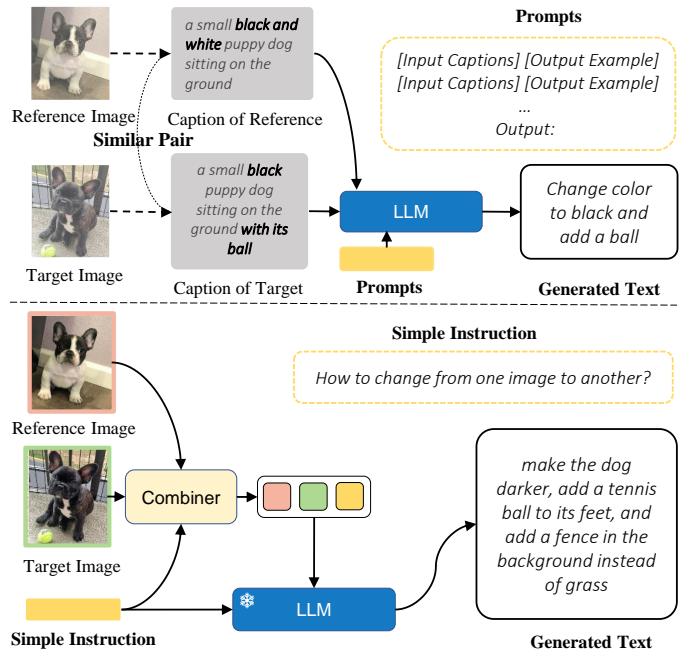


Fig. 1: Comparison of existing automatic modifier generation method and ours. Above: Existing methods utilize indirect information, such as labels and captions of images to generate modifiers, which often results in poorer notation quality, limited text length, and a lack of diversity in textual form. Below: Our proposed method combines the image pair and instruction, mapping them into a frozen LLM for text generation. This allows for a more flexible description of the details between images and generates high-quality modification text with controllable length.

target image within an image gallery. In contrast to the conventional image retrieval tasks reliant solely on textual information [1], [2] or tag information [3], the CIR model demands superior feature extraction, fusion, and inference capabilities. Images encapsulate rich visual information and intuitive perceptions, while text provides precise descriptions and semantic understanding of image content. The fusion of these two modalities for retrieval purposes can support the identification of the target image more accurately. As a result, CIR is regarded as a meaningful and promising research area and extensive efforts have been dedicated to this task [4]–[12]. Common practices typically utilize the well-annotated triplets to train the models. However, the annotation of triplets requires remarkable human efforts and triplets cannot be collected from social web-like image-text pairs.

Motivated by the achievements of AIGC [13]–[21], a

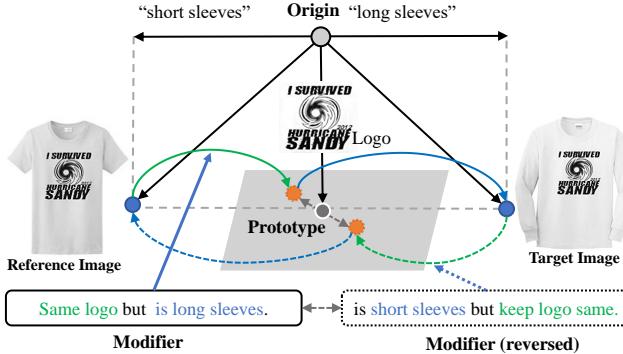


Fig. 2: **An intuitive prototype example** of the reference image’s prototype reached by the modifier and the target image’s prototype reached by the reversed modifier. The green dashed line represents the content preserved in the image based on the corresponding modification text. The blue dashed line represents the modifications and additions made to the prototype based on the corresponding modified text.

straightforward solution for this issue is the synthesis of triplets. Predominantly, the procedure can be bifurcated into two methodologies: (1) One might resort to leveraging image editing models [16] to manipulate the reference images based on the provided modification text. However, there is a domain gap between the real image and the synthesized one. (2) Some alternative approaches [22]–[24] utilize labels or captions of two related images and combine that information with some delicate prompts or templates as input to large language models to generate modification texts indirectly (*see Figure 1*). This approach often results in poorer notation quality, limited text length, and a lack of diversity in textual form.

With the above considerations, we resort to generating the modification text using the reference and target image as input. This solution, on the one hand, produces the textual modality, whose gap is smaller than the visual modality. On the other hand, sufficient triplets can be generated by simply feeding two images, which is similar to the manual labeling process. To generate reliable modification text, we leverage the annotated triplets as the training source and train a modification text generator by tuning the multimodal large-scale models. Subsequently, we mine a substantial number of relevant image pairs using similar labels or image sets. With the trained generator and image pairs, we can freely generate the Modification Text-oriented Synthetic Triplets (MTST) to augment the original benchmarks, supporting a large scale pretraining. The final model can be harvested by following the standard CIR learning protocol for fine-tuning, which merges query pairs and aligns the queries with target images.

Nonetheless, a key distinction lies in the nature of the modification text in CIR, which acts less as a content descriptor and more akin to a two-part instructional guide. The information in the modification text for one reference image can be divided into two parts: one part pertains to the elements or features of the reference image that need to be preserved, which are usually expressed implicitly, and the other part includes new additions or changes. Figure 2 shows one example to illustrate this characteristic of the modifier. If we start from the origin

and add two orthogonal pieces of information, [logo] and “short sleeves” result in the reference image, while [logo] and “long sleeves” point to the target image. Therefore, the modifier, “Same logo but is long sleeves”, which transitions from the reference to the target image, can be decoupled into two steps. The first step is to retain identical or similar information from the reference image, corresponding to “same logo”, along with other implicitly unmentioned details such as the collar style and color. We denote these preserved traits as the implicit prototype, generated based on the reference image and the corresponding modification text. The second step combines the implicit prototype with new or altered content from the modifier, such as “longer sleeves”, to form a composite query to retrieve the most suitable target image.

Motivated by this observation, we design a Prototypical Two-Hop Alignment (PTHA) strategy to progressively bridge the semantic gap between the multi-modal query and the target image. In particular, PTHA decouples the alignment into two steps: the first step generates the reverse modifier using the MTST generator and learns the implicit prototype preserving the sharing clues, while the second step combines the implicit prototype with the modifier to align with the target image. In a nutshell, the contribution can be summarized as follows:

- Considering the triplet scarcity in composed image retrieval (CIR), we contribute a generation framework of Modification Text-oriented Synthetic Triplets (MTST) to augment the existing benchmarks with high-quality synthetic triplets, supporting effective pre-training for CIR.
- We build two large-scale pretraining datasets for nature and fashion domains with our trained modification text generator, containing 800K and 580K triplets with expressive modification texts.
- A Prototypical Two-Hop Alignment (PTHA) strategy is proposed, which decouples the CIR problem as a two-step alignment paradigm to gradually bridge the gap between the multimodal query and the target image.
- Benefiting from the generated high-quality triplets and our devised PTHA network, we achieve comparable results, with improvements of +2.39 in Avg. on the CIRR benchmark from nature and +1.57 in Avg. on the FashionIQ benchmark.

II. RELATED WORK

A. Vision Language Models

In recent years, significant progress has been made in the foundational research of Vision Language Models (VLMs). Typical models like CLIP [25] and ALIGN [26] achieve cross-modal understanding by leveraging contrastive learning on large-scale image and text pairs. Li *et al.* [27] introduce image-text matching and masked language modeling (MLM) tasks during training to enhance fine-grained matching. BLIP [28] equip the pre-trained models with text generation capabilities by language modeling (LM). With a similar spirit, some recent works further fine-tune the cross-modality model for different downstream tasks, such as text-based person retrieval [29] and drone localization [30]. The emergence of various Large Language Models (LLMs) [15], [17], [18], [31]–[35] has also

influenced the development of visual language models, as they possess vast knowledge and powerful text generation capabilities. LLaVa [36] directly maps visual features to LLMs and aligns spaces through fintuning. BLIP2 [1] establishes a bridge between vision language base models and various open-source LLMs by deploying a Q-Former on filtered data. InstructBLIP [37] further improves performance using instruction tuning and exhibits enhanced text generation capabilities while reducing training costs through LLM freezing. We deploy instruction tuning to composed image retrieval and make it possible to generate modifiers using related images.

