VQA4CIR: Boosting Composed Image Retrieval with Visual Question Answering

Chun-Mei Feng[1], Yang Bai[1*], Tao Luo[1], Zhen Li[2], Salman Khan[3,4]
Wangmeng Zuo[5], Rick Siow Mong Goh[1], Yong Liu[1]
[1]Institute of High Performance Computing (IHPC),
Agency for Science, Technology and Research (A*STAR), Singapore
[2]SSE, The Chinese University of Hong Kong, Shenzhen (CUHK), China
[3]Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), UAE
[4]Australian National University, Canberra ACT, Australia
[5]Harbin Institute of Technology, Harbin, China
fengcm.ai@gmail.com

Abstract

Albeit progress has been made in Composed Image Retrieval (CIR), we empirically find that a certain percentage of failure retrieval results are not consistent with their relative captions. To address this issue, this work provides a Visual Question Answering (VQA) perspective to boost the performance of CIR. The resulting VQA4CIR is a post-processing approach and can be directly plugged into existing CIR methods. Given the top-C retrieved images by a CIR method, VQA4CIR aims to decrease the adverse effect of the failure retrieval results being inconsistent with the relative caption. To find the retrieved images inconsistent with the relative caption, we resort to the "QA generation $\rightarrow$ VQA" self-verification pipeline. For QA generation, we suggest fine-tuning LLM (*e.g.*, LLaMA) to generate several pairs of questions and answers from each relative caption. We then fine-tune LVLM (*e.g.*, LLaVA) to obtain the VQA model. By feeding the retrieved image and question to the VQA model, one can find the images inconsistent with relative caption when the answer by VQA is inconsistent with the answer in the QA pair. Consequently, the CIR performance can be boosted by modifying the ranks of inconsistently retrieved images. Experimental results show that our proposed method outperforms state-of-the-art CIR methods on the CIRR and Fashion-IQ datasets.

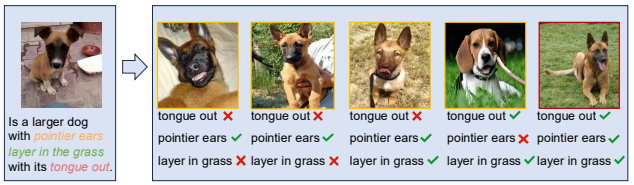
Code — https://github.com/chunmeifeng/VQA4CIR

Introduction

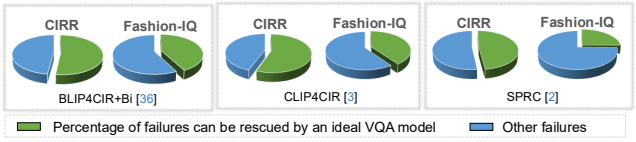
Composed Image Retrieval (CIR) (Liu et al. 2021; Vo et al. 2019; Baldrati et al. 2022b) is a challenging retrieval task, where a reference image together with a relative caption are combined to retrieve the desired target image. Benefiting from its dual-modal query nature, CIR can offer a nuanced depiction of the desired image, allowing for interactive refinements by tweaking the reference image and description. Such capabilities make CIR especially suited for applications like e-commerce and digital search (Feng et al. 2020).

In the recent few years, considerable progress has been made in CIR. For example, many late fusion methods (Anwaar, Labintcev, and Kleinsteuber 2021; Dodds et al. 2020; Liu et al. 2021; Vo et al. 2019) have been developed to integrate the information from reference image and relative caption. Pseudo-word embedding (Gal et al. 2022; Baldrati et al. 2022b) is shown to convert reference image into description embedding for combining with the relative caption to retrieve the target image. Taking both reference images and relative captions into account, sentence-level prompts have also been proposed to enhance CIR (Bai et al. 2024).

However, for most existing CIR methods, we empirically find that a certain percentage of retrieval results are not consistent with their relative captions. For example, in Fig. 1 (a), given the relative caption "`Is a larger dog with pointier ears laying in the grass with its tongue out`", the rank-1 image retrieved by SPRC (Bai et al. 2024) does not satisfy the attributes



(a) Examples of failed retrieval that can be identified by simply comparing the relative caption with candidate images.



(b) Statistical analysis on the failure cases for BLIP4CIR+Bi [36], CLIP4CIR [3], and SPRC [2].

Figure 1: **Failure cases analysis**. **(a)** Examples of failure retrieval results of SPRC (Bai et al. 2024) on CIRR dataset, which one can see that they can be identified by comparing with relative caption, **(b)** is the statistical analysis of the percentage of failures from BLIP4CIR+Bi (Liu et al. 2023b), CLIP4CIR (Baldrati et al. 2022a), and SPRC (Bai et al. 2024) on the CIRR and Fashion-IQ datasets, where one can see that a *certain percentage* of failure cases can be ascribed to the *inconsistency* between retrieved images and relative captions.

*Corresponding author.

lying in the grass, and tongue out. In Fig. 1 (b), we summarize the percentage of causes of failures from randomly selected 150 failure cases of the state-of-the-art methods BLIP4CIR+Bi (Liu et al. 2023b), CLIP4CIR (Baldrati et al. 2022a), and SPRC (Bai et al. 2024) from the validation set on CIRR and Fashion-IQ datasets. Among these, 52.2%, 55.8% and 48.2% of failures on CIRR, 42.5%, 38.3% and 25.3% of the failures on Fashion-IQ, can be attributed to the inconsistency between the retrieved images and relative captions. Motivated by the above result, this paper aims to present a post-processing method to find the retrieved images that are inconsistent with the relative captions, and to modify their ranks for boosting CIR performance.

