PatternCIR Benchmark and TisCIR: Advancing Zero-Shot Composed Image Retrieval in Remote Sensing

Zhechun Liang[1], **Tao Huang**[1*], **Fangfang Wu**[1], **Shiwen Xue**[1], **Zhenyu Wang**[1], **Weisheng Dong**[1,3], **Xin Li**[2] and **Guangming Shi**[1,3]

[1]Xidian University

[2]State University of New York at Albany

[3]Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education

Abstract

Remote sensing composed image retrieval (RSCIR) is a new vision-language task that takes a composed query of an image and text, aiming to search for a target remote sensing image satisfying two conditions from intricate remote sensing imagery. However, the existing attribute-based benchmark Patterncom in RSCIR has significant flaws, including the lack of query text sentences and paired triplets, thus making it unable to evaluate the latest methods. To address this, we propose the *Zero-Shot Query Text Generator* (ZS-QTG) that can generate full query text sentences based on attributes, and then, by capitalizing on ZS-QTG, we develop the *PatternCIR* benchmark. PatternCIR rectifies Patterncom's deficiencies and enables the evaluation of existing methods. Additionally, we explore zero-shot composed image retrieval methods that do not rely on massive pre-collected triplets for training. Existing methods use only the text during retrieval, performing poorly in RSCIR. To improve this, we propose *Text-image Sequential Training of Composed Image Retrieval* (TisCIR). TisCIR undergoes sequential training of multiple self-masking projection and fine-grained image attention modules, which endows it with the capacity to filter out conflicting information between the image and text, enhancing the retrieval by utilizing both modalities in harmony. TisCIR outperforms existing methods by 12.40% to 62.03% on PatternCIR, achieving *state-of-the-art* performance in RSCIR. The data and code are available here.

1 Introduction

In recent years, remote sensing(RS) for earth observation (EO) has become a popular research topic. The large remote sensing data obtained provides research material for many computer vision tasks. However, managing and extracting relevant images from remote sensing data has become an urgent challenge. The ability to efficiently and quickly retrieve remote sensing images is crucial.

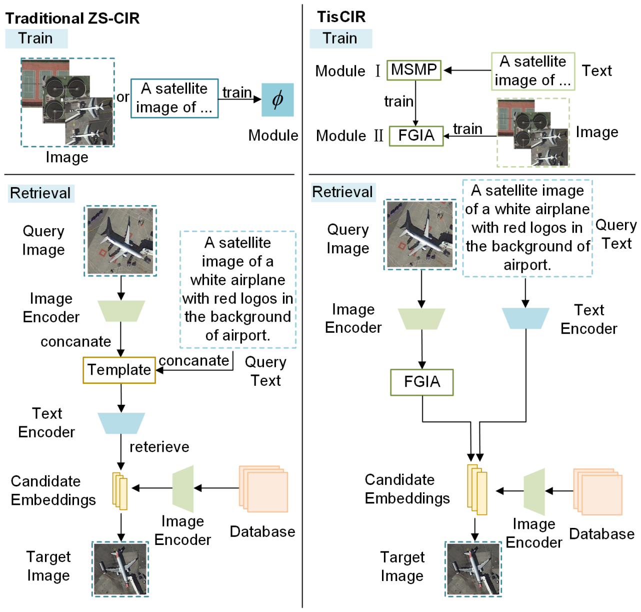

Figure 1: The workflow of our proposed TisCIR and traditional ZS-CIR methods. Unlike other ZS-CIR methods, TisCIR utilizes text and image for sequential training, therefore using both text and image simultaneously during the retrieval process.

Composed image retrieval (CIR) addresses this challenge by enabling the search and retrieval of images from RS image archives. Unlike traditional image retrieval, which relies on single-modality queries, CIR is a more challenging vision-language (VL) task that involves paired images and text. The goal is to retrieve the target image by adjusting the existing image through text. Since language is the most natural way to encode human interaction, CIR enhances the retrieval process, allowing users to refine their queries through language and obtain the desired images more effectively.

In natural image CIR, datasets consist of triplets $\langle x_{qi}, x_{qt}, x_{it}\rangle$, where x_{qi} is the query image, x_{qt} is the query text, and x_{it} is the target image. Many CIR methods [Huang *et al.*, 2023; Yang *et al.*, 2023; Song *et al.*, 2024; Chen *et al.*, 2024; Levy *et al.*, 2024; Wang *et al.*, 2024;

[*] Corresponding author

Qi *et al.*,] rely on these triplets for supervised training. However, obtaining massive triplets for training is difficult and expensive. Some datasets choose to use manual labeling, such as CIRR [Liu *et al.*, 2021] and FashionIQ [Wu *et al.*, 2021], but these datasets offer a limited variety of image classes. Consequently, CIR models trained on them lack generalization capability.

To address these limitations, zero-shot composed image retrieval (ZS-CIR) has emerged as a promising solution. ZS-CIR eliminates the reliance on pre-collected triplets, allowing models to perform retrieval without needing manually annotated datasets. Several methods, including projection-based ZS-CIR approaches. [Saito *et al.*, 2023a; Gu *et al.*, 2023; Baldrati *et al.*, 2023a] have demonstrated this approach.As shown in Fig.1, these methods leverage pre-trained models like CLIP [Radford *et al.*, 2021] to train a lightweight projection ϕ that represents the query image x_{qi} as words, and retrieval is then performed by concatenating the projected image representation with the query text x_{qt} according to the template. Although these methods demonstrate better generalization ability, the projection ϕ only captures coarse information from the image, neglecting the remaining fine-grained details. Additionally, only text modality is used during retrieval, which results in suboptimal performance when applied to RS images.

