Discovering and Mitigating Visual Biases through Keyword Explanation

Younghyun Kim*[1] Sangwoo Mo*[2] Minkyu Kim[3] Kyungmin Lee[1] Jaeho Lee[4] Jinwoo Shin[1]
[1]KAIST [2]University of Michigan [3]KRAFTON [4]POSTECH
younghyun.kim@kaist.ac.kr swmo@umich.edu

Abstract

Addressing biases in computer vision models is crucial for real-world AI deployments. However, mitigating visual biases is challenging due to their unexplainable nature, often identified indirectly through visualization or sample statistics, which necessitates additional human supervision for interpretation. To tackle this issue, we propose the Bias-to-Text (B2T) framework, which interprets visual biases as keywords. Specifically, we extract common keywords from the captions of mispredicted images to identify potential biases in the model. We then validate these keywords by measuring their similarity to the mispredicted images using a vision-language scoring model. The keyword explanation form of visual bias offers several advantages, such as a clear group naming for bias discovery and a natural extension for debiasing using these group names. Our experiments demonstrate that B2T can identify known biases, such as gender bias in CelebA, background bias in Waterbirds, and distribution shifts in ImageNet-R/C. Additionally, B2T uncovers novel biases in larger datasets, such as Dollar Street and ImageNet. For example, we discovered a contextual bias between "bee" and "flower" in ImageNet. We also highlight various applications of B2T keywords, including debiased training, CLIP prompting, and model comparison.[1]

1. Introduction

Biased datasets can induce failures in image classifiers, potentially harming model performance and raising fairness concerns [76]. These model failures may manifest as spurious correlations, where specific groups contribute to model errors [71], or as distribution shifts, where the test distribution differs from the training distribution [63]. For instance, in a face dataset, if blond images are predominantly associated with women, the image classifier may misclassify blond faces as women, resulting in fairness issues [6]. Moreover, this bias can impact model performance when evaluated in different scenarios, such as a gender-balanced dataset of blonds [21]. Therefore, extensive efforts have been devoted to recognizing and addressing biases in models [9, 48].

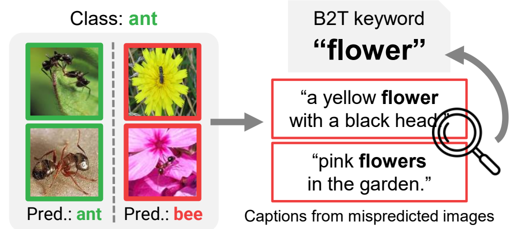


Figure 1. **Concept.** Our Bias-to-Text (B2T) framework reveals visual biases of image classifiers in a keyword explanation form. For example, B2T identified novel biases in ImageNet [14]. Specifically, the keyword "flower" implies that the classifier associates "ant" images with "flower" as "bees," indicating contextual bias.

Previous research has attempted to identify visual biases by analyzing problematic samples [44, 55, 74] or problematic attributes [29, 72, 73]. However, these methods define biases indirectly, often relying on visualization or sample groups with specific statistics, and they require human supervision to express them in an explainable form. To address this issue, recent research aimed at interpreting biases using vision-language models [59]. Nonetheless, these studies have limitations in discovering and mitigating novel biases. Some studies [17, 30] retrieve the closest word from a predefined vocabulary, limiting their discovery to known biases. Others analyze neurons [27] or images synthesized by generative models [80] to comprehend biases. However, they focus on generating detailed captions explaining activated neurons or failure examples, which can help understand individual cases but hard to utilize for debiasing.

Instead, our main idea is to explain visual biases as *keywords* by aggregating common traits from the language descriptions[2] of problematic images. Figure 1 illustrates our concept, with the keyword "flower" capturing the distinctive attributes of "ant" class images mispredicted as "bee." This keyword form offers several advantages, providing a natural name for each bias group and easily incorporating with debiasing techniques using these group names.

*Equal Contribution
[1]Code: https://github.com/alinlab/b2t

[2]We used a pre-trained captioning model, but other language descriptions like hard prompt optimization [79] can also be applied.

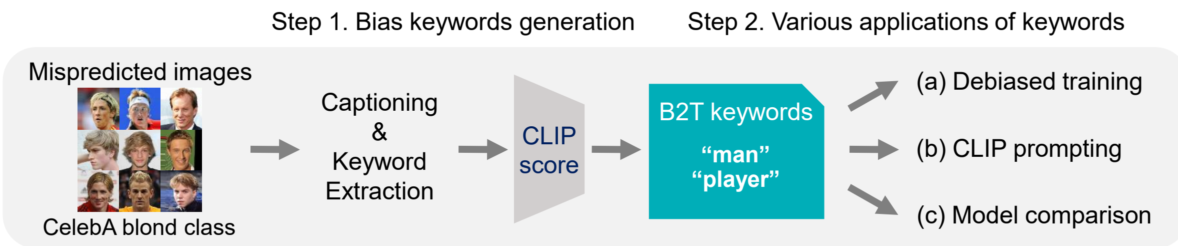


Figure 2. **Method.** (Step 1) B2T generates language descriptions from mispredicted images and extracts common keywords. We then verify whether these keywords indicate bias by measuring their similarity to the mispredicted images using a vision-language model like CLIP [59]. (Step 2) The discovered keywords have various applications, including debiased training, CLIP prompting, and model comparison.

Contribution. We introduce the *Bias-to-Text* (B2T) framework, which identifies visual biases as keywords. To achieve this, we first generate language descriptions from mispredicted images and extract common keywords from these descriptions, suggesting potential biases. We then validate whether these keywords represent bias by measuring their similarity to the mispredicted images using a vision-language scoring model such as CLIP [59]. By ensuring that the keywords align more closely with misclassified images rather than the correct ones, we can confirm they are biases.