B. Composed Image Retrieval

Image retrieval is an important research task in multi-modal field. It aims to retrieve target images from a gallery based on a given query. This can be done by solely using text descriptions or by using images [38], [39] to retrieve similar or related images. However, single-modal tasks such as text-based image retrieval or image-based image retrieval cannot accurately and conveniently meet the specific retrieval needs of certain scenarios. To address this issue, the Composed Image Retrieval task has been proposed [4], [5], [7], [8], [40], which involves integrating the reference image feature and supplementing or modifying the textual feature to retrieve the target image. There have been efforts in training lightweight connection layers to obtain fused features from image and text representations. ARTEMIS [41] combines triplets through explicit matching and implicit similarity, and Baldrati *et al.* [42] proposes a combiner to leverage CLIP visual and textual representations. Liu *et al.* [43] proposes a re-ranking method after the first selection. In very recent works, many existing works [44]–[48] leverage large amounts of external data to achieve zero-shot CIR capabilities. MagicLens [48] achieves strong performance in zero-shot CIR while also making progress in richer relations beyond image similarity. SPRC [49] utilizes Q-Former to extract sentence-level prompts and guides sentence-level prompt generation aligned with an auxiliary text prompt. In addition, there have been works that enhance task performance by introducing additional datasets for pre-training [22]–[24]. These datasets provide extra training examples and diverse data distributions, allowing the models to learn more comprehensive and robust representations. In our work, we design a prototypical two-hop alignment network to decompose CIR into an implicit prototype learning module and fusion module. In the implicit prototype learning module, we utilize generated reversed modifiers to benefit implicit prototype learning.

C. Composed Image Retrieval Triplet Generation

In previous works, Composed Image Retrieval(CIR) triplet generation has been primarily achieved through manual and automatic methods: **Manual Annotated**. The CIRR dataset [4] consists of manually annotated textual triplets, which are derived from a subset of images from the NLVR2 dataset [50], representing real-world scenarios. The FashionIQ dataset [40] comprises manually selected pairs of similar fashion images,

along with human-annotated textual triplets, specifically curated for the fashion domain. **Automatic Annotated**. Han *et al.* [8] employed the differences in manually annotated attribute labels of fashion200k dataset images to generate modified texts using a triplet template. In recent years, there have been proposed methods that leverage automatic generation techniques from other tasks and models. LaSCo [24] utilized VQA 2.0 [51] to construct triplets by using different answers for similar images and the same question, employing GPT 3.0 [52], followed by manual quality control. CompoDiff [22] built triplets based on InstructPix2Pix [16], using text descriptions collected from both human annotators and large-scale model generation and generating images using Stable Diffusion [13]. CoVR [24] employed similar video captions to filter similar image pairs and trained an MTG-LLM to generate a modifier using two similar captions, forming triplets. Compared to existing methods, we incorporate images into the training of modifier generation and map visual features into the space of a large model. We propose a lightweight text generation method that is more flexible, diverse, and controllable in length while maintaining low training costs.

In previous works, CIR triplet generation has been primarily achieved through manual and automatic methods: **Manual Annotated**. The CIRR dataset [4] consists of manually annotated textual triplets, which are derived from a subset of images from the NLVR2 dataset [50], representing real-world scenarios. The FashionIQ dataset [40] comprises manually selected pairs of similar fashion images, along with human-annotated textual triplets, specifically curated for the fashion domain. **Automatic Annotated**. Han *et al.* [8] employed the differences in manually annotated attribute labels of fashion200k dataset images to generate modified texts using a triplet template. In recent years, there have been proposed methods that leverage automatic generation techniques from other tasks and models. LaSCo [24] utilized VQA 2.0 [51] to construct triplets by using different answers for similar images and the same question, employing GPT 3.0 [52], followed by manual quality control. CompoDiff [22] built triplets based on InstructPix2Pix [16], using text descriptions collected from both human annotators and large-scale model generation and generating images using Stable Diffusion [13]. CoVR [24] employed similar video captions to filter similar image pairs and trained an MTG-LLM to generate a modifier using two similar captions, forming triplets. Compared to existing methods, we incorporate images into the training of modifier generation and map visual features into the space of a large model. We propose a lightweight text generation method that is more flexible, diverse, and controllable in length while maintaining low training costs.

III. MODIFICATION TEXT-ORIENTED SYNTHETIC TRIPLETS (MTST) GENERATION

A. MTST Generator

1) Architecture: Modification Text-oriented Synthetic Triplets (MTST) generator takes the reference and target images as input and outputs the modification text. To produce high-quality text, we follow the multimodal large model [1],

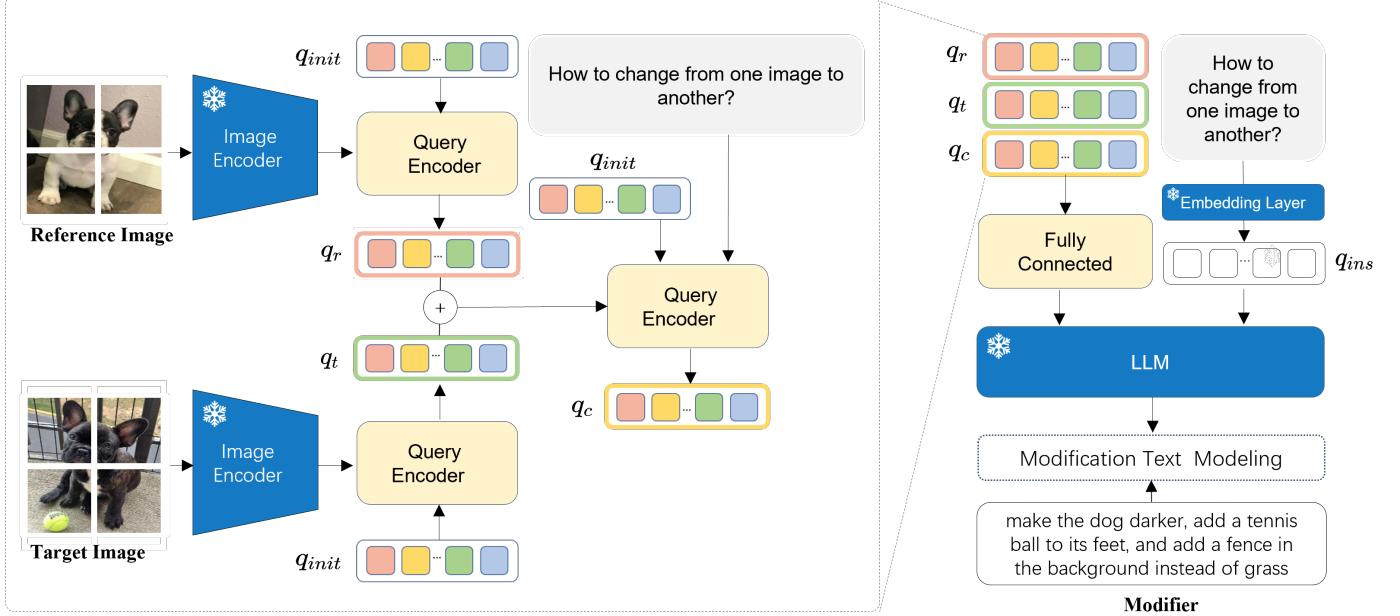


Fig. 3: An overview of MTST Generator architecture. The paired reference images and target images are fed into an Image Encoder and a trainable Query Encoder to get their respective features q_r, q_t . These representations are then concatenated and fused with the instruction using the same Query Encoder to obtain a fusion representation q_c . Representations q_r, q_t, q_c along with the instruction embedding are concatenated and fed into a frozen LLM to generate modification text.

[37] to design our MTST generator, *i.e.*, image encoder extracts the image features and a Large Language Model (LLM) is followed to give reliable text output. Figure 3 depicts the overall architecture, consisting of an Image Encoder and a Query Encoder for vision feature extraction, a Fully Connected Layer bridging the vision feature to the LLM, and a LLM as the text prediction module. We keep the Image Encoder and LLM frozen and only train the Query Encoder and Fully Connected Layer to map the learned prompt feature to LLM space.

2) *Generator Training:* The manually annotated (training) triplets in the standard benchmark like CIRR [4] are taken as the training samples. To guide the network learning, we also provide an instruction, “How to change from one image to another?”, which is incorporated into the Query Encoder to clarify the purpose. In specific, the reference image and target image are separately input to the Image Encoder and Query Encoder to get their respective query features q_r, q_t from initial query tokens q_{init} .

To acquire the task-oriented features, we next feed the instruction and vision features q_r, q_t into Query Encoder to obtain a composed representation q_c . Subsequently, we pass q_r, q_t , and q_c through the projection layer FC_{llm} to produce a comprehensive vision feature, which is then concatenated with the instruction embedding q_{ins} and fed into LLM for text generation:

$$\text{input}_{llm} = FC_{llm}(q_r \oplus q_t \oplus q_c) \oplus q_{ins}, \quad (1)$$

For modification text modeling, we deploy a language generative task auto-regressively to predict the next token of

the modification text by maximizing the conditional likelihood:

$$\mathcal{L}_{gen} = -\mathbb{E}_{(I_r, I_t, t) \sim \mathcal{D}} \left[\sum_{m=1}^M \log P(t_m | \text{input}_{llm}, t_0, \dots, t_{m-1}) \right], \quad (2)$$

where M denotes the length of modification text t and t_m denotes the m^{th} token of t . \mathcal{D} is the distribution of triplets. I_r and I_t denote the reference and the target image, respectively.

