Discovering and Mitigating Visual Biases through Keyword Explanation

Appendix

Contents

A. Implementation Details

Computation cost. With a single RTX 3090 GPU, it took approximately 30 minutes to generate captions for the CelebA validation set, which contains 19,867 images. Extracting keywords took 5 seconds, and deriving CLIP scores took 33 seconds.

A.1. Bias discovery

A.1.1 Dataset details

CelebA blond. The CelebA [46] dataset contains 19,867 validation images, and we use the ResNet-50 [22] classifiers from the DRO repository.[6] Specifically, we use the ERM and DRO models trained with a learning rate of 0.0001 and batch size of 128, achieving accuracies of 95.44% and 90.40% for the blond class, respectively.

Waterbirds. The Waterbirds [66] dataset contains 1,199 validation images, and we use the ResNet-50 classifiers from the DRO repository. Specifically, we use the ERM and DRO models trained with a learning rate of 0.001 and batch size of 128. ERM achieved accuracies of 86.66% and 91.24% for the waterbird and landbird classes, respectively.

Dollar Street. The Dollar Street [64] dataset contains a snapshot of the original web page on July 30th, 2019,[7] and we use the ResNet-50 classifier trained on ImageNet for evaluation. We convert the class names of Dollar Street to ImageNet names using a mapping shown in Table 6.

Table 6. Conversion of class names from Dollar Street to ImageNet.

Dollar Street	ImageNet
books	bookcase
computers	desktop computer
cups	tea cup
diapers	diaper
dish_racks	plate rack
dishwashers	dishwasher
necklaces	necklace
stoves	stove
tables_with_food	dining table
toilet_paper	toilet paper
toilets	toilet seat
wall_clocks	wall clock
wardrobes	wardrobe
wheel_barrows	barrow
wrist_watches	digital watch

ImageNet and variants. ImageNet [14] has 1,281,167 training images across 1,000 classes. We use CLIP zero-shot classifier with 80-prompts ensemble strategy and class names, following the CLIP paper. Specifically, with ResNet-50 architecture, the classifier achieves an accuracy of 60.56% for vanialla ImageNet. We apply B2T to the most challenging classes, where the classifier exhibits the lowest accuracy.

ImageNet-R [26] consists of 30,000 validation images, representing artistic renditions of ImageNet classes. We use the full set of 1,000 classes to infer classifiers, while ImageNet-R samples belong to a subset of 200 ImageNet classes. ImageNet-C [25] contains corrupted versions of the ImageNet validation set, including snow and frost corruptions. Each corrupted dataset has 50,000 images, corresponding to vanilla ImageNet. To extract B2T keywords, we sample 10% of each validation set and combine them with an equal number of samples from the original vanilla ImageNet.

We use the ResNet-50 classifier trained on vanilla ImageNet with the classic training recipe (V1) from the PyTorch model hub, which achieved 76.15% accuracy for vanilla ImageNet. It achieves 52.8%, 64.6%, and 67.7% accuracy for ImageNet-R, ImageNet-C snow, and ImageNet-C frost, respectively.

[6] https://github.com/kohpangwei/group_DRO

[7] https://github.com/greentfrapp/dollar-street-images

A.1.2 Inferring bias labels

Domino. Domino [17] identifies underperforming subgroups, referred to as "slices", by employing a Gaussian mixture model (GMM) in the CLIP embedding space. Following the parameters suggested in the paper, we use a log-likelihood weight of 10 for y and $\hat{y}$, and set the number of slices to 2. We train slicing functions on the validation data for each class and then apply these learned slicing functions to the test data, resulting in soft slice assignments. The soft slice assignments are utilized to construct the AUROC curve.

Failure Direction. Failure Direction [30] distills model failure modes using a linear support vector machine (SVM) to identify error patterns and represents them as directions within the CLIP feature space. We train class-wise SVMs on the validation data to obtain decision values for the test data, which are then used to construct the AUROC curve.

B2T (ours). For the CelebA dataset, we determine whether a training sample belongs to the "man" group or not. For the Waterbirds dataset, we determine the background of a training sample as either "land" or "water". To effectively utilize the zero-shot classifier, we employ several techniques. Firstly, we use the general templates provided in the official CLIP ImageNet zero-shot classification[8]. Secondly, we incorporate dataset-specific templates for improved information extraction. Lastly, we employ various B2T keywords as group names for classification. The prompts are generated in the format of `"[general template]+[dataset-specific template]+[group name],"` such as "a photo of a bird in a forest". We use the CLIP ResNet-50 model. Table 7 presents the complete list of prompt templates and group names used.

A.2. Debiasing classifiers

Debiased DRO training. We train DRO-B2T models following the protocol of [66]. We utilize the SGD optimizer with a momentum of 0.9 to train ImageNet pre-trained ResNet-50 models on both datasets. For the CelebA dataset, we use a batch size of 64, a learning rate of 1e-5, a weight decay of 0.1, a group adjustment of 0, and train for 50 epochs. For the Waterbirds dataset, we use a batch size of 128 and train for 300 epochs. We sweep the hyperparameters (learning rate, weight decay, group adjustment) in the search space $\{(1e-3, 1e-4, 0), (1e-4, 0.1, 0), (1e-5, 1.0, 0), (1e-5, 1.0, 1), (1e-5, 1.0, 2), (1e-5, 1.0, 3), (1e-5, 1.0, 4), (1e-5, 1.0, 5)\}$ with validation worst-group accuracy. We report the average and worst-group test accuracies at the epoch with the best validation worst-group accuracy.

