Reason-before-Retrieve: One-Stage Reflective Chain-of-Thoughts for Training-Free Zero-Shot Composed Image Retrieval

Yuanmin Tang[1,2*] Jue Zhang[3] Xiaoting Qin[3] Jing Yu[4] Gaopeng Gou[1] Gang Xiong[1†]
Qingwei Lin[3] Saravan Rajmohan[3] Dongmei Zhang[3] Qi Wu[5]
[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]Microsoft, [4]Minzu University of China, [5]University of Adelaide
{tangyuanmin,gougaopeng,xionggang}@iie.ac.cn, {juezhang,xiaotingqin}@microsoft.com,
jing.yu@muc.edu.cn, qi.wu01@adelaide.edu.au

Abstract

Composed Image Retrieval (CIR) aims to retrieve target images that closely resemble a reference image while integrating user-specified textual modifications, thereby capturing user intent more accurately. Existing training-free zero-shot CIR (ZS-CIR) methods often employ a two-stage process: they first generate a caption for the reference image and then use Large Language Models for reasoning a target description. However, these methods suffer from missing critical visual details and limited reasoning capabilities, leading to suboptimal retrieval performance. To address these challenges, we propose a novel, training-free one-stage method, One-Stage Reflective Chain-of-Thought Reasoning (OSrCIR) for ZS-CIR, which employs Multimodal Large Language Models to retain essential visual information in a single-stage reasoning process, eliminating the information loss in two-stage methods. Our Reflective Chain-of-Thought framework further improves interpretative accuracy by aligning manipulation intent with contextual cues from reference images. OSrCIR achieves performance gains of 1.80% to 6.44% over existing training-free methods across multiple tasks, setting new state-of-the-art results in ZS-CIR and enhancing its utility in vision-language applications. Our code is available at https://github.com/microsoft/ACV/tree/main/OSrCIR.

1. Introduction

Composed Image Retrieval (CIR) [53] aims to retrieve a target image that is visually similar to a reference image while incorporating modifications specified by user-provided manipulation text. Unlike traditional content-based image retrieval [8], which relies solely on single-modality features, CIR leverages both visual and textual data to capture user intent more accurately, as shown in Figure 1. This dual-modality approach allows users to specify desired changes to reference images, improving search precision and enabling a clearer articulation of user intent. Consequently, CIR has garnered increasing interest in internet search and e-commerce [7, 40], where it facilitates tasks such as scene image search with object manipulation or product recommendations with attribute modification.

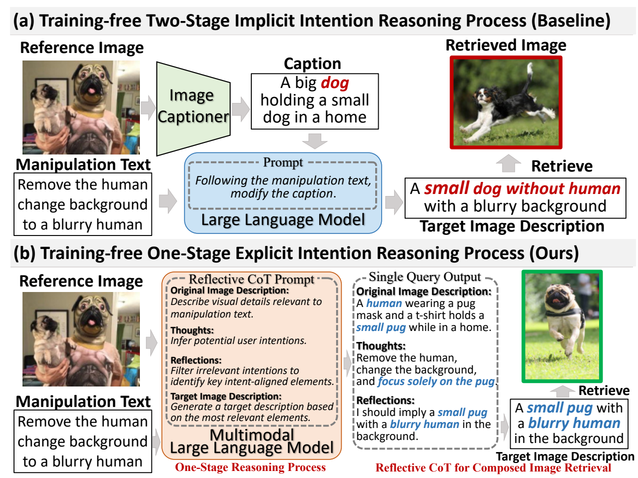

Figure 1. Illustration of our motivation. (a) Two-stage implicit intention reasoning of the baseline CIReVL method. (b) Our one-stage approach OSrCIR with explicit intention reasoning.

CIR faces two fundamental challenges: (1) user intent spans both visual and textual modalities, necessitating a common semantic space for effective cross-modal reasoning, and (2) understanding user intent demands deep reasoning, as it is often implicitly conveyed, particularly through

*Work is done during an internship at Microsoft
†Corresponding author

reference images. While supervised methods have been proposed to tackle these issues [3, 30], they rely on extensive annotated triplets (*i.e.,* reference image, manipulation text, target image) CIR datasets to train task-specific models, which is labor-intensive and limits generalizability.

Zero-Shot Composed Image Retrieval (ZS-CIR) has emerged as a solution to these limitations [4, 40, 50], utilizing the pre-trained large-scale Vision-Language Models (VLMs), *i.e.,* CLIP [39], to reframe ZS-CIR as a text-based image retrieval task. It encodes reference image content into language and combines them with manipulation text to obtain query captions for target retrieval within CLIP's shared semantic space. Query generation methods in ZS-CIR can be implicit or explicit. Implicit methods, like textual inversion [4, 40, 50], are often training-dependent, using large image-caption datasets to train a mapping network that converts images into text tokens. A static template then combines these tokens with textual modifications to create query captions. However, even with large-scale VLMs, these implicit ZS-CIR methods are limited by CLIP capacity for human intention reasoning, which restricts the accurate interpretation of manipulation intent.

Alternatively, recent research [21, 46] explores training-free ZS-CIR methods that utilize Large Language Models (LLMs) for explicit query inference. As illustrated in Figure 1(a), current explicit training-free methods follow a two-stage process: an image captioner (*e.g.,* BLIP-2 [24]) first encodes the reference image into text, followed by LLM-based reasoning to derive a target image description for retrieval. Despite this progress, current two-stage LLM-based methods for ZS-CIR still face two limitations:

(1) **Missing Visual Information.** The initial captioning process is not informed by manipulation text, so critical visual details needed for query composition are often missing. For instance, in Figure 1, without explicit emphasis on the term "human" in manipulation text, the caption fails to include the term "human holds pug". Thus, even with a large-scale retrieval model, this problem remains unresolved.

(2) **Limited Exploitation of LLM Reasoning Capabilities.** Although LLMs offer strong reasoning capabilities, current methods often rely on simple reasoning prompts like *following <Manipulation Text>, modify <Caption>* [21], which restricts LLMs' full reasoning potential and may lead to suboptimal inferences. As seen in Figure 1, the true user intent of "a blurry human in the background" is misinterpreted as "without human with a blurry background".

