

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

Supplementary Material

Table of Contents

• A Complete Template for Reflective CoT	1
• B Algorithm of OSrCIR’s Process	1
• C Vision-by-Language In-Context Learning Details	3
• D More Ablation Study on One-Stage Reasoning and Backbone Models	3
• E Detailed Comparison of Computational Cost	5
• F Experiments on ImageNet and COCO	5
• G More Qualitative Analysis of Reflective CoT	5
• H More Implementation Details	5
– H.1 Evaluation Datasets Details	6
– H.2 Evaluation Tasks Details	9

We provide additional details and discussions on the components of the main paper. Specifically, we highlight Sections A to C for details on OSrCIR’s CIR process, and Sections D to H for in-depth analyses of our design choices with additional quantitative and qualitative experiments.

A. Complete Template for Reflective CoT

The complete template of our reflective CoT prompt is shown in Figure 1. 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.

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.

Thoughts. Given the intention-relevant visual details and manipulation text, the MLLM then seeks to capture the user’s intention. 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, 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 and identify the most relevant manipulated elements. 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.*

Target Image Description. Given the manipulated visual elements most relevant to the user’s intention, the AI agent needs to generate a target description that associates those manipulated visual elements for retrieval. We simply ask the MLLM to *generate a target image description that only contains the target image content.*

Input and Output. As shown in Figure 1, the input to the LLM is a concatenated prompt as $T_t = \Psi_M(p_c \circ I_r \circ T_m)$ comprising the base CoT prompt p_c , the base64-encoded image URL of the reference image I_r (prepended with “Original Image Context”), and the manipulation intent text T_m (prepended with “Manipulation Text”). This task-agnostic prompt format allows for application across various CIR tasks. The output is provided as a JSON file containing “Original Image Description”, “Thoughts”, “Reflections”, and “Target Image Description”. The “Target Image Description” is selected as the final output, while the additional information can serve as valuable reference data for LLM-based ensemble methods [27], potentially boosting performance at the cost of efficiency.

B. Algorithm of OSrCIR’s Process.

Algorithm 1 outlines OSrCIR’s process for training-free ZS-CIR. Given the target image description T_t , the model encodes the image-search database \mathcal{D} and T_t using a frozen pre-trained CLIP. The retrieved target image I_t is selected based on cosine similarity $\cos(\Psi_I(I_c), \Psi_T(T_t))$, where I_t is the image most similar to the generated description T_t . This retrieval process is modular and performed after combining the reference image and manipulation text, allowing for flexible substitution of retrieval systems to balance efficiency and effectiveness. The approach creates a human-understandable ZS-CIR pipeline, fully expressing reasoning in the language domain while keeping the retrieval process

You are an image description expert. You are given an original image and manipulation text. Your goal is to generate a target image description that reflects the changes described based on manipulation intents while retaining as much image content from the original image as possible.

Guidelines on generating the Original Image Description

- Ensure the original image description is thorough, capturing all visible objects, attributes, and elements.
- The original image description should be as accurate as possible, reflecting the content of the image.

Guidelines on generating the Thoughts

- In your Thoughts, explain your understanding of the manipulation intents and how you formulated the target image description.
- Provide insight into how you interpreted the manipulation intent in detail in the manipulation text.
- Discuss how the manipulation intent influenced which elements of the original image you focused.

Guidelines on generating the Reflections

- In your Reflections, summarize how the manipulation intent influenced your approach to transforming the original image description.
- Explain how the changes made reflect the specific semantic, Highlight key decisions that were made to preserve the coherence and context of the original image while meeting the manipulation intent.
- Reflect on the impact these changes have on the overall appearance or narrative of the image.
- Ensure that your reflections provide a concise yet insightful summary of the considerations and strategies applied in crafting the target description, offering a logical connection between the original and final content.

Guidelines on generating Target Image Description

- The target image description you generate should be complete and can cover various semantic aspects.
- The target image description only contains the target image content and needs to be as simple as possible. Minimize aesthetic descriptions as much as possible.