CLIP zero-shot prompting. We augment prompt templates by adding B2T-inferred bias keywords to the end. Additionally, we utilize general templates provided for ImageNet zero-shot classification and dataset-specific templates to leverage the CLIP zero-shot classifier. Table 8 presents the complete augmented templates with bias keywords that have positive CLIP scores. For example, a prompt for the landbird class in the Waterbirds dataset is "a photo of a landbird in the forest." We generate ensembles of all possible prompt combinations while inferring the group labels. We use a pre-trained CLIP model with a ResNet-50 image encoder.

A.3. Ablation studies

Captioning models. We use the ClipCap[9] [52] model trained on Conceptual Captions [69] without beam search as our captioning model if not specified. We employ the BLIP [40] base captioning model trained on COCO and BLIP-2 utilizing the OPT-2.7b architecture from the LAVIS repository [10]. For CoCa [84], we use ViT-L-14 backbone pretrained on the LAION-2b dataset from the open CLIP repository [11], and for LLaVA [45], we use v1.5-13B that was trained in September 2023.

Scoring models. We use the CLIP model with the ViT-L/14 backbone from the CLIP repository [12]. We employ OpenCLIP [10] with the ViT-L/14 backbone trained on the LAION-2b dataset [67], and the base version of BLIP [40] and the pretrain version of BLIP-2 [41] from the LAVIS repository [13].

Keyword extraction. We apply the YAKE[14] [7] algorithm to extract bias keywords from a corpus of mispredicted or generated samples. The maximum n-gram size is 3, and we select up to 20 keywords with a deduplication threshold of 0.9. For high-frequency words, we lemmatize each word using WordNet [49] to count words.

[8] https://github.com/openai/CLIP/blob/main/notebooks/Prompt_Engineering_for_ImageNet.ipynb

[9] https://github.com/rmokady/CLIP_prefix_caption

[10] https://github.com/salesforce/LAVIS

[11] https://github.com/mlfoundations/open_clip

[12] https://github.com/openai/CLIP

[13] https://github.com/salesforce/LAVIS

[14] https://github.com/LIAAD/yake

Table 7. Prompt designs for inferring group labels.

Dataset	Dataset-wise Template	Group Name
CelebA	• `[group name]` • `[group name] celebrity`	1. Male • `man` • `male` 2. Non-male • Empty string `""`
Waterbirds	• `[group name]` • `bird on [group name]` • `bird on a [group name]` • `bird and a [group name]` • `fowl on [group name]` • `fowl on a [group name]` • `fowl and a [group name]`	1. Land background • `forest` • `woods` • `tree` • `branch` 2. Water background • `ocean` • `beach` • `surfer` • `boat` • `dock` • `water` • `lake`

Table 8. Prompt designs for debiaisng zero-shot classifiers.

Dataset	Dataset-wise Template	Class Name
CelebA	• [class name] • [class name] man • [class name] player • [class name] person • [class name] artist • [class name] comedy • [class name] film • [class name] actor • [class name] face	1. Blond • blond hair • celebrity of blond hair 2. Non blond • non blond hair • celebrity of non blond hair
Waterbirds	• [class name] • [class name] on the forest • [class name] with woods • [class name] on a tree • [class name] on a branch • [class name] in the forest • [class name] on the tree • [class name] on the ocean • [class name] on a beach • [class name] on the lake • [class name] with a surfer • [class name] on the water • [class name] on a boat • [class name] on the dock • [class name] on the rocks • [class name] in the sunset • [class name] with a kite • [class name] on the sky • [class name] is on flight • [class name] is on flies	1. Landbird • landbird 2. Waterbird • waterbird

B. Extension to Generative Models

We extend the B2T framework to text-to-image (T2I) generative models [65]. Here, we define biases as spurious correlations between input conditions and generated attributes [56], i.e., unintended attributes not explicitly specified through prompts.

B.1. B2T for text-to-image (T2I) generative models

T2I generative models produce an image $x \in \mathcal{X}$ from a given text description $y \in \mathcal{Y}$. Our goal is to identify a biased attribute $a \in \mathcal{A}$ that is spuriously correlated with the input prompts, i.e., the generated images x contain the biased attribute a even though it is not explicitly described in y. For example, a generative model may produce only female images when conditioned on blond, suggesting that the attribute "woman" is spuriously correlated with the prompt "blond."

Bias keywords. To identify the biased attribute a, we extract common keywords from the captions of the *generated* images, rather than the mispredicted ones for classifiers. The keywords that appear in the generated images can be either the intended text y or unintended bias a, and the user can infer the candidate set of biased keywords. In the case of the generative model conditioned on the prompt "blond," the keywords "woman" (as well as "blond") will frequently appear.

