Pattern Recognition 179 (2026) 113724

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

BenchCIR: Benchmarking robustness in composed image retrieval across modalities

Shitong Sun [a], Qilei Li [b,c,*], Shaogang Gong [a], Weitong Cai [a], Philip Torr [d], Jindong Gu [d]

[a] Queen Mary University of London, London, E1 4NS, UK
[b] Laboratory for Artificial Intelligence and New Forms of Education, Central China Normal University, Wuhan, 430079, Hubei, China
[c] National Engineering Research Center of Educational Big Data, Faculty of Artificial Intelligence in Education, Central China Normal University, Wuhan, 430079, Hubei, China
[d] University of Oxford, Oxford, OX1 2JD, UK

ARTICLE INFO

Keywords:
Composed image retrieval
Robustness evaluation
Vision-language models
Multimodal retrieval
Query composition

ABSTRACT

Composed image retrieval aims to retrieve images based on a query that consists of a reference image and text describing desired modifications to that image. It has recently attracted attention for its ability to tailor image retrieval to user intentions by combining information-rich reference images with concise natural language instructions. Despite its current success, the robustness of composed image retrieval methods to either (1) common corruptions or (2) variations of the textual descriptions have never been systematically evaluated. In this paper, we perform the first robustness study of composed image retrieval, establishing three new benchmarks for a systematic evaluation of robustness to common corruption (in both the textual and visual domains) and robustness in text understanding. For analysis of natural image corruption, we introduce two new large-scale benchmark datasets, CIRR-C and FashionIQ-C, for the open domains and fashion domains respectively—both of which feature 75 visual corruptions and 35 textual corruptions. To facilitate robust evaluation of text understanding, we introduce a new diagnostic dataset CIRR-D by expanding the CIRR dataset with synthetic data, specifically probing text understanding across variations in: numerical, attribute, object removal, and background. We introduce BenchCIR, a testbed for evaluating composed image retrieval model robustness with standardized evaluation protocols. Through benchmarking ten published models in the testbed, we reveal insights into how the composition of visual and textual modalities affects model robustness. The code is in https://suntongtongtong.github.io/BenchCIR/

1. Introduction

Composed image retrieval aims to retrieve an image of interest from a gallery of images through a composed query consisting of a reference image and its corresponding modified text. For example, the single word 'dog' can map to thousands of images showing different breeds, poses, and scenarios of dogs, indicating that natural language is sparse in semantic spaces while the image domain is dense. Therefore, using *both* images and text together provides the advantage of expressing queries using the unique properties of each modality. Furthermore, the resulting retrieval requests are often much more semantically rich and nuanced than is possible by querying with a single modality alone. This method holds potential in a variety of real-world applications, including fashion domain e-commerce [1] and open domain internet search [2].

However, existing composed image retrieval methods have only been evaluated on clean data, without systematic evaluation of model robustness under distribution shifts [3], including textual typos, image corruptions (e.g., weather changes), and variation of modified text. Meanwhile, existing works on multimodal robustness primarily focus on scenarios where one modality serves as input and the other as output, studying their correspondence relationships. In contrast, the stability of inter-modality interactions when both modalities serve as input simultaneously remains unexplored. In this work, we take the first step towards providing a thorough evaluation of the robustness of composed image retrieval by building three new large-scale robustness benchmarks in our proposed testbed, BenchCIR, on both fashion and open domains. We study the following two questions: **Q1:** *How robust are composed image retrieval models to common corruption in both the*

* Correspondence to: Central China Normal University, Wuhan, 430079, Hubei, China.
E-mail address: qilei.li@ccnu.edu.cn (Q. Li).

https://doi.org/10.1016/j.patcog.2026.113724
Received 8 November 2025; Received in revised form 19 February 2026; Accepted 10 April 2026
Available online 18 April 2026

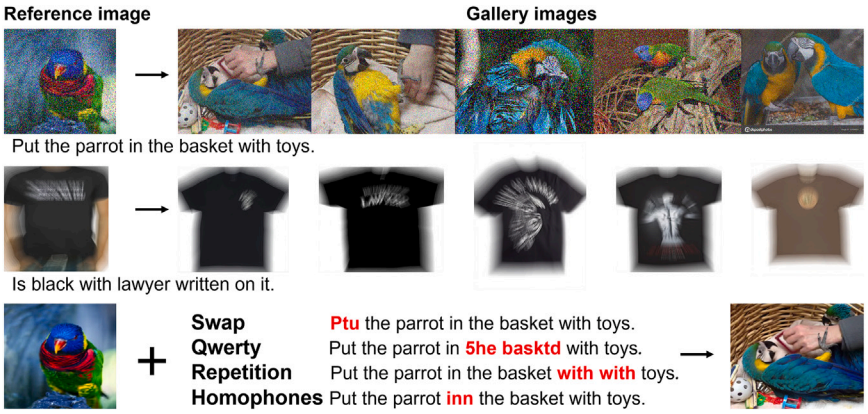


Fig. 1. Sample visualization of proposed benchmarks under common corruption with both visual and textual. **Top:** CIRR-C with impulse noise image corruption **Middle:** FashionIQ-C with zoom blur image corruption; **Bottom:** CIRR-C under character-level (Swap and Qwerty) and word-level (Repetition and Homophones) textual corruptions. Gallery images are shown without a particular order.

visual and textual domains? and **Q2:** *How robust are composed image retrieval models at text understanding?*

To address the first question, we introduce two benchmark datasets on the composed image retrieval task. We propose our two benchmark datasets: FashionIQ-C and CIRR-C. These are based on the FashionIQ [4] and CIRR [5] datasets in the fashion and open domains respectively as shown in Fig. 1. Both datasets incorporate 75 visual corruptions and 35 textual corruptions, providing a comprehensive evaluation of model robustness against common corruptions in both images and text. To answer the second question, we construct a new diagnostic dataset CIRR-D to probe text understanding abilities across variations of four fundamental scenarios: numerical, attribute, object removal, and background, as shown in Fig. 4. Through extensive experiments on ten recently published methods, this benchmark is suitable for robustness analysis against common corruption in both image and text, as well as for probing text understanding abilities.

Our contributions are as follows: (1) We pioneer the analysis of the robustness of composed image retrieval methods against common corruption (in both the visual and textual domains) and understanding under four categories of textual variation. (2) We introduce three new large-scale benchmarks: two benchmark datasets, FashionIQ-C and CIRR-C, to evaluate robustness against common corruption in both image and text, and one diagnostic benchmark, CIRR-D, to probe text understanding robustness. (3) We provide BenchCIR, an open-source testbed for easily evaluating composed image retrieval models. Through extensive experiments, we study the effects of various perturbations and explore how textual and visual modalities contribute to model robustness.

2. Related works

Composed image retrieval. Traditional composed image retrieval models implement separate independent image and text encoders, whose features are combined with late fusion. For example, TIRG [6] and ARTEMIS [7] implement separate pre-trained ResNet as the image encoder and LSTMs as the text encoder. With the development of recent large multimodal models, composed image retrieval models have achieved noticeable improvements. For example, CLIP4CIR [8] leverages the power of CLIP's unified multimodal space [9] by implementing a lightweight adapter for image-text late fusion and further tuning it in target domains. SPRC [10], which is based on BLIP2 [11], achieves noticeable improvements by converting images into a series of tokens and further reasoning using language models. Recent advancements in CIR have further diversified in both methodology and task scope. To capture nuanced multimodal interactions, global-local alignment [12], composition-decomposition learning [13], and geometric matching [14] have been introduced for refined feature correspondence. To address inherent ambiguity and data noise, uncertainty regularization [15], multi-order adversarial learning [16], and noisy triplet correspondence [17] have been proposed. Moreover, the integration of generative models has emerged as a prominent trend; for instance, latent diffusion is leveraged by CompoDiff [18] for versatile retrieval, while generative zero-shot frameworks [19] and data roaming [20] are explored to enhance training quality. The task has also been extended to the video domain via web-caption-based composed retrieval [21]. Despite these algorithmic strides, the robustness of these models, especially under visual and textual perturbations, remains underexplored. Although some of these latest models are not yet integrated into our current evaluation due to code availability or task-specific constraints, our BenchCIR framework provides an open-source and extensible testbed designed to facilitate the systematic assessment of these and future CIR methods as the field evolves.