On the input format <Input>

- Input consist of two parts: The original image and the manipulation text.

```
{  
  "Original Image": <image_url>,  
  "Manipulation text": <manipulation_text>.  
}
```

Guidelines on determining the response <Response>

- Responses include the Original Image Context, Target Image Description, and Thoughts.

```
{  
  "Original Image Description": <original_image_description>,  
  "Thoughts": <thoughts>,  
  "Reflections": <reflections>,  
  "Target Image Description": <target_image_description>  
}
```

Here are some examples for reference:

...

Figure 1. The complete template of our reflective Chain-of-Thought process for Training-free ZS-CIR.

Algorithm 1 Computing Process of OSrCIR

Input: Reference image I_r , manipulation text T_m , reflective CoT prompt p_c , image-search database \mathcal{D} .

Parameters: Frozen MLLM Ψ_M , frozen CLIP vision encoder Ψ_I , language encoder Ψ_T .

Output: Retrieved target image I_t .

- 1: Initialize pre-trained and frozen models Ψ_M, Ψ_I, Ψ_T .
- 2: Generate target image description:

$$T_t = \Psi_M(p_c \circ I_r \circ T_m)$$

- 3: Compute normalized text embedding:

$$\hat{e}_T = \frac{\Psi_T(T_t)}{\|\Psi_T(T_t)\|}$$

- 4: **for** each image I_i in \mathcal{D} **do**
- 5: Compute normalized image embedding:

$$\hat{e}_{I_i} = \frac{\Psi_I(I_i)}{\|\Psi_I(I_i)\|}$$

- 6: Compute similarity score: $s_i = \hat{e}_{I_i}^\top \hat{e}_T$
 - 7: **end for**
 - 8: Retrieve target image: $I_t = \operatorname{argmax}_{I_i \in \mathcal{D}} s_i$
 - 9: **return** I_t
-

independent, without additional training or mapping modules.

C. Vision-by-Language In-Context Learning Details

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 [17, 25, 29].

To ensure a zero-shot setting in ZS-CIR, we propose a vision-by-language In-Context Learning (ICL) approach. As illustrated in Figure 2, our vision-by-language ICL provides a few expected MLLM outputs (*i.e.*, three samples) in text form as examples, without requiring a reference image to guide the MLLM through the reasoning process at each step. Notably, each sample uses the same placeholder “<image_url>” instead of an actual reference image URL, guiding the MLLM formatting of the input and output.

For instance, consider the manipulation text(sample 1): “Change to a large fancy white carriage, facing the opposite direction, must include man in a black suit and hat in-

stead of a woman.” The language-based description of the reference image is: “The image shows a woman in a black outfit and a large hat decorated with pink flowers, driving a small, wooden, two-wheeled carriage pulled by a miniature horse.” Following the Reflective CoT steps:

- **Original Image Description:** The MLLM captures all visible objects and attributes relevant to the manipulation text, ensuring fine-grained details are included. In this case, it notes the woman in a black outfit with a large hat, the small wooden carriage, the miniature horse, and the outdoor setting with a white fence and trees.
- **Thoughts:** The MLLM interprets the manipulation intent by explaining that the scene should be transformed into one featuring a large, fancy white carriage facing the opposite direction, and the woman replaced with a man in a black suit and hat. This step demonstrates the model’s understanding of the required changes and how they influence the focused elements in the original image.
- **Reflections:** The MLLM reflects on key decisions to preserve coherence while fulfilling the manipulation intent. It acknowledges that changing multiple components—such as the carriage’s appearance, the direction it faces, and the driver—introduces a more sophisticated aesthetic and alters the visual narrative. The reflection offers a logical connection between the original content and the final description.
- **Target Image Description:** The MLLM generates a refined description containing only the target image content: “A large, fancy white carriage facing the opposite direction, with a man in a black suit and hat driving it.”

This example illustrates how our vision-by-language in-context learning approach guides the MLLM through each step of the Reflective CoT process, enabling it to produce accurate and coherent descriptions for the target image without direct visual input. By providing language-based examples, the MLLM can internalize the reasoning pattern and apply it to new instances, ensuring consistency and effectiveness in zero-shot settings without reference images.