SD score. To validate whether the keywords represent biases, we define a score analogous to the CLIP score. Our score relies on the underlying generative model being a T2I diffusion model [28], but it could be extended to other generative models in principle. In our experiments, we use Stable Diffusion (SD) [65] and refer to our metric as the SD score.

The SD score measures the diffusion score between generated images and the original prompts or bias keywords, ensuring that only keywords that are already present in the generated images (and thus possibly associated with biased attributes) have a low SD score. To calculate this score, we compare the diffusion scores of generated images x conditioned on the original prompt y or bias keywords a. Intuitively, the diffusion score for the conditions y and a will be similar if the generated image x already reflects the bias keyword. The SD score is given by:

$$s_{\mathsf{SD}}(a; y) := \frac{1}{|\mathcal{D}_y|} \sum_{x \in \mathcal{D}_y} ||\mathsf{score}(x; a) - \mathsf{score}(x; y)||. \tag{3}$$

Here, $\mathcal{D}_y$ is the set of generated images conditioned on text y, $\mathsf{score}(x; y)$ is the diffusion score of an image x conditioned on text y (i.e., the gradient on the data space to update an image x to follow the condition y), and $|| \cdot ||$ denotes the ℓ_2-norm. The SD score uses the diffusion score of the generative model itself and is thus not affected by the bias in off-the-shelf captioning models. Additionally, the SD score can be interpreted as a classifier that uses the T2I diffusion model to compare the classification confidence of an image x towards the classes y and a, as explored in [12, 39].

B.2. Experimental results

Bias discovery. We apply B2T on Stable Diffusion [65] using the prompts from [19, 47], resulting in the unfair generation of images. B2T recovers known biases, such as spurious correlations between occupations and gender or race. For instance, as shown in Figure 8, Stable Diffusion associates nurses with "women" and construction workers with "men," indicating gender bias, and maids with "Asians," indicating racial bias. Moreover, B2T uncovers unknown biases from the same prompts, such as the association of nurses with "stethoscope," construction workers with "hat," and Native Americans with "feathers," suggesting that the model exhibits stereotypes based on the appearance of certain occupations and ethnicities.

Debiasing T2I diffusion. We use the bias keywords to debias a T2I diffusion model, Stable Diffusion [65]. To achieve this, we apply the Fair Diffusion [19] algorithm, which adjusts the diffusion score that used to update images during generation, in order to regulate the effects of the specified keywords. Figure 9 demonstrates that Fair Diffusion, utilizing the bias keywords discovered by B2T, effectively eliminates the biases mentioned earlier. Our approach balances the unfair generation of images. We believe that B2T can facilitate the desirable use of fair T2I generative models.

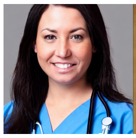
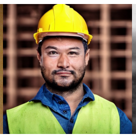
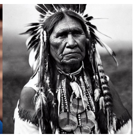
Prompt	“a photo of a face of a **nurse**”	“a photo of a face of a **maid**”	“a photo of a face of a **construction worker**”	“a photo of a face of a **native American**”
Generated images				
B2T keywords	woman, stethoscope	woman, girl, young, asian	man, hardhat, site	man, Indian, feathers

Figure 8. **Discovering biases in T2I generative models.** Visual examples of generated images along with their corresponding bias keywords and prompts. B2T successfully uncovers known biases, such as gender and race, that spuriously link to occupations [19, 47]. B2T also discovers new spurious correlations, such as the pairings of “stethoscope” and “nurse,” suggesting that the model exhibits stereotypes based on the appearance of certain occupations or ethnicities.

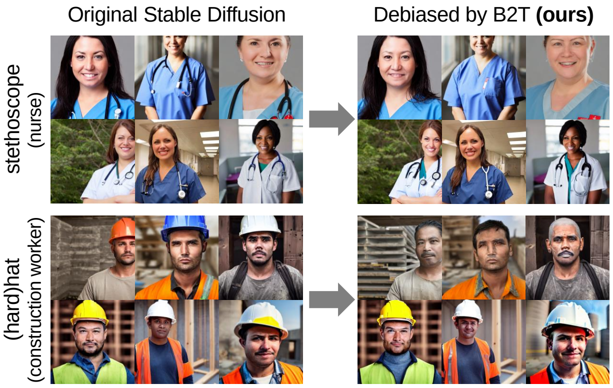

Original Stable Diffusion
Debiased by B2T **(ours)**
stethoscope (nurse)
(hard)hat (construction worker)

Figure 9. **Debiasing T2I diffusion models.** We use the bias keywords discovered by B2T to debias the spurious correlations in Stable Diffusion. B2T effectively balances the generation of the unfair attribute “stethoscope” or “(hard)hat.”

C. Additional DRO Results

C.1. Multi-class debiasing

We conduct an additional experiment on datasets of more classes. We use the 2- and 10-class setups from the MetaShift [81] dataset, which aims to address spurious correlations between the cat and dog classes, associated with the indoor and outdoor attributes, respectively. First, we apply B2T to the ERM classifier and obtain outdoor keywords like “street” and “parked” for cats, as well as indoor keywords like “room” and “sleeping” for dogs. We then perform DRO training using these keywords and compare it with the baseline ERM and the oracle DRO using ground-truth labels. The table below displays the worst-group accuracies, with variations in the weights of minority subgroups (lower p indicates stronger bias). DRO-B2T (ours) performs well for both the 2-class and 10-class scenarios.

