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

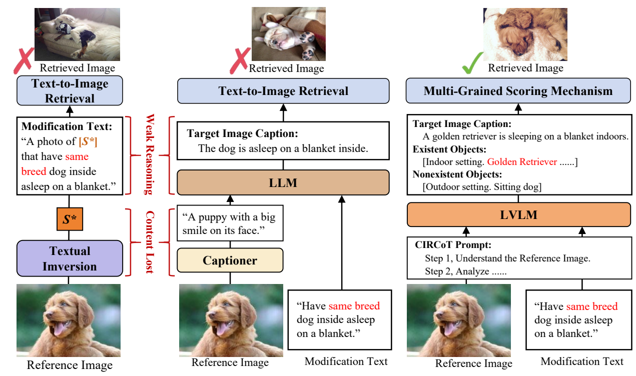
Zelong Sun*, Dong Jing* , Zhiwu Lu †
Gaoling School of Artificial Intelligence
Renmin University of China, Beijing, China
zelongsun@ruc.edu.cn, luzhiwu@ruc.edu.cn

Abstract

Zero-Shot Composed Image Retrieval (ZS-CIR) aims to retrieve target images by integrating information from a composed query (reference image and modification text) without training samples. Existing methods primarily combine caption models and Large Language Models (LLMs) to generate target captions from composed queries but face various issues such as incompatibility, visual information loss, and insufficient reasoning. In this work, we propose CoTMR, a training-free framework with novel Chain-of-thought (CoT) and Multi-scale Reasoning. Instead of relying on caption models for modality transformation, CoTMR directly employs the Large Vision-Language Model (LVLM) to achieve unified understanding and reasoning of composed queries. To enhance reasoning reliability, we devise CIRCoT, which guides the LVLM to perform step-by-step reasoning by following predefined subtasks. Additionally, while most existing approaches focus solely on global-level reasoning, CoTMR introduces fine-grained predictions about the presence or absence of key elements at the object scale for more comprehensive reasoning. Furthermore, we design a Multi-Grained Scoring (MGS) mechanism, which integrates CLIP similarity scores of the above reasoning outputs with candidate images to realize precise retrieval. Extensive experiments demonstrate that our CoTMR not only drastically outperforms previous methods across four prominent benchmarks but also offers appealing interpretability.

1. Introduction

Zero-Shot Composed Image Retrieval (ZS-CIR) [3, 36, 40] aims to retrieve the target image by integrating information from a reference image and a modification text, without training with annotated triplets data. In contrast to traditional image retrieval tasks [9–11, 24], CIR queries necessitate precise "editing" to the reference image based on the modification text. Therefore, successfully completing the CIR task entails (1) **Advanced multimodal composed understanding abilities** to accurately interpret the visual context and user's modification intent in modification text, and (2) **Robust multimodal reasoning abilities** to implement the modifications appropriately.



(a) Pseudo-Token (b) LLM-based Method (c) **Ours (CoTMR)**

Figure 1. **Flowcharts of existing ZS-CIR methods and our proposed CoTMR.** Methods (a) and (b) face serious issues of visual information loss and insufficient reasoning. In contrast, our method (c) fully perceives image content, enhances reasoning process with CIRCoT, and augments multi-grained descriptions with multi-scale reasoning.

As shown in Figure 1 (a), previous methods [3, 36] primarily propose a textual inversion module to generate pseudo-tokens from the reference image and concatenate it with the modification text. However, these methods still require extensive data for training, and the relatively short length of pseudo-tokens limits the model's ability to fully represent the reference image. Notably, these statistical methods lack enough logical reasoning for the CIR task. Recent works [17, 47] leverage Large Language Models (LLMs) to identify the user's modification intent. As shown in Figure 1 (b), these methods use pre-trained caption models to generate a caption for the reference image, and then employ the LLM to edit this caption based on the modifi-

*Equal Contribution
†Corresponding Author

cation text. However, the cascading combination of different models introduces several challenges. (1) *Component Incompatibility:* There are domain gaps in language style and way of thought between caption models and LLMs; (2) *Visual Information Loss:* During the caption generation process, some detailed information about the reference image is inevitably lost; (3) *Single-scale Reasoning:* Existing methods focus solely on image-scale reasoning, neglecting fine-grained details; (4) *Insufficient Reasoning:* As a key component, current approaches have not fully leveraged the reasoning capability of LLM. Therefore, as shown in Figure 1, the aforementioned methods make it hard to preserve the "golden retriever" characteristics in the reference image.

In this work, we propose **CoTMR**, a training-free and highly interpretable framework with novel Chain-of-Thought (CoT) and Multi-scale Reasoning. As shown in Figure 1 (c), instead of relying on the combination of caption models and LLMs, CoTMR employs the Large Vision-Language Model (LVLM) to achieve unified understanding and reasoning for composed queries. This framework offers several appealing benefits, including rich visual information, unified reasoning, and simplified workflow. Furthermore, we propose a novel CoT method, named **CIRCoT**, to further enhance the reasoning capability and interpretability of the LVLM. Unlike previous works [50] that entirely delegate the task decomposition process to the model, CIRCoT pre-divides the CIR task into multiple subtasks and allows the model to reason each pre-defined subtask step-by-step. Additionally, a few examples can also be included for reference in CoT [45]. This structured reasoning process not only guides the LVLM through a step-by-step inference process but also provides high-level interpretability, allowing users to intervene for more precise retrieval when necessary.

With this structured reasoning process, we further propose **Multi-Scale Reasoning** to predict both the global description and fine-grained details of target image from composed query. As shown in Figure 1 (c), in addition to reasoning at the image scale, we further conduct object-scale reasoning to emphasize key objects and attributes. Notably, aligning with the requirement of CIR, we should not only infer the objects that should be present in the target image ("existent objects") but also naturally take those that should not be present ("nonexistent objects") into account. The existent objects further supplement the target image caption, while nonexistent objects are used to reduce distracting information. Subsequently, **Multi-Grained Scoring (MGS)** mechanism is designed to enable a precise retrieval process. This module comprehensively considers the characteristics of multi-grained outputs and separately calculates their similarity scores with the candidate images via CLIP [33]. Ultimately, MGS integrates these scores to achieve a balanced evaluation by rewarding the presence of relevant content while penalizing irrelevant or conflicting content.