To address these limitations, we propose a novel training-free ***O**ne-**S**tage **r**eflective chain-of-thought reasoning for zero-shot **C**omposed **I**mage **R**etrieval* (**OSrCIR**). As shown in Figure 1(b), in this one-stage reasoning process, we leverage Multimodal Large Language Models (MLLMs) that handle visual and textual inputs simultaneously, thereby avoiding the intrinsic information loss seen in two-stage methods. Our Reflective Chain-of-Thought (CoT) framework further enhances reasoning by interpreting nuanced manipulation intents from both the manipulation text and contextual cues in the reference images, allowing the model to more accurately locate and apply relevant visual details. This approach is inspired by human cognitive processes, particularly iterative refinement and reasoning, enhancing both model performance and interpretability.

The main contributions are summarized as follows: (1) We propose a one-stage reasoning method based on MLLMs, which fully retains the visual information of the reference image. This approach helps unleash the model's reasoning ability in CIR, thereby improving the accuracy and efficiency of training-free ZS-CIR. (2) We designed a Reflective CoT reasoning approach to address the current model's insufficient understanding of manipulation intention. This approach interprets visual intent based on visual information and accurately identifies relevant visual elements during reasoning, significantly enhancing model performance and interpretability. (3) Our model improves from 1.80% to 6.44% across four tasks on ViT-L/14 while maintaining inference efficiency, setting new state-of-the-art results in ZS-CIR, further impacting a broader range of vision and language applications.

2. Related works

Composed Image Retrieval. Composed Image Retrieval (CIR) involves combining image and text features for retrieval [53], using late fusion to integrate visual and textual features while requiring extensively annotated triplets CIR datasets [3, 30, 59]. Zero-shot CIR models [4, 10, 13, 18, 21, 26, 40, 47–51] eliminate the need for large-scale CIR datasets enabling CIR without extensive labeled data. Textual inversion ZS-CIR methods [3, 4, 50] leverages image-text pairs during training, using pre-trained CLIP language encoder for reasoning. However, these methods often struggle to interpret implicit human intent embedded in manipulation text. Training-free ZS-CIR approaches [21, 46, 57], such as CIReVL [21], leverage LLM to infer manipulation intent. However, their two-stage process often results in inaccuracies, as critical visual details and implicit intent are missed. To address these challenges, we propose a one-stage approach that directly reasons about user intent using complete image content. Unlike diffusion-based [12] or ensemble-based methods like LDRE [57], which introduce substantial computational overhead, our model achieves greater efficiency and faster inference times.

Vision and Language Pre-training Models. Vision and Language Pre-training (VLP) models, such as CLIP [39], leverage large-scale image-text pairs to align visual and textual data implicitly. Recent advancements in VLP [44, 64] have employed static models that merge encoded image and text features, enabling a variety of zero-shot tasks

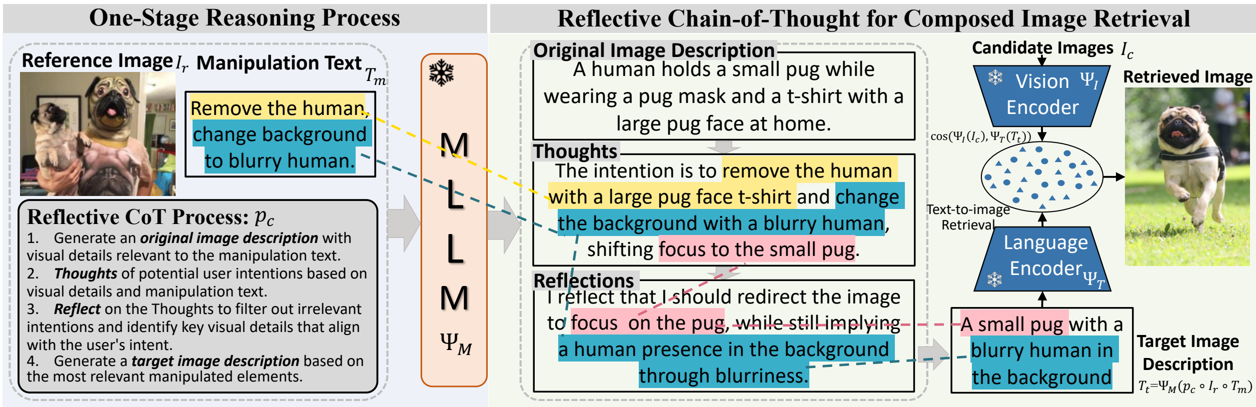

Figure 2. An overview of our model. An MLLM processes the reference image and the manipulation text to generate a description of the desired target image by reflective CoT. To obtain the desired image, we use a vision-language model and perform text-to-image retrieval. Different colors denote the reasoning outcomes of each intention.

[2, 16, 23, 24, 41, 42, 44]. More recent work has focused on integrating vision and language processing within the architecture of large pre-trained language models, leading to the development of state-of-the-art Multimodal Large Language Models (MLLMs) such as LLaVA [28] and GPT-4 [34, 35], which offer enhanced multimodal capabilities. Additionally, methods [6, 19, 25, 45] like ComCLIP [19] have further enhanced the cross-modal retrieval capabilities of multimodal models, pushing the boundaries of image-text retrieval tasks. Our work demonstrates that an MLLM alone, when combined with vision-language retrieval models, can suffice for effective CIR without additional training.

Reasoning Capability of LLMs and MLLMs. LLMs demonstrate strong reasoning abilities, enabled mainly by in-context learning (ICL) [5], where prompted examples and contextual cues improve model performance. Chain-of-Thought prompting [55] further enhances reasoning by guiding LLMs to generate intermediate reasoning steps in complex reasoning tasks. Studies show that LLMs benefit from both crafted demonstrations [55] and zero-shot prompting [22]. Furthermore, self-reflection techniques [43] have proven effective in enhancing reasoning. However, MLLMs face challenges in reasoning due to the gap between visual and textual data. To address this gap, recent research has developed advanced training [2, 31, 37, 65] and prompting methods [15, 32, 60, 61, 63]. Several studies [11, 14, 32, 54, 58, 62, 63] have adapted CoT for multimodal reasoning tasks, showing that CoT can significantly enhance visual reasoning in MLLMs. Building on these advancements, our work is the first to apply CoT to ZS-CIR, extending CoT's impact to a new multimodal domain.

