

# CoTMR: Chain-of-Thought Multi-Scale Reasoning for Training-Free Zero-Shot Composed Image Retrieval

## Supplementary Material

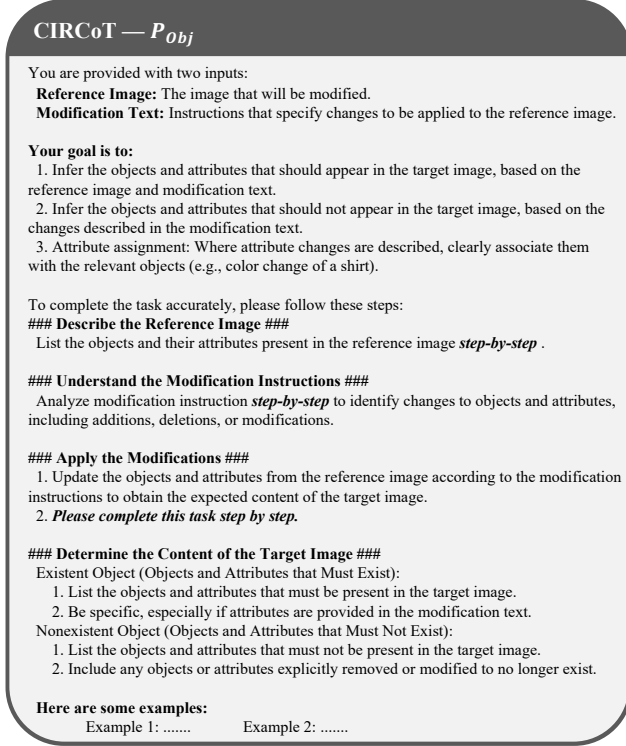


Figure 7. **Illustration of CIRCoT in object-scale reasoning ( $P_O$ )**, which includes four predefined subtasks and allows LVLM to reason step-by-step within each subtasks.

## 8. CIRCoT in Object-Scale Reasoning

In Figure 7, we show the details of CIRCoT used in object-scale reasoning ( $P_O$ ). Similar to image-scale reasoning, we divide the task of generating “existent objects” and “nonexistent objects” into four subtasks: (1) *Describe the Reference Image*; (2) *Understand the Modification Instructions*; (3) *Apply the Modifications* and (4) *Determine the Content of the Target Image*. These four fundamental tasks serve to structure the overall reasoning process of the LVLM. For each subtask, we allow the model to reason step-by-step. Additionally, we add several reasoning examples to further stimulate the model’s reasoning capabilities.

## 9. Example of User Interventions

CIRCoT enables a highly transparent and interpretable reasoning process, which facilitates error tracking and correction through user intervention when necessary. We demonstrate this capability through illustrative cases in Figure 8,

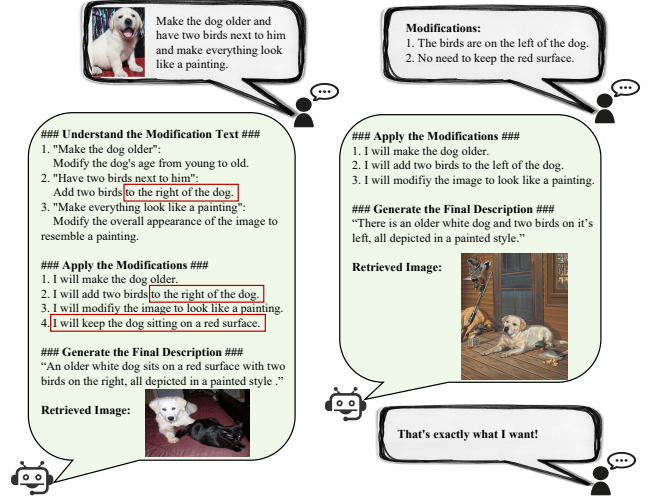


Figure 8. **The demonstration of making user interventions to enhance ZS-CIR performance with CoTMR.** For instance, by fixing the mistakes in the reasoning process, users are able to correctly retrieve the desired image with further dialogue.

where initial reasoning processes led to suboptimal retrieval results. The structured nature of our reasoning framework allows users to precisely identify problematic reasoning steps and initiate corrective interactions with the LVLM. As illustrated in Figure 8, we present instances where users successfully identified and addressed two reasoning errors: *the bird is to the right of the dog* and *red surface*". Through subsequent dialogue-based refinement, the model’s retrieval accuracy was effectively improved, highlighting the practical value of our interpretable reasoning approach.