Our main contributions can be summarized as follows: (1) We propose CoTMR, a novel training-free LVLM-based framework for ZS-CIR. (2) We present multi-scale reasoning and a novel scoring module to provide multi-grained descriptions and evaluations. (3) We design a novel CIRCoT, which standardizes the LVLM's reasoning process, allowing it to focus on specific goals at each subtask. (4) Extensive experiments demonstrate that our CoTMR not only significantly outperforms state-of-the-art methods across four prominent benchmarks but also offers appealing interpretability for CIR.

2. Related Work

2.1. Zero-Shot Composed Image Retrieval

CIR [2, 5, 20, 43] integrates concepts from compositional learning [16, 27] and cross-modal retrieval [31, 32]. To mitigate the high cost and time-consuming nature of training dataset annotation for CIR, ZS-CIR has recently been introduced. Currently, two prominent directions exist: one approach [3, 13, 36] trains a textual inversion module using only image-caption data, representing the reference image with a single pseudo-token that is then concatenated with the reference caption. This method not only requires training but also is limited by the length of the pseudo-token, which constrains the representation of the reference image. The other approach [17, 39, 47] forms a training-free method by cascading multiple off-the-shelf tools. It first converts reference image into a textual description using a captioning model and then edits this caption according to the modification text by a LLM. Finally, the edited caption is used to compute CLIP scores with candidate images for retrieval. However, such methods face several challenges, including component incompatibility, visual information loss, and insufficient reasoning. In this work, we propose a unified, training-free, and interpretable framework with CIRCoT and Multi-Scale Reasoning.

2.2. Vison-Language Model

There are two main types of Vision-Language Models (VLMs). The first type, including models like CLIP [33] and BLIP [21], is pre-trained on large-scale image-caption datasets, enabling them to map images and text into a shared embedding space for cross-modal retrieval [4, 8, 34] or open-vocabulary classification [30, 41]. In this work, we use CLIP for the multimodal retrieval process. The second type is LVLM [7, 23, 44], which are pre-trained to integrate visual information into LLM and are post-trained to align with users. Thus, LVLMs could understand user intent and process various visual tasks, such as image captioning [1, 22], VQA [12, 14], and OCR [26, 37]. In this work, we utilize the LVLM to inference both the global description and fine-grained details based on composed queries.

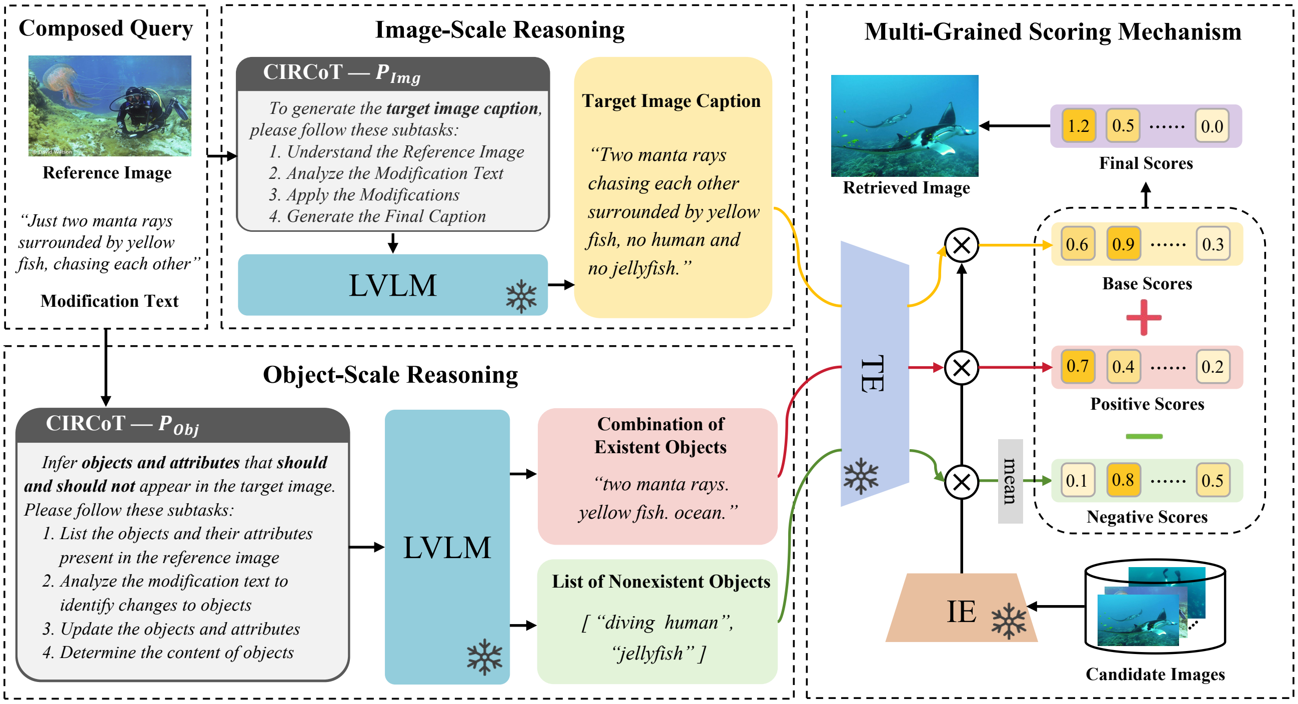


Figure 2. **Overview architecture of CoTMR:** (1) The LVLM equipped with CIRCoT, P_{Img} and P_{Obj}, performs reasoning on the composed query at both image and object scales, to provide multi-grained outputs. (2) The Multi-Grained Scoring Mechanism combines the similarities of the three outputs with candidate images in the CLIP space through a reward-penalty calculation. IE and TE represent the image encoder and text encoder of CLIP, respectively.

