Benchmarking Robustness of Text-Image Composed Retrieval

Anonymous authors
Paper under double-blind review

Abstract

Text-image composed retrieval aims to retrieve a target image through a composed query, specified as an image coupled with some text describing desired modifications to the input image. It has recently attracted attention for leveraging both information-rich images and concise language to precisely express requirements for the target image. However, the robustness of these approaches against real-world corruptions or their understanding of text variations has never been studied. In this paper, we perform the first robustness study and establish three new diversified benchmarks for systematic analysis of text-image composed retrieval against natural corruptions in both vision and text, and further probe textural understanding. For natural corruption analysis, we introduce two new large-scale benchmark datasets, CIRR-C and FashionIQ-C for open domain and fashion domain respectively, both of which apply 15 visual corruptions and 7 textural corruptions. For textural understanding analysis, we introduce a new diagnostic dataset CIRR-D by expanding the original raw CIRR dataset with synthetic data, which contains modified text to better probe textual understanding ability including numerical variation, attribute variation, object removal, background variation, and fine-grained evaluation.

1 Introduction

Text-image composed retrieval, known as composed image retrieval or text-guided image retrieval, aims to retrieve an image of interest from gallery images through a composed query consisting of a reference image and its corresponding modified text. As a single word (e.g., 'dog') can correspond to thousands of images depicting dogs in various breeds, poses, and scenarios, language is considered discrete and sparse, while images are regarded as dense and continuous. Using both images and text as queries enables the effective utilization of the continuous and dense nature of images to accurately express the requirements while leveraging discrete and sparse text to bridge semantic gaps beyond what the images alone can capture. This method holds potential in a variety of real-world applications, including fashion domain e-commerce Han et al. (2022b; 2023); Goenka et al. (2022); Han et al. (2022a); Chen et al. (2020) and open domain internet search Liu et al. (2021); Baldrati et al. (2022); Gu et al. (2023); Saito et al. (2023). However, existing text-image composed retrieval methods mostly are primarily tested on clean data, while real world models may naturally encounter distribution shifts Wang et al. (2021c), such as text typos and image corruptions owing to weather changes. Furthermore, there is currently no analysis of whether the model understands the meaning of the text, rather than solely relying on finding correspondences with main objects as a shortcut to the text-image composed retrieval task. For example, with a source image of a dog and its modified text 'change to two dogs on the table', the model might retrieve the target image by merely recognizing the words 'dog' and 'table' without the ability of numerical counting. Whether text-image composed retrieval models are robust in real-world applications, where natural corruption exists in both images and text, remains unexplored. Additionally, the question of whether these models are robust across diverse textural understanding requirements remains an unexplored area in this domain.

In this work, we make the first attempt to evaluate the robustness of text-image composed retrieval by building three new large-scale robustness benchmarks on both fashion and open domains. We raise the following two questions: **Q1:** *How robust are text-image composed retrieval models to natural corruption including both visual and textual?* Further to evaluate the text understanding ability, we have the second question: **Q2:** *How robust are text-image composed retrieval models in text understanding?*

To address the first question, we introduce two benchmark datasets on text-image composed retrieval task. Based on two widely used datasets FashionIQ Wu et al. (2021) in the fashion domain and CIRR Liu et al. (2021) in the open domain, we propose our benchmark datasets, namely FashionIQ-C and CIRR-C. Both datasets incorporate 15 visual corruptions and 7 textual corruptions, providing a comprehensive evaluation of model robustness against natural corruptions in both images and text. To answer the second question, we introduce a new diagnostic dataset CIRR-D to probe text understanding abilities across five fundamental scenarios: numerical variation, attribute manipulation, object removal, background variation, and fine-grained variation. In detail, the diagnostic dataset is constructed through the generation of synthetic triplets via image editing using a reference image and various text captions. This includes variations of both main captions and extended captions of the existing CIRR validation set. Our experiments show that the new benchmarks we introduced are suitable for robustness analysis against natural corruption on both image and text, as well as for probing text understanding abilities.

Our contributions are: (1) We pioneer the analysis of the robustness of text-image composed retrieval methods against natural corruption (including visual and textual) and understanding against textual variations (including five foundamental variations). (2) We introduce three new large-scale benchmarks including two benchmark datasets (FashionIQ-C and CIRR-C) to evaluate robustness against natural corruption in both image and text, and one diagnostic benchmark CIRR-D to probe text understanding robustness. (3) We provide an empirical analysis and conduct extended experiments to validate our findings.

2 Related Works

Robustness analysis. Quantifying robustness aims to evaluate the model stability to defend against corruption including natural corruption Hendrycks & Dietterich (2019); Chantry et al. (2022); Wang et al. (2021b), adversarial attacks Croce et al. (2020); Wang et al. (2021a;b), or to probe certain ability such as logical reasoning Sanyal et al. (2022) and visual content manipulation Li et al. (2020a). Traditional works about robustness analysis mainly focus on single modality involving visual modality-based tasks like image classification Hendrycks & Dietterich (2019), face detection Dooley et al. (2022); textual modality based task like text classification Zeng et al. (2021) and audio modality based task like speech recognition Mitra et al. (2017). Recently, robustness analysis against 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. For example, Li et al. Li et al. (2020a) take the first step to systematically analyze the robustness of a multimodal task, Visual Question Answering (VQA), against 4 generic robustness including linguistic variation and visual content manipulation. However, it is limited to VQA tasks and doesn't introduce benchmarks to pinpoint sophisticated reasoning abilities. Schiappa et al. Chantry et al. (2022) introduce natural corrupted visual and textual benchmarks on text-to-video retrieval. However, the robustness analysis of the multimodal underlying hypothesis, which aims to generalize textual semantic and reasoning ability to visual space, is not discussed. We consider the analysis of both natural corruption in image and text and further underlying text understanding and take the first step to conduct an extensive analysis of the natural corruption and text understanding of the robustness of deep neural networks in text-image composed retrieval.