B. Grounded MTST Generation

In this paper, we focus on two common domains, *i.e.*, the nature and the fashion, for MTST generation. We take the popular CIRR and FahsionIQ benchmarks as the source data. To generate triplets resembling the real world, we carefully design the image sampling strategies for the triplet generation.

CIRR_{MTST}. We employ two strategies for selecting image pairs from the CIRR dataset: (1) The CIRR training dataset comprises 3,345 image sets, each featuring six analogous images, from which the training triplets are sampled. Pairwise combinations of images within these clusters yield a total of 100,350 unique pairs. (2) We create new image sets by combining images with the same category from the NLVR2 dataset [50], which is the source dataset for CIRR images. We then pair these newly created image sets, resulting in 707,745 image pairs. Therefore, by extending MTST on these image pairs, we generate a total of 808,095 triplets on the CIRR dataset.

FashionIQ_{MTST} For the FashionIQ dataset, each image has multiple labels that describe its style, such as ‘short sleeves’, ‘v-neck’, ‘hoodie’. We classify the images based on their labels, and images with the same label, share certain characteristics and similarities. From these images, we can

TABLE I: Statistics of existing CIR datasets and our generated dataset: We expand the triplets of CIRR and FashionIQ datasets using our MTST generator. The table compares the number of triplets, unique images, unique words, and the average length of modification text.

Name	Domain	Image Source	#Triplets	#Unique images	#Average length	#Unique Words
FashionIQ (train) [40] CIRR (train) [4]	Fashion Nature	FashionIQ (train) CIRR (train)	16,914 36,761	23,813 21,185	54.90 59.51	4,253 7,129
SynthTriplets 18M [22] LaSCo [23] WebVid-CoVR [24]	Nature Nature Nature	Synthetic VQA2.0 [51] WebVid2M, WebVid10M [53]	18,000,000 389,305 1,648,789	- 121,479 130,775	- 30.7 23.36	- 13,488 19,163
FashionIQ_{MTST} CIRR_{MTST}	Fashion Nature	FashionIQ [40] (train) NLVR2 [50] (train)	579,114 808,096	26,048 103,170	61.66 113.03	5,212 19,681

select the image pairs we need. However, some labels usually correspond to a large number of images. For example, the label “long sleeve” corresponds to 1,006 images. If we pair the images with same label all together, we would end up with 1,011,030 image pairs for only one category. This could result in weak image correlations and an imbalanced dataset. To address this, we impose a limit to the number of image pairs, ensuring it does not exceed three times the number of images in its respective category. As a result, we generated a total of 579,114 triplets from the dress, shirt, and top tee categories in FashionIQ. Table I presents the statistics of our final datasets and the comparison with the existing CIR datasets. Our proposed MTST exhibits several salient advantages:

- **Narrower domain gap to the annotated triplets.** By resorting to text generation, we effectively mitigate domain discrepancies compared to image generation methods. Furthermore, benefiting from the paradigm of directly utilizing real images as input for text generation, we bypass visual domain gaps inherent in other approaches.
- **Expressive modification text.** Leveraging the capabilities of Large Language Models (LLMs), MTST yields modification texts that are more verbose than those found in existing benchmarks. This characteristic allows for richer and more expressive content representation.
- **Greater flexibility in triplet generation.** The MTST framework necessitates only two images for generating corresponding text, thus demonstrating a high degree of flexibility. This enables large-scale triplet generation without compromising efficiency or diversity.

IV. PROTOTYPICAL TWO-HOP ALIGNMENT NETWORK

As illustrated in Figure 4, our prototypical two-hop alignment (PTHA) network comprises an image encoder and a text encoder for image and text feature extraction respectively, and a multimodal encoder to combine the multimodal query pair. During training, we first generate the reversed modifier using MTST and then apply the proposed PTHA to learn the network. When inference, we utilize the fusion feature to compute cosine similarity with the image feature extracted from candidates in the image gallery to perform retrieval.

A. Pre-training with MTST

Before optimizing our PTHA network, we first adopt MTST to perform pre-training, pursuing a better initialization for the subsequent learning. Formally, we first encode the reference

image I_r adopting a frozen image encoder \mathcal{E}_I , and fuse the resulted representation with modification text T_{r2t} with the multi-modal encoder \mathcal{E}_M :

$$f_{r2t} = \mathcal{E}_M(\mathcal{E}_I(I_r), T_{r2t}). \quad (3)$$

Following SPRC [49], we take the multimodal feature f_{r2t} as the textual prompt, which is then fed into the text encoder \mathcal{E}_T with modification text to produce the fusion feature of the reference image and modification text:

$$f_q = \mathcal{E}_T(f_{r2t}, T_{r2t}). \quad (4)$$

We utilize the same Image Encoder \mathcal{E}_I to encode target image I_t and the same multi-modal encoder \mathcal{E}_M to produce the target feature of query pair:

$$f_t = \mathcal{E}_M(\mathcal{E}_I(I_t)). \quad (5)$$

Subsequently, we deploy contrastive learning loss query feature f_q and target image feature f_t to train the network:

$$\mathcal{L}_{q2t} = -\frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} \log \frac{\exp(\text{sim}(f_{qi}, f_{ti})/\tau)}{\sum_{j=1}^{|\mathcal{B}|} \exp(\text{sim}(f_{qi}, f_{tj})/\tau)}, \quad (6)$$

where \mathcal{B} denotes the input batch, f_{qi} and f_{ti} denotes the i -th fusion feature and target feature in \mathcal{B} respectively, and τ is a learnable temperature parameter. For similarity calculation, we adopt the [CLS] token in f_q , i.e., f_{qcls} , to query the target image embeddings and take the max pool as a similarity estimation:

$$\text{sim}(f_q, f_t) = \max_{k \in [1, N]} \frac{f_{qcls} \cdot f_t[k]}{\|f_{qcls}\| \cdot \|f_t[k]\|}, \quad (7)$$

where $f_t[k]$ means the k -th token embedding in f_t .

B. Prototype-bridged Two-Hop Fine-Tuning

Motivated by the intuition in Figure 2, we perform a two-hop alignment strategy during the fine-tuning phase. **Implicit Prototype Learning via Reverse Text.** In the first step, we target to learn the implicit prototype preserving the shared information between the images. Particularly, we force the two images to approach each other via the respective text guidance, to reach a feature (implicit prototype) containing shared information in the two images. To support this, we first synthesize the modifiers from the target image to the reference image adopting the trained MTST generator.

In specific, let T_{t2r} be the reverse modification text, then similar to Eq. 3, the representation f_{t2r} of the reversed pair