Robustness analysis. The quantification of robustness aims to evaluate models' ability to defend against common corruption [22], or adversarial attacks [23], Traditional robustness analysis predominantly targets single-modality settings: visual tasks such as face detection [24] or textual tasks such as text classification [25]. Recently, robustness analysis for multimodal tasks (which is closer to real life and attempts to take a step towards a reliable system) has appeared, but is still in its infancy. Li et al. [26] take the first step to systematically analyze the robustness of the multimodal task of Visual Question Answering (VQA) against 4 generic robustness measures. Schiappa et al. [27] introduce natural corrupted visual and textual benchmarks on text-to-video retrieval. However, the composition of visual and textual modalities, which seeks to extend textual semantics and reasoning abilities to the visual domain, has not been discussed. In contrast, we consider the analysis of common corruption to both image and text – further studying the underlying model textual understanding –and take the first step to conduct an extensive analysis of robustness of deep neural networks in composed image retrieval.

Diagnostic analysis. Recently, several benchmarks for visual understanding have been introduced, including image captioning [28], visual question answering [29], and visio-linguistic composed reasoning [30]. For composed image retrieval, the benchmarks can be categorized into sythetic-based datasets by cubes [6] or natural scenes [18], fashion-based datasets [31], and open domain datasets [5]. Among them, the majority of the textual descriptions are limited to predefined attributes [6,31]. For this issue, FashionIQ [4] and CIRR [5] leverage the flexibility of natural language, producing the most widely used benchmarks in the fashion domain and open domain respectively. Based on CIRR, we developed a diagnostic dataset CIRR-D to probe specific text understanding across four fundamental scenarios: numerical variation, attribute variation, object removal, and background variation.

3. Robustness of composed image retrieval

3.1. Problem formulation

Given a reference image I_r and modified text T_m as input query, the aim of composed image retrieval is to retrieve the target image I_t from the gallery set $\{I_t^n\}_{n=1}^N$, where N is the number of images in the gallery set. CIR addresses fundamental cross-modal asymmetries where a single semantic word corresponds to thousands of images, while visually similar images may represent divergent semantics. Effective CIR requires three core capabilities: (1) Image representation to provide a precise anchor in the continuous visual space; (2) Text representation to provide subtle or significant differences between various visual contents, providing an unprecise target direction the model can generalize to; (3) Generalize modified text attributes to reference images to precisely predict the target visual content through the fusion of the vision and text modalities.

3.2. Definition of robustness in composed image retrieval

According to the foundation of composed image retrieval above, a robust model should demonstrate stable image feature extraction, text feature extraction, and modality fusion. In light of this, the robustness of composed image retrieval can be defined in twofold: *robustness against common corruption* for both text and image and *robustness in text understanding* for consistent reasoning between textual and visual modalities.

Definition 1 (*Robustness Against Common Corruption*)**.** A CIR model M is considered robust if the performance degradation under a set of visual corruptions C_v or textual corruptions C_t remains within an acceptable bound ϵ: $\Delta Perf(M, C_v, C_t) < \epsilon$ where $\Delta Perf$ denotes the relative performance change. To assess robustness against visual and textual corruptions, the evaluation includes 75 image corruptions-categorized into noise, blur, weather, and digital effects following [22]—and 35 textual corruptions divided into character-level and word-level variations.

Definition 2 (*Robustness in Text Understanding*)**.** A CIR model M is considered robust in text understanding if its performance degradation $\Delta Perf$ remains stable across semantic reasoning variations $\mathcal{V}$. Specifically, $\Delta\text{Perf}(M, v) < \epsilon$ for all $v \in \mathcal{V}$, where $\mathcal{V}$ encompasses:

- Numerical Variation: Changes in the quantity of objects.
- Attribute Variation: Changes in object properties like color, size, or material.
- Object Removal: Negation or removal of specific entities.
- Background Variation: Changes in the environmental context of the scene.

This is measured by evaluating whether the model maintains consistent retrieval accuracy when text instructions involve complex linguistic reasoning rather than simple descriptive alignment. For textual understanding, we evaluate linguistic reasoning through modified text selected based on specific keywords, categorized into numerical variation, attribute variation, object removal, and background variation.

4. Proposed benchmarks

We propose three new benchmarks for our composed image retrieval experiments based on two existing datasets: FashionIQ [4] in the fashion domain and CIRR [5] in the open domain. Both datasets include human-generated captions that distinguish image pairs. FashionIQ is a fine-grained dataset, with each image containing a single subject positioned centrally with a clean background. CIRR comprises real-life images extracted from NLVR2 , containing rich visual content in diverse backgrounds. As shown in Fig. 1, the benchmark is built and composed image retrieval models are evaluated on text and image common corruption robustness. Additionally, as shown in Fig. 4, the CIRR dataset is expanded and text understanding robustness is evaluated.

4.1. CIRR-C and FashionIQ-C

To evaluate the robustness of the composed image retrieval models against common corruption in both image and text, we create our robustness benchmark CIRR-C and FashionIQ-C with 75 visual corruptions and 35 textual corruptions. For visual corruption, 15 standard common corruptions are implemented following [22], categorized into noise, blur, weather, and digital, each with severities from 1 to 5. For textual corruption, 7 relevant corruptions are implemented following [43], including 4 character-level corruptions and 3 word-level corruptions. The definition of the textual corruptions are as follows:

- `Swap`: Randomly shuffles two characters within a word.
- `Qwerty`: Simulates errors made while writing on a QWERTY-type keyboard. Characters are swapped for their neighbors on the keyboard
- `RemoveChar`: Randomly removes characters from words.
- `RemoveSpace`: Removes a space from text, merging two words.
- `Misspelling`: Misspells words appearing in the Wikipedia list of commonly misspelled English words.
- `Repetition`: Randomly repeat words.
- `Homophone`: Changes words into their homophones from the Wikipedia list of common homophones. The list contains around 500 pairs or triples of homophonic words.

4.2. CIRR-D: Diagnostic benchmark

Following the current methods [5,8] reporting the results on the validation set, we expand and build our probing datasets CIRR-D based on the validation set of CIRR to pinpoint text understanding ability. We hypothesize that the model's corresponding reasoning capabilities can be evaluated when the modified text involves descriptions such as numbers, attributes, objects removal, or changing the background. In light of this, we build the triplets (reference image, modified text, and target image) for our diagnostic dataset according to the appearances of specific keywords in the modified text: "zero" to "ten", "number" for the numerical query; color, shape and size for attribute query, "remove" for object removal query; "background" for background variation query.

Construction pipeline. The construction of CIRR-D dataset involves four probing categories from three sources as follows: (1) Existing Validation Set of CIRR: Comprising 2297 images and 4181 triplets, this set is widely utilized. (2) Auxiliary captions of the CIRR validation set: Although supplied, these captions have not been used in the conventional evaluations. These captions highlight differences in removed content or background changes between image pairs, but they may not provide sufficient information to locate the target image precisely. Consequently, we manually eliminated triplets that resulted in an excessive number of target images. (3) Synthetic image generation through Visual ChatGPT : To integrate language reasoning with visual recognition, we augment the validation set by generating images through the process of image editing. The augmentation of the current distribution incorporates diverse variations in object quantity, color, shape, size, and existence. This can be regarded as a natural distribution shift occurring in real-world scenarios. To initiate this process, image captions for the CIRR validation set are generated by Visual ChatGPT. Subsequently, we create ten variants of the captions using ChatGPT, including four for numerical variants, three for color variants, two for size variants, and one for object removal. Afterward, leveraging the reference image and caption variants, Visual ChatGPT utilizes

Table 1
Details of the compared models. Models listed above the dashed line do not involve task-specific fine-tuning, whereas those below the line are fine-tuned for the target task.