D. More Ablation Study

Table 1 presents additional ablation analyses. **(1) Models ‘2-4’ assess the significance of the one-stage reasoning strategy.** Using GPT-4o as the captioner with manipulation text to enhance the reference image captioning process (model ‘2’) results in a 3.62% performance decline, while incorporating GPT-4o with our Reflective CoT process (model ‘3’) leads to a 2.46% decline. These results highlight the necessity of our one-stage reasoning process for capturing complete reference image content and the importance of multimodal CoT for effective manipulation intention understanding. Incorporating manipulation text into caption generation in the two-stage approach (model ‘4’) achieves similar performance but introduces additional

```

## Example 1
<Input>
{
  "Original Image": <image_url>, (each sample uses the same placeholder "<image_url>" instead of an actual image URL)
  "Manipulation text": "Change to a large fancy white carriage, facing the opposite direction, must include man in a black suit and hat instead of a woman."
}
<Response>
{
  "Original Image Description": "The image shows a woman in a black outfit and a large hat decorated with pink flowers, driving a small, wooden, two-wheeled carriage pulled by a miniature horse. The scene takes place outdoors, with a white fence in the background and green trees visible beyond the fence. The horse is trotting along a paved surface, and the woman holds the reins in one hand while guiding the horse in a forward direction.",
  "Thoughts": "The manipulation intent is to transform the scene by changing the carriage to a large, fancy white one and altering the direction it faces. Additionally, the woman is to be replaced by a man in a black suit and hat. This involves a significant change in both the vehicle and the driver, shifting the visual style from simple and rustic to elegant and formal. The direction change also alters the spatial orientation of the scene.",
  "Reflections": "I reflect that the manipulation involved changing multiple components, such as the carriage's appearance, the direction of the scene, and the person driving. These changes introduce a more sophisticated and refined aesthetic, emphasizing the luxurious look of the new carriage and creating a different visual narrative. Additionally, replacing the subject from a woman to a man modifies the image's context while still keeping the same core activity—driving a horse-drawn carriage.",
  "Target Image Description": "A large, fancy white carriage facing the opposite direction, with a man in a black suit and hat driving it."
}

## Example 2
<Input>
{
  "Original Image": <image_url>, (each sample uses the same placeholder "<image_url>" instead of an actual image URL)
  "Manipulation text": "is dark and has more words and Desired item is blue and references Pink Floyd".
}
<Response>
{
  "Original Image Description": "The image features a person wearing a black T-shirt with a prominent graphic design. The shirt displays the iconic "Star Wars" logo at the top in bold, stylized letters. Below the logo, there is a colorful illustration that includes several well-known characters from the Star Wars franchise.",
  "Thoughts": "The manipulation intent is to alter specific visual and textual elements of the T-shirt while maintaining its thematic focus on the Star Wars franchise and incorporating an additional reference to Pink Floyd. In the original image, the T-shirt is black with a "Star Wars" logo and a colorful illustration of characters from the series. The manipulation text specifies that the shirt should be dark and have more words, and specifically be blue and reference Pink Floyd. These instructions suggest a change to a darker tone of blue for the shirt itself and the addition of text related to Pink Floyd, thereby marrying elements from two different cultural icons but maintaining the graphical theme of Star Wars.",
  "Reflections": "I reflect that the manipulation intent effectively combines elements from two cultural icons, adjusting color and adding textual references, while maintaining the original Star Wars theme and ensuring the central focus of the graphic design remains intact. By focusing on these specifications, the target image description actively integrates the new textual references and color change, bringing Pink Floyd into a traditionally Star Wars-themed T-shirt, while ensuring the design's central focus—Star Wars characters—remains intact.",
  "Target Image Description": "The shirt is a dark blue t-shirt featuring the "Star Wars" logo at the top, accompanied by additional text referencing Pink Floyd."
}

```

Figure 2. Examples of our vision-by-language in-context learning. Notably, each sample uses the same placeholder “<image_url>” instead of an actual reference image URL, guiding the MLLM formatting of the input and output.