Moreover, in remote sensing composed image retrieval (RSCIR), the only benchmark, Patterncom [Psomas *et al.*, 2024], currently supports retrieval based solely on image attribute. It does not include query text sentences or paired triplets, making it incompatible with most CIR methods designed for fine-grained retrieval in natural image datasets. As a result, many advanced CIR techniques cannot be effectively applied to RSCIR.

To address the problems in existing research, firstly, we propose a model for expanding the attribute words into full-text query sentences, called the Zero-Shot Query Text Generator (ZS-QTG). ZS-QTG utilizes the keywords-to-sentence language model CBART [He, 2021a] to generate the sentences. After several fine-tunings, ZS-QTG obtains the text query sentence that describes the target image while retaining the attribute words.

Secondly, based on Patterncom, we constructed the triplets $\langle x_{qi}, x_{qt}, x_{it}\rangle$ following the CIRR benchmark. We named the new benchmark PatternCIR. Unlike Patterncom, since PatternCIR constructs subsets and triplets, the latest CIR methods can be evaluated on PatternCIR. Specifically, Resnet50 is used to extract image features and calculate similarity. Next, image pairs $\langle x_{qi}, x_{it}\rangle$ are selected according to the criteria of CIRR, and text query statements x_{qt} are obtained using ZS-QTG. Finally, the dataset is organized into the CIRR format, enabling CIR methods designed for natural images to be conveniently evaluated on PatternCIR.

Lastly, we propose a ZS-CIR method called Text-image sequential training of Composed Image Retrieval (TisCIR). TisCIR first trains a Multiple Self-Masking Projection (MSMP) module $\hat{\phi}$ using only text queries. The MSMP module $\hat{\phi}$ is designed to extract the conflicting information between query text x_{qt} and query image x_{qi}. Next, images and $\hat{\phi}$ are used to further train a Fine-Grained Image Attention (FGIA) module, enabling it to filter the conflicting information and retain fine-grained information in image embedding. Finally, unlike all ZS-CIR methods, TisCIR can use both text embedding and image embedding simultaneously during the retrieval process. A weighted sum of the distances to the text embedding and the filtered image embedding is computed, serving as the final distance for ranking.

In the experiments, our TisCIR demonstrated excellent performance in RSCIR, surpassing ZS-CIR methods applied to natural images, achieving state-of-the-art performance. Additionally, we validated the effect of the FGIA module by visualizing the image embedding attention. We confirmed that the module can effectively remove the information that needs to be replaced while retaining the fine-grained information in the rest of the image. Our contributions are summarized as follows:

- We propose a model for generating text query, called the *Zero-Shot Query Text Generator* (ZS-QTG). Using ZS-QTG, we constructed *PatternCIR*, the first fine-grained composed image retrieval benchmark in RSCIR. Compared to the existing benchmark, PatternCIR contains subsets of similar images and triplets, which enable the evaluation of all the latest CIR methods.
- We propose *Text-image sequential training of Composed Image Retrieval* (TisCIR), a new zero-shot composed image retrieval method for RSCIR. TisCIR uses sequential training of images and text to avoid the use of labeled triplets. Unlike existing ZS-CIR methods, TisCIR allows the simultaneous use of both image and text embeddings during retrieval, utilizing more fine-grained information.
- TisCIR improves upon the latest ZS-CIR methods by 22.95% to 62.03% in RSCIR, achieving state-of-the-art performance. The effectiveness of the TisCIR modules has also been demonstrated and visualized in the ablation study.

2 Related Work

2.1 Vision-langauge Model

Vision-language model (VLM) is pre-trained on large-scale datasets containing hundreds of millions of image-caption pairs, such as CLIP [Radford *et al.*, 2021], ALIGN [Li *et al.*, 2021]. These models utilize an image and language encoder pair that can project both images and text to embedding vectors in the same space, thus bridging the modality gap between text and image. Once trained, VLM can be applied to various VL tasks with minimal or no additional annotation cost. As a result, many researchers are using image datasets in different domains to train VLM to enhance their ability to understand images. For example, RemoteCLIP [Liu *et al.*, 2024] is the latest VLM trained with remote sensing data, demonstrating excellent performance in the field of remote sensing images. In this paper, we use VLM in both the generation of query text statements and the CIR task.