Model	Venue	Params(M)	Base	Image encoder	Text encoder	Fusion
Pic2Word [32]	CVPR23	428.7	CLIP [9]	ViT-L/14	GPT-2 [33]	pseudo token
SEARLE [34]	ICCV23	441.8	CLIP [9]	ViT-L/14	GPT-2 [33]	pseudo token
TIRG [6]	CVPR19	30.7	–	ResNet50	LSTM	concat
MAAF [35]	Arxiv20	34.6	–	ResNet50	LSTM	transformer
CIRPLANT [5]	ICCV21	155.5	Oscar [36]	ResNet152	BERT [37]	transformer
ARTEMIS [7]	ICLR22	29.9	–	ResNet50	LSTM	attention
CLIP4CIR [8]	CVPR22	237.3	CLIP [9]	ResNet50 × 4	GPT-2 [33]	concat
FashionViL [38]	ECCV22	135.0	–	ResNet50	BERT [37]	transformer
BIBLIP4CIR [39]	WACV24	278.8	BLIP [40]	ViT-B/16	BERT [37]	concat
SPRC [10]	ICLR24	473.8	BLIP2 [11]	ViT-L/14	BERT [37]	pseudo token
VLM2Vec [41]	ICLR25	8300	Qwen2-VL [42]	ViT	Qwen2-7B	transformer

groundingDINO [44] for object detection, segment anything [45] for mask generation and stable diffusion [46] for target image generation. To ensure high dataset quality, we manually removed implausible images based on two criteria: consistency and local semantic alignment. First, regarding consistency, the generated image should modify only the regions explicitly specified by the text prompt while preserving all other elements of the source image. For background-change tasks, the foreground subjects must remain intact; for other localized edit tasks, the original background should be preserved without introducing unintended changes or hallucinated content. Second, regarding local semantic alignment. The applied modifications should be spatially localized and accurately reflect the semantics described in the text prompt. This includes verifying correct attribute binding (e.g., color or quantity) and ensuring that the edited regions do not exhibit structural distortions or visual artifacts that violate basic physical plausibility. (See Table 3.)

5. Experimental setting

Evaluation metrics. To evaluate the performance of models in composed image retrieval, we adopt the standard evaluation metric in retrieval, namely Recall@K denoted by R@K for short. Further, to measure robustness, we adopt relative robustness metrics $\gamma = 1 - (R_c - R_p)/R_c$ following the previous works [22,27], where R_c and R_p are the R@K under clean data and corrupted data, respectively. Additionally, in order to facilitate fair comparison among different models, we expand the codebase of [7] and established a unified testing platform for the convenient integration of various models. In detail, we set the gallery as the whole validation set as in [7,8], which includes more distractors and results in a larger need for discrimination, instead of setting the gallery the same as the query set as in [47,48]. Specifically for evaluating the fashionIQ dataset, we combine the two captions in a single query as [8,35] instead of combining the two modified captions in forward and reverse direction as [48]. All the evaluated models are trained in three categories jointly and tested individually for dress, shirt, and toptee categories. The reported results for fashionIQ are the average of the three categories.

Evaluated models. The proposed testbed integrates ten published composed image retrieval models detailed in Table 1, which are selected for their representativeness and reproducibility. These models can be categorized into the following overlapping categories: (1) Training-strategies view: Models above the dashed line are zero-shot, lacking task-specific fine-tuning, while models below are supervised with task-specific triplets. For the zero-shot models, we implement the three single-modality query models for comparison. Notably, the Image-only (CLIP) and Text-only (CLIP) share the same image and text encoders as CLIP4CIR. (2) Model base view: TIRG, MAAF and ARTEMIS are three initial models based on ResNet50 and LSTM, each with fewer than 50M parameters. Recent models are based on Vision-Language Models (VLM) including clip-based models (Pic2Word, SEARLE, CLIP4CIR) and blip-based models (BIBLIP4CIR on BLIP [40] and SPRC on BLIP2 [11]). (3) Modality fusion view: Late fusion includes concatenation-base fusion, light attention-based fusion and transformer-based fusion. Early fusion includes converting the reference image into a single pseudo token (Pic2Word, SEARLE) or multiple pseudo tokens (SPRC).

Implementation details. To ensure the fairness in evaluation, a standardized testbed is established for the compared models, unifying the evaluation process. Models or dataset selection can be easily managed through input parameters. Additional models can be integrated into our testbed by providing model structure files with the necessary interfaces. These interfaces include image feature extraction, text feature extraction, feature composition, and distance comparison. FashionViL [38] is tested in the fashion domain, CIRPLANT in the open domain, and all the remaining published models both domains. All the experiments are conducted with NVIDIA A100 GPUs.

6. Results and analysis

6.1. Robustness against common corruption

To evaluate the studied models through the lens of robustness against common corruptions, we conduct experiments involving 75 visual corruptions, further categorized into noise, blur, weather, and digital corruptions, on both the fashion domain and open domain. Table 2 presents the relative robustness γ under the highest severity of each common visual corruption, which shares the same trend across other corruption severities. To evaluate the robustness against textual corruption, experiments are conducted under 35 textual corruptions categorized as character-level and word-level shown in Table 4. More detailed results, including concrete value and performance on CIRCO and COCO under common corruptions, are provided in the Appendix.

Will textual modality help robustness? We explore this question by comparing retrieval using either a single visual modality or both visual and textual modalities. As shown in Fig. 3(b), we visualize the average relative robustness γ under visual corruptions and recall@10 under clean conditions. Among the ResNet50-based methods, models that leverage both visual and textual modalities tend to achieve superior recall performance but at the expense of robustness. However, in the case of CLIP-based methods, the multimodal CLIP4CIR achieves both superior recall and robustness compared to its corresponding single visual encoder, Image-only (CLIP). This can be attributed to the alignment of text and image embeddings in a unified space during the pretraining process for CLIP, whereas ResNet-based methods utilize unaligned visual and textual features Thus, we speculate that *features from a shared vision-language space can help improve robustness, while text features from independent spaces will damage the model robustness.*

Does Task-specific fine-tuning help robustness? Considering that text from an aligned space can boost robustness, we further evaluate the different ways of leveraging textual features: through text-image pairs and task-specific triplets. As shown in Table 2, we observe

Table 2
Relative robustness score for text-image composed retrieval under 15 natural image corruptions in CIRR-C Recall@10, FashionIQ-C Recall@10. Recall@10 performance under clean conditions on the left..

		Noise			Blur				Weather				Digital			
CIRR-C	Clean	Gauss.	Shot	Implu.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contr.	Elast.	Pixel	JPEG
Image-only(RN50)	50.4	0.57	0.55	0.58	0.68	0.28	0.82	0.45	0.38	0.34	0.64	0.86	0.20	0.48	0.76	0.88
Image-only(CLIP)	36.2	0.56	0.55	0.58	0.66	0.32	0.83	0.49	0.52	0.45	0.77	0.91	0.24	0.41	0.78	0.91
Text-only(CLIP)	51.2	0.79	0.76	0.81	0.85	0.29	1.0	0.55	0.65	0.70	0.89	1.0	0.19	0.40	0.96	1.0
Pic2word [32]	64.8	0.89	0.89	0.89	0.88	0.64	0.96	0.71	0.85	0.77	0.92	0.93	0.48	0.76	0.95	0.93
SEARLE [34]	67.2	0.89	0.89	0.88	0.91	0.66	0.98	0.71	0.85	0.79	0.93	0.97	0.43	0.77	0.98	0.97
TIRG [6]	55.1	0.34	0.36	0.34	0.48	0.21	0.70	0.43	0.31	0.22	0.40	0.70	0.12	0.47	0.74	0.84
MAAF [35]	49.9	0.50	0.49	0.50	0.62	0.26	0.80	0.41	0.36	0.31	0.50	0.74	0.11	0.48	0.83	0.87
ARTEMIS [7]	59.0	0.39	0.42	0.38	0.51	0.25	0.70	0.44	0.31	0.26	0.45	0.71	0.10	0.47	0.75	0.86
CIRPLANT [5]	68.8	0.70	0.69	0.71	0.77	0.28	0.89	0.51	0.44	0.43	0.66	0.88	0.17	0.56	0.85	0.92
CLIP4CIR [8]	80.3	0.68	0.68	0.69	0.77	0.28	0.90	0.52	0.55	0.60	0.80	0.91	0.16	0.39	0.91	0.92
BIBLIP4CIR [39]	79.1	0.66	0.65	0.65	0.87	0.28	0.92	0.49	0.64	0.63	0.78	0.93	0.2	0.53	0.88	0.85
SPRC [10]	88.9	0.94	0.93	0.94	0.94	0.67	0.99	0.75	0.87	0.83	0.95	0.98	0.44	0.77	0.99	0.98
VLM2VEC [41]	70.1	0.83	0.85	0.88	0.94	0.17	1.00	0.72	0.54	0.58	0.82	1.00	0.66	0.44	0.97	0.97