2.3. Chain of Thought

Recently, zero-shot [18] and few-shot [35, 45] multi-step reasoning prompts have shown significant enhancement to the reasoning capabilities of LLMs. Consequently, CoT strategy raises increasing research attention and is also extended into multimodal domains. MM-CoT [49] designs a two-stage framework where the model initially learns to generate rationales based on real annotations and then uses all available information to produce the final answer. DD-CoT [50] focuses on text understanding, breaking down questions into sub-questions for step-by-step responses. CCoT [28], on the other hand, is based on image understanding, generating scene graphs of images to provide answers. However, several works [38, 48] suggest that CoT seems to work effectively only in some specific domains. In this work, we propose CIRCoT, which pre-divides the task into multiple subtasks and allows the model to reason these subtasks step-by-step.

3. Methodology

3.1. Preliminary

Given a composed query $Q = \{I_r, T_m\}$, where I_r denotes the reference image and T_m denotes the modification text, and a candidate set $D = \{I_t^1, I_t^2, ..., I_t^{N_D}\}$ consisting of N_D images, the goal of CIR is to identify the k target images from the candidate set D that are most relevant to the query Q, with $k \ll N_D$. ZS-CIR further requires that no training data triplets be used.

Different from the traditional multi-modal retrieval task [9–11, 24], CIR requires the model to retrieve images that both preserve the key features of the reference image and satisfy the modifications described in the modification text. Successfully completing the CIR task requires: (1) correctly understanding the content of the reference image and the modification text, (2) accurately applying the modifications, and (3) an effective score mechanism for the retrieval. Therefore, CIR methods should possess advanced multimodal composed understanding and reasoning capabilities, as well as a comprehensive score mechanism.

3.2. Overall Architecture

The proposed CoTMR is an effective, training-free and interpretable CIR framework based on public pre-trained VLMs. As shown in Figure 2, our CoTMR consists of two steps: reasoning the composed query by LVLM and retrieving the target image by CLIP. In the reasoning process, to enhance the interpretability and reliability of reasoning, we first propose CIRCoT, a novel CoT strategy with predefined subtask divisions tailored for CIR. Moreover, we conduct image-scale and object-scale reasoning, both with

CIRCoT, to obtain the global description and fine-grained details for the target image. For the retrieval process, we design a novel Multi-Grained Scoring (MGS) mechanism that comprehensively considers the characteristics of the above reasoning outputs at different scales via a reward-penalized formulation. We describe the three modules below.

3.3. CIRCoT

CIR requires precise understanding and reasoning of the multi-modal composed query, making it a complex task. To achieve a more accurate and reliable reasoning process, we propose using CoT to facilitate multi-step reasoning of LVLM. However, we find that traditional CoT approaches, such as DDCoT [50], which typically rely on LVLM itself to independently develop problem-solving and task decomposition strategies, tend to work effectively only in a few specific domains [38]. Given the certainty of CIR inputs (reference image and modification text) and the clarity of CIR task (editing the reference image according to the modification text), we thus propose CIRCoT, which decomposes the CIR task into multiple subtasks in advance.

As illustrated in Figure 3, we divide the task of generating the target image caption using the LVLM into four key subtasks: (1) *Image understanding*; (2) *Modification text understanding*; (3) *Modification implementation* and (4) *Target image caption generation*. These four fundamental subtasks structure the overall reasoning process of the LVLM. For each subtask, we adhere to the traditional CoT approach, enabling the model to reason in a step-by-step manner (as represented by italicized prompts in Figure 3). Additionally, we randomly select several reasoning examples from training dataset to further stimulate the LVLM's reasoning capability like [45]. We emphasize that CIRCoT not only capitalizes on the LVLM's reasoning capability but also offers significant interpretability. Users can clearly follow the LVLM's inference process and, if needed, intervene to modify it. More details can be found in Appendix 9.

Combined with the multi-scale reasoning strategy to be introduced, we have two CIRCoT prompts in this work, denoted as P_{Img} and P_{Obj}, which are applied at image scale and object scale, respectively. Here, we take P_{Img} as an example shown in Figure 3, and P_{Obj} follows a similar process (see Appendix 8 for details).

3.4. Mutil-Scale Reasoning

When applying LVLM to address ZS-CIR task, a straightforward approach is to directly infer the target image caption based on the composed query. However, this global description presents several challenges: (1) The dense semantic content in the generated caption overshadows the key objects and attributes that need more attention. (2) In complex scenarios, the model may be confused by irrelevant details in the reference image (e.g. the "human" and "jellyfish" in Figure 2).

CIRCoT — P_{Img}

Your task is to modify the reference image based on the modification instructions and generate the updated image description. The description should be complete and can cover various semantic aspects, such as cardinality, addition, negation, direct addressing, compare & change, comparative, conjunction, spatial relations & background, viewpoint.

To complete the task accurately, please follow these steps:
Understand the Reference Image
1. Identify all the objects, attributes, and their relationships in the image.
2. Especially pay attention to the content related to the modification instructions.
3. ***Please complete this task step by step.***

Analyze the Modification Instructions
1. Break down the modification instructions into separate modification steps.
2. Determine which objects or attributes need to be modified and how.
3. Pay attention to any additions, deletions, or changes to attributes.
4. ***Please complete this task step by step.***

Apply the Modifications###
1. Apply the modifications ***step by step*** to update the content of the reference image.

Generate the Target Image Caption
1. Write a coherent and concise image caption.
2. Ensure the caption accurately reflects all the modifications.
3. The edited caption needs to be as simple as possible.
4. Do not mention the content that will not be present in the target image.

Here are some examples:
Example 1: .... Example 2: ....

Figure 3. **Illustration of CIRCoT in image-scale reasoning** (P_{Img}), which includes four predefined subtasks and allows LVLM to reason step-by-step within each subtask. CIRCoT in object-scale reasoning (P_{Obj}) follows a similar process (see appendix 8 for details).