Table 9. **Multi-class debiasing.** DRO-B2T (ours) also works with multi-class debiasing scenarios.

		2 Class			**10 Class**		
	GT	p=12%	p=6%	p=1%	p=12%	p=6%	p=1%
ERM	-	50.00	47.92	37.50	68.58	67.01	63.19
DRO-B2T (ours)	-	74.54	69.91	51.62	70.08	69.33	65.16
DRO	✓	77.78	70.60	52.55	68.75	70.66	66.32

C.2. Nonsensical groups

Defining DRO subgroups by keywords does not pose a problem without human oversight: 1) keywords with high CLIP scores represent minorities, thus defining meaningful subgroups without supervision, and 2) even if the keywords are nonsensical, the subgroups become randomly sampled subsets, not affecting the outcome of DRO. To verify this, we perform an extra DRO experiment on the CelebA blond dataset, using a nonsensical keyword “face” alongside the bias keyword “man.” The table shows that the nonsensical keyword has no impact on the results. Lastly, human monitoring is still necessary due to the subjective nature of bias, and our goal is to assist rather than replace them.

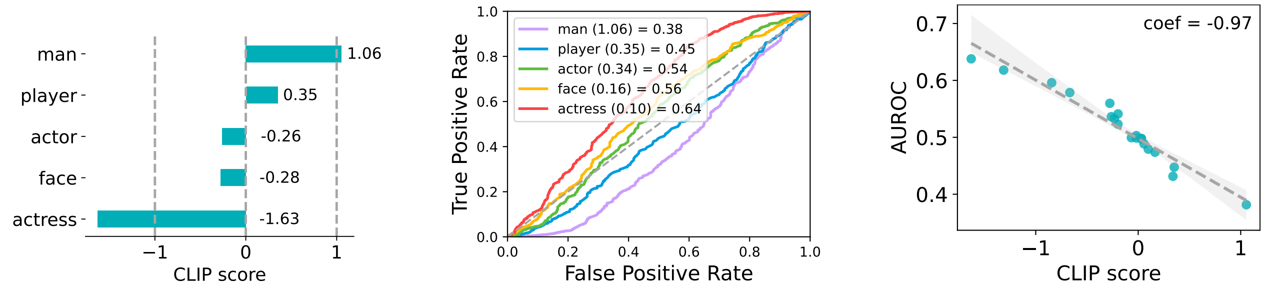
Table 10. **Nonsensical groups.** DRO-B2T (ours) also works with nonsensical group keywords.

Keyword	**Worst-group**	**Average**
man	90.37±0.32	93.02±0.31
man+face	90.00±0.96	93.15±0.20

D. Additional Analyses

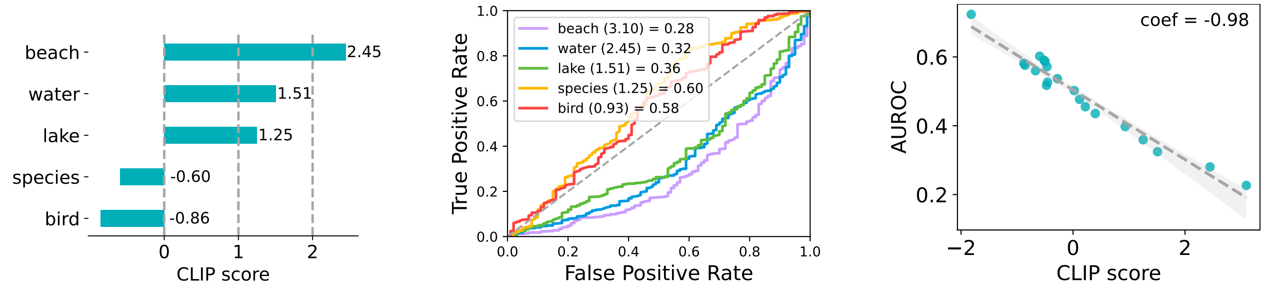
D.1. Validation of the CLIP score

We demonstrate the effect of the CLIP score using the blond class of the CelebA dataset in figure 10 and the landbird class of the Waterbirds dataset in figure 11.



(a) CLIP score (b) ROC curve of subgroup accuracy (c) Correlation of CLIP score and AUROC

Figure 10. **Effect of the CLIP score (blond class in CelebA).** We can observe similar trends with the waterbird class.