Diagnostic analysis. Recently, a range of benchmarks for visual understanding have been proposed, including datasets for image captioning Shekhar et al. (2017), visual question answering Johnson et al. (2017), visual reasoning Zerroug et al. (2022) and visio-linguistic compositional reasoning Thrush et al. (2022); Yuksekgonul et al. (2022); Ma et al. (2023). For text-image composed retrieval, the benchmarks can be categorized into sythetic-based datasets by cubes Vo et al. (2019) or natural scenes Gu et al. (2023), fashion-based datasets Han et al. (2017); Berg et al. (2010); Wu et al. (2021), object-state dataset Isola et al. (2015) and open domain dataset Liu et al. (2021). Among them, the majority of the textual descriptions are limited by predefined attributes Han et al. (2017); Vo et al. (2019); Isola et al. (2015). To overcome this limitation, FashionIQ Wu et al. (2021) and CIRR Liu et al. (2021) leverage the flexibility of natural language and becomes the most widely used benchmarks in fashion domain and open domain respectively. We expand and categorize the validation set of the CIRR benchmark using both its main annotation and previously unused extended annotations. This expansion allows us to probe specific text understanding in five foundamental scenarios: numerical variation, attribute variation, object removal, background variation, and fine-grained variation. Similar to ours, many diagnostic datasets are also synthetic. CLVER Johnson

et al. (2017) is a synthetic dataset to probe elementary vision reasoning including color, shape, and spatial relationships. CVR Zerroug et al. (2022) generates irregular shape, location, color, etc, and is designed for detecting the outliers from a small set of generated images. However, these are all simulated images and are not generated by imitating natural scenes. CasualVQA Agarwal et al. (2020) and CompoDiff Gu et al. (2023) generate images imitating natural scenes. However, CasualVQA is designed for visual question answering tasks and the generated images include noticeable artifacts. While ComoDiff is designed for text-image composed retrieval, it generates images by replacing only the objects (noun) rather than attributes (numerical, adjectival, manipulation instructions) like ours. As a result, it cannot precisely pinpoint the target reasoning abilities. To assess the compositional abilities, visio-linguistic compositional reasoning diagnostic benchmarks introduce variations in objects Ma et al. (2023); Yuksekgonul et al. (2022), attributes Ma et al. (2023); Yuksekgonul et al. (2022), or text order Yuksekgonul et al. (2022); Thrush et al. (2022). However, these benchmarks only provide image-text pairs for single-modality queries rather than image-text-image triplets for multi-modality queries. Additionally, their text composition Ma et al. (2023); Yuksekgonul et al. (2022) may lack corresponding images, such as in the case of 'grass eat horse'. In comparision, our CIRR-D introduces the first visio-linguistic compositional reasoning benchmark tailored for text-image composed retrivel task in natural scenes.

Text-image composed retrieval. Composed image retrieval aims to retrieve the target image, where the input query is specified in the form of an image plus other interactions, such as relative attributes Parikh & Grauman (2011), natural language Chen et al. (2020); Vo et al. (2019), and spatial layout Mai et al. (2017), to describe the desired modifications. Among them, natural language as the most pervasive interaction between humans and computers to convey intricate specifications has attracted increasing attention, which has often led '*composed image retrieval*' to become interchangeable with '*text-guided image retrieval*' in the literature. To provide clarity on the composition of the query, we term this task as text-image composed retrieval. Traditional text-image composed retrieval models implement separate independent image and text encoders, whose features are combined with late fusion. For example, TIRG Vo et al. (2019) and Artemis Delmas et al. (2022) implement separate pre-trained ResNet as image encoder and LSTM as text encoder. Until recently, with the power of unified multimodal space CLIP Radford et al. (2021), current text-image composed retrieval models achieved a noticeable improvement. For example, CLIP4CIR Baldrati et al. (2022) implements a light adapter as image-text late fusion and further tune it in target domains. Further based on CLIP, FAME Han et al. (2023) and CASE Levy et al. (2023) separately implement early cross attention between text and image, which shows obvious improvement.

3 Robustness Criteria for Text-Image Composed Retrieval

Foundation of text-image composed retrieval. Given a reference image I_r and modified text T_m as input query, the aim of text-image composed 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. In other words, text-image composed retrieval aims to retrieve the target visual content through dense continuous images guided by sparse discrete text. We discuss the concept of 'dense' and 'sparse' in the semantic space, where a single semantic word can correspond to thousands of images. Therefore, text-image composed query can overcome the limitations of singular modality image retrieval, where text-image retrieval suffers from the imprecise descriptions and unlimited correct targets, and image-image retrieval suffers from expression limitation without the ability to generalize to different visual content. In light of this, the foundational abilities of text-image composed retrieval are threefold: (1) Image representation to provide a precise anchor in the dense 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 sparse modified text attributes to dense reference images to precisely predict the target visual content through the fusion of vision and text modality.

