# LDRE: LLM-based Divergent Reasoning and Ensemble for Zero-Shot Composed Image Retrieval

Zhenyu Yang, Dizhan Xue, Shengsheng Qian, Weiming Dong,
Changsheng Xu

July 14–18, 2024

## Abstract

In recent years, zero sample synthetic image retrieval (ZS-CIR) has received increasing attention, with the aim of retrieving target images based on queries composed of reference images and modified text without training samples. Specifically, the modified text describes the difference between the two images. In order to perform ZS-CIR, mainstream methods use pre trained image to text models to convert query images and text into a single text, and then project it into a common feature space through CLIP to retrieve the target image. However, these methods overlook that ZS-CIR is a typical fuzzy retrieval task, where the semantics of the target image are not strictly defined by the query image and text. To overcome this limitation, this paper proposes a ZS-CIR divergent inference and ensemble (LDRE) method based on untrained LLM to capture various possible semantics of the combined results. Firstly, the author uses a pre trained subtitle model to generate dense subtitles for the reference image, with a focus on the different semantic perspectives of the reference image. Then, the author suggests that Large Language Models (LLMs) use dense headings and modified text for divergent combination reasoning, resulting in divergent edited headings that cover the possible semantics of the combined target. Finally, the author designed a divergent subtitle ensemble to obtain integrated subtitle features weighted by semantic relevance scores, which were then used to retrieve target images in the CLIP feature space. Extensive experiments on three public data sets show that the LDRE proposed by the author achieves the latest and most advanced performance. **Keywords:** Composed image retrieval, zero-shot learning, multi-modal retrieval.

## 1 Introduction

Composed Image Retrieval (CIR) [5, 43, 48] requires a sophisticated integration of image content and textual query semantics in order to retrieve a novel image that precisely encapsulates the pertinent image elements and modifications specified in the textual query. To accomplish this objective, prior research requires meticulously curated triplets consisting of a reference image, modification text, and a target image for training. However, annotating such triplets poses considerable challenges and requires extensive labor. To address this issue, the Zero-Shot Composed Image Retrieval (ZS-CIR) task [3, 36, 39] has recently been introduced to improve the generalization of CIR models without relying on annotated triplets.

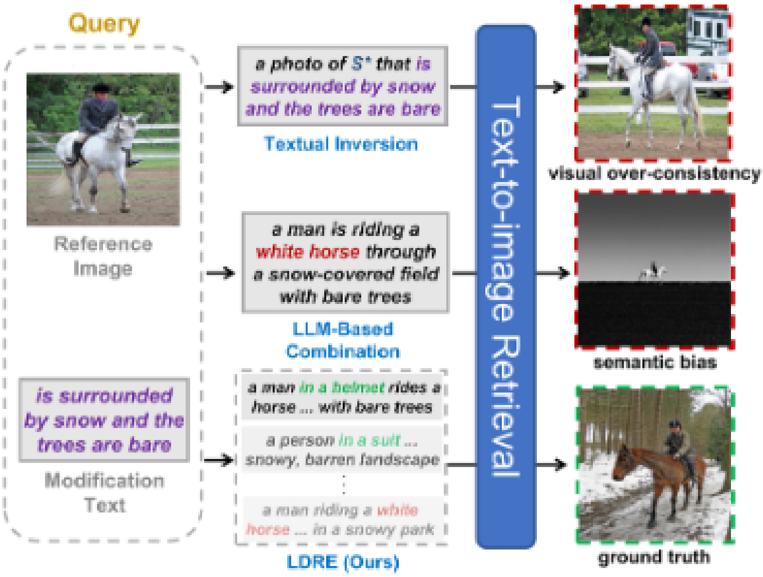

Query
Reference Image
is surrounded by snow and the trees are bare
Modification Text
a photo of $S^*$ that is surrounded by snow and the trees are bare
Textual Inversion
a man is riding a white horse through a snow-covered field with bare trees
LLM-Based Combination
a man in a helmet rides a horse ... with bare trees
a person in a suit ... snowy, barren landscape
⋮
a man riding a white horse ... in a snowy park
LDRE (Ours)
Text-to-image Retrieval
visual over-consistency
semantic bias
ground truth

Figure 1. Comparison of composed image retrieval among existing methods. Red words are semantically irrelevant to the target image and green words semantically match the target image. The color depth of the text boxes represents the weight considered by our LDRE.