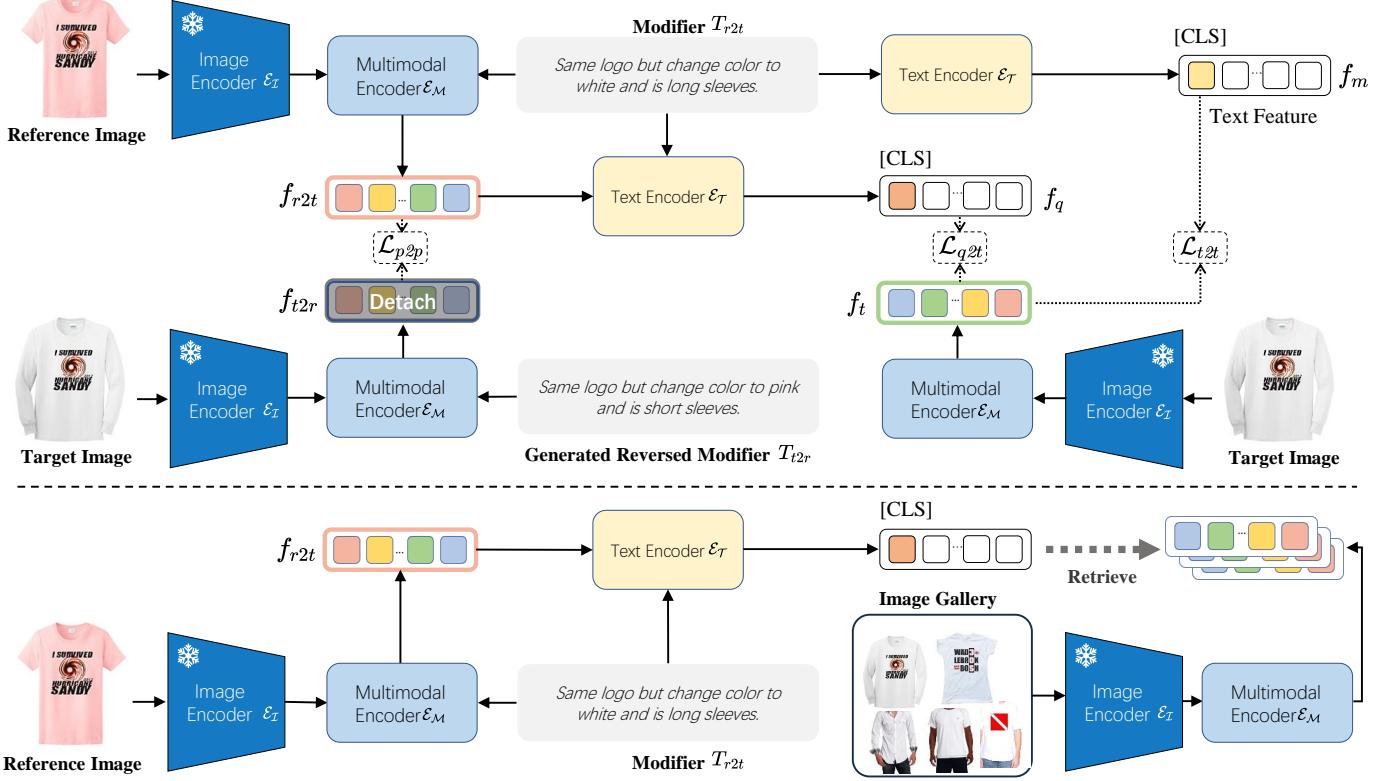


Fig. 4: **Top:** Overview of the training phase. We first employ implicit prototype learning between f_{r2t} and detached f_{t2r} . f_{t2r} is extracted from the target image and generated reversed text using the MTST generator. In the second step, we utilize contrastive loss between fusion feature f_q and target image feature f_t , text-only feature f_m , and target image feature f_t . **Bottom:** Overview of the inference phase. We leverage the fusion feature f_q to compute similarity with the features extracted from the image gallery to perform retrieval.

can be obtained from the target image I_t and the generated reverse modification text T_{t2r} .

Subsequently, we learn the implicit prototype by making the two multi-modal features step towards each other:

$$\mathcal{L}_{p2p} = \frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} (\bar{f}_{r2t_i} - \bar{f}_{t2r_i})^2, \quad (8)$$

where \bar{f} denotes the mean of N tokens of feature f . During training, we stop the gradient propagation of f_{t2r} . The gradient update is halted for the reversed multimodal pair due to two considerations: (1) The reversed text is only generated during training for L_{p2p} calculation, consequently, we maintain the integrity of implicit prototype details in the real query pair's features to align across both training and inference stages; (2) Mitigating the likelihood of model degeneration or collapse.

Implicit Prototype-bridge Alignment. In the second step, we further fuse the learned implicit prototype f_{r2t} with the modifier T to combine the necessary modifications or additions, yielding the composite feature f_q for retrieval, which is then aligned with the target image with a contrastive procedure similar to eq. (6). Besides, following SPRC [49], we also align the modification text and the target image as an auxiliary constraint \mathcal{L}_{t2t} , which matches the content in the modification text feature f_m to the target image feature f_t to aid the learning of dominated constraints, on the other hand, narrows the semantic gap between the modification text and the target

image, such that ease the alignment of query pair and target image in feature space. The constraint is also specified as a contrastive procedure similar to eq. (6) with the features of the target image and the modification text. Our final objective is formulated as:

$$\mathcal{L} = \mathcal{L}_{q2t} + \mathcal{L}_{t2t} + \alpha \mathcal{L}_{p2p}, \quad (9)$$

where α is a non-negative trade-off hyper-parameter.

V. EXPERIMENT

A. Experimental Setup

1) *Datasets and Evaluation Metrics:* Following previous work [42], [49], [54], [55], the real-world dataset CIRR and the fashion domain dataset FashionIQ are considered. Both are manually annotated CIR datasets based on real images. CIRR is created from 21k real-world images from NLVR2 [50], producing 36k triples. FashionIQ primarily contains images from three categories in the fashion domain: dress, shirt, and toptee. It comprises a total of 77k fashion product images and 30k triplets, each pair of images contains two different relative captions annotated by two individuals. Following the metric settings of the standard evaluation experiment, we report the average recall at rank K ($R@K$). For FashionIQ, we report the $R@10$ and $R@50$ on the val set in three categories. For CIRR, we disclose the $R@1$, 5 , 10 , and 50 as well as $Recall_{subset}@1$,

TABLE II: Comparison on CIRR test set. "Pretraining data" is the dataset for model pretraining. "Avg." means $(\text{Recall}@5 + \text{Recall}_{\text{subset}}@1)/2$. The best result is indicated in **bold**, while the second best is underlined. Our proposed PTHA with pretraining on CIRRMST outperforms the previous method in all metrics. * indicates that the method deploys an extra re-ranking strategy. † indicates the utilizing of extra training method SPN4CIR [55].

Method	backbone	Pretraining Data	Recall@K				Recall _{subset} @K			Avg.
			K=1	K=5	K=10	K=50	K=1	K=2	K=3	
MAAF [57]	w/o VLP	-	10.31	33.03	48.30	80.06	21.05	41.81	61.60	27.04
TIRG [5]	w/o VLP	-	14.61	48.37	64.08	90.03	22.67	44.97	65.14	35.52
ARTEMIS [41]	w/o VLP	-	16.96	46.10	61.31	87.73	39.99	62.20	75.67	43.05
CIRPLANT w/OSCAR [4]	w/o VLP	-	19.55	52.55	68.39	92.38	39.20	63.03	79.49	45.88
ComqueryFormer [11]	CLIP	-	25.76	61.76	75.90	95.13	51.86	76.26	89.25	56.81
NSFSE [12]	CLIP	-	20.70	52.50	67.96	90.74	44.20	65.53	78.50	48.35
Compdiff [22]	CLIP	SynthTriplets	22.35	54.36	73.41	91.77	35.84	56.11	76.60	29.10
CLIP4CIR [42]	CLIP	-	38.53	69.98	81.86	95.93	68.19	85.64	94.17	69.09
BLIP4CIR+Bi [54]	BLIP	-	40.15	73.08	83.88	96.27	72.10	88.27	95.93	72.59
CASE [23]	BLIP	LaSCo	49.35	80.02	88.75	97.47	76.48	90.37	95.71	78.25
CoVR-BLIP [24]	BLIP	WebVid-CoVR	49.69	78.60	86.77	94.31	75.01	88.12	93.16	80.81
Reranking* [43]	BLIP	-	50.55	81.75	89.78	97.18	80.04	91.90	96.58	80.90
SPRC [49]	BLIP-2	-	51.96	82.12	89.74	97.69	80.65	92.31	96.60	81.39
SPRC ² * [49]	BLIP-2	-	54.15	83.01	90.39	98.17	82.31	92.68	96.87	82.66
SPRC† [49] [55]	BLIP-2	-	<u>55.06</u>	83.83	90.87	<u>98.29</u>	81.54	92.65	97.04	82.69
Baseline	BLIP-2	-	51.39	81.95	89.92	97.90	78.98	91.78	96.36	80.46
PTHA (Ours)	BLIP-2	-	51.85	82.1	89.93	97.98	80.32	92.36	96.70	81.21
PTHA (Ours)	BLIP-2	CIRRMST	54.70	<u>84.05</u>	<u>90.89</u>	98.26	81.64	<u>93.30</u>	<u>97.30</u>	<u>82.85</u>
PTHA (Ours)† [55]	BLIP-2	CIRRMST	56.43	84.92	91.74	98.43	<u>82.12</u>	<u>93.35</u>	<u>97.42</u>	83.52

2, 3 on test set. Recall_{subset}@K serves as a benchmark for fine-grained matching, considering each image set composed of 6 similar images as the search space.

2) *Implementation Details: MTST Generator.* The MTST Generator is fine-tuned on the CIR task based on the InstructBlip-Vicuna-7b [37] base model, with only the Q-Former fine-tuned. The number of query tokens in the Q-Former [1] is set to 32. The frozen vicuna checkpoint is specified as vicuna-7b-v1.1 [32]. The model weights are loaded from a pre-trained model available from InstructBLIP [37].

PTHA. The model architecture in PTHA is the same as SPRC [49]. The multimodal encoder is the query encoder in BLIP-2 [1], the text encoder is the text encoder in BLIP-2. The vision encoder is the frozen ViT-g/14 from EVA-CLIP [56]. We employ two-stage strategy. The batch size is set to 128 and the number N of query tokens is set to 32 for both stages. The initial model weights of pretraining is from BLIP2-pretrained [1]. In the fine-tuning stage, we follow all the settings of SPRC [49]. During our finetuning stage on CIRR after pretraining on CIRRMST, we replace the [CLS] token $f_{q_{cls}}$ with the average embedding of N query tokens and the [CLS] token in f_q , i.e., $f_{q_{avg}}$ in eq. (7).