## 10. Ablation Study on Multi-Scale Reasoning

Tables 5 and 6 present a comparative analysis of single-process versus dual-process approaches in the multi-scale reasoning module, evaluated on the Fashion-IQ and CIRR datasets using the ViT-B/32 CLIP model. “One process” refers to generating all three responses with LVLM simultaneously in a single inference pass. “Two processes” represents our default methodology, which conducts reasoning separately at different scales through independent inference processes. Analysis of Table 5 reveals that employing a single process not only compromises the effectiveness of image-scale reasoning but also diminishes the performance gains typically achieved through object-scale reasoning integration. In the more challenging CIRR dataset, as shown in Table 6, while concurrent reasoning of target

	Method	Shirt		Dress		Tops&Tee		Avg.		$R_{mean}$
		R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	
One process	image-scale	29.44	47.11	26.82	48.79	32.33	54.77	29.53	50.22	39.87
	multi-scale	32.14	50.25	30.04	51.81	35.34	59.10	32.50	53.72	43.11
Two processes	image-scale	30.03	48.58	26.57	48.69	34.12	56.35	30.24	51.21	40.72
	multi-scale	<b>33.42</b>	<b>53.93</b>	<b>31.09</b>	<b>54.54</b>	<b>38.40</b>	<b>61.14</b>	<b>34.30</b>	<b>56.54</b>	<b>45.42</b>

Table 5. **Ablation study on the impact of process quantity in Multi-Scale Reasoning on Fashion-IQ dataset.** All experiments are performed with the ViT-B/32 CLIP model.

	Method	Recall@k				Recall <sub>sub</sub> @k			Avg.
		k=1	k=5	k=10	k=50	k=1	k=2	k=3	
One process	image-scale	30.76	59.01	70.75	90.34	66.08	83.74	91.96	62.54
	multi-scale	29.56	58.69	70.27	89.72	65.61	83.52	91.68	62.15
Two processes	image-scale	30.11	58.10	70.58	89.95	65.08	83.07	91.41	61.59
	multi-scale	<b>31.88</b>	<b>61.27</b>	<b>72.90</b>	<b>91.03</b>	<b>67.85</b>	<b>85.00</b>	<b>92.68</b>	<b>64.56</b>

Table 6. **Ablation study on the impact of process quantity in Multi-Scale Reasoning on CIRR val dataset.** All experiments are performed with the ViT-B/32 CLIP model.

image caption and key objects enhances image-scale reasoning accuracy, the incorporation of object-scale reasoning results yields a marginal performance degradation. We attribute these observations to two primary factors: (1) The utilization of identical reasoning logic across both scales potentially limits the semantic richness of the reasoning outcomes. (2) Qwen2-VL’s current capabilities in managing multiple concurrent tasks may be insufficient, where the increased cognitive load adversely affects the precision of the results.

## 11. GeneCIS Full Results

Table 7 shows the full results of the comparison methods on GeneCIS. In the table, we observe that CoTMR shows significantly better performances than others, especially for “Focus” tasks and “Object” tasks. We attribute this superior performance to two key factors. First, CoTMR’s ability to simultaneously process both reference images and modification text enables more fine-grained attention to desired content in reference images, reducing information loss and thereby enhancing performance on “Focus” tasks. Second, our proposed multi-scale reasoning mechanism enables object-level reasoning about the presence or absence of objects in reference images, leading to improved performance on “Object” tasks.

## 12. Integration with LinCIR

Currently, there are two mainstream approaches in ZS-CIR: pseudo-token-based methods [3, 13, 36] and textual caption-based methods (including LLM-based [17, 39, 47] methods and LVLM-based methods, such as our CoTMR). In this section, we compare CoTMR with LinCIR [13], a state-of-the-art pseudo-token-based method, and prelim-

inarily explore the potential for their collaboration.