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

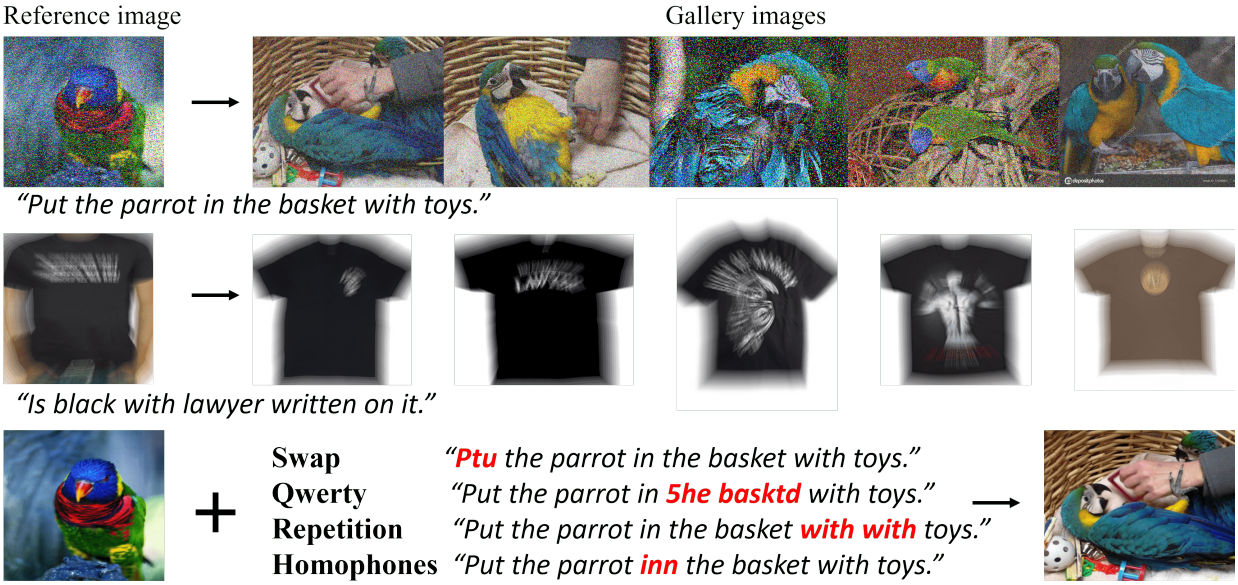

Figure 1: Sample visualization of proposed benchmarks under natural corruption with both visual and textual. **Top:** CIRR-C with impluse noise image corruption **Middle:** FashionIQ-C with zoom blur image corruption; **Bottom:** CIRR-C under character-level (Swap and Qwerty) and word-level (Repetition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

frequently encountered in real-life scenarios in both visual and textual domains. The evaluation involves 15 standard image corruptions categorized into noise, blur, weather, and digital following Hendrycks & Dietterich (2019). Additionally, we evaluate 7 text corruptions categorized into character-level and word-level variations. Furthermore, for robustness against textual understanding, we evaluate common linguistic reasoning by selecting modified text with specific keywords or gallery set, categorized into numerical variation, attributes variation, object removal, background variation, and fine-grained variation, respectively.

Evaluation metrics. To evaluate the performance of models in text-image composed retrieval, we adopt the standard evaluation metric in retrieval, namely Recall@K donated as 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 Chantry et al. (2022); Hendrycks & Dietterich (2019), 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 have expanded Delmas et al. (2022) and established a unified testing platform for the convenient integration of various models. In detail, we set gallery the whole validation set as in Delmas et al. (2022); Baldrati et al. (2022), which includes more distractors and results to higher discriminative requirement, instead of setting gallery the same as the query set as in Chen et al. (2020); Lee et al. (2021). Specifically for evaluating the fashionIQ dataset, we combine the two captions in a single query as Baldrati et al. (2022); Dodds et al. (2020) instead of combining the two modified captions in forward and reverse direction as Lee et al. (2021). 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 average of the three categories.

Evaluation datasets. We utilize three new benchmarks for our text-image composed retrieval experiments, which are generated based on two existing datasets: FashionIQ Wu et al. (2021) in the fashion domain and CIRR Liu et al. (2021) in the open domain. Both datasets include human-generated captions that distinguish image pairs. FashionIQ is based on the fashion domain containing 77,684 garment images, which can be divided into three categories: dress, shirt, and toptee. Each image in FashionIQ contains a single subject positioned centrally with a clean background. CIRR is composed of 21,552 real-life images extracted from NLVR2 Suhr et al. (2018), which contains rich visual content in diverse backgrounds. As shown in Figure. 1, we build our benchmark and evaluate text-image composed retrieval models on text and image

Figure 2: Sample visualization of proposed benchmarks probing text understanding. **Row 1:** CIRR-D with attribute variations; **Row 2:** CIRR-D with number variations; **Row 3:** CIRR-D with object removal. **Row 4:** CIRR-D with background variations; **Row 5:** CIRR-D with fine-grained variations.

natural corruption robustness. Further as shown in Figure. 2, we expand the CIRR dataset and evaluate text understanding robustness. The creation of our dataset can be found in supplementary **??** and more visualization of these three benchmarks can be found in the supplementary **??**.

Natural vision and text corruption. To evaluate the robustness of the text-image composed retrieval model against natural corruption in both image and text, we create our robustness benchmark CIRR-C and FashionIQ-C with 15 visual corruptions and 7 textual corruptions. For vision corruption, we follow Hendrycks & Dietterich (2019) to implement 15 standard natural corruptions which fall into four categories: noise, blur, weather, and digital, each having a severity from 1 to 5. For text corruption, we follow Rychalska et al. (2019) and implement the most related seven corruptions including four character-level corruptions and three word-level corruptions respectively.