Methods	CIRCO			Fashion-IQ	
	k=5	k=10	k=25	k=10	k=50
1. Full model (GPT-4o)	23.87	25.33	27.84	33.26	54.37
Significance of the one stage reasoning strategy					
2. two-stage+enhance captioner	20.93	21.34	23.27	30.14	50.87
3. two-stage+CoT	21.73	22.78	24.47	31.16	52.22
4. two-stage+enhance captioner+CoT	23.24	24.97	27.04	32.54	53.47
Impact of different backbone models					
5. BLIP	23.93	25.47	27.53	32.10	53.69
6. long clip	23.73	25.12	26.91	31.77	53.02
Alternative solutions for Reflective CoT					
7. CoT (w/o Reflective)	22.04	22.74	25.32	32.05	52.11
8. Simple prompt	20.86	21.40	23.34	30.27	51.06
9. Advanced prompt (DDCoT)	19.93	20.74	22.88	29.85	50.39

Table 1. More Ablation study on CIRCO and FashionIQ.

MLLM queries, reducing efficiency, and is therefore unnecessary. **(2) Models ‘5-6’ evaluate different backbone retrieval models.** OSrCIR with BLIP ViT-L/16 [13] and Long-CLIP ViT-L/14 [28] achieves results comparable to the CLIP backbone, demonstrating the generalizability and robustness of OSrCIR across different CLIP-based backbones. **Model ‘7-9’, we evaluate the effect of alternative solutions for Reflective CoT.** Specifically, we compare our Reflective CoT (Model ‘1’) with (i) a standard CoT (Model ‘7’) without reflective reasoning and (ii) a simpler prompt (Model ‘8’) without Reflective CoT. We observe performance drops of 2.08% and 3.55%, respectively. Additionally, (iii) replacing Reflective CoT with DDCoT [29] (Model ‘9’), a widely-used two-stage CoT that can first automatically divide each CIR query into sub-problems before reasoning the target image caption, leading to a significant 4.18% drop and even slightly below the simple prompt. These results validate the effectiveness of Reflective CoT in capturing user manipulation intent in CIR tasks.

E. Detailed Comparison of Computational Cost

In Table 5, we conducted a thorough comparison of query latency (averaged over 100 samples), GPU memory for retrieval, API cost per call, and average performance (ViT-L/14) on CIRCO and FashionIQ across our OSrCIR (MLLM-based) variants and two baselines, Context-I2W (CLIP-based) and CIReVL (LLM-based).

While CLIP-based models like Context-I2W have faster latency ($\sim 0.02s$), their GPU memory usage (16 GB) matches our MLLM-based OSrCIR models. LLM-based models like CIReVL require more memory (40 GB) due to an additional image captioning module (e.g., BLIP-2). Among our models, GPT-4o-mini offers a favorable trade-off with lower latency ($\sim 0.5 \pm 0.05s$) and API cost ($\sim \$0.002$) while maintaining similar performance to GPT-4o (31.29 vs. 32.27). These results highlight OSrCIR’s effi-

Model	LLM	Latency	GPU Memory	API Cost	Performance
Context-I2W	*	$\sim 0.02s$	16 GB	\$0	22.94
CIReVL	GPT-3.5	$\sim 1s$	40 GB	$\sim \$0.001$	26.23
OSrCIR	GPT-4o-mini	$\sim 0.5 \pm 0.05s$	16 GB	$\sim \$0.002$	31.29
OSrCIR	GPT-4o	$\sim 0.7 \pm 0.08s$	16 GB	$\sim \$0.004$	32.27

Table 5. Required Comparison of Computational Cost.

ciency and effectiveness, making it practical for real-world applications. We will update the “Effectiveness and Efficiency Analysis” section in the revised manuscript.

