Generating a Paracosm for Training-Free Zero-Shot Composed Image Retrieval

Tong Wang[1], Yunhan Zhao[2], and Shu Kong[1,3,✉]

[1] University of Macau, Macau 999078, China
[2] Uiversity of California at Irvine, Irvine, CA 92697, USA
[3] Institute of Collaborative Innovation, University of Macau, Macau 999078, China
website and code: `https://github.com/leowangtong/Paracosm/`

Abstract. Composed Image Retrieval (CIR), a task towards personalizing visual intelligence, aims to retrieve a target image from a database based on users' multimodal query, which contains a reference image and a modification text. The text specifies how to alter the reference image to form a "mental image", based on which CIR should find the target image in the database. The fundamental challenge of CIR is that this "mental image" is not physically available and is only implicitly defined by the query. The contemporary literature pursues zero-shot methods and uses a Large Multimodal Model (LMM) to generate a textual description for a given multimodal query, and then employs a Vision-Language Model (VLM) for textual-visual matching to search for the target image. In contrast, we address CIR from first principles by directly generating the "mental image" for more accurate matching. Particularly, we prompt an LMM to generate a "mental image" for a given multimodal query and propose to use this "mental image" to search for the target image. As the "mental image" has a synthetic-to-real domain gap with real images, we also generate a synthetic counterpart for each real image in the database to facilitate matching. In this sense, our method uses LMM to construct a "paracosm", where it matches the multimodal query and database images. Hence, we call this method Paracosm. Notably, Paracosm is a training-free zero-shot CIR method. It significantly outperforms existing zero-shot methods on challenging benchmarks, achieving state-of-the-art performance for zero-shot CIR.

Keywords: Training-Free Zero-Shot Composed Image Retrieval · Foundation Models · Synthetic-to-Real Domain Gap

1 Introduction

Composed Image Retrieval (CIR) formulates new scenarios in search [3, 12, 25] and e-commerce [50, 67], where it retrieves a target image from a database based on a user-provided multimodal query, which consists of a reference image and modification text [64]. The modification text describes how to alter the reference

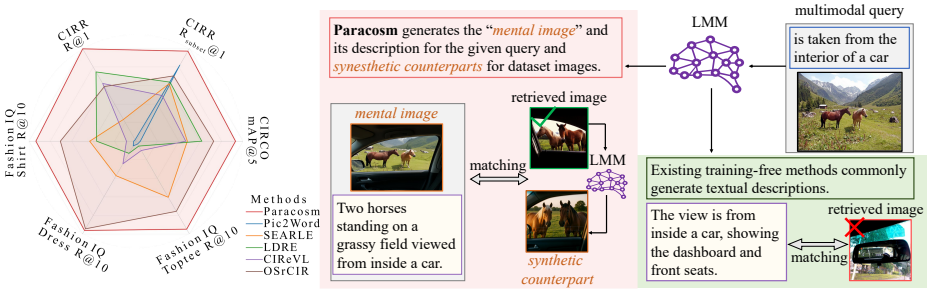

Fig. 1: Overview of our method and benchmarking results. Unlike existing training-free methods [24, 57, 70] generating descriptions for multimodal queries, which use an LMM to generate descriptions for multimodal queries, we use it to generate "mental images" for multimodal queries and synthetic counterparts of database images. Matching them effectively mitigates synthetic-to-real domain gaps and boosts CIR performance. Our final training-free zero-shot method *Paracosm* (Fig. 2) significantly outperforms existing zero-shot CIR methods, as summarized in the radar chart on standard benchmarks. Detailed results are provided in Section E.

image for which CIR algorithms should find the best-matching image from the database [2, 35, 67]. It exemplifies an effort for personalized visual intelligence.

Status Quo. CIR was addressed through supervised learning [7, 11, 15, 27, 36] over annotated triplets *<reference image, modification text, target image>*. As curating large-scale triplet data is prohibitively expensive, recent works have developed zero-shot CIR (ZS-CIR) approaches [6, 43, 57, 70]. Notably, "zero-shot" does not necessarily mean non-learned methods [39, 48, 49, 76] but emphasizes not directly training models on target dataset. Rather, they can train models in an indirect way [18, 21, 55, 56, 65, 72]. For example, many methods [6, 43] train textual inversion networks on image-text pairs to map images to text tokens for cross-modality matching; some [32, 65] synthesize a pseudo image for the multimodal query using pretrained image generative models [21, 41, 44, 77] to assist with cross-modality matching. To contrast training-based ZS-CIR approaches, training-free methods [24, 57, 70] propose to exploit Large Multimodal Models (LMMs) [10, 38, 63] to generate a description for the multimodal query and use it for text-to-image (T2I) matching for CIR.

Insights. The fundamental challenge of CIR is that the multimodal query only implicitly defines a "mental image" that does not physically exist to be used for retrieving the corresponding target image. We aspire to address this challenge from first principles by generating a "mental image" for a given query to enable more accurate retrieval (Fig. 1). While existing works have exploited LMMs to generate textual descriptions, we leverage LMMs for image generation based on multimodal queries [66]. Yet, as the generated "mental image" has synthetic-to-real domain gaps compared to the real images in the dataset, we propose generating a synthetic counterpart for each dataset image, which is used to facilitate matching. As our method essentially uses LMMs to define a synthetic or virtual space, like a "paracosm", we name our method **Paracosm**. Notably, *Paracosm is a training-free zero-shot CIR method.* To validate Paracosm, we follow the established experimental protocols that use pretrained Vision-Language

Models (VLMs) [22, 40] for matching. Extensive experiments on challenging benchmarks demonstrate that Paracosm significantly outperforms existing ZS-CIR methods (ref. a summary in Fig. 1).

Contributions. We make three major contributions:

- We solve CIR from first principles with the training-free method Paracosm. It generates "mental images" for multimodal queries to facilitate matching with database images.
- We mitigate the synthetic-to-real domain gaps of mental images by generating synthetic counterparts of database images. Matching them together boosts CIR performance.
- On standard benchmarks, we demonstrate that Paracosm significantly outperforms existing ZS-CIR approaches, achieving state-of-the-art performance.

2 Related Work

Composed Image Retrieval (CIR) extends traditional retrieval tasks [45,46,78], such as text-to-image retrieval and image-to-image retrieval, by allowing users to use multimodal queries in retrieval [64]. CIR was initially approached by supervised learning methods [7, 11, 15, 27, 36], i.e., training models over annotated data triplets *<reference image, modification text, target image>*. These methods [7, 20, 61, 68, 73, 74] propose to train a Transformer or an MLP atop pretrained backbones over the annotated data triplets, intending to fuse the reference image and the modification text in a feature space and to allow matching with database images. Some methods [28, 36, 37] finetune a pretrained Vision-Language Model (VLM) [29, 40] for better matching between multimodal queries and database images. However, as supervised learning methods require costly curation of triplet data, recent methods explore zero-shot CIR (ZS-CIR) [6, 18, 24, 32, 43, 57, 65, 70]. We explore *training-free* ZS-CIR and introduce a rather simple method that rivals some recent supervised methods.

Zero-Shot Composed Image Retrieval (ZS-CIR) aims to solve CIR without directly training on annotated triplet data [43]. It does not mean developing training-free methods but allows training in an indirect manner [6,18,43,55,56,72]. For example, many methods train a textual inversion network on existing image datasets [42,47], mapping the reference image to a pseudo-word token. This token, combined with the modification text, is then used in cross-modality matching with database images. A couple of recent methods [32, 65] use pretrained generative models [41, 44, 77] to synthesize a synthetic image (similar to our "mental image"), termed pseudo-target images, for a given multimodal query. They use these images to augment the textual feature (i.e., output by the textual inversion network) to compute similarity scores with database images. Importantly, training-free ZS-CIR methods [14, 24, 52, 57, 70] have emerged that leverage an LMM to generate a description for a given multimodal query. This avoids training a separate textual inversion network. Notably, existing ZS-CIR methods have not considered generating either descriptions or synthetic images for database images. In contrast, our work does so, motivated to address ZS-CIR from first

principles. Specifically, we propose to generate the "mental image" for the query and synthetic counterparts of real database images, mitigating synthetic-to-real domain gaps for better performance.