We demonstrate that B2T can **discover biases** in image classifiers trained on various datasets (Section 4):

- *Known bias.* B2T detects popular biases, such as gender bias in CelebA [46], background bias in Waterbirds [66], and distribution shifts in ImageNet-R [26] and ImageNet-C [25]. B2T keywords provide more fine-grained information for each bias, such as "bamboo," representing the land background in Waterbirds. Moreover, these keywords can infer sample-wise bias labels, surpassing previous bias discovery approaches [17, 30].
- *Novel bias.* B2T uncovers novel biases in larger datasets, such as geographic bias in Dollar Street [64] and contextual bias in ImageNet [14]. For example, images with the keyword "flower" are predicted as "bee" instead of "ant" in ImageNet, indicating a contextual bias where bees are more commonly associated with flowers than ants.

We then showcase that the bias keywords can be used for **various applications** (Section 5):

- *Debiased training.* The keywords can be used to infer bias labels for each sample using the CLIP classifier. These labels are then used for debiased training, such as distributionally robust optimization (DRO) [66], and it outperforms previous debiasing approaches.
- *CLIP prompting.* The keywords can be used to improve the CLIP zero-shot classifier. Prompting with fine-grained B2T keywords (e.g., "bamboo") outperforms the previous strategy using group names (e.g., "land").
- *Model comparison.* The keywords can be used to compare the failure of different models. For example, ResNet [22] struggles more with complex scenes compared to ViT [16], as indicated by abstract keywords like "work out."
- *Label diagnosis.* B2T can detect issues in labels, such as mislabeling or label ambiguities. For example, we found that "bee" is often mislabeled as "fly" in ImageNet.

Lastly, we emphasize the robustness and versatility of our bias-to-text approach. First, we confirm that B2T keywords exhibit reasonable robustness across different captioning and similarity scoring models (Section 6), yet could be improved using advanced vision-language models like GPT-4 [58]. Additionally, B2T can be extended beyond image classification, such as text-to-image generative models (Appendix B) and other computer vision tasks like object detection.

2. Related Work

Bias and fairness. Biases in datasets and models have long been issued in computer vision and machine learning [48]. Our goal is to study classifier failures for specific attributes or groups, known as spurious correlations [71]. These failures are closely related to fairness concerns, as models often perform poorly on particular gender [6, 24, 90] or race [32, 37]. Such biases result from various sources, such as dataset imbalance [33, 75], or representational bias [3, 77, 89], which is further exacerbated during model training. B2T aims to identify these fairness issues, providing "man" as a bias keyword for the "blond" class in CelebA [66].

Not only is bias related to fairness, but it also significantly impacts generalization, particularly in the presence of distribution shifts [54]. The ratio of majority to minority samples can vary, making the model susceptible to changes in their composition. This is closely connected to shortcut learning, where the model excessively relies on spurious features rather than core features [21]. Various types of shortcuts exist, including texture bias [20], background bias [83], and scene bias [50]. B2T could discover various types of shortcuts, as exemplified by "illustration" in ImageNet-R, "forest" in Waterbirds, "flower" for "ant" class in ImageNet.

Bias discovery. Previous studies attempted to identify biases by analyzing problematic samples [4, 5, 15, 31, 36, 38, 61, 82]. Specifically, they detected biased samples by simply retrieving the mispredicted images [44] or utilizing embeddings or gradients [1, 55, 74]. To uncover unknown biases, prior works iteratively trained a discoverer and classifier [43] or selected confident samples using two auxiliary biased models [88]. Another line of research analyzed problematic attributes to interpret spurious correlations and visualized them by highlighting specific regions [29, 72, 73], or generating traversal images alongside the attribute [42]. However, these methods still require human supervision to comprehend the common traits among failure cases, unlike B2T, which provides a practical keyword explanation.

Bias discovery with language. Recent works describe bias using pre-trained vision-language models like CLIP [59]. They define bias as an outlier (or slices) in the joint image-text embedding space [17, 30, 87]. However, they match the outliers to a pre-defined bias vocabulary, limiting their ability to detect a single known bias. In contrast, B2T directly generates captions from images, potentially containing more detailed descriptions than the encoder embeddings. Thus, B2T effectively discovers multiple and fine-grained biases without the need for an iterative discovery procedure.

Other works analyze neurons [27] or images synthesized by generative models [80] to understand biases. In particular, Wiles et al. [80] extract captions from synthesized images, similar to B2T. However, they provide detailed sentence descriptions, which are informative but not straightforward for debiasing. In contrast, the keyword explanation of B2T is more practical, as demonstrated in our applications, such as debiasing. Additionally, they need to specify a pair of true and mispredicted target classes, which may be challenging to scale if there are many classes. In contrast, B2T can find bias keywords for all failure cases simultaneously.

Debiasing classifier. Numerous efforts have been made to mitigate biases of classifiers. DRO [60, 66] is a popular approach that minimizes the loss over all bias groups. However, DRO requires bias annotations for all samples. Some works addressed this issue by inferring the bias group labels in an unsupervised manner [44, 55]. We demonstrate that the keyword explanations of B2T can infer bias labels using a zero-shot classifier like CLIP. This enables more accurate bias group estimation and improved debiased training compared to previous methods, as shown in Section 5.1.

Moreover, we demonstrate a prompting strategy to debias CLIP using the B2T keywords, which are more fine-grained than those in prior work [85], as shown in Section 5.2.

3. Bias-to-Text (B2T) Framework

In this section, we begin by defining the biases we aim to address (3.1). Then, we introduce the Bias-to-Text (B2T) framework, which provides bias keywords using a captioning model and validates them with a scoring model (3.2). Finally, we validate the effect of the scoring model, showing that keywords with high scores tend to exhibit stronger bias (3.3).

3.1. Problem formulation

Image classifiers predict a class $y \in \mathcal{Y}$ for an image $x \in \mathcal{X}$. If images with attribute a are frequently misclassified from their original class y, we refer to attribute a as a bias associated with class y. Our goal is to identify this biased attribute a in the *keyword* explanation form.

The bias attributes include spurious correlation [71] or distribution shifts [70]. Spurious correlations lead models to rely on unintended decision rules (e.g., associating "blond" hair color with "man"), resulting in incorrect predictions when the rule does not apply [66]. On the other hand, distribution shifts (e.g., style transfers like "illustration") can impede model generalization in unseen samples [26].