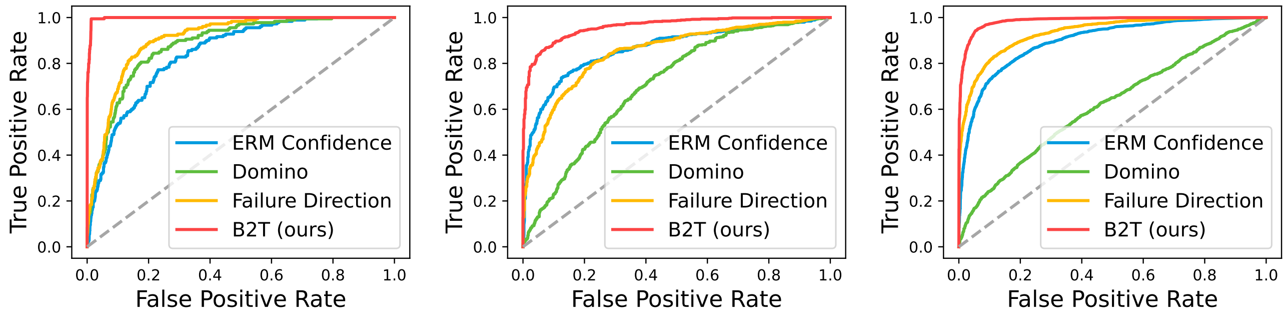


(a) CLIP score (b) ROC curve of subgroup accuracy (c) Correlation of CLIP score and AUROC

Figure 11. **Effect of the CLIP score (landbird class).** We can observe similar trends with the waterbird class.

D.2. Comparison of bias discovery methods

We compare B2T with prior unsupervised bias discovery methods: JTT [44], Domino [17], and Failure Direction [30]. Figure 12 illustrates that B2T significantly outperforms prior methods, achieving near-optimal performance across all considered scenarios.



(a) CelebA blond (b) Waterbird (c) Landbird

Figure 12. **Comparison of bias discovery methods.** The AUROC curves for (a) CelebA blond (male), (b) Waterbird (waterbird on land), and (c) Landbird (landbird on water) classes, with parentheses indicating the corresponding minority groups. B2T outperforms prior works by a large margin.

D.3. Keyword extraction

We compare the YAKE algorithm with a simple high-frequency and another popular keyword extraction algorithm, FRAKE. As depicted in Table 11, the extracted keywords are mostly shared across different methods. We observe significant biases such as "man" in CelebA blond or "forest" and "water" for Waterbirds waterbird and landbird class, respectively, across all keyword extraction methods. The 20 keywords for each method are reported in Table 11.

Table 11. Different keyword extraction methods

(a) High Frequency

	Keywords
CelebA blond	actor, person, hair, film, premiere, player, actress, face, model, comedy, former, love, woman, like, artist, style, man, want, first, contestant
Waterbird	specie, biological, bird, tree, garden, person, forest, saw, prey, one, wood, bamboo, wild, rainforest, paradise, pond, rock, wall, selected, art
Landbird	specie, biological, beach, bird, person, water, fly, seagull, rock, sky, dog, seen, lake, city, pond, parrot, yellow, one, saw, sunset

(b) FRAKE

	Keywords
CelebA blond	actor person, actor premiere comedy film, person model actress, actor, person, want hair like, hair, player, film, premiere, actress, model, face, comedy, love, man, like, style, artist, contestant
Waterbird	biological species bird prey, biological bamboo forest, species bamboo forest, biological, species, bird tree, bird, tree, person, rainforest, saw, garden, forest, photo, wild, bamboo, trees, pond, prey, woods
Landbird	biological species beach, bird flies water, bird beach, person beach, species, biological, bird, beach, person, water, flies, seagull, sits, sky, sunset, sea, paraglider, rocks, flight, city

E. Additional Model Comparisons

Multimodal learning: ERM vs. CLIP. Table 12 presents a comparison of bias keywords obtained from ImageNet-R using ViT-B models trained by ERM and CLIP. ERM identifies distribution shifts like "illustration" and "drawing" as bias keywords, which have high CLIP scores. In contrast, CLIP identifies different bias keywords such as "dog" and exhibits low CLIP scores. This suggests that CLIP is less affected by distribution shifts compared to ERM.

Table 12. Comparison of ERM vs. CLIP.

(a) ERM

	Score
hand drawn illustration	2.02
drawing	1.61
hand drawn	1.42
vector illustration	1.38
tattoo	1.27
white vector illustration	1.22
illustration	1.09
sketch	1.02
step by step	0.53
digital art	0.31

(b) CLIP

	Score
dog	0.64
art	0.55
art selected	0.53
person	0.48
tattoo	0.48
drawing	0.45
painting	0.42
step by step	0.36
made	0.31
digital art selected	0.30

Self-supervised learning: ERM vs. DINO vs. MAE. Table 13 presents a comparison of bias keywords obtained from ImageNet-R using ViT-B models trained by DINO [8] and MAE [23], along with ERM mentioned earlier. DINO provides similar bias keywords to ERM, while MAE provides keywords with low CLIP scores. Intuitively, both ERM and DINO demonstrate less robustness to distribution shifts than MAE.

Table 13. Comparison of DINO vs. MAE.