To conduct ZS-CIR, most works leverage the cross-modal alignment ability of large-scale pre-trained Vision-Language Models (VLMs) (e.g., CLIP [34]). Some textual inversion methods [3, 9, 36] use image-caption pairs to learn to map images into text pseudo tokens. However, their method still relies on training images, which may compromise the generalization in the test scenarios. Furthermore, the presence of excessive visual features in pseudo tokens results in the composed caption being overly saturated with details, content, texture, and style of the reference image. As shown in Figure 1, since the target image typically does not strictly align with the reference image, this visual over-consistency leads to sub-optimal retrieval performance. In contrast to textual inversion methods, there is a new exploration of employing LLMs for combination. Recently, Karthik et al. [21] propose a simple method that captions the reference image using a pre-trained captioning model and employs an LLM to recompose the caption based on the modification text for subsequent retrieval. To sum up, the existing methods for ZS-CIR mostly generate a single edited caption based on the reference image and modification text, which is subsequently projected into the CLIP feature space to retrieve the target image.

However, most of the existing methods neglect that ZS-CIR (as well as CIR) is a typical fuzzy retrieval [6, 8, 37] task, where the semantics of the target image are not strictly defined by the query image and text. Such fuzziness results from the heterogeneity of the multi-modal query. While the modification text is given to modify the semantics of the reference image, the query does not explicitly specify which visual objects should be modified, which ones should be retained, and which ones can be omitted. Therefore, as exemplified in Figure 1, such semantic fuzziness can lead to diverse possible semantics of the retrieval target. For instance, the modification text does not explicitly state whether the horse in the target image should remain white. However, the LLM-based method may describe a white horse in the target image, which can introduce semantic irrelevance. Therefore, merely generating an edited caption may fail to capture the diversity of possible composed results, leading to compromised retrieval performance for ZS-CIR. Inspired by the above observation, we aim to address **Challenge 1**: How to effectively generate a variety of edited captions that can cover diverse possible semantics of the composed results?

Moreover, once diverse edited captions are obtained, integrating them for the final image retrieval is not a trivial problem. Since the existing pre-trained aligned VLMs typically have a short input text limitation (e.g., CLIP [34] can only encode a maximum of 75 textual tokens), simply concatenating all edited captions and encoding the whole text is infeasible for ZS-CIR. Furthermore, diverse edited captions can simultaneously bring the noise problem, where the quality and significance of different edited captions are uneven. As exemplified in Figure **??**, the last caption "a man riding a white horse ... in a snowy park" in our LDRE lacks matching images in the database, where the semantic irrelevance to the target image caused by the phrase "white horse" can be regarded as noisy. Therefore, simply combining all edited captions equally may trivialize the salient semantic information and generate a degraded semantic feature. Based on the above discussion, we have to tackle **Challenge 2**: How to effectively integrate diverse edited captions while capturing important semantic information and filtering out noisy ones?

To deal with the above challenges, we propose a novel LLM-based Divergent Reasoning and Ensemble (LDRE) method for Zero-Shot Composed Image Retrieval (ZS-CIR), capturing the diverse possible semantics of the composed target. For **Challenge 1**, we propose an LLM-based divergent compositional reasoning to generate diverse edited captions, covering possible semantics of the composed results for fuzzy retrieval. Our approach draws inspiration from divergent thinking [10,29,35] extensively studied in psychology, which refers to a thought process used to generate creative ideas by exploring many possible solutions. To conduct the divergent compositional reasoning, we generate dense captions for the reference image with focuses on different semantic perspectives. Then, we prompt an LLM to infer divergent edited captions based on the modification text, describing possible composed images of diverse semantics. For **Challenge 2**, we propose a divergent caption ensemble to integrate complementary information in divergent edited captions and filter out noise. We compute the semantic relevance score for every edited caption to measure the relatively reduced similarity incurred by the semantic edition toward the images in the database. The semantic relevance scores can be regarded as the distinguishing scores of every edited caption, weighted by which an ensemble caption feature is obtained by CLIP. Finally, the composed image can be retrieved based on the cosine similarity in the CLIP feature space. Extensive experiments on three benchmark datasets for ZS-CIR indicate significant performance improvements of our proposed LDRE compared with the state-of-the-art methods.