3.2. Discovering bias keywords

Bias keywords. Our core idea is to extract keywords that represent biases. To achieve this, we extract common keywords from the language descriptions of class-wise *mispredicted* images. Minority subgroups are those misclassified from the original class y and thus often appear in these descriptions. For example, in the case of the blonds vs. not-blonds classifier, the keyword "man" would frequently appear in the mispredicted images of the "blond" class.[3] We employ a pre-trained captioning model [41, 84] to generate descriptions and extract common keywords. We chose ClipCap [52] as our default captioning model because of its strong performance and fast inference speed (see Table 4), and apply the YAKE [7] algorithm to extract keywords.

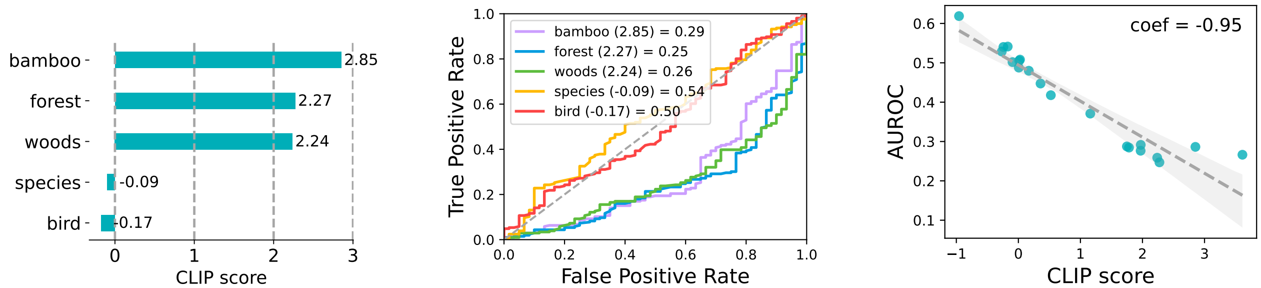
CLIP score. We validate whether the keywords represent bias. To do this, we use a vision-language scoring model like CLIP [59] that measures the similarities between keywords and the mispredicted images. The CLIP score ensures that keywords associated with a biased concept have a high CLIP score, while others have a low score. Specifically, we compare the CLIP embedding similarities between the keyword a and images x from $\mathcal{D}_{\textsf{wrong}}$ and $\mathcal{D}_{\textsf{correct}}$. These subsets of the class-wise validation set $\mathcal{D}$ correspond to the incorrect and correct predictions by the classifier, respectively. Formally, the CLIP score is given by:

$$s_{\textsf{CLIP}}(a; \mathcal{D}) := \textsf{sim}(a, \mathcal{D}_{\textsf{wrong}}) - \textsf{sim}(a, \mathcal{D}_{\textsf{correct}}). \quad (1)$$

Here, $\textsf{sim}(a, \mathcal{D})$ is the similarity between the keyword a and the dataset $\mathcal{D}$, computed as the average cosine similarity between normalized embeddings of a word $f_{\textsf{text}}(a)$ and images $f_{\textsf{image}}(x)$ for $x \in \mathcal{D}$, where

$$\textsf{sim}(a, \mathcal{D}) := \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} f_{\textsf{image}}(x)\, f_{\textsf{text}}(a). \quad (2)$$

[3] Technically, the keywords discovered by B2T are the *opposite* of the biased concept. For example, B2T finds the keyword "man" for the blond class in a hair classifier. Here, "woman" is a bias-aligned (as an opposite of bias-conflicting) attribute following the definition of Nam et al. [55].



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

Figure 3. **Effect of the CLIP score (waterbird class).** (a) The CLIP score can identify incorrect bias keywords, showing low CLIP scores near zero for non-bias keywords like "species." (b) The ROC curve represents subgroup accuracy, which defines the subgroup based on images with high CLIP similarity to specific keywords while varying the thresholds. The legend displays the B2T keywords alongside their corresponding CLIP scores in parentheses, with the AUROC of their respective curves denoted after the equal sign. Keywords with high CLIP scores tend to exhibit low subgroup accuracies, indicating they are biases. (c) Colored dots illustrate the negative correlation between the CLIP score and AUROC of subgroup accuracy over B2T keywords, indicating that a higher CLIP score implies stronger bias.

We referred to this as the CLIP score because we used CLIP as our default choice, but note that other vision-language models [10, 40] also work well (see Table 5).

Further experimental details are provided in Appendix A. While we primarily focus on image classifiers in this paper, our principle of interpreting visual biases as keywords can extend to other computer vision tasks, such as text-to-image generative models, as discussed in Appendix B.

3.3. Validation of the CLIP score

We demonstrate the effect of the CLIP score in validating whether a keyword represents bias. Figure 3 displays several analyses on the CLIP score using the waterbird class in the Waterbirds [66] dataset. Panel (a) illustrates how the CLIP score identifies incorrect bias keywords. For instance, when the captioning model generates terms like "species" or "bird," the CLIP score categorizes them as non-bias keywords due to their presence in both correctly and incorrectly predicted images, resulting in a low CLIP score.

Panel (b) displays the subgroup accuracy for each keyword. We use the CLIP similarity of individual samples associated with each keyword to define the subgroup. Subgroup accuracy is defined here as the AUROC calculated across different thresholds of CLIP similarity. Keywords with high CLIP scores (in parentheses) exhibit lower subgroup accuracies (after equal signs). For example, the keyword "bamboo" has a CLIP score of 2.85 and a subgroup accuracy of 0.29. In contrast, common keywords with CLIP scores near 0 (e.g., "bird") demonstrate performance similar to random guessing (grey dotted line), suggesting that they are not biased.

Panel (c) visualizes the correlation between the CLIP score and subgroup accuracy (AUROC) for B2T keywords. These metrics have a high correlation coefficient of -0.95, indicating that the CLIP score reflects the severity of bias in keywords. See Appendix D for further evaluations.