Foundation Models (FMs), pretrained on various formats of web-scale data across multiple modalities, demonstrate unprecedented performance on downstream tasks in a zero-shot manner. Consequently, FMs are extensively utilized in the CIR literature [8, 9, 31, 34, 69]. First, Vision-Language Models (VLMs), pretrained on web-scale image-text pairs [22,40], are commonly employed in CIR to extract features from modification texts and images, enabling cross-modality matching [7, 24, 32, 39, 53, 57]. Second, Large Language Models (LLMs), pretrained on massive text corpora [4, 10, 54, 62], are utilized in CIR to generate a target image description by incorporating the reference image caption and the modification text [32, 70]. Third, Large Multimodal Models (LMMs) [5, 38, 58, 66], which are pretrained on web-scale multimodal data and typically larger than VLMs in parameter size, are used in CIR to generate image captions [24, 32, 70] for multimodal queries. Earlier CIR works leverage LMMs to create triplet data for supervised training [34, 55, 72]. In contrast, most recent training-free ZS-CIR methods [24,57,70] employ LMMs to generate textual descriptions for multimodal queries and utilize a VLM to match these descriptions with database images, transforming the CIR task into a text-to-image retrieval problem. Building upon this foundation, we extensively exploit LMMs not only to generate "mental images" for multimodal queries but also to create synthetic counterparts of real database images. Matching them effectively mitigates synthetic-to-real domain gaps, thereby significantly enhancing CIR performance.

3 Methodology

We begin by defining the ZS-CIR problem and its development protocol. We then introduce our *training-free* method, Paracosm. Despite its conceptual simplicity, Paracosm achieves state-of-the-art performance among zero-shot methods. Fig. 2 illustrates the complete pipeline.

3.1 Preliminaries of Zero-Shot CIR

Task Definition. Let $(\mathbf{I}_{ref}, \mathbf{t}_{mod})$ represent the reference image and modification text of a given multimodal query. $\mathbf{t}_{mod}$ describes how to alter $\mathbf{I}_{ref}$ to match what the user wants to search for, i.e., a target image $\mathbf{I}_{target}$ from a given database consisting of n images $\{\mathbf{I}^1, \mathbf{I}^2, ..., \mathbf{I}^n\}$. ZS-CIR aims to develop methods to retrieve target images for multimodal queries without directly training on annotated triplets $(\mathbf{I}_{ref}, \mathbf{t}_{mod}, \mathbf{I}_{target})$.

Methodology Development. While ZS-CIR eschews a training set of annotated triplet data, it still allows exploiting data available in the open world and pretrained FMs therein. Following this protocol, existing ZS-CIR methods leverage LMMs in different ways, e.g., using a VLM for cross-modality matching [6,

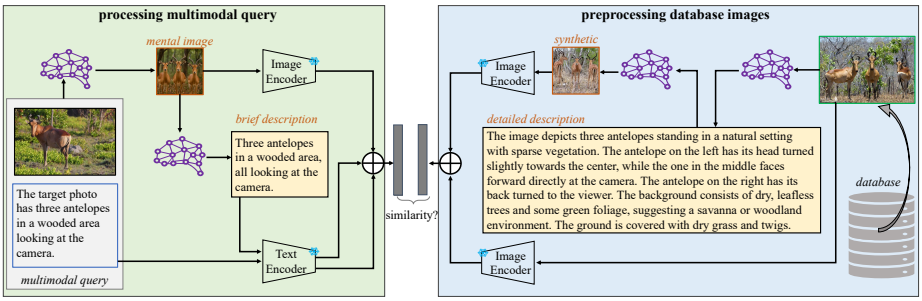

Fig. 2: Flowchart of our training-free zero-shot CIR method Paracosm. Given a multimodal query that consists of a reference image and a modification text, we feed it to an LMM to generate a "mental image". We further generate a brief description for it. Both the "mental image" and description, as well as the modification text, are used as feature representations for the query. As the "mental image" is synthetic, we mitigate synthetic-to-real domain gaps by generating synthetic counterparts of database images. To do so, we use the LMM to generate *detailed* descriptions, which are used as prompts for image generation. For a database image, we use both itself (i.e., the real photo) and its synthetic counterpart as representation for retrieval. In sum, our method uses LMMs to create a virtual **paracosm**, where it matches the query and database images.

18, 43] and an LMM for generating descriptions [24, 57, 70]. By exploiting open-world data [42, 47], some methods train necessary models for fusing $\mathbf{I}_{ref}$ and $\mathbf{t}_{mod}$ into features [6, 18, 43], or for generating pseudo images for the multimodal query [32, 65]. In this work, we aspire to develop a *training-free* method by extensively leveraging LMMs, without training any new models. Next, we present our method Paracosm in detail.

3.2 The Proposed Method: Paracosm

Paracosm is a rather simple method that processes multimodal queries and database images using LMMs. Below, we describe how it processes them and matches them for retrieval.

Processing Multimodal Query. For a multimodal query consisting of a reference image and modification text, Paracosm first generates the "mental image" $\mathbf{I}_{mental}$ using an LMM [66]. Based on this mental image $\mathbf{I}_{mental}$, we further generate a brief textual description $\mathbf{t}_{query}$, which can be thought of as the description of the target image $\mathbf{I}_{target}$. For this step, we use a prompt that instructs the LMM [60] to focus solely on visual content while minimizing aesthetic details, producing single-sentence descriptions. Fig. 3 illustrates the prompt templates used for processing a multimodal query. Since different datasets provide modification texts in varying formats, we adapt the prompt template structure accordingly to ensure proper input formatting for the LMM. Given a multimodal query, Paracosm uses both the generated mental image and short description for retrieval. This is different from some recent methods [24, 57, 70], which exclusively use generated descriptions for retrieval. For example, some of such methods first generate a description for the reference image and then revise

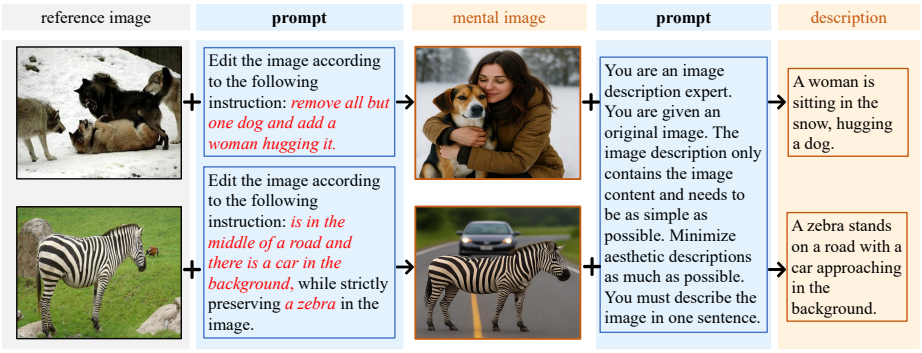

Fig. 3: Illustration of processing a multimodal query. Two random examples from CIRR and CIRCO datasets are displayed in the two rows, respectively. For a multimodal query consisting of a reference image and *modification texts*, we employ a prompt incorporating the latter to edit the former, generating a mental image representing this query. To adapt to the varying modification text across benchmarks, where a modification text can contain a relative caption and a shared concept (ref. CIRCO in the second row), we design the prompt to incorporate both. Further, for the mental image, we prompt an LMM to generate a a concise, single-sentence description, exclusively focusing on its visual content while minimizing aesthetic details. We use both mental image and short description to retrieve the target image from the database.

the description with the modification text using an LLM. Fig. 4 demonstrates that solely relying on generated descriptions misses crucial visual information, whereas our mental images retain rich information, facilitating CIR.

