AI-GENERATED IMAGES INTRODUCE INVISIBLE RELEVANCE BIAS TO TEXT-IMAGE RETRIEVAL

Shicheng Xu[1]*, **Danyang Hou**[1]*, **Liang Pang**[1]†, **Jingcheng Deng**[1], **Jun Xu**[2], **Huawei Shen**[1], **Xueqi Cheng**[1]
[1]CAS Key Laboratory of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences
[2]Gaoling School of Artificial Intelligence, Renmin University of China
{xushicheng21s,houdanyang18b,pangliang,dengjingcheng23s}@ict.an.cn

ABSTRACT

With the application of foundation 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. Findings in this paper reveal the potential impact of AI-generated images on text-image retrieval and have implications for further research.

1 INTRODUCTION

With the advancement of foundation models, the quality of AI-generated content (AIGC) has been increasingly improved Yang et al. (2023); Brown et al. (2020). The utilization of AI for content generation has transformed the way of content creation. It not only reduces the cost of content generation but also enhances the efficiency, leading to a rapid influx of large amounts of AI-generated content onto the internet Ai et al. (2023); Dai et al. (2023). Information retrieval (IR) is an important way for people to obtain the target information from massive data Manning (2009). 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, how will AI-generated content influence the ranking results of search? Since the internet is replete with a substantial number of images, serving as crucial sources for IR systems, our paper performs the investigation of this challenge to text-image retrieval models.

Firstly, **we construct 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. To solve the above problem, we propose an image over-sampling and selection strategy based on the merged caption. Experimental results and human evaluation show that our proposed method can successfully construct semantically

*Equal Contributions
†Corresponding author

similar AI-generated images for real images. 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 as **invisible relevance bias**. 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 and fusion encoder retrieval models.

Another important point is that AI-generated content does not only have an impact on the inference stage of the retrieval models. Due to their wide distribution 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 to explore the impact. 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)**, the retrieval model exhibits a greater inclination to rank generated images higher. This phenomenon reveals that 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 Shumailov et al. (2023), leading to more serious bias and forming a vicious cycle (Figure 1). We also propose an effective method to alleviate this bias and identify the causes of this bias (Appendix B and C). We reveal the impact of AI-generated images on text-image retrieval and has implications for further research.

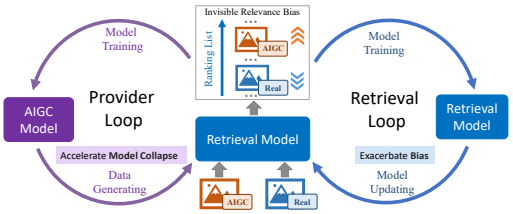


Figure 1: IR models tend to rank AI-generated images higher than real images. This bias increases the likelihood of the generated images being exposed from massive data, which makes them more likely to be mixed into the training of AIGC and retrieval models, leading to more serious bias and forming a vicious cycle.

2 Benchmark Construction

We construct a retrieval corpus consisting of real images and AI-generated images for bias assessment. The corpus should meet four critical requirements. ($\mathcal{R}$-1) Firstly, 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. We select two widely used text-image retrieval datasets including Flickr30k Plummer et al. (2015) and MSCOCO Lin et al. (2014) for construction. Details of construction of the benchmark are introduced in Appendix A.

3 Bias Assessment

3.1 Text-Image Retrieval Models

Our experiments assess both fusion and dual encoder models and models that have been pre-trained on massive image-text pairs and trained from scratch: (1) **NAAF** Zhang et al. (2022) (fusion encoder), (2) **VSE** Chen et al. (2021) (dual encoder), (3) **VILT** (fusion encoder, pre-trained). (4) **FLAVA** (dual encoder, pre-trained), (5) **ALIGIN** (dual encoder, pre-trained) (6) **BEIT-3** (dual encoder, pre-trained).

Table 1: Retrieval models on the benchmark consisting of both real and AI-generated images. Relative$\triangle < 0$ means retrieval models rank generated images higher than real images.