3. Methodology

Given a reference image I_r and a manipulation text T_m describing the user's intention of hypothetical semantic changes on the reference image, ZS-CIR retrieves images from an image database $\mathcal{D}$ that are visually similar to I_r while incorporating the modifications specified in T_m. Figure 2 illustrates our model. We introduce a new approach to explicitly reasoning a target image description T_t as the composed query based on a Multimodal Large Language Model (MLLM) Ψ_M, which contains pre-trained knowledge to understand the user's intention embedded in manipulation text. To ensure that Ψ_M reasons T_t in a human-understandable manner, we introduce a Reflective Chain-of-Thought prompt p_c. The obtained target image description T_m is then used for image retrieval via CLIP, with the associated pre-trained text encoder Ψ_T embedding both the target image description T_t and candidate images I_c into a shared, searchable space. The matching score is computed using cosine similarity $\cos(\Psi_I(I_c), \Psi_T(T_t))$.

3.1. One-Stage Reasoning Process

The conventional two-stage structure of training-free ZS-CIR restricts the ability of image captioners to capture essential visual details, thereby constraining the reasoning capacity of LLMs. To overcome this limitation, we propose a streamlined one-stage approach that eliminates the need for a separate image captioning stage, which does not include user-provided manipulation intent. As shown in Figure 2 (left), we aim to leverage Ψ_M's inherent multimodal understanding to capture the reference image's details directly. This enables reasoning a target image description T_t, modeling the user's intention of hypothetical manipulation of T_m on the reference image I_r as a transformation in the resulting target description T_t without additional training. Formally, given an MLLM Ψ_M, we generate a target image description T_t contains the user's manipulation intent T_m on the reference image I_r as follows:

$$T_t = \Psi_M(p_c \circ I_r \circ T_m), \tag{1}$$

where the LLM is queried with a concatenated prompt composed of the base CoT prompt p_c (see Section 3.2 for details), the reference image I_r (prepended with "`Original Image Context`"), and T_m, the manipulation intent text (prepended with "`Manipulation text`"). This prompt format is largely task-agnostic, enabling its application across various CIR tasks.

3.2. Reflective Chain-of-Thought for ZS-CIR

Each image-intention input pair comprises a reference image and manipulation text that implicitly conveys the user's intention to modify the reference image. To generate the target image description T_t, the adopted MLLM needs to understand this manipulation intention accurately. Existing methods rely on simple prompts (*e.g.,* `Following` T_m`, modify reference image caption`) to extract these intentions, but this approach is insufficient for accurately inferring user's implicit intention embedded in T_m (see Section 4.2). To address this limitation, we introduce a Reflective CoT prompt p_c, which guides the MLLM to progressively reason about user intent across both the reference image and manipulation text, ensuring accurate ZS-CIR.

Specifically, as shown in Figure 2 (right), the Reflective CoT prompt instructs the following progressive reasoning steps: First, the *Original Image Description* step highlights visual details relevant to the user's intention in the reference image. The *Thoughts* step then captures the user's intention and reasoning for potentially manipulated visual elements. In the *Reflections* step, these elements are further evaluated to identify those mostly aligned with the user's intent. Finally, the *Target Image Description* step generates a refined description based on the most intention-relevant visual modifications for target retrieval. Notably, all steps are included in a **single** prompt for MLLM, ensuring both efficiency and interpretability. We illustrate each reasoning step below using the example in Figure 2 while providing the complete prompt template in Supplementary.

Original Image Description. During this step, the MLLM is asked to *capture all visible objects, attributes, and elements relevant to the manipulation text*, and to *reflect on the content and context of the image* to ensure retention of fine-grained details. In Figure 2, the intention-irrelevant visual details (*e.g.,* a table, lights, or photos) are excluded in the caption while relevant elements (*e.g.,* human holding a small pug) are preserved to align with the manipulation text.

Thoughts. Given the intention-relevant visual details and manipulation text, the MLLM then seeks to capture the user's intention (*e.g.,* "Remove the human, change the background"). We first prompt the MLLM to *explain its understanding of the manipulation intent*. Since the user's intentions are often implicit, requiring reference image context for interpretation (*e.g.,* "Removing the human to focus on the pug"), we further ask the MLLM to *discuss how the manipulation intent influences the choice of focused elements in the original image*.

Reflections. Given the manipulation intent and reference image, the MLLM needs to filter out incorrect intentions (*e.g.,* removing the human) and identify the most relevant manipulated elements (*e.g.,* the small pug, a blurry human background). We ask the MLLM to *highlight key decisions made to preserve the coherence and context of the original image while fulfilling the manipulation intent* and to *offer a logical connection between the original content and the final description.* This step also alleviates hallucination issues present in the Thoughts step (See Figure 5).

Target Image Description. Given the filtered manipulated elements, the MLLM finally generates a target description based on the manipulated elements mostly relevant to user intent. We simply ask the MLLM to *generate a target image description that only contains the target content.*

Vision-by-Language In-Context Learning. Simply providing guidelines for the Reflective CoT process is insufficient for MLLMs to understand the CoT process required at each step. To address this, we leverage in-context learning, a technique widely used in LLM and MLLM CoT methods [33, 55, 63]. To ensure a zero-shot setting in ZS-CIR, we propose a vision-by-language in-context learning (ICL) approach. This method provides a few expected MLLM outputs in text form as examples, without requiring a reference image, to guide the MLLM through the reasoning process at each step. Refer to our Supplementary for more details.

Composed Image Retrieval. Given the target image description T_t, our model encodes the image-search database $\mathcal{D}$ alongside T_t using a frozen pre-trained CLIP. The retrieved target image I_t is determined as follows:

$$I_t = \underset{I_r \in \mathcal{D}}{\texttt{argmax}} \frac{\Psi_I(I_r)^\intercal \Psi_T(T_t)}{||\Psi_I(I_r)||\,||\Psi_T(T_t)||}, \tag{2}$$

where the selected target image I_t is the one most similar to the generated target image description. The retrieval process is modular, performed only after combining the reference image and manipulation text, allowing flexibility to substitute different retrieval systems based on practical needs and the desired trade-off between efficiency and effectiveness. Our approach enables a human-understandable ZS-CIR pipeline, where reasoning is fully expressed in the language domain, and the retrieval process is clearly separated, requiring no additional training or mapping modules.