In summary, our contributions can be summarized as follows:

- We propose a novel LLM-based Divergent Reasoning and Ensemble (LDRE) method for Zero-Shot Composed Image Retrieval (ZS-CIR), which can utilize out-of-shelf tools to accurately retrieve a composed image based on a reference image and a modification text without training.

- We propose LLM-based divergent compositional reasoning to generate diverse edited captions from different semantic perspectives of the reference image. The divergent edited captions can effectively cover the possible semantics of the composed results, overcoming the fuzzy nature of ZS-CIR.

- We propose a divergent caption ensemble to integrate complementary information in divergent edited captions and filter out noise. Specifically, we design a semantic relevance scorer to measure the distinguishing scores of edited captions, weighted by which a single CLIP feature is computed for retrieval.

- Extensive experiments conducted on three benchmark datasets demonstrate significant performance improvements of the proposed LDRE compared with the state-of-the-art methods for ZS-CIR.

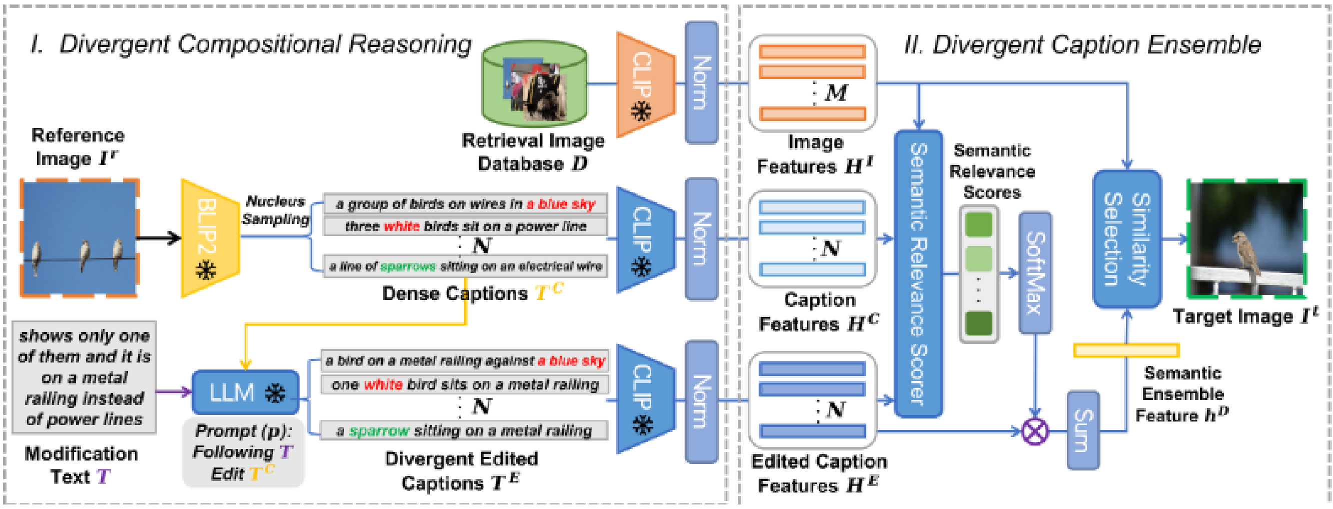

I. Divergent Compositional Reasoning
Reference Image $I^r$
BLIP2
Nucleus Sampling
a group of birds on wires in a blue sky
three white birds sit on a power line
N
a line of sparrows sitting on an electrical wire
Dense Captions $T^C$
Retrieval Image Database $D$
CLIP
Norm
shows only one of them and it is on a metal railing instead of power lines
Modification Text $T$
LLM
Prompt (p): Following $T$ Edit $T^C$
a bird on a metal railing against a blue sky
one white bird sits on a metal railing
N
a sparrow sitting on a metal railing
Divergent Edited Captions $T^E$
II. Divergent Caption Ensemble
M
Image Features $H^I$
N
Caption Features $H^C$
N
Edited Caption Features $H^E$
Semantic Relevance Scorer
Semantic Relevance Scores
SoftMax
Sum
Semantic Ensemble Feature $h^D$
Similarity Selection
Target Image $I^t$