Pre-Processing Database Images. As the generated mental images are synthetic and have synthetic-to-real domain gaps, directly matching them with real images from the database is suboptimal. To mitigate this gap, we create synthetic counterparts for database images and use such counterparts to assist matching [17,19,75]. Specifically, Paracosm first leverages an LMM [60] to generate a *detailed* description for each database image. We employ a description prompt that emphasizes capturing all visible objects, attributes, spatial relationships, and fine-grained visual elements to ensure maximum fidelity. Fig. 5 illustrates this prompt template. The detailed description then serves as a prompt for a text-to-image (T2I) generation model [66], generating a synthetic counterpart $\mathbf{I}^i_{syn}$ for each database image $\mathbf{I}^i$. Both real database images and synthetic counterparts are used jointly for matching multimodal queries.

Matching for Retrieval. After transforming both multimodal queries and database images into the virtual paracosm, we construct features using a pre-trained VLM, which consists of a visual encoder $V(\cdot)$ and a text encoder $T(\cdot)$. Specifically, the query feature $\mathbf{q}$ and the i^{th} database image feature $\boldsymbol{\phi}^i$ are computed as follows:

$$\begin{aligned} \mathbf{q} &= \lambda(V(\mathbf{I}_{mental}) + T(\mathbf{t}_{query})) + (1-\lambda)T(\mathbf{t}_{mod}) \\ \boldsymbol{\phi}^i &= V(\mathbf{I}^i) + V(\mathbf{I}^i_{syn}) \end{aligned} \tag{1}$$

Fig. 4: Comparison of qualitative results between OSrCIR [57] and our Paracosm. We show four examples from the CIRCO dataset [2] in the first column, followed by generated descriptions and top-4 retrievals by OSrCIR, and the mental images and top-4 retrievals by Paracosm. For each multimodal query, OSrCIR uses an LMM to generate a description, uses it to match database images, and returns top-ranked ones. Instead, Paracosm uses an LMM to generate a "mental image" for each query, which contains much richer information than a description, allowing image-to-image matching for better retrieval. Consequently, Paracosm yields better retrievals than OSrCIR.

where λ is a hyperparameter controlling the contribution of incorporating the modification text $\mathbf{t}_{mod}$. The weights of features are discussed in detail in Section 4.2 and Section C. Finally, Paracosm computes the cosine similarity score and returns the index of the potential target image:

$$i^* = \underset{i=1\cdots n}{\operatorname{argmax}} \frac{\mathbf{q}^T \boldsymbol{\phi}^i}{\|\mathbf{q}\|_2 \|\boldsymbol{\phi}^i\|_2}$$

3.3 Remarks

Paracosm makes better use of LMMs. Existing ZS-CIR approaches [24, 57, 70] also leverage LMMs but focus on generating descriptions for multimodal queries and transforming the CIR problem into a text-to-image retrieval problem. However, text descriptions alone cannot sufficiently capture the rich visual information crucial to CIR and hence often lead to incorrectly retrieved images, as demonstrated in Fig. 4. A couple of recent works [32, 65] realize the importance of incorporating pseudo-target images to facilitate CIR, e.g., by exploiting a pretrained text-to-image (T2I) generative model [77] to synthesize a pseudo-target image based on a description. Differently, we leverage an LMM that has image editing capability, i.e., editing the reference image based on the modification text into a "mental image". As shown in Table 1, editing reference images yields better CIR results than T2I generation of pseudo-target images. Moreover, Paracosm exploits an LMM to generate detailed descriptions for each database image and uses these descriptions as prompts to generate synthetic counterparts.

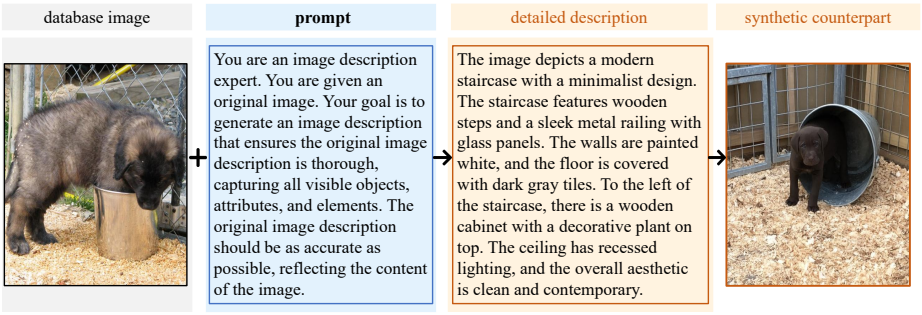

Fig. 5: Illustration of processing database images. For a database image, we first prompt an LMM to generate a comprehensive description about its visual content, capturing all visible objects, attributes, and visual elements. Using this description, we prompt a text-to-image generation model to produce a synthetic counterpart for this database image. For a database image, we use both itself (i.e., the real photo) and its synthetic counterpart as representation for retrieval.

Is CIR still needed given the high-quality edited images? This question emerges in front of the high quality of the generated "mental images" based on multimodal queries (Fig. 4). We argue that CIR is still important in real-world visual search applications. For example, in e-commerce and the fashion industry, users might want to start from a photo of clothes and search for a different genre or style in an e-shop by specifying how to alter the photo. That said, no matter how photorealistic a mental image is, image generation cannot replace a real product in inventory.

Computation cost. As Paracosm extensively exploits LMMs, it has a high computation cost, especially for generating images, i.e., the mental images for the query and synthetic counterparts of database images. Yet, this is not a flaw pertaining only to Paracosm, as existing works also rely on generative models for retrieval. For example, prevailing ZS-CIR methods use LMMs to generate descriptions [24, 32, 57, 70], and some turn to generative models to synthesize images [32, 65]. Importantly, like these methods, Paracosm can process database images ahead of time and does not require computing features for them on the fly during inference. Thus, the major computational overhead is on the process of a given multimodal query during inference, including generating the corresponding mental images and their descriptions. As efficient inference in generative models is an important topic and has been greatly advanced through model optimization [23, 71] and optimized implementation [16, 26], Paracosm would not suffer from computation in the long run.

4 Experiments

We conduct extensive experiments to validate the proposed Paracosm, comparing it against existing methods and ablating its components. We start by introducing datasets, metrics, implementation details, and compared methods.

Table 1: Comparing methods for mental image generation. We adopt the CLIP ViT-B/32 VLM and report results on the CIRR test set. "T2I Generation" means that we first generate pseudo-target descriptions and then use them to generate mental images, as done in [32]. "Image Edit" is our final design choice, for generating mental images in Paracosm, which directly edits the reference image based on the corresponding modification text. Clearly, the latter performs better. Moreover, for "Image Edit", we compare using Qwen-Image-Edit vs. LongCat-Image-Edit as the image generator. Results show stable performance of our Paracosm.