4. Experiments

Datasets and Baselines. We utilize four commonly used datasets in CIR: CIRR [30], CIRCO [4], FashionIQ [56], and GeneCIS [52]. CIRR is the first natural image dataset for CIR, although it can include false negatives [4], where several images could be potential ground truths but are not labeled as such. The CIRCO dataset addresses this by providing multiple annotated ground truths to reduce false negatives. GeneCIS, built from MS-COCO [27] and Visual

CIRCO + CIRR →		CIRCO				CIRR					
Metric		mAP@k				Recall@k			$\text{Recall}_{\text{Subset}}$@k		
Arch	Method	k=5	k=10	k=25	k=50	k=1	k=5	k=10	k=1	k=2	k=3
ViT-B/32	SEARLE	9.35	9.94	11.13	11.84	24.00	53.42	66.82	54.89	76.60	88.19
	CIReVL	14.94	15.42	17.00	17.82	23.94	52.51	66.00	60.17	80.05	90.19
	CIReVL*	16.02	16.69	17.77	18.89	24.25	52.83	66.32	60.43	80.35	90.51
	OSrCIR	**18.04**	**19.17**	**20.94**	**21.85**	**25.42**	**54.54**	**68.19**	**62.31**	**80.86**	**91.13**
ViT-L/14	Pic2Word	8.72	9.51	10.64	11.29	23.90	51.70	65.30	-	-	-
	SEARLE	11.68	12.73	14.33	15.12	24.24	52.48	66.29	53.76	75.01	88.19
	LinCIR	12.59	13.58	15.00	15.85	25.04	53.25	66.68	57.11	77.37	88.89
	Context-I2W	13.04	14.62	16.14	17.16	25.60	55.10	68.50	-	-	-
	CIReVL	18.57	19.01	20.89	21.80	24.55	52.31	64.92	59.54	79.88	89.69
	CIReVL*	18.92	19.32	21.15	22.14	24.83	52.68	65.28	59.82	80.15	89.98
	OSrCIR	**23.87**	**25.33**	**27.84**	**28.97**	**29.45**	**57.68**	**69.86**	**62.12**	**81.92**	**91.10**
ViT-G/14	LinCIR	19.71	21.01	23.13	24.18	35.25	64.72	76.05	63.35	82.22	91.98
	CIReVL	26.77	27.59	29.96	31.03	34.65	64.29	75.06	67.95	84.87	93.21
	CIReVL*	27.12	28.01	30.35	31.39	34.98	64.68	75.41	68.37	85.23	93.24
	OSrCIR	**30.47**	**31.14**	**35.03**	**36.59**	**37.26**	**67.25**	**77.33**	**69.22**	**85.28**	**93.55**

Table 1. **Comparison on CIRCO and CIRR Test Data.** On CIRCO, OSrCIR significantly outperforms even adaptive methods across retrieval models, while it achieves competitive results on CIRR despite the noise in the benchmark. Grey lines represent the training-free ZS-CIR methods. CIReVL* uses the GPT4o [1] in two-stage. **Bold** and '_' denote the best and second-best result, respectively.

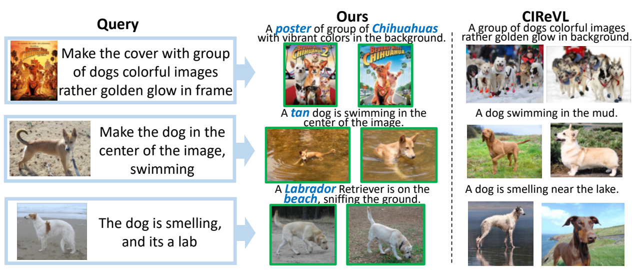

Figure 3. Results on the object manipulation on the CIRR.

Attributes in the Wild [38], offers four task variations, enabling retrieval or modification tasks around specific objects or attributes. FashionIQ focuses specifically on fashion-related retrieval. These datasets cover distinct CIR tasks: CIRCO and CIRR for object manipulation (using reference images to guide object or background manipulation), GeneCIS for object and attribute composition (with various object and attribute labels used to combine with cropped query images for retrieval), and FashionIQ for attribute manipulation (offering descriptive sentences to modify image attributes). Following the original benchmarks, we use Recall@k (R@k) as the evaluation metric for CIRR, GeneCIS, and FashionIQ, and mean average precision (mAP@k) for CIRCO to account for multiple positives. We also evaluate CIRR in a subset setting, where $\text{Recall}_{\text{Subset}}$@k measures retrieval performance within a limited selection of images relevant to the query in the database.

We compare OSrCIR with several commonly benchmarked ZS-CIR methods, categorized as textual inversion or training-free approaches. The textual inversion methods are training-dependent and include: 1) **Pic2Word** [40]: maps the visual features of a reference image into a pseudo-word token. 2) **SEARLE** [4]: combines the pseudo-word token with the GPT-generated caption [5] and applies distillation for efficiency. 3) **Context-I2W** [50]: selectively maps text description-relevant visual information from the reference image. 4) **LinCIR** [13]: masks subjects in captions to enhance training efficiency.

The training-free baseline methods are as follows: 1) **CIReVL** [21], a two-stage approach where a pre-trained image captioner generates a reference image caption, followed by an LLM composing a target image description based on manipulation text; and 2) **CIReVL***, following CIReVL's two-stage process but employing the same MLLM used in OSrCIR for both reference image captioning and target image description generation. To ensure a fair comparison, we present results without using LLM-based ensemble methods like LDRE [57] or diffusion-based models like CompoDiff [12], as these approaches add substantial computational overhead in inference or training. We evaluate our method across three backbones (ViT-B/32, ViT-L/14, and ViT-G/14) but focus primarily on ViT-L/14 for baseline comparisons. This choice is driven by its balance of inference efficiency and retrieval quality, which is widely reported by other baselines and is more practical for real-world applications.

Implementation Details. The default MLLM used in OSrCIR is GPT-4o [1], while we also perform ablations with GPT-4o-mini, GPT-4V, and open-source MLLMs including LLaVA [29] and MiniGPT4 [66]. GPT APIs are used with a temperature setting of 0, while all other parameters remain at their default values. The retrieval module, built in PyTorch [36] based on the codebase from [20], performs all computations on a single NVIDIA A100 GPU. For the CLIP-based ViT variants [9], we adopt weights from