F. More Quantitative Experiments

In Tables 6 and 7, we evaluate OSrCIR on two tasks: ImageNet domain conversion and object composition, as proposed in [21]. For the domain conversion task, images from 200 classes of the original ImageNet dataset [5] are used as queries to retrieve images of the same object in a specified domain from ImageNet-R [9]. Details on the datasets and experimental settings are provided in Section H.1. Our model surpasses the state-of-the-art (SoTA) model, Context-I2W [23], in domain conversion, achieving an 18.2% average performance improvement on ViT-L/14. In the COCO object combination task, OSrCIR also demonstrates a notable average improvement of 6.17% on ViT-L/14. These results underscore that an MLLM, when combined with pre-trained vision-language models (e.g., CLIP), can leverage extensive pre-trained knowledge to enhance performance in domain conversion and object composition tasks significantly.

G. More Qualitative Analysis of Reflective CoT

To further explore the benefits of Reflective CoT in interpreting user intent, we present additional case studies from the CIRr validation set (Figure 3) and the FashionIQ validation set (Figure 4). Reflective CoT effectively captures user intent, as demonstrated in Figure 3 (Row 1), where it identifies the intent to “change the swing holder from wooden to white”. Additionally, Reflective CoT filters out irrelevant elements, as seen in Figure 4 (Row 1), where it infers the transition from a t-shirt to a ball cap based on both the manipulation text and reference image content. This ability to focus on relevant details enhances model robustness and likely contributes to its strong performance on CIR tasks.

H. More 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 [15] and MiniGPT4 [30]. 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 [18] based on

Backbones	Methods	Conferences	Cartoon		Origami		Toy		Sculpture		Average	
			R10	R50	R10	R50	R10	R50	R10	R50	R10	R50
ViT-L/14	Pic2Word [†]	CVPR 2023	8.0	21.9	13.5	25.6	8.7	21.6	10.0	23.8	10.1	23.2
	SEARLE	ICCV 2023	9.6	24.9	16.1	27.3	7.6	25.4	11.3	26.4	11.2	26.0
	LinCIR	CVPR 2024	9.4	24.2	15.7	26.9	10.8	27.0	11.7	27.9	11.9	26.5
	Context-I2W [†]	AAAI 2024	10.2	26.1	17.5	28.7	11.6	27.4	12.1	28.2	12.9	27.6
	OSrCIR	–	24.0	48.4	27.4	49.1	33.8	45.3	27.8	51.4	28.3	48.6
ViT-G/14	LinCIR	CVPR 2024	13.7	30.2	19.5	32.9	13.8	30.2	15.2	34.0	15.5	31.8
	OSrCIR	–	26.3	53.5	36.7	45.8	29.2	52.5	31.8	52.6	31.0	51.1

Table 6. Results on ImageNet for domain conversion. [†] indicates results from the original paper.

Backbones	Methods	R1	R5	R10
ViT-L/14	Pic2Word [†]	11.5	24.8	33.4
	SEARLE	13.3	28.3	37.6
	LinCIR	11.7	24.9	34.2
	Context-I2W [†]	13.5	28.5	38.1
	OSrCIR	17.3	35.4	45.9
ViT-G/14	LinCIR	14.8	30.6	40.5
	OSrCIR	19.0	36.4	48.2

Table 7. Results on COCO for object composition. [†] indicates results from the original paper.

the codebase from [11], performs all computations on a single NVIDIA A100 GPU. For the CLIP-based ViT variants [6], we adopt weights from the official CLIP implementation [20] while using OpenCLIP [10] for ViT-G/14. LLaVA weights are taken from LLaVA-1.6-13B [15], and MiniGPT-4 weights from MiniGPT-4 (Vicuna 13B) [30]. Performance metrics are averaged over three trials to ensure reliability.