(a) DINO

	Score
hand drawn illustration	2.13
drawn vector illustration	2.06
cartoon illustration	1.86
white vector illustration	1.70
vector art illustration	1.63
vector illustration	1.53
tattoo	1.48
white background vector	1.38
art	0.97
digital art	0.90

(b) MAE

	Score
drawn vector illustration	1.69
cartoon illustration	1.61
tattoo	1.45
white vector illustration	1.31
vector art illustration	1.27
vector illustration	1.20
drawing	1.20
white background vector	1.05
art	0.84
person	0.78

F. Further Discussion of Limitations

We discover the bias of image classifiers using captioning (e.g., ClipCap [52]) and scoring (e.g., CLIP [59]) models. However, there is a potential risk that these models themselves may be biased [2, 18]. Thus, users should not fully rely on the extracted captions, and the involvement of human juries remains essential in the development of fair machine learning systems.

For instance, ClipCap and CLIP are mostly trained on natural images, and are less effective for specialized domains [51] such as medical or satellite. To check this, we apply B2T to the ChestX-ray14 [78] and FMoW [11] datasets. We use classifiers publicly released in the ChexNet[15] [62] and WILDS[16] [35] codebases, utilizing the ERM classifier seed 0 for FMoW.

Figure 13 visualizes the images and their corresponding captions. ClipCap generates nonsensical captions, such as "broken nose" for chest images or trivial captions like "city from the air" for aerial-view images. Consequently, one must train a specialized captioning model to apply B2T effectively.

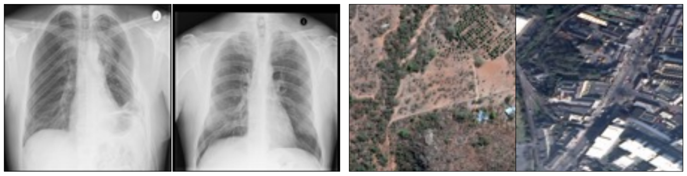
	(a) ChestX-14ray		(b) FMoW	
Samples				
Actual	no disease	disease	crop field	Parking lot or garage
Pred.	disease	no disease	debris or rubble	place of worship
Caption	a picture of a patient with a broken nose.	a picture of a woman with a broken neck.	a small village in the middle of the desert.	a city from the air.

Figure 13. Visual examples of (a) ChestX-ray14 and (b) FMoW datasets.

[15] https://github.com/arnoweng/CheXNet
[16] https://worksheets.codalab.org/worksheets/0xa96b8749679944a5b4e4e7cf0ae61dc9

G. Additional Visual Examples

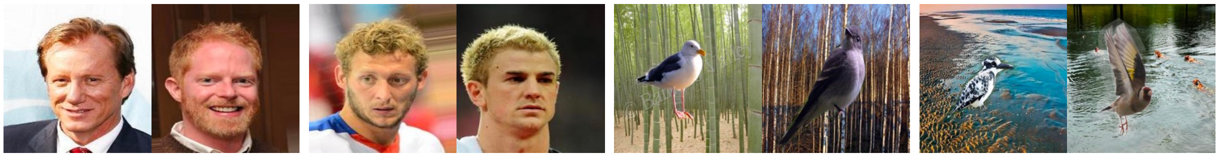
	(a) CelebA blond				(b) Waterbirds			
Keyword	Man		Player		Bamboo	Woods	Beach	Water
Samples								
Actual	blond	blond	blond	blond	waterbird	waterbird	landbird	landbird
Pred.	not blond	not blond	not blond	not blond	landbird	landbird	waterbird	waterbird
Caption	actor is a **man** of many talents.	actor, the **man** behind the voices.	the most important **player** in the history of hockey.	football **player** has been named the player of the year.	biological species in a **bamboo** forest.	biological species - i saw one of these in the **woods**.	biological species on the **beach**.	a bird in the **water**.

Figure 14. Additional visual examples of CelebA and Waterbirds.

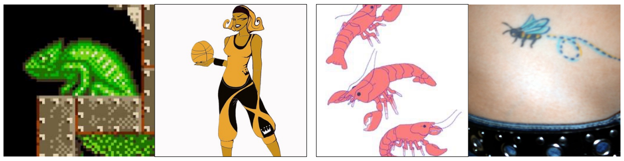
Keyword	Illustration		Drawing	
Samples				
Actual	African chameleon	basketball	American lobster	bee
Pred.	oscilloscope	knee pad	handkerchief	necklace
Caption	vector **illustration** of a frog.	cartoon **illustration** of a basketball […]	a **drawing** of a crab.	a **drawing** of a bee.

Figure 15. Additional visual examples of ImageNet-R.

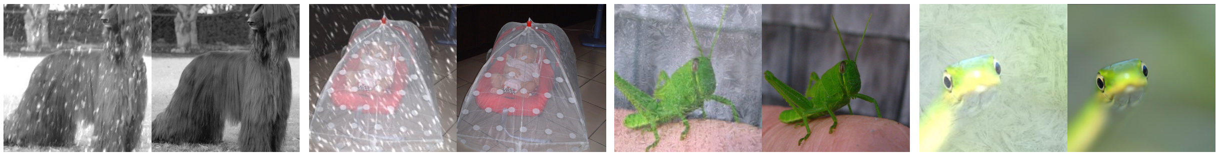
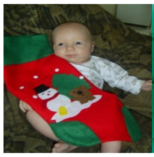
	(a) ImageNet-C snow				(b) ImageNet-C frost			
Keyword	Snow		Rain		Window		Glass	
Samples								
Actual	Afghan hound	Afghan hound	mosquito net	mosquito net	grasshopper	grasshopper	green snake	green snake
Pred.	fountain	Afghan hound	shower cap	mosquito net	African chameleon	grasshopper	rock beauty	green snake
Caption	a horse in the **snow**.	person, the dog of the day.	the umbrella in the **rain**.	the baby in the tent.	a green chameleon on a **window** sill.	a green grasshopper on my finger.	a frog in a **glass** of water.	a green frog in the jungle.