Figure 2. Architecture of the proposed LDRE method: (1) The LLM-based divergent compositional reasoning is designed to generate diverse edited captions from different semantic perspectives of the reference image, which can effectively cover the possible semantics of the composed result, in response to the fuzzy nature of ZS-CIR; (2) The divergent caption ensemble is designed to integrate complementary information in divergent edited captions and filter out noise for the final retrieval.

## 2 Related works

### 2.1 Composed Image Retrieval

The task of Composed Image Retrieval (CIR) [7,13,23,42], which involves retrieving a target image based on a query composed of a reference image and a modification text, expands upon cross-modal retrieval [31–33] and has seen a surge of interest in recent years. CIR integrates the concepts of compositional learning [19,22] and image retrieval, creating a unique and challenging task, which has found substantial applications in conditional search [5] and fashion styling [44]. Text-Image Residual Gating (TIRG) [42], for instance, leverages ResNet [16] for image feature extraction and LSTM [17] for text feature extraction. Subsequently, a gating and residual module synthesizes these multi-modal query features and adaptively generates the composed query feature. However, traditional CIR methods rely on annotated data, which is complicated and requires extensive labor in CIR.

Therefore, Zero-Shot Composed Image Retrieval (ZS-CIR) [3, 36, 39] has recently attracted intensive attention, with the object of developing generalized CIR models without annotated data. Existing ZS-CIR methods typically convert the image modality into a text modality using methods such as a captioning model or textual inversion. Some existing methods [3,36] use image-caption pairs to train textual inversions mapping images to text tokens. Recently, Karthik et al. [21] propose a training-free method that captions the reference image using a pre-trained VLM and employs an LLM to recompose the caption based on the modification text for subsequent retrieval. However, merely generating one edited caption may fail to capture the diversity of possible composed results, leading to compromised retrieval performance for ZS-CIR. To overcome this problem, we propose an LLM-based Divergent Reasoning and Ensemble (LDRE) method to generate and integrate diverse edited captions, covering possible semantics of the composed results for fuzzy retrieval.

### 2.2 Vision-Language Models for CIR

The success of the pre-trained BERT [12] model has inspired pretrained Vision-Language Models (VLM) [25, 28, 38] and promoted various downstream tasks [45, 46]. The goal is to create Transformer-based [40] models trained on large-scale image-text triplets to produce vision-and-language representations that can be applied to various tasks. For CIR, to map images and text into a shared embedding space, many methods harness large pre-trained multi-modal models, such as CLIP [34], as the backbone for feature extraction. These models [4, 15, 21] have recently gained popularity due to their exceptional ability to handle multi-modal data, resulting in superior performance. For example, Baldrati et al. [4] utilize CLIP to extract both image and text features and then employ a simple combiner module to amalgamate the multi-modal query features, achieving commendable retrieval performance. Recently, Han et al. [15] further advance this approach by designing a unified visual-linguistic model capable of managing multiple multi-modal learning tasks, including CIR. They adapt the model to various tasks using different cross-attention adaptors and achieve leading performance in CIR by capitalizing on the large model and multi-task learning.

Further advancements have been made with models such as BLIP [24] and CoCa [47]. These models move beyond shared space projection to tackle other vision-language tasks, such as captioning [41] and visual question answering [2]. While these models have been indirectly utilized for CIR via specialized modules [5, 11, 42] and through fine-tuning [14], our research demonstrates that when standalone vision-language models are combined with an LLM, they can effectively carry out CIR without requiring extra training.