		Flicker30k+AI						MSCOCO+AI					
		NDCG			Recall			NDCG			Recall		
		@1	@3	@5	@1	@3	@5	@1	@3	@5	@1	@3	@5
Models trained from scratch													
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$\triangle$	-17.81	-9.00	-8.05	-17.81	-5.8	-4.36	-13.53	-3.64	-2.22	-13.53	-1.29	-1.24
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$\triangle$	-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													
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$\triangle$	-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$\triangle$	-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$\triangle$	-0.72	-2.17	-1.41	-0.72	-2.97	-1.44	0.62	-3.90	-3.01	0.62	-5.50	-3.72
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$\triangle$	-13.38	-2.69	-1.69	-13.38	0.97	2.25	-11.51	-4.90	-4.32	-11.51	-0.32	0.28

3.2 Experimental Settings and Metrics

As the neural networks tend to fit the data domian in training Xu et al. (2022), 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, 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 retrieval datasets. The metric follows Dai et al. (2023) to measure the difference between the ranking of real and AI-generated images in the retrieved results as:

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

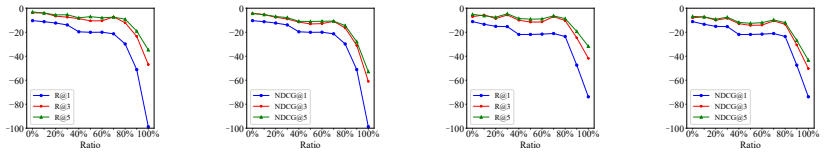
in which Metric can be the metrics for IR such as NDCG@k and R@k. Relative$\triangle > 0$ means retrieval models rank real images higher than AI-generated images, Relative$\triangle < 0$ means retrieval models rank AI-generated images higher than real images. The absolute value of Relative$\triangle$ indicates the degree of the bias Dai et al. (2023).

3.3 Invisible Relevance Bias

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

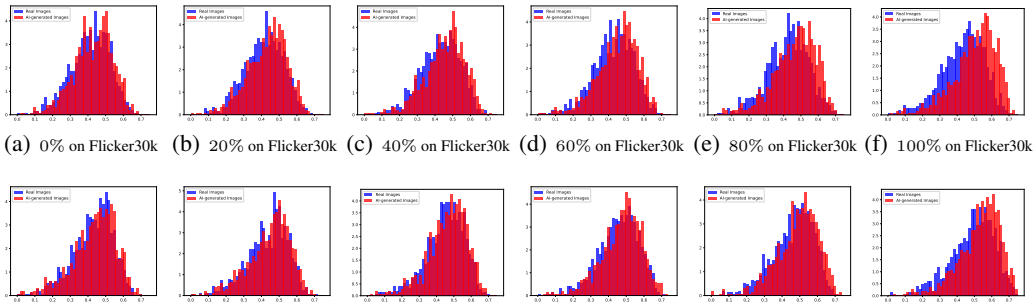
3.4 More Serious Bias Caused by Training

Due to 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



(a) R@k of Flicker30k (b) NDCG of Flicker30k (c) R@k of MSCOCO (d) NDCG of MSCOCO

Figure 2: Assessment results (Relative△) on the training set miexed with AI-generated images. We change the ratio of AI-generated images in the datasets (X-axis). The model is tested on the test set of Flicker30k+AI (in-domain) and MSCOCO+AI (out-of-domain) respectively.



(a) 0% on Flicker30k (b) 20% on Flicker30k (c) 40% on Flicker30k (d) 60% on Flicker30k (e) 80% on Flicker30k (f) 100% on Flicker30k

(g) 0% on MSCOCO (h) 20% on MSCOCO (i) 40% on MSCOCO (j) 60% on MSCOCO (k) 80% on MSCOCO (l) 100% on MSCOCO

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.

delves deeper into the impact on 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 generated images into the training data. To ensure an accurate assessment, our experiments focus on the model trained from scratch (VSE). We reconstruct the training set of Flicker30k by replacing a certain ratio (α) of real images with AI-generated images. This means that in our reconstructed training data, the paired images for α percentage of captions are AI-generated 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. In both in-domain and out-of-domain settings, the following conclusions can be drawn: (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.

4 Conclusion

This paper is dedicated to exploring the impact of images generated by foundation models on search. We construct a reasonable benchmark to simulate the retrieval scenarios comprising both real 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 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 that exacerbates the bias.

REFERENCES

Qingyao Ai, Ting Bai, Zhao Cao, Yi Chang, Jiawei Chen, Zhumin Chen, Zhiyong Cheng, Shoubin Dong, Zhicheng Dou, Fuli Feng, et al. Information retrieval meets large language models: A strategic report from chinese ir community. *AI Open*, 4:80–90, 2023.

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. Language models are few-shot learners, 2020.

