# Invisible Relevance Bias: Text-Image Retrieval Models Prefer AI-Generated Images

Shicheng Xu[*]
CAS Key Laboratory of AI Safety,
Institute of Computing Technology,
Chinese Academy of Sciences
University of Chinese Academy of
Sciences, Beijing, China
xschit@163.com

Danyang Hou[*]
CAS Key Laboratory of AI Safety
Institute of Computing Technology,
Chinese Academy of Sciences
University of Chinese Academy of
Sciences, Beijing, China
houdanyang18b@ict.ac.cn

Liang Pang[†]
CAS Key Laboratory of AI Safety,
Institute of Computing Technology,
Chinese Academy of Sciences
Beijing, China
pangliang@ict.ac.cn

Jingcheng Deng
CAS Key Laboratory of AI Safety,
Institute of Computing Technology,
Chinese Academy of Sciences
University of Chinese Academy of
Sciences, Beijing, China
dengjingcheng23s@ict.ac.cn

Jun Xu
Gaoling School of Artificial
Intelligence, Renmin University of
China
Beijing, China
junxu@ruc.edu.cn

Huawei Shen
CAS Key Laboratory of AI Safety,
Institute of Computing Technology,
Chinese Academy of Sciences
Beijing, China
shenhuawei@ict.ac.cn

Xueqi Cheng
CAS Key Laboratory of AI Safety, Institute
of Computing Technology, Chinese
Academy of Sciences, Beijing, China
cxq@ict.ac.cn

## ABSTRACT

With the application of generation models, internet is increasingly inundated with AI-generated content (AIGC), causing both real and AI-generated content indexed in corpus for search. This paper explores the impact of AI-generated images on text-image search in this scenario. Firstly, we construct a benchmark consisting of both real and AI-generated images for this study. In this benchmark, AI-generated images possess visual semantics sufficiently similar to real images. Experiments on this benchmark reveal that text-image retrieval models tend to rank the AI-generated images higher than the real images, even though the AI-generated images do not exhibit more visually relevant semantics to the queries than real images. We call this bias as **invisible relevance bias**. This bias is detected across retrieval models with different training data and architectures. Further exploration reveals that mixing AI-generated images into the training data of retrieval models exacerbates the invisible relevance bias. These problems cause a vicious cycle in which AI-generated images have a higher chance of exposing from massive data, which makes them more likely to be mixed into the training of retrieval models and such training makes the invisible

relevance bias more and more serious. To mitigate this bias and elucidate the potential causes of the bias, firstly, we propose an effective method to alleviate this bias. Subsequently, we apply our proposed debiasing method to retroactively identify the causes of this bias, revealing that the AI-generated images induce the image encoder to embed additional information into their representation. This information makes the retriever estimate a higher relevance score. We conduct experiments to support this assertion.

Findings in this paper reveal the potential impact of AI-generated images on retrieval and have implications for further research. Code is released at https://github.com/xsc1234/Invisible-Relevance-Bias.

## CCS CONCEPTS

• **Information systems** → **Retrieval models and ranking**.

## KEYWORDS

Text-Image Retrieval, AIGC, Bias and Fairness

---

[*]Equal Contributions
[†]Corresponding author

## 1 INTRODUCTION

With the advancement of generation models, the quality of AI-generated content (AIGC) has been increasingly improved [7, 39]. The utilization of AI for content generation has transformed the
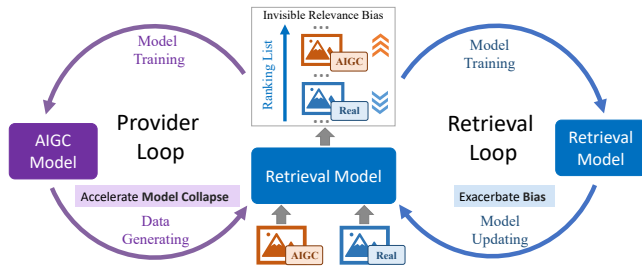
**Figure 1: Bias found in our paper. IR models tend to rank AI-generated images higher than real images even though they have very similar semantics. This bias increases the likelihood of the generated images being exposed from massive data of internet, which makes them more likely to be mixed into the training of AIGC and retrieval models, leading to more serious bias and forming a vicious cycle.**

way of content creation. It not only reduces the cost of content generation but also enhances efficiency, leading to a rapid influx of large amounts of AI-generated content onto the internet [2, 12].

Information retrieval (IR) is an important way for people to obtain the target information from massive data [24]. However, the rapid proliferation of AI-generated content (AIGC) presents a significant new challenge to IR: As the internet becomes increasingly inundated with AI-generated content, the corpus for search contains both real and AI-generated content, so, how will AI-generated content influence the ranking of search? In response to this challenge, [12] conducted a study on text modality and found that neural information retrieval models exhibit a preference for texts generated by Large Language Models. They refer to this category of biases in neural retrieval models towards the LLM-generated text as the **source bias**. However, beyond textual information, the internet is replete with a substantial number of images, serving as crucial sources for IR systems. Based on this, our paper extends the investigation of source bias in AI-generated content to text-image retrieval models. A pivotal question emerges: What impact will AI-generated images have on existing text-image retrieval models?

A crucial distinction from research on text modality lies in the methodology for **constructing a benchmark to simulate retrieval scenarios comprising both AI-generated and real images for IR models (§ 2).** A reasonable scenario for assessing the potential bias requires that the generated images and the real images have sufficiently similar visual semantics. This can avoid increasing or decreasing some semantic associations between the generated images and the query caused by the image generation. In other words, the IR model preferring (or rejecting) an AI-generated image that is more (or less) semantically relevant to the query than the real image cannot conclusively prove the existence (or nonexistence) of the bias. Study on text modality prompts LLMs to rewrite each real text in the corpus, thereby constructing an LLM-generated text for each real text without introducing additional semantic information. However, this approach is unsuitable for constructing the data in image modality. It is because we find mainstream image generation models such as diffusion models [27] are not good at following prompts to rewrite images while maintaining semantics.

Besides, the rewriting paradigm diverges from the prevalent content generation scenario that receives a description as a prompt and generates the texts or images. To solve the above problem, we propose an image over-sampling and selection strategy based on the merged caption (§ 2.1). Experimental results and human evaluation show that our proposed method can successfully construct semantically similar AI-generated images for real images (§ 2.2). The mixture of these images provides an effective simulation benchmark for investigating text-image retrieval models within scenarios featuring a mix of AI-generated and real images.

Our experiments on the constructed benchmark reveal that **text-image retrieval models tend to rank the AI-generated images higher than the real images, even though the AI-generated images do not exhibit more visually relevant semantics to the queries than real images (§ 3.3).** We define this bias as **invisible relevance bias** introduced by AI-generated images. This bias widely exists in retrieval models with different training data and architectures, including models trained from scratch and models pre-trained on massive image-text pairs, and encompassing dual-encoder and fusion-encoder retrieval models.

Another important point that is not considered in previous work is that AI-generated content does not only have an impact on the inference stage of the retrieval models. Due to the wide distribution of AI-generated images on the internet, they are very likely to be mixed into the training data of retrieval models in the future. Therefore, we further mix the AI-generated images into the training data of the retrieval model and explore the rules of its impact on the retrieval models by adjusting the mixing ratio. Our experiments show that **as the mixing ratio of AI-generated images in training data increases, the invisible relevance bias becomes more serious (§ 3.4)**, and the retrieval model exhibits a greater inclination to rank generated images higher. This phenomenon reveals a vicious cycle (Figure 1): Invisible relevance bias causes the generated images to have a higher chance of being obtained from massive data, which makes them more likely to be mixed into the training of generation and retrieval models and causes the model collapse [29], leading to more serious bias and forming a cycle.