## 3 Method

As mentioned in the problem statement, the input of CIR is a multi-modal query $q = (I^r, T)$, where $I^r$ and $T$ denote the reference image and the modification text, respectively. We leverage a combination of visual and textual information for retrieval. The reference image serves as the visual feature, while the modification text provides additional context or constraints for the retrieval process. Our objective is to find the target images that not only match the modification text but also capture the essence of the reference image. To this end, we propose an LLM-based Divergent Reasoning and Ensemble (LDRE) method for ZS-CIR. The overall architecture of LDRE is shown in Figure **??**, which consists of divergent compositional reasoning and divergent caption ensemble.

### 3.1 Divergent Compositional Reasoning

To ensure comprehensive coverage of possible semantics of the composed results for fuzzy retrieval, we propose LLM-based divergent compositional reasoning to generate diverse edited captions, which includes a dense caption generator for generating dense captions and a multi-prompt editing reasoner for reasoning and editing.

### 3.1.1 Dense Caption Generator

We employ a language-for-vision approach for composed image retrieval. Instead of relying on visual features, we leverage a pre-trained captioning model $\Psi(\cdot)$, such as GIT [47], CoCa [52], or BLIP [25], to generate natural language captions of the reference image. In order to obtain dense captions, we employ nucleus sampling [19] during the caption generation process to enhance the diversity of the generated captions. The generation of dense captions can be formulated as follows:

$$\mathcal{T}^C = \{T_1^C, \ldots, T_N^C\} = \Psi(I^r),$$

where $T_i^C$ is the $i$-th generated caption of the reference image, $\mathcal{T}^C$ denotes the set of dense captions, and $N$ is the number of generated captions.

### 3.1.2 Multi-Prompt Editing Reasoner

Directly relying on the modification text as the sole textual input is insufficient, as it merely provides relative modification information towards the target image. It lacks the inclusion of essential reference image details and context. Similarly, depending solely on the reference image caption is inadequate, as it fails to incorporate the crucial contextual information provided by the modification text. Although simple combination templates, such as "a photo of {reference caption} that {modification text}", can combine them in a fixed manner, they lack the flexibility to accommodate diverse forms of modification text and determine the most suitable caption format for a specific query. To address this issue, we harness the reasoning capabilities of existing LLMs. Instead of merging the reference image caption and modification text in a fixed template, our objective is to derive cohesive, unified, and divergent edited captions by LLMs.

Formally, given the dense captions $\mathcal{T}^C$ of the reference image and modification text $T$, we design a simple prompt template $pt(\cdot,\cdot)$ inspired by [29], combining the dense captions and modification text to create the full prompts $\mathcal{P}$ for LLM:

$$\mathcal{P} = \{p_i = pt(T_i^C, T) \,|\, 0 \leq i < N\},$$

where $p_i$ represents the prompt that combines the reference image caption and modification text. We fill these two parts into the template to get the full prompt. Then, we input the prompts into the LLM for reasoning and obtain the generated edited captions $\mathcal{T}^E$:

$$\mathcal{T}^E = \{T_i^E = \text{LLM}(p_i) \,|\, 0 \leq i < N\}.$$

### 3.1.3 Alignment Feature Extractor

To retrieve the target image from the image database using edited captions, we need to extract features from both modalities and align them in a shared feature space. To accomplish this, we use the image and text encoders from large-scale vision-language models (e.g., CLIP) to process the candidate image database and edited captions, respectively.

# 4 Implementation details

## 4.1 Comparing with the released source codes

1. Availability of Source Codes

The source code of the original method is publicly available on GitHub https://github.com/yzy-bupt/LDRE. We used this code as a baseline for our implementation.

2. Our Work

- **Reproduction**: We faithfully reproduced the original code structure based on the released code and the paper. However, we identified several issues in the original implementation, such as *raw GPT prompt cannot generate accurate text* and *raw method cannot attend to global information*, which we attempted to address in our work.
- **Improvements**:
  - We improved the prompt of GPT to generate more accurate text.
- **Creative Additions**:
  - We proposed a novel code for Open-Vocabulary detection to enhance the attention to global information, which was not considered in the original work.