To alleviate the negative impact of unclear key features and irrelevant information contained in the global caption, in addition to global caption generation, we propose to reason at the object scale to obtain supplementary fine-grained details. As shown in Figure 2, at image-scale reasoning, we utilize the LVLM to reason the editing process and generate the target image caption, which is formulated as:

$$T_{tc} = LVLM(I_r, T_m, P_{Img}) \tag{1}$$

where T_{tc} is the target image caption, $LVLM(\cdot)$ denotes the reasoning process with LVLM and P_{Img} denotes the CIRCoT prompt at image-scale reasoning.

In object-scale reasoning, we let LVLM focus on specific objects and their attributes, specifying the set of objects that should be present in the target image (*"existent objects"*), and those should not be present (*"nonexistent objects"*). This process is expressed as:

$$EO, NEO = LVLM(I_r, T_m, P_{Obj}) \tag{2}$$

Here, $EO = [T_{eo}^i]_{i=0}^{L_e}$ denotes the list of *"existent objects"*, where T_{eo}^i represents the $i-th$ object that should be present and L_e is the total number of these existent objects. Similarly, $NEO = [T_{neo}^i]_{i=0}^{L_u}$ denotes the list of *"nonexistent*

Backbone	Method	Training-free	Shirt		Dress		Tops&Tee		Avg.		
			R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	R_{mean}
ViT-B/32	PALAVRA	✗	21.49	37.05	17.25	35.94	20.55	38.76	19.76	37.25	28.50
	SEARLE	✗	24.44	41.61	18.54	39.51	25.70	46.46	22.89	42.53	32.71
	CIReVL	✓	28.36	47.84	25.29	46.36	31.21	53.85	28.29	49.35	38.82
	LDRE	✓	27.38	46.27	19.97	41.84	27.07	48.78	24.81	45.63	35.22
	CoTMR	✓	**33.42**	**53.93**	**31.09**	**54.54**	**38.40**	**61.14**	**34.30**	**56.54**	**45.42**
ViT-L/14	Pic2Word	✗	26.2	43.6	20.00	40.2	27.9	47.40	24.70	43.70	34.20
	SEARLE	✗	26.89	45.58	20.48	43.13	29.32	49.97	25.56	46.23	35.89
	CIReVL	✓	29.49	47.40	24.79	44.76	31.36	53.65	28.55	48.57	38.56
	LDRE	✓	31.04	51.22	22.93	46.76	31.57	53.64	28.51	50.54	39.52
	LinCIR	✗	28.81	46.61	20.72	41.70	29.07	49.77	26.21	46.03	36.12
	CoTMR	✓	**35.43**	**54.91**	**31.18**	**55.04**	**38.55**	**61.33**	**35.05**	**57.09**	**46.50**
ViT-G/14	CIReVL	✓	33.71	51.42	27.07	49.53	35.80	56.14	32.19	52.36	42.27
	LDRE	✓	35.94	58.58	26.11	51.12	35.42	56.67	32.49	55.46	43.97
	LinCIR	✗	**44.08**	**62.56**	**38.48**	**60.63**	**48.58**	**69.10**	**43.71**	**64.10**	**53.91**
	CoTMR	✓	38.32	62.24	34.51	57.36	41.90	64.30	38.25	61.32	49.78

Table 1. **Comparison with the state-of-the-art methods on the Fashion-IQ dataset.** R_{mean} indicates the average results across all the metrics. The best results are in boldface, while the second-best results are underlined.

objects". T^i_{neo} and L_u represents the $i-th$ nonexistent object and the total number of these objects. P_{Obj} denotes the CIRCoT prompt used at object-scale reasoning.

As shown in Figure 2, the *"existent objects"* further emphasize the key elements that require extra attention ("Two manta rays", "yellow fish" and "ocean"), while the *"nonexistent objects"* mitigate the influence of irrelevant information from the reference image ("human" and "jellyfish").

3.5. Multi-Grained Scoring

After obtaining the above outputs at multiple scales (target image caption, existent objects, and nonexistent objects), we further design this MGS mechanism to comprehensively consider their impact on the final retrieval process.

Specifically, as illustrated in Figure 2, we first compute similarities between the "target image caption" and candidate images using CLIP as the base scores S_{base}:

$$S_{base} = CLIP(T_{tc}, D) \tag{3}$$

where D denotes the set of candidate images, and $CLIP(\cdot,\cdot)$ computes the similarity between text and images in the CLIP space.

At object scale, considering that "existent objects" typically have inherent correlations, they should be treated as a whole to collectively influence the matching result. Therefore, we concatenate these objects into one string and then compute its similarities with candidate images to obtain positive scores S_{pos}:

$$S_{pos} = CLIP(Concat([T^i_{eo}]^{L_e}_{i=0}), D) \tag{4}$$

where $Concat(\cdot)$ denotes the concatenation of strings.

In contrast, "nonexistent objects" usually have no inherent correlation with each other. Thus, we first calculate their similarities with candidate images individually, and then average their similarities to derive the negative scores S_{neg}:

$$S_{neg} = Avg([CLIP(T^i_{neo}, D)]^{L_u}_{i=0}) \tag{5}$$

where $Avg(\cdot)$ denotes the average of scores. This strategy ensures an equal contribution of each undesired object.

Finally, we combine base scores, positive scores, and negative scores using weighted aggregation to obtain the final scores S served as selection criteria:

$$S = S_{base} + \lambda \cdot S_{pos} - \mu \cdot S_{neg} \tag{6}$$

where λ and μ are the weights assigned to the positive score, and negative score, respectively.

4. Experiments

4.1. Implementation Details

For the LVLM, we use Qwen2-VL-72B [44]. For the retrieval model, we experiment with different CLIP variants, including ViT-B/32, ViT-L/14, and ViT-G/14 CLIP from OpenCLIP [15]. The hyperparameter λ and μ are set to 1 and 0.5 for the FashionIQ dataset, 1 and 0.3 for the CIRR dataset, and 0.5 and 0.3 for the CIRCO dataset, respectively. The entire model is implemented using PyTorch [29] on 8 NVIDIA A800 GPUs. We deployed the model using vLLM [19] to achieve real-time response capabilities, with each inference taking only 1s.