To mitigate the bias, **we propose an effective training method for debiasing (§ 4).** Specifically, we introduce a contrastive loss between real and AI-generated images during training. This contrastive loss measures the invisible relevance bias introduced by the AI-generated images for retrieval models. Optimization of this loss can reduce the invisible relevance bias, thereby alleviating the preference to AI-generated images in text-image retrieval models. In addition, we also introduce the sampling probability to enable our debiasing method to dynamically adjust the tolerance to AI-generated images of text-image retrieval models.

Last but not least, we use our proposed debiasing method to reversely determine that **the causes of invisible relevance bias is the generated images causing the image encoder to embed additional information into their representation (§ 5)**. This additional information has a certain consistency in different generated images and can make the retriever estimate a higher relevance score. We also design experiments to support this.

The main contributions of this paper can be concluded as:

(1) We reveal the potential impact of AI-generated images on the ranking results of text-image retrieval systems in the context of the

rapid development of AIGC. To reasonably study the impact, we construct a reasonable benchmark to simulate text-image retrieval in scenarios consisting of both real and AI-generated images.

(2) Experimental results show that AI-generated images introduce the **invisible relevance** to text-image retrieval models, causing the bias that retrieval models prefer ranking AI-generated images higher even though the AI-generated images do not exhibit more visually relevant semantics to the queries than real images. This bias widely exists in retrieval models with different training data and architectures. We also reveal that the loop formed by training and inference causes the retrieval model to fall into a vicious cycle of increasingly serious bias.

(3) We propose an effective method to mitigate the bias by measuring and optimizing the invisible relevance bias introduced by AI-generated images in the training of retrieval models.

(4) We determine the causes of invisible relevance bias is that AI-generated images make the image encoder in the retriever embed additional information to their representations. This additional information is invisible but can amplify the relevant information to get a higher relevance score.

## 2 BENCHMARK CONSTRUCTION

This paper focuses on revealing the potential impact of AI-generated images on the ranking results of text-image retrieval systems. The prerequisite for all this is to construct a retrieval corpus consisting of real images and AI-generated images. This corpus should meet four critical requirements. ($\mathcal{R}$-1) Firstly, from the perspective of fair assessment, a reasonable scenario for assessing the potential bias requires that the generated images and the real images must have sufficiently similar visual semantics. This can avoid increasing or decreasing some semantic associations between the generated images and the queries caused by image generation. That is, the IR model preferring (or rejecting) an AI-generated image that is more (or less) semantically relevant to the query than the real image cannot prove the existence (or nonexistence) of the bias. ($\mathcal{R}$-2) Secondly, retrieval performance on the corpus containing only generated images should not change significantly compared to retrieval performance on real images only. This can further ensure that no additional visual semantics relevant (or irrelevant) to the query are introduced during the image generation. ($\mathcal{R}$-3) Thirdly, the image generation pipeline should be in line with the most common scenario of content generation that receives a description as prompt and generates the texts or images, so that the obtained bias assessment results can be more consistent with the actual scene. ($\mathcal{R}$-4) Fourthly, the number of generated images and real images should be equal to achieve fair comparison.

### 2.1 Image Generation

We select two widely used text-image retrieval datasets including Flickr30k [25] and MSCOCO [23] as the basis for our benchmark construction. In these two datasets, each image is annotated with five captions that describe the content of the image. For each real image, we aim to generate a corresponding AI-generated image. We propose four image generation methods and use the above four requirements to select the optimal image generation method. Details will be introduced in the following.

***Generation Based on Single Caption.*** In this method, we randomly select one caption from the five captions of each real image and use the selected caption as the prompt of stable diffusion model[1] to get the AI-generated image corresponding to the real image.

***Generation Based on Merged Caption.*** Since a single caption may not contain the complete visual semantics of the image, we propose to merge five captions to obtain a relatively complete semantic description of the image. Specifically, for each real image ($I_r$), we use *gpt-3.5-turbo* to combine the five captions to get a newly merged caption $C_m$. And input the merged caption to a stable diffusion model **M** to generate the image ($I_g$). The intention of this design is that different captions of an image can be the descriptions from different perspectives of the image. Therefore, merging these captions can obtain an overall description of the image, thereby making stable diffusion generate images that have sufficiently similar visual semantics to the real image.

***Generation Based on Image Rewriting.*** In addition to the merged caption $C_m$, the real image $I_r$ is also used as the initial image to condition the generation of the new image. Generation constrained by the input real image can output the generated image that has higher similarity to the real image. But it also requires the generation model to have higher multi-modal understanding ability.

***Generation Based on Image Over Sampling and Selection.*** To further narrow the semantic similarity between the generated images and real images, we propose an image over-sampling strategy. Specifically, for a real image $I^r$ we use stable diffusion to perform multiple times generation with different random seeds and get $n$ generated images $\{I_1^g, I_2^g, ..., I_n^g\}$. Then, we use the vision encoder $v(\cdot; \theta)$ of a powerful open source pre-trained vision-language model[2] to get the embedding $e_r$ for $I^r$ and set of embeddings $E = \{e_1^g, e_2^g, ..., e_n^g\}$ for $\{I_1^g, I_2^g, ..., I_n^g\}$. We calculate the cosine similarity between each embedding in $E$ and $e_r$ to get the similarity between the generated images and the real image $I^r$. Finally, we select the generated image with the maximum similarity between $I^r$ as the final generated image $I^g$ for the real image.

***Comparison between Different Generation Methods.*** The comparison between different generation methods under the above-mentioned four requirements is shown in Table 1. Taking these results into account, we choose generation based on merged caption and image over-sampling and selection as the image generation method to construct our benchmark. Using this method, we can get an AI-generated image for each real image in Flicker30k and MSCOCO. In the following content, Flicker30k and MSCOCO indicate the datasets that only contain real images, Flicker30k (AI) and MSCOCO (AI) indicate the datasets that only contain AI-generated images, Flicker30k+AI and MSCOCO+AI indicate the datasets that consist of both real and AI-generated images.