Figure 16. Additional visual examples of ImageNet-C.

Keyword	-	Cave	-	Fire	-	Bucket	-	Hole
Samples								
Actual	wardrobe	wardrobe	stove	stove	plate rack	plate rack	toilet seat	toilet seat
Pred.	wardrobe	poncho	stove	caldron	plate rack	oil filter	toilet seat	wheelbarrow
Caption	the back of the wardrobe.	the **cave** is full of surprises.	a stove for the kitchen.	a **fire** in the kitchen.	a man is putting a lot of plates in the dishwasher.	a **bucket** of water and a few tools.	a toilet in the bathroom.	the **hole** in the ground.
Country (Income)	Romania ($6256/month)	Tanzania ($32/month)	United States ($855/month)	Togo ($321/month)	India ($2499/month)	Cote d'Ivoire ($8/month)	Mexico ($898/month)	Cameroon ($137/month)

Figure 17. Visual examples of Dollar Street classes.

Keyword	Baby		Red		Pendant		Fighter	
Samples								
Actual	Christmas stocking	Christmas stocking	mushroom	mushroom	chain	chain	military aircraft	military aircraft
Pred.	baby bib	baby bib	agaric	agaric	necklace	necklace	airplane wing	airplane wing
Caption	how to make a christmas sweater for your **baby**.	a **baby** in a red christmas jumper	a **red** mushroom in the grass.	**red** mushroom in the forest.	a bracelet made from a recycled **pendant**.	this is a sterling silver **pendant**.	a **fighter** jet in flight.	a jet **fighter** in flight.

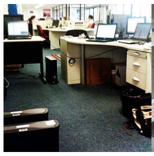
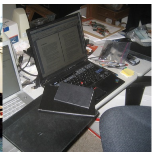
Keyword	Street		Office		Dish		Concert	
Samples								
Actual	plastic bag	plastic bag	notebook	notebook	broccoli	broccoli	electric guitar	electric guitar
Pred.	poncho	paddle	desk	laptop	plate	plate	stage	stage
Caption	a homeless man begging on the **street**s.	actor reading a book on the **street**.	the **office** of person, who is now.	the laptop in my **office**.	the finished **dish** with a side of broccoli.	the finished **dish** with the rice.	the band performs a live **concert**.	person performs on stage during the **concert**.

Figure 18. Additional visual examples of ImageNet classes.

H. Complete Lists of the B2T Keywords

Bias keywords from image classifiers, and their corresponding CLIP scores and subgroup accuracies. Higher CLIP scores and lower subgroup accuracies indicate more significant biases.

Table 14. Candidates of bias keywords for CelebA blond.

(a) Blond (base acc.: 86.0)

	Score	Acc.
man	1.22	38.2
player	0.42	27.8
person	0.17	79.8
artist	0.16	69.6
comedy	0.16	88.2
film	0.13	88.3
actor	0.08	88.2
face	0.06	88.5
love	0.06	91.3
clothing	0.05	93.5
outfit	0.05	93.5
hair	0.02	91.2
style	0.00	92.2
weight	-0.06	93.6
clothing style	-0.08	93.5
model	-0.19	95.5
premiere	-0.52	89.1
premiere of comedy	-0.63	86.2
model and actress	-1.00	82.7
actress	-1.28	83.3

(b) Not blond (base acc.: 97.2)

	Score	Acc.
model	0.50	96.9
favorite outfit	0.34	94.8
hair	0.33	94.4
love	0.17	96.7
style	0.14	94.7
premiere	0.11	98.0
clothing style	0.09	94.8
outfit	0.08	94.8
favorite	0.08	94.8
feet size	0.06	94.8
clothing	0.06	94.8
film	0.00	98.3
weight	-0.03	94.8
face	-0.05	97.3
feet	-0.06	94.8
size	-0.08	94.8
comedy	-0.25	96.5
person	-0.28	97.5
bob	-0.50	93.2
actor	-0.98	97.5

Table 15. Candidates of bias keywords for Waterbirds.