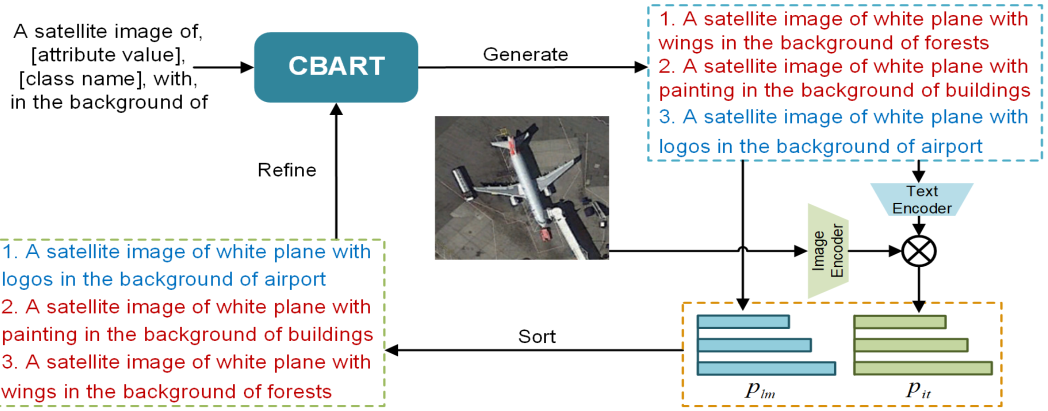
A satellite image of, [attribute value], [class name], with, in the background of
CBART
Generate
1. A satellite image of white plane with wings in the background of forests
2. A satellite image of white plane with painting in the background of buildings
3. A satellite image of white plane with logos in the background of airport
Text Encoder
Image Encoder
Refine
1. A satellite image of white plane with logos in the background of airport
2. A satellite image of white plane with painting in the background of buildings
3. A satellite image of white plane with wings in the background of forests
Sort
p_{lm}
p_{it}

Figure 2: **The overview of ZS-QTG.** ZS-QTG generates candidate sentences by expanding the template *"A satellite image of, [attribute value], [class name], with, in the background of."*. Then sort the sentences with fluency and image guidance. The one with the highest score **(blue one)** is selected for refining, and others **(red ones)** are abolished. After rounds of refinement, the sentence with the highest score is the query text x_{qt}.

2.2 Composed Image Retrieval

Composed Image Retrieval (CIR) is proposed to retrieve the target image x_{ti} using a query pair consisting of an image query x_{qi} and a query text x_{qt} [Vo *et al.*, 2019]. This topic has been explored in the field of natural images, such as fashion and scene composition.

Mainstream CIR methods generally fall into two categories: supervised and zero-shot. Supervised methods mainly rely on labeled datasets to train fusion modules for the image and text embeddings extracted by VLM. These methods are heavily dependent on labeled datasets and lack generalization ability.

Zero-shot methods do not require labeled paired triplets during training. Currently, most zero-shot methods train projection ϕ with text or image separately. ϕ projects the embedding of a query image x_{qi} into a token embedding, which is then concatenated with the embedding of query text x_{qt} according to a template, thus ultimately retrieving only in the text modality. Pic2Word [Saito *et al.*, 2023b] trains the projection module ϕ by minimizing the contrastive loss between the image latent embedding and the textual latent embedding of the predefined template "a photo of [$]". SEARLE [Baldrati *et al.*, 2023b] employs a similar approach to Pic2Word. LinCIR [Gu *et al.*, 2023] proposed using only text to train the projection ϕ and mitigates the modality gap between image and text by adding noise.

However, the information extracted by the ϕ in current zero-shot methods depends on the predefined template and mask, which can lead to the loss of fine-grained information in the image during the cross-modal process. In contrast to these methods, our approach can preserve the Fine-Grained information in the image and perform simultaneous retrieval using both the text and image modalities.

2.3 Keywords-to-sentence LM

To generate a fluent, visually-related query text x_{qt} based on image attribute words, we utilize the pre-trained lexically constrained language model, CBART [He, 2021b]. In the use of CBART for image understanding, there has already been some research [Zeng *et al.*,]. Specifically, CBART is designed to generate a sentence $S = (x_1, ..., x_n)$ by maximizing the conditional probability, given the ordered set of K keywords

$$S = \underset{S}{argmax}\ P(x_1, ..., x_n|c_{i\,i=1}^{T}) \tag{1}$$

Specifically, CBART consists of an action encoder and a language decoder. In the inference process, the encoder predicts the action (copy, replacement, or insertion) to be taken for each word at the t-th iteration.

Copy means the current word remains unchanged. **Replacement** suggests the current word should be replaced. Specifically, CBART uses a mask token $\langle M\rangle$ to replace the current word and sample a new word based on the language model's conditional probability p_{lm}. **Insertion** indicates the decoder should insert a word before the current word. Similar to the replacement action, CBART inserts a $\langle M\rangle$ token before the current word and then samples a word from p_{lm}.

Accordingly, the decoder can refine the sentence from S_t to S_{t+1}. Therefore, the complete encoder-decoder sentence refinement by CBART at the t-th iteration can be formulated as

$$L_t = LM_{encoder}(S_t) \tag{2}$$

$$S_{t+1} = LM_{decoder}(S_t, L_t) \tag{3}$$

where L_t is actions predicted by encoder at t-th iteration.

In this paper, although CBART can generate the sentence S_t while preserving attribute words, there is no constraint to ensure that accurately reflects the content of the target image x_{ti}. Our ZS-QTG improves upon this and manages to generate accurate query text x_{qt}.

3 Methodology

In this section, we first introduce our model for generating query text, the Zero-Shot Query Text Generator (ZS-QTG), and the overview of the model is shown in Fig.2. Then, we discuss the expansion and improvement of the Patterncom dataset, explaining how subsets and triplets $\langle x_{qi}, x_{qt}, x_{it} \rangle$ are constructed. Finally, we introduce our proposed new ZS-CIR method, Text-image Sequential Training of Composed Image Retrieval (TisCIR), with the model training and inference process illustrated in Fig.3, Fig.4 and Fig.5. TisCIR incorporates two modules, Multiple Self-Masking Projection (MSMP) and Fine-Grained Image Attention (FGIA), for sequential training, enabling the use of both text and image information during retrieval inference.