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

Sunhao Dai, Yuqi Zhou, Liang Pang, Weihao Liu, Xiaolin Hu, Yong Liu, Xiao Zhang, and Jun Xu. Llms may dominate information access: Neural retrievers are biased towards llm-generated texts. *arXiv preprint arXiv:2310.20501*, 2023.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Christopher D Manning. *An introduction to information retrieval*. Cambridge university press, 2009.

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

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arxiv:2305.17493*, 2023.

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

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

Shicheng Xu, Liang Pang, Huawei Shen, and Xueqi Cheng. 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*, CIKM '22, pp. 2290–2300, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392365. doi: 10.1145/3511808.3557388. URL https://doi.org/10.1145/3511808.3557388.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications, 2023.

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

A BENCHMARK CONSTRUCTION

A.1 IMAGE GENERATION

We select two widely used text-image retrieval datasets including Flickr30k Plummer et al. (2015) and MSCOCO Lin et al. (2014) 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. And input the merged caption to a stable diffusion model $\mathbf{M}$ to generate the image (I_g), which can be described as:

$$C_m = \text{GPT-3.5}(C_1, C_2, ..., C_n), \quad I^g = \mathbf{M}(C_m).$$

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. Besides, this also avoids the relevance bias in assessment caused by the generated image being overly inclined to one certain caption.

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 have much 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 $\boldsymbol{e_r}$ for I^r and set of embeddings $\boldsymbol{E} = \{\boldsymbol{e_1^g}, \boldsymbol{e_2^g}, ..., \boldsymbol{e_n^g}\}$ for $\{I_1^g, I_2^g, ..., I_n^g\}$. We calculate the cosine similarity between each embedding in $\boldsymbol{E}$ and $\boldsymbol{e_r}$ 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:

$$I^g = \arg\max_i \{cosine(\boldsymbol{e^r}, \boldsymbol{e_i^g}) | \boldsymbol{e_i^g} \in \boldsymbol{E}\}. \tag{2}$$

Comparison between Different Generation Methods. The comparison between different generation methods under the above-mentioned four requirements is shown in Table 2. 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

[1] https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0

[2] https://huggingface.co/laion/CLIP-ViT-H-14-laion2B-s32B-b79K

image for each real image in Flicker30k and MSCOCO. For the convenience of description, 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.

Table 2: 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	$\lvert\triangle\text{NDCG@5}\rvert = 10.8$	text to image (✓)	1:1 (✓)
Merged Caption	0.5348	$\lvert\triangle\text{NDCG@5}\rvert = 3.72$	text to image (✓)	1:1 (✓)
Merged Caption + Image Rewriting	0.5503	$\lvert\triangle\text{NDCG@5}\rvert = 7.22$	text-image to image (×)	1:1 (✓)
Merged Caption + Image Rewriting + Over-Sampling & Selection	0.5845	$\lvert\triangle\text{NDCG@5}\rvert = 5.31$	text-image to image (×)	1:1 (✓)
Merged Caption + Over-Sampling & Selection	0.6470	$\lvert\triangle\text{NDCG@5}\rvert = 1.44$	text to image (✓)	1:1 (✓)

A.2 Assessing the Quality of Generated Images

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 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 further evaluate the semantic information from human vision. Specifically, we let humans determine whether the generated images have more visual semantics relevant to the queries compared to the real images.

A.2.1 Retrieval Performance

The experimental results in Table 3 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 Singh et al. (2022), ALIGN Jia et al. (2021), and BEIT-3 Wang et al. (2022) 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 Wang et al. (2022), we use these models directly for retrieval in zero-shot setting.

A.2.2 Human Evaluation

Table 4 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 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. The triple consisting of a caption (i.e., the query), a real image, and its corresponding AI-generated image is the basic unit in this evaluation. We ask the five humans to select which image (real or AI-generated) has more relevant semantics to the caption. We count the proportion of selections made by humans on our benchmark and the results are shown in Table 4.

Table 3: 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

Table 4: 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%

B Our Method in debiasing

In this section, we propose an effective method to alleviate the invisible relevance bias. The details will be introduced in the following.