4.2. Datasets and Baselines

We make performance evaluations on four CIR benchmarks, including a fashion-domain dataset **Fashion-IQ** [46], as well as three open-domain datasets **CIRR** [25], **CIRCO** [3] and **GeneCIS** [42]. For Fashion-IQ and GenCIS, we adopt Recall@K as the evaluation metric. For CIRR, we additionally report Recall_{subset}@K. For CIRCO, since there are multiple positives, we use the mean average precision@k (mAP@k) as the metric.

For baseline models, we select several textual inversion methods (PALAVRA [6], Pic2Word [36], SEARLE [3] and LinCIR [13]), and some LLM-based, training-free methods (CIReVL [17] and LDRE [47]). Among them, CIReVL serves as our most direct baseline.

Benchmark			CIRCO				CIRR							
	Metric		mAP@k				Recall@k				Recall$_{sub}$@k			Avg.
Backbone	Method	Training-free	k=5	k=10	k=25	k=50	k=1	k=5	k=10	k=50	k=1	k=2	k=3	
ViT-B/32	PALAVRA	✗	4.61	5.32	6.33	6.80	16.62	43.49	58.51	83.95	41.61	65.30	80.94	42.55
	SEARLE	✗	9.35	9.94	11.13	11.84	24.00	53.42	66.82	89.78	54.89	76.60	88.19	54.15
	CIReVL	✓	14.94	15.42	17.00	17.82	23.94	52.51	66.00	86.95	60.17	80.05	90.19	56.34
	LDRE	✓	17.96	18.32	20.21	21.11	25.69	55.13	69.04	89.90	60.53	80.65	90.70	57.83
	CoTMR	✓	**22.23**	**22.78**	**24.68**	**25.74**	**31.50**	**60.80**	**73.04**	**91.06**	**66.61**	**84.50**	**92.55**	**63.71**
ViT-L/14	Captioning	✗	1.65	1.96	2.42	2.71	4.05	15.88	25.69	49.21	20.87	40.60	60.89	18.37
	Pic2Word	✗	8.72	9.51	10.64	11.29	23.90	51.70	65.30	87.80	-	-	-	-
	SEARLE	✗	11.68	12.73	14.33	15.12	24.24	52.48	66.29	88.84	53.76	75.01	88.19	53.12
	CIReVL	✓	18.57	19.01	20.89	21.80	24.55	52.31	64.92	86.34	59.54	79.88	89.69	55.92
	LDRE	✓	23.35	24.03	26.44	27.50	26.53	55.57	67.54	88.50	60.43	80.31	89.90	58.00
	LinCIR	✗	12.62	13.40	14.81	15.69	25.08	53.63	67.30	88.72	56.36	76.96	88.57	54.99
	CoTMR	✓	**27.61**	**28.22**	**30.61**	**31.70**	**35.02**	**64.75**	**76.18**	**92.51**	**69.39**	**85.75**	**93.33**	**67.07**
ViT-G/14	CIReVL	✓	26.77	27.59	29.96	31.03	34.65	64.29	75.06	91.66	67.95	84.87	93.21	66.12
	LDRE	✓	31.12	32.24	34.95	36.03	36.15	66.39	77.25	93.95	68.82	85.66	93.76	67.60
	LinCIR	✗	19.71	20.79	22.99	24.00	35.34	65.08	76.28	93.22	63.73	82.62	92.12	64.41
	CoTMR	✓	**32.23**	**32.72**	**35.60**	**36.83**	**36.36**	**67.52**	**77.82**	**93.99**	**71.19**	**86.34**	**93.87**	**69.36**

Table 2. **Comparison with the state-of-the-art methods on CIRCO and CIRR test sets.** Avg. indicates the average results of Recall@5 and Recall$_{sub}$@1. The best results are in boldface, while the second-best results are underlined.

Backbone	Method	Training-free	Avg. R@1	Avg. R@2	Avg. R@3
ViT-B/32	SEARLE	✗	14.4	25.3	35.4
	CIReVL	✓	15.8	26.8	36.8
	CoTMR	✓	**17.9**	**29.6**	**39.5**
ViT-L/14	SEARLE	✗	14.4	25.3	34.9
	CIReVL	✓	15.9	27.1	36.3
	LinCIR	✗	12.2	22.8	32.4
	CoTMR	✓	**17.6**	**29.6**	**39.8**
ViT-G/14	CIReVL	✓	17.4	29.8	39.5
	LinCIR	✗	13.7	24.6	33.5
	CoTMR	✓	**19.1**	**31.4**	**41.0**

Table 3. **Average results on GeneCIS test set.** The full table is in Appendix 11. The best results are in boldface.

4.3. Comparison with Bselines

Fashion-IQ. Table 1 presents the comparative results on the Fashion-IQ dataset. We have the following observations: (1) Compared to pseudo-word-based methods, CoTMR achieves impressive performance across multiple metrics even without any training. This indicates that generating target captions with LVLM provides semantic information that is more suitable for CLIP's text encoding. (2) Compared to LDRE, which also uses LLMs to model the modified images, our approach achieves significant improvements. This is attributed to CoTMR's superior preservation of image semantics, more refined reasoning process, and finer-grained feature recognition. (3) Across all metrics with different CLIP, CoTMR consistently outperforms most baseline methods. Using ViT-B/32 as an example, our method relatively outperforms LDRE by 9.49% in average R@10. These results strongly support CoTMR's effectiveness. (4) We observe that CoTMR significantly outperforms LinCIR when utilizing ViT-L/14, whereas LinCIR exhibits better results with ViT-G/14 CLIP. Notably, as detailed in Appendix 12, we find that CoTMR can achieve even better performance through integration with LinCIR.