Diagnostic dataset. Following the current methods Liu et al. (2021); Baldrati et al. (2022) 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; and the ability to deal with fine-grained variations can be evaluated

Table 1: Details of modality fusion for the evaluated models in this study. R_i, M_t, T_i represent reference image, modified text and target image feature respectively. C donates the composed of the two features. <,> represents cosine similarity. The compared methods include TIRG Vo et al. (2019), MAAF Dodds et al. (2020), ARTEMIS Delmas et al. (2022), CIRPLANT Liu et al. (2021), CLIP4CIR Baldrati et al. (2022), FashionViL Han et al. (2022b), Pic2word Saito et al. (2023) and SEARLE Baldrati et al. (2023).

Model	Image encoder	Text encoder	C_{RiMt}	C_{RiTi}	C_{MtTi}	Distance
TIRG	ResNet50	LSTM	cat+residual	-	-	$<C_{RiMt}, T_i>$
MAAF	ResNet50	LSTM	self attn+ cross attn	-	-	$<C_{RiMt}, T_i>$
Artemis	ResNet50	LSTM	dot product	dot product	dot product	$<C_{RiMt},C_{TiMt}> + <C_{TiMt},M_t>$
CIRPLANT	ResNet152	BERT Devlin et al. (2018)	OSCAR Li et al. (2020b)	-	-	$<C_{RiMt}, T_i>$
CLIP4CIR	ResNet50x4	GPT-2 Radford et al. (2019)	cat + residual	-	-	$<C_{RiMt}, T_i>$
FashionViL	ConvNet Huang et al. (2020)	BERT Devlin et al. (2018)	transformer	-	-	$<C_{RiMt}, T_i>$
Pic2word	ViT-L/14	GPT-2 Radford et al. (2019)	textual inversion	-	-	$<C_{RiMt}, T_i>$
SEARLE	ViT-L/14	GPT-2 Radford et al. (2019)	textual inversion	-	-	$<C_{RiMt}, T_i>$

when the gallery images are highly similar following Liu et al. (2021). 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 modified text: "zero" to "ten", "number" for numerical query; color, shape and size for attribute query, "remove" for object removal query; "background" for background variation query. The detailed statistic of the CIRR-D can be found in Table **??**. The construction of CIRR-D dataset involves five probing categories from three sources as follows: (1) Existing Validation Set of CIRR: Comprising 2297 images and 4181 triplets, this set is widely utilized. Each image includes a subset of 6 highly similar images as a gallery to enhance the detection of fine-grained discriminative ability. (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 removal content or background changes between image pairs, but they may not provide sufficient information to precisely locate the target image. Consequently, we manually eliminated triplets that resulted in an excessive number of target images. (3) Synthetic image generation through Visual ChatGPT Wu et al. (2023): 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 the 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 groundingDINO Liu et al. (2023) for object detection, segment anything Kirillov et al. (2023) for mask generation and stable diffusion Rombach et al. (2021) for target image generation. We conduct manual eliminated of implausible generated images, ensuring that our synthetic image pairs maintain the original background and only modify the specific areas mentioned in the text.

Evaluated models. We conduct experiments on six text-image composed retrieval models. The modality fusion of these approaches is summarized in supplementary Table 1 including Vo et al. (2019); Liu et al. (2021); Baldrati et al. (2022); Delmas et al. (2022); Han et al. (2022b); Dodds et al. (2020), which can be categorized into overlapped categories: (1) Large pretrained models: FashionViL, CIRPLANT, CLIP4CIR, whose pretrained dataset size are 1.35 million, 6.5 million, and 400 million image-text pairs respectively; (2) Multi-task model: FashionViL, which is pretrained with four tasks simultaneously; (3) Light attention-based methods: ARTEMIS; (4) Transformer-based models: MAAF, CIRPLANT, CIRPLANT and FashionViL; (5) Lightweight models: MAAF, TIRG, and ARTEMIS (all with ResNet50 image encoder and LSTM text encoder for fair comparison). We additionally design some methods with single-modality queries to gain better insights: (6) Single-modality models: (i) Image-only (RN50): Queried with images embedded by ResNet50, the same as evaluated TIRG, MAAF, ARTEMIS. (ii) Image-only (CLIP): Queried with images embedded by CLIP image encoder RN50x4, the same as evaluated CLIP4CIR. (iii) Text-only: Queried with text embedded using the CLIP text encoder. These methods were selected because the reproduced results match originally reported results. We test FashionViL Han et al. (2022b) in the fashion domain, CIRPLANT in the open domain, and all the rest published models in both the fashion domain and open domain.

Table 2: Relative robustness score for text-image composed retrieval under 15 natural image corruptions in CIRR-C Recall@10, FashionIQ-C Recall@10 and CIRCO-C mAP@10. Recall@10 performance under clean conditions on the left. **Bold** is the highest relative robustness for the compared composed retrieval methods: Pic2word Saito et al. (2023), SEARLE Baldrati et al. (2023), TIRG Vo et al. (2019), MAAF Dodds et al. (2020), ARTEMIS Delmas et al. (2022), FashionViL Han et al. (2022b), CIRPLANT Liu et al. (2021), CLIP4CIR Baldrati et al. (2022).