We conduct a preliminary investigation into the potential of combining pseudo-token-based and textual caption-based methods to achieve superior performance. Specifically, we first convert the reference image into a pseudo-token following LinCIR. Then, for the multi-scale reasoning outputs in CoTMR, we concatenated the pseudo-token with both the “*target image caption*” and the combination of “*existent objects*”, while maintaining the “*nonexistent objects*” unchanged. We then retrieved target images using the same scoring mechanism as CoTMR. As shown in the last row of Table 8, CoTMR with additional pseudo-tokens achieves substantial improvements (e.g., a **6.42%** increase in  $R_{mean}$ ). Similarly, the multi-grained descriptions generated through CoTMR’s reasoning process help enhance LinCIR’s performance (e.g., a **1.83%** improvement in  $R_{mean}$ ). This experiment suggests a promising direction for ZS-CIR research: optimizing the text that concatenated with pseudo-tokens with LVLM. We leave this exploration for future work.

## 13. Qualitative Example of CIRCoT

In Figure 9, we illustrate the reasoning process generated by the LVLM when using CIRCoT at both image and object scale. During image-scale reasoning, the LVLM analyzes the global content of the reference image to ensure comprehensive information coverage. By incrementally breaking down the modification text and executing the modification process, each user modification intent is accurately and completely executed. At object-scale reasoning, the LVLM focuses on the objects and their attributes in the reference image, accurately reasoning which objects and attributes should or should not be present by executing the modification process step-by-step. As a result, the LVLM success-

Backbone	Method	Training-free	Focus Attribute			Change Attribute			Focus Object			Change Object			Avg. R@1
			R@1	R@2	R@3	R@1	R@2	R@3	R@1	R@2	R@3	R@1	R@2	R@3	
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
	CoTMR	✓	<b>21.3</b>	<b>34.3</b>	<b>45.9</b>	<b>15.3</b>	<b>27.0</b>	<b>36.3</b>	<b>16.8</b>	<b>26.9</b>	<b>35.5</b>	<b>18.0</b>	<b>30.3</b>	<b>40.3</b>	<b>17.9</b>
ViT-L/14	SEARLE	✗	17.1	29.6	40.7	<b>16.3</b>	25.2	34.2	12.0	22.2	30.9	12.0	24.1	33.9	14.4
	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
	LinCIR	✗	16.9	30.0	41.5	16.2	<b>28.0</b>	<b>36.8</b>	8.3	17.4	26.2	7.4	15.7	25.0	12.2
	CoTMR	✓	<b>21.0</b>	<b>35.3</b>	<b>45.8</b>	15.6	26.0	36.6	<b>15.1</b>	<b>26.6</b>	<b>36.2</b>	<b>18.8</b>	<b>30.6</b>	<b>40.6</b>	<b>17.6</b>
ViT-G/14	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
	LinCIR	✗	19.1	33.0	42.3	<b>17.6</b>	<b>30.2</b>	38.1	10.1	19.1	28.1	7.9	16.3	25.7	13.7
	CoTMR	✓	<b>22.4</b>	<b>35.4</b>	<b>45.1</b>	17.2	30.0	<b>39.7</b>	<b>17.2</b>	<b>28.5</b>	<b>37.8</b>	<b>19.5</b>	<b>31.5</b>	<b>41.4</b>	<b>19.1</b>

Table 7. Comparison with the state-of-the-art methods on GeneCIS test set. The best results are in boldface.

Backbone	Method	Training-free	Shirt		Dress		Tops&Tee		Avg.	
			R@10	R@50	R@10	R@50	R@10	R@50	R@10	R <sub>mean</sub>
ViT-G/14	LinCIR	✗	44.08	62.56	38.48	60.63	48.58	<b>69.10</b>	43.71	53.91
	CoTMR	✓	38.32	62.24	33.96	56.22	40.90	64.30	37.72	49.32
	CoTMR + LinCIR	✗	<b>46.24</b>	<b>67.95</b>	<b>40.06</b>	<b>63.11</b>	<b>49.59</b>	68.03	<b>45.13</b>	<b>55.74</b>

Table 8. Results of the integratio of CoTMR and LinCIR [13] on the Fashion-IQ dataset.  $R_{mean}$  indicates the average results across all the metrics. We reproduce the results of LinCIR. The best results are in boldface.