GeneCIS →		Focus Attribute			Change Attribute			Focus Object			Change Object			**Average**
Backbone	Method	R@1	R@2	R@3	R@1	R@2	R@3	R@1	R@2	R@3	R@1	R@2	R@3	R@1
ViT-B/32	SEARLE	18.9	30.6	41.2	13.0	23.8	33.7	12.2	23.0	33.3	13.6	23.8	33.3	14.4
	CIReVL	17.9	29.4	40.4	14.8	25.8	35.8	14.6	24.3	33.3	16.1	27.8	37.6	15.9
	CIReVL*	18.2	29.7	40.7	15.1	26.1	36.1	14.9	24.5	33.6	16.4	28.1	37.9	16.2
	OSrCIR	**19.4**	**32.7**	**42.8**	**16.4**	**27.7**	**38.1**	**15.7**	**25.7**	**35.8**	**18.2**	**30.1**	**39.4**	**17.4**
ViT-L/14	SEARLE	17.1	29.6	40.7	16.3	25.2	34.2	12.0	22.2	30.9	12.0	24.1	33.9	14.4
	LinCIR	16.9	30.0	41.5	16.2	28.0	36.8	8.3	17.4	26.2	7.4	15.7	25.0	12.2
	Context-I2W	17.2	30.5	41.7	16.4	28.3	37.1	8.7	17.9	26.9	7.7	16.0	25.4	12.7
	CIReVL	19.5	31.8	42.0	14.4	26.0	35.2	12.3	21.8	30.5	17.2	28.9	37.6	15.9
	CIReVL*	19.7	32.1	42.3	14.8	26.2	35.4	12.5	22.1	30.7	17.3	29.1	37.9	16.1
	OSrCIR	**20.9**	**33.1**	**44.5**	**17.2**	**28.5**	**37.9**	**15.0**	**23.6**	**34.2**	**18.4**	**30.6**	**38.3**	**17.9**
ViT-G/14	LinCIR	19.1	33.0	42.3	17.6	30.2	38.1	10.1	19.1	28.1	7.9	16.3	25.7	13.7
	CIReVL	20.5	34.0	44.5	16.1	28.6	39.4	14.7	25.2	33.0	18.1	31.2	41.0	17.4
	CIReVL*	20.9	34.4	44.9	16.5	29.0	39.8	15.1	25.6	33.4	18.5	31.6	41.4	17.8
	OSrCIR	**22.7**	**36.4**	**47.0**	**17.9**	**30.8**	**42.0**	**16.9**	**28.4**	**36.7**	**21.0**	**33.4**	**44.2**	**19.6**

Table 2. **Comparison on GeneCIS Test Data.** OSrCIR is able to significantly outperform adaptive methods across all GeneCIS sub-benchmarks, with its inherent modularity allowing for further simple scaling to achieve additional large gains. Grey lines represent the training-free ZS-CIR methods. CIReVL* uses the GPT4o in two-stage. **Bold** and '_' denotes the best and second-best result, respectively.

the official CLIP implementation [39] while using OpenCLIP [17] for ViT-G/14. Performance metrics are averaged across three trials to ensure reliability.

4.1. Quantitative and Qualitative Results

Our main quantitative experimental results are presented in Tables 1, 2 and 3, while Figures 3 and 4 show qualitative comparisons between our model and the baseline CIReVL.

In Table 1, we show the comparison results for the CIRCO and CIRR datasets, which evaluate our model's capability in foreground and background differentiation as well as fine-grained image editing through object and scene manipulation tasks. Performances are evaluated on the hidden test sets of CIRCO and CIRR, accessible via the submission servers [4, 40]. For all different CLIP-based ViT variants for retrieval, our approach significantly outperforms existing methods, including training-free and textual inversion. For instance, on the default ViT-L/14 in CIRCO, which contains clean annotations of manipulation text with multiple target images, our model achieves a mAP@5 of 23.87%, notably surpassing the 18.92% obtained by the best training-free method (CIReVL*) and nearly doubling the 13.04% achieved by the top textual inversion method (Context-I2W). Furthermore, in CIRR, where the manipulation text is less explicit and noisier [4, 21], our model still shows a significant 3.23% average improvement across all evaluation metrics over the best training-free method, CIReVL*. Note that although CIReVL* outperforms CIReVL, the difference is marginal, suggesting that simply adopting a better MLLM does not address the limitations of the two-stage approach.

Qualitatively, as illustrated in Figure 3, our method, OSrCIR, generates target image descriptions that align with user intent and capture intricate visual details. In comparison, CIReVL misses critical elements, such as the image type "poster" and dog breed "Chihuahuas" in Row 1, the dog's "tan" color in Row 2, and the contextual details of the "beach" setting and dog breed "Labrador" in Row 3.

We further evaluate our model's capability on object and attribute composition using the GeneCIS dataset, with the results detailed in Table 2. Unlike CIRCO and CIRR, GeneCIS uses single-word manipulation texts with varied interpretations depending on the task, such as focusing on or changing a specific attribute or object. Consequently, user intent is often abstract and ambiguous, requiring our model to interpret intent precisely based on the reference image. For a fair comparison, we adopt the same output format in our reflective CoT process as the recent work [21]. Specifically, for the "Focus" tasks, we direct the MLLM to *retain the attribute or object specified in the instruction*. For the "Change" tasks, we prompt it to *replace the corresponding object*. For the ViT-L/14 retrieval backbone, our method achieves a 1.8% improvement in Average $R@1$ over the best training-free method (CIReVL*) and outperforms the best textual inversion method (Context-I2W) by 5.2%. Similar improvements are also observed for the other two backbones, underscoring the effectiveness of our reflective CoT process in capturing the user's implicit intent.

Lastly, Table 3 presents our model's performance on attribute manipulation tasks using the FashionIQ validation set, requiring accurate localization of specific fashion attributes (*e.g.,* style, color, pattern). The results show that OSrCIR surpasses existing ZS-CIR models using the ViT-B/14 and ViT-L/14 backbones. For instance, on ViT-L/14, our method outperforms the best training-free model (CIReVL*) and the leading textual inversion model (Context-I2W) by 4.74% and 5.47% on average, respectively. On ViT-G/14, our method achieves a notable 4.6% improvement over the best training-free model, CIReVL*, yet still falls short of the best-performing textual inversion approach, LinCIR. This discrepancy may stem from LinCIR's training process being aligned with the CLIP model