4. Discovering Biases in Image Classifiers

We demonstrate that B2T discovers visual biases in image classifiers trained on various datasets. First, we illustrate that B2T can identify known biases in benchmark datasets (4.1). Then, we show how bias keywords infer sample-wise bias labels using the CLIP classifier, outperforming previous methods (4.2). Finally, we explore the capacity of B2T to uncover novel biases in larger datasets (4.3).

4.1. Can B2T identify the known biases?

Spurious correlation. We use B2T to analyze gender and background biases in the CelebA [46] and Waterbirds [66] datasets. CelebA contains facial images of celebrities with attribute annotations. Following Sagawa et al. [66], we focus on classifying hair color as "blond" or "not blond." Waterbirds comprise images of waterbirds and landbirds on land or water backgrounds. Here, we apply B2T to the empirical risk minimization (ERM) classifiers [66], which are known to be affected by spurious correlations.

Figure 4 (a, b) displays the bias keywords. B2T captures "man" for CelebA blond and "forest" and "ocean" for Waterbirds, revealing gender and background biases in each dataset. Furthermore, B2T finds fine-grained keywords like "bamboo," providing more detailed information than the original "land" background annotations.

Distribution shifts. B2T can detect distribution shifts in ImageNet variants: ImageNet-R (rendition) [26], which contains artistic images of ImageNet classes, and ImageNet-C (corruption) [25], which contains noisy images of ImageNet classes. We use a supervised ResNet-50 [22] classifier trained on ImageNet, which often struggles to generalize to these datasets, indicating its bias towards the training data. We apply B2T to the union of ImageNet and each variant (not class-wise) to identify the failures of the classifier.

Figure 4 (c, d) displays the bias keywords. For ImageNet-

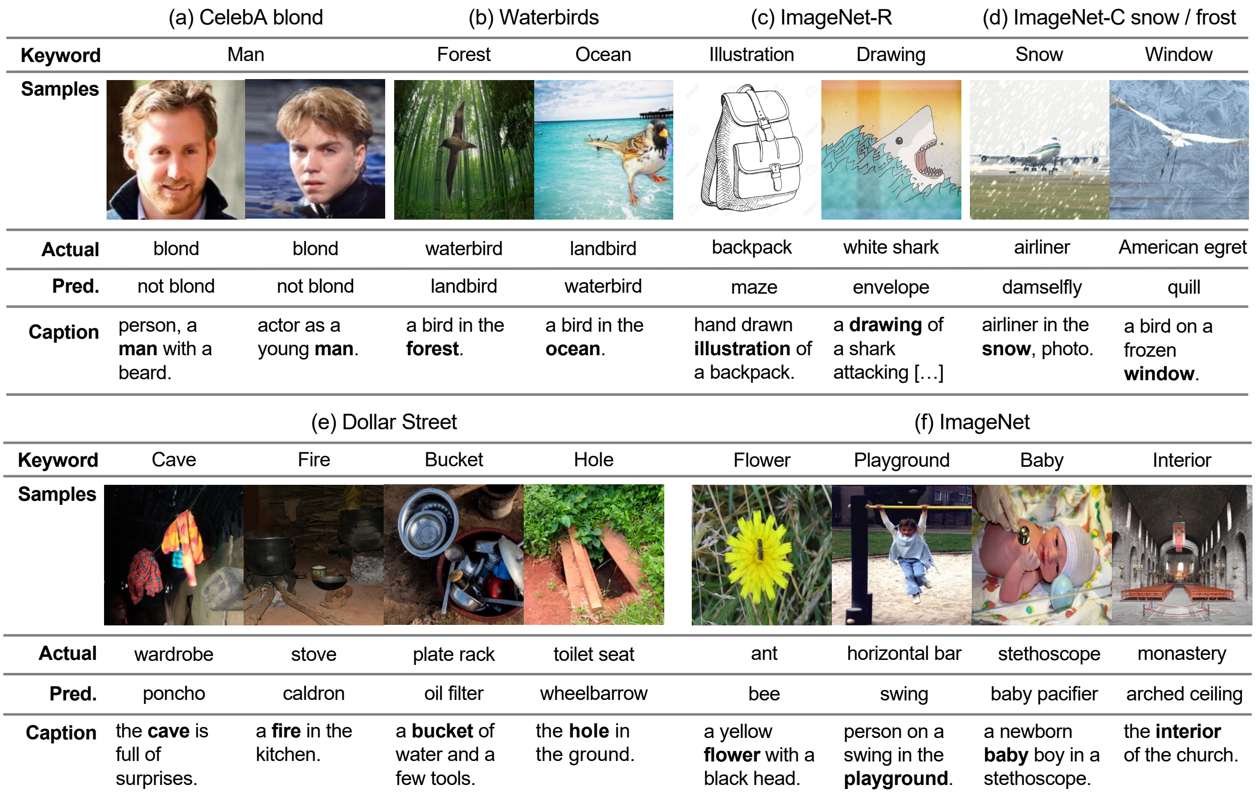


Figure 4. **Discovered biases in image classifiers.** Visual examples of mispredicted images, along with their corresponding bias keywords, captions, actual classes, and predicted classes. B2T successfully identified known biases, such as (a) gender bias in CelebA blond, (b) background bias in Waterbirds, and distribution shifts in (c) ImageNet-R with different styles, and (d) ImageNet-C with natural corruptions. B2T also uncovered novel biases in larger datasets, such as the spurious correlations between (e) the keyword "cave" and the wardrobe class, indicating geographical bias in Dollar Street, and (f) the keyword "flower" and the ant class, indicating contextual bias in ImageNet.

R, B2T captures keywords like "illustration" and "drawing," with more detailed information such as "hand-drawn" and "vector art." For ImageNet-C, B2T captures keywords like "snow" for snow corruption, and "window" for frost corruption. Here, the keyword "window" implies that the frozen images visually resemble those behind the window.

4.2. Sample-wise bias labeling

We can infer sample-wise bias (or group) labels by applying the bias keywords to the CLIP zero-shot classifier. Specifically, we create prompts like "a photo of a [group]," where "[group]" represents the bias keywords, and we assign the label of each image to the nearest group.[4]

We then evaluate this sample-wise bias labeling in CelebA and Waterbirds, where ground-truth bias labels are available. We compare B2T with prior unsupervised bias discovery methods: JTT [44], Domino [17], and Failure Direction [30]. These methods use ERM confidence, GMMs, and SVMs to predict the bias labels, respectively. Figure 5 illustrates that B2T significantly outperforms prior methods, achieving near-optimal performance across all considered scenarios.