Method	R@1	R@5	R@10	R@50
T2I Generation	31.71	61.37	73.59	91.54
Image Edit w/ Qwen	**32.27**	**62.60**	**75.16**	**92.60**
Image Edit w/ LongCat	32.12	62.20	74.43	92.24

Datasets. Following the literature [6, 14, 32, 43, 70], we use established benchmarks: CIRR [35], CIRCO [2], and Fashion IQ [67]. These datasets are publicly available for non-commercial research and educational purposes. CIRCO is released under the CC-BY-NC 4.0 license; CIRR is licensed under CC-BY 4.0; Fashion-IQ is distributed under the Community Data License Agreement (CDLA). As we focus on ZS-CIR, we do not exploit their training data to develop models. For Fashion IQ, which only releases its validation set, we benchmark methods on this val-set; for other datasets, we benchmark methods on their official test sets. Moreover, Fashion IQ has multiple subsets for testing, we report results on each of them. We provide more details in the supplementary Section A.

Metrics. Following prior arts [18, 24, 57, 70], we use Recall@k (R@k) as the metric for CIRR and Fashion IQ. For CIRCO, we use mean average precision (mAP@k) as it has more target images for queries.

Implementation Details. Following the literature [24, 57, 70], we report the results of all the methods using CLIP ViT-L/14 [40] and OpenCLIP ViT-G/14 [22] to compute image and text features. The LMMs used in Paracosm are Qwen2.5-VL-7B-Instruct [60], Qwen-Image [66], and Qwen-Image-Edit [66], for description generation, database image synthesis, and "mental image" generation. For all image generation tasks (mental images and synthetic counterparts), we set the output resolution to 512×512 pixels, with all other parameters following the default values recommended by the official implementations. We preprocess database images with a computing cluster of 16× NVIDIA A100 GPUs. We run individual methods on a single NVIDIA A100 GPU. Figures 3 and 5 show the detailed generation process. Paracosm (Fig. 2) employs a hyperparameter λ (Eq. 1) to control the contribution of modification texts from the query. We tuned λ on the CIRR validation set using CLIP ViT-B/32 and set $\lambda = 0.3$ throughout our work. Fig. 6 studies its effects on the final CIR performance over the benchmarks. Interestingly, $\lambda = 0.3$ consistently produces the highest numeric metrics on all the benchmarks.

Compared Methods. We compare our Paracosm with representative and recent ZS-CIR methods, which can be categorized into training-based and training-free groups. We also compare baseline methods.

Table 2: Benchmarking results on CIRCO and CIRR test sets. We evaluate Paracosm against state-of-the-art zero-shot CIR methods using ViT-L/14 and ViT-G/14 backbones. Metrics include Recall@k/Recall$_{\text{Subset}}$@k for CIRR and mAP@k for CIRCO (which contains multiple target images per query). Best results are in **bold**, second-best are underlined. Paracosm achieves state-of-the-art performance among zero-shot methods across all benchmarks and backbones, demonstrating the effectiveness of our generation approach. Notably, our training-free method even surpasses several supervised approaches, highlighting its practical potential.

			CIRR						CIRCO			
			Recall@k			Recall$_{\text{Subset}}$@k			mAP@k			
Backbone	**Method**	venue&year	k=1	k=5	k=10	k=1	k=2	k=3	k=5	k=10	k=25	k=50
supervised methods	Combiner [7]	CVPR'22	33.59	65.35	77.35	62.39	81.81	92.02	–	–	–	–
	BLIP4CIR [36]	WACV'24	40.17	71.81	83.18	72.34	88.70	95.23	–	–	–	–
	CLIP-ProbCR [30]	ICMR'24	23.32	54.36	68.64	54.32	76.30	88.88	–	–	–	–
ViT-L/14	Image-only	baseline	7.47	23.86	34.10	20.82	41.88	61.23	1.80	2.43	3.04	3.45
	Text-only	baseline	22.05	45.64	57.54	61.69	80.31	90.41	2.99	3.18	3.68	3.92
	Image+Text	baseline	10.58	32.65	45.69	31.08	55.71	73.90	3.89	4.79	5.93	6.47
	Pic2Word [43]	CVPR'23	23.90	51.70	65.30	53.76	74.46	87.07	8.72	9.51	10.64	11.29
	SEARLE [6]	ICCV'23	24.24	52.48	66.29	53.76	75.01	88.19	11.68	12.73	14.33	15.12
	LinCIR [18]	CVPR'24	25.04	53.25	66.68	57.11	77.37	88.89	12.59	13.58	15.00	15.85
	LDRE [70]	SIGIR'24	26.53	55.57	67.54	60.43	80.31	89.90	23.35	24.03	26.44	27.50
	CIReVL [24]	ICLR'24	24.55	52.31	64.92	59.54	79.88	89.69	18.57	19.01	20.89	21.80
	IP-CIR + LDRE [32]	CVPR'25	29.76	58.82	71.21	62.48	81.64	90.89	26.43	27.41	29.87	31.07
	CIG + SEARLE [65]	CVPR'25	26.72	55.52	68.10	57.95	77.81	89.45	12.84	13.64	15.32	16.17
	Paracosm	ours	**31.95**	**61.56**	**72.96**	**64.68**	**82.89**	**91.47**	**30.24**	**31.51**	**34.29**	**35.42**
ViT-G/14	Pic2Word [43]	CVPR'23	30.41	58.12	69.23	68.92	85.45	93.04	5.54	5.59	6.68	7.12
	SEARLE [6]	ICCV'23	34.80	64.07	75.11	68.72	84.70	93.23	13.20	13.85	15.32	16.04
	LinCIR [18]	CVPR'24	35.25	64.72	76.05	63.35	82.22	91.98	19.71	21.01	23.13	24.18
	LDRE [70]	SIGIR'24	36.15	66.39	77.25	68.82	85.66	93.76	31.12	32.24	34.95	36.03
	CIReVL [24]	ICLR'24	34.65	64.29	75.06	67.95	84.87	93.21	26.77	27.59	29.96	31.03
	CIG + LinCIR [65]	CVPR'25	36.05	66.31	76.96	64.94	83.18	91.93	20.64	21.90	24.04	25.20
	CoTMR [52]	ICCV'25	36.36	67.52	77.82	**71.19**	86.34	93.87	32.23	32.72	35.60	36.83
	OSrCIR [57]	CVPR'25	37.26	67.25	77.33	69.22	85.28	93.55	30.47	31.14	35.03	36.59
	Paracosm	ours	**39.30**	**70.41**	**80.39**	70.82	**86.92**	**94.46**	**39.82**	**40.86**	**43.96**	**45.05**

- *Baseline methods* include "image-only" which uses features of reference images for matching database images, "text-only" which uses modification text for cross-modality matching, and "image+text" which sums the image and text features for matching.
- *Training-based methods* include Pic2Word [43], SEARLE [6], LinCIR [18], CIG [65], and IP-CIR [32]. All these methods rely on textual features, e.g., training a textual inversion network to map reference images into pseudo-word tokens and merge them with textual tokens of modification texts. Notably, IP-CIR [32] and CIG [65] generate pseudo-target images for multimodal queries using a pretrained generative model. Yet, they run along with other methods to ensure good CIR performance.
- *Training-free methods.* CIReVL [24] uses an LMM to generate a description of a given reference image, and then an LLM to modify this description based on the modification text. It uses the modified description for cross-modality retrieval in the image database. LDRE [70] proposes to generate multiple diverse descriptions to improve CIR performance. AutoCIR [14] employs automatic multi-agent collaboration to iteratively refine retrieval queries through self-correction. CoTMR [52] leverages an LMM to generate both global captions and fine-grained object-level descriptions combined with a multi-grained scoring that significantly improves the retrieval performance.

Table 3: Benchmarking results on the Fashion IQ validation set. We compare Paracosm against zero-shot CIR methods using ViT-L/14 and ViT-G/14 backbones. Metrics include Recall@10 (R@10) and Recall@50 (R@50) across three categories (Shirt, Dress, Toptee) and their average. Best results are in **bold**, second-best are underlined. Paracosm achieves state-of-the-art performance among zero-shot methods across all benchmarks and backbones.