		Noise			Blur				Weather				Digital			
FashionIQ-C	Clean	Gauss.	Shot	Implu.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contr.	Elast.	Pixel	JPEG
Pic2word	24.7	0.61	0.61	0.59	0.57	0.38	0.79	0.61	0.54	0.54	0.66	0.73	0.38	0.43	0.85	0.82
SEARLE	25.6	0.64	0.64	0.63	0.62	0.37	0.85	0.63	0.56	0.56	0.69	0.78	0.39	0.44	0.87	0.92
TIRG	23.8	0.28	0.26	0.23	0.34	0.22	0.61	0.57	0.32	0.27	0.37	0.61	0.12	0.64	0.85	0.85
MAAF	23.4	0.31	0.27	0.25	0.44	0.21	0.67	0.53	0.29	0.24	0.31	0.54	0.13	0.54	0.83	0.83
ARTEMIS	24.9	0.24	0.24	0.20	0.38	0.26	0.65	0.60	0.36	0.25	0.38	0.55	0.14	0.63	0.86	0.87
FashionViL	23.4	0.26	0.28	0.25	0.40	0.31	0.82	0.67	0.33	0.31	0.34	0.70	0.15	0.86	1.09	1.06
CLIP4CIR	35.9	0.44	0.42	0.44	0.54	0.21	0.72	0.50	0.46	0.43	0.60	0.70	0.22	0.37	0.74	0.83
BIBLIP4CIR	31.3	0.44	0.37	0.37	0.62	0.31	0.73	0.57	0.44	0.45	0.65	0.71	0.21	0.55	0.70	0.71
SPRC	41.0	0.76	0.73	0.74	0.71	0.46	0.88	0.70	0.65	0.66	0.81	0.78	0.39	0.60	0.90	0.91
VLM2VEC	18.5	0.71	0.68	0.71	0.67	0.13	0.90	0.61	0.46	0.46	0.75	0.80	0.47	0.50	0.92	0.92

Table 3
Details of CIRR-D dataset. The first column is the number of images. The rest columns contain the number of triplets for four probing abilities.

	Images	Numerical	Attribute	Removal	Background
Val.	2297	820	1397	233	358
Extend caption	–	–	–	505	812
Synthetic	1245	986	3858	140	–
Total	3542	1806	5255	878	1170

that among the CLIP-based methods, zero-shot methods (Pic2Word, SEARLE), even without task-specific triplet training, exhibit higher robustness than CLIP4CIR, albeit at the expense of recall performance. This phenomenon also applies to ResNet-based models, where task-specific training methods demonstrate superior performance but lower robustness. These empirical observations suggest a trade-off between task-specific performance and robustness, where *fine-tuning on task-specific triplets tends to correlate with improved performance but reduced relative robustness.*

Do larger pre-training datasets bring higher robustness? As shown in Fig. 3(a), we visualize the performance of the compared models and the sizes of their pre-trained datasets. Among the models compared on CIRR-C, four models (TIRG, ARTEMIS, MAAF, and CIRPLANT) were pretrained on datasets with fewer than 10 million samples, whereas five other models were pretrained on datasets with over 100 million samples. We can observe a trend that the models with higher robustness and performance tend to be associated with larger pre-trained datasets. We expand the CLIP4CIR [8] with ViT-L/14 pretrained on LAION400M, LAION2B and DataComp1B respectively, as shown in Fig. 2. We observe that models are more robust with larger dataset sizes within the same dataset quality. Notably, the model pre-trained on DataComp1B, demonstrates superior robustness than the model pre-trained on LAION2B. This implies that models with large pre-trained datasets of the same quality result in better robustness against visual corruptions, aligning with the findings from Paul et al. [49]. Furthermore, *a model trained on a high-quality dataset can be more robust than one trained on a larger, noisier dataset.*

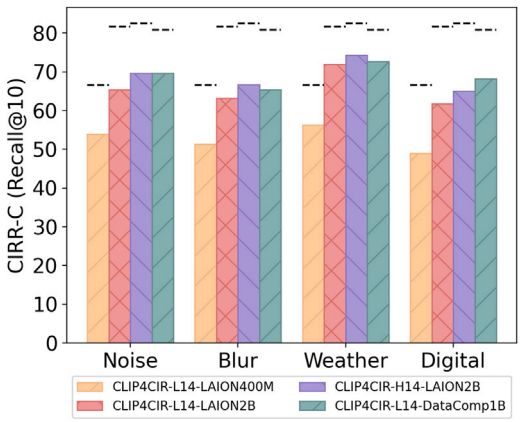
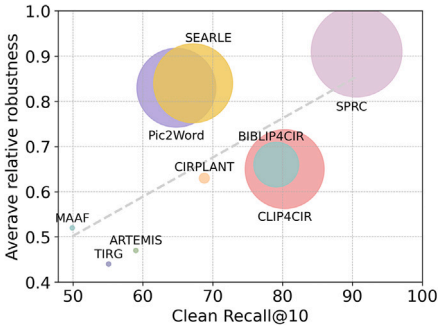


Fig. 2. Performance for visual corruptions on CIRR-C. Dashes are Recall@10 on clean and bars are on corrupted. Models are robust with large and high-quality datasets.

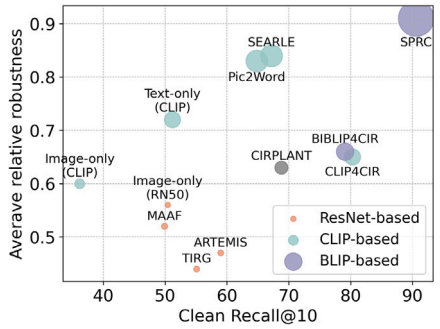
Coarse dataset vs. Fine-grained dataset. For both visual as shown in Table 2 and textual corruption as shown in Table 4, we observe that models under the FashionIQ-C dataset perform worse than those under the CIRR-C dataset. The FashionIQ-C dataset consists of clothing images with fine-grained differences, whereas the CIRR-C dataset is open-domain with more obvious content variations. This implies

Table 4
Relative robustness score for composed image retrieval under textual corruptions in CIRR-C and FashionIQ-C. Recall@10 performance under clean conditions is on the left. Models above the dashed line lack task-specific fine-tuning, while models below have it..