### 2.2 Assessing the Quality of Benchmark

Ensuring sufficiently similar visual semantics between generated and real images to avoid increasing or decreasing relevance to the query is a prerequisite for assessing the bias. In this section, we propose two methods to evaluate this. Firstly, we perform retrieval on

---

[1]https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0
[2]https://huggingface.co/laion/CLIP-ViT-H-14-laion2B-s32B-b79K

**Table 1: Compare generation methods based on the four requirements outlined in Section 2. Similarity is the cosine between embeddings of real and generated images encoded by OpenClip. Retrieval Performance is the difference in retrieval performance of BEIT-3 on the corpora only containing generated images and only containing real images respectively.**

| | Similarity ($\mathcal{R}$-1) | Retrieval Performance ($\mathcal{R}$-2) | Generation Pipeline ($\mathcal{R}$-3) | Ratio of Number ($\mathcal{R}$-4) |
|---|---|---|---|---|
| Single Caption | 0.5275 | \|△NDCG@5\| = 10.8 | text to image (✓) | 1:1 (✓) |
| Merged Caption | 0.5348 | \|△NDCG@5\| = 3.72 | text to image (✓) | 1:1 (✓) |
| Merged Caption + Image Rewriting | 0.5503 | \|△NDCG@5\| = 7.22 | text-image to image (×) | 1:1 (✓) |
| Merged Caption + Image Rewriting + Over-Sampling & Selection | 0.5845 | \|△NDCG@5\| = 5.31 | text-image to image (×) | 1:1 (✓) |
| Merged Caption + Over-Sampling & Selection | 0.6470 | \|△NDCG@5\| = 1.44 | text to image (✓) | 1:1 (✓) |

**Table 2: Retrieval performance (NDCG) on the corpora that only contain real images and only contain AI-generated images. Performance on AI-generated images is not significantly better than the performance on real images can demonstrate the success of our benchmark construction.**

| | | Flicker30k or Flicker30k (AI) | | | MSCOCO or MSCOCO (AI) | | |
|---|---|---|---|---|---|---|---|
| | | N@1 | N@3 | N@5 | N@1 | N@3 | N@5 |
| FLAVA | Real | 38.70 | 45.72 | 47.26 | 36.65 | 44.45 | 46.27 |
| | AI | 41.59 | 47.45 | 48.72 | 36.49 | 43.29 | 45.17 |
| ALIGIN | Real | 45.43 | 50.62 | 51.65 | 38.13 | 44.71 | 46.33 |
| | AI | 43.47 | 49.38 | 50.42 | 36.34 | 43.32 | 45.06 |
| BEIT-3 | Real | 47.45 | 52.15 | 52.87 | 41.24 | 47.16 | 48.63 |
| | AI | 45.31 | 50.49 | 51.43 | 38.33 | 44.76 | 46.19 |

the corpora that only contain real images and only contain generated images and count the retrieval performance respectively. The intention for this is that if the generated images have more semantics relevant to their corresponding queries than the real images, distinguishing images becomes easier, resulting in significantly higher retrieval performance on the generated images corpus than on the real images corpus. Secondly, we introduce human evaluation to determine whether the generated images have more visual semantics relevant to the queries.

*2.2.1 Retrieval Performance.* The experimental results in Table 2 show that retrieval performance on the corpus containing only AI-generated images is not significantly greater than retrieval performance on the corpus containing only real images. It indicates that the AI-generated images in our benchmark do not introduce more visual semantics relevant to the queries. Specifically, we use three open-source and powerful vision-language pre-trained models including FLAVA [30], ALIGN [19], and BEIT-3 [34] to perform retrieval on the corpora that only contain real images and only contain generated images respectively. Since text-image retrieval datasets are composed of real images, the training on these images will introduce additional bias. Therefore, considering the excellent zero-shot text-image retrieval performance of the three models [34], we use these models directly for retrieval in zero-shot setting.

*2.2.2 Human Evaluation.* Table 3 shows humans think that in most samples of our benchmark, real images have more or equal visual semantics relevant to the queries than AI-generated images. This further guarantees the fairness of our assessment of invisible relevance bias. If the AI-generated images do not have more relevant

**Table 3: Proportion of the selections made by humans.**

| Flicker30k+AI | | | MSCOCO+AI | | |
|---|---|---|---|---|---|
| Which image is more relevant to the query? | | | | | |
| Real | AI | Equal | Real | AI | Equal |
| 46.25% | 13.75% | 40% | 45.35% | 12.15% | 42.5% |

visual semantics than real images, while the text-image retrieval model still tends to rank AI-generated images higher than the real images, the invisible relevance bias does exist. Specifically, we invite five humans with master's degrees to participate in the evaluation. We randomly shuffle the dataset into five parts and assign each part randomly to the volunteers. Volunteers are shown with a triple consisting of a caption (i.e., the query), a real image, and its corresponding AI-generated image. We ask them to select which image (real or AI-generated) has more relevant semantics to the caption. We instruct volunteers how to visually judge the relevance between images and captions, and calculate the pass rate through cross-validation between two different volunteers. The formal human evaluation begins when each volunteer's pass rate reaches 95%. We count the proportion of selections made by humans on our benchmark and the results are shown in Table 3.

## 3 BIAS ASSESSMENT

Based on the benchmark constructed above, we can assess the impact of AI-generated images on text-image retrieval. We not only study the impact when AI-generated images appear in the corpus, but also further discuss the impact when AI-generated images are mixed into the training data of retrieval models.

### 3.1 Text-Image Retrieval Models

As for training data, our assessment includes not only the retrieval models trained from scratch on supervised text-image pairs but also vision-language models that have been pre-trained on massive image-text pairs. As for model architecture, our assessment includes both dual-encoder-based and fusion-based models. The specific models used in the assessment include: (1) **NAAF** [41] is a fusion encoder text-image retrieval model that exploits both the positive effect of matched fragments and the negative effect of mismatched fragments to jointly infer text-image similarity. (2) **VSE** [10] is a dual-encoder text-image retrieval model that learns to automatically adapt the best pooling strategy for visual semantic embedding. (3)

**Table 4: Performance of the retrieval models on the benchmark we constructed consisting of both real and AI-generated images. Relative△ > 0 means retrieval models rank real images higher than AI-generated images, Relative△ < 0 means retrieval models rank AI-generated images higher than real images. The absolute value of Relative△ indicates the degree of the bias.**