Fashion-IQ →		Shirt		Dress		Toptee		**Average**	
Backbone	Method	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
ViT-B/32	SEARLE	24.44	41.61	18.54	39.51	25.70	46.46	22.89	42.53
	CIReVL	28.36	47.84	25.29	46.36	31.21	53.85	28.29	49.35
	CIReVL*	28.83	48.36	25.82	46.89	31.73	54.34	28.79	49.86
	OSrCIR	**31.16**	**51.13**	**29.35**	**50.37**	**36.51**	**58.71**	**32.34**	**53.40**
ViT-L/14	Pic2Word	26.20	43.60	20.00	40.20	27.90	47.40	24.70	43.70
	SEARLE	26.89	45.58	20.48	43.13	29.32	49.97	25.56	46.23
	LinCIR	29.10	46.81	20.92	42.44	28.81	50.18	26.28	46.49
	Context-I2W	29.70	48.60	23.10	45.30	30.60	52.90	27.80	48.90
	CIReVL	29.49	47.40	24.79	44.76	31.36	53.65	28.55	48.57
	CIReVL*	29.98	47.92	25.29	45.28	31.89	54.13	29.05	49.11
	OSrCIR	**33.17**	**52.03**	**29.70**	**51.81**	**36.92**	**59.27**	**33.26**	**54.37**
ViT-G/14	LinCIR	**46.76**	**65.11**	**38.08**	**60.88**	**50.48**	**71.09**	**45.11**	**65.69**
	CIReVL	33.71	51.42	27.07	49.53	35.80	56.14	32.19	52.36
	CIReVL*	34.01	51.92	27.56	50.04	36.29	56.63	32.62	52.86
	OSrCIR	38.65	54.71	33.02	54.78	41.04	61.83	37.57	57.11

Table 3. **Comparison on FashionIQ Validation Data.** OSrCIR is able to significantly outperform adaptive methods across all sub-benchmarks, with its inherent modularity allowing for further simple scaling to achieve additional large gains. Grey lines represent the training-free ZS-CIR methods. CIReVL* uses the GPT4o in two-stage. **Bold** and '_' denotes the best and second-best result, respectively.

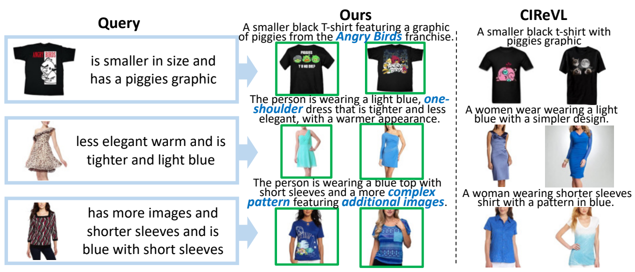

Figure 4. Results of attribute manipulation on the FashionIQ.

used in retrieval, unlike our training-free approach, which lacks this specific alignment. The limitation is particularly evident in the fashion domain, where CLIP may have limited domain-specific knowledge. For instance, terms like "sequined bodice" in the target description are challenging for CLIP to interpret without training-based alignment, leading to reduced performance. Conversely, in the natural image domain, such as CIRCO, where MLLM/LLM outputs are more comprehensible to CLIP, our training-free method substantially outperforms all textual inversion techniques. Future work might explore enhancing the alignment between reasoning and retrieval modules to improve model performance in specialized domains.

Qualitative comparison results of our method and the baseline method CIReVL are presented in Figure 4. OSrCIR accurately identifies and manipulates the attribute-relevant visual elements of "Angry Birds" (Row 1), a "one-shoulder" dress (Row 2), and a tee with a complex pattern featuring more images (Row 3).

4.2. Ablation Study and Performance Analysis

Similar to [13, 21, 57], we examine the contributions of core components in OSrCIR using a ViT-L/14 backbone on CIRCO and Fashion-IQ (Table 4). **(1) Models '2-3' assess the impact of key modules in OSrCIR.** Adapting CIReVL's second stage with our reflective CoT process (model '2') results in a 2.46% average performance drop compared to our method (model '1'), highlighting the necessity of our one-stage reasoning process for capturing complete reference image content. Removing Reflective CoT (model '3') causes a 3.55% performance decline, indicating the importance of our multimodal CoT for effective manipulation intention understanding. We choose not to conduct an ablation integrating manipulation text into caption generation with MLLM in the two-stage approach, as it is methodologically closely aligned with OSrCIR but adds an additional MLLM query, which is unnecessary and reduces efficiency. **(2) Models '4-7' evaluate each Reflective CoT step.** Skipping the generation of the original image description guided by manipulation text (model '4') causes a 1.44% performance decline, emphasizing the need to remove irrelevant visual information. Similarly, without reasoning about user intentions (model '5') or filtering irrelevant ones (model '6'), performance drops by 2.60% and 2.08%, respectively, underscoring the importance of capturing user intentions and identifying relevant visual elements. Removing our vision-by-language in-context samples (model '7') results in a 1.29% decline, showing the benefit of ICL for guiding the reflective CoT. **(3) In models '8-11', we analyze the impact of the choice of MLLM.** Open-source models, such as LLaVA (model '8') and MiniGPT-4 (model '9'), achieve results close to the best training-free ZS-CIR method, CIReVL, but there remains a gap of 2.89% and 3.96% compared to GPT-4o (model '1'). Notably, GPT-4o-mini (model '10') performs comparably well, with only a 0.97% decline, while being more efficient than GPT-4o. Please refer to our Supplementary (Section D) for more ablation studies.

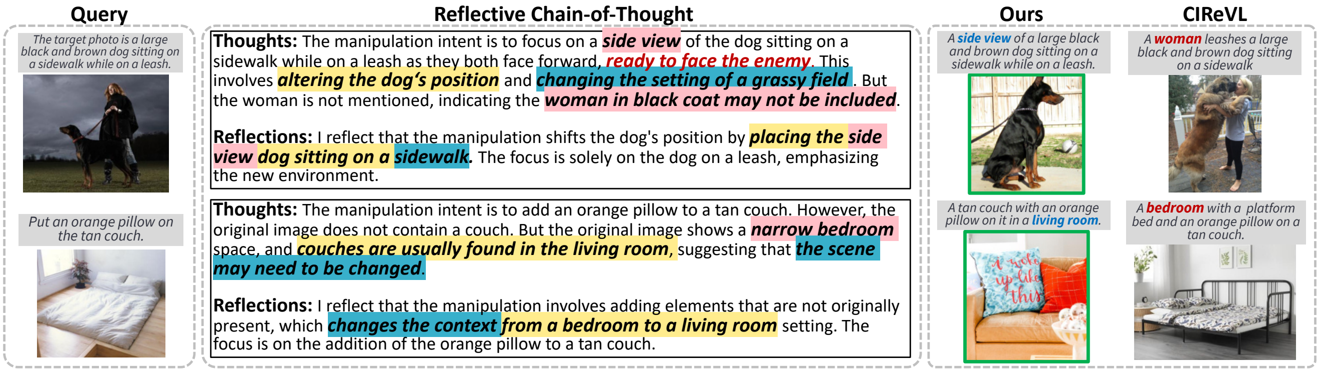

Figure 5. Visualization of Reflective CoT samples. We compare the top 1 retrieval results of ours and CIReVL. Different colors denote the reasoning outcomes of each intention. Our Reflective CoT effectively filters out elements irrelevant to user intention.