H.1. Evaluation Datasets Details

We evaluate our model on four ZS-CIR datasets in our main paper, two ZS-CIR datasets in our supplementary, *i.e.*, COCO [14] and GeneCIS [24] for object composition, ImageNet [5, 9] for domain conversion, CIRR [16] and CIRCO [3] for object/scene manipulation, and Fashion-IQ [26] for attribute manipulation. Following the original benchmarks, we use Recall@k (R@k) as the evaluation metric for CIRR, GeneCIS, and FashionIQ, COCO, ImageNet and mean average precision (mAP@k) for CIRCO to account for multiple positives. We also evaluate CIRR in a subset setting, where Recall_{Subset}@k measures retrieval performance within a limited selection of images relevant to the query in the database. The evaluation datasets are preprocessed, as explained in the main paper, we describe the details of the dataset, *i.e.*, number of query images and candidate images used for evaluation.

FashionIQ [26] is a dataset of fashion-related images across three categories: Shirt, Dress, and Toptee, comprising 30,134 triplets from 77,684 images. The dataset was curated by collecting image attributes and then tasking human annotators to write captions describing highly related

images based on those attributes. FashionIQ simulates realistic user interactions, as captions were generated via a chat-based visual interface to mimic online shopping queries. The dataset is divided into training (60%), validation (20%), and test (20%) splits. For zero-shot CIR, we use only the validation split, as the test set annotations are not publicly available.

CIRR [16] contains 21,552 real-world images sourced from NLVR² [22]. The dataset includes training, validation, and test splits, with the latter evaluated via a remote server. Our analysis focuses on the validation split for model selection. Unlike FashionIQ, which targets fashion-specific queries, CIRR encompasses diverse domains with complex descriptions. The dataset was built by identifying visually similar images using ResNet-152 [8] pretrained on ImageNet [5] and employing human annotators to describe differences between paired images. However, CIRR suffers from two key issues: (1) image pairs identified by ResNet often lack true visual similarity, as they were not verified by human annotators; and (2) captions are often unrealistic or ambiguous, including unnecessary details. These limitations reduce CIRR’s practical relevance compared to FashionIQ. Additionally, CIRR uses a small subset retrieval task (*e.g.*, five items) to mitigate noise, but this approach is problematic, as the target image often relates only to the text condition rather than the reference image. Previous studies [3, 7, 21], have noted the prevalence of false negatives (FNs) in CIRR, complicating evaluation accuracy, as seen in other cross-modal retrieval tasks [4, 31].

Notably, both FashionIQ and CIRR face challenges from FN instances. While each query has a single labeled positive, multiple valid matches may exist in the dataset. FashionIQ mitigates this by reporting Recall@K with larger K values (*e.g.*, 10 or 50), whereas CIRR employs subset retrieval. However, these approaches fail to fundamentally resolve the FN issue, particularly for CIRR’s noisy annotations.

CIRCO [3] builds on the COCO dataset [14], addressing the FN problem by including an average of 4.53 ground truths per query. This design enables more reliable eval-