| | | | Flicker30k+AI | | | | | | MSCOCO+AI | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | NDCG@1 | NDCG@3 | NDCG@5 | R@1 | R@3 | R@5 | NDCG@1 | NDCG@3 | NDCG@5 | R@1 | R@3 | R@5 |
| | | | Models trained from scratch | | | | | | | | | | | |
| Dual-encoder | VSE | Real | 16.18 | 26.93 | 29.26 | 26.40 | 56.10 | 65.32 | 11.85 | 20.19 | 22.87 | 19.34 | 42.66 | 53.24 |
| | | AI-generated | 19.59 | 29.68 | 31.86 | 31.96 | 59.78 | 68.34 | 13.56 | 20.93 | 23.37 | 22.12 | 43.21 | 53.90 |
| | | Relative△ | -17.81 | -9.00 | -8.05 | -17.81 | -5.8 | -4.36 | -13.53 | -3.64 | -2.22 | -13.53 | -1.29 | -1.24 |
| Fusion-encoder | NAAF | Real | 13.40 | 23.39 | 26.14 | 21.86 | 49.41 | 60.28 | 10.61 | 17.73 | 20.45 | 17.30 | 37.26 | 48.02 |
| | | AI-generated | 17.04 | 26.04 | 28.31 | 27.79 | 52.70 | 61.70 | 10.75 | 17.87 | 20.33 | 17.54 | 37.50 | 47.24 |
| | | Relative△ | -23.57 | -10.63 | -7.86 | -23.57 | -6.45 | -2.31 | -1.13 | -0.73 | 0.62 | -1.13 | -0.66 | 1.63 |
| | | | Pre-trained Vision-Language Models | | | | | | | | | | | |
| Dual-encoder | FLAVA | Real | 5.44 | 18.44 | 21.79 | 8.88 | 44.92 | 58.14 | 12.59 | 25.98 | 29.02 | 20.54 | 57.30 | 69.34 |
| | | AI-generated | 37.61 | 44.86 | 46.36 | 61.33 | 81.34 | 87.26 | 27.01 | 36.81 | 38.87 | 44.06 | 70.99 | 79.12 |
| | | Relative△ | -148.85 | -83.78 | -72.44 | -148.85 | -58.32 | -40.69 | -72.81 | -34.49 | -29.00 | -72.81 | -21.36 | -13.21 |
| | ALIGIN | Real | 21.92 | 37.20 | 39.05 | 35.76 | 7696 | 84.22 | 18.82 | 31.42 | 33.89 | 30.70 | 64.98 | 74.76 |
| | | AI-generated | 25.48 | 39.10 | 40.91 | 41.56 | 78.38 | 85.44 | 21.31 | 33.23 | 35.49 | 34.76 | 67.24 | 76.16 |
| | | Relative△ | -14.6 | -4.95 | -4.59 | -14.6 | -1.93 | -1.49 | -12.41 | -5.65 | -4.63 | -12.41 | -3.48 | -1.88 |
| | BEIT-3 | Real | 24.37 | 38.67 | 40.50 | 39.76 | 78.22 | 85.46 | 21.38 | 33.26 | 35.57 | 34.88 | 67.11 | 76.22 |
| | | AI-generated | 24.40 | 39.54 | 41.12 | 39.80 | 80.50 | 86.68 | 21.24 | 34.55 | 36.63 | 34.64 | 70.86 | 79.08 |
| | | Relative△ | -0.72 | -2.17 | -1.41 | -0.72 | -2.97 | -1.44 | 0.62 | -3.90 | -3.01 | 0.62 | -5.50 | -3.72 |
| Fusion-encoder | VILT | Real | 17.53 | 29.63 | 32.16 | 28.60 | 61.90 | 71.90 | 16.30 | 29.71 | 32.08 | 26.60 | 63.10 | 72.50 |
| | | AI-generated | 20.04 | 30.43 | 32.71 | 32.70 | 61.30 | 70.30 | 18.29 | 31.21 | 33.50 | 29.85 | 63.30 | 72.30 |
| | | Relative△ | -13.38 | -2.69 | -1.69 | -13.38 | 0.97 | 2.25 | -11.51 | -4.90 | -4.32 | -11.51 | -0.32 | 0.28 |

**VILT** is a fusion encoder text-image matching model based on the interaction between image and text in Transformer [33]. It is a vision-language model that has been pre-trained on massive text-image pairs. (4) **FLAVA** learns strong representations from multimodal. It is a vision-language model pre-trained on massive text-image pairs and can be used as the dual encoder model for text-image retrieval. (5) **ALIGIN** is also a dual encoder vision-language that has been pre-trained on over one billion image-text pairs. (6) **BEIT-3** is a multimodal foundation model that has been pre-trained on hundreds of millions of text-image pairs and massive texts and images. It can be used as a dual-encoder text-image retrieval model.

## 3.2 Experimental Settings and Metrics

As the neural networks tend to fit the data domain in training [37], our assessment is performed under the out-of-domain setting to try to mitigate potential bias introduced by the domain of the training data. Specifically, for the models that need to train from scratch on supervised text-image pairs (NAAF and VSE), we train them on Flicker30k (MSCOCO) and evaluate their performance on MSCOCO (Flicker30k). For the models that have been pre-trained on massive real text-image pairs and show excellent zero-shot performance in text-image retrieval, we directly use these pre-trained models to perform retrieval on the test datasets. An exception is that even though VILT has been pre-trained, it needs to be combined with a specific multi-layer perceptron to complete the text-image matching task in text-image retrieval. So we fine-tune VILT on supervised text-image retrieval datasets just like NAAF and VSE.

The metric follows [12] to measure the difference between the ranking of real and AI-generated images in the retrieved results as:

$$\text{Relative}\triangle = \frac{2(\text{Metric}_{\text{real}} - \text{Metric}_{\text{AI-generated}})}{\text{Metric}_{\text{real}} + \text{Metric}_{\text{AI-generated}}} \times 100\%, \quad (1)$$

in which Metric can be the metrics for IR such as NDCG@k and R@k. Relative△ > 0 means retrieval models rank real images higher than AI-generated images, Relative△ < 0 means retrieval models

rank AI-generated images higher than real images. The absolute value of Relative△ indicates the degree of the bias [12].

## 3.3 Invisible Relevance Bias

The experimental results are shown in Table 4. Overall, invisible relevance bias widely exists in text-image retrieval models, that is, text-image retrieval models tend to rank AI-generated images higher than real images even though they have very similar visual semantics. Specifically, the following conclusions can be made: (1) The invisible relevance bias exists in both the models trained from scratch and the vision-language models that have been pre-trained on massive supervised text-image pairs. (2) The invisible relevance bias exists in both dual-encoder-based and fusion-encoder-based retrieval models. (3) The invisible relevance bias has a relatively greater impact on the Top-1 retrieved image. In the retrieved list, the Top-1 item is most likely to be clicked by users, which means that invisible relevance bias introduced by AI-generated images will have a huge impact on users' actual search and click results.

## 3.4 More Serious Bias Caused by Training

Due to the wide distribution of AI-generated images on the internet and the bias in Section 3.3, AI-generated images are very likely to be mixed into the training data of retrieval models. This section delves deeper into the impact on retrieval performance and invisible relevance bias when AI-generated images are mixed into the training of retrieval models. The experimental results reveal a vicious cycle of falling into more serious invisible relevance bias. Specifically, the invisible relevance bias of text-image retrieval models causes the AI-generated images to have a higher chance of being obtained from massive data, which makes them more likely to be mixed into the training of retrieval models, leading to more serious bias and forming a vicious cycle. This ultimately results in users' search results being surrounded by AI-generated images.

Specifically, we explore the impact of the training mixed with generated images on retrieval by incorporating varying ratios of
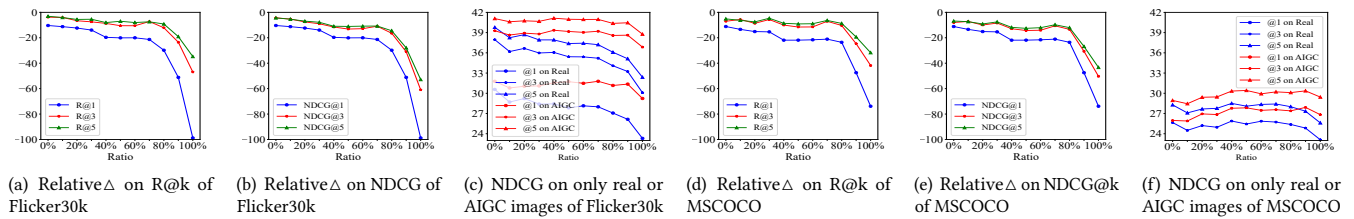
**Figure 2: Assessment results on the training set mixed with AI-generated images. We change the ratio of AI-generated images in the datasets (X-axis) while keeping the total number of training samples unchanged. The model is tested on the test set of Flicker30k+AI (in-domain) and MSCOCO+AI (out-of-domain) respectively that we constructed in Section 2.**
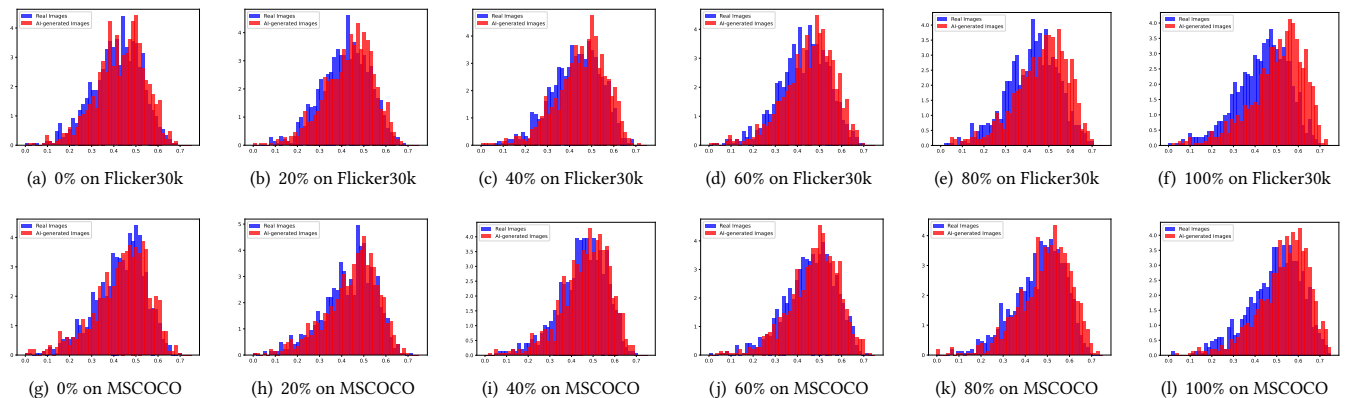