		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	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	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	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	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	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	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	**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**

		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

		Noise			Blur				Weather				Digital			
CIRCO-C	Clean	Gauss.	Shot	Implu.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contr.	Elast.	Pixel	JPEG
Image-only (CLIP)	2.79	0.3	0.32	0.38	0.32	0.04	0.53	0.10	0.22	0.18	0.56	0.8	0.01	0.05	0.63	0.62
Text-only (CLIP)	2.51	0.6	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	0.8
Pic2word	8.5	0.65	0.7	0.67	0.78	0.19	0.90	0.42	0.6	0.38	0.65	0.83	0.21	0.29	0.84	0.77
SEARLE	15.1	0.64	0.69	0.64	0.7	0.18	0.86	0.34	0.52	0.37	0.61	0.81	0.1	0.3	0.86	0.81

Evaluation settings. To ensure the fairness of the evaluation, we establish a standardized testbed for various models, excluding FashionViL, to unify the evaluation process. Additionally, to reproduce the original performance, we implement the official pre-trained weights for FashionViL and CLIP4CIR. We retrain the models and achieve similar results as reported for MAAF, TIRG, ARTEMIS, and CIRPLANT. In detail, we extend the existing ARTEMIS code framework to provide a convenient interface of different trained models, where TIRG and ARTEMIS are already implemented. For image input among these models, CIRPLANT is based on frozen ResNet152 pre-trained features while other models take raw images as input. We implement a frozen ResNet152 image encoder so that we can introduce corruption to the raw image directly. For all the 15 image corruptions, we apply the highest severity of corruption to observe obvious performance differences. Regarding text input among these models, TIRG, MAAF, and ARTEMIS build the vocabulary based on appearance words in target evaluation caption. In contrast, FashionViL, CIRPLANT, and CLIP4CIR implement their vocabulary from a large pretraining dataset. We implement textual corruption by directly modifying raw text. For a concise presentation, we report the results of FashionIQ by showing the average of the three categories: dress, shirt, and toptee.

Implementation details We supply 5 models (TIRG, MAAF, ARTEMIS, CIRPLANT and CLIP4CIR) in the same testbed to have a fair comparison with different benchmark datasets. The selection of models or datasets can be easily accomplished through input parameters. Further models can be implemented in our testbed by simply providing model structure files with the necessary interface. In detail, the necessary interfaces include image feature extraction, text feature extraction, feature composing process and distance comparison. Our testbed is currently compatible with two environments and five models. CIRPLANT is implemented with Python(3.1) and Pytorch(1.8.1). TIRG, MAAF, CLIP4CIR and ARTEMIS were implemented with Python(3.8) and Pytorch(2.0). All the experiments are conducted and tested on NVIDIA A100 GPUs. We will maintain our code for benchmarking and testbed open source.

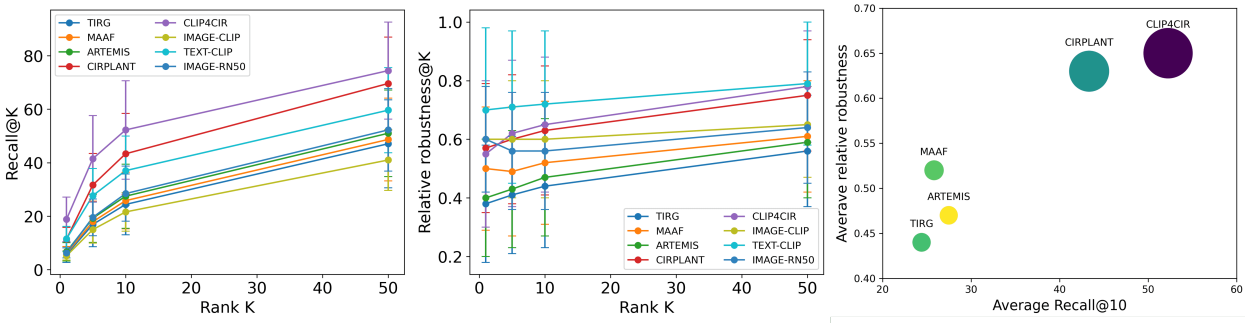

Figure 3: Models average performance in CIRR under 15 vision corruptions. Left: Recall vs. rank K. Middle: γ vs. rank K. Right: γ vs. recall@10, circle size indicates number of model parameters.

Table 3: Relative robustness score for text-image composed retrieval under 7 natural text corruptions in CIRR-C recall@10 and FashionIQ-C recall@10 on average of three categories. Recall@10 performance under clean conditions on the left. **Bold** are the highest relative robustness among the compared methods.

		Character				Word		
CIRR-C	Clean	Swap	QWERTY	RemoveC.	RemoveS.	Misspelling	Repetition	Homophone
Text-only(CLIP)	51.2	0.75	0.74	0.78	1.0	0.99	0.98	0.92
Pic2word Saito et al. (2023)	64.8	0.87	0.85	0.87	1.0	0.99	0.99	0.94
SEARLE Baldrati et al. (2023)	67.2	0.89	0.88	0.89	1.0	1.0	1.0	0.96
TIRG Vo et al. (2019)	55.1	0.77	0.76	0.80	1.0	0.98	1.0	0.89
MAAF Dodds et al. (2020)	49.9	**0.95**	**0.97**	**0.96**	**1.0**	**1.0**	**1.0**	**0.97**
ARTEMIS Delmas et al. (2022)	59.0	0.61	0.58	0.65	**1.0**	0.98	0.98	0.82
CIRPLANT Liu et al. (2021)	68.8	0.92	0.93	0.93	**1.0**	**1.0**	**1.0**	**0.97**
CLIP4CIR Baldrati et al. (2022)	**80.3**	0.89	0.89	0.90	**1.0**	**1.0**	0.99	**0.97**
		Character				Word		
FashionIQ-C	Clean	Swap	QWERTY	RemoveC.	RemoveS.	Misspelling	Repetition	Homophone
Pic2word Saito et al. (2023)	24.7	0.51	0.51	0.56	0.69	0.69	0.68	0.64
SEARLE Baldrati et al. (2023)	29.1	0.53	0.54	0.57	0.70	0.69	0.67	0.65
TIRG Vo et al. (2019)	23.8	0.26	0.20	0.29	0.66	0.63	0.61	0.52
MAAF Dodds et al. (2020)	23.4	0.40	0.39	0.39	0.70	0.68	0.68	0.62
ARTEMIS Delmas et al. (2022)	24.9	0.25	0.20	0.31	0.70	0.67	0.67	0.55
FashionViL Han et al. (2022b)	23.4	**0.55**	**0.59**	**0.60**	**0.86**	**0.84**	**0.85**	**0.76**
CLIP4CIR Baldrati et al. (2022)	**35.9**	0.52	0.51	0.54	0.71	0.70	0.69	0.67