To find the retrieved images being inconsistent with relative captions, we resort to the "QA generation $\rightarrow$ VQA" self-verification pipeline, resulting in our VQA4CIR method (see Fig. 2). In QA generation, we aim to generate question $\mathbf{Q}_i$ and answer $\mathbf{A}_i$ pairs from relative caption. Motivated by the unprecedented success of large language models (LLMs), we adopt the open-sourced LLaMA (Touvron et al. 2023) and fine-tune it to fulfill our requirements. To construct the instruction dataset, we use GPT-4 (OpenAI 2023) to generate QA pairs, and modify parts of low-quality QA pairs in a handcrafted manner. Using the instruction dataset, we use the pre-defined instruction prompt, and adopt LoRA (Hu et al. 2021) for fine-tuning LLaMA (Touvron et al. 2023). In VQA, we feed each question $\mathbf{Q}_i$ and the retrieved image into the VQA model to generate an answer $\mathbf{A}'_i$. For training the VQA model, we fine-tune large vision-language models (*e.g.*, LLaVA (Liu et al. 2023a)) using the instruction data. When all $\mathbf{A}'_i$s are equal to the corresponding $\mathbf{A}_i$s, the retrieved image will be regarded to be consistent with relative caption. By modifying the ranks of the inconsistent retrieved images, we rerank the retrieved images to boost CIR performance.

Our proposed VQA4CIR is a post-processing approach and can be directly plugged into existing CIR methods for better CIR performance. Using the top-4 retrieved results by SPRC (Bai et al. 2024) in Fig. 1 (a) as an example, from the relative caption, one can use LLM to generate three QA pairs in Fig. 2. Obviously, for each of the three top-rank retrieved images, at least one of the answers is not consistent with the answer in the QA pairs. For all questions, the answers are consistent in the target images and QA pairs. Thus, we can modify the ranks of the first three retrieved images to be later compared with the target image, thereby improving the CIR performance. Extensive experiments are conducted on the CIRR and Fashion-IQ datasets. The results show that our VQA4CIR can be incorporated with different CIR methods and outperforms the state-of-the-art CIR methods.

To sum up, the contributions of this work are three-fold:

- By analyzing the failure CIR retrieved results, we suggest a *VQA perspective* for boosting the performance of existing CIR approaches, resulting in our VQA4CIR method.
- Following the "*QA generation $\rightarrow$ VQA*" *self-verification pipeline*, we fine-tune LLM to generate QA pairs from the relative caption, and fine-tune LVLM to answer the questions. By finding and modifying the ranks of the inconsistent retrieved images, we rerank the retrieved images to attain better CIR performance.
- Experimental results show that our VQA4CIR outperforms the state-of-the-art CIR methods and can be directly plugged into existing CIR methods.

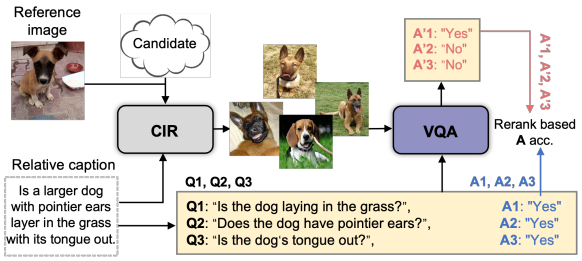


Figure 2: **Illustration** of the ***main idea*** of VQA4CIR, which converts the relative captions into multiple QA pairs, and uses the VQA model to respond to each candidate image. Finally, *reranking* is conducted on the candidate images by comparing the answers of the VQA model and QA pairs.

Related Work

Composed Image Retrieval. Early CIR techniques primarily employed a late fusion strategy to combine features from a reference image with its relative caption. Then, the merged features were compared with all candidate image features to retrieve the most matched image among all candidates in an extensive image corpus (Anwaar, Labintcev, and Kleinsteuber 2021; Dodds et al. 2020; Liu et al. 2021; Vo et al. 2019). Subsequent developments introduced various feature fusion methods (Vo et al. 2019) and attention mechanisms (Dodds et al. 2020), and exhibited impressive performance in CIR tasks. Recently, many CIR techniques began to leverage the abilities of pre-trained models to enhance image and text features (Ventura et al. 2023; Baldrati et al. 2022a; Liu et al. 2023c; Gu et al. 2023a; Ray et al. 2023; Liu et al. 2023b; Goenka et al. 2022). Another line of techniques adopted the 'text inversion' approach to transform a reference image into pseudo-word embeddings, which are then combined with their relative captions, facilitating text-to-image retrieval. Bai *et al*. employed both the reference image and relative caption to derive sentence-level prompts, aiming to enrich the caption by offering a proper description of pertinent elements within the reference image (Bai et al. 2024). However, existing CIR methods are limited in retrieving the target images, and a certain percentage of failure retrieval results even are not consistent with their relative captions (see Fig. 1). Thus, we suggest using VQA4CIR to find these inconsistent images and rerank the retrieved images. While Liu *et al*. also aims to mitigate errors from the first stage through a re-ranking mechanism (Liu et al. 2023c), it merely concatenates two traditional CIR methods without addressing the underlying issue of CIR.

LLMs and LVLMs. In the recent few years, LLMs have demonstrated remarkable capabilities in language genera-

tion, contextual learning, world knowledge, and reasoning. The GPT familes, *e.g.*, ChatGPT (OpenAI 2022), GPT-4 (OpenAI 2023), and InstructGPT (Ouyang et al. 2022), stands as the most significant achievements in LLMs. Open-source models like OPT (Zhang et al. 2022), LLaMA (Touvron et al. 2023), MOSS (Sun et al. 2023), Alpaca (Taori et al. 2023), and Vicuna (Chiang et al. 2023) served as valuable resources that allowed fine-tuning for specific domains or tasks. Most recently, a plethora of research has focused on extending LLMs into LVLMs, *e.g.*, Minigpt-4 (Zhu et al. 2023), LLaVA (Liu et al. 2023a), instructBLIP (Dai et al. 2023). Typically, LVLMs comprise a visual encoder, a language encoder (*i.e.*, LLM), and a cross-modal alignment network. Many efficient vision-text interactions (Li et al. 2023a), efficient training methods (Gao et al. 2023; Zhang et al. 2023a), and instruction tuning (Liu et al. 2023a; Zhu et al. 2023; Dai et al. 2023) methods have been proposed. LVLMs are adept at intricate reasoning surrounding these objects, culminating in enhanced outcomes in diverse multimodal challenges by Visual Question Answering (VQA). As such, we resort to the powerful performance of LLM and LVLM to conduct complex reasoning on relative captions and retrieved images through QA generation and VQA, thereby offering a new perspective for improving VQA performance.