## 4.2 Experimental environment setup

### 4.2.1 Dataset

We compared our LDRE with the latest baselines from three public datasets: CIRCO [3] and CIRR [27], which have been widely used for CIR. CIRR is the first natural image dataset specifically designed for CIR. It is based on real images in open domains. CIRCO is based on the unlabeled set of real-world images from COCO 2017 [26], and is the first dataset provided by CIR with multiple fundamental facts. On average, CIRCO provided 123,403 images in the retrieval database, with 4.53 basic facts per query, providing a more robust and comprehensive evaluation for the CIR model.

### 4.2.2 Evaluation Indicators

On CIRCO, assuming that each query has multiple target images as ground truth values, we use Mean Average Precision (mAP), which is a finer-grained metric to consider the ranking of retrieval results. On CIRR, following the initial benchmark testing, we use Recall@$k$ ($k \in \{1, 5, 10, 50\}$) as the primary metric to represent the percentage of target images included in the top $k$ lists.

### 4.2.3 Implementation Details

For the subtitle model, we use pre-trained Florence-2. We use nucleus sampling [18] in the subtitle generation process to generate dense subtitles, and set the number of subtitles $N$ to 15. For VLM, we use OpenCLIP's

ViT-G/14 CLIP [20]. For LLM, we default to using GPT-4 [1]. The temperature $\tau$ in adaptive semantic integration is set to 0.04. For ViT-G/14, the dimension of the feature space is 1024. The entire model is implemented using PyTorch [30] and NVIDIA 3090 GPU.

## 4.3 Main contributions

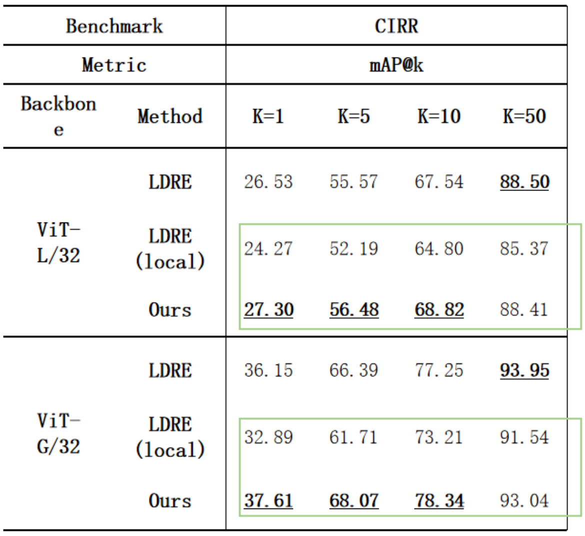
### 4.3.1 Improvement 1

Due to the lack of accuracy in the text description of the target text generated by the prompt in the original paper, it paid too much attention to useless background information, which resulted in assigning weights to insignificant background information in the image-text matching process, leading to an increase in the matching result error. Therefore, we have modified the prompt mode of GPT to focus more on generating textual descriptions of the main entities and ignore irrelevant background information.

### 4.3.2 Improvement 2

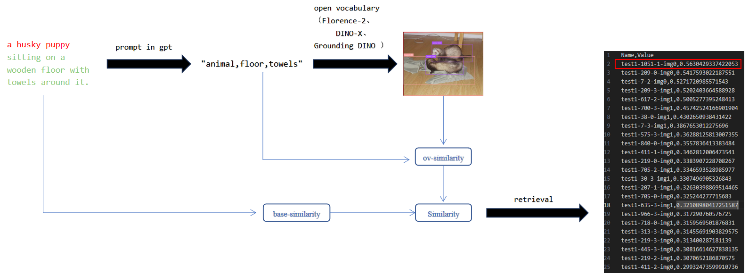
For the original method, due to the low alignment between text and image dimensions, the matching process naturally focuses on entity information with higher weights and ignores global entity information, resulting in matching results biased towards a certain feature, which has a significant impact on accurate matching. Therefore, we are trying to find a method that can enhance global matching while continuing to ensure the model's untrained inference characteristics. To address this, we attempt to use open-vocabulary object detection methods to perform open-domain object detection on entities appearing in text descriptions, penalize samples that detect missing entities, and filter out matching results with average global information to avoid errors caused by a high matching degree of a certain feature.