(a) Waterbird (base acc.: 75.6)

	Score	Acc.
forest	2.12	61.5
woods	1.94	62.5
tree	1.45	41.7
branch	1.20	35.7
prey	0.20	70.0
wild	0.19	75.0
bird of prey	-0.03	66.7
species	-0.05	74.2
area	-0.09	0.0
biological species	-0.11	74.2
bird in flight	-0.27	50.0
biological	-0.28	74.2
bird	-0.36	62.5
person	-0.41	81.3
bird flying	-0.42	75.0
eagle	-0.69	95.5
bald	-0.69	60.0
snow	-0.80	66.7
great bird	-0.80	0.0
large bird	-1.05	50.0

(b) Landbird (base acc.: 89.9)

	Score	Acc.
ocean	3.41	44.4
beach	2.83	74.7
surfer	2.73	55.6
boat	2.16	64.7
dock	1.56	75.0
water	1.38	75.0
lake	1.17	80.0
rocks	1.02	76.5
sunset	0.88	70.0
kite	0.67	64.6
sky	0.28	84.2
flight	0.23	62.5
flies	-0.17	73.3
person	-0.38	86.9
pond	-0.47	87.0
biological species	-0.48	95.5
biological	-0.55	93.4
species in flight	-0.92	44.4
species	-0.97	93.4
bird	-1.64	93.8

Table 16. Candidates of bias keywords for ImageNet-R and ImageNet-C.

(a) ImageNet-R (base acc.: 52.8)

	Score	Acc.
hand drawn illustration	2.02	19.4
drawing	1.61	29.2
hand drawn	1.42	20.2
vector illustration	1.38	26.1
tattoo	1.27	12.2
white vector illustration	1.22	29.2
illustration	1.09	20.5
sketch	1.02	16.2
step by step	0.53	25.8
digital art	0.31	23.2

(b) ImageNet-C snow (base acc.: 64.6)

	Score	Acc.
snow falling	3.05	27.9
rain falling	2.58	0.9
rain drops falling	2.52	26.1
rain drops	2.25	26.7
rain	2.14	51.6
snow	1.83	54.2
water drops	1.52	32.3
falling	1.33	27.5
water	1.02	51.1
day	0.53	67.9

(c) ImageNet-C frost (base acc.: 67.7)

	Score	Acc.
room	0.97	53.4
glass	0.83	47.1
window	0.81	55.9
snow	0.70	70.8
water	0.58	65.5
person playing	0.52	65.3
tree	0.39	72.2
person	0.33	65.7
dogs playing	0.31	50.0
car	0.31	62.4

Table 17. Candidates of bias keywords for Dollar Street.

(a) Wardrobe (base acc.: 60.7)

	Score	Acc.
cave	1.83	0.0
laundry	1.05	33.3
man	0.67	0.0
pile	0.34	50.0
sleeps	0.30	0.0
living	-0.01	0.0
shed	-0.48	0.0
clothes	-0.99	72.7
full	-1.10	71.4
room	-1.22	58.3

(b) Stove (base acc.: 50.0)

	Score	Acc.
burns	0.90	0.0
fire	0.80	0.0
fireplace	0.05	0.0
cat	0.04	0.0
sits	-0.17	0.0
room	-0.17	0.0
small	-0.51	50.0
sink	-0.62	0.0
kitchen	-1.60	50.0
stove	-1.64	61.5

(c) Plate rack (base acc.: 24.3)

	Score	Acc.
bucket	0.78	3.8
water	0.78	3.0
small	0.13	25.0
sink	0.03	29.5
food	-0.02	11.8
full	-0.51	21.6
laundry	-0.52	22.2
kitchen	-1.13	32.3
dishes	-1.41	25.0
collection	-1.42	21.1

(d) Toillet seat (base acc.: 46.0)

	Score	Acc.
hole	0.65	0.0
house	0.10	62.3
property	0.04	80.0
basement	-0.09	42.9
man	-0.17	42.9
image	-1.03	81.6
small	-1.03	23.5
room	-1.50	58.1
bathroom	-3.50	59.6
toilet	-4.70	71.4

Table 18. Candidates of bias keywords for ImageNet.

(a) Ant (base acc.: 30.0)

	Score	Acc.
flowers	1.08	14.7
flower	1.03	20.9
bee	0.99	12.9
tree	0.86	19.1
spider	0.78	29.5
fly	0.75	24.2
beetle	0.58	30.4
leaf	0.32	27.3
close	0.12	33.3
black	0.10	18.1

(b) Horizontal bar (base acc.: 70.8)

	Score	Acc.
swings	7.01	6.3
playground	5.09	9.5
park	4.63	3.6
swing	4.31	12.5
child	3.47	27.7
plays	2.83	20.7
girl	2.52	22.0
playing	2.14	4.1
person	1.35	65.2
boy	1.10	20.0

(c) Stethoscope (base acc.: 69.1)

	Score	Acc.
baby	1.23	24.4
boy	1.23	28.0
girl	0.71	36.2
person	0.51	36.2
student	0.44	30.4
nurse	0.01	72.9
doctor	-0.81	88.4
hospital	-0.87	56.1
medical	-0.99	88.3
stethoscope	-3.04	93.3

(d) Monastery (base acc.: 53.0)

	Score	Acc.
interior	1.12	17.6
built	0.53	54.2
cathedral	0.35	36.7
person	0.29	60.5
century	0.06	58.8
city	0.03	56.7
church	-0.01	53.3
temple	-0.16	46.9
courtyard	-0.50	58.5
town	-0.64	60.6