B. Quantitative Results

1) *CIRR.*: Table II shows the comparison result on CIRR. It is worth mentioning that the results of pretraining on our CIRRMST and fine-tuning using our proposed PTHA framework outperforms all existing methods across all metrics. Compared to the SOTA model without an extra re-ranking strategy, i.e., SPRC, we achieve improvements of **+2.74**, **+1.93**, **+1.15**, and **+0.57** in Recall@1, 5, 10, and 50, respectively, and obtain an overall average improvement of **+1.46**. Additionally,

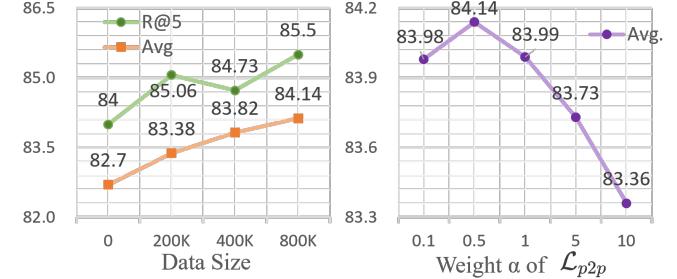


Fig. 5: Left: Ablation studies on different pre-training CIRRMST data size. We report Recall@5 and Avg. metric on CIRR validation set by fine-tuning with eq. (9). Right: We deploy different \mathcal{L}_{p2p} weight α on fine-tuning stage with the same pre-trained model.

we further fine-tuned our method using SPN4CIR [55], achieving better performance. Compared to SPRC with SPN4CIR [55], we achieve an improvement of +0.83 on average.

2) *FashionIQ.*: We then evaluate our FashionIQ_{MST} and PTHA on FashionIQ. As shown in Table III, we observe a similar upward trend in performance improvement as it in CIRR, indicating consistent progress. Apart from the R@10 metric on Toptee, we surpass the second-best method, i.e., SPRC in all other metrics, achieving an average improvement of **+0.79**. It indicates that our method and pretraining strategy are similarly applicable in the FashionIQ dataset.

C. Ablation Studies

1) *Effect of PTHA Learning.*: Table IVa presents an analysis of the loss terms employed in PTHA learning, elucidating their contributions. Results from both the pretraining and finetuning phases confirm the effectiveness of all three losses. For instance, using only the query-to-target contrastive loss

TABLE III: Comparison on FashionIQ validation set. “Avg.” means (Recall@10 + Recall@50)/2. * indicates that the method deploys extra re-ranking strategy.

Method	baseline	Pretraining Data	Dress		Shirt		Toptee		Average		Avg.
			R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	
TIRG [5]	w/o VLP	-	14.87	34.66	18.26	37.89	19.08	39.62	17.40	37.39	27.45
CIRPLANT w/OSCAR [4]	w/o VLP	-	17.45	40.41	17.53	38.81	61.64	45.38	18.87	41.53	30.20
MAAF [57]	w/o VLP	-	23.8	48.6	21.3	44.2	27.9	53.6	24.3	48.8	36.6
CurlingNet [58]	w/o VLP	-	26.15	53.24	21.45	44.56	30.12	55.23	25.90	51.01	34.36
CosMo [59]	w/o VLP	-	25.64	50.30	24.90	49.18	29.21	57.46	26.58	52.31	39.45
ARTEMIS [41]	w/o VLP	-	25.68	51.25	28.59	55.06	21.57	44.13	25.25	50.08	37.67
NSFSE [12]	CLIP	-	31.12	55.73	24.58	45.85	31.93	58.37	29.17	53.24	41.26
MUR [60]	CLIP	-	32.61	61.34	33.23	62.55	41.40	72.51	35.75	65.47	50.61
Css-Net [61]	CLIP	-	33.65	63.16	35.96	61.96	42.65	70.70	37.42	65.27	51.35
CLIP4CIR [42]	CLIP	-	33.81	59.40	39.99	60.45	41.41	65.37	38.82	61.74	50.03
ComqueryFormer [11]	CLIP	-	33.86	61.08	35.57	62.19	42.07	69.30	37.17	64.19	50.68
CompoDiff [22]	CLIP	SynthTriplets	40.88	53.06	35.53	49.56	41.15	54.12	39.05	52.34	46.31
FAME-VIL [62]	CLIP	-	42.19	67.38	47.64	68.79	50.69	73.07	46.84	69.75	58.29
BLIP4CIR+Bi [54]	BLIP	-	42.09	67.33	41.76	64.28	46.61	70.32	43.49	67.31	55.40
CoVR-BLIP [24]	BLIP	WebVid-CoVR	44.55	69.03	48.43	67.42	52.60	74.31	48.53	70.25	59.39
CASE [23]	BLIP	LaSCo	47.77	69.36	48.48	70.23	50.18	72.24	48.79	70.68	59.74
Ranking* [43]	BLIP	-	48.14	71.34	50.15	71.25	55.23	76.80	51.17	73.13	62.15
SPRC [49]	BLIP2	-	49.18	72.43	55.64	73.89	59.35	78.58	54.92	74.97	64.85
Baseline	BLIP2	-	48.04	72.65	53.54	73.91	57.37	78.85	52.98	75.13	64.07
PTHA (Ours)	BLIP2	-	49.54	72.81	55.50	73.97	57.96	78.96	54.33	75.24	64.79
PTHA (Ours)	BLIP2	FashionIQ_{MTST}	50.77	73.78	55.91	75.36	58.28	79.70	54.99	76.28	65.64

(\mathcal{L}_{q2t}) without pretraining yields an average performance of 81.64. The inclusion of the text-to-target image term (\mathcal{L}_{12t}) enhances this to 82.15. The simultaneous deployment of all three losses optimizes performance to its peak. Under the cases of learning with pretraining, the contribution of each loss is sufficiently validated, and simultaneously applying three losses promotes our performance to the new state-of-the-art.

2) *Effect of MTST pretraining:* By comparing the results of the cases with and without pretraining, it can be observed that regardless of the type of the applied loss function, the models that have undergone pretraining consistently exhibit significant performance improvement. In the application of four different combinations of loss functions, the pretraining respectively leads to an improvement of 1.52, 1.57, 1.86, and 1.43 in “Avg.”. The effect of pretraining on enhancing Recall@K is notably substantial. We also give a comprehensive discussion in terms of the size and the image source of pretraining MTST. The results are reported in Figure 5, left. As the size of the data increases from 0 to 800K, the performance of the model shows an upward trend.

D. MTST Assessment

1) *Scalability of MTST Pre-training:* To further validate the effectiveness of MTST Pre-training, we pre-train four more methods [5], [41], [42], [49] using CIRR_{MTST}700K, which doesn’t contain the image pairs from the original image set. The results are reported in Tab.IVb. MTST pre-training brings clear performance improvements across all four methods, especially TIRG [5] and ARTEMIS [41].

2) *Generated Data Quality:* As shown in Table IVd, by directly using our CIRR_{MTST} for training without further finetuning, and excluding image pairs from the CIRR training set, we demonstrate competitive results directly on the validation set. This suggests a high similarity between CIRR_{MTST} and those of CIRR. To further assess the quality of generated data,

we generate 4K modifiers from the CIRR validation image pairs. Our evaluation includes 3 parts:

(1) Direct evaluation. The generated modifier text can achieve a 0.25 ROUGE-1 Score and 0.19 METEOR Score. However, due to the modifiers’ diversity, we affirm these metrics don’t fully measure the quality.

(2) Indirect Evaluation. Substituting generated text for CIRR validation leads to much higher scores, highlighting our data’s robust quality and pre-training suitability (see Table IVc).

(3) User study. We ask five experts to assess 100 randomly selected image pairs in CIRR_{MTST} to choose the better text between the real modifier and the generated modifier. We randomly shuffle the order of these two types of text. The findings suggest a comparable preference for generated text over real text, with favorable scores of 43%, 40%, 50%, 46%, and 45%, respectively. It is worth noting that we provide users with an option to report any factual errors in the two modification texts corresponding to each image pair, and we report the average error reporting rate, which is calculated as: the number of reported errors in the generated modification texts / (the total number of generated modification texts × the number of test users). The final average error reporting rate of generated text is only 4%.

3) *Comparison with other generated data:* We compare the gains of the CIRR_{MTST}700K vs. SynthTriplets18M [22] on ARTEMIS [41]. We separately utilize these two data to pretrain and CIRR training set to finetune. Figure 6 shows that with much less data, our performance enhancement on all markers is more comprehensive and significant than SynthTriplets18M.

E. PTHA Assessment

Our method shares the same baseline with [49]. We observe PTHA’s comparable performance in Table II, row 16. Notably, our PTHA, combined with pre-training, shows more than additive effectiveness, especially on Recall@1. When finetuned with an identical pre-trained model, PTHA outperforms

TABLE IV: Ablation studies. (a): Comparison of different loss combinations on Recall@5 and Recall_{subset}@1 metrics of CIRR validation set. “✓” denotes the loss in the column is applied. We report the results on CIRR validation set. (b): Performance comparison of four methods w/ and w/o pre-training using CIRR_{MTST}700k. MTST pre-training brings clear performance improvements across all four methods. (c): Comparison of the results using original text and generated text for triplets. (d): Comparison of the results of PTHA and SPRC [49] on CIRR val dataset. (e): We achieve admirable performance after first phase’s pre-training on CIRR_{MTST}700K by only utilizing a simple contrastive learning loss \mathcal{L}_{q2t} . (f):Zero-shot CIR performance comparison on CIRCO [54] test set.