# 5 Results and analysis



| Benchmark | | CIRR | | | |
|---|---|---|---|---|---|
| Metric | | mAP@k | | | |
| Backbone | Method | K=1 | K=5 | K=10 | K=50 |
| ViT-L/32 | LDRE | 26.53 | 55.57 | 67.54 | 88.50 |
| | LDRE (local) | 24.27 | 52.19 | 64.80 | 85.37 |
| | Ours | 27.30 | 56.48 | 68.82 | 88.41 |
| ViT-G/32 | LDRE | 36.15 | 66.39 | 77.25 | 93.95 |
| | LDRE (local) | 32.89 | 61.71 | 73.21 | 91.54 |
| | Ours | 37.61 | 68.07 | 78.34 | 93.04 |

Figure 3. Experimental results: LDRE refers to the method in the original paper, LDRE (local) refers to the results we reproduced, and ours refers to the results achieved through our improvements.



a husky puppy sitting on a wooden floor with towels around it.
prompt in gpt
"animal,floor,towels"
open vocabulary（Florence-2、DINO-X、Grounding DINO ）
ov-similarity
base-similarity
Similarity
retrieval

```
Name,Value
test1-1051-1-img0,0.5630429337422053
test1-209-0-img0,0.5417593022187551
test1-7-2-img0,0.527172098557154 3
test1-209-3-img1,0.5202403664588928
test1-617-2-img1,0.5005277395248413
test1-700-3-img1,0.45742524166901904
test1-38-0-img1,0.4302650938431422
test1-7-3-img1,0.3867653012275696
test1-575-3-img1,0.36288125813007355
test1-840-0-img0,0.3557836413383484
test1-411-1-img0,0.3462812006473541
test1-219-0-img0,0.3383907228708267
test1-705-2-img1,0.3346593528985977
test1-30-3-img1,0.330749690532 6843
test1-207-1-img1,0.32630398869514465
test1-705-0-img0,0.325244277715683
test1-635-3-img1,0.3210898041725158 7
test1-966-3-img0,0.317290760576725
test1-718-0-img1,0.3159569501876831
test1-313-3-img0,0.31455691903829575
test1-219-3-img0,0.313400287181139
test1-445-3-img0,0.30816614627838135
test1-219-2-img1,0.3070652186870575
test1-411-2-img0,0.29932473599910736
```

Figure 4. Open-Vocabulary Detection: The above is the process of conducting experiments using open vocabulary detection method on samples that are difficult to distinguish with the original method. The results showed that the highest matching degree was achieved on this sample.

## 6 Conclusion and future work

Due to various reasons, our replication work did not fully achieve the results in the original paper. However, we made improvements based on the replication results, and the results showed an improvement of about 4% compared to before the improvement, as shown in Figure 3. Therefore, it can be proven that our method can effectively alleviate the troubles caused by redundant background information. For the problem of insufficient global attention, we conducted a large number of experiments using open vocabulary detection methods. The process and results of the experiment on difficult to match samples are shown in Figure 4. The samples that could not be successfully matched before achieved the top matching degree under this method. Therefore, our method is effective in some samples. However, due to the significant interference of background information or dataset data, it is difficult for us to obtain completely accurate textual descriptions on the overall dataset. Therefore, we cannot achieve breakthroughs at the dataset level through open vocabulary detection accurately. Currently, we are exploring solutions.

## References

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

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015.

[3] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *arXiv preprint arXiv:2303.15247*, 2023.

[4] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Conditioned and composed image retrieval combining and partially fine-tuning clip-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4959–4968, 2022.

[5] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21466–21474, 2022.

[6] Gloria Bordogna and Gabriella Pasi. A fuzzy linguistic approach generalizing boolean information retrieval: A model and its evaluation. *Journal of the American Society for Information Science*, 44(2):70–82, 1993.

[7] Yanbei Chen, Shaogang Gong, and Loris Bazzani. Image search with text feedback by visiolinguistic attention learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3001–3011, 2020.

[8] Yixin Chen and James Ze Wang. A region-based fuzzy feature matching approach to content-based image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1252–1267, 2002.