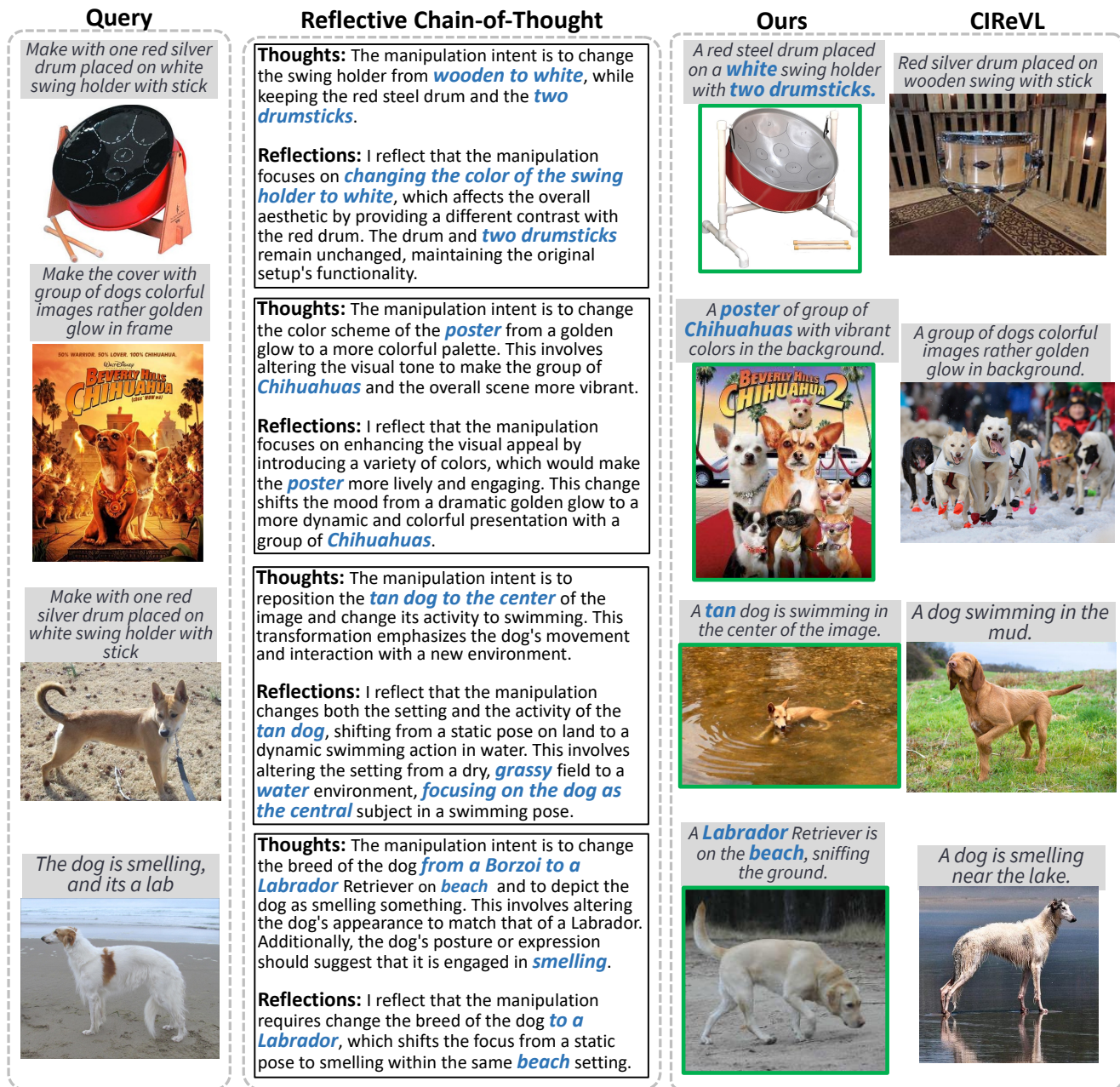


Figure 3. Visualization of Reflective CoT samples on CIRR validation set. We compare the top 1 retrieval results of ours and CIReVL. Our Reflective CoT effectively captures the user's intent and filters out elements irrelevant to user intention.



Figure 4. Visualization of Reflective CoT samples on FashionIQ validation set. We compare the top 1 retrieval results of ours and CIReVL. Our Reflective CoT effectively captures the user’s intent and filters out elements irrelevant to user intention.

uation using metrics like mAP. CIRCO contains no training split and provides validation (220 queries) and test (800 queries) splits, with the latter evaluated remotely.

GeneCIS [24] defines conditional retrieval tasks focusing on attributes (*e.g.*, “focus on an attribute”, “change an attribute”) and objects (*e.g.*, “focus on an object”, “change an object”). Attribute tasks use VisualGenome [12] and VAW [19], while object tasks are based on COCO [14]. Each task comprises around 2,000 queries with a small gallery size (*e.g.*, 15 images, 10 for “focus on an attribute”) to limit FNs. Text queries correspond to attributes or objects (*e.g.*, “color”, “backpack”).

COCO [14] contains images with corresponding lists of object classes and instance masks of query images. Following Pic2Word, we randomly crop one object and mask its background using its instance mask to create a query for each image. The list of object classes is used as text specification.