CIRR. When applied to the open-domain dataset CIRR, CoTMR still shows compelling results, summarized in the right section of Table 2. We have the following observations: (1) Notably, the CIRR dataset is quite noisy, with minimal correlation between the reference image and the target image. Therefore, CoTMR's ability to capture rich information from reference images also means it may receive more distracting information. Despite this challenge, CoTMR consistently outperforms all baseline metrics. These findings highlight the robustness of CoTMR, demonstrating its ability to deliver significant results even in the presence of noisy data. (2) CIRR also provides another evaluation, where the task is to retrieve the correct image from six curated samples. In this evaluation, our approach also significantly surpasses previous methods (Our method outperforms LDRE by 6.08% in Recall$_{sub}$@1 when using ViT-B/32 CLIP). This shows the versatility of our method, enabling it to perform well across different contexts. (3) Using ViT-L/14 as an example, compared to our most direct baseline, CIReVL, our method achieves significant improvement. This further shows the effectiveness of our proposed modules, such as CIRCoT and multi-scale reasoning.

CIRCO. In the left section of Table 2, we present the competitive results of CoTMR. Based on the results, we make the following observations: (1) Since CIRCO uses mAP as the evaluation metric, the incorrect selection of negative samples has a significant impact on the results. CoTMR, by introducing a negative scoring mechanism, effectively eliminates incorrect samples, achieving optimal performance across multiple metrics. (2) Thanks to multi-scale reasoning and the pre-defined subtask decomposition, CoTMR shows substantial improvements over LLM-based methods such as CIReV and LDRE across several metrics. Using ViT-B/32 as an example, our method relatively outperforms LDRE

Benchmark	FashionIQ-Avg		CIRCO			
Metric	Recall@k		mAP@k			
Method	k=10	k=50	k=5	k=10	k=25	k=50
A. Multi-Grained Scoring						
A.1 Base	33.99	56.34	26.40	27.98	30.35	31.43
A.2 Pos + Neg	30.50	52.65	14.92	16.49	18.33	19.12
A.3 Base + Pos	35.62	58.39	27.54	29.59	32.24	33.26
A.4 Base + Neg	34.42	56.95	27.28	28.30	30.63	31.82
A.5 Full	**37.72**	**60.92**	**28.87**	**30.61**	**33.30**	**34.32**
B. Chain of Thought						
B.1 No COT	31.03	51.01	20.07	21.12	23.43	24.44
B.2 DDCOT	29.21	48.25	17.41	18.84	21.25	22.17
B.3 w/o task-1	32.62	52.89	23.15	24.46	26.63	27.52
B.4 w/o task-2	32.32	53.02	21.62	22.72	25.11	26.17
B.5 w/o task-3	33.02	52.86	21.58	22.14	24.85	25.87
B.6 ZS_CIRCoT	33.41	53.50	23.54	24.88	27.42	28.40
B.7 CIRCoT	**33.99**	**56.34**	**26.40**	**27.98**	**30.35**	**31.43**
C. Scoring for Objects						
C.1 Pos + mean	35.16	56.48	28.57	30.46	33.00	34.14
C.2 Neg + concat	37.25	59.29	28.71	30.46	33.04	34.06
C.3 Normal	**37.72**	**60.92**	**28.87**	**30.61**	**33.30**	**34.32**
D. Model Variants						
D.1 LLaVa-Onevision	32.51	53.48	17.76	18.95	21.05	22.15
D.2 GPT-4o	37.27	60.29	28.40	30.22	32.73	33.72
D.3 Qwen2-VL-2B	20.27	36.68	4.90	5.50	6.15	6.34
D.4 Qwen2-VL-7B	33.35	54.02	16.10	17.05	19.45	19.41
D.5 Qwen2-VL-72B	**37.72**	**60.92**	**28.87**	**30.61**	**33.30**	**34.32**

Table 4. **Ablation study results for the proposed components on Fashion-IQ val set and CIRCO val sets.** All experiments are performed with the ViT-G/14 CLIP model.

by 4.27% and CIReVL by 7.27% in mAP@5. This further demonstrates the effectiveness of CoTMR for ZS-CIR.

GeneCIS. In Table 3, we present the average R@K results for four subtasks on GeneCIS. This benchmark is particularly challenging due to its single-word modification text design, requiring precise object and attribute identification in reference images. As discussed in Appendix 11, our CoTMR shows compelling results especially in "Focus" task and "Object" tasks. We attribute that to CoTMR's ability to simultaneously perceive the image and text and object-level reasoning, which can effectively help CoTMR focus on specific objects/attributes.

5. Ablation Study

Effects of Multi-Grained Scoring Mechanism. In Table 4 A, we investigate the impact of three scores. Our observations are as follows: (1) Compared to using only the base score S_{base} (**A.1**), only rely on the object-level reasoning output (**A.2** S_{pos} and S_{neg}) leads to a significant decline of CoTMR's performance. This indicates that the logical relationships between objects are still essential for effective retrieval. (2) When S_{base} is combined with either S_{pos} (**A.3**) or S_{neg} (**A.4**), the model's performance improves. This further confirms the effectiveness of object-level reasoning. (3)

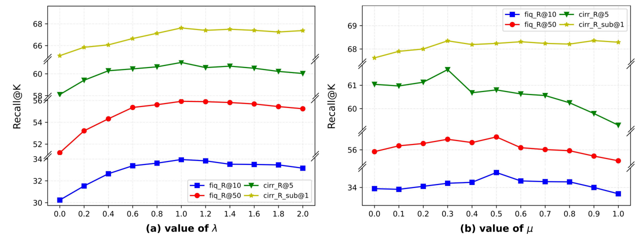


Figure 4. **Ablation study on the value of λ and μ on Fashion-IQ val set and CIRR val set.** All experiments are performed with the ViT-B/32 CLIP model.

When both S_{pos} and S_{neg} are used with S_{base} (**A.5**), the performance significantly improves compared to using only one of them. This suggests that S_{pos} and S_{neg} have complementary roles in the retrieval process.

Effects of CIRCoT. In Table 4 B, taking image-level reasoning as an example, we first compare four different approaches for subtask decomposition: without using CoT (**B.1**), using DDCoT [50], CIRCoT without examples (**B.6**) and our proposed CIRCoT (**B.7**). The results show that DDCoT caused slight performance degradation compared to not using CoT, indicating that the reasoning process constructed by the model itself may cause confusion in the CIR task. In contrast, when using our CIRCoT, the model's performance improves significantly, which shows the effectiveness of the predefined subtasks. Moreover, adding step-wise reasoning examples also provides assistance in helping the model understand task requirements.