Downstreaming Applications of LLMs and LVLMs. LLMs and LVLMs have demonstrated exceptional and innovative performance in various downstream few-shot vision learning scenarios (Chen et al. 2022; Feng et al. 2023c, 2024, 2023a). For example, in the realm of object detection, leveraging prompt embeddings to fine-tune the LVLMs via prompt learners can lead to cutting-edge results (Gu et al. 2023c). In image segmentation, LVLMs not only enhance the performance of open vocabulary but also allow the segmentation model to inherit and utilize the language generation capabilities of LLM (Bangalath et al. 2022; Liang et al. 2023). For video understanding, LVLMs can help boost the capability of understanding and generating human-like conversations about videos (Li et al. 2023b; Maaz et al. 2023). For 3D representations, LVLMs features can be used to encode semantics in 3D representations (Gu et al. 2023b). Inspired by the outstanding performance of LLMs and LVLMs in various domains, this work explores the potential of leveraging LLM and LVLM to enhance the CIR performance.

Methodology

Overview. In CIR, the multimodal composite query $\{\boldsymbol{I}_r, \boldsymbol{t}\}$ involves a reference image $\boldsymbol{I}_r$ and a relative caption $\boldsymbol{t}$. The essence of CIR lies in retrieving the target image from an extensive image corpus $\mathcal{D}$, relying on the content of both the reference image and relative caption. Given that the target image needs to capture the changes in objects and attributes described in relative caption while preserving visual resemblances to the reference image, this positions CIR as more challenging than conventional text-to-image retrieval.

In this work, we suggest a new perspective on CIR through the lens of VQA, resulting in our VQA4CIR. VQA4CIR can be incorporated with any existing CIR methods. First, by finetuning LLM, we obtain a QA generation model to generate QA pairs $\{(\mathbf{Q}_1, \mathbf{A}_1), ..., (\mathbf{Q}_K, \mathbf{A}_K)\}$ from the relative caption $\boldsymbol{t}$. Using any CIR method, we can get its top-C rank candidate images $\{\boldsymbol{I}_1, ..., \boldsymbol{I}_c, ..., \boldsymbol{I}_C\}$. Then, by finetuning LVLM, we obtain the VQA model. For each retrieved image $\boldsymbol{I}_c$ and question $\mathbf{Q}_k$, the VQA model takes $(\boldsymbol{I}_c, \mathbf{Q}_k)$ as the input to generate the answer $\mathbf{A}'_k$. When $\mathbf{A}'_k$ is not equal to $\mathbf{A}_k$, we can treat $\boldsymbol{I}_c$ to be inconsistent with relative caption and modify its rank. In this way, the adverse effect of inconsistent retrieved images can be suppressed and better CIR performance can be attained. In the following, we will introduce the QA generation, VQA, and the inference process in detail.

Generating QAs from Relative Caption

LLM is powerful in many natural language processing tasks but should be finetuned to match the requirements of VQA4CIR in generating QA pairs from relative captions. To this end, we construct an instruction dataset and finetune LLaMA as follows.

Construction of Instruction Dataset. To train LLaMA (Touvron et al. 2023) for generating QA pairs from relative captions, we require a substantial number of question-and-answer pairs $\mathbf{x}_{\texttt{QAs}}$ as training data. Motivated by the great success of GPT models in text generation (Gilardi, Alizadeh, and Kubli 2023), we employ GPT-4 to ease the cost of labor-intensive manual annotations of QA pairs. In general, the generated QA pairs should cover all the contents of the relative caption and should have a small number. We also recommend the answer to be '`yes`' or '`no`' for clarity. To fulfill these requirements, we formulated a set of prompts for GPT-4 (refer to the *Suppl.* for further details). Nonetheless, a small percentage of QA pairs generated by GPT-4 are of poor quality. And we further conducted a manual review to refine and remove them. The resulting instruction dataset is presented in **`JSON`** format, with '`QA Pairs`' as the main key, 'Q' for questions, and 'A' for answers, ensuring that all data originates from the relative caption. In our experiments, we selected $5,000$ and $3,000$ samples to construct the instruction data for the CIRR and Fashion-IQ datasets, respectively.

Fine-tune LLaMA. LLaMA (Touvron et al. 2023) is an open-source language model and thus can be finetuned to enhance performance on specialized tasks. With the instruction datasets, we are equipped to refine LLaMA (Touvron et al. 2023), enabling it to generate QA pairs from relative captions. As illustrated in Fig 3 (a), we adopt the handcrafted prompt $\boldsymbol{p}$ from GPT-4 and keep it frozen during finetuning.

Following (Zhang et al. 2023b), we leverage LoRA (Hu et al. 2021) to perform efficient fine-tuning, where the backbone is frozen while only the LoRA module is learnable (Feng et al. 2023b). Formally, we have

$$\mathbf{y}_{\texttt{QA}} = \mathcal{F}_{\texttt{LLaMA}}(\boldsymbol{p}, \boldsymbol{t}), \tag{1}$$

where $\mathbf{y}_{\texttt{QA}}$ denotes the generated QA pairs from relative caption $\boldsymbol{t}$. Denote by $\mathbf{x}_{\texttt{QA}} = \{(\mathbf{Q}_1, \mathbf{A}_1), ..., (\mathbf{Q}_K, \mathbf{A}_K)\}$ the ground-truth of QA pairs. Supervised instruction tuning can then be adopted to finetune LLaMA for generating QA pairs.