4 Results and Analysis

4.1 Natural corruption analysis

To evaluate whether the text-image composed retrieval models are robust under natural corruptions, we conduct experiments involving 15 visual corruptions, further categorized into **noise**, **blur**, **weather**, and **digital** corruptions on both fashion domain and open domain. Table 2 presents the relative robustness γ under the highest severity of each natural visual corruption, which shares the same trend across other corruption severities. To evaluate the robustness against textual corruption, experiments are conducted under 7 textual corruptions categorized as **character-level** and **word-level**. More experimental results can be found in the supplementary **??**.

Pretraining. Among the compared models, FashionViL, CIRPLANT and CLIP4CIR are pretrained on large datasets, with respective sizes of 1.35 million, 6.5 million, and 400 million image-text pairs, while another three compared models are based on ImageNet pretrained ResNet50 as image encoder and random initialized LSTM as text encoder. As shown in Figure 3, the models with large pretrained datasets consistently show

better robustness in both open domain and fashion domain. *This implies that models with large pretrained datasets may result in better robustness against visual corruptions, which is in alignment with the statement from Paul et al. Paul & Chen (2022).*

Bottleneck of robustness. In Figure 3, we present visualizations of the recall performance with rank K improvement on the left and relative robustness on the right. With ranking K improves, both recall and relative robustness improve for all of models. Additionally, the ResNet50-based image-only search shows superior accuracy as well as robustness compared to TIRG, ARTEMIS, and MAAF, which share the same image encoder, ResNet50. According to the three foundations of text-image composed retrieval in Sec. 3, both the image encoder and modality fusion module can be vulnerable to vision corruption. We observe that the ResNet50 backbone shows relatively high robustness, while the modality fusion of TIRG, MAAF, and ARTEMIS models exacerbates the instability of the model. However, this observed phenomenon does not apply to the CLIP feature, whose text and image embedding are aligned in a unified space in the pretraining process. Comparing Image-only (CLIP) and CLIP4CIR, which query with CLIP image embedding and CLIP text-image composed embedding respectively, we can find out CLIP4CIR consistently performs better recall performance as well as robustness in Figure 3. *Thus, we speculate that text features from aligned space can help boost the robustness, while text features from independent space will damage the model robustness.*

Further to better pinpoint the vulnerability in variant model fusion modules, we compare TIRG, MAAF and ARTEMIS with the same text LSTM and image backbone ResNet50 but different fusion methods. As shown in Table 2 in open domain, MAAF shows the trend performing the most robust, while ARTEMIS performs the second best over TIRG. The modality fusion modules of these three models, TIRG, ARTEMIS, and MAAF are concatenation-based, light attention, and transformer, respectively. Among them, MAAF utilizes modality-agnostic attention which extracts word and image tokens to conduct thorough merge through self-attention and cross-attention. *We hypothesize that more sufficient cross-modal interactions, such as cross-attention, can better promote robustness.*

4.2 Textual robustness against natural corruption

Comparing the relative robustness against textual corruption in Table 3 and visual corruption in Table 2, we can observe that the robustness is higher against textual corruptions. Among the compared models, MAAF and FashionViL show the highest robustness in open domain and fashion domain respectively. This aligns with the findings we discovered regarding vision corruptions, where large pretrained models (FashionViL with fashion-specific pretraining) and models with sufficient modality fusion result to higher robustness. Additionally, comparing CLIP4CIR and the Text-only retrieval method implementing CLIP text embedding, we find out the robustness can be boosted after fusion with vision modality. However, with images corrupted, CLIP4CIR shows lower robustness than text-only model, from which *we speculate that aligned clean image feature can boost the robustness, while the corrupted image feature will impair robustness.*

4.3 Robustness against text understanding

In this section, we analyze model reasoning ability through variation of modified text on numerical variation, attributes variation, object removal, background variation, and fine-grained variation, which are provided by our proposed CIRR-D dataset. We evaluate the performance for each query type as shown in Table 4. To compare the model's understanding of different text inputs, we evaluate the performance on the CIRR-D gallery set across various queries involving variations in numerical, attribute, object removal, and background. To establish a baseline mixed of different query categories, we utilize the 4181 queries from the origin CIRR dataset and evaluate their performance in our extended CIRR-D dataset. This baseline includes diverse reasoning instructions and aims to represent the models' average performance across various types of instructions. The comparison between the baseline and the specific query types is regarded as the understanding ability of each specific type, involving variations in numerical aspects, attributes, object removal and background. However, unlike the above four categories, the gallery set for fine-grained is a subset, which is composed of six highly similar images following Liu et al. (2021). Detailed analyses are discussed below.

Table 4: Recall of CIRR-D dataset. The red and green arrows indicate the performance increase or decrease compared with CIRR queries. **Bold** and underline are the largest decrease and increase.