Building upon **B.6**, we conducted ablation studies by removing Tasks 1, 2, and 3 respectively (**B.3-5**) to investigate their individual impact on performance. We observed that Task-1 "Reference Image Observation" had minimal impact on the CIRCO dataset but significantly affected FashionIQ performance. Meanwhile, both Task-2 "Modification Text Analysis" and Task-3 "Modification Execution" showed substantial influence on both datasets' performance.

Scoring Methods for Objects: In Table 4 C, we conduct experiment on two different scoring strategies for S_{pos} and S_{neg}: (**C.1**) computing individual scores for each object in EO and then averaging them, and (**C.2**) concatenating all objects in NEO for scoring. Our observations indicate that scoring objects in EO individually results in a significant performance decline compared to the full model. We assume that because of the intrinsic correlations among "existent objects"—treating them separately overlooks these dependencies. On the other hand, concatenating the uncorrelated objects in NEO may introduce potential bias that causes the decline in performance, as the overall score may be disproportionately affected by certain individual objects.

Impact of LVLM Architectures and Scales: We conduct experiments in Table 4 D to investigate the impact of different LVLMs (**D.1** LLaVA-OneVision-7B and **D.2** GPT-4o) and various scales of Qwen2-VL (**D.3-5**) on the

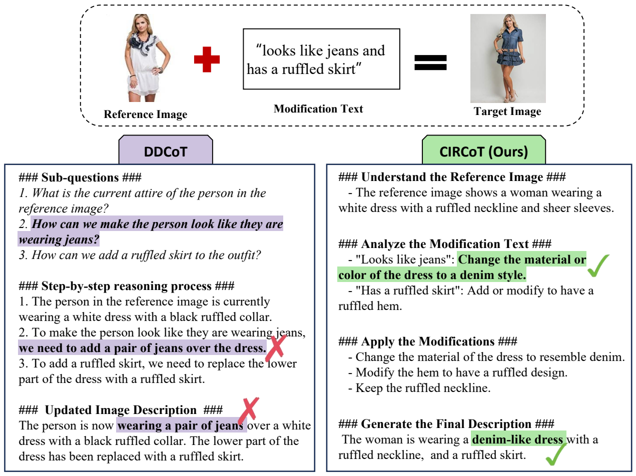


Figure 5. **Comparison between DDCoT and CIRCoT prompting strategies from Fashion-IQ val set.**

results. Several key observations emerge: **(1)** Different LVLMs demonstrate substantial variations in performance, with GPT-4o and Qwen2VL-72B exhibiting notably superior capabilities compared to other models. **(2)** Models of similar scale achieve comparable performance, as evidenced by the similar results obtained from the 7B-parameter versions of LLaVA-OneVision and Qwen2-VL. **(3)** Model scale significantly influences performance, with Qwen2-VL showing dramatic improvements as its parameter scale increases from 2B to 72B.

Impact of hyperparameter λ and μ: To analyze the sensitivity of the hyperparameters in CoTMR, we conduct controlled experiments as shown in Figure 4. First, we set μ to 0 to better demonstrate the effect of λ. As shown in Figure 4 (a), when the value of λ increases from 0, all four metrics show a rapid rise, stabilizing and slightly declining when λ reaches 1. Next, we fix λ at 1 to explore the effect of μ. As shown in Figure 4 (b), for the Fashion-IQ dataset, the impact of μ is relatively mild, with the metrics reaching their peak at $\mu = 0.5$. However, for the CIRR dataset, due to the higher noise in the data, increasing μ too much leads to a significant drop in the R@5 metric. Therefore, for CIRR, the best average performance is achieved when $\mu = 0.3$.

6. Qualitative Results

Comparison of DDCoT and CIRCoT In Figure 5, we compare the reasoning processes of DDCoT and CIRCoT. It can be observed that when using DDCoT, the definition of subtasks is entirely left to the LVLM, which sometimes results in confusing subtasks, such as “*How can we make the person look like they are wearing jeans?*” in Figure 5. Such subtasks can mislead the LVLM into providing incorrect answers, as seen in Figure 5, where the LVLM decides on “*a pair of jeans*” to correspond to the modification text’s requirement of “*looks like jeans*”. This example highlights

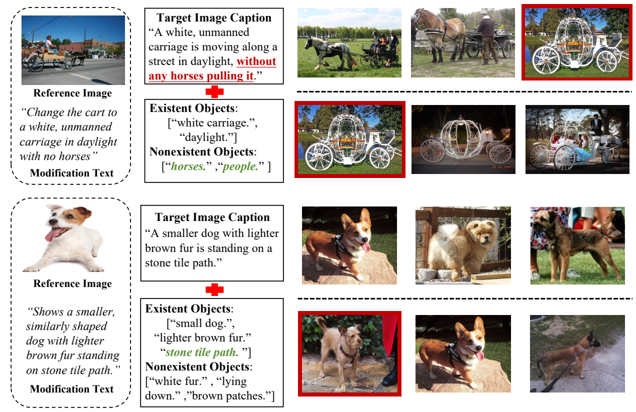


Figure 6. **Successful retrieval examples with muti-scale reasoning from CIRR val set.** The ground-truth image is highlighted with the red box. Red underlined text indicates distracting information that causes mistake retrieval, while green italicized text represents key objects that help in correct retrieval.

the critical importance of subtask definition; inappropriate subtasks can lead to erroneous logical reasoning and cause confusion. In contrast, our proposed CIRCoT, with its four predefined subtasks, offers a stable and correct reasoning process, leading to more accurate outputs.