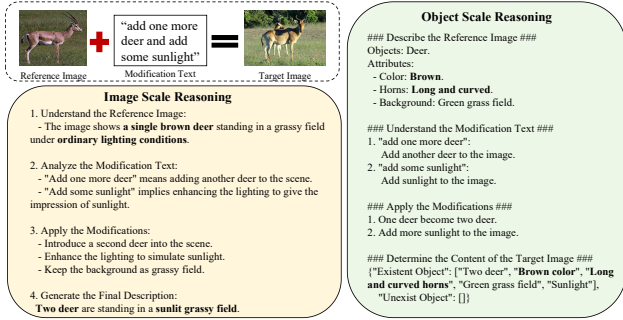


Figure 9. An example of a reasoning process with CIRCoT from CIRR val set. The LVLM focuses on specific objectives in each subtask within CIRCoT and gradually completes the overall reasoning goal.

fully noticed the key object, i.e., "long and curved horns". This predefined structured reasoning process standardizes the model’s reasoning path, preventing user modification intents from being overlooked or incorrectly propagated.

## 14. More Qualitative Examples

Figure 10 visualizes more cases where the combination of image-scale and object-scale reasoning leads to successful retrievals on both Fashion-IQ and CIRR datasets. (1) In the first example, the top two images retrieved using the target image caption overlooked the semantics of "Asian-inspired design". However, after emphasizing this part with existent objects, the target image was successfully ranked first. (2) In the second example, the target image caption contained distracting information (religious message and cross design), causing the top three images to include some religious elements. By incorporating nonexistent objects, our model

successfully reduced the impact of this distracting information. (3) In the third example, the reference image had very little relevance to the target image, meaning the model could easily be misled by distracting information from the reference image, such as "in a basket". Object-scale reasoning can reduce such distractions because it doesn’t need to consider the logical relationships between objects. Thus, our model successfully categorized "Blue basket" as a nonexistent object. (4) In the fourth example, the target image caption was similarly affected by the "Pepsi logos" in the reference image. Our model mitigated the distracting influence by emphasizing "Green bottles" and successfully ranked the target image first.




 <p><b>Reference Image</b></p> <p><i>"Is longer and more Asian-inspired and shiny black."</i></p> <p><b>Modification Text</b></p>	<p><b>Target Image caption</b></p> <p>"The woman is wearing a long, shiny black dress with an Asian-inspired design and short sleeves."</p> <p>+</p> <p><b>Existent Objects:</b></p> <p>[Long dress, Shiny black, <i>Asian-inspired elements.</i>]</p> <p><b>Nonexistent Objects:</b></p> <p>[Purple dress , Knee-length]</p>	
 <p><b>Reference Image</b></p> <p><i>"is white colored and is less religious and more humorous."</i></p> <p><b>Modification Text</b></p>	<p><b>Target Image caption</b></p> <p>"A white t-shirt with a humorous graphic or text, <u>replacing the original religious message and cross design.</u>"</p> <p>+</p> <p><b>Existent Objects:</b></p> <p>[White t-shirt, <i>Humorous design.</i>]</p> <p><b>Nonexistent Objects:</b></p> <p>[Gray t-shirt, <i>cross design, Religious design.</i>]</p>	
 <p><b>Reference Image</b></p> <p><i>"Show the brown dogs with water and snow behind them."</i></p> <p><b>Modification Text</b></p>	<p><b>Target Image caption</b></p> <p>"Brown dogs are <u>in a basket</u> with water and snow behind them."</p> <p>+</p> <p><b>Existent Objects:</b></p> <p>[Brown dogs, Water, Snow. ]</p> <p><b>Nonexistent Objects:</b></p> <p>[Crabs , <i>Blue basket</i> , Netting. ]</p>	
 <p><b>Reference Image</b></p> <p><i>"the bottles are on a table, and they are green."</i></p> <p><b>Modification Text</b></p>	<p><b>Target Image caption</b></p> <p>"Several green glass bottles with <u>Pepsi logos</u> are placed on a table."</p> <p>+</p> <p><b>Existent Objects:</b></p> <p>[<i>Green bottles</i>, Table.]</p> <p><b>Nonexistent Objects:</b></p> <p>[Clear glass bottles. ]</p>	

Figure 10. **Successful retrieval examples with multi-scale reasoning from Fashion-IQ and CIRRR val set.** The ground-truth image is highlighted with the red box. Red underlined text indicates distracting information that causes mistake retrieval, while green italicized text represents key objects that help in correct retrieval.