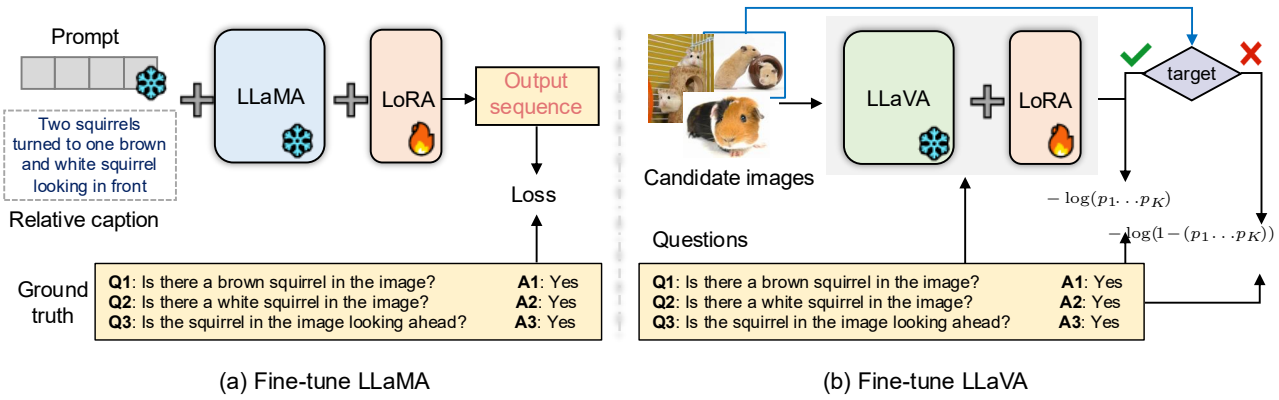


Figure 3: **Overview** of the *training stage* of our **VQA4CIR**. **(a)** Fine-tuning the LLaMA (Touvron et al. 2023) with the instruction data, where the prompt and backbone are frozen while the LoRA is learnable. **(b)** Fine-tuning the LLaVA (Liu et al. 2023a) with the training data, where the backbone is frozen and only the LoRA is learnable.

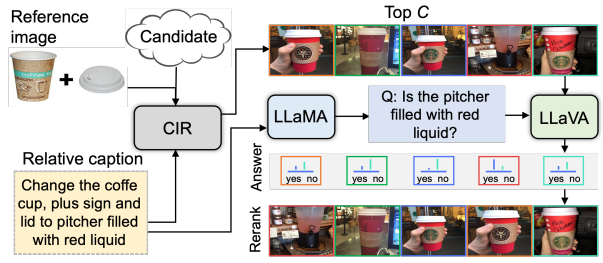


Figure 4: **Overview** of the *inference process*. For any given CIR model, one can select its top C retrieved images, and send the relative caption to finetuned LLaMA (Touvron et al. 2023) for generating QA pairs. By feeding the candidate images and generated questions to finetuned LLaVA (Liu et al. 2023a), we acquire a set of answers, and rerank the candidate images by comparing these answers with the ground truth in QA pairs.

VQA for Boosting CIR

Training Data. To train the VQA model, we construct a training set by using the top-C (*e.g.*$C = 5$) retrieved images $\{\boldsymbol{I}_1, ..., \boldsymbol{I}_c, ..., \boldsymbol{I}_C\}$ for a CIR method (*e.g.*, SPRC), and generated QA pairs $\mathbf{x}_{\texttt{QA}}$. We further introduce an indicator variant y_c, where $y_c = 1$ if $\boldsymbol{I}_c$ is the target image, else $y_c = -1$. Then, the training set for training the VQA method can be represented as $\{(\boldsymbol{I}_c, y_c, \mathbf{x}_{\texttt{QA}})\}$

Finetune LLaVA. As shown in Fig. 3 (b), using LLaVA as an example, we froze the backbone model while leveraging the LoRA trainable (Lai et al. 2023). LLaVA takes a candidate image $\boldsymbol{I}_c$ and a question $\mathbf{Q}_k$ as the input, and outputs the prediction of the answer $\mathbf{A}'_k$. For finetuning LLaVA, we also predict the probability p_k of $\mathbf{A}'_k = \mathbf{A}_k$, *i.e.*,

$$p_k = \mathcal{F}_{\texttt{LLaVA}}\left(\boldsymbol{I}_c, \mathbf{Q}_k; \mathbf{A}_k\right). \tag{2}$$

To finetune LLaMA, the training loss is defined by considering all the QA pairs. When $\boldsymbol{I}_c$ is the target image (*i.e.*, $y_c = 1$), we require LLaVA to correctly answer all the questions, *i.e.*, $p_k \simeq 1$. In contrast, when $\boldsymbol{I}_c$ is not the target image (*i.e.*, $y_c = -1$), LLaVA is expected to incorrectly answer at least one question. Thus, the training loss can be written as

$$-\log\left\{\frac{1-y_c}{2} + y_c\left(p_1 \cdot p_2 \cdot \ldots p_K\right)\right\}. \tag{3}$$

In this manner, we can obtain a VQA model that can also be used to distinguish target images from failure-retrieved images, thereby benefiting the CIR performance.

Inference Process

With the fine-tuned LLaMA and LLaVA as the VQA models, we can re-rank the output candidates of any CIR model. As shown in Fig. 4, given the relative caption, finetuned LLaMA is used to generate K QA pairs $\{(\mathbf{Q}_1, \mathbf{A}_1), ..., (\mathbf{Q}_K, \mathbf{A}_K)\}$. Then, the top-$C$ candidate images $\mathcal{I} = \{\boldsymbol{I}_1, \boldsymbol{I}_2, ..., \boldsymbol{I}_C\}$ are first obtained using a CIR model,

$$\mathcal{I} = \mathcal{F}_{\texttt{CIR}}\left(\boldsymbol{I}_r, \boldsymbol{t}\right). \tag{4}$$

For each question $\mathbf{Q}_k$, we use the finetuned LLaVA to predict the probability p_k of $\mathbf{A}'_k = \mathbf{A}_k$. Then, the product of all p_ks are used to indicate the consistency between the retrieved image and relative caption,

$$p^A = p_1 \cdot p_2 \cdot \ldots p_K. \tag{5}$$

For an ideal VQA model, one can safely reject the candidate images with $p^A = 0$, and simply only keep those with $p^A = 1$ during reranking. However, as shown in Fig. 5, on the CIRR validation set, the p^A distributions of target images and failure retrieved images are overlapped. Thus, we present a soft reranking scheme. For a candidate image $\boldsymbol{I}_c$, where its original rank is c, we modify its rank to $c + R(p^A)$. Obviously, $R(p^A)$ should be larger when p^A is small and should be near zero when $p^A \simeq 1$. To this end, we define $R(p^A)$ as follows,

$$R(p^A) = \alpha \mathbf{e}^{-\beta p^A}, \tag{6}$$

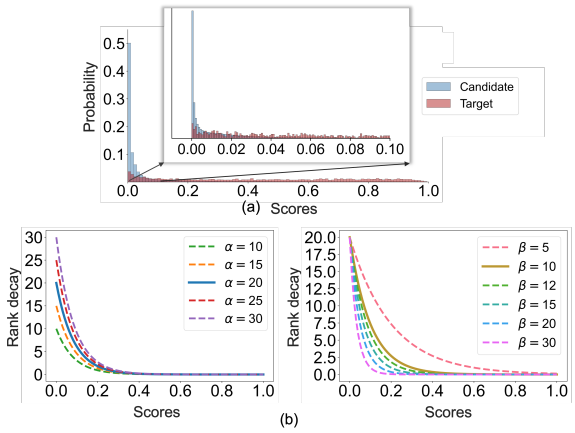


Figure 5: **(a) Distribution** visualization of the target predictions and non-target predictions on the CIRR validation sets. **(b)** Visualization of the **curves** under different values of α and β.