	Methods	CIRCO k=5	CIRCO k=10	CIRCO k=25	Fashion-IQ k=10	Fashion-IQ k=50
1.	Full model (GPT-4o)	23.87	25.33	27.84	33.26	54.37
Significance of key modules of `OSrCIR`						
2.	w/o one-stage reasoning	21.73	22.78	24.47	31.16	52.22
3.	w/o Reflective CoT	20.86	21.40	23.34	30.27	51.06
Necessity of each step in our Reflective CoT						
4.	w/o Original Description	22.56	23.57	26.02	32.37	52.97
5.	w/o Thoughts	21.46	22.07	25.06	31.59	51.47
6.	w/o Reflections	22.04	22.74	25.32	32.05	52.11
7.	w/o ICL	22.97	23.50	26.55	32.03	53.17
Impact of different MLLMs						
8.	LLaVA	20.89	22.30	24.88	30.75	51.42
9.	MiniGPT-4	19.85	21.30	23.90	29.36	50.47
10.	GPT-4o-mini	23.10	24.47	26.73	32.19	53.32
11.	GPT-4V	22.15	23.58	25.24	31.55	52.60

Table 4. Ablation study on CIRCO and FashionIQ.

Qualitative Analysis of Reflective CoT. To further examine the benefits of reflective CoT on interpreting user intent, we present additional case studies in Figure 5 alongside the example in Figure 2. For instance, in Row 1, reflective CoT effectively filters out elements irrelevant to user intent, such as "the woman in a black coat" and the hallucinated thought (*i.e.,* "ready to face the enemy"). Notably, reflective CoT also demonstrates accuracy in interpreting intent even when the connection between the reference image and manipulation text is weak, as shown in Row 2. Although this situation technically falls outside CIR, it reflects real user behavior, where users may not closely align manipulation text with the reference image. In Row 2, reflective CoT uses common sense (*e.g.*, recognizing that couches are uncommon in small bedrooms) to infer the user's intention of transitioning from a bedroom to a living room. This filtering of irrelevant details enhances model robustness and likely underlies its strong performance on the CIR task.

Analysis of Common Failure Cases. To gain insights into failure cases of `OSrCIR`, we analyzed 300 failure cases from the FashionIQ validation set using ViT-G/14. As shown in Figure 6, we identify two main issues: (1) *Difficulty with reasoning terms* (49%): The retrieval model (*i.e.,* CLIP) often misreads reasoning terms (*e.g.,* comparisons) like interpreting "less flowy" incorrectly (Row 1) while substituting "stiffer" corrected this. (2) *Misalignment concepts between MLLM and retrieval model* (34%): the retrieval model struggles to interpret fashion-specific terms from MLLM, like "Hawaiian style" and "floral and geometric" (Row 2). Replacing them with simpler terms ("tropical style", "flower shapes") improved retrieval accuracy.

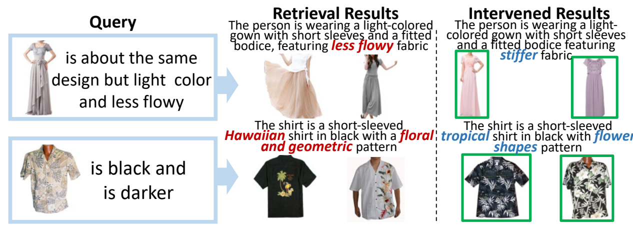

Figure 6. Visualization of common failure cases in the FashionIQ validation set. The top 2 retrieval results are shown.

Effectiveness and Efficiency Analysis. Our approach not only outperforms the best training-free ZS-CIR method (CIReVL) on four CIR tasks but also has a faster inference time, taking about 0.6 second per query, which is 66.67% faster than CIReVL. Compared to textual inversion methods, while our performance surpasses them without training, our inference speed remains 30× slower. As MLLM API calls account for 97% of the total time in `OSrCIR`, we believe that faster APIs may resolve this issue in the future. For details, please refer to our Supplementary (Section E).

5. Conclusion

In this paper, we propose a one-stage reflective chain-of-thought reasoning approach that leverages MLLMs to simultaneously process visual and textual inputs, reducing information loss found in two-stage training-free ZS-CIR methods. By capturing nuanced manipulation intents from text and image cues, `OSrCIR` demonstrates strong generalization and significantly outperforms existing methods on four diverse tasks, achieving comparable inference times. This work advances intention-based image retrieval and has broad implications for vision-language applications.

Acknowledgements

This paper is supported by National Key Laboratory of Science and Technology on Blind Signal Processing (No. 23007522).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 5

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 23716–23736, 2022. 3

[3] 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. 2

[4] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. *arXiv:2303.15247*, 2023. 2, 4, 5, 6

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pages 1877–1901. Curran Associates, Inc., 2020. 3, 5

[6] Chi Chen, Ruoyu Qin, Fuwen Luo, Xiaoyue Mi, Peng Li, Maosong Sun, and Yang Liu. Position-enhanced visual instruction tuning for multimodal large language models, 2023. 3

[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. 1

[8] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2):1–60, 2008. 1

[9] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[10] Yongchao Du, Min Wang, Wengang Zhou, Shuping Hui, and Houqiang Li. Image2sentence based asymmetrical zero-shot composed image retrieval. *arXiv preprint arXiv:2403.01431*, 2024. 2

[11] Timin Gao, Peixian Chen, Mengdan Zhang, Chaoyou Fu, Yunhang Shen, Yan Zhang, Shengchuan Zhang, Xiawu Zheng, Xing Sun, Liujuan Cao, and Rongrong Ji. Cantor: Inspiring multimodal chain-of-thought of mllm. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 9096–9105, New York, NY, USA, 2024. Association for Computing Machinery. 3

[12] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, HeeJae Jun, Yoohoon Kang, and Sangdoo Yun. Compodiff: Versatile composed image retrieval with latent diffusion. *arXiv preprint arXiv:2303.11916*, 2023. 2, 5

[13] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, , Yoohoon Kang, and Sangdoo Yun. Language-only efficient training of zero-shot composed image retrieval. In *Conference on Computer Vision and Pattern Recognition*, 2024. 2, 5, 7

[14] Liqi He, Zuchao Li, Xiantao Cai, and Ping Wang. Multi-modal latent space learning for chain-of-thought reasoning in language models, 2023. 3