			Shirt		Dress		Toptee		Average	
Backbone	**Method**	venue&year	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
supervised methods	Combiner [7]	CVPR'22	36.36	58.00	31.63	56.67	38.19	62.42	35.39	59.03
	PL4CIR [74]	SIGIR'22	33.22	59.99	46.17	68.79	46.46	73.84	41.98	67.54
	Uncertainty retrieval [13]	ICLR'24	32.61	61.34	33.23	62.55	41.40	72.51	35.75	65.47
ViT-L/14	Image-only	baseline	10.45	20.76	5.21	13.49	8.01	18.05	7.89	17.43
	Text-only	baseline	20.26	34.10	15.12	33.71	21.98	39.98	19.12	35.93
	Image+Text	baseline	19.14	32.63	14.38	31.09	20.50	36.26	18.01	33.33
	Pic2Word [43]	CVPR'23	26.20	43.60	20.00	40.20	27.90	47.40	24.70	43.70
	SEARLE [6]	ICCV'23	26.89	45.58	20.48	43.13	29.32	49.97	25.56	46.23
	LinCIR [18]	CVPR'24	29.10	46.81	20.92	42.44	28.81	50.18	26.28	46.49
	CIReVL [24]	ICLR'24	29.49	47.40	24.79	44.76	31.36	**53.65**	28.55	48.57
	CIG + LinCIR [65]	CVPR'25	28.90	47.25	21.12	43.88	29.78	50.54	26.60	47.22
	Paracosm	ours	**31.80**	**49.51**	**24.99**	**47.45**	**31.82**	52.83	**29.45**	**49.93**
ViT-G/14	Pic2Word [43]	CVPR'23	33.17	50.39	25.43	47.65	35.24	57.62	31.28	51.89
	SEARLE [6]	ICCV'23	36.46	55.35	28.16	50.32	39.83	61.45	34.82	55.71
	CIReVL [24]	ICLR'24	33.71	51.42	27.07	49.53	35.80	56.14	32.19	52.36
	LDRE [70]	SIGIR'24	35.94	**58.58**	26.11	51.12	35.42	56.67	32.49	55.46
	AutoCIR [14]	KDD'25	36.36	55.84	26.18	47.69	37.28	60.38	33.27	54.63
	OSrCIR [57]	CVPR'25	38.65	54.71	33.02	54.78	41.04	61.83	37.57	57.11
	Paracosm	ours	**40.48**	57.80	**33.17**	**55.18**	**42.58**	**64.20**	**38.74**	**59.06**

OSrCIR [57] carefully designs the prompt for an LMM with Reflective Chain-of-Thought to improve the output description quality for a multimodal query.

4.1 Benchmarking Results

Quantitative Results. Table 2 and 3 display benchmarking results of all the compared methods across different VLM backbones, with respect to different metrics on all the test sets. Convincingly, Paracosm significantly outperforms all compared methods. Generally, with more powerful VLM models, e.g., ViT-G/14 > ViT-L/14, methods yield better CIR performance. Paracosm even rivals recent supervised learning CIR methods [13, 74].

Qualitative Results. Fig. 4 visually compares the retrieved images by Paracosm and the recent method OSrCIR [57], along with their generated mental images and descriptions, respectively. Clearly, generated descriptions by OSrCIR insufficiently represent the given multimodal queries, whereas mental images by Paracosm preserve and capture rich visual details pertaining to the multimodal queries. This helps explain why Paracosm outperforms OSrCIR.

Efficiency Comparisons. For a comprehensive analysis, we evaluate the computational costs of Paracosm and some recent zero-shot methods under the same experimental conditions in Table 4. To ensure a fair comparison, all methods are evaluated using the same VLM backbone (OpenCLIP ViT-L/14 [22]) for feature extraction and matching and are run on NVIDIA A100 GPUs. Regarding LMM backbones, AutoCIR [14] and OSrCIR [57] use GPT [38], whereas CoTMR [52] and our Paracosm use Qwen [5]. Computational costs are measured on the CIRCO benchmark [2], which comprises over 123K database

Table 4: Efficiency comparisons. Computational costs are measured on the CIRCO benchmark. We break down the costs into (1) offline one-time database processing (w.r.t. wall-clock time, and storage for synthetic images and features), and (2) online per-query inference (w.r.t. GPU memory consumption and latency). Notably, synthetic database images are generated only during pre-processing for feature extraction and do not need to be stored permanently; only compact features (0.38 GB) are retained. While Paracosm requires additional offline computation to mitigate synthetic-to-real domain gaps, it achieves state-of-the-art performance.

Method	time	img. stg.	feat. stg.	inference		CIRCO	CIRR	FashionIQ
	hrs	GB	GB	GB	sec	mAP@5	R@1	R@10
AutoCIR [14]	0.1	n/a	0.38	2.3	24	24.05	31.81	30.68
CoTMR [52]	0.1	n/a	0.38	70.2	1	27.61	35.02	35.05
OSrCIR [57]	0.1	n/a	0.38	2.3	3	23.87	29.45	33.26
Paracosm	12.9	41.4	0.38	2.7	14	37.40	38.24	36.45

images. The results show that (1) Paracosm incurs substantial wall time hours to process database images, (2) but its inference efficiency is comparable to AutoCIR, and (3) importantly, Paracosm significantly outperforms these methods!

A Warning. It is worth noting that some results of recent studies are not reproducible probably because of misreported backbones in the papers. We find that using OpenCLIP [22] achieves better ZS-CIR performance than CLIP [40], which is reported to be used for experiments in some papers such as OSrCIR [57]. To rigorously analyze this issue, in Table 5, we copy the results reported by OSrCIR and reproduce it in the same setting and under different settings. Clearly, the results achieved by using OpenCLIP in our implementation match what are reported in [57], which, however, reports using CLIP. Our analysis supports the doubts in the community, as seen in the Issues of their official GitHub repositories [1]. Moreover, results in Table 5 show that using GPT-4o achieves better performance than Qwen2.5-VL for OSrCIR, while our Paracosm, which uses Qwen2.5-VL, outperforms OSrCIR regardless of using Qwen2.5-VL or GPT-4o.

4.2 Analyses and Ablation Study

Reference image editing vs. Text-to-Image generation. When generating mental images for multimodal queries, our Paracosm leverages the LMM to directly edit the reference images based on modification texts. In comparison, another workaround is T2I generation, as done in the recent work [32]: generating descriptions for reference images using an LMM, then modifying the descriptions based on modification texts using an LLM, and lastly generating mental images based on the modified descriptions. Table 1 compares them, clearly showing that Paracosm's design choice leads to better performance.

Different image generator. We compare the results of Paracosm using Qwen-Image-Edit [66] and another open-source image generator, LongCat-Image-Edit [59], on the CIRR test set without tuning hyperparameters or prompts. The results in Table 1 show stable performance of Paracosm with different generators.

Table 5: Rigorous comparisons between using OpenCLIP vs. CLIP backbones. We copy the results of OSrCIR [57] from its original paper (marked in grey shade), which reports that it uses the CLIP ViT-B/32 backbone and GPT-4o in experiments. We reproduce it in the same setting but achieve notably worse performance. We also adapt OSrCIR by using the OpenCLIP backbone with either GPT-4o or Qwen2.5-VL. Notably, the results achieved with OpenCLIP (in blue shade) match what are reported in OSrCIR [57]. This suggests that the backbone used by OSrCIR is likely to be OpenCLIP, instead of CLIP. Our study helps answer questions regarding the performance gaps in the community, as seen in the Issues of the OSrCIR's official GitHub repository [1]. Refer to the main text for more analyses.