	CIRR-C			FashionIQ-C		
	Clean (R@10)	Character ($\overline{\gamma} \pm \sigma_\gamma$)	Word ($\overline{\gamma} \pm \sigma_\gamma$)	Clean (R@10)	Character ($\overline{\gamma} \pm \sigma_\gamma$)	Word ($\overline{\gamma} \pm \sigma_\gamma$)
Text-only (CLIP)	51.2	0.82 ± 0.11	0.96 ± 0.03	17.6	0.43 ± 0.09	0.56 ± 0.01
Pic2Word [32]	64.8	0.90 ± 0.06	0.97 ± 0.02	24.7	0.57 ± 0.07	0.67 ± 0.02
SEARLE [34]	67.2	0.92 ± 0.05	0.99 ± 0.02	29.1	0.59 ± 0.07	0.67 ± 0.02
TIRG [6]	55.1	0.83 ± 0.10	0.96 ± 0.05	23.8	0.35 ± 0.18	0.59 ± 0.05
MAAF [35]	49.9	0.97 ± 0.02	0.99 ± 0.01	23.4	0.47 ± 0.13	0.66 ± 0.03
ARTEMIS [7]	59.0	0.71 ± 0.17	0.93 ± 0.08	24.9	0.37 ± 0.20	0.63 ± 0.06
CIRPLANT [5]	68.8	0.95 ± 0.03	0.99 ± 0.01	–	–	–
CLIP4CIR [8]	80.3	0.92 ± 0.05	0.99 ± 0.01	35.0	0.57 ± 0.08	0.69 ± 0.01
FashionViL	–	–	–	23.4	0.65 ± 0.12	0.82 ± 0.04
BIBLIP4CIR [39]	79.1	0.91 ± 0.05	0.97 ± 0.02	31.3	0.72 ± 0.12	0.90 ± 0.01
SPRC [10]	88.9	0.95 ± 0.03	0.99 ± 0.01	41.0	0.63 ± 0.08	0.75 ± 0.02



(a) Circle size is pretrained dataset size.



(b) Circle size indicates model parameters.

Fig. 3. The average relative robustness versus the clean Recall@10 of compared models under visual corruptions in CIRR-C. Note that in this setting, only the image modality is perturbed, while the textual input remains fixed.

that *composed image retrieval is more sensitive to common corruption in fine-grained datasets than in coarse datasets.*

How to do modality composition for robust CIR? To better pinpoint the causes of low robustness in various model fusion modules, we compare TIRG, MAAF and ARTEMIS by isolating the fusion mechanism as the sole variable under a unified configuration of a ResNet50 image backbone and an LSTM text encoder. As shown in Table 2 in the open domain, MAAF demonstrates the most robust performance, while ARTEMIS performs the second best over TIRG. This trend is intrinsically linked to the increasing granularity of cross-modal interaction: while TIRG relies on basic concatenation-based fusion representing global-level merging, ARTEMIS introduces channel-wise attention for modality-specific gating, and MAAF utilizes a transformer-based multi-head attention mechanism to facilitate dense alignment between visual and textual features through self- and cross-attention. A similar correlation between structural complexity and robustness is observed in recent models, such as the transition from single pseudo-token representations in SEARLE [34] and Pic2Word [32] to the multiple-token interaction space in SPRC [10]. Therefore, we hypothesize that *more sufficient cross-modal interactions, such as cross-attention, can better promote robustness.*

To better pinpoint the structural factors influencing model robustness across various fusion modules, we conducted a comparative study on TIRG, ARTEMIS, and MAAF by isolating the fusion mechanism as the sole variable under a unified configuration of a ResNet50 image backbone and an LSTM text encoder. As reported in Table 2, a clear performance hierarchy emerges in the open domain, where MAAF demonstrates the most robust performance, followed by ARTEMIS and TIRG. This trend is intrinsically linked to the increasing granularity of cross-modal interaction: while TIRG relies on basic concatenation-based fusion representing global-level static merging, ARTEMIS introduces channel-wise attention for modality-specific gating, and MAAF utilizes a transformer-based multi-head attention mechanism to facilitate dense, token-level alignment between visual and textual features through self- and cross-attention. A similar correlation between structural complexity and robustness is observed in SOTA models, such as the transition from single pseudo-token representations in SEARLE and Pic2Word to the multiple-token interaction space in SPRC. These findings collectively support the hypothesis that more fine-grained and sufficient cross-modal interactions establish a more resilient semantic mapping, thereby serving as a decisive factor in promoting model robustness against complex logical perturbations.

6.2. Robustness in text understanding

The modified text allows users to conveniently describe their intentions regarding the relationship between reference and target images in natural language. In this section, the models' understanding of different text inputs is analyzed, and the performance of the CIRR-D is evaluated across various queries involving variations in numerical, attribute, object removal, and background. The quantitative performance for each query type is shown in Table 5. The qualitative results for vision-language-based models and single-modality query models are visualized in Fig. 4.

Do text and image modalities equally contribute in the composed query? As shown in Table 5, models show significant variations in performance across different query types. Specifically, VLM-based composed image retrieval models perform a clear trend of excelling at numerical and attribute queries, while showing a substantial drop of around 30% when changing to removal and background queries.

Table 5
R@5 of CIRR-D dataset. Models above the dash lack task-specific fine-tuning, while models below have it.

Model	Numerical	Attribute	Removal	Background
Image-only(CLIP)	24.20	27.36	27.90	25.64
Text-only(CLIP)	43.02	45.86	11.62	11.62
Pic2Word [32]	54.98	51.78	27.56	24.02
SEARLE [34]	55.15	53.63	29.54	33.93
TIRG [6]	36.82	33.68	30.41	32.82
MAAF [35]	32.50	33.53	31.09	34.27
ARTEMIS [7]	37.71	35.78	33.26	35.56
CLIP4CIR [8]	64.62	66.36	31.66	41.88
BIBLIP4CIR [39]	65.28	62.87	37.93	48.80
SPRC [10]	80.56	80.49	48.97	58.12

In contrast, image-only models, like RN50, show less variability and actually perform better on removal and background queries than on numerical and attribute ones. This suggests that *for composed queries: the text modality is more crucial for numerical and attribute queries, while the image modality is more important for removal and background queries.*

When will text modality be beneficial in the composed query? In this section, the reasons text modality excels in numerical and attribute variations yet underperforms in object removal and background changes are explored. For numerical and attribute variations, the modified text tends to be a description of the visual content of the target image, providing detailed discriminative information. Since VLM models are pre-trained on large-scale image-caption pairs (400M pairs for CLIP-based models and 130M for BLIP-based models), the descriptive-type modified text can be well-aligned with the target images in the feature space. However, for object removal, the modified text is an instruction to manipulate the reference image. As shown in Row 2 of Fig. 4, the Text-only model does not recognize the modification but is heavily weighted on the noun as a shortcut. For background variation, modifications are typically limited to changes in background color or blurring. Relying solely on these changes, as demonstrated by the Text-only model in Row 3 of Fig. 4, can lead to unrelated targets and cause distractions. It is suggested that *the modified text is beneficial when it accurately describes the target image. It may result in worse result when involving instructions to manipulate the reference image or lead to a larger number of satisfactory candidate images.*

How to do modality composition for better reasoning? Composed image retrieval models perform better when the modified text is a direct description but perform worse when it involves instruction reasoning. Proper interaction between the image and text modalities is crucial. Typical modality fusion methods include two-stream late fusion [6–8, 35,39] or single-stream early fusion [10,32,34]. For late fusion models, composed reasoning is conducted through the fusion module by training an additional adapter. In single-stream early fusion models, composed reasoning is facilitated by the pre-trained language model, which has reasoning capabilities. Among early fusion models, Pic2Word and SEARLE convert reference images into a single token, which limits sufficient interaction. In contrast, SPRC, which uses multiple pseudo tokens, shows better performance. Based on these results, we speculate that *early fusion models with sufficient cross-modality interaction and task-specific fine-tuning can better activate the reasoning capabilities of the language model, leading to improved reasoning performance.*

7. Conclusion

In this work, we proposed three robustness benchmarks for composed image retrieval in a testbed–including two for common corruption (in both images and text) and one for probing textual understanding. Specifically, we first introduced two benchmark datasets, CIRR-C and FashionIQ-C, with common corruption in the open and fashion domains respectively. Additionally, we created the benchmark dataset CIRR-D to assess textual understanding: including under variations to numerical object count, different attributes, object removal, and background change. Based on our results, we provide the following observations about enhancing model robustness in composed image retrieval: (1) Text features from an aligned space can help boost the robustness, while text features from independent spaces will reduce robustness, (2) Fine-tuning with task-specific datasets improves performance but may reduce robustness. (3) A model trained on a high-quality dataset can be more robust than one trained on a larger, noisier dataset. (4) Enhanced cross-modal interactions, such as cross-attention, can boost robustness. Additionally, in text understanding: (1) Modifying the textual query boosts discriminative ability when it minimizes the number of feasible targets (and will harm performance when producing more candidate responses). (2) Early fusion models with sufficient cross-modality interaction can better leverage the reasoning capabilities of the language model, leading to improved reasoning performance. We suggest that these findings have the potential to boost the robustness of composed image retrieval in future work.