	R@5					Rsub@1
	CIRR	Numerical	Attribute	Removal	Background	Fine grained
Image-only(RN50)	31.55	31.47 $\downarrow$ (0.08)	32.57 $\uparrow$ (1.02)	35.99 $\uparrow$ (4.44)	39.15 $\uparrow$ (7.60)	20.25
Image-only(CLIP)	22.51	24.80 $\uparrow$ (2.29)	29.09 $\uparrow$ (6.58)	27.90 $\uparrow$ (5.39)	25.64 $\uparrow$ (3.13)	20.02
Text-only(CLIP)	39.02	42.84 $\uparrow$ (3.82)	49.45 $\uparrow$ (10.43)	11.62 $\downarrow$ (27.4)	11.62 $\downarrow$ (**27.4**)	53.73
TIRG Vo et al. (2019)	36.35	39.64 $\uparrow$ (3.29)	37.77 $\uparrow$ (1.42)	30.41 $\downarrow$ (5.94)	32.82 $\downarrow$ (3.53)	35.90
MAAF Dodds et al. (2020)	32.19	32.53 $\uparrow$ (0.34)	35.57 $\uparrow$ (3.38)	31.09 $\downarrow$ (1.10)	34.27 $\uparrow$ (2.08)	28.63
ARTEMIS Vo et al. (2019)	40.05	39.56 $\downarrow$ (0.49)	42.68 $\uparrow$ (2.63)	33.26 $\downarrow$ (6.79)	35.56 $\downarrow$ (4.49)	40.80
CIRPLANT Liu et al. (2021)	48.82	45.07 $\downarrow$ (**3.75**)	47.73 $\downarrow$ (**1.09**)	41.12 $\downarrow$ (7.70)	45.98 $\downarrow$ (2.84)	38.19
CLIP4CIR Baldrati et al. (2022)	62.94	64.18 $\uparrow$ (1.24)	69.15 $\uparrow$ (6.21)	31.66 $\downarrow$ (**31.28**)	41.88 $\downarrow$ (21.06)	62.66

Numerical variation. To probe the ability of numerical variation, the modified text contains either a precise value of the number from zero to ten or an estimated value by comparison like '*reduce/increase the number*'. Comparing numerical specific query with CIRR query as shown in Table 4, we can not observe significant variation which may result from the long-tailed distribution. Namely, the numerical set has a large number of samples in the range of 1 to 3, while a small number of samples in the range of 4 to 10. More analysis can be found in the supplementary **??**. For now, we speculate that *numerical modification may not be the bottleneck of the current text-image composed retrieval.*

Attributes variation. To evaluate the model's discriminative ability when querying elementary attributes, the modified text includes variations of color, shape and size. As observed in Table 4, all of the methods (except CIRPLANT) achieve higher performance with attribute queries than with CIRR queries. Additionally, the performance of CLIP based image-only model and CLIP4CIR have an obvious increment of over 6% compared with their performance with CIRR queries, which have a strong ability of attribute recognition including color, shape, and size. This implies that *attribute is the one of main focuses during training and models gain strong attribute discriminative ability.*

Object removal. Object removal is a convenient approach to describe the differences between images but is universally overlooked by current methods in text-image composed retrieval. To probe the ability of object removal through CIRR-D, the modified text of the query explicitly contains the word *'remove'*. As shown in Table 4, all of the five compared methods achieve their lowest performance in object removal with an average decrement of 10.6% compared with the CIRR query. In particular, CLIP4CIR has a drop of over 30%, which may be a result of its static pretraining process by aligning only image text pairs without comparison between images. Surprisingly, image-only methods can have an increment over CIRR queries, which illustrates that visual similarity can boost the robustness over object removal but the text condition over guidance the model decision. This aligns with the foundation of the task: images are dense and continuous while text is sparse and discrete. *In the case of object removal, text guidance expands the possibility of the targets, which distracts the model and results in lower performance.*

Background variation. To probe the robustness against background modifications, the modified text of the query explicitly includes the word "background". We observe a similar phenomenon as in object removal, where the performance of compared models (except MAAF) decreases but the performance of image-only models increases compared to CIRR queries. As the CIRR-D sample visualization in Figure 2, we can observe that the background modification method is limited such as changing the background color or making the background blur, which can lead to unrelated targets by relying solely on the text itself. We further speculate that *a modified text leading to more satisfactory candidates may result in impaired outcomes.*

Fine-grained variation. To probe the fine-grained variation discriminative ability, we utilize the subset in the CIRR dataset, where each image is retrieved from its subset composed of another five highly similar images. As the gallery is different from the above reasoning function, the recall cannot be compared with CIRR query performance directly. We can observe from Table 4 that image-only models perform similarly to random guessing and text-only by CLIP embedding can achieve an acceptable result. Among the five compared methods, TIRG achieve the lowest performance which hypothesizes the slight adjustment in visual space is sufficient rather than exploring text deeply and establishing a fusion space. This phenomenon

indicates that it is difficult to distinguish between two states in continuous visual space. In contrast, text can precisely define subtle differences due to its discrete nature. We also speculate that *a modified text offers accurate information while minimizing the number of feasible targets can enhance the model's discriminative ability.*

5 Conclusion

In this work, we proposed three robustness benchmarks for text-image composed retrieval including two for natural corruption in both image and text and one for probing textual understanding. Concretely, we first introduced two benchmark datasets, CIRR-C and FashionIQ-C with natural corruption (both image and text) in the open domain and fashion domain respectively. Further, we create benchmark CIRR-D to assess the text understanding including numerical, attribute, object removal, background, and fine-grained variation. Based on our observation, we provide the following suggestions to enhance model robustness in text-image composed retrieval: 1) Models pretrained on large datasets with little distribution shift will lead to better robustness, 2) Text features from an aligned space can help boost the robustness, while text features from independent space will damage the model robustness, 3) A modified text is more likely to enhance the model's discriminative ability when it minimizes the number of feasible targets and will distract the model when it leads to more satisfactory candidates. These findings have the potential to boost the robustness of text-image composed retrieval in the future.