3.1 Zero-Shot Query Text Generator

The overview of ZS-QTG is shown in Fig.2. To ensure that the sentence S_t generated by CBART is highly relevant to the given target image x_{ti}, we need visual guidance to select candidate words. We choose to use the contrastive score of the RemoteCLIP model, a VL model in remote sensing, for evaluating the visual-text similarity. By combining conditional probability p_{lm} and the contrastive score p_{it}, we can adjust CBART's original word prediction to align with the content of x_{it}.

Specifically, when sampling a word x_i at position i, CBART first predicts a conditional probability p_{lm} and select top-K_w candidate words $\{x_i^k\}_{k=1}^{K_w}$ with the corresponding fluent score, as:

$$p_{lm}(x_i^k) = p_{lm}(x_i^k | x_{-i}^k), k = 1, ..., K_w \tag{4}$$

where x_{-i}^k denotes words that remain unchanged. Then K_w candidate sentences $\{s_k = (x_1, ..., x_i^k, ..., x_n)\}_{k=1}^{K_w}$ are formed by combining candidate word x_i^k with the context x_{-i}^k.

Then the the contrastive score is denoted as p_{it}, which can be computed by taking candidate sentences $\{s_k\}_{k=1}^{K_w}$and the target image x_{ti} as input to calculate the image-text cross-modality similarity as

$$p_{it}(x_i^k) = cos(E_i(x_{ti}), E_t(s_k)) \tag{5}$$

where E_i and E_t denote the image and text encoder of RemoteCLIP. Then, after a weighted sum of Eq.(4) and Eq.(5) we have the final score as

$$p(x_i^k) = \alpha p_{lm}(x_i^k) + \beta p_{it}(x_i^k), k = 1, ..., K_w \tag{6}$$

As a result, when sampling i-th word for replacement or insertion in CBART for our model, our ZS-QTG selects the candidate word with the highest score $p(x_i^k)$.

3.2 Construct PatternCIR Dataset

In RSCIR, the existing Patterncom dataset only provides image attributes as labels and does not construct subsets or triplets $\langle x_{qi}, x_{qt}, x_{it} \rangle$, making it unsuitable for evaluating fine-grained CIR methods. To construct our own PatternCIR dataset based on Patterncom, we first follow the approach used in the CIRR dataset, utilizing ResNet152 to extract image features. Then, we construct image subsets based on similarity, and subsequently, select image pairs from the subsets. Finally, using ZS-QTG, we generate to further construct triplets.

Image subset construction

The nature of the CIR task requires a set of negative images with high visual similarity, on which triplets are constructed. Without this, distinguishing between reference images and target images would be an easy task. Additionally, subset recall $R_{subset}@K$ is an important metric for evaluating the performance of CIR methods, so constructing image subsets is a necessary step in building a CIR dataset. Following the approach used in CIRR, for each attribute, we construct multiple subsets of six images that are visually similar, denoted as $S = \{I_1, ...I_6\}$.

Here, to construct a subset within a set of images for a given attribute D, we select one image from the image set randomly, $I_1 \in D$. We sort the remaining images in D according to their cosine similarity to I_1 using ResNet152 [He *et al.*, 2016] image feature vectors pre-trained on ImageNet [Krizhevsky *et al.*, 2017], which is denoted as k_i for image I_i. We then pick five images to construct the subset as follows: First, we remove images with $k_i \geq 0.94$ to avoid near-identical images to I_1.Then for the rest images, we add each image based on similarity ranking, skipping an image if its similarity is within 0.002 of the last image added. If a subset of size 6 images cannot be created or the number of overlapping images with the existing subset exceeds 4, the entire set is discarded.

Image pairing and annotations

In the constructed subset S, we aim to select image pairs to form triplets as supervisory data. First, we choose a query image and a target image to form an image pair $\langle x_{qi}, x_{it} \rangle$. To this end, we designate the initial image I_1 of each subset S as the query image x_{qi} and sequentially pair it with the remaining 5 images to generate image pairs. Subsequently, we use ZS-QTG to generate the query text x_{qt} for each image pair $\langle x_{qi}, x_{it} \rangle$, thereby constructing triplets $\langle x_{qi}, x_{qt}, x_{it} \rangle$.

While generating query text x_{qi}, we choose a fixed initial template as the input to ZS-QTG, which is *"A satellite image of, [attribute value], [class name], with, in the background of."*. Here, *[attribute value]* corresponds to the target image x_{it}'s attribute values (e.g., white, dense), and *[class name]* corresponds to its object-of-interest class names (e.g., airplane, tennis court). ZS-QTG expands this template into a full sentence while preserving the initial template words. Phrases separated by commas are treated as a whole, neither split nor having their order disrupted.

Finally, we have obtained a total of 17700 triplets, which are divided into test, validation, and training sets in a ratio of 1.5:1.5:7. The dataset is organized according to the CIRR format, enabling existing CIR methods to be easily evaluated on it. More details of PatternCIR are in the appendix.