where α is the step size of ranking descent, and β is the rate of ranking decline. A larger α value indicates a greater step size in the descent, and vice versa. A larger β indicates a faster rate of decline, refer to Fig. 5 (b). Detailed discussion regarding these two hyperparameters can be seen in Fig. 6. Finally, we sort all $c + R(p^A)$ values to give the ranks after reranking.

Experiments

Experimental Setup

Our VQA4CIR is implemented with Pytorch on NVIDIA RTX A100 GPUs with 40GB of memory per card. To preserve the generalization ability of the pre-trained models, *i.e.*, LLaMA (Touvron et al. 2023) and LLaVA (Liu et al. 2023a), we leverage LoRA (Hu et al. 2021) to fine-tune them while keeping the backbones frozen, *i.e.*, LLaVA-v1.5-13B and Vicuna-13B-v1.5. We note the word embeddings of the LLaMA (Touvron et al. 2023) and LLaVA (Liu et al. 2023a) are also frozen. During training, we randomly adopt $5,000$ and $3,000$ samples from the CIRR dataset and Fashion-IQ training data, respectively, to fine-tune LLaMA (Touvron et al. 2023) and LLaVA (Liu et al. 2023a). The AdamW (Loshchilov and Hutter 2017) is adopted as the optimizer with a weight decay of 0.05 across all the experiments. We adopt WarmupDecayLR as the learning rate scheduler with warmup iterations of $1,000$. For LLaVA (Liu et al. 2023a), the learning rate is initialized at 2e-5, while for LLaMA (Touvron et al. 2023), it is initialized at 3e-4. The hyperparameter of α is respectively set to 20 and 30 on the CIRR and Fashion-IQ datasets, while β is empirically set to 10 and 12. We evaluate our method on two CIR benchmarks, *i.e.*, CIRR (Suhr et al. 2018) and Fashion-IQ (Wu et al. 2021). The detailed setups follow previous works (Suhr et al. 2018; Wu et al. 2021).

Comparison with State-of-the-arts

CIRR. In Table 1 lists the results on the CIRR dataset. To evaluate whether VQA can assist CIR, we selected the most classical method, CLIP4CIR* (Baldrati et al. 2023), and the currently best-performing method, SPRC (Bai et al. 2024), as our CIR base models. As can be seen, our method achieves noticeable improvements across all the metrics. Although the baseline SPRC has the highest performance among the competing methods, our VQA4CIR can further improve its recall values. Benefited from VQA4CIR, our method, by adopting CLIP4CIR as the base model, achieves the second-best performance under Recall@K=10 and RecallSubset@K=2. Compared to the re-ranking method (Liu et al. 2023c), our method attains a notable improvement on all metrics.

Fashion-IQ. Table 2 lists the results of competing methods on the Fashion-IQ dataset. As can be seen, our VQA4CIR achieves consistent performance gains across eight evaluation metrics in three different categories, *i.e.*, it has the highest recall values, as indicated by the bold results in the table. Although SPRC (Bai et al. 2024) has the highest performance among all the existing baseline methods, our method can further improve its recalls, and there is also a significant improvement in the average recall, *i.e.*, 64.76 ***vs.*** **65.41**. For CLIP4CIR (Baldrati et al. 2022a), we used its enhanced version, CLIP4CIR* (Baldrati et al. 2023). By incorporating VQA4CIR with CLIP4CIR, our method has a significant improvement, *e.g.*, in class `Dress`, R@10: 39.46 to **40.91**, and R@50: 64.55 to **65.13**, and there is also a significant improvement in the average recalls. Moreover, the re-ranking method (Liu et al. 2023c) still falls below ours, *e.g.*, in average recalls. Although VQA4CIR also involves a stage of re-ranking, it is essentially different from (Liu et al. 2023c).

Ablation Studies

Top C for Re-ranking. Our method involves re-ranking the output of existing CIR models, and we thus discuss the effect of the number of the top-C retrieved images by the base CIR method, *i.e.*, the effect of C on CIR performance. Using the CIRR dataset, we provide the results of different C values on the validation sets of CIRR in Table 3. One can see that as C increases, there is an improvement in recall, possibly ascribing to that the larger C is, the greater the coverage of ground truth, thereby yielding higher performance. It is noteworthy that our method achieves the highest improvement when $C = 15$. The smaller the value of C, the more pronounced the improvement, *e.g.*, under the Recall@K metric with $C = 1$, recalls improve from top 0: 53.94 to top 15: **56.15**, and under the $\text{Recall}_{\text{Subset}}$ metric with $C = 1$, recalls improve from top 0: 79.78 to top 15: **82.56**. Our method achieves a noticeable improvement on this metric, indicating that the VQA mechanism can effectively aid the model in recognizing the candidate image being inconsistent with relative caption. In contrast, Liu *et al.* shows almost no change on this metric (Liu et al. 2023c). We also note that the inference time cost is positively correlated with the value of top C. In this paper, considering the results in Table 3, as well as the inference costs, we choose top 70 across all the experiments.