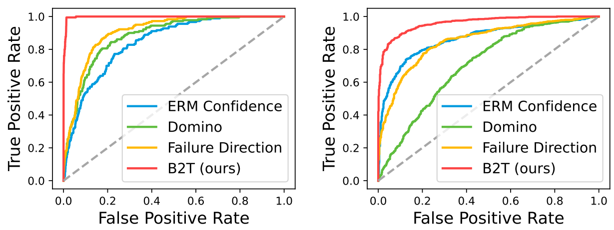
[4]The prompt template depends on the dataset. For example, we use "a photo of a bird in the [group]" for Waterbirds. We select the bias keywords with CLIP scores larger than 1.0, like "forest" or "woods" for the waterbird class. See Table 7 in Appendix for the detailed prompt templates.

4.3. Exploring novel biases in larger datasets

We apply B2T to discover novel biases in larger datasets. Note that B2T generates keywords from captioning models in a zero-shot manner, thus not requiring a pre-defined set of potential bias keywords, unlike prior works [17, 30].

Dollar Street. Dollar Street [64] includes object images from countries with varying income levels. Previous studies have shown that classifiers perform poorly on objects from low-income countries [13]. We aim to examine this geographic bias further by applying B2T to the validation set of Dollar Street using the ImageNet [14] classifier. The classifier correctly predicted labels for objects from high-income countries but failed for low-income countries.

Figure 4 (e) displays the bias keywords. Here, B2T discovers bias keywords like "cave" for "wardrobe," and "fire" for "stove" classes. Wardrobes from low-income countries are often in dark places resembling caves, and stoves from



(a) CelebA blond (b) Waterbird

Figure 5. **Comparison of bias discovery methods.** The AUROC curves for (a) CelebA blond (male) and (b) Waterbirds (waterbirds on land), with parentheses indicating the corresponding minority groups. B2T outperforms prior works by a large margin.

low-income countries often have a traditional design using fire. The keyword "bucket" for "plate rack" class suggests that buckets can be commonly used for stacking plates, and "hole" for "toilet seat" class suggests that the classifier is not familiar with squat toilets. These distinctions in objects across countries lead to geographical bias.

ImageNet. We apply B2T to the ImageNet [14] training set using the CLIP [59] zero-shot classifier. We employ the 80-prompts ensemble strategy following the CLIP paper. We focus on investigating the highly confusing classes frequently misclassified as a specific class.

Figure 4 (f) displays the bias keywords. We discover contextual biases between objects in the scene. For instance, the classifier predicts "ant" with the keyword "flower" as "bee," indicating a stronger association between flowers and bees than ants. The keyword "playground" implies that the classifier confuses a "horizontal bar" in the playground as a "swing." The classifier confuses a "stethoscope" with the keyword "baby" as a "baby pacifier," which is reasonable due to their similar appearances. In addition, we found the keywords "street" for the "plastic bag" class and "office" for the "notebook" class, suggesting that the classifier struggles in complex scenes with multiple objects.

More examples. We provide additional visual examples in Appendix G and the lists of B2T keywords in Appendix H.

5. Applications of the B2T Keywords

We showcase that the keyword form of B2T offers various applications, including debiased DRO training, CLIP zero-shot prompting, model comparison, and label diagnosis.

5.1. Debiased DRO training

Bias keywords can be used to train a debiased classifier. To be specific, we infer sample-wise bias labels as described in Section 4.2 and apply them for training with DRO [66]. We compare our DRO-B2T with various baselines, including ERM, DRO using the ground-truth (GT) bias labels, and debiased training methods that infer the group labels in an unsupervised manner: LfF [55], GEORGE [74], JTT [44], and CNC [86]. We excerpt values from the CNC paper.

Table 1 presents the worst-group and average accuracies. DRO-B2T outperforms the previous methods that infer group labels in an unsupervised manner, confirming the impact of B2T keywords. DRO-B2T also surpasses DRO using GT labels, possibly because of the noise in GT annotations. Check Appendix C for additional DRO experiments.

Table 1. **Debiased DRO training.** Worst-group and average accuracies (%) of our debiased classifier (DRO-B2T) and prior works. GT denotes the usage of ground-truth bias labels for training, and bold denotes the best worst-group accuracy. B2T keywords enable accurate bias label prediction, facilitating effective DRO training.

		CelebA blond		Waterbirds	
Method	GT	Worst	Avg.	Worst	Avg.
ERM	-	47.7±2.1	94.9	62.6±0.3	97.3
LfF [55]	-	77.2	85.1	78.0	91.2
GEORGE [74]	-	54.9±1.9	94.6	76.2±2.0	95.7
JTT [44]	-	81.5±1.7	88.1	83.8±1.2	89.3
CNC [86]	-	88.8±0.9	89.9	88.5±0.3	90.9
DRO-B2T (ours)	-	**90.4**±0.9	93.2	**90.7**±0.3	92.1
DRO [66]	✓	90.0±1.5	93.3	89.9±1.3	91.5

Table 2. **CLIP zero-shot prompting.** Worst-group and average accuracies (%) of the CLIP zero-shot classifier using the base prompt or augmented ones: with the base group names (group) or B2T keywords with positive (B2T-pos) or negative (B2T-neg) CLIP scores. Bold indicates the best worst-group accuracy. B2T-pos improves worst-group accuracy, while B2T-neg harms. This implies that augmenting proper keywords to the prompts enhances the debiased accuracy of CLIP zero-shot inference.