[15] Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. Cogagent: A visual language model for gui agents. *arXiv preprint arXiv:2312.08914*, 2023. 3

[16] Thomas Hummel, Shyamgopal Karthik, Mariana-Iuliana Georgescu, and Zeynep Akata. Egocvr: An egocentric benchmark for fine-grained composed video retrieval. *arXiv preprint arXiv:2407.16658*, 2024. 3

[17] 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. 6

[18] Young Kyun Jang, Dat Huynh, Ashish Shah, Wen-Kai Chen, and Ser-Nam Lim. Spherical linear interpolation and text-anchoring for zero-shot composed image retrieval. *arXiv preprint arXiv:2405.00571*, 2024. 2

[19] Kenan Jiang, Xuehai He, Ruize Xu, and Xin Wang. ComCLIP: Training-free compositional image and text matching. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6639–6659, Mexico City, Mexico, 2024. Association for Computational Linguistics. 3

[20] Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. Vision-by-language for training-free compositional image retrieval. *arXiv preprint arXiv:2310.09291*, 2023. 5

[21] Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. Vision-by-language for training-free compositional image retrieval. In *ICLR*, 2024. 2, 5, 6, 7

[22] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are

zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022. 3

[23] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, pages 12888–12900, 2022. 3

[24] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 2, 3

[25] Yongqi Li, Wenjie Wang, Leigang Qu, Liqiang Nie, Wenjie Li, and Tat-Seng Chua. Generative cross-modal retrieval: Memorizing images in multimodal language models for retrieval and beyond. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11851–11861, Bangkok, Thailand, 2024. Association for Computational Linguistics. 3

[26] 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 International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 240–250. ACM, 2024. 2

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 4

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, pages 34892–34916. Curran Associates, Inc., 2023. 3

[29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 5

[30] 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. 2, 4

[31] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 10034–10043, 2020. 3

[32] Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14420–14431, 2024. 3

[33] Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14431, 2024. 4

[34] OpenAI. Gpt-4v. `https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4`, 2024. Accessed: 2024-11-10. 3

[35] OpenAI. Gpt-4o. `https://platform.openai.com/docs/models/gpt-4o`, 2024. Accessed: 2024-11-10. 3

[36] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32, 2019. 5

[37] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 3

[38] Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. Learning to predict visual attributes in the wild. In *CVPR*, 2021. 5

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763, 2021. 2, 6

[40] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19305–19314, 2023. 1, 2, 5, 6

[41] Jiangming Shi, Yachao Zhang, Xiangbo Yin, Yuan Xie, Zhizhong Zhang, Jianping Fan, Zhongchao Shi, and Yanyun Qu. Dual pseudo-labels interactive self-training for semi-supervised visible-infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11218–11228, 2023. 3

[42] Jiangming Shi, Xiangbo Yin, Yeyun Chen, Yachao Zhang, Zhizhong Zhang, Yuan Xie, and Yanyun Qu. Multi-memory matching for unsupervised visible-infrared person re-identification. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XVIII*, pages 456–474, 2024. 3

[43] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 8634–8652. Curran Associates, Inc., 2023. 3

[44] Haoyu Song, Li Dong, Wei-Nan Zhang, Ting Liu, and Furu Wei. Clip models are few-shot learners: Empirical studies on vqa and visual entailment, 2022. 2, 3

[45] Siqi Sun, Yen-Chun Chen, Linjie Li, Shuohang Wang, Yuwei Fang, and Jingjing Liu. LightningDOT: Pre-training visual-semantic embeddings for real-time image-text retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 982–997, Online, 2021. Association for Computational Linguistics. 3

[46] Shitong Sun, Fanghua Ye, and Shaogang Gong. Training-free zero-shot composed image retrieval with local concept reranking. *arXiv preprint arXiv:2312.08924*, 2023. 2

[47] Yucheng Suo, Fan Ma, Linchao Zhu, and Yi Yang. Knowledge-enhanced dual-stream zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26951–26962, 2024. 2

[48] Yuanmin Tang, Jing Yu, Keke Gai, Gang Xiong, Gaopeng Gou, and Qi Wu. Manipulation intention understanding for accurate zero-shot composed image retrieval. 2024.

[49] Yuanmin Tang, Jing Yu, Keke Gai, Jiamin Zhuang, Gaopeng Gou, Gang Xiong, and Qi Wu. Denoise-i2w: Mapping images to denoising words for accurate zero-shot composed image retrieval. *arXiv preprint arXiv:2410.17393*, 2024.

[50] Yuanmin Tang, Jing Yu, Keke Gai, Jiamin Zhuang, Gang Xiong, Yue Hu, and Qi Wu. Context-i2w: Mapping images to context-dependent words for accurate zero-shot composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5180–5188, 2024. 2, 5

[51] Yuanmin Tang, Jing Yu, Keke Gai, Jiamin Zhuang, Gang Xiong, Gaopeng Gou, and Qi Wu. Missing target-relevant information prediction with world model for accurate zero-shot composed image retrieval, 2025. 2

[52] Sagar Vaze, Nicolas Carion, and Ishan Misra. Genecis: A benchmark for general conditional image similarity. In *CVPR*, 2023. 4

[53] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval - an empirical odyssey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6439–6448, 2019. 1, 2

[54] Lei Wang, Yi Hu, Jiabang He, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 19162–19170, 2024. 3

[55] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 3, 4

[56] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11307–11317, 2021. 4

[57] 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. 2, 5, 7

[58] Daoan Zhang, Junming Yang, Hanjia Lyu, Zijian Jin, Yuan Yao, Mingkai Chen, and Jiebo Luo. Cocot: Contrastive chain-of-thought prompting for large multimodal models with multiple image inputs, 2024. 3

[59] Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. Magiclens: Self-supervised image retrieval with open-ended instructions, 2024. 2

[60] Yuechen Zhang, Shengju Qian, Bohao Peng, Shu Liu, and Jiaya Jia. Prompt highlighter: Interactive control for multi-modal llms. *arXiv preprint arXiv:2312.04302*, 2023. 3

[61] Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. What makes good examples for visual in-context learning? *Advances in Neural Information Processing Systems*, 2023. 3

[62] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models, 2023. 3

[63] Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:5168–5191, 2023. 3, 4

[64] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 2

[65] Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. Seqtr: A simple yet universal network for visual grounding. In *European Conference on Computer Vision*, pages 598–615. Springer, 2022. 3

[66] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 5