				CIRCO				Fashion IQ	
				mAP@k				Average	
Backbone	**Method**	**LMMs**	venue&year	k=5	k=10	k=25	k=50	R@10	R@50
CLIP ViT-B/32	OSrCIR	GPT-4o	CVPR'25	18.04	19.17	20.94	21.85	32.34	53.40
	OSrCIR	GPT-4o	reproduced	14.91	15.34	16.79	17.60	19.15	36.88
	OSrCIR	Qwen2.5-VL	reproduced	11.84	12.39	13.53	14.27	18.75	36.05
	Paracosm	Qwen2.5-VL	ours	26.10	27.02	29.29	30.45	26.14	46.39
CLIP ViT-L/14	OSrCIR	GPT-4o	CVPR'25	23.87	25.33	27.84	28.97	33.26	54.37
	OSrCIR	GPT-4o	reproduced	17.70	18.56	20.52	21.47	23.01	41.39
	OSrCIR	Qwen2.5-VL	reproduced	14.06	14.93	16.65	17.50	20.95	39.49
	Paracosm	Qwen2.5-VL	ours	30.24	31.51	34.29	35.42	28.76	49.32
OpenCLIP ViT-B/32	OSrCIR	GPT-4o	reproduced	18.77	19.33	21.04	22.02	31.41	52.26
	OSrCIR	Qwen2.5-VL	reproduced	14.47	15.23	16.70	17.55	26.54	47.35
	Paracosm	Qwen2.5-VL	ours	31.31	32.48	35.19	36.33	35.14	56.04
OpenCLIP ViT-L/14	OSrCIR	GPT-4o	reproduced	22.75	23.51	25.73	26.78	31.55	51.90
	OSrCIR	Qwen2.5-VL	reproduced	17.88	18.72	20.51	21.39	27.84	47.16
	Paracosm	Qwen2.5-VL	ours	37.40	38.64	41.57	42.73	36.45	57.60

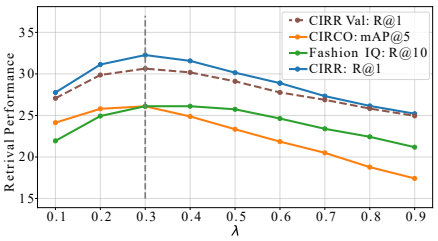

Fig. 6: Analysis of λ which controls the importance of incorporating modification text in Eq. 1. We set $\lambda = 0.3$ based on the results on the CIRR validation set. Interestingly, on all datasets, setting $\lambda = 0.3$ consistently yields the highest numeric metrics reported for all the datasets.

Ablation Study. We ablate core components in Paracosm with the CLIP ViT-B/32 VLM on CIRR and CIRCO. Table 6 summarizes key results; supplementary Section E provides comprehensive details.

- *Incorporating mental images boosts performance.* While existing training-free methods generate descriptions $\mathbf{t}_{query}$, aka pseudo-target descriptions [57], Paracosm further incorporates its generated mental images $\mathbf{I}_{mental}$. This results in clear performance gains.
- *Incorporating synthetic counterparts of database images boosts performance.* As $\mathbf{I}_{mental}$ has synthetic-to-real domain gaps with database images, Paracosm further generates synthetic counterparts $\mathbf{I}^i_{syn}$ for them and uses such for matching (Eq. 1). Clearly, doing this significantly boosts the final performance.
- *Incorporating modification texts boosts performance.* Intuitively, modification texts $\mathbf{t}_{mod}$ contain crucial keywords that can be leveraged for CIR. Indeed,

Table 6: Ablation study. We ablate key components of Paracosm with the CLIP ViT-B/32 VLM on the CIRCO and CIRR test sets. Per Eq. 1, we focus on constructing features for multimodal queries and database images. $\mathbf{I}_{mental}$, $\mathbf{t}_{query}$, and $\mathbf{t}_{mod}$ indicate the generated mental image, description of the target image based on the query, and modification text, respectively. $\mathbf{I}^i$ and $\mathbf{I}^i_{syn}$ represent database images and their synthetic counterparts. Best results are highlighted in **bold**. Clearly, incorporating $\mathbf{I}_{mental}$, $\mathbf{I}_{syn}$, and $\mathbf{t}_{mod}$ significantly boosts CIR performance.

multimodal query			database images		CIRR							CIRCO			
					Recall@k				$\text{Recall}_{\text{Subset}}$@k			mAP@k			
$\mathbf{t}_{query}$	$\mathbf{I}_{mental}$	$\mathbf{t}_{mod}$	$\mathbf{I}^i$	$\mathbf{I}_{syn}$	k=1	k=5	k=10	k=50	k=1	k=2	k=3	k=5	k=10	k=25	k=50
✓			✓		17.21	43.49	55.90	81.16	52.00	72.31	84.92	14.91	15.34	16.79	17.60
✓	✓		✓		18.80	44.96	58.84	82.65	50.39	72.80	83.93	13.71	13.89	15.19	15.92
✓	✓	✓	✓		27.93	57.11	70.29	90.31	61.88	80.70	90.70	18.29	18.69	20.45	21.27
✓				✓	17.21	43.93	56.29	80.10	50.94	71.81	84.53	13.58	13.92	15.44	16.30
✓	✓			✓	19.95	46.29	59.98	82.34	51.42	72.17	84.80	14.07	14.43	15.76	16.42
✓	✓	✓		✓	25.95	54.24	67.16	88.84	60.80	79.71	89.74	17.43	17.98	19.62	20.46
✓			✓	✓	22.46	50.99	63.59	84.58	53.88	74.80	86.27	16.72	17.41	19.19	20.11
✓	✓		✓	✓	24.60	52.68	65.86	86.53	54.84	74.92	86.82	16.57	17.10	18.59	19.30
✓	✓	✓	✓	✓	**32.27**	**62.60**	**75.16**	**92.60**	**65.16**	**83.25**	**92.34**	**26.10**	**27.02**	**29.29**	**30.45**

our ablation results demonstrate that incorporating modification texts greatly boosts CIR performance. This observation aligns with what is reported in prior works [6, 18, 65].

5 Impacts and Limitations

Societal Impacts. CIR is an important component of search services in many real-world applications, such as e-commerce and the fashion industry. The proposed Paracosm, by extensively leveraging LMMs, greatly advances the CIR area, as demonstrated by its state-of-the-art performance on standard benchmark datasets. It is worth noting that LMMs' large-scale pretraining datasets might contain poisonous, offensive, or biased data. This could cause the LMMs to generate inaccurate, biased, or even harmful text descriptions and synthetic images. Moreover, Paracosm does not have an alerting mechanism for inappropriate multimodal queries, which may contain malicious modification texts or intend to retrieve inappropriate target images. If forced to retrieve target images based on such queries, the intermediate mental images and retrieval outputs may lead to negative societal impacts.