	CelebA blond		Waterbirds	
	Worst	Avg.	Worst	Avg.
CLIP zero-shot	76.2	85.2	50.3	72.7
+ Group prompt [85]	76.7	87.0	53.7	78.0
+ B2T-neg prompt	72.9	88.0	45.4	70.8
+ B2T-pos prompt (ours)	**80.0**	87.2	**61.7**	76.9

5.2. CLIP zero-shot prompting

Bias keywords can improve the CLIP zero-shot classifier by integrating them into prompts. In the original CLIP, the prompt template is "a photo of a [class]." We modify the prompt by adding a keyword, such as "a photo of a [class] in the [group]," where the keywords represent group names, as in the case of Waterbirds.[5] Here, we calculate the average text embeddings of prompts across all groups to get class embeddings and assign the image to the nearest class.

We augment the prompt with different sets of keywords to assess their importance. Specifically, we use B2T keywords with positive or negative CLIP scores, which we refer to as

[5]See Table 8 in Appendix A for the detailed prompt templates.

Keyword	Work		Supermarket	
Samples				
ViT-B	O	O	O	O
RN50	O	X	O	X
Actual (RN50)	dumbbell	dumbbell	shopping basket	shopping basket
Pred (RN50)	dumbbell	horizontal bar	shopping basket	grocery store
Caption	a set of dumbbells with weights.	person **works** out in the gym.	a basket full of food.	woman shopping in a **supermarket**.

Figure 6. **Model comparison: ResNet vs. ViT.** We compare the predictions made by ResNet and ViT, both trained and evaluated on ImageNet. We report their predicted labels and B2T keywords from ResNet. ViT excels at understanding global contexts and handling fine-grained classes than ResNet. For example, ResNet struggles with complex images whose B2T keywords represent abstract contexts like "work out" and "supermarket."

B2T-pos and B2T-neg, respectively. For instance, in the case of Waterbirds, we use "ocean" for B2T-pos and "bird" for B2T-neg. B2T-pos keywords represent the minor subgroups, which would aid in recognizing them. We compare this approach with using the base group names, such as "water" background, as suggested in Zhang and Ré [85].

Table 2 presents the worst-group and average accuracies of the CLIP classifier. B2T-pos keywords enhance both worst-group and average accuracies. In contrast, base group names provide less assistance, and B2T-neg keywords even decrease the worst-group accuracy. This suggests that augmenting appropriate keywords to the prompts improves the debiased accuracy of CLIP zero-shot inference.

5.3. Model comparison

Bias keywords can be used to analyze and compare different classifiers based on their keywords.

Architecture: ResNet vs. ViT. We compare ResNet [22] and ViT [16] architectures. Recent studies claim that ViT is better than ResNet in understanding object shapes [57]. We further investigate this by examining bias keywords. We train and evaluate the models on ImageNet.

Figure 6 demonstrates the comparison results. ViT excels in understanding global contexts and fine-grained classes compared to ResNet. For instance, ViT successfully predicts complex images with abstract bias keywords like "work out." We attribute this to the global self-attention of ViT, which allows for broader context consideration.

Debiased training: ERM vs. DRO. We compare biased and debiased training methods: ERM and DRO [66], on CelebA and Waterbirds. We list the bias keywords from ERM with

Table 3. **Model comparison: ERM vs. DRO.** We compare biased (ERM) and debiased (DRO) classifiers on CelebA and Waterbirds. We present the CLIP scores for ERM, DRO, and the gap between them. We mark ✗ if the bias keyword is not found. In DRO, either the bias keyword is absent or its score is reduced; for example, the keyword "man" is no longer present in CelebA blond.

	Keyword	ERM	DRO	Gap
CelebA blond	man	1.06	✗	✗
Waterbird	bamboo forest	3.61	✗	✗
	bamboo	2.85	✗	✗
	forest	2.27	1.97	-0.30
	woods	2.24	1.88	-0.36
Landbird	seagull	3.10	1.85	-1.24
	beach	2.45	1.15	-1.30
	water	1.51	0.67	-0.84
	lake	1.25	✗	✗

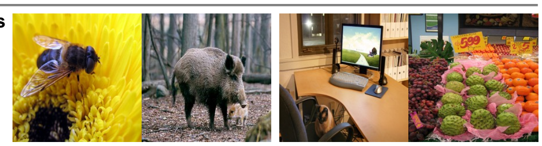
Keyword	Bee	Boar	Desk	Market
Samples				
Label	fly	pig	computer mouse	custard apple
Pred.	bee	wild boar	desktop computer	grocery store
Caption	a **bee** on a yellow flower.	wild **boar** in the forest.	the **desk** in the office.	fruit and vegetables at the **market**.

Figure 7. **Label diagnosis.** We identify labeling errors, such as mislabeling and label ambiguities, in ImageNet using bias keywords. For example, the keyword "bee" implies that the images labeled as "fly" class are actually mislabeled. On the other hand, the keyword "desk" indicates that the images contain multiple objects, including both a "computer mouse" and a "desktop computer" on the desk, making it difficult to assign the appropriate class.

CLIP scores higher than 1.0.

Table 3 illustrates the CLIP scores of ERM, DRO, and their gap. We mark ✗ if the keyword is not found. DRO indeed yields fewer bias keywords. For example, the keyword "man" is absent in the CelebA blond class, and the CLIP scores of highly biased keywords are reduced, such as from 3.10 to 1.85 for "seagull" in the landbird class.

Additional model comparisons. We present additional results in Appendix E, investigating the robustness of classifiers to distribution shifts. Our findings demonstrate that in multimodal learning, CLIP is more robust than ERM, and in self-supervised learning, MAE [23] exhibits better robustness, while DINO [8] shows similarity to ERM.

5.4. Label diagnosis

B2T can diagnose common labeling errors, such as mislabeling and label ambiguities. Previous studies have shown that ImageNet contains label errors [68]. We analyze these errors following the setup in Section 4.3.

Figure 7 visualizes examples. We found mislabeled images, such as "bee" and "boar" images labeled as "fly" and "pig," respectively. We also found images with ambiguous labels, indicated by keywords "desk" and "market." They are ambiguous as the scene contains multiple objects.

6. Ablation Study

We study the effect of using different captioning and scoring models in our B2T framework. Details such as the architecture used for each model are stated in Appendix A.