3.3 Text-image Sequential Training of Composed Image Retrieval

Existing ZS-CIR methods perform poorly in Remote Sensing Cross-Image Retrieval (RSCIR) when only text embedding is used during retrieval. To address this issue, we propose the Text-image Sequential Training of Composed Image Retrieval (TisCIR). We design TisCIR to utilize both the query text x_{qt}'s and query image x_{qi}'s embeddings simultaneously

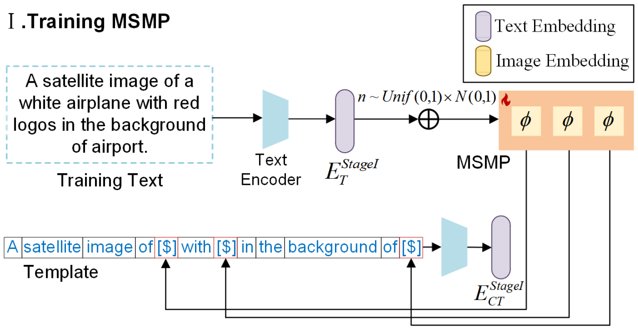
Ⅰ.Training MSMP
Text Embedding
Image Embedding
A satellite image of a white airplane with red logos in the background of airport.
Training Text
Text Encoder
E_T^{StageI}
$n \sim Unif(0,1) \times N(0,1)$
MSMP
ϕ ϕ ϕ
A satellite image of [$] with [$] in the background of [$]
Template
E_{CT}^{StageI}

Figure 3: **Stage I:Training MSMP.** A multiple of a random scalar and a vector drawn from Gaussian distribution is added to the text embedding E_T^{stageI} to better bridge the modality gap. Subsequently, the modified text embedding is projected onto the text in the masked positions of the template. After that, the embedding E_{CT}^{stageI} of the concatenated text is obtained. Finally, the loss between them is computed according to (7).

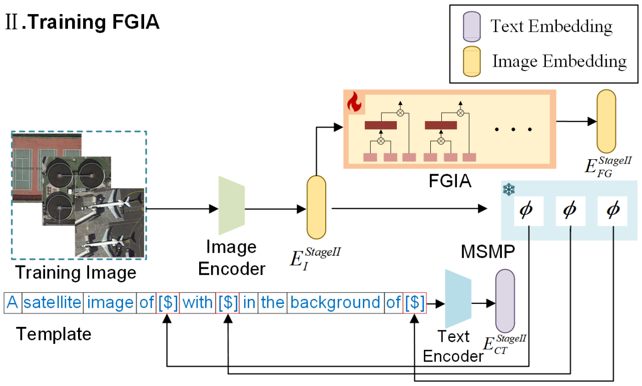

Figure 4: **Stage II:Training FGIA.** FGIA filters the initial image embedding $E_I^{stageII}$ to obtain the fine-grained information embedding $E_{FG}^{stageII}$. Next, the trained MSMP projects $E_I^{stageII}$ onto text and concatenates the projected text with the template, resulting in the concatenated text $E_{CT}^{stageII}$. Finally, the loss between them is computed according to (8), guaranteeing that FGIA filters out conflicting information.

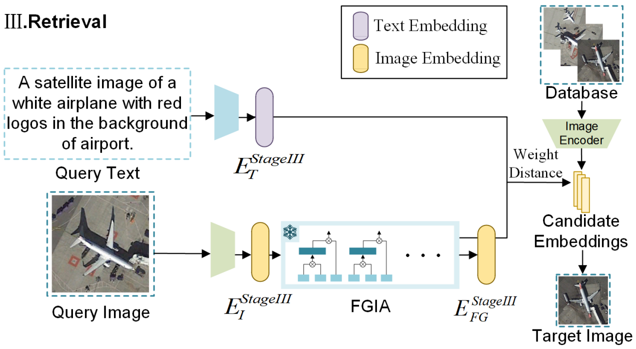

Figure 5: **Stage III:Retrieval.** FGIA filters the initial image embedding $E_I^{stageIII}$ to obtain fine-grained information embedding $E_{FG}^{stageIII}$. The, the weighted distance between the embeddings of database images E_I and them is calculated according to (9) as the basis for retrieval.

during retrieval. However, challenges arise as some information in the two embeddings is conflicting. The query text is mainly used to describe the target image x_{it} rather than the query image x_{qi}, often being adjusted through human interaction. Therefore, we aim to preserve the information in the query text embedding while removing the conflicting information in the query image embedding.

Specifically, the training of TisCIR is divided into two stages. In the first stage, only the query text is used. Considering that text information has a structural nature, we propose a Multiple Self-Masking Projection (MSMP) module to extract its information in conjunction with a masked template. Then, in the second stage of training, only the query image and the trained MSMP are employed to further train the Fine-Grained Image Attention (FGIA) module. This enables the FGIA to filter out conflicting information from the image embedding and focus on fine-grained details. Finally, during retrieval, both the query text embedding and the filtered query image embedding are utilized for retrieval. Since the paired query text and query image $\langle x_{qt}, x_{qi} \rangle$ are not used during training, TisCIR is considered a zero-shot method.