(a) Ablation of Loss Functions and Pretraining

	Losses			Recall@5	Recall _{subset} @1	Avg.
	\mathcal{L}_{q2t}	\mathcal{L}_{t2t}	\mathcal{L}_{p2p}			
w/o pretrain	✓			83.76	79.52	81.64
	✓	✓		82.87	81.43	82.15
	✓		✓	83.74	80.27	82.00
	✓	✓	✓	84.0	81.39	82.70
w/ pretrain	✓			85.12	81.2	83.16
	✓	✓		85.52	81.92	83.72
	✓		✓	85.29	82.42	83.86
	✓	✓	✓	85.55	82.71	84.13

(c) Comparison of the results using original text and generated text for triplets.

Method	Modifier	Recall@K		Recall _{subset} @K		Avg.
		K=1	K=5	K=1	K=2	
PTHA	origin	56.80	85.55	82.71	94.00	84.13
PTHA	generated	69.17	93.63	88.14	96.22	90.89

(e) Two stages’ result comparison of PTHA and SPRC [49]

Method	Query	Recall@K		Recall _{subset} @K		Avg.
		K=1	K=5	K=1	K=2	
SPRC [49]	implicit prototype f_{r2t}	48.51	79.81	74.67	90.17	77.24
PTHA	implicit prototype f_{r2t}	51.21	82.09	76.27	90.74	79.18
SPRC [49]	final feature f_q	56.32	85.50	82.09	93.49	83.80
PTHA	final feature f_q	56.80	85.55	82.71	94.00	84.13

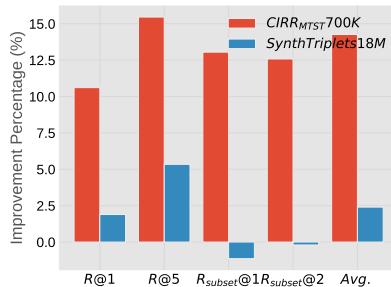


Fig. 6: Two pre-training data, CIRR_{MTST}700K and SynthTriplets18M benefit on ARTEMIS after same finetuning on CIRR. With much less data, CIRR_{MTST}700K’s performance enhancement on all markers is more comprehensive and significant than SynthTriplets18M’s.

SPRC on CIRR (See Table IVe, rows 3-4). This is due to the consistency of the generated data in the pre-training and fine-tuning stages, as well as the supervisory role of the \mathcal{L}_{p2p} . Further more, PTHA has a better intermediate feature quality. We directly utilize the features extracted solely through the multi-modal encoder (i.e., the **implicit prototype**, f_{r2t}) as the query for validation. The results, as shown in Table IVe, rows 1-2, clearly indicate that our first-phase feature quality

(b) Effectiveness of Pretraining

Method	Pre-train with CIRR _{MTST}	Recall@K		Recall _{subset} @K		Avg.
		K=1	K=5	K=1	K=2	
TIRG [5]	-	10.62	38.36	39.47	61.05	38.91
TIRG [5]	✓	18.60	53.54	51.58	72.61	52.56 ↑ 13.65
ARTEMIS [41]	-	17.47	47.31	40.70	61.91	44.00
ARTEMIS [41]	✓	28.08	62.77	53.75	74.49	58.26 ↑ 14.26
CLIP4CIR [42]	-	42.17	76.11	69.70	87.42	72.91
CLIP4CIR [42]	✓	44.69	77.57	71.80	88.14	74.68 ↑ 1.77
SPRC [49]	-	53.67	82.87	81.44	92.97	82.16
SPRC [49]	✓	55.30	85.05	81.46	93.22	83.31 ↑ 1.15

(d) Comparable performance after pretraining

CIRR	CIRR _{MTST}	Recall@K				Recall _{subset} @K	Avg.	
		K=1	K=5	K=10	K=50			
✓	✓	53.03	83.76	90.60	97.96	79.52	92.71	96.82
		51.28	79.86	87.75	97.22	77.01	91.05	96.17

(f) Zero-shot CIR performance on CIRCO [44] test set.

Arch	Pretraining Method	Finetuning Method	mAP@K		
			K=5	K=10	K=25
ViT-g/14	CompoDiff [22]	-	15.33	17.71	19.45
ViT-g/14	-	SPRC [49]	19.68	20.73	22.63
ViT-g/14	CIRR _{MTST}	PTHA	20.06	21.05	23.01
ViT-g/14	CIRR _{MTST}	-	22.8	24.02	26.17
ViT-g/14	LinCIR [45]	-	20.34	21.85	23.98
ViT-g/14	LDRE+IP-CIR [63]	-	32.75	34.26	36.86
CoCa-L	MagicLens-L [48]	-	34.1	35.4	38.1

is superior to SPRC’s.

We set the weight of \mathcal{L}_{q2t} to 1 and the weight of \mathcal{L}_{t2t} to 0.4 following SPRC [49]. We conduct experiment with different weight α of \mathcal{L}_{p2p} (See Figure 5, right) on pre-trained model using CIRR_{MTST}700k. Weighting \mathcal{L}_{p2p} with $\alpha = 0.5$ leads to the best results.

F. Zero-shot ability on CIRCO [44]

As shown in Table IVf, we compare the performance of our pre-trained model, fine-tuned model, SPRC [49], and other pre-trained models [22], [45], [48], [63] on CIRCO [44] in a zero-shot setting. It can be seen that our fine-tuned model outperforms SPRC, and our pre-trained model shows stronger generalization capabilities with higher zero-shot performance.

G. Qualitative Results

As shown in Figure 7, we qualitatively present the triplets generated in CIRR_{MTST} and FashionIQ_{MTST}. Particularly in CIRR_{MTST}, our generated modification text encompasses both a description of the target image and the changes observed by comparing the two images. We have observed that the model demonstrates strong descriptive capabilities in capturing changes in features such as color, condition, quantity, and the



Fig. 7: **Selected examples of generated triplets in CIRR_{MTST}(row 1-4) and FashionIQ_{MTST}(row5).** The **blue box** represents the reference image, while the **green box** indicates the target image. We leverage these two images as input to generate modified text. The **blue text** represents the information derived from the reference image, while the **green text** represents new additions or changes specific to the target image.

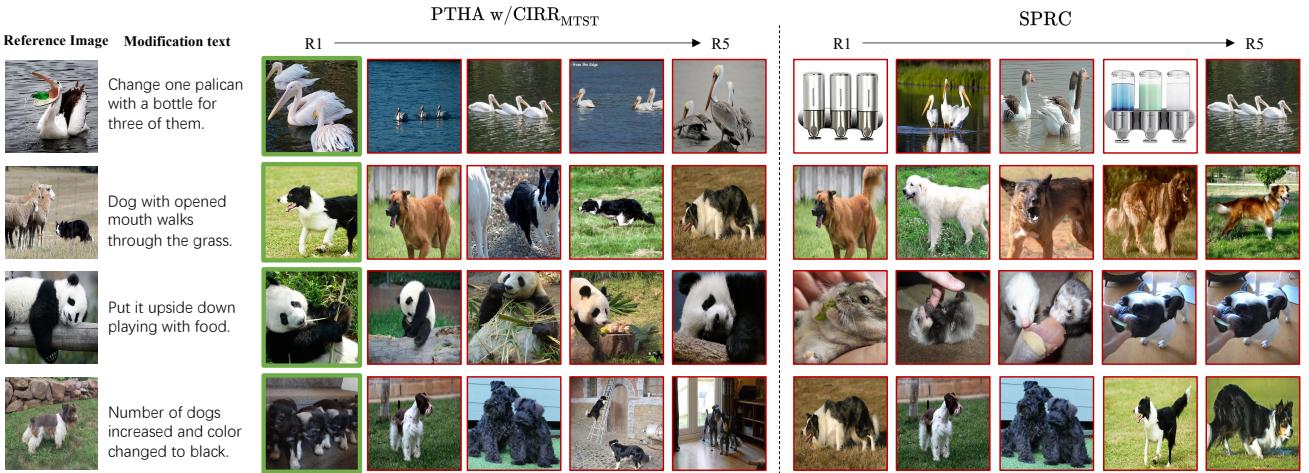


Fig. 8: Qualitative CIR results of our methods and SPRC, placed in descending order from right to left based on similarity. The **green boxes** indicate the correct matches, and the images in the **red boxes** are the wrong matches.

addition or removal of objects. Furthermore, we also compare the retrieval results of our method with the state-of-the-art approach, SPRC [49] on several examples (see Figure 8). We can observe that during the retrieval process, our method effectively preserves the relevant implicit prototypes of the reference image based on the modification text. The target images largely retain these implicit prototypes. For instance, in the first row, when the description might mislead the model, we accurately preserve the “pelican” prototype. In the second to fourth rows, we implicitly retain the characteristics of animals from the reference image.