**Figure 3: Distribution of the caption-image relevance scores estimated by retrieval models that are trained on the datasets mixed with different ratios of AI-generated images. Flicker30k is in-domain and MSCOCO is out-of-domain.**

generated images into the training data. To ensure an accurate assessment, our experiments focus on the model trained from scratch (VSE). This is because pre-trained vision-language models have been pre-trained on massive real text-image pairs, which introduce additional biases in evaluating the impact of specific ratios of generated images in the training data. We reconstruct the training set of Flicker30k by replacing a certain ratio of real images with AI-generated images. For each real image ($I_r$) in the training set, we use the method in Section 2 to generate its corresponding AI-generated image ($I_g$). Then, we introduce the ratio $\alpha$, which means that in our reconstructed training data, the paired images for $\alpha$ percentage of captions are AI-generated images, and for $(100 - \alpha)$ percentage of captions are real images. We change the ratio while keeping the total number of training samples unchanged. We evaluate the performance of the trained model on the test set of Flicker30k+AI (in-domain setting) and MSCOCO+AI (out-of-domain setting).

Figure 2 illustrates that as the ratio increases, the ranking disparity between generated images and real images widens, with the retrieval model exhibiting a greater inclination to rank generated images higher (Relative△ decreasing). Meanwhile, Figure 3 demonstrates that with an increasing ratio, the discrepancy in score distribution between generated images and real images increases, and the scores of AI-generated images gradually become greater and greater than those of real images. In both in-domain and out-of-domain settings, the following conclusions can be drawn from

these results: (1) Retrieval models trained on the datasets mixed with AI-generated images exhibit more serious invisible relevance bias. (2) The invisible relevance bias tends to become more serious as the ratio of AI-generated images in the training data increases. (3) As the ratio of AI-generated images in the training data increases, the retrieval performance on real images is gradually damaged.

## 4 OUR METHOD IN DEBIASING

This section proposes an effective method to alleviate the bias.

### 4.1 Design of Debiasing Method

Our method alleviates the retrieval model's preference bias for generated images by measuring and optimizing the additional relevance score of generated images in training. Given a caption-image pair $(C_i, I_i^r)$ in the training set, the training process for both dual-encoder and fusion-encoder text-image retrieval models can be formulated as estimating the relevance score $s$ between $C_i$ and $I_i^r$, and using contrastive loss or regression loss as the optimization objective to adjust $s$, which can be described as:

$$s = R(C_i, I_i^r; \theta), \quad \theta_{\text{optimal}} = \arg\min_{\theta} \mathcal{L}(s, y; \theta),$$

in which $R(\cdot; \theta)$ is the relevance scoring function of the retrieval model such as cosine similarity between representations in the dual-encoder model and neural networks in the fusion-encoder model,

**Table 5: Performance of the retrieval models on the benchmark we constructed consisting of both real and AI-generated images with different sampling probability $\beta$ in our debiasing method.** Relative$\triangle > 0$ **means retrieval models rank real images higher than AI-generated images,** Relative$\triangle < 0$ **means retrieval models rank AI-generated images higher than real images. When** Relative$\triangle < 0$, **the absolute value of** Relative$\triangle$ **indicates the value of this bias.**

| | Flicker30k+AI (In-domain) | | | | | | | MSCOCO+AI (Out-of-domain) | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | w/o debias | $\beta = 50\%$ | $\beta = 60\%$ | $\beta = 70\%$ | $\beta = 80\%$ | $\beta = 90\%$ | $\beta = 100\%$ | w/o debias | $\beta = 50\%$ | $\beta = 60\%$ | $\beta = 70\%$ | $\beta = 80\%$ | $\beta = 90\%$ | $\beta = 100\%$ |
| Relative$\triangle$ on NDCG@1 | -10.35 | -1.406 | 31.42 | 62.77 | 91.71 | 112.06 | 129.20 | -13.53 | -1.384 | 45.35 | 80.67 | 114.63 | 140.23 | 154.43 |
| Relative$\triangle$ on NDCG@3 | -4.31 | -0.656 | 15.08 | 32.85 | 50.08 | 65.44 | 77.31 | -3.64 | -0.354 | 27.61 | 53.11 | 78.92 | 101.67 | 114.31 |
| Relative$\triangle$ on NDCG@5 | -4.37 | -0.876 | 13.13 | 27.84 | 42.28 | 55.68 | 65.31 | -2.22 | -0.214 | 23.47 | 46.42 | 69.62 | 90.28 | 102.89 |
| NDCG@1 on only real images | 30.57 | 33.44 | 33.15 | 33.26 | 33.12 | 33.09 | 33.20 | 18.50 | 21.09 | 21.48 | 21.32 | 20.52 | 20.43 | 20.01 |
| NDCG@1 on only real images | 37.95 | 40.44 | 40.32 | 40.53 | 40.38 | 40.13 | 40.31 | 25.66 | 28.92 | 29.05 | 28.78 | 28.32 | 28.05 | 27.65 |
| NDCG@5 on only real images | 39.78 | 42.29 | 41.98 | 42.18 | 42.05 | 41.93 | 42.10 | 28.28 | 31.55 | 31.69 | 31.43 | 30.97 | 30.56 | 30.02 |