Training multiple self-masking projection

The original approach [Saito *et al.*, 2023b] involves training a single SMP [Gu *et al.*, 2023] for all multiple masks at different positions, which yielded poor performance. This is because the single SMP might not be able to handle the diverse semantic information from different masks effectively. In particular, when masks are located at different positions, they may carry different semantic nuances that a single SMP fails to capture comprehensively. Therefore, we propose MSMP, where a separate SMP is trained for each mask, illustrated in Fig.3. During training, only the corresponding mask is retained, whereas the other parts of the text remain unchanged. This approach allows each SMP to focus on extracting the semantic information of a single mask, thereby avoiding interference from other parts of the text. As a result, this approach leads to better performance by providing more accurate and focused semantic extraction. Specifically, by isolating the processing of each mask, we can ensure that the semantic extraction is not diluted by irrelevant information, leading to more reliable and useful embeddings. The loss function of MSMP is as follows:

$$L_{MSMP} = ||E_T^{stageI} - E_{CT}^{stageI}||_2^2 \tag{7}$$

We use the MSE loss between E_T^{stageI} and E_{CT}^{stageI} to enable MSMP to extract information from the masked parts of the template. To more effectively bridge the modality gap between the image and text modalities, a vector is sampled from a Gaussian distribution, multiplied by a random scalar, and the resulting product is added to the embedding.

Training fine-grained image attention

Building on MSMP, we further train the Fine-Grained Image Attention (FGIA) module, with the training process shown in the Fig.4. The structure of FGIA is a multihead self-attention mechanism applied to embeddings. Specifically, the multi-

	$R@1$	$R@5$	$R@10$	$R@20$	$R@50$	$R_{subset}@1$	$R_{subset}@2$	$R_{subset}@3$
Pic2word (ViT-L)	5.92	17.69	27.18	40.31	64.20	36.39	58.27	76.27
LinCIR (ViT-L)	4.53	14.18	22.12	33.31	57.05	34.93	57.05	75.09
Searle (ViT-B)	1.99	7.42	12.96	21.53	39.92	26.31	47.80	67.76
Searle-xl (ViT-L)	5.97	17.97	26.50	38.93	62.94	39.46	62.75	78.97
WEICOM(ViT-L)	3.76	13.57	21.97	32.63	53.91	39.24	62.19	78.75
FTI4CIR(ViT-L)	6.17	18.21	26.53	40.21	63.64	37.27	59.81	77.61
LDRE(ViT-L)	6.53	18.23	26.84	41.14	63.40	39.21	60.10	79.91
Ours (ViT-L)	**7.34**	**19.01**	**27.76**	**41.34**	**64.51**	**44.81**	**67.83**	**82.94**

Table 1: **Main results.** The comparison of recall metric between our TisCIR and other ZS-CIR competing methods.

MSMP	FGIA	$R@1$	$R@5$	$R@10$	$R@20$	$R@50$	$R_{subset}@1$	$R_{subset}@2$	$R_{subset}@3$
		4.32	12.6	18.43	21.35	38.21	28.24	49.34	69.67
	✓	5.63	17.92	27.21	39.13	60.12	33.14	52.81	71.25
✓		4.53	14.18	22.1	33.31	57.05	34.93	57.5	75.09
✓	✓	**7.34**	**19.01**	**27.76**	**41.34**	**64.51**	**44.81**	**67.83**	**82.94**

Table 2: **Ablation study.** The comparison of recall metric between TisCIR without FGIA, with FGIA trained by SMP and with FGIA trained by MSMP on PatternCIR.

head self-attention mechanism allows FGIA to focus on different aspects of the embeddings simultaneously. In order to enable FGIA to filter out information that conflicts with the text, we employ the following fine-grained loss for constraint:

$$L_{FGIA} = ||E_I^{stageII} - E_{CT}^{stageII} - E_{FG}^{stageII}||_2^2 \quad (8)$$

where $E_I^{stageII}$ contains all the information from the initial query image x_{qi}, $E_{FG}^{stageII}$ contains the fine-grained information in x_{qi} and $E_{CT}^{stageII}$ contains the conflicting information between image and text extracted by MSMP. With the FGIA loss, $E_{FG}^{stageII}$ is adjusted to align with the difference between $E_I^{stageII}$ and $E_{CT}^{stageII}$. This means that $E_{FG}^{stageII}$ represents the fine-grained information in the image, excluding the conflicting information.

Retrieval

After obtaining FGIA, we can perform retrieval by simultaneously using the query text x_{qt} and query image x_{qi}. The retrieval process is illustrated in the Fig.5. Specifically, during the retrieval process, a weighted sum distance D is calculated. This distance is based on the cosine distances between the embeddings of database images E_i and both the text embedding $E_T^{stageIII}$ and the fine-grained information embedding $E_{FG}^{stageIII}$. The formula is as follows:

$$D = Dist(E_{FG}^{stageIII}, E_i) + \lambda Dist(E_T^{stageIII}, E_i) \quad (9)$$

$$Dist(A, B) = 1 - \frac{A \cdot B}{||A||||B||} \quad (10)$$

where λ is the retrieval distance weight.

4 Experiments

In this section, we discuss the evaluation of TisCIR on the PatternCIR dataset and compare it with the latest ZS-CIR methods in natural image tasks. Subsequently, We conduct ablation experiments on the MSMP and FGIA modules to verify the effectiveness of our proposed improvements. Finally, to further demonstrate the effectiveness of the FGIA module, we visualize the similarity among the embeddings before FGIA filtering, the embeddings after FGIA filtering, and the embedding of the original image in Fig.6.