Limitations and Future Work. Paracosm relies on the quality of LMM-generated mental images for queries and synthetic counterparts of database images, as well as text descriptions for them. This means that the performance of Paracosm is inherently contingent upon the generative capabilities of the underlying LMMs. These generated images might look visually plausible at first sight but often lack factual fidelity and fine-grained details. For instance, in the first row of Figure 7, the mental image contains a cartoon-style duck, which does not correspond to any actual items in the real world and causes retrieval to fail. Future work can focus on improving methods for image generation or editing and developing methods to handle erroneously generated visuals. Moreover, to achieve state-of-the-art performance, Paracosm uses slightly different prompt

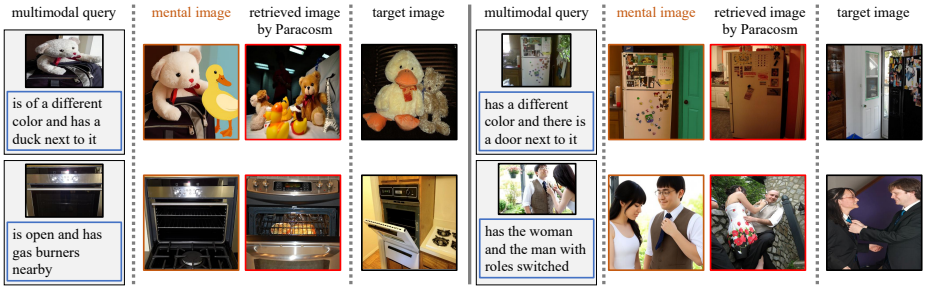

Fig. 7: Failure cases on the CIRCO dataset. Paracosm can fail due to limitations of generative models that can generate implausible and counterfactual mental images. Four examples, respectively, demonstrate different failures of Paracosm. (1) It incorrectly generates a cartoon-style duck, making it fail to return the correct database image, which captures a plush toy duck. (2) It generates a counterfactual oven that has gas burners on its door, making it incorrectly retrieve an image that captures a burning oven. (3) It fails to edit the door color of the specified refrigerator and hence fails to return the correct target image. (4) It fails to comprehend the multimodal query, resulting in an incorrect mental image and hence failing to return the target image.

templates in the benchmarking datasets. This is primarily due to variations in the format and specificity of modification texts across datasets, as shown in Figure 3. For example, a modification text can contain a relative caption and a shared concept(ref. CIRCO). To overcome this, future work could develop adaptive prompt generation methods that leverage LMMs to automatically construct optimal prompts for diverse types of multimodal queries, reducing the need for manual tuning and enhancing robustness across varying dataset distributions.

6 Conclusions

We present Paracosm, a novel and rather simple training-free zero-shot CIR method. Different from most existing ZS-CIR methods, which rely on generated text descriptions of multimodal queries, Paracosm leverages LMMs to generate mental images for multimodal queries, offering rich visual information that is beneficial to CIR. Moreover, as the synthetic mental images have synthetic-to-real domain gaps with real database images, it further generates synthetic counterparts of database images. By incorporating the generated visuals as well as the generated descriptions, it achieves state-of-the-art performance on popular benchmarking datasets. Paracosm significantly outperforms existing ZS-CIR methods and even rivals supervised learning approaches.

Acknowledgments

This work was supported by the Science and Technology Development Fund of Macau (0058/2025/RIA2, 0067/2024/ITP2), the University of Macau (SRG2023-00044-FST), the Institute of Collaborative Innovation, and the CK Foundation. Authors thank Hanxin Wang, Di Wu, and Nan Deng for helpful discussions.

References

1. Github. https://github.com/Pter61/osrcir/issues/5, 2025.07.20
2. Agnolucci, L., Baldrati, A., Bimbo, A.D., Bertini, M.: isearle: Improving textual inversion for zero-shot composed image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **47**(11), 10801–10817 (2025)
3. Arabi, A.A., Cai, C.Y., Lu, R., Kong, S., Kim, J.: Eurexa: End-user reconfiguration of environment with explainable augmentation for generative fabrication. In: CHI Conference on Human Factors in Computing Systems (2026)
4. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
5. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
6. Baldrati, A., Agnolucci, L., Bertini, M., Del Bimbo, A.: Zero-shot composed image retrieval with textual inversion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15338–15347 (October 2023)
7. Baldrati, A., Bertini, M., Uricchio, T., Del Bimbo, A.: Effective conditioned and composed image retrieval combining clip-based features. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
8. Bao, T., Liu, C., Xu, D., Zheng, Z., Xu, T.: Mllm-i2w: Harnessing multimodal large language model for zero-shot composed image retrieval. In: International Conference on Computational Linguistics (COLING) (2025)
9. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18392–18402 (June 2023)
10. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in Neural Information Processing Systems (NeurIPS) **33**, 1877–1901 (2020)
11. Chen, Y., Bazzani, L.: Learning joint visual semantic matching embeddings for language-guided retrieval. In: European Conference on Computer Vision (ECCV). pp. 136–152. Springer (2020)
12. Chen, Y., Gong, S., Bazzani, L.: Image search with text feedback by visiolinguistic attention learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3001–3011 (2020)
13. Chen, Y., Zheng, Z., Ji, W., Qu, L., Chua, T.S.: Composed image retrieval with text feedback via multi-grained uncertainty regularization. In: International Conference on Learning Representations (ICLR) (2024)
14. Cheng, Z., Ma, Y., Lang, J., Zhang, K., Zhong, T., Wang, Y., Zhou, F.: Generative thinking, corrective action: User-friendly composed image retrieval via automatic multi-agent collaboration. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD). pp. 334–344 (2025)
15. Delmas, G., Rezende, R.S., Csurka, G., Larlus, D.: Artemis: Attention-based retrieval with text-explicit matching and implicit similarity. In: International Conference on Learning Representations (ICLR) (2022)
16. Di, Z., Zhu, G., Duan, Z., Chu, Z., Chen, Y., Lu, W.: Diffsynth-engine: a high-performance diffusion inference engine. https://github.com/modelscope/diffsynth-engine (2025)
17. Gao, J., Zhang, J., Liu, X., Darrell, T., Shelhamer, E., Wang, D.: Back to the source: Diffusion-driven adaptation to test-time corruption. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)

18. Gu, G., Chun, S., Kim, W., Kang, Y., Yun, S.: Language-only training of zero-shot composed image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13225–13234 (June 2024)
19. Guo, J., Zhao, J., Du, C., Wang, Y., Ge, C., Ni, Z., Song, S., Shi, H., Huang, G.: Everything to the synthetic: Diffusion-driven test-time adaptation via synthetic-domain alignment. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) (2025)
20. Hu, Z., Zhu, X., Tran, S., Vidal, R., Dhua, A.: Provla: Compositional image search with progressive vision-language alignment and multimodal fusion. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
21. Huynh, C., Yang, J., Tawari, A., Shah, M., Tran, S., Hamid, R., Chilimbi, T., Shrivastava, A.: Collm: A large language model for composed image retrieval. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). pp. 3994–4004 (2025)
22. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: Openclip (Jul 2021). https://doi.org/10.5281/zenodo.5143773
23. Imagen-Team-Google: Imagen 3. arXiv preprint arXiv:2408.07009 (2024)
24. Karthik, S., Roth, K., Mancini, M., Akata, Z.: Vision-by-language for training-free compositional image retrieval. In: International Conference on Learning Representations (ICLR) (2024)
25. Kwon, N., Lu, Q., Qazi, M.H., Liu, J., Oh, C., Kong, S., Kim, J.: Accesslens: Auto-detecting inaccessibility of everyday objects. In: CHI Conference on Human Factors in Computing Systems (2024)
26. Labs, B.F.: FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2 (2025)
27. Lee, S., Kim, D., Han, B.: Cosmo: Content-style modulation for image retrieval with text feedback. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 802–812 (June 2021)
28. Levy, M., Ben-Ari, R., Darshan, N., Lischinski, D.: Data roaming and quality assessment for composed image retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). vol. 38, pp. 2991–2999 (2024)
29. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning (ICML). pp. 12888–12900. PMLR (2022)
30. Li, M., Zhao, Z., Jiang, X., Jiang, Z.: Clip-probcr: Clip-based probability embedding combination retrieval. In: Proceedings of the 2024 International Conference on Multimedia Retrieval (ICMR). pp. 1104–1109 (2024)
31. Li, W., Fan, H., Wong, Y., Yang, Y., Kankanhalli, M.: Improving context understanding in multimodal large language models via multimodal composition learning. In: Forty-first International Conference on Machine Learning (ICML) (2024)
32. Li, Y., Ma, F., Yang, Y.: Imagine and seek: Improving composed image retrieval with an imagined proxy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3984–3993 (June 2025)
33. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision (ECCV). pp. 740–755. Springer (2014)
34. Liu, Y., Yao, J., Zhang, Y., Wang, Y., Xie, W.: Zero-shot composed text-image retrieval. arxiv:2306.07272 (2024)