8. Limitations and future work

We discuss the limitations and future work of the proposed benchmarks in this section. For benchmarking natural corruption in CIRR-C and FashionIQ-C, the method of simulating real-world corruption with the noise still has limitations. Our study focuses on a systematic empirical evaluation of existing methods rather than proposing new defense mechanisms. Exploring whether robustness-oriented techniques, such as adversarial training, can mitigate these issues remains a promising direction for future research. Furthermore, we acknowledge the potential for subjective bias in our data filtering process. Due to the large scale of the dataset, this filtering was conducted sequentially by two annotators without overlapping cross-validation. While a shared annotation standard was established and agreed upon beforehand to maintain consistency, the lack of formal inter-rater reliability metrics means that individual subjective judgment could still influence the sample selection. Some environmental corruptions (e.g., snow and fog) included in FashionIQ-C are less representative of studio-based fashion product photography. They are incorporated primarily to maintain a standardized corruption taxonomy across benchmarks, and their direct applicability to certain fashion e-commerce scenarios may therefore be limited.

9. License

All the models in this study are available to the public. The model code for TIRG [6] and MAAF [35] have the Apache License Version 2.0, ARTEMIS [7] has CC BY-NC-SA 4.0 License, CIRPLANT [5] has MIT license and FashionViL [38] has BSD License. We will provide CIRR-C, FashionIQ-C and CIRR-D publicly. These datasets are based on existing CIRR [5] and FashionIQ [4]. For CIRR-C and FashionIQ-C, we did not add any new images or text sources. For CIRR-D, we further generate synthetic images and text to expand the original CIRR dataset. All of these datasets are available to the public and we apply similar licenses to our testbed code and our proposed benchmarks.

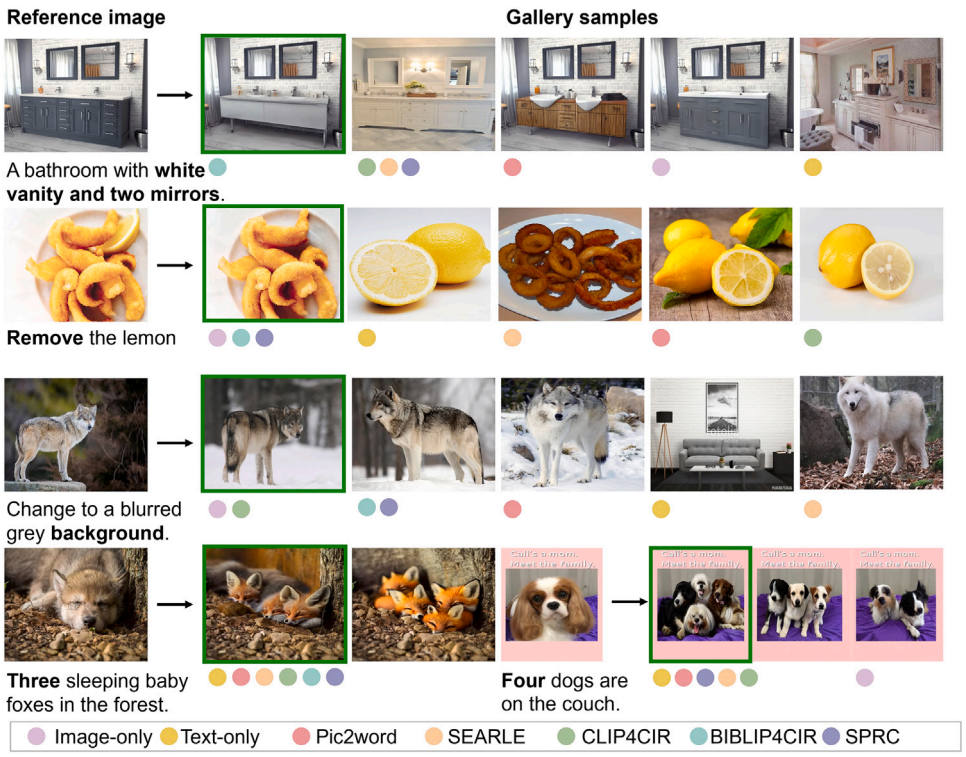


Fig. 4. Visualization of results for compared methods on the proposed CIRR-D benchmark probing textual understanding. The green boxes indicate the ground-truth target images. Other gallery images correspond to failure cases retrieved by different models, as indicated by the color-coded dots shown in the legend below. We observe that many failure cases retrieve images sharing the same noun category (e.g., object type) but violating the compositional constraint specified in the text, illustrating the tendency of models to rely on noun shortcuts. Images from: Row 1: CIRR-D with attribute variations; Row 2: CIRR-D with object removal; Row 3: CIRR-D with background variations; Row 4: CIRR-D with number variations.

CRediT authorship contribution statement

Shitong Sun: Writing – original draft, Methodology. **Qilei Li:** Validation, Formal analysis. **Shaogang Gong:** Supervision. **Weitong Cai:** Validation. **Philip Torr:** Supervision. **Jindong Gu:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the China Postdoctoral Science Foundation (Grant No. 2025M781597), the Hubei Provincial Natural Science Foundation of China (Grant No. JCZRMS202601596), the Fundamental Research Funds for the Central Universities, China (Grant No. XJ2026000901), and the Academy of Frontier Interdisciplinary Research at Central China Normal University. We thank the anonymous reviewers for their valuable comments and suggestions.

Appendix A. Control experiment on domain shift vs. Text understanding

To investigate whether performance differences on CIRR-D stem from domain shift introduced by synthetic images rather than text understanding, we conduct a control experiment over an aligned subset of 47 reference images shared across all three settings, using the same gallery throughout.

Experimental settings

- **Exp1 (Real vs. Real):** A subset from the original CIRR validation set, serving as the real-image baseline.
- **Exp2 (Real vs. Synthetic, Matched Variation):** The same reference images paired with synthetic target images whose semantic change direction matches the instruction (*e.g.*, both describe a numerical or attribute change of the same type).
- **Exp3 (Real vs. Synthetic, Other Variation):** The same reference images paired with synthetic target images of a different semantic change direction.

A.1. Results and analysis

We make two observations from the results reported in Table A.6. First, the Exp2 vs. Exp3 comparison reflects text understanding ability: larger models (CLIP4CIR, SPRC) show stable and consistent performance across both settings, while smaller models (ARTEMIS, MAAF) exhibit larger fluctuation, consistent with their generally weaker text comprehension observed in the main CIRR-D benchmark. Second, the Exp1 vs. Exp2 comparison serves as an approximation of domain shift: for most models, the performance gap between Exp1 and Exp2 is comparable in magnitude to that between Exp2 and Exp3, suggesting that the impact of domain shift is limited and does not account for the primary source of performance variation on CIRR-D.

Appendix B. Analysis of COCO-C and CIRCO-C

To evaluate the compared models on a more general domain, we implement our image corruptions on the validation set of COCO [50], represented by COCO-C. We set the masked bounding box as the

Table A.6
Recall@10 across three experimental settings on an aligned subset of 47 reference images.

Model	Exp1 Real vs. Real	Exp2 Real vs. Syn. (Matched)	Exp3 Real vs. Syn. (Other)
ARTEMIS	48.65	46.43	34.78
TIRG	40.54	45.54	45.45
MAAF	44.59	44.64	48.22
CLIP4CIR	75.68	87.50	91.30
SPRC	93.24	92.86	92.10

Table B.7
Relative robustness score for text-image composed retrieval under 15 natural image corruptions in COCO-C Recall@10.