$\theta$ is the set of parameters of the model, $\mathcal{L}$ is the loss function such as contrastive loss or regression loss, $y$ is the label. In our method, for each real image $I_i^r$ in the training data, we use the method in Section 2 to generate its corresponding AI-generated image $I_i^g$. Then we can get the caption-real-AI triple as $(C_i, I_i^r, I_i^g)$, in which $C_i$ is the paired caption for the real image $I_i^r$. We introduce the contrastive loss to get the difference in relevance scores between $I_i^r$ and $I_i^g$ with respect to the caption $C_i$ as:

$$\triangle s(I_i^g, I_i^r) = R(C_i, I_i^g; \theta) - R(C_i, I_i^r; \theta),$$

which can measure the additional invisible relevance introduced by the AI-generated image $I_i^g$ for $C_i$ compared with the real image $I_i^r$. Therefore, this can be used as a part of the optimization objective to mitigate the invisible relevance bias. It is because minimizing the difference between $R(C_i, I_i^g; \theta)$ and $R(C_i, I_i^r; \theta)$ in training can make the retrieval model eliminate the additional score estimated for $I_i^g$. In all triples in the training data, we consider the caption-real-AI triples $(C_i, I_i^r, I_i^g)$ whose $\triangle s(I_i^r, I_i^g)$ is greater than 0 and perform Bernoulli sampling from these triples with probability $\beta$ to get the target triple set $\mathbb{B}$ for debiasing optimization as:

$$\mathbb{B} = \{(C_i, I_i^r, I_i^g) \mid (C_i, I_i^r, I_i^g) \sim \text{Bernoulli}(\beta) \cdot \mathbb{I}(\triangle s(I_i^g, I_i^r) > 0)\}.$$

The reason why we only sample triples with probability $\beta$ is to adjust the tolerance of the retrieval models to the AI-generated images. The higher the probability $\beta$, the more likely the retrieval models are to rank the AI-generated images to a lower position. The total optimization objective in training is:

$$\theta_{\text{optimal}} = \arg\min_{\theta}\left( \sum_{s_i, y_i \in \mathbb{A}} \mathcal{L}(s_i, y_i; \theta) + \sum_{I_i^r, I_i^g \in \mathbb{B}} \triangle s(I_i^g, I_i^r) \right),$$

in which $\mathbb{A}$ is the set of all samples in the training data. For the sample $i$, $s_i$ is the estimated score and $y_i$ is the label.

## 4.2 Evaluation of Debiasing

Evaluation of our debiasing method focuses on: (1) How it affects the ranking difference between real and AI-generated images. (2) How it affects the distribution of caption-image relevance scores.

Table 5 shows how the Relative$\triangle$ and retrieval performance changes with the sampling probability $\beta$. The results indicate that: (1) Our method not only effectively alleviates the retrieval model's preference for AI-generated images, but also makes real images ranked significantly higher than AI-generated images. (2) As the sampling probability $\beta$ increases, real images are ranked higher and

higher than AI-generated images. (3) When $\beta$ is 0.5, retrieval models can achieve a fair ranking between real and generated images with little bias. (4) Our method improves the retrieval performance. It is because, in our method, AI-generated images can be seen as the hard negatives in training, which enhances the ability of the retrieval model to distinguish the images with very similar semantics.

Figure 4 shows the distribution of the caption-image relevance scores estimated by retrieval models with different sampling probability $\beta$. The results indicate that our method effectively reduces the relevance scores between captions and generated images estimated by the retrieval model. With the increase in sampling probability ($\beta$), the disparity in score distribution between generated images and real images expands, and the scores of real images gradually become greater and greater than those of generated images.

## 5 CAUSES OF INVISIBLE RELEVANCE BIAS

In this section, we use the debiased model proposed in Section 4 to reversely analyze the causes of the invisible relevance bias. Specifically, we extract the transformations performed by the debiased retriever on AI-generated images. The reverse process of these transformations can be seen as the cause of invisible relevance bias. We also design the experiments to support this point of view.

## 5.1 Transformations in Debiasing

To simplify the analysis process, we focus on the dual-encoder retrieval model that estimates the relevance score by computing vector similarity between text and image representations. For the same text, image representation can directly affect the estimation of the relevance, so we use image representation as the main object in our analysis, which can help us to find the most direct causes of invisible relevance bias. Specifically, we analyze the difference between image representations encoded by the original retriever and debiased retriever and extract the transformations of debiasing from this difference. Then, we reverse these transformations to explore the causes of the invisible relevance bias.

Given the image encoder $v(\cdot; \theta)$ of original retriever, the image encoder $v^d(\cdot; \theta')$ of our debiased retriever, and an AI-generated image $I^g$, we can get the representations of $I^g$ encoded by $v(\cdot; \theta)$ and $v^d(\cdot; \theta')$ respectively as $\boldsymbol{r} = v(I^g; \theta), \boldsymbol{r}^d = v^d(I^g; \theta')$. For the normalized text representation $\boldsymbol{t} = [t_1, t_2, ..., t_n]$, image representation $\boldsymbol{r} = [r_1, r_2, ..., r_n]$ from the original retriever and image representation $\boldsymbol{r}^d = [r_1^d, r_2^d, ..., r_n^d]$ from the debiased retriever, the relevance
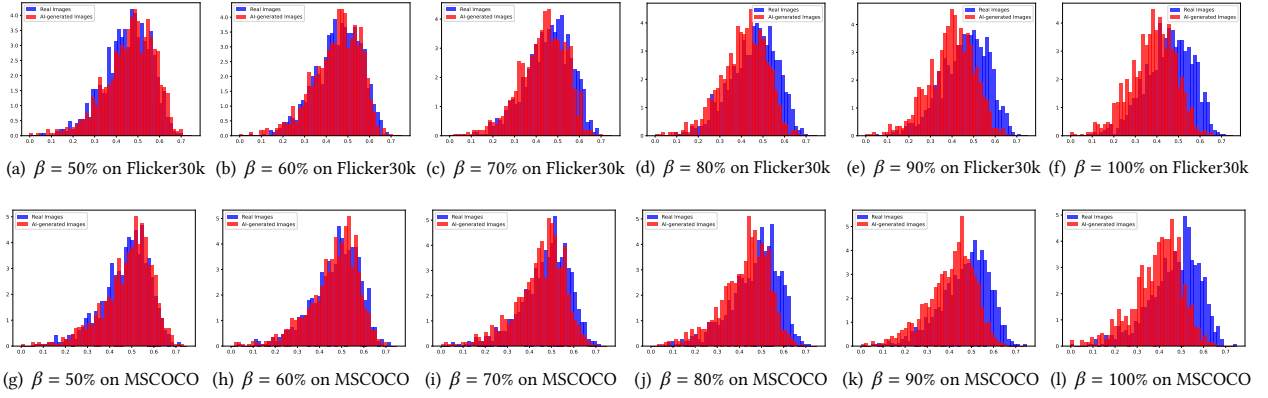
(a) $\beta = 50\%$ on Flicker30k (b) $\beta = 60\%$ on Flicker30k (c) $\beta = 70\%$ on Flicker30k (d) $\beta = 80\%$ on Flicker30k (e) $\beta = 90\%$ on Flicker30k (f) $\beta = 100\%$ on Flicker30k

(g) $\beta = 50\%$ on MSCOCO (h) $\beta = 60\%$ on MSCOCO (i) $\beta = 70\%$ on MSCOCO (j) $\beta = 80\%$ on MSCOCO (k) $\beta = 90\%$ on MSCOCO (l) $\beta = 100\%$ on MSCOCO

**Figure 4: Distribution of the caption-image relevance scores estimated by retrieval models with different sampling probability $\beta$ in our debiasing method. Flicker30k is in-domain and MSCOCO is out-of-domain.**