[9] Niv Cohen, Rinon Gal, Eli A. Meirom, Gal Chechik, and Yuval Atzmon. "this is my unicorn, fluffy": Personalizing frozen vision-language representations. In *European Conference on Computer Vision*, pages 558–577, 2022.

[10] A.J. Cropley. Divergent thinking and science specialists. *Nature*, 215(5101):671–672, 1967.

[11] Ginger Delmas, Rafael Sampaio de Rezende, Gabriela Csurka, and Diane Larlus. Artemis: Attention-based retrieval with text-explicit matching and implicit similarity. In *arXiv preprint arXiv:2203.08101*, 2022.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *arXiv preprint arXiv:1810.04805*, 2018.

[13] Chunbin Gu, Jiajun Bu, Zhen Zhang, Zhi Yu, Dongfang Ma, and Wei Wang. Image search with text feedback by deep hierarchical attention mutual information maximization. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4600–4609, 2021.

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

[15] Xiao Han, Xiatian Zhu, Licheng Yu, Li Zhang, Yi-Zhe Song, and Tao Xiang. Fame-vil: Multi-tasking vision-language model for heterogeneous fashion tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2669–2680, 2023.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[17] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[18] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *arXiv preprint arXiv:1904.09751*, 2019.

[19] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 584–600, 2020.

[20] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021.

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

[22] Jongseok Kim, Youngjae Yu, Hoeseong Kim, and Gunhee Kim. Dual compositional learning in interactive image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1771–1779, 2021.

[23] Seungmin Lee, Dongwan Kim, and Bohyung Han. Cosmo: Content-style modulation for image retrieval with text feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 802–812, 2021.

[24] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900, 2022.

[25] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. In *arXiv preprint arXiv:1908.03557*, 2019.

[26] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. *CoRR*, abs/1405.0312, 2014.

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

[28] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[29] Robert R. McCrae. Creativity, divergent thinking, and openness to experience. *Journal of Personality and Social Psychology*, 52(6):1258, 1987.

[30] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[31] Shengsheng Qian, Dizhan Xue, Quan Fang, and Changsheng Xu. Adaptive label-aware graph convolutional networks for cross-modal retrieval. *IEEE Transactions on Multimedia*, 24:3520–3532, 2021.

[32] Shengsheng Qian, Dizhan Xue, Quan Fang, and Changsheng Xu. Integrating multi-label contrastive learning with dual adversarial graph neural networks for cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4794–4811, 2022.

[33] Shengsheng Qian, Dizhan Xue, Huaiwen Zhang, Quan Fang, and Changsheng Xu. Dual adversarial graph neural networks for multi-label cross-modal retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2440–2448, 2021.

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021.

[35] Mark A. Runco. *Divergent Thinking*. Ablex Publishing Corporation, Norwood, NJ, 1991.

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

[37] Valiollah Tahani. A fuzzy model of document retrieval systems. *Information Processing & Management*, 12(3):177–187, 1976.

[38] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *arXiv preprint arXiv:1908.07490*, 2019.

[39] Yuanmin Tang, Jing Yu, Keke Gai, Zhuang Jiamin, Gang Xiong, Yue Hu, and Qi Wu. Context-i2w: Mapping images to context-dependent words for accurate zero-shot composed image retrieval. In *arXiv preprint arXiv:2309.16137*, 2023.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[41] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):652–663, 2016.

[42] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval–an empirical odyssey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6439–6448, 2019.

[43] Haokun Wen, Xian Zhang, Xuemeng Song, Yinwei Wei, and Liqiang Nie. Target-guided composed image retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 915–923, 2023.

[44] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11307–11317, 2021.

[45] Dizhan Xue, Shengsheng Qian, Quan Fang, and Changsheng Xu. Mmt: Image-guided story ending generation with multimodal memory transformer. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 750–758, 2022.

[46] Dizhan Xue, Shengsheng Qian, and Changsheng Xu. Variational causal inference network for explanatory visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2515–2525, 2023.

[47] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. In *arXiv preprint arXiv:2205.01917*, 2022.

[48] Gangjian Zhang, Shikui Wei, Huaxin Pang, and Yao Zhao. Heterogeneous feature fusion and cross-modal alignment for composed image retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5353–5362, 2021.