Captioning models. We study the robustness of the bias keywords across different captioning models: ClipCap [52], BLIP [40], BLIP-2 [41], CoCa [84], and LLaVA [45]. Table 4 shows the results. Different captioning models agree on severe biases while offering diverse fine-grained keywords. For example, all models capture major keywords like "man", "forest", and "beach". Conversely, different models provide diverse fine-grained keywords, such as "rainforest" or "lake"; thus ensembling a few captioning models can diversify the discovered biased keywords. We use ClipCap as our default choice, given its strong performance and fast inference time. However, one could opt for advanced models like GPT-4 [58] for improved captioning.

Scoring models. We study the robustness of the CLIP score across different vision-language models. Specifically, we test CLIP trained on different datasets: OpenCLIP [10] trained on the LAION [67] dataset, and models with different architectures: BLIP [40] and BLIP-2 [41]. Table 5 shows the results. The scoring models provide consistent rankings, with high scores for keywords like "man" or "bamboo forest." We use CLIP as our default choice, but one could also consider the advanced models.

Keyword extraction. We use YAKE [7] in our experiments, but other keyword extraction strategies, such as high-frequency words also perform well (see Appendix D).

7. Conclusion

We propose B2T, a framework for identifying and mitigating biases through keyword explanation. The use of keywords offers several advantages, such as debiased training and model comparison. We hope that our B2T framework could assist in the responsible use of image recognition.

Limitations. B2T relies on the recent advances in vision-language models, harnessing pre-trained captioning and scoring models. However, these models may not be perfect. For example, captioning models trained on web-crawled data may not generate informative descriptions in uncommon domains like medical and satellite images. Similarly, scoring models may not adequately capture image-text similarity due to limitations in their training data. Nevertheless, both models perform well in various scenarios, highlighting the practical merits of our B2T framework. Further discussions of limitations can be found in Appendix F.

Table 4. **Ablation on different captioning models.** B2T keywords discovered by different captioning models. We report the average inference time to extract a caption from a single image (in seconds on an RTX 3090 GPU) alongside the model names. The models consistently capture highly biased keywords such as "man," "forest," and "beach," while different models may find diverse fine-grained keywords such as "rainforest" or "lake."

		ClipCap	BLIP	CoCa	BLIP-2	LLaVA
	Inference time	0.13 sec	0.20 sec	0.34 sec	0.56 sec	1.90 sec
CelebA blond	man	O	O	O	O	O
Waterbird	forest	O	O	O	O	O
	bamboo	O	O	O	O	O
	woods	O	-	-	O	-
	rainforest	O	-	-	-	-
Landbird	beach	O	O	O	O	O
	ocean	-	O	O	O	-
	boat	-	O	O	O	O
	lake	O	-	-	-	-

Table 5. **Ablation on different scoring models.** B2T keywords alongside their scores using different scoring models. The models provide consistent rankings, with high scores for keywords like "man" or "bamboo forest," supporting their reliability.

		CLIP	OpenCLIP	BLIP	BLIP-2
CelebA blond	man	1.06	2.23	1.19	4.04
	player	0.35	1.30	0.74	2.67
	face	-0.28	0.44	0.49	1.46
	actress	-1.63	-2.48	-1.68	-4.25
Waterbird	bamboo forest	3.61	4.68	5.22	9.85
	woods	2.24	4.43	3.47	7.08
	bird	-0.09	0.67	-0.03	-0.70
	pond	-0.27	-0.63	-0.92	-1.69

Broader impacts. Bias and fairness research inherently have potential negative social impacts. We emphasize that B2T does not aim to *fully automate* the discovery of biases but to *assist* humans in making decisions based on the bias keywords. The final judgment is left to the users, who should also be monitored by a cross-verification system.

We illustrate sensitive examples - gender and geographic biases. It is crucial to note that our intention is to raise awareness and mitigate potential risks in real-world data.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2019-0-00075, Artificial Intelligence Graduate School Program(KAIST); No. 2022-0-00184, Development and Study of AI Technologies to Inexpensively Conform to Evolving Policy on Ethics). We thank Eunji Kim for constructive discussion.

References

[1] Chirag Agarwal, Daniel D'souza, and Sara Hooker. Estimating example difficulty using variance of gradients. In *CVPR*, 2022. 3

[2] Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. Evaluating clip: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*, 2021. 12

[3] Jerone Andrews, Dora Zhao, William Thong, Apostolos Modas, Orestis Papakyriakopoulos, and Alice Xiang. Ethical considerations for responsible data curation. In *NeurIPS*, 2024. 2

[4] Andrei Atanov, Andrei Filatov, Teresa Yeo, Ajay Sohmshetty, and Amir Zamir. Task discovery: Finding the tasks that neural networks generalize on. In *NeurIPS*, 2022. 2

[5] Yujia Bao and Regina Barzilay. Learning to split for automatic bias detection. *arXiv preprint arXiv:2204.13749*, 2022. 2

[6] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NeurIPS*, 2016. 1, 2

[7] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289, 2020. 3, 8

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 7, 11

[9] Simon Caton and Christian Haas. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*, 2020. 1

[10] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. *arXiv preprint arXiv:2212.07143*, 2022. 4, 8, 3

[11] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *CVPR*, 2018. 12

[12] Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero-shot classifiers. *arXiv preprint arXiv:2303.15233*, 2023. 6

[13] Terrance De Vries, Ishan Misra, Changhan Wang, and Laurens Van der Maaten. Does object recognition work for everyone? In *CVPR*, 2019. 5

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1, 2, 5, 6

[15] Greg d'Eon, Jason d'Eon, James R Wright, and Kevin Leyton-Brown. The spotlight: A general method for discovering systematic errors in deep learning models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022. 2

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 2, 7

[17] Sabri Eyuboglu, Maya Varma, Khaled Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Ré. Domino: Discovering systematic errors with cross-modal embeddings. In *ICLR*, 2022. 1, 2, 3, 5, 9

[18] Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. Data determines distributional robustness in contrastive language image pre-training (clip). *arXiv preprint arXiv:2205.01397*, 2022. 12

[19] Felix Friedrich, Patrick Schramowski, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Sasha Luccioni, and Kristian Kersting. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893*, 2023. 6, 7