Method	Recall@K				Recall$_{Subset}$			Average
	K=1	K=5	K=10	K=50	K=1	K=2	K=3	
CLIP4CIR (Baldrati et al. 2022a)	38.53	69.98	81.86	95.93	68.19	85.64	94.17	69.09
BLIP4CIR+Bi (Liu et al. 2023b)	40.15	73.08	83.88	96.27	72.10	88.27	95.93	72.59
DRA (Jiang et al. 2023)	39.93	72.07	83.83	96.43	71.04	87.74	94.72	71.55
CoVR-BLIP (Ventura et al. 2023)	49.69	78.60	86.77	94.31	75.01	88.12	93.16	80.81
Re-ranking (Liu et al. 2023c)	50.55	81.75	89.78	97.18	80.04	91.90	96.58	80.90
CLIP4CIR* (Baldrati et al. 2023)	44.82	77.04	86.65	97.90	73.16	88.84	95.59	75.10
+ **VQA4CIR**	51.40	81.71	<u>89.83</u>	<u>98.10</u>	80.15	<u>92.63</u>	<u>96.75</u>	80.93
+ **improvements**	(6.6)↑	(4.7)↑	(3.2)↑	(0.2)↑	(7.0)↑	(3.8)↑	(1.1)↑	(5.8)↑
SPRC (Bai et al. 2024)	<u>51.96</u>	<u>82.12</u>	89.74	97.69	<u>80.65</u>	92.31	96.60	<u>81.39</u>
+ **VQA4CIR**	**54.00**	**84.23**	**91.85**	**98.10**	**82.07**	**93.45**	**97.08**	**83.15**
+ **improvements**	(2.0)↑	(2.1)↑	(2.1)↑	(0.4)↑	(1.4)↑	(1.1)↑	(0.5)↑	(1.8)↑

Table 1: **Quantitative comparison** in terms of recalls across competing methods on the `test` set of CIRR, where best and second-best results are highlighted in **bold** and <u>underlined</u>, respectively.

Method	Dress		Shirt		Toptee		Average		Avg.
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	
CoVR-BLIP (Ventura et al. 2023)	44.55	69.03	48.43	67.42	52.60	74.31	48.53	70.25	59.39
CLIP4CIR (Baldrati et al. 2022a)	33.81	59.40	39.99	60.45	41.41	65.37	38.32	61.74	50.03
BLIP4CIR+Bi (Liu et al. 2023b)	42.09	67.33	41.76	64.28	46.61	70.32	43.49	67.31	55.04
Re-ranking (Liu et al. 2023c)	<u>48.14</u>	71.43	50.15	71.25	55.23	76.80	51.17	73.13	62.15
CompoDiff (Gu et al. 2023a)	40.65	57.14	36.87	57.39	43.93	61.17	40.48	58.57	49.53
CLIP4CIR* (Baldrati et al. 2023)	39.46	64.55	44.41	65.26	47.48	70.98	43.78	66.93	55.35
+ **VQA4CIR**	40.91	65.13	45.62	65.68	49.21	71.22	45.24	67.34	56.29
+ **improvements**	(1.5)↑	(0.6)↑	(1.2)↑	(0.4)↑	(1.7)↑	(0.2)↑	(1.5)↑	(0.4)↑	(0.9)↑
SPRC (Bai et al. 2024)	47.80	<u>72.70</u>	<u>55.84</u>	<u>74.37</u>	<u>58.89</u>	<u>78.99</u>	<u>54.17</u>	<u>75.35</u>	<u>64.76</u>
+ **VQA4CIR**	**49.18**	**73.06**	**56.79**	**74.52**	**59.67**	**79.30**	**55.21**	**75.62**	**65.41**
+ **improvements**	(1.4)↑	(0.5)↑	(1.0)↑	(0.2)↑	(0.8)↑	(0.3)↑	(1.0)↑	(0.3)↑	(0.7)↑

Table 2: **Quantitative comparison** in terms of recalls of various methods on the validation set of the Fashion-IQ dataset.

Method	Recall@K				Recall$_{Subset}$			Average
	K=1	K=5	K=10	K=50	K=1	K=2	K=3	
Top 0 (Baseline)	53.94	84.33	90.91	98.13	79.78	92.46	96.89	82.06
Top 15	56.15	85.24	92.01	98.13	82.56	93.27	97.27	83.90
Top 50	56.15	85.31	92.23	98.13	83.06	93.42	97.32	84.19
Top 70	56.11	85.27	92.23	98.18	83.40	93.59	97.34	84.33
Top 150	55.94	85.29	92.20	98.28	82.99	93.49	97.37	84.14

Table 3: **Ablation studies** in terms of recalls with regard to *different Top C values* on the validation set of CIRR dataset.

Analysis of α and β. As mentioned in Sec. , a larger α value indicates a greater step size in the descent, and a larger β indicates a faster rate of decline. Fig. 6 shows the changes in the average recalls for CIRR under different α values and β values. As can be seen from Fig. 6 (a), the recalls improve with an increase in α, reaching a maximum when α is 25 upon CLIP4CIR (Baldrati et al. 2023) and 20 upon SPRC (Bai et al. 2024). When α is greater than 25, a decrease in recalls occurs. Analogously, as the β value increases, refer to Fig. 6 (b), the recalls on the two methods improve, reaching a maximum when β equals 5 and 10, respectively. Nonetheless, when β becomes larger, the recalls decrease.

Question-wise *vs.* Caption-wise Re-ranking. We further discuss the effect of our method when employing a question-wise approach versus a caption-wise approach for re-ranking during the inference stage. The caption-wise approach multiplies the prediction results for all questions to recognize whether the candidate image is consistent with the relative caption, whereas the question-wise approach re-ranks the predictions for each question directly. We have recorded the differences in recall values between the question-wise and caption-wise approaches for the validation set of CIRR datasets of both CLIP4CIR+VQA4CIR

Method	Recall@K				Recall$_{Subset}$			Average
	K=1	K=5	K=10	K=50	K=1	K=2	K=3	
LLaVA	53.53	84.05	91.33	99.26	79.92	93.58	96.76	81.98
InstructBLIP	52.19	82.99	91.06	97.53	78.19	93.00	95.99	80.59
Fine-Tune LLaVA	56.11	85.27	92.23	98.18	83.40	93.59	97.34	84.33

Table 4: **Ablation studies** in terms of recalls with regard to *different VQA models* on the validation set of the CIRR dataset.

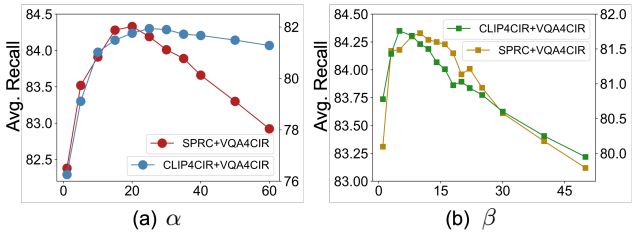


Figure 6: **Ablation** studies of different values of **(a)** α and **(b)** β on CIRR dataset.