References

Vedika Agarwal, Rakshith Shetty, and Mario Fritz. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9690–9698, 2020.

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21466–21474, 2022.

Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion, 2023.

Tamara L Berg, Alexander C Berg, and Jonathan Shih. Automatic attribute discovery and characterization from noisy web data. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part I 11*, pp. 663–676. Springer, 2010.

Madeline Chantry, Shruti Vyas, Hamid Palangi, Yogesh Rawat, and Vibhav Vineet. Robustness analysis of video-language models against visual and language perturbations. *Advances in Neural Information Processing Systems*, 35:34405–34420, 2022.

Yanbei Chen, Shaogang Gong, and Loris Bazzani. Image search with text feedback by visiolinguistic attention learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3001–3011, 2020.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.

Ginger Delmas, Rafael Sampaio de Rezende, Gabriela Csurka, and Diane Larlus. Artemis: Attention-based retrieval with text-explicit matching and implicit similarity. *arXiv preprint arXiv:2203.08101*, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Eric Dodds, Jack Culpepper, Simao Herdade, Yang Zhang, and Kofi Boakye. Modality-agnostic attention fusion for visual search with text feedback. *arXiv preprint arXiv:2007.00145*, 2020.

Samuel Dooley, George Z Wei, Tom Goldstein, and John Dickerson. Robustness disparities in face detection. *Advances in Neural Information Processing Systems*, 35:38245–38259, 2022.

Sonam Goenka, Zhaoheng Zheng, Ayush Jaiswal, Rakesh Chada, Yue Wu, Varsha Hedau, and Pradeep Natarajan. Fashionvlp: Vision language transformer for fashion retrieval with feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14105–14115, 2022.

Geonmo Gu, Sanghyuk Chun, Wonjae Kim, HeeJae Jun, Yoohoon Kang, and Sangdoo Yun. Compodiff: Versatile composed image retrieval with latent diffusion. *arXiv preprint arXiv:2303.11916*, 2023.

Xiao Han, Sen He, Li Zhang, Yi-Zhe Song, and Tao Xiang. Uigr: unified interactive garment retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2220–2225, 2022a.

Xiao Han, Licheng Yu, Xiatian Zhu, Li Zhang, Yi-Zhe Song, and Tao 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*, pp. 634–651. Springer, 2022b.

Xiao Han, Xiatian Zhu, Licheng Yu, Li Zhang, Yi-Zhe Song, and Tao Xiang. Fame-vil: Multi-tasking vision-language model for heterogeneous fashion tasks. *arXiv preprint arXiv:2303.02483*, 2023.

Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. Automatic spatially-aware fashion concept discovery. In *Proceedings of the IEEE international conference on computer vision*, pp. 1463–1471, 2017.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020.

Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1383–1391, 2015.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

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

Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Data roaming and early fusion for composed image retrieval. *arXiv preprint arXiv:2303.09429*, 2023.

Linjie Li, Zhe Gan, and Jingjing Liu. A closer look at the robustness of vision-and-language pre-trained models. *arXiv preprint arXiv:2012.08673*, 2020a.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu 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*, pp. 121–137. Springer, 2020b.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

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

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10910–10921, 2023.

Long Mai, Hailin Jin, Zhe Lin, Chen Fang, Jonathan Brandt, and Feng Liu. Spatial-semantic image search by visual feature synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4718–4727, 2017.

Vikramjit Mitra, Horacio Franco, Richard M Stern, Julien Van Hout, Luciana Ferrer, Martin Graciarena, Wen Wang, Dimitra Vergyri, Abeer Alwan, and John HL Hansen. Robust features in deep-learning-based speech recognition. *New Era for Robust Speech Recognition: Exploiting Deep Learning*, pp. 187–217, 2017.

Devi Parikh and Kristen Grauman. Relative attributes. In *2011 International Conference on Computer Vision*, pp. 503–510. IEEE, 2011.

Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2071–2081, 2022.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

Barbara Rychalska, Dominika Basaj, Alicja Gosiewska, and Przemysław 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*, pp. 235–247. Springer, 2019.

Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. *arXiv preprint arXiv:2302.03084*, 2023.

Soumya Sanyal, Zeyi Liao, and Xiang Ren. Robustlr: A diagnostic benchmark for evaluating logical robustness of deductive reasoners. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9614–9631, 2022.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. Foil it! find one mismatch between image and language caption. *arXiv preprint arXiv:1705.01359*, 2017.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2018.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022.

Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6439–6448, 2019.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. *arXiv preprint arXiv:2111.02840*, 2021a.

Xiao Wang, Qin Liu, Tao Gui, Qi Zhang, Yicheng Zou, Xin Zhou, Jiacheng Ye, Yongxin Zhang, Rui Zheng, Zexiong Pang, et al. Textflint: Unified multilingual robustness evaluation toolkit for natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pp. 347–355, 2021b.

Xuezhi Wang, Haohan Wang, and Diyi Yang. Measure and improve robustness in nlp models: A survey. *arXiv preprint arXiv:2112.08313*, 2021c.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. The fashion iq dataset: Retrieving images by combining side information and relative natural language feedback. *CVPR*, 2021.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2022.

Jiehang Zeng, Xiaoqing Zheng, Jianhan Xu, Linyang Li, Liping Yuan, and Xuanjing Huang. Certified robustness to text adversarial attacks by randomized [mask]. *arXiv preprint arXiv:2105.03743*, 2021.

Aimen Zerroug, Mohit Vaishnav, Julien Colin, Sebastian Musslick, and Thomas Serre. A benchmark for compositional visual reasoning. *arXiv preprint arXiv:2206.05379*, 2022.