$score(t, r)$ and $score(t, r^d)$ can be calculated as:

$$score(t, r) = t_1 r_1 + t_2 r_2 + t_3 r_3 + ... + t_n r_n,$$

$$score(t, r^d) = t_1 r_1^d + t_2 r_2^d + t_3 r_3^d + ... + t_n r_n^d.$$

Therefore, the adjustment of relevance score between text $t$ and AI-generated image $I^g$ in the debiased model is essentially changing the values of each element ($r_i$) in $r$, which can be described as:

$$score(t, r^d) = t_1 r_1^d + t_2 r_2^d + ... + t_n r_n^d$$
$$= t_1(r_1 + \triangle r_1) + t_2(r_2 + \triangle r_2) + ... + t_n(r_n + \triangle r_n).$$

The transformations $p$ in the debiased model can be represented by a vector with the same dimensions as $r$ and $r^d$:

$$p = [\triangle r_1, \triangle r_2, ..., \triangle r_n] = [p_1, p_2, ..., p_n],$$

Then, we perform two-dimensional visualization of the $r$, $r^d$ and $p$ of all images in datasets to try to find the patterns from them. The T-SNE visualization in Figure 5 shows that compared with the scattered image representations, the transformations vector $p$ shows an obvious aggregation phenomenon. This indicates that there is consistency in the transformations performed by the debiased retriever on AI-generated images with very different semantics.

### 5.2 Reversing the Transformations

Debiased retriever modifies each element ($r_i$) of the representation $r$ from the original retriever according to the value of the corresponding element $p_i$ in transformations vector $p$ and gets the debiased representation $r^d$, which can be described as:

$$r_i^d = r_i + p_i, r_i \in r, p_i \in p, r_i^d \in r^d.$$

We can reverse this process to get the causes of the invisible relevance bias. This is because the transformation made by a debiased retriever to make a biased AI-generated image become unbiased is exactly the difference between the biased AI-generated images and the real images. That is, the reason why an unbiased representation $r^d$ becomes the representation $r$ with bias is that the reverse transformation ($-p_i$) is done on $r_i^d$ ($r_i = r_i^d - p_i$). Therefore, we conclude that an unbiased representation $r^d$ becomes the representation $r$ with bias because $r^d$ is added bitwisely by a vector
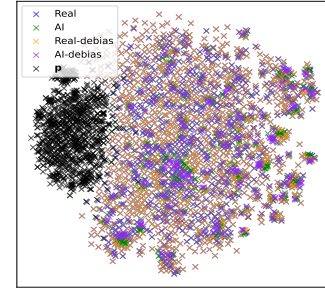


**Figure 5: T-SNE visualization of image representations and transformations vector $p$.**

$-p = [-p_1, -p_2, -p_3, ..., -p_n]$. Combining the phenomenon shown in Figure 5 that there is consistency in the transformations vector $p$ on AI-generated images with very different semantics, we can get the causes of invisible relevance bias in AI-generated images: **AI-generated images cause the image encoder in the retriever to embed additional information to their representations.** This additional information is the direct cause of invisible relevance bias and can be obtained by the difference in image representations between original and debiased retrievers. This information has the following three characteristics: **(1)** This information cannot be reflected in a visible way, but can only be embedded by neural network-based models. **(2)** When this information is embedded into the representation of the image, it can amplify the query-image relevance to produce a higher score. **(3)** This information has a certain consistency in AI-generated images with different semantics. This information may be like the watermark that is universal information for the image generation model and can be expressed by neural network visual models such as image encoders. We design the experiments to support these three points.

### 5.3 Supporting Experiments

This section aims to use experimental results to support three characteristics in the causes of invisible relevance bias mentioned above.

**Table 6: Effect of reverse transformations vector $-p$ on test set of Flicker30k. The retrieval model is VSE trained on Flicker30k without any debiasing training.**

| | NDCG@1 | NDCG@3 | NDCG@5 | R@1 | R@3 | R@5 |
|---|---|---|---|---|---|---|
| | | Relative △ on | | | | |
| Original | -10.35 | -4.31 | -4.37 | -10.35 | -4.72 | -4.06 |
| Add $-p$ to Real | 17.85 | 4.54 | 2.99 | 17.85 | -0.28 | -1.17 |

For the first point, human evaluation in Section 2.2.2 and retrieval performance in Section 2.2.1 have shown that AI-generated images do not introduce additional visual semantics compared to their real images, indicating that the additional information is invisible. Besides, the ranking bias detected in Section 3.3 and Section 3.4 shows that this additional information can be embedded by the image encoder and produce higher relevance score than real images.

For the second point, we design a direct experiment to support it. We apply the bitwise addition of the reverse transformations vector $-p$ to the representations of real images encoded by the original, non-debiased retriever and detect whether the bias can be eliminated. The intention for this is that if this additional information ($-p$) is indeed the cause of the higher ranking of AI-generated images, then by incorporating this information into the representation of real images, the real images will similarly attain a higher ranking. Consequently, this would mitigate the ranking disparity between real and generated images. The experimental results are shown in Table 6. It is very surprising that the ranking advantage of generated images over real images caused by invisible relevance bias is not only eliminated but reversed by simply bitwisely adding $-p$ to the representation of the real images without any debiasing training. This proves that the reverse transformations vector $-p$ we found is an important cause of the invisible relevance bias. It is implicit in the AI-generated images and can be embedded into the representations by the image encoder.

For the third point, T-SNE visualization of image representations and transformations vector $p$ in Figure 5 has shown that compared with the scattered image representations, the transformations vector $p$ show an obvious aggregation phenomenon. This proves that for AI-generated images with different semantics, the debiased model only needs roughly consistent transformations on representations to remove the bias, which means that there is a certain consistency in the additional information for AI-generated images encoded by the image encoder.

## 6 RELATED WORK

### 6.1 Artificial Intelligence Generated Content

With the development of generation models, AI-generated content (AIGC) becomes more realistic and less discernible [1, 8, 36]. The use of AI to generate content has greatly changed the way of content generation. It reduced the cost of content generation and improved the efficiency than traditional compared to traditional content generation pipeline biased on humans [18, 31]. For example, Stable Diffusion [11, 39] and DALL-E-3 [4] can generate high-quality images by simply following a short text description of uses' input. However, as AIGC is gradually applied to various content production pipelines in society and the internet, some potential

risk concerns are also worthy of consideration. Some studies have found that AI-generated content has risks in many aspects, such as discrimination, privacy leakage, ethics, and safety [9, 13, 20, 32, 42]. A recent study finds that texts generated by AI introduce bias in text retrieval, causing the retriever to rank them higher [12]. This paper extends the study of this bias to cross-modal retrieval.

### 6.2 Bias and Fairness in Information Retrieval

Due to the influence of training data, indexes, models, etc., the ranking results of information retrieval will produce certain deviations [16]. These biases are specifically reflected in various aspects such as gender [3], cognition [6], understandability [43], and retrievability [35]. Kulshrestha et al. [22] formulated a framework for quantifying bias, addressing its origins in both the data source and the ranking system. Beutel et al. [5], Kuhlman et al. [21], and Yao and Huang [40] introduced pairwise comparisons focusing on utility and prediction errors. Geyik et al. [17] and Yang along with Stoyanovich [38] suggested fairness measures based on distances for ranked outputs, grounded in statistical parity. Gao [15] advocated for a unified evaluation metric tailored for fairness-aware ranking algorithms. Diaz et al. [14] proposed metrics specifically designed for assessing fairness in stochastic rankings. Numerous studies have conducted comparisons among existing fairness metrics [14, 17, 26, 28]. In this paper, we focus on the ranking bias caused by AI-generated Images in text-image retrieval.