VI. CONCLUSION

In this paper, we focus on alleviating the scarcity of training triplets in composed image retrieval. To this end, we train a modification text generator that produces synthetic, high-quality modification-oriented triplets. Our generator inputs

two images and outputs versatile, descriptive modifications to form realistic-like triples. With the trained generator, we benefit from the learning of CIR in both the pretraining and fine-tuning stages. In the pretraining stage, we generate the large-scale triplets to perform pretraining. In the fine-tuning stage, we first synthesize the reversed modification text, supporting us design a two-step alignment mechanism to gradually address the gap between the multimodal query and the target image. We first learn the implicit prototype with the real triplet and its reverse counterpart and combine the implicit prototype with the modification text to align with the target image. Extensive experimentation on benchmark datasets from both natural and fashion domains demonstrates that our method achieves a comparable performance with state-of-the-art approaches.

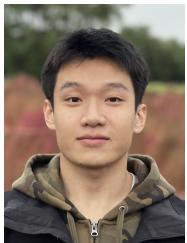
LIMITATIONS AND FUTURE WORK

Our work primarily offers a paradigm for expanding training sets on CIR. Focusing on improving performance on specific datasets, CIRR [4] and FashionIQ [40], the model is pre-trained on the extended data of such dataset. Therefore, training on specific domain results in insufficient effectiveness and generalization compared to previous methods that utilize large-scale triplet pre-training, as shown in zero-shot CIR performance on CIRCO [44] in Table IVf. Furthermore, applying MTST generation strategy to MLLMs [64], [65] that already possess the ability to distinguish between two images leads to a decrease in the model's generalization capability. Future work could benefit from employing powerful MLLMs to further explore a more rapid and efficient approach to domain adaptation, along with a method for generating text with higher quality and finer-grained modifications.

REFERENCES

- [1] J. Li, D. Li, S. Savarese, and S. C. H. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International Conference on Machine Learning*, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:256390509>
- [2] T. Karthikeyan, P. Manikandaprabhu, and S. Nithya, "A survey on text and content based image retrieval system for image mining," *International Journal of Engineering*, vol. 3, 2014.
- [3] D. Lu, X. Liu, and X. Qian, "Tag-based image search by social re-ranking," *IEEE Transactions on Multimedia*, vol. 18, no. 8, pp. 1628–1639, 2016.
- [4] Z. Liu, C. Rodriguez-Opazo, D. Teney, and S. Gould, "Image retrieval on real-life images with pre-trained vision-and-language models," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 2105–2114.
- [5] N. Vo, L. Jiang, C. Sun, K. Murphy, L.-J. Li, L. Fei-Fei, and J. Hays, "Composing text and image for image retrieval - an empirical odyssey," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2019. [Online]. Available: <http://dx.doi.org/10.1109/cvpr.2019.00660>
- [6] F. Zhang, M. Xu, Q. Mao, and C. Xu, "Joint attribute manipulation and modality alignment learning for composing text and image to image retrieval," in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 3367–3376. [Online]. Available: <https://doi.org/10.1145/3394171.3413917>
- [7] F. Zhang, M. Yan, J. Zhang, and C. Xu, "Comprehensive relationship reasoning for composed query based image retrieval," in *Proceedings of the 30th ACM International Conference on Multimedia*, ser. MM '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 4655–4664. [Online]. Available: <https://doi.org/10.1145/3503161.3548126>
- [8] 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 *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. [Online]. Available: <http://dx.doi.org/10.1109/iccv.2017.163>
- [9] H. Pang, S. Wei, G. Zhang, S. Zhang, S. Qiu, and Y. Zhao, "Heterogeneous feature alignment and fusion in cross-modal augmented space for composed image retrieval," *IEEE Transactions on Multimedia*, vol. 25, pp. 6446–6457, 2023.
- [10] G. Zhang, S. Wei, H. Pang, S. Qiu, and Y. Zhao, "Enhance composed image retrieval via multi-level collaborative localization and semantic activeness perception," *IEEE Transactions on Multimedia*, vol. 26, pp. 916–928, 2024.
- [11] Y. Xu, Y. Bin, J. Wei, Y. Yang, G. Wang, and H. T. Shen, "Multi-modal transformer with global-local alignment for composed query image retrieval," *IEEE Transactions on Multimedia*, vol. 25, pp. 8346–8357, 2023.
- [12] Y. Wang, L. Liu, C. Yuan, M. Li, and J. Liu, "Negative-sensitive framework with semantic enhancement for composed image retrieval," *IEEE Transactions on Multimedia*, vol. 26, pp. 7608–7621, 2024.
- [13] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2022. [Online]. Available: <http://dx.doi.org/10.1109/cvpr52688.2022.01042>
- [14] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *ArXiv*, vol. abs/2204.06125, 2022. [Online]. Available: <https://api.semanticscholar.org/CorpusID:248097655>
- [15] O. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H. Chung, D. Cummings, and J. Currier, "Gpt-4 technical report," Dec 2023.
- [16] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," *arXiv preprint arXiv:2211.09800*, 2022.
- [17] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.
- [18] Baichuan, "Baichuan 2: Open large-scale language models," *arXiv preprint arXiv:2309.10305*, 2023. [Online]. Available: <https://arxiv.org/abs/2309.10305>
- [19] Z. Li, Y. Guo, K. Wang, F. Liu, L. Nie, and M. S. Kankanhalli, "Learning to agree on vision attention for visual commonsense reasoning," *IEEE Transactions on Multimedia*, vol. 26, pp. 1065–1075, 2024.
- [20] Z. Li, Y. Guo, K. Wang, X. Chen, L. Nie, and M. S. Kankanhalli, "Do vision-language transformers exhibit visual commonsense? an empirical study of VCR," in *Proceedings of the 31st ACM International Conference on Multimedia*. ACM, 2023, pp. 5634–5644.
- [21] Z. Li, Y. Guo, K. Wang, Y. Wei, L. Nie, and M. S. Kankanhalli, "Joint answering and explanation for visual commonsense reasoning," *IEEE Transactions on Image Processing*, vol. 32, pp. 3836–3846, 2023.
- [22] G. Gu, S. Chun, W. Kim, H. Jun, Y. Kang, and S. Yun, "Compodiff: Versatile composed image retrieval with latent diffusion," *Transactions on Machine Learning Research*, 2024, expert Certification. [Online]. Available: <https://openreview.net/forum?id=mKtlzW0bWc>
- [23] M. Levy, R. Ben-Ari, N. Darshan, and D. Lischinski, "Data roaming and quality assessment for composed image retrieval," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 4, pp. 2991–2999, Mar. 2024. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/28081>
- [24] L. Ventura, A. Yang, C. Schmid, and G. Varol, "Covr: Learning composed video retrieval from web video captions," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, p. 5270–5279, Mar. 2024. [Online]. Available: <http://dx.doi.org/10.1609/aaai.v38i6.28334>
- [25] A. Radford, J. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Amanda, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," *Cornell University - arXiv; Cornell University - arXiv*, Feb 2021.
- [26] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International conference on machine learning*. PMLR, 2021, pp. 4904–4916.
- [27] J. Li, R. R. Selvaraju, A. D. Gotmare, S. Joty, C. Xiong, and S. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," in *NeurIPS*, 2021.
- [28] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 17–23 Jul 2022, pp. 12 888–12 900. [Online]. Available: <https://proceedings.mlr.press/v162/lj22n.html>
- [29] S. Yang, Y. Zhou, Z. Zheng, Y. Wang, L. Zhu, and Y. Wu, "Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 4492–4501.
- [30] M. Chu, Z. Zheng, W. Ji, T. Wang, and T.-S. Chua, "Towards natural language-guided drones: Geotext-1652 benchmark with spatial relation