COCO-C	Noise			Blur				Weather				Digital			
	Gauss.	Shot	Implu.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contr.	Elast.	Pixel	JPEG
TIRG [6]	0.19	0.21	0.14	0.42	0.25	0.62	0.58	0.35	0.21	0.51	0.89	0.05	0.40	0.40	0.72
ARTEMIS [7]	0.14	0.16	0.08	0.43	0.22	0.72	0.52	0.41	0.32	0.45	1.06	0.05	0.40	0.48	0.70
CLIP4CIR [8]	0.52	0.58	0.52	0.65	0.12	0.85	0.36	0.51	0.49	0.71	0.90	0.10	0.24	0.77	0.77

Table B.8
Relative robustness score for text-image composed retrieval under 15 natural image corruptions in CIRCO-C mAP@10. mAP@10 performance under clean conditions on the left..

CIRCO-C	Clean	Noise			Blur				Weather				Digital			
		Gauss.	Shot	Implu.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contr.	Elast.	Pixel	JPEG
Image-only (CLIP)	2.79	0.30	0.32	0.38	0.32	0.04	0.53	0.10	0.22	0.18	0.56	0.80	0.01	0.05	0.63	0.62
Text-only (CLIP)	2.51	0.60	0.54	0.59	0.53	0.08	0.88	0.45	0.45	0.41	0.6	0.79	0.06	0.11	0.84	0.80
Pic2word	8.50	0.65	0.70	0.67	0.78	0.19	0.90	0.42	0.60	0.38	0.65	0.83	0.21	0.29	0.84	0.77
SEARLE	15.1	0.64	0.69	0.64	0.70	0.18	0.86	0.34	0.52	0.37	0.61	0.81	0.10	0.30	0.86	0.81

Table C.9
Relative robustness score for text-image composed retrieval under 15 natural image corruptions in FashionIQ-C Recall@10 for dress, shirt, and toptee respectively.

	Noise			Blur				Weather				Digital			
FashionIQ-C Dress	Gauss.	Shot	Implu.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contr.	Elast.	Pixel	JPEG
TIRG [6]	0.21	0.18	0.17	0.37	0.22	0.64	0.58	0.33	0.25	0.35	0.63	0.12	0.62	0.82	0.85
MAAF [35]	0.30	0.24	0.22	0.42	0.19	0.65	0.56	0.28	0.21	0.32	0.58	0.10	0.54	0.78	0.81
ARTEMIS [7]	0.23	0.22	0.18	0.38	0.24	0.66	0.62	0.39	0.26	0.37	0.59	0.14	0.67	0.85	0.9
FashionViL [38]	0.21	0.22	0.23	0.38	0.34	0.84	0.72	0.29	0.29	0.3	0.79	0.13	0.88	1.1	1.1
CLIP4CIR [8]	0.44	0.38	0.44	0.54	0.24	0.74	0.52	0.41	0.36	0.55	0.68	0.16	0.42	0.75	0.82
BIBLIPCIR [39]	0.43	0.31	0.34	0.59	0.32	0.71	0.61	0.39	0.38	0.63	0.72	0.16	0.58	0.7	0.67
SPRC [10]	0.74	0.72	0.73	0.8	0.55	0.89	0.68	0.64	0.69	0.78	0.72	0.32	0.71	0.87	0.91
FashionIQ-C Shirt	Gauss.	Shot	Implu.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contr.	Elast.	Pixel	JPEG
TIRG [6]	0.33	0.32	0.27	0.28	0.20	0.57	0.54	0.32	0.28	0.37	0.51	0.15	0.60	0.86	0.81
MAAF [35]	0.33	0.30	0.27	0.46	0.20	0.67	0.50	0.30	0.27	0.34	0.47	0.16	0.57	0.84	0.79
ARTEMIS [7]	0.27	0.28	0.25	0.39	0.26	0.62	0.61	0.36	0.24	0.38	0.54	0.16	0.61	0.84	0.88
FashionViL [38]	0.29	0.34	0.26	0.38	0.26	0.77	0.6	0.33	0.32	0.37	0.63	0.17	0.83	1.09	1.02
CLIP4CIR [8]	0.47	0.48	0.45	0.50	0.18	0.65	0.48	0.51	0.50	0.65	0.71	0.27	0.31	0.69	0.82
BIBLIP4CIR [39]	0.46	0.42	0.4	0.62	0.29	0.72	0.51	0.51	0.51	0.7	0.71	0.26	0.54	0.68	0.72
SPRC [10]	0.8	0.79	0.81	0.75	0.53	0.91	0.7	0.66	0.71	0.84	0.77	0.61	0.66	0.89	0.94
FashionIQ-C Toptee	Gauss.	Shot	Implu.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contr.	Elast.	Pixel	JPEG
TIRG [6]	0.30	0.28	0.25	0.36	0.24	0.63	0.58	0.32	0.27	0.39	0.58	0.10	0.69	0.88	0.88
MAAF [35]	0.30	0.28	0.27	0.45	0.24	0.71	0.52	0.28	0.23	0.28	0.56	0.14	0.52	0.88	0.88
ARTEMIS [7]	0.21	0.23	0.18	0.37	0.28	0.68	0.57	0.33	0.25	0.38	0.53	0.13	0.66	0.88	0.82
FashionViL [38]	0.28	0.28	0.27	0.44	0.32	0.85	0.69	0.38	0.33	0.36	0.69	0.15	0.88	1.09	1.06
CLIP4CIR [8]	0.42	0.4	0.42	0.58	0.21	0.76	0.49	0.46	0.44	0.60	0.71	0.24	0.39	0.78	0.84
BIBLIP4CIR [39]	0.44	0.38	0.37	0.64	0.31	0.76	0.58	0.42	0.45	0.62	0.71	0.2	0.54	0.71	0.73
SPRC [10]	0.76	0.75	0.78	0.82	0.54	0.92	0.7	0.68	0.71	0.79	0.8	0.49	0.67	0.91	0.91

reference image, the raw image as the target image, and the labels of objects as modified text the following [32,51]. The three compared models are trained on the CIRR dataset and evaluated on the validation set of COCO with 5000 images. The results in Table B.7 show that the large pretrained model CLIP4CIR has higher robustness than smaller models TIRG and ARTEMIS, which follow the same conclusion in paper Section 5.1. Similarly, we evaluated the corrupted version of CIRCO (denoted as CIRCO-C). As shown in Table B.8, both zero-shot methods demonstrate comparable robustness against corruptions.

Appendix C. Subcategories analysis of FashionIQ-C

For detailed results in the FashionIQ-C dataset, we report the results on the three categories, namely dress, shirt and toptee respectively, as shown in Table C.9. Overall, a similar trend is observed across the three categories, with SPRC consistently exhibiting the highest relative robustness. In the shirt category, overall robustness tends to be slightly higher than in the dress and, toptee categories. According to the definition of relative robustness: $\gamma = 1 - \left(R_c - R_p\right)/R_c$ following [22],

lower recall performance under clean condition R_c will lead to higher relative robustness γ.

Data availability

Data will be made available on request.

References

[1] F. Liu, D. Chen, X. Du, R. Gao, F. Xu, MEP-3M: A large-scale multi-modal E-commerce product dataset, Pattern Recognit. 140 (2023) 109519.

[2] Z. Wu, B. Ma, Prototype-guided text-based person search on rich Chinese descriptions, Pattern Recognit. 169 (2026) 111781.

[3] S. Maldonado, R. Saltos, C. Vairetti, J. Delpiano, Mitigating the effect of dataset shift in clustering, Pattern Recognit. 134 (2023) 109058.

[4] H. Wu, Y. Gao, X. Guo, Z. Al-Halah, S. Rennie, K. Grauman, R. Feris, The fashion IQ dataset: Retrieving images by combining side information and relative natural language feedback, in: CVPR, 2021.