B.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, θ 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 β 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 β is to adjust the tolerance of the retrieval models to the AI-generated images. The higher the probability β, the more likely the retrieval models are to rank the AI-generated images to a lower position. The total optimization objective in training is:

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

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

	Flicker30k+AI (In-domain)						
	w/o debias	$\beta = 50\%$	$\beta = 60\%$	$\beta = 70\%$	$\beta = 80\%$	$\beta = 90\%$	$\beta = 100\%$
$\mathrm{Relative}\triangle$ on NDCG@1	-10.35	-1.406	31.42	62.77	91.71	112.06	129.20
$\mathrm{Relative}\triangle$ on NDCG@3	-4.31	-0.656	15.08	32.85	50.08	65.44	77.31
$\mathrm{Relative}\triangle$ on NDCG@5	-4.37	-0.876	13.13	27.84	42.28	55.68	65.31
NDCG@1 on only real images	30.57	33.44	33.15	33.26	33.12	33.09	33.20
NDCG@1 on only real images	37.95	40.44	40.32	40.53	40.38	40.13	40.31
NDCG@5 on only real images	39.78	42.29	41.98	42.18	42.05	41.93	42.10

Table 6: Performance of the retrieval models on the benchmark we constructed consisting of both real and AI-generated images with different sampling probability β in our debiasing method. $\mathrm{Relative}\triangle > 0$ means retrieval models rank real images higher than AI-generated images, $\mathrm{Relative}\triangle < 0$ means retrieval models rank AI-generated images higher than real images. When $\mathrm{Relative}\triangle < 0$, the absolute value of $\mathrm{Relative}\triangle$ indicates the value of this bias.

	MSCOCO+AI (Out-of-domain)						
	w/o debias	$\beta = 50\%$	$\beta = 60\%$	$\beta = 70\%$	$\beta = 80\%$	$\beta = 90\%$	$\beta = 100\%$
$\mathrm{Relative}\triangle$ on NDCG@1	-13.53	-1.384	45.35	80.67	114.63	140.23	154.43
$\mathrm{Relative}\triangle$ on NDCG@3	-3.64	-0.354	27.61	53.11	78.92	101.67	114.31
$\mathrm{Relative}\triangle$ on NDCG@5	-2.22	-0.214	23.47	46.42	69.62	90.28	102.89
NDCG@1 on only real images	18.50	21.09	21.48	21.32	20.52	20.43	20.01
NDCG@1 on only real images	25.66	28.92	29.05	28.78	28.32	28.05	27.65
NDCG@5 on only real images	28.28	31.55	31.69	31.43	30.97	30.56	30.02

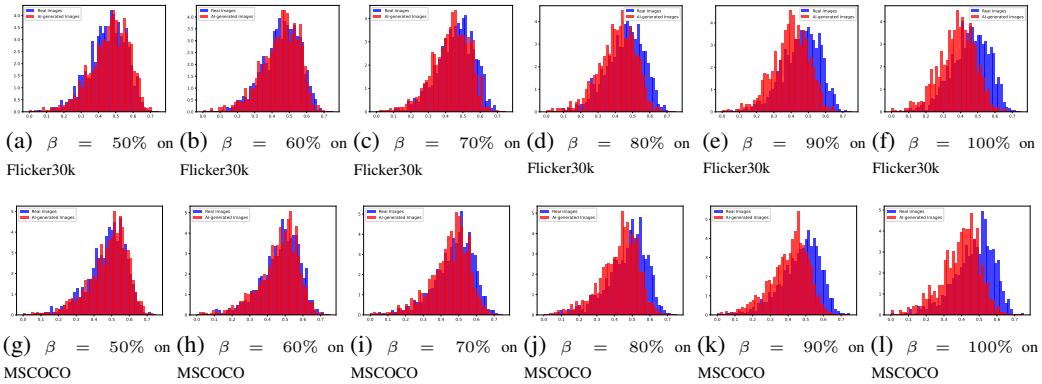
in which $\mathbb{A}$ is the set of all samples in the training data. For the sample i, s_i is the estimated score and y_i is the label.

B.2 Evaluation of Debiasing

We evaluate the effect of our debiasing method from three perspectives: (1) How our method affects the ranking difference between real images and AI-generated images. (2) How our method affects the distribution of caption-image relevance scores. (3) How our method affects the representations of real and AI-generated images. All of these three evaluations are performed in both in-domain (Flicker30k) and out-of-domain (MSCOCO) settings.

Table 5 and 6 shows how the $\mathrm{Relative}\triangle$ and retrieval performance changes with the sampling probability β. 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 β increases, real images are ranked higher and higher than AI-generated images. (3) When β 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.

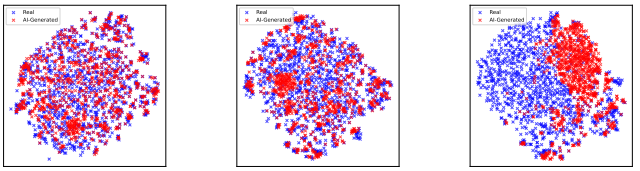
Figure 4 shows the distribution of the caption-image relevance scores estimated by retrieval models with different sampling probability β. The results indicate that our method effectively reduces the relevance scores between captions and generated images estimated by the retrieval model. With the increase in sampling probability (β), 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.