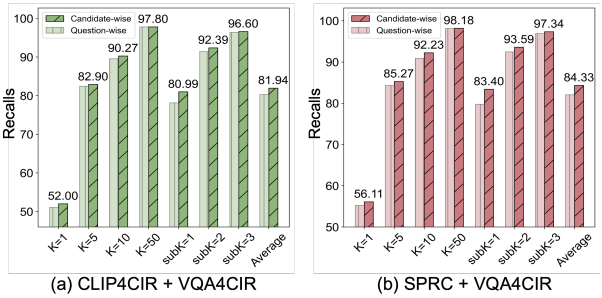


Figure 7: **Ablation** studies of *different reranking mechanisms*, *i.e.*, question-wise versus caption-wise on the validation set of CIRR dataset.

and SPRC+VQA4CIR in Fig. 7. It can be seen that the caption-wise approach performs better than the question-wise approach because it can accurately reflect whether a candidate image is consistent with relative caption. In light of this, we adopted the caption-wise approach for re-ranking during the inference stage in our experiments.

Fine-tune LLaVA in Question-wise *vs.* Caption-wise. To train LLaVA to adapt to CIR, we designed a loss to iteratively check whether the answers to all questions are correct. Here, we discuss the differences between training LLaVA in a question-wise manner and a caption-wise manner. We summarize the performance of these two methods on the validation sets of CIRR in Fig. 8. It can be seen that both methods show differences in CIRR, the average recall values of CLIP4CIR+VQA4CIR and SPRC+VQA4CIR differ by 2.62% and 3.27% respectively. The caption-wise method provides a higher recall because it can accurately reflect whether a candidate image is consistent with relative caption. Thus, we adopt the caption-wise method to train LLaVA in our experiments.

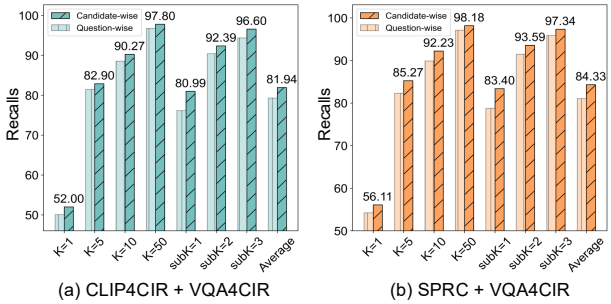


Figure 8: **Ablation** studies of *different LLaVA fine-tune mechanisms*, *i.e.*, question-wise versus caption-wise on the validation set of CIRR.

Discussion on Different VQA Models. We employ different LVLM models, *i.e.*, the pre-trained LLaVA (Liu et al. 2023a) and instructBLIP (Dai et al. 2023), as VQA models to assess the effect of LVLM. We provide the performance of the two variants on the validation sets of two datasets in Table 4. One can see that both approaches significantly underperform our fine-tuned LLaVA because directly using pre-trained LVLMs does not generalize well to our task. Moreover, using the pre-trained LLaVA and instructBLIP directly do not show a substantial difference in recall values, *i.e.*, average recalls: LLaVA: 81.98 ***vs.*** InstructBLIP: 80.59. In contrast, our method, which only fine-tunes a small number of parameters with LoRA can achieve noticeably improved performance, with average recalls: 81.98 & 80.59 *vs.* **84.33**.

Conclusion

This paper provides a VQA perspective for boosting CIR performance and suggest the VQA4CIR method. Our VQA4CIR is built upon the fact that a certain percentage of failure retrieval results in existing CIR methods are not consistent with their relative captions. To find the inconsistent images, we adopted the "QA generation → VQA" self-verification pipeline. First, we leverage fine-tuned LLaMA to generate QA pairs from relative caption. Then, fine-tuned LLaVA was adopted as the VQA model for finding the images inconsistent with relative caption. The inconsistent images are then reranked to enhance CIR performance. Our VQA4CIR can be incorporated with most existing CIR methods. Experiments show that our VQA4CIR outperforms the state-of-the-art methods on the CIRR and Fashion-IQ datasets.

Acknowledgements

This work was supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-TC-2021-003), Agency for Science, Technology and Research (A*STAR) through its AME Programmatic Funding Scheme Under Project A20H4b0141, A*STAR Central Research Fund "A Secure and Privacy Preserving AI Platform for Digital Health", and Agency for Science, Technology and Research (A*STAR) through its RIE2020 Health and Biomedical Sciences (HBMS) Industry Alignment Fund Pre-Positioning (IAF-PP) (grant no. H20C6a0032) and Shenzhen-Hong Kong Joint Funding No. SGDX20211123112401002, by Shenzhen General Program No. JCYJ20220530143600001, and NSFC No. 62302399, China Association for Science and Technology Youth Care Program.

References

Anwaar, M. U.; Labintcev, E.; and Kleinsteuber, M. 2021. Compositional learning of image-text query for image retrieval. In *Proceedings of the IEEE/CVF Winter conference on Applications of Computer Vision*, 1140–1149.

Bai, Y.; Xu, X.; Liu, Y.; Khan, S.; Khan, F.; Zuo, W.; Goh, R. S. M.; and Feng, C.-M. 2024. Sentence-level Prompts Benefit Composed Image Retrieval. *International Conference on Learning Representations*.

Baldrati, A.; Bertini, M.; Uricchio, T.; and Del Bimbo, A. 2022a. 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*, 4959–4968.

Baldrati, A.; Bertini, M.; Uricchio, T.; and Del Bimbo, A. 2022b. Effective conditioned and composed image retrieval combining clip-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21466–21474.

Baldrati, A.; Bertini, M.; Uricchio, T.; and Del Bimbo, A. 2023. Composed image retrieval using contrastive learning and task-oriented clip-based features. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(3): 1–24.

Bangalath, H.; Maaz, M.; Khattak, M. U.; Khan, S. H.; and Shahbaz Khan, F. 2022. Bridging the gap between object and image-level representations for open-vocabulary detection. *Advances in Neural Information Processing Systems*, 35: 33781–33794.

Chen, J.; Guo, H.; Yi, K.; Li, B.; and Elhoseiny, M. 2022. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18030–18040.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. arXiv:2305.06500.