[5] Z. Liu, C. Rodriguez-Opazo, D. Teney, S. Gould, Image retrieval on real-life images with pre-trained vision-and-language models, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 2125–2134.

[6] N. Vo, L. Jiang, C. Sun, K. Murphy, L.-J. Li, L. Fei-Fei, J. Hays, Composing text and image for image retrieval-an empirical odyssey, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 6439–6448.

[7] G. Delmas, R.S. de Rezende, G. Csurka, D. Larlus, Artemis: Attention-based retrieval with text-explicit matching and implicit similarity, 2022, arXiv preprint arXiv:2203.08101.

[8] A. Baldrati, M. Bertini, T. Uricchio, A. Del Bimbo, Effective conditioned and composed image retrieval combining CLIP-based features, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 21466–21474.

[9] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.

[10] Y. bai, X. Xu, Y. Liu, S. Khan, F. Khan, W. Zuo, R.S.M. Goh, C.-M. Feng, Sentence-level prompts benefit composed image retrieval, in: The Twelfth International Conference on Learning Representations, 2024.

[11] J. Li, D. Li, S. Savarese, S. Hoi, Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023, arXiv preprint arXiv:2301.12597.

[12] Y. Xu, Y. Bin, J. Wei, Y.M. Yang, G. Wang, H.T. Shen, Multi-modal transformer with global-local alignment for composed query image retrieval, IEEE Trans. Multimed. 25 (2023) 8346–8357.

[13] Y. Xu, Y. Bin, J. Wei, Y. Yang, G. Wang, H.T. Shen, Align and retrieve: Composition and decomposition learning in image retrieval with text feedback, IEEE Trans. Multimed. 26 (2024) 9936–9948.

[14] Z. Wang, Z. Gao, Y. Yang, G. Wang, C. Jiao, H.T. Shen, Geometric matching for cross-modal retrieval, IEEE Trans. Neural Networks Learn. Syst. 36 (3) (2024) 5509–5521.

[15] Y. Xu, J. Wei, Y. Bin, Y. Yang, Z. Ma, H.T. Shen, Set of diverse queries with uncertainty regularization for composed image retrieval, IEEE Trans. Circuits Syst. Video Technol. 34 (10) (2024) 10494–10506.

[16] Z. Fu, X. Chen, J. Dong, et al., Multi-order adversarial representation learning for composed query image retrieval, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2021, pp. 1685–1689.

[17] S. Li, C. He, X. Liu, et al., Learning with noisy triplet correspondence for composed image retrieval, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025, pp. 19628–19637.

[18] G. Gu, S. Chun, W. Kim, H. Jun, Y. Kang, S. Yun, CompoDiff: Versatile composed image retrieval with latent diffusion, 2023, arXiv preprint arXiv:2303.11916.

[19] L. Wang, W. Ao, V.N. Boddeti, et al., Generative zero-shot composed image retrieval, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025, pp. 29690–29700.

[20] M. Levy, R. Ben-Ari, N. Darshan, D. Lischinski, Data roaming and quality assessment for composed image retrieval, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, (4) 2024, pp. 2991–2999.

[21] L. Ventura, A. Yang, C. Schmid, G. Varol, Covr: Learning composed video retrieval from web video captions, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, (6) 2024, pp. 5270–5279.

[22] D. Hendrycks, T. Dietterich, Benchmarking neural network robustness to common corruptions and perturbations, in: Proceedings of the International Conference on Learning Representations, 2019.

[23] F. Croce, M. Andriushchenko, V. Sehwag, E. Debenedetti, N. Flammarion, M. Chiang, P. Mittal, M. Hein, Robustbench: a standardized adversarial robustness benchmark, 2020, arXiv preprint arXiv:2010.09670.

[24] L. Shi, J. Zhang, Z. Ji, J. Bai, S. Shan, Real face foundation representation learning for generalized deepfake detection, Pattern Recognit. 161 (2025) 111299.

[25] X. Liu, Y. Gao, L. Zong, W. Liang, B. Xu, Guiding prototype networks with label semantics for few-shot text classification, Pattern Recognit. 164 (2025) 111497.

[26] L. Li, Z. Gan, J. Liu, A closer look at the robustness of vision-and-language pre-trained models, 2020, arXiv preprint arXiv:2012.08673.

[27] M. Chantry, S. Vyas, H. Palangi, Y. Rawat, V. Vineet, Robustness analysis of video-language models against visual and language perturbations, Adv. Neural Inf. Process. Syst. 35 (2022) 34405–34420.

[28] R. Shekhar, S. Pezzelle, Y. Klimovich, A. Herbelot, M. Nabi, E. Sangineto, R. Bernardi, Foil it! find one mismatch between image and language caption, 2017, arXiv preprint arXiv:1705.01359.

[29] J. Johnson, B. Hariharan, L. Van Der Maaten, L. Fei-Fei, C. Lawrence Zitnick, R. Girshick, Clevr: A diagnostic dataset for compositional language and elementary visual reasoning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2901–2910.

[30] T. Thrush, R. Jiang, M. Bartolo, A. Singh, A. Williams, D. Kiela, C. Ross, Winoground: Probing vision and language models for visio-linguistic compositionality, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5238–5248.

[31] X. Han, Z. Wu, P.X. Huang, X. Zhang, M. Zhu, Y. Li, Y. Zhao, L.S. Davis, Automatic spatially-aware fashion concept discovery, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1463–1471.

[32] K. Saito, K. Sohn, X. Zhang, C.-L. Li, C.-Y. Lee, K. Saenko, T. Pfister, Pic2word: Mapping pictures to words for zero-shot composed image retrieval, 2023, arXiv preprint arXiv:2302.03084.

[33] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI Blog 1 (8) (2019) 9.

[34] A. Baldrati, L. Agnolucci, M. Bertini, A.D. Bimbo, Zero-shot composed image retrieval with textual inversion, 2023, arXiv:2303.15247.

[35] E. Dodds, J. Culpepper, S. Herdade, Y. Zhang, K. Boakye, Modality-agnostic attention fusion for visual search with text feedback, 2020, arXiv preprint arXiv:2007.00145.

[36] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, et al., Oscar: Object-semantics aligned pre-training for vision-language tasks, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16, Springer, 2020, pp. 121–137.

[37] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.

[38] X. Han, L. Yu, X. Zhu, L. Zhang, Y.-Z. Song, T. Xiang, FashionViL: Fashion-focused vision-and-language representation learning, in: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV, Springer, 2022, pp. 634–651.

[39] Z. Liu, W. Sun, Y. Hong, D. Teney, S. Gould, 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, pp. 5753–5762.

[40] J. Li, D. Li, C. Xiong, S. Hoi, BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in: ICML, 2022.

[41] Z. Jiang, R. Meng, X. Yang, S. Yavuz, Y. Zhou, W. Chen, Vlm2vec: Training vision-language models for massive multimodal embedding tasks, 2024, arXiv preprint arXiv:2410.05160.

[42] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, et al., Qwen technical report, 2023, arXiv preprint arXiv:2309.16609.

[43] B. Rychalska, D. Basaj, A. Gosiewska, P. Biecek, Models in the wild: On corruption robustness of neural nlp systems, in: Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part III 26, Springer, 2019, pp. 235–247.

[44] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, et al., Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2023, arXiv preprint arXiv:2303.05499.

[45] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A.C. Berg, W.-Y. Lo, P. Dollár, R. Girshick, Segment anything, 2023, arXiv:2304.02643.

[46] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, 2021, arXiv:2112.10752.

[47] Y. Chen, S. Gong, L. Bazzani, Image search with text feedback by visiolinguistic attention learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3001–3011.

[48] S. Lee, D. Kim, B. Han, Cosmo: Content-style modulation for image retrieval with text feedback, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 802–812.

[49] S. Paul, P.-Y. Chen, Vision transformers are robust learners, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, 2022, pp. 2071–2081.

[50] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, Springer, 2014, pp. 740–755.

[51] A. Neculai, Y. Chen, Z. Akata, Probabilistic compositional embeddings for multimodal image retrieval, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4547–4557.