- matching,” in *European Conference on Computer Vision*. Springer, 2025, pp. 213–231.
- [31] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, “Chain-of-thought prompting elicits reasoning in large language models,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022.
- [32] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, “Judging LLM-as-a-judge with MT-bench and chatbot arena,” in *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. [Online]. Available: <https://openreview.net/forum?id=uccHPGDIao>
- [33] G. Penedo, Q. Malartic, D. Hesslow, R. Cojocaru, A. Cappelli, H. Alobeidli, B. Pannier, E. Almazrouei, and J. Launay, “The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only,” *arXiv preprint arXiv:2306.01116*, 2023.
- [34] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin *et al.*, “Opt: Open pre-trained transformer language models,” *arXiv preprint arXiv:2205.01068*, 2022.
- [35] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma *et al.*, “Scaling instruction-finetuned language models,” *arXiv preprint arXiv:2210.11416*, 2022.
- [36] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” 2023.
- [37] W. Dai, J. Li, D. Li, A. Tiong, J. Zhao, W. Wang, B. Li, P. N. Fung, and S. Hoi, “Instructclip: Towards general-purpose vision-language models with instruction tuning,” in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 49 250–49 267. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/9a6a435e75419a836fe47ab6793623e6-Paper-Conference.pdf
- [38] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, “Deepfashion: Powering robust clothes recognition and retrieval with rich annotations,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016. [Online]. Available: <http://dx.doi.org/10.1109/cvpr.2016.124>
- [39] F. Huang, Y. Cheng, C. Jin, Y. Zhang, and T. Zhang, “Deep multimodal embedding model for fine-grained sketch-based image retrieval,” in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Aug 2017. [Online]. Available: <http://dx.doi.org/10.1145/3077136.3080681>
- [40] H. Wu, Y. Gao, X. Guo, Z. Al-Halah, S. Rennie, K. Grauman, and R. Feris, “Fashion iq: A new dataset towards retrieving images by natural language feedback,” in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 11 302–11 312.
- [41] G. Delmas, R. S. Rezende, G. Csurka, and D. Larlus, “Artemis: Attention-based retrieval with text-explicit matching and implicit similarity,” in *International Conference on Learning Representations*, 2022.
- [42] A. Baldriati, M. Bertini, T. Uricchio, and A. D. Bimbo, “Composed image retrieval using contrastive learning and task-oriented clip-based features,” *ACM Transactions on Multimedia Computing, Communications and Applications*.
- [43] Z. Liu, W. Sun, D. Teney, and S. Gould, “Candidate set re-ranking for composed image retrieval with dual multi-modal encoder,” *Transactions on Machine Learning Research*, 2024. [Online]. Available: <https://openreview.net/forum?id=fjAwemcvpl>
- [44] A. Baldriati, L. Agnolucci, M. Bertini, and A. D. Bimbo, “Zero-shot composed image retrieval with textual inversion,” 2023.
- [45] G. Gu, S. Chun, W. Kim, , Y. Kang, and S. Yun, “Language-only training of zero-shot composed image retrieval,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [46] K. Saito, K. Sohn, X. Zhang, C.-L. Li, C.-Y. Lee, K. Saenko, and T. Pfister, “Pic2word: Mapping pictures to words for zero-shot composed image retrieval,” in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 19 305–19 314.
- [47] S. Vaze, N. Carion, and I. Misra, “Genecis: A benchmark for general conditional image similarity,” in *CVPR*, 2023.
- [48] K. Zhang, Y. Luan, H. Hu, K. Lee, S. Qiao, W. Chen, Y. Su, and M.-W. Chang, “MagicLens: Self-supervised image retrieval with open-ended instructions,” in *Proceedings of the 41st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, Eds., vol. 235. PMLR, 21–27 Jul 2024, pp. 59 403–59 420. [Online]. Available: <https://proceedings.mlr.press/v235/zhang24an.html>
- [49] Y. Bai, X. Xu, Y. Liu, S. Khan, F. Khan, W. Zuo, R. S. M. Goh, and C.-M. Feng, “Sentence-level prompts benefit composed image retrieval,” in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: <https://openreview.net/forum?id=m3ch3kJL7q>
- [50] A. Suhr, S. Zhou, A. Zhang, I. Zhang, H. Bai, and Y. Artzi, “A corpus for reasoning about natural language grounded in photographs,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Jan 2019. [Online]. Available: <http://dx.doi.org/10.18653/v1/p19-1644>
- [51] Y. Goyal, T. Khot, A. Agrawal, D. Summers-Stay, D. Batra, and D. Parikh, “Making the v in vqa matter: Elevating the role of image understanding in visual question answering,” *International Journal of Computer Vision*, p. 398–414, Apr 2019. [Online]. Available: <http://dx.doi.org/10.1007/s11263-018-1116-0>
- [52] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, “Language models are few-shot learners,” *CoRR*, vol. abs/2005.14165, 2020. [Online]. Available: <https://arxiv.org/abs/2005.14165>
- [53] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, “Frozen in time: A joint video and image encoder for end-to-end retrieval,” in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2021. [Online]. Available: <http://dx.doi.org/10.1109/iccv48922.2021.00175>
- [54] Z. Liu, W. Sun, Y. Hong, D. Teney, and S. Gould, “Bi-directional training for composed image retrieval via text prompt learning,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2024, pp. 5753–5762.
- [55] Z. Feng, R. Zhang, and Z. Nie, “Improving composed image retrieval via contrastive learning with scaling positives and negatives,” in *Proceedings of the 32nd ACM International Conference on Multimedia*, ser. MM ’24. New York, NY, USA: Association for Computing Machinery, 2024, p. 1632–1641. [Online]. Available: <https://doi.org/10.1145/3664647.3680808>
- [56] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, and Y. Cao, “Eva: Exploring the limits of masked visual representation learning at scale,” 2022. [Online]. Available: <https://arxiv.org/abs/2211.07636>
- [57] E. Dodds, J. Culpepper, S. Herdade, Y. Zhang, and K. Boakey, “Modality-agnostic attention fusion for visual search with text feedback,” *arXiv preprint arXiv:2007.00145*, 2020.
- [58] Y. Yu, S. Lee, Y. Choi, and G. Kim, “Curlingnet: Compositional learning between images and text for fashion iq data,” *arXiv preprint arXiv:2003.12299*, 2020.
- [59] S. Lee, D. Kim, and B. Han, “Cosmo: Content-style modulation for image retrieval with text feedback,” in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2021. [Online]. Available: <http://dx.doi.org/10.1109/cvpr46437.2021.00086>
- [60] Y. Chen, Z. Zheng, W. Ji, L. Qu, and T.-S. Chua, “Composed image retrieval with text feedback via multi-grained uncertainty regularization,” in *International Conference on Learning Representations (ICLR)*, 2024.
- [61] X. Zhang, Z. Zheng, L. Zhu, and Y. Yang, “Collaborative group: Composed image retrieval via consensus learning from noisy annotations,” *Know-Based Syst.*, vol. 300, no. C, Nov. 2024. [Online]. Available: <https://doi.org/10.1016/j.knosys.2024.112135>
- [62] X. Han, X. Zhu, L. Yu, L. Zhang, Y.-Z. Song, and T. Xiang, “Fame-vil: Multi-tasking vision-language model for heterogeneous fashion tasks,” in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 2669–2680.
- [63] Z. Yang, D. Xue, S. Qian, W. Dong, and C. Xu, “Ldre: Llm-based divergent reasoning and ensemble for zero-shot composed image retrieval,” in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024, pp. 80–90.
- [64] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, and J. Lin, “Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution,” *arXiv preprint arXiv:2409.12191*, 2024.
- [65] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu *et al.*, “Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 24 185–24 198.



Yinan Zhou received the B.S. degree from School of software Engineering, Xi'an Jiaotong University, Xi'an, China, in 2021. He is now a Ph.D. student in the School of Software Engineering at Xi'an Jiaotong University. His research interests include composed image retrieval, cross-modal retrieval, and multi-modal large language model.



Xiaoxiong Wang received the B.S. degree from Lanzhou University, Lanzhou, China, in 2015, and Ph.D. degree at School of software Engineering, Xi'an Jiaotong University, Xi'an, China, in 2021. He is now a associate professor in Hefei University of Technology. His research interests include cross-modal retrieval, image generation, semantic segmentation, and ReID.



Haokun Lin received the B.S. degree from School of Software Engineering, Huazhong University of Science and Technology, Wuhan, China, in 2021. He is now a Ph.D. student in Institute of Automation, Chinese Academy of Sciences. His research interests include multi-modal learning, model compression and large language models.



Chen Ma got his PhD from the School of Computer Science, McGill University, supervised by Prof. Xue (Steve) Liu. In the meantime, he also closely worked with Prof. Mark Coates. Before joining McGill, he received his MS and BS degrees in Software Engineering from Beijing Institute of Technology. He is currently an Assistant Professor in the Department of Computer Science, at the City University of Hong Kong since August 2021. His main research interests include natural language processing, recommender systems and data mining.



Li Zhu received the B.S. degree from Northwestern Polytechnical University, Xi'an, China, in 1989, and the M.S. and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, in 1995 and 2000, respectively..He is currently a Professor with the School of Software, Xi'an Jiaotong University. His main research interests include multimedia processing and communication, parallel computing, and networking.



Zhedong Zheng is an Assistant Professor with the University of Macau. He received the Ph.D. degree from the University of Technology Sydney in 2021 and the B.S. degree from Fudan University in 2016. He was a postdoctoral research fellow at the School of Computing, National University of Singapore. He received the IEEE Circuits and Systems Society Outstanding Young Author Award of 2021. His research interests include robust learning for image retrieval, generative learning for data augmentation, and unsupervised domain adaptation. He served as the senior PC for IJCAI and AAAI, and the area chair for ACM MM'24 and ICASSP'25.