Dodds, E.; Culpepper, J.; Herdade, S.; Zhang, Y.; and Boakye, K. 2020. Modality-agnostic attention fusion for visual search with text feedback. *arXiv preprint arXiv:2007.00145*.

Feng, C.-M.; He, Y.; Zou, J.; Khan, S.; Xiong, H.; Li, Z.; Zuo, W.; Goh, R. S. M.; and Liu, Y. 2024. Diffusion-Enhanced Test-time Adaptation with Text and Image Augmentation. arXiv:2412.09706.

Feng, C.-M.; Li, B.; Xu, X.; Liu, Y.; Fu, H.; and Zuo, W. 2023a. Learning federated visual prompt in null space for mri reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8064–8073.

Feng, C.-M.; Yu, K.; Liu, N.; Xu, X.; Khan, S.; and Zuo, W. 2023b. Towards instance-adaptive inference for federated learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23287–23296.

Feng, C.-M.; Yu, K.; Liu, Y.; Khan, S.; and Zuo, W. 2023c. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2704–2714.

Feng, F.; Niu, T.; Li, R.; Wang, X.; and Jiang, H. 2020. Learning Visual Features from Product Title for Image Retrieval. In *Proceedings of the 28th ACM International Conference on Multimedia*, 4723–4727.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Gao, P.; Han, J.; Zhang, R.; Lin, Z.; Geng, S.; Zhou, A.; Zhang, W.; Lu, P.; He, C.; Yue, X.; et al. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.

Gilardi, F.; Alizadeh, M.; and Kubli, M. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.

Goenka, S.; Zheng, Z.; Jaiswal, A.; Chada, R.; Wu, Y.; Hedau, V.; and Natarajan, P. 2022. Fashionvlp: Vision language transformer for fashion retrieval with feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14105–14115.

Gu, G.; Chun, S.; Kim, W.; Jun, H.; Kang, Y.; and Yun, S. 2023a. CompoDiff: Versatile Composed Image Retrieval With Latent Diffusion. *arXiv preprint arXiv:2303.11916*.

Gu, Q.; Kuwajerwala, A.; Morin, S.; Jatavallabhula, K. M.; Sen, B.; Agarwal, A.; Rivera, C.; Paul, W.; Ellis, K.; Chellappa, R.; et al. 2023b. ConceptGraphs: Open-Vocabulary 3D Scene Graphs for Perception and Planning. *arXiv preprint arXiv:2309.16650*.

Gu, Z.; Zhu, B.; Zhu, G.; Chen, Y.; Tang, M.; and Wang, J. 2023c. AnomalyGPT: Detecting Industrial Anomalies using Large Vision-Language Models. *arXiv preprint arXiv:2308.15366*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jiang, X.; Wang, Y.; Wu, Y.; Wang, M.; and Qian, X. 2023. Dual Relation Alignment for Composed Image Retrieval. *arXiv preprint arXiv:2309.02169*.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2023. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Li, K.; He, Y.; Wang, Y.; Li, Y.; Wang, W.; Luo, P.; Wang, Y.; Wang, L.; and Qiao, Y. 2023b. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Liang, F.; Wu, B.; Dai, X.; Li, K.; Zhao, Y.; Zhang, H.; Zhang, P.; Vajda, P.; and Marculescu, D. 2023. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7061–7070.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023a. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Liu, Z.; Rodriguez-Opazo, C.; Teney, D.; and Gould, S. 2021. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2125–2134.

Liu, Z.; Sun, W.; Hong, Y.; Teney, D.; and Gould, S. 2023b. Bi-directional Training for Composed Image Retrieval via Text Prompt Learning. *arXiv preprint arXiv:2303.16604*.

Liu, Z.; Sun, W.; Teney, D.; and Gould, S. 2023c. Candidate Set Re-ranking for Composed Image Retrieval with Dual Multi-modal Encoder. *arXiv preprint arXiv:2305.16304*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Maaz, M.; Rasheed, H.; Khan, S.; and Khan, F. S. 2023. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. *arXiv preprint arXiv:2306.05424*.

OpenAI. 2022. Introducing chatgpt. *https://openai.com/blog/chatgpt*.

OpenAI. 2023. Gpt-4 technical report.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.

Ray, A.; Radenovic, F.; Dubey, A.; Plummer, B. A.; Krishna, R.; and Saenko, K. 2023. COLA: How to adapt vision-language models to Compose Objects Localized with Attributes? *arXiv preprint arXiv:2305.03689*.

Suhr, A.; Zhou, S.; Zhang, A.; Zhang, I.; Bai, H.; and Artzi, Y. 2018. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*.

Sun, T.; Zhang, X.; He, Z.; Li, P.; Cheng, Q.; Yan, H.; Liu, X.; Shao, Y.; Tang, Q.; Zhao, X.; Chen, K.; Zheng, Y.; Zhou, Z.; Li, R.; Zhan, J.; Zhou, Y.; Li, L.; Yang, X.; Wu, L.; Yin, Z.; Huang, X.; and Qiu, X. 2023. MOSS: Training Conversational Language Models from Synthetic Data.

Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Stanford alpaca: An instruction-following llama model.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ventura, L.; Yang, A.; Schmid, C.; and Varol, G. 2023. CoVR: Learning Composed Video Retrieval from Web Video Captions. *arXiv preprint arXiv:2308.14746*.

Vo, N.; Jiang, L.; Sun, C.; Murphy, K.; Li, L.-J.; Fei-Fei, L.; and Hays, J. 2019. Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6439–6448.

Wu, H.; Gao, Y.; Guo, X.; Al-Halah, Z.; Rennie, S.; Grauman, K.; and Feris, R. 2021. 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*, 11307–11317.

Zhang, A.; Fei, H.; Yao, Y.; Ji, W.; Li, L.; Liu, Z.; and Chua, T.-S. 2023a. Transfer visual prompt generator across llms. *arXiv preprint arXiv:2305.01278*.

Zhang, L.; Zhang, L.; Shi, S.; Chu, X.; and Li, B. 2023b. LoRA-FA: Memory-efficient low-rank adaptation for large language models fine-tuning. *arXiv preprint arXiv:2308.03303*.

Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X. V.; et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.