ImageNet [14] consists of 200 classes across diverse domains with domain annotations. Unlike previous benchmarks, the task involves retrieving an image in the specified domain for the same semantic object category (*e.g.*, retrieving a cartoon goldfish given a natural goldfish reference image and the modifier “cartoon”). This task requires no reasoning over image semantics, as the modifier independently specifies a domain change. Significant improvements over Pic2Word and Context-I2W can be achieved by leveraging the final description format, “a domain of a caption”.

Table 8. The number of images used for evaluation in each dataset.

Dataset	Query images	Candidate images
CIRR (Test)	4,148	2,315
CIRCO (Test)	800	123,403
Fashion (Dress)	2,017	3,817
Fashion (Shirt)	2,038	6,346
Fashion (TopTee)	1,961	5,373
GeneCIS (Focus Attribute)	2000	10
GeneCIS (Change Attribute)	2112	15
GeneCIS (Focus Object)	1960	15
GeneCIS (Change Object)	1960	15
COCO	4,766	4,766
ImageNet	10,000	16,983

H.2. Evaluation Tasks Details

(1) Object/Attribute composition. We evaluate the GeneCIS [24] test split and the validation split (5000 images) of COCO [14], which dataset contains images with corresponding lists of object classes and instance mask of

query images. Following Pic2Word, we randomly crop one object and mask its background using its instance mask to create a query for each image. The list of object classes is used as text specification. Similarly, the GeneCIS dataset introduces four task variations, such as changing a specific attribute or object.

(2) Object/scene manipulation by text description. In this setup, a reference image is provided alongside a text description containing instructions for manipulating either an object or the background scene depicted in the reference image. This composition of the reference image and text description enables the retrieval of manipulated images. We evaluate the test split of CIRR [16] and CIRCO [3] using the standard evaluation protocol.

(3) Attribute manipulation. We employ Fashion-IQ [26], which includes various modification texts related to image attributes. These attribute manipulations are given as a sentence. In evaluation, we employ the validation set, following previous works [2, 3, 21, 23].

(4) Domain conversion. This setup evaluates the ability to compose real images and domain information to retrieve corresponding domain-specific images. We utilize ImageNet [5] and ImageNet-R [9], which comprises 200 classes with diverse domains and has domain annotations. Following Pic2Word, we pick cartoon, origami, toy, and sculpture as the evaluation target to avoid noise in the annotations. With this selection, we have 16,983 images as candidates.

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] 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. 9
- [3] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. *arXiv:2303.15247*, 2023. 6, 9
- [4] 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. 6
- [5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition*, pages 248–255, 2009. 5, 6, 9
- [6] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6
- [7] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, , Yoohoon Kang, and Sangdoon Yun. Language-only efficient training

- of zero-shot composed image retrieval. In *Conference on Computer Vision and Pattern Recognition*, 2024. 6
- [8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 6
- [9] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 5, 6, 9
- [10] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hananeh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip. 6
- [11] Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. Vision-by-language for training-free compositional image retrieval. *arXiv preprint arXiv:2310.09291*, 2023. 6
- [12] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, page 32–73, 2017. 9
- [13] 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. 5
- [14] 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 *European Conference on Computer Vision*, pages 740–755, 2014. 6, 9
- [15] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 5, 6
- [16] 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. 6, 9
- [17] 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. 3
- [18] 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
- [19] 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. 9
- [20] 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. 6
- [21] 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. 5, 6, 9
- [22] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Hua-jun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2018. 6
- [23] 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. 5, 9
- [24] Sagar Vaze, Nicolas Carion, and Ishan Misra. Genecis: A benchmark for general conditional image similarity. In *CVPR*, 2023. 6, 9
- [25] 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
- [26] 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. 6, 9
- [27] 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. 1
- [28] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip, 2024. 5
- [29] Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibeil 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, 5
- [30] 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, 6

- [31] Jason Zhu, Yanling Cui, Yuming Liu, Hao Sun, Xue Li, Markus Pelger, Tianqi Yang, Liangjie Zhang, Ruofei Zhang, and Huasha Zhao. Textgnn: Improving text encoder via graph neural network in sponsored search. In *Proceedings of the Web Conference*, pages 2848–2857, 2021. [6](#)