4.1 Implementation Details

Settings

For the MLP in MSMP, we use the same settings as in [Gu *et al.*, 2023]: LN [Ba *et al.*, 2016]-Linear-GeLU [Hendrycks and Gimpel, 2016]-Linear-GeLU-Linear-LN. We use the AdamW optimizer [Loshchilov and Hutter, 2017] with a fixed learning rate of 0.0001, a weight decay of 0.01, and a batch size of 256 for the first stage of training, whereas a batch size of 32 is used for the second stage. Dropout with a probability of 50% is applied for regularization. During the first stage of training, we use all the query texts from the PatternCIR training set as training texts, and during the second stage, we employ all the images from the training set as training images.

Evaluation metrics

In CIR, recall rate $R@K$ represents the probability of retrieving the target image among the top K images. Among which we select $R@1$ as the reference for choosing the optimal training weights on the validation set. Additionally, the subset recall rate $R_{subset}@K$ refers to the probability of retrieving the target image within the top K images of a subset. Compared to $R@K$, $R_{subset}@K$ reduces the impact of false negatives, thereby making it an important reference metric in the CIR field.

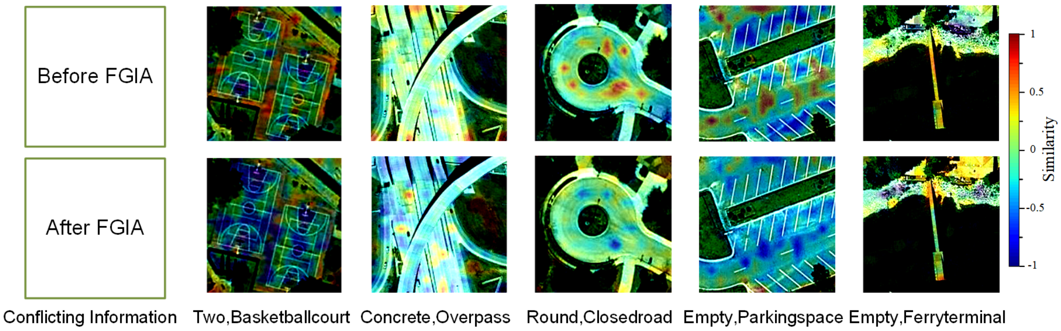

Figure 6: **Visulization of FGIA.** We calculated the cosine similarity between the embeddings before and after being filtered by FGIA and the patch embeddings of the original image. We then plotted a heatmap overlaying the original image and, below, listed the attributes of the images that need to be filtered out by FGIA due to conflicts with the query text.

Comparison methods

We compare TisCIR with the latest ZS-CIR methods, namely Pic2Word [Saito *et al.*, 2023b], LinCIR [Gu *et al.*, 2023], SEARLE [Baldrati *et al.*, 2023b], WEICOM[Psomas *et al.*, 2024], FTI4CIR[Lin *et al.*, 2024], and LDRE[Yang *et al.*, 2024]. All methods employ the ViT-L CLIP backbone for training, except for SEARLE, which uses the ViT-B backbone. Therefore, we also compare TisCIR with its variant SEARLE-XL, which uses the ViT-L backbone.

4.2 Main Results

The main experimental results are shown in Tab.1, from which it can be seen that TisCIR outperforms all ZS-CIR methods in both the recall rate $R@K$ and the subset recall rate $R_{subset}@K$. Among the methods using ViT-L backbone, TisCIR improves $R@1$ by 22.95% to 62.03% and $R_{subset}@1$ by 13.56% to 28.29%, thereby achieving state-of-the-art performance on RSCIR. SEARLE, which uses the ViT-B backbone, exhibits poor performance, indicating that the choice of backbone has a significant impact on RSCIR. Additionally, LinCIR, which only uses text for training, performs the worst among all CIR methods using ViT-L backbone, suggesting that image information is more significant in RSCIR.

4.3 Analysis

In this section, we discuss the roles of our two improvement modules, MSMP and FGIA. We first conduct ablation experiments on both MSMP and FGIA. Subsequently, we visualize the patch similarity map of the embeddings before and after FGIA filtering, which intuitively demonstrates the effect of FGIA.

Ablation study

The results of the ablation experiments are presented in the Tab.2. We separately validate the effectiveness of the MSMP and FGIA modules. The method utilizing both MSMP and FGIA performs the best. This indicates the effectiveness of our improvements over the original method.

FGIA visualization

Visualizing CLIP image embeddings has always been a challenge due to the absence of an inverse mapping capable of mapping the image embedding back to the image. However, we can still visualize the information within the image embedding by employing certain methods, and the similarity map is the approach we chose. When CLIP extracts image embeddings, it first divides the image into patches based on the the model setup, thereby obtaining multiple patch embeddings. Subsequently, it uses a linear layer to obtain the global embeddings as the final image embedding. In this paper, we calculate the cosine similarity among the image embeddings before and after filtering by FGIA and the original image patch embeddings. We then construct a similarity map based on the patch positions and overlay it onto the original image, enabling us to observe the parts of the image on which each image embedding focuses.