Examples of successful retrieval. Figure 6 visualizes cases where the combination of image-scale and object-scale reasoning leads to successful retrievals. In the first example, the target image caption includes “*without any horses*”, which, while meeting user requirements, is detrimental to CLIP retrieval. However, the undesired objects (“*horses*” and “*people*”) identified through object-scale reasoning eliminates this interference, successfully retrieving the target image. In the second example, we observe that the retrieval results initially overlooked the attribute “*stone tile path*”. Object-scale reasoning, however, highlighted this attribute with existent objects, leading to the successful retrieval of the target image. These examples clearly demonstrate that object-scale reasoning can supplement emphasis and eliminate distracting information. More examples can be found in Appendix 14.

7. Conclusion

In this work, we propose CoTMR, an effective, training-free, and interpretable method for ZS-CIR. It provides a unified understanding and reasoning framework for composed queries, utilizing a step-by-step process guided by CIRCoT. To incorporate fine-grained details, multi-scale reasoning (alongside a novel scoring mechanism) is devised for multi-grained generation and evaluation. Extensive experiments demonstrate the effectiveness of our CoTMR. Moreover, thanks to CIRCoT, our method also offers appealing interpretability for user intervention.

Acknowledgements

This work is partially supported by National Natural Science Foundation of China (62376274, 62437002) and Beijing Natural Science Foundation (L233008).

References

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *IEEE/CVF International Conference on Computer Vision*, pages 8948–8957, 2019. 2

[2] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Conditioned and composed image retrieval combining and partially fine-tuning clip-based features. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4959–4968, 2022. 2

[3] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *IEEE/CVF International Conference on Computer Vision*, pages 15338–15347, 2023. 1, 2, 5

[4] Simion-Vlad Bogolin, Ioana Croitoru, Hailin Jin, Yang Liu, and Samuel Albanie. Cross modal retrieval with querybank normalisation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5194–5205, 2022. 2

[5] Yanbei Chen, Shaogang Gong, and Loris Bazzani. Image search with text feedback by visiolinguistic attention learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3001–3011, 2020. 2

[6] Niv Cohen, Rinon Gal, Eli A Meirom, Gal Chechik, and Yuval Atzmon. "this is my unicorn, fluffy": Personalizing frozen vision-language representations. In *European conference on computer vision*, pages 558–577. Springer, 2022. 5

[7] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[8] Zhiwu Lu Dong Jing, Xiaolong He et al. Fineclip: Self-distilled region-based clip for better fine-grained understanding. *Advances in Neural Information Processing Systems*, 2024. 2

[9] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *Advances in Neural Information Processing Systems*, 26:2121–2129, 2013. 1, 3

[10] Dehong Gao, Linbo Jin, Ben Chen, Minghui Qiu, Peng Li, Yi Wei, Yi Hu, and Hao Wang. Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval. In *SIGIR*, pages 2251–2260, 2020.

[11] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *European Conference on Computer Vision*, pages 241–257, 2016. 1, 3

[12] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017. 2

[13] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, Yoohoon Kang, and Sangdoo Yun. Language-only training of zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13225–13234, 2024. 2, 5, 3

[14] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6700–6709, 2019. 2

[15] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, et al. Openclip. 2021. 5

[16] Shyamgopal Karthik, Massimiliano Mancini, and Zeynep Akata. Kg-sp: Knowledge guided simple primitives for open world compositional zero-shot learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9336–9345, 2022. 2

[17] Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. Vision-by-language for training-free compositional image retrieval. *arXiv preprint arXiv:2310.09291*, 2023. 1, 2, 5

[18] 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

[19] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023. 5

[20] Seungmin Lee, Dongwan Kim, and Bohyung Han. Cosmo: Content-style modulation for image retrieval with text feedback. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 802–812, 2021. 2

[21] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900, 2022. 2

[22] 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. Springer, 2014. 2

[23] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 2

[24] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1096–1104, 2016. 1, 3

[25] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *IEEE/CVF International Conference on Computer Vision*, pages 2125–2134, 2021. 5

[26] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, pages 947–952. IEEE, 2019. 2

[27] Ishan Misra, Abhinav Gupta, and Martial Hebert. From red wine to red tomato: Composition with context. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1792–1801, 2017. 2

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

[29] 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. *Advances in Neural Information Processing Systems*, 32, 2019. 5

[30] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *IEEE/CVF International Conference on Computer Vision*, pages 15691–15701, 2023. 2

[31] Shengsheng Qian, Dizhan Xue, Quan Fang, and Changsheng Xu. Adaptive label-aware graph convolutional networks for cross-modal retrieval. *IEEE Transactions on Multimedia*, 24: 3520–3532, 2021. 2

[32] Shengsheng Qian, Dizhan Xue, Quan Fang, and Changsheng Xu. Integrating multi-label contrastive learning with dual adversarial graph neural networks for cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4794–4811, 2022. 2

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 2

[34] Karsten Roth, Oriol Vinyals, and Zeynep Akata. Integrating language guidance into vision-based deep metric learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16177–16189, 2022. 2

[35] Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633*, 2021. 3

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

[37] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *European Conference on Computer Vision*, pages 742–758. Springer, 2020. 2

[38] Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. *arXiv preprint arXiv:2409.12183*, 2024. 3, 4

[39] 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

[40] 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 *AAAI Conference on Artificial Intelligence*, pages 5180–5188, 2024. 1

[41] Vishaal Udandarao, Ankush Gupta, and Samuel Albanie. Sus-x: Training-free name-only transfer of vision-language models. In *IEEE/CVF International Conference on Computer Vision*, pages 2725–2736, 2023. 2

[42] Sagar Vaze, Nicolas Carion, and Ishan Misra. Genecis: A benchmark for general conditional image similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6862–6872, 2023. 5

[43] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6439–6448, 2019. 2

[44] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2, 5

[45] 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. 2, 3, 4

[46] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11307–11317, 2021. 5

[47] 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 *SIGIR*, pages 80–90, 2024. 1, 2, 5

[48] Zihan Yu, Liang He, Zhen Wu, Xinyu Dai, and Jiajun Chen. Towards better chain-of-thought prompting strategies: A survey. *arXiv preprint arXiv:2310.04959*, 2023. 3

[49] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023. 3

[50] 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. 2, 3, 4, 7