## 7 CONCLUSION

This paper explores the impact on text-image retrieval introduced by AI-generated images. We construct a reasonable benchmark to simulate the retrieval scenarios comprising both real images and AI-generated images. Experiments on this benchmark underscore that AI-generated images tend to be ranked higher by retrieval models, despite lacking more visually relevant semantics to the queries than real images. We define this bias as **invisible relevance bias**. This bias is prevalent across retrieval models with varying training data and architectures. Moreover, mixing AI-generated images into the training data makes the bias more serious, causing a vicious cycle where AI-generated images gain more exposure from massive data, increasing their likelihood of being mixed into retrieval model training, and exacerbating the bias further. To solve this, we propose an effective debiasing method to mitigate the invisible relevance bias. Then, we use our proposed debiasing method to reversely determine that the cause of invisible relevance is that the AI-generated images cause the image encoder to introduce additional information into their representation. This additional information can make the retriever estimate a higher relevance score. Findings in this paper reveal the potential impact of AI-generated images on text-image retrieval systems in the context of the rapid development of AIGC and have implications for further research.

# REFERENCES

[1] Jorge Agnese, Jonathan Herrera, Haicheng Tao, and Xingquan Zhu. 2020. A survey and taxonomy of adversarial neural networks for text-to-image synthesis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10, 4 (2020), e1345.

[2] Qingyao Ai, Ting Bai, Zhao Cao, Yi Chang, Jiawei Chen, Zhumin Chen, Zhiyong Cheng, Shoubin Dong, Zhicheng Dou, Fuli Feng, et al. 2023. Information Retrieval Meets Large Language Models: A Strategic Report from Chinese IR Community. *AI Open* 4 (2023), 80–90.

[3] Leif Azzopardi. 2021. Cognitive biases in search: a review and reflection of cognitive biases in Information Retrieval. In *Proceedings of the 2021 conference on human information interaction and retrieval.* 27–37.

[4] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions.

[5] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H Chi, et al. 2019. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining.* 2212–2220.

[6] Amin Bigdeli, Negar Arabzadeh, Morteza Zihayat, and Ebrahim Bagheri. 2021. Exploring gender biases in information retrieval relevance judgement datasets. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43.* Springer, 216–224.

[7] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL]

[8] Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S Yu, and Lichao Sun. 2023. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *arXiv preprint arXiv:2303.04226* (2023).

[9] Canyu Chen and Kai Shu. 2023. Can LLM-Generated Misinformation Be Detected? *arXiv preprint arXiv:2309.13788* (2023).

[10] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. 2021. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 15789–15798.

[11] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. 2023. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

[12] Sunhao Dai, Yuqi Zhou, Liang Pang, Weihao Liu, Xiaolin Hu, Yong Liu, Xiao Zhang, and Jun Xu. 2023. LLMs may Dominate Information Access: Neural Retrievers are Biased Towards LLM-Generated Texts. *arXiv preprint arXiv:2310.20501* (2023).

[13] Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335* (2023).

[14] Fernando Diaz, Bhaskar Mitra, Michael D Ekstrand, Asia J Biega, and Ben Carterette. 2020. Evaluating stochastic rankings with expected exposure. In *Proceedings of the 29th ACM international conference on information & knowledge management.* 275–284.

[15] Ruoyuan Gao. 2021. *Toward a fairer information retrieval system.* Ph. D. Dissertation. Rutgers The State University of New Jersey, School of Graduate Studies.

[16] Ruoyuan Gao and Chirag Shah. 2021. Addressing bias and fairness in search systems. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval.* 2643–2646.

[17] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining.* 2221–2231.

[18] Hans WA Hanley and Zakir Durumeric. 2023. Machine-Made Media: Monitoring the Mobilization of Machine-Generated Articles on Misinformation and Mainstream News Websites. *arXiv preprint arXiv:2305.09820* (2023).

[19] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning.* PMLR, 4904–4916.

[20] Bohan Jiang, Zhen Tan, Ayushi Nirmal, and Huan Liu. 2023. Disinformation Detection: An Evolving Challenge in the Age of LLMs. *arXiv preprint arXiv:2309.15847* (2023).

[21] Caitlin Kuhlman, MaryAnn VanValkenburg, and Elke Rundensteiner. 2019. Fare: Diagnostics for fair ranking using pairwise error metrics. In *The world wide web conference.* 2936–2942.

[22] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P Gummadi, and Karrie Karahalios. 2017. Quantifying search bias: Investigating sources of bias for political searches in social media. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing.* 417–432.

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. 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, 740–755.

[24] Christopher D Manning. 2009. *An introduction to information retrieval.* Cambridge university press.

[25] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision.* 2641–2649.

[26] Amifa Raj and Michael D Ekstrand. 2020. Comparing fair ranking metrics. *arXiv preprint arXiv:2009.01311* (2020).

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 10684–10695.

[28] Piotr Sapiezynski, Wesley Zeng, Ronald E Robertson, Alan Mislove, and Christo Wilson. 2019. Quantifying the impact of user attentionon fair group representation in ranked lists. In *Companion proceedings of the 2019 world wide web conference.* 553–562.

[29] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. The Curse of Recursion: Training on Generated Data Makes Models Forget. *arXiv preprint arxiv:2305.17493* (2023).

[30] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 15638–15650.

[31] Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. 2023. AI model GPT-3 (dis) informs us better than humans. *arXiv preprint arxiv:2301.11924* (2023).

[32] Jinyan Su, Terry Yue Zhuo, Jonibek Mansurov, Di Wang, and Preslav Nakov. 2023. Fake news detectors are biased against texts generated by large language models. *arXiv preprint arXiv:2309.08674* (2023).

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[34] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. 2022. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442* (2022).

[35] Colin Wilkie and Leif Azzopardi. 2014. A retrievability analysis: Exploring the relationship between retrieval bias and retrieval performance. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management.* 81–90.

[36] Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Hong Lin. 2023. Ai-generated content (aigc): A survey. *arXiv preprint arXiv:2304.06632* (2023).

[37] Shicheng Xu, Liang Pang, Huawei Shen, and Xueqi Cheng. 2022. Match-Prompt: Improving Multi-Task Generalization Ability for Neural Text Matching via Prompt Learning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (Atlanta, GA, USA) *(CIKM '22).* Association for Computing Machinery, New York, NY, USA, 2290–2300. https://doi.org/10.1145/3511808.3557388

[38] Ke Yang and Julia Stoyanovich. 2017. Measuring fairness in ranked outputs. In *Proceedings of the 29th international conference on scientific and statistical database management.* 1–6.

[39] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. 2023. Diffusion Models: A Comprehensive Survey of Methods and Applications. arXiv:2209.00796 [cs.LG]

[40] Sirui Yao and Bert Huang. 2017. New fairness metrics for recommendation that embrace differences. *arXiv preprint arXiv:1706.09838* (2017).

[41] Kun Zhang, Zhendong Mao, Quan Wang, and Yongdong Zhang. 2022. Negative-aware attention framework for image-text matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 15661–15670.

[42] Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Exploring ai ethics of chatgpt: A diagnostic analysis. *arXiv preprint arXiv:2301.12867* (2023).

[43] Guido Zuccon. 2016. Understandability biased evaluation for information retrieval. In *European Conference on Information Retrieval.* Springer, 280–292.