The final visualization results are shown in the Fig.6. We can see that embeddings before being filtered by FGIA focuses more on the main object in the image, which conflicts with the query text. After filtering, the embeddings remove this information and focus more on the fine-grained details in the background. This visually demonstrates that our FGIA has achieved the intended design effect.

5 Conclusion

In this paper, we propose a novel model for generating CIR query text statements, namely *Zero-Shot Query Text Generator* (ZS-QTG) , and use ZS-QTG to construct the RSCIR benchmark dataset *PatternCIR*, which encompasses subsets and triplets of similar images. PatternCIR enables the evaluation of most existing CIR methods. Finally, we introduce a novel ZS-CIR method, *Sequential Training of Composed Image Retrieval* (TisCIR), which is distinct from existing ZS-CIR methods. During training, TisCIR leverages both image and text sequences to train the MSMP and FGIA modules, thereby enabling the concurrent utilization of both modalities during retrieval. TisCIR outperforms all state-of-the-art ZS-CIR methods in the RSCIR task, achieving state-of-the-art performance.

References

[Ba *et al.*, 2016] Jimmy Ba, Jamie Kiros, and GeoffreyE. Hinton. Layer normalization. *arXiv: Machine Learning,arXiv: Machine Learning*, Jul 2016.

[Baldrati *et al.*, 2023a] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and AlbertoDel Bimbo. Zero-shot composed image retrieval with textual inversion. Mar 2023.

[Baldrati *et al.*, 2023b] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and AlbertoDel Bimbo. Zero-shot composed image retrieval with textual inversion. Mar 2023.

[Chen *et al.*, 2024] Yanzhe Chen, Huasong Zhong, Xiangteng He, Yuxin Peng, Jiahuan Zhou, and Lele Cheng. Fashionern: Enhance-and-refine network for composed fashion image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1228–1236, 2024.

[Gu *et al.*, 2023] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, and YoohoonKangandSangdoo Yun. Language-only efficient training of zero-shot composed image retrieval. Dec 2023.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016.

[He, 2021a] Xingwei He. Parallel refinements for lexically constrained text generation with bart. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Jan 2021.

[He, 2021b] Xingwei He. Parallel refinements for lexically constrained text generation with bart. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Jan 2021.

[Hendrycks and Gimpel, 2016] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *Cornell University - arXiv,Cornell University - arXiv*, Jun 2016.

[Huang *et al.*, 2023] Fuxiang Huang, Lei Zhang, Xiaowei Fu, and Suqi Song. Dynamic weighted combiner for mixed-modal image retrieval. Dec 2023.

[Krizhevsky *et al.*, 2017] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, page 84–90, May 2017.

[Levy *et al.*, 2024] Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Data roaming and quality assessment for composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2991–2999, 2024.

[Li *et al.*, 2021] Junnan Li, RamprasaathR. Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and StevenC.H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Cornell University - arXiv,Cornell University - arXiv*, Jul 2021.

[Lin *et al.*, 2024] Haoqiang Lin, Haokun Wen, Xuemeng Song, Meng Liu, Yupeng Hu, and Liqiang Nie. Fine-grained textual inversion network for zero-shot composed image retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 240–250, 2024.

[Liu *et al.*, 2021] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2021.

[Liu *et al.*, 2024] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[Loshchilov and Hutter, 2017] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *Learning,Learning*, Nov 2017.

[Psomas *et al.*, 2024] Bill Psomas, Ioannis Kakogeorgiou, Nikos Efthymiadis, Giorgos Tolias, Ondrej Chum, Yannis Avrithis, and Konstantinos Karantzalos. Composed image retrieval for remote sensing. May 2024.

[Qi *et al.*,] Daiqing Qi, Handong Zhao, and Sheng Li. Easy regional contrastive learning of expressive fashion representations. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[Radford *et al.*, 2021] Alec Radford, JongWook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Askell Amanda, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *Cornell University - arXiv,Cornell University - arXiv*, Feb 2021.

[Saito *et al.*, 2023a] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. 0pic2word: Mapping pictures to words for zero-shot composed image retrieval. Feb 2023.

[Saito *et al.*, 2023b] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. 0pic2word: Mapping pictures to words for zero-shot composed image retrieval. Feb 2023.

[Song *et al.*, 2024] Chull Hwan Song, Taebaek Hwang, Jooyoung Yoon, Shunghyun Choi, and Yeong Hyeon Gu. Syncmask: Synchronized attentional masking for fashion-centric vision-language pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13948–13957, 2024.

[Vo *et al.*, 2019] 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 *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2019.

[Wang *et al.*, 2024] Yifan Wang, Wuliang Huang, Lei Li, and Chun Yuan. Semantic distillation from neighborhood for composed image retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5575–5583, 2024.

[Wu *et al.*, 2021] 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 *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2021.

[Yang *et al.*, 2023] Xingyu Yang, Daqing Liu, Heng Zhang, Yong Luo, Chaoyue Wang, and Jing Zhang. Decompose semantic shifts for composed image retrieval. Sep 2023.

[Yang *et al.*, 2024] Zhenyu Yang, Dizhan Xue, Shengsheng Qian, Weiming Dong, and Changsheng 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*, pages 80–90, 2024.

[Zeng *et al.*,] Zequn Zeng, Yan Xie, Hao Zhang, Chiyu Chen, Bo Chen, and Zhengjue Wang. Meacap: Memory-augmented zero-shot image captioning.