35. Liu, Z., Rodriguez-Opazo, C., Teney, D., Gould, S.: Image retrieval on real-life images with pre-trained vision-and-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
36. Liu, Z., Sun, W., Hong, Y., Teney, D., Gould, S.: Bi-directional training for composed image retrieval via text prompt learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2024)
37. Liu, Z., Sun, W., Teney, D., Gould, S.: Candidate set re-ranking for composed image retrieval with dual multi-modal encoder. Transactions on Machine Learning Research (TMLR) (2024)
38. OpenAI: Gpt-4 technical report. arxiv preprint arxiv:2303.08774 (2024)
39. Parashar, S., Lin, Z., Liu, T., Dong, X., Li, Y., Ramanan, D., Caverlee, J., Kong, S.: The neglected tails in vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
40. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML). pp. 8748–8763. PmLR (2021)
41. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
42. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015)
43. Saito, K., Sohn, K., Zhang, X., Li, C.L., Lee, C.Y., Saenko, K., Pfister, T.: Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19305–19314 (June 2023)
44. Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial diffusion distillation. In: European Conference on Computer Vision (ECCV) (2024)
45. Shafique, S., Kong, B., Kong, S., Fowlkes, C.: Creating a forensic database of shoeprints from online shoe-tread photos. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 858–868 (2023)
46. Shafique, S., Kong, S., Fowlkes, C.: Crisp: Leveraging tread depth maps for enhanced crime-scene shoeprint matching. In: European Conference on Computer Vision (2024)
47. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Annual Meeting of the Association for Computational Linguistics (ACL) (2018)
48. Shen, Q., Zhao, Y., Kwon, N., Kim, J., Li, Y., Kong, S.: A high-resolution dataset for instance detection with multi-view object capture. Advances in Neural Information Processing Systems **36**, 42064–42076 (2023)
49. Shen, Q., Zhao, Y., Kwon, N., Kim, J., Li, Y., Kong, S.: Solving instance detection from an open-world perspective. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9901–9910 (2025)
50. Song, X., Feng, F., Han, X., Yang, X., Liu, W., Nie, L.: Neural compatibility modeling with attentive knowledge distillation. In: ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR) (2018)
51. Suhr, A., Zhou, S., Zhang, A., Zhang, I., Bai, H., Artzi, Y.: A corpus for reasoning about natural language grounded in photographs. In: Annual Meeting of the Association for Computational Linguistics (ACL) (2019)

52. Sun, Z., Jing, D., Lu, Z.: Cotmr: Chain-of-thought multi-scale reasoning for training-free zero-shot composed image retrieval. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2025)
53. Sun, Z., Fang, Y., Wu, T., Zhang, P., Zang, Y., Kong, S., Xiong, Y., Lin, D., Wang, J.: Alpha-clip: A clip model focusing on wherever you want. In: IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR) (2024)
54. Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490 (2019)
55. Tang, Y., Yu, J., Gai, K., Zhuang, J., Xiong, G., Gou, G., Wu, Q.: Missing target-relevant information prediction with world model for accurate zero-shot composed image retrieval. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). pp. 24785–24795 (2025)
56. Tang, Y., Yu, J., Gai, K., Zhuang, J., Xiong, G., Hu, Y., Wu, Q.: Context-i2w: Mapping images to context-dependent words for accurate zero-shot composed image retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). vol. 38, pp. 5180–5188 (2024)
57. Tang, Y., Zhang, J., Qin, X., Yu, J., Gou, G., Xiong, G., Lin, Q., Rajmohan, S., Zhang, D., Wu, Q.: Reason-before-retrieve: One-stage reflective chain-of-thoughts for training-free zero-shot composed image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
58. Team, G.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
59. Team, M.L., Ma, H., Tan, H., Huang, J., Wu, J., He, J.Y., Gao, L., Xiao, S., Wei, X., Ma, X., et al.: Longcat-image technical report. arXiv preprint arXiv:2512.07584 (2025)
60. Team, Q.: Qwen2.5-vl (January 2025), https://qwenlm.github.io/blog/qwen2.5-vl/
61. Tian, Y., Newsam, S., Boakye, K.: Fashion image retrieval with text feedback by additive attention compositional learning. In: Proceedings of the Winter Conference on Applications of Computer Vision (WACV). pp. 1011–1021 (2023)
62. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
63. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
64. Vo, N., Jiang, L., Sun, C., Murphy, K., Li, L.J., Fei-Fei, L., Hays, J.: Composing text and image for image retrieval - an empirical odyssey. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
65. Wang, L., Ao, W., Boddeti, V.N., Lim, S.N.: Generative zero-shot composed image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
66. Wu, C., Li, J., Zhou, J., Lin, J., et al.: Qwen-image technical report. arxiv:2508.02324 (2025)
67. Wu, H., Gao, Y., Guo, X., Al-Halah, Z., Rennie, S., Grauman, K., Feris, R.: 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 (CVPR). pp. 11307–11317 (June 2021)

68. Xu, Y., Bin, Y., Wei, J., Yang, Y., Wang, G., Shen, H.T.: Multi-modal transformer with global-local alignment for composed query image retrieval. IEEE Transactions on Multimedia (TMM) **25**, 8346–8357 (2023)
69. Yang, Y., Wang, Y., Wang, Y.: Sda: Semantic discrepancy alignment for text-conditioned image retrieval. In: Findings of the Association for Computational Linguistics (ACL). pp. 5250–5261 (2024)
70. Yang, Z., Xue, D., Qian, S., Dong, W., Xu, C.: Ldre: Llm-based divergent reasoning and ensemble for zero-shot composed image retrieval. In: ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR) (2024)
71. Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T., Park, T.: One-step diffusion with distribution matching distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6613–6623 (2024)
72. Zhang, K., Luan, Y., Hu, H., Lee, K., Qiao, S., Chen, W., Su, Y., Chang, M.W.: MagicLens: Self-supervised image retrieval with open-ended instructions. In: International Conference on Machine Learning (ICML) (2024)
73. Zhang, X., Zheng, Z., Zhu, L., Yang, Y.: Collaborative group: Composed image retrieval via consensus learning from noisy annotations. Knowledge-Based Systems (KBS) **300**, 112135 (2024)
74. Zhao, Y., Song, Y., Jin, Q.: Progressive learning for image retrieval with hybrid-modality queries. In: ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR) (2022)
75. Zhao, Y., Kong, S., Shin, D., Fowlkes, C.: Domain decluttering: Simplifying images to mitigate synthetic-real domain shift and improve depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
76. Zhao, Y., Ma, H., Kong, S., Fowlkes, C.: Instance tracking in 3d scenes from egocentric videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21933–21944 (2024)
77. Zhou, D., Li, Y., Ma, F., Zhang, X., Yang, Y.: Migc: Multi-instance generation controller for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
78. Zhou, S., Kong, S.: Dare to plagiarize? plagiarized painting recognition and retrieval. In: 2025 IEEE International Conference on Advanced Visual and Signal-Based Systems (AVSS) (2025)