(a) β = 50% on Flicker30k (b) β = 60% on Flicker30k (c) β = 70% on Flicker30k (d) β = 80% on Flicker30k (e) β = 90% on Flicker30k (f) β = 100% on Flicker30k

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

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



(a) β = 60% on Flicker30k (b) β = 80% on Flicker30k (c) β = 100% on Flicker30k

(d) β = 60% on MSCOCO (e) β = 80% on MSCOCO (f) β = 100% on MSCOCO

Figure 5: T-SNE visualization of the embeddings of real and AI-generated images obtained from the image encoder in VSE.

Figure 5 shows the T-SNE of the embeddings of real and AI-generated images encoded by image encoder in VSE. This reveals that with the increase in the sampling probability (β), the representations of generated and real images become more distinguishable.

C Causes of Invisible Relevance Bias

In this section, we use the debiased model proposed in Section B to reversely analyze the causes of the invisible relevance bias. Specifically, we discern the transformations performed by the debiased retriever on AI-generated images through a comparative analysis of the difference in image representations between the original retriever and the debiased retriever. 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.

C.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

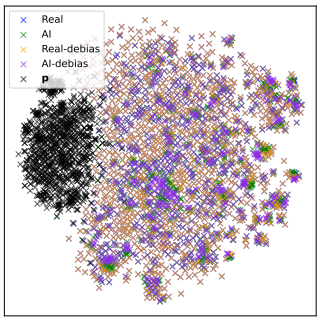


Figure 6: T-SNE visualization of image representations and transformations vector $\boldsymbol{p}$.

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:

$$\boldsymbol{r} = v(I^g;\theta), \quad \boldsymbol{r^d} = v^d(I^g;\theta').$$

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 $score(\boldsymbol{t}, \boldsymbol{r})$ and $score(\boldsymbol{t}, \boldsymbol{r^d})$ can be calculated as:

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

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 $\boldsymbol{r}$, which can be described as:

$$\begin{aligned} score(\boldsymbol{t}, \boldsymbol{r^d}) &= t_1 r_1^d + t_2 r_2^d + ... + t_n r_n^d \\ &= t_1(r_1 + \triangle r_1) + t_2(r_2 + \triangle r_2) + ... + t_n(r_n + \triangle r_n). \end{aligned}$$

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

$$\begin{aligned} \boldsymbol{p} &= [\triangle r_1, \triangle r_2, ..., \triangle r_n] \\ &= [p_1, p_2, ..., p_n], \end{aligned}$$

Then, we perform two-dimensional visualization of the $\boldsymbol{r}$, $\boldsymbol{r^d}$ and $\boldsymbol{p}$ of all images in datasets to try to find the patterns from them. The T-SNE visualization in Figure 6 shows that compared with the scattered image representations, the transformations vector $\boldsymbol{p}$ 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.

C.2 Reversing the Transformations

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

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

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

Table 7: Effect of reverse transformations vector $-\boldsymbol{p}$ on test set of Flicker30k. The retrieval model is VSE trained on Flicker30k without any debiasing training.

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

reason why an unbiased representation $\boldsymbol{r^d}$ becomes the representation $\boldsymbol{r}$ with bias is that the reverse transformation $(-p_i)$ is done on r_i^d, which can be described as:

$$r_i = r_i^d - p_i.$$

Therefore, we conclude that an unbiased representaion $\boldsymbol{r^d}$ becomes the representaion $\boldsymbol{r}$ with bias because $\boldsymbol{r^d}$ is added bitwisely by a vector $-\boldsymbol{p} = [-p_1, -p_2, -p_3, ..., -p_n]$. Combining the phenomenon shown in Figure 6 that there is consistency in the transformations vector $\boldsymbol{p}$ 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.

C.3 Supporting Experiments

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

For the first point, human evaluation in Section A.2.2 and retrieval performance in Section A.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 $-\boldsymbol{p}$ 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 $(-\boldsymbol{p})$ 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 7. 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 $-\boldsymbol{p}$ to the representation of the real images without any debiasing training. This proves that the reverse transformations vector $-\boldsymbol{p}$ 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 $\boldsymbol{p}$ in Figure 6 has shown that compared with the scattered image representations, the transformations vector $\boldsymbol{p}$ 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.