[20] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2019. 2

[21] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 1, 2

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2, 4, 7

[23] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 7, 11

[24] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *ECCV*, 2018. 2

[25] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019. 2, 4

[26] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, 2021. 2, 3, 4

[27] Evan Hernandez, Sarah Schwettmann, David Bau, Teona Bagashvili, Antonio Torralba, and Jacob Andreas. Natural language descriptions of deep visual features. In *ICLR*, 2022. 1, 3

[28] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 6

[29] Saachi Jain, Hadi Salman, Eric Wong, Pengchuan Zhang, Vibhav Vineet, Sai Vemprala, and Aleksander Madry. Missingness bias in model debugging. In *ICLR*, 2022. 1, 3

[30] Saachi Jain, Hannah Lawrence, Ankur Moitra, and Aleksander Madry. Distilling model failures as directions in latent space. In *ICLR*, 2023. 1, 2, 3, 5, 9

[31] Nikita Jaipuria, Katherine Stevo, Xianling Zhang, Meghana L Gaopande, Ian Calle, Jinesh Jain, and Vidya N Murali. deep-pic: Deep perceptual image clustering for identifying bias in vision datasets. In *CVPR*, 2022. 2

[32] Ajil Jalal, Sushrut Karmalkar, Jessica Hoffmann, Alex Dimakis, and Eric Price. Fairness for image generation with uncertain sensitive attributes. In *ICML*, 2021. 2

[33] Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1): 1–54, 2019. 2

[34] Hyunwoo Kang, Sangwoo Mo, and Jinwoo Shin. Oamixer: Object-aware mixing layer for vision transformers. *arXiv preprint arXiv:2212.06595*, 2022.

[35] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *ICML*, 2021. 12

[36] Arvindkumar Krishnakumar, Viraj Prabhu, Sruthi Sudhakar, and Judy Hoffman. Udis: Unsupervised discovery of bias in deep visual recognition models. In *BMVC*, 2021. 2

[37] Nicol Turner Lee. Detecting racial bias in algorithms and machine learning. *Journal of Information, Communication and Ethics in Society*, 2018. 2

[38] Seongmin Lee, Zijie J Wang, Judy Hoffman, and Duen Horng Polo Chau. Viscuit: Visual auditor for bias in cnn image classifier. In *CVPR*, 2022. 2

[39] Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. *arXiv preprint arXiv:2303.16203*, 2023. 6

[40] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 4, 8, 3

[41] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 3, 8

[42] Zhiheng Li and Chenliang Xu. Discover the unknown biased attribute of an image classifier. In *ICCV*, 2021. 3

[43] Zhiheng Li, Anthony Hoogs, and Chenliang Xu. Discover and mitigate unknown biases with debiasing alternate networks. In *ECCV*, 2022. 3

[44] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *ICML*, 2021. 1, 2, 3, 5, 6, 9

[45] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 8, 3

[46] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 2, 4

[47] Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Analyzing societal representations in diffusion models. *arXiv preprint arXiv:2303.11408*, 2023. 6, 7

[48] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35, 2021. 1, 2

[49] George Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 3

[50] Sangwoo Mo, Hyunwoo Kang, Kihyuk Sohn, Chun-Liang Li, and Jinwoo Shin. Object-aware contrastive learning for debiased scene representation. In *NeurIPS*, 2021. 2

[51] Sangwoo Mo, Minkyu Kim, Kyungmin Lee, and Jinwoo Shin. S-clip: Semi-supervised vision-language pre-training using few specialist captions. *arXiv preprint arXiv:2305.14095*, 2023. 12

[52] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 3, 8, 12

[53] Seung Jun Moon, Sangwoo Mo, Kimin Lee, Jaeho Lee, and Jinwoo Shin. Masker: Masked keyword regularization for reliable text classification. In *AAAI*, 2021.

[54] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *ICML*, 2013. 2

[55] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. In *NeurIPS*, 2020. 1, 3, 6

[56] Junhyun Nam, Sangwoo Mo, Jaeho Lee, and Jinwoo Shin. Breaking the spurious causality of conditional generation via fairness intervention with corrective sampling. *arXiv preprint arXiv:2212.02090*, 2022. 6

[57] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *NeurIPS*, 2021. 7

[58] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 8

[59] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2, 3, 6, 12

[60] Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019. 3

[61] Nazneen Rajani, Weixin Liang, Lingjiao Chen, Meg Mitchell, and James Zou. Seal: Interactive tool for systematic error analysis and labeling. In *EMNLP*, 2022. 2

[62] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017. 12

[63] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019. 1

[64] William A Gaviria Rojas, Sudnya Diamos, Keertan Ranjan Kini, David Kanter, Vijay Janapa Reddi, and Cody Coleman. The dollar street dataset: Images representing the geographic

and socioeconomic diversity of the world. In *Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 2, 5

[65] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 6

[66] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *ICLR*, 2020. 2, 3, 4, 6, 7

[67] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022. 8, 3

[68] Vaishaal Shankar, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, and Ludwig Schmidt. Evaluating machine accuracy on ImageNet. In *ICML*, 2020. 8

[69] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 3

[70] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000. 3

[71] Herbert A Simon. Spurious correlation: A causal interpretation. *Journal of the American statistical Association*, 49(267): 467–479, 1954. 1, 2, 3

[72] Sahil Singla and Soheil Feizi. Salient imagenet: How to discover spurious features in deep learning? In *ICLR*, 2022. 1, 3

[73] Sahil Singla, Besmira Nushi, Shital Shah, Ece Kamar, and Eric Horvitz. Understanding failures of deep networks via robust feature extraction. In *CVPR*, 2021. 1, 3

[74] Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. In *NeurIPS*, 2020. 1, 3, 6

[75] Rohan Taori and Tatsunori B. Hashimoto. Data feedback loops: Model-driven amplification of dataset biases. In *ICML*, 2022. 2

[76] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR*, 2011. 1

[77] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *ICCV*, 2019. 2

[78] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *CVPR*, 2017. 12

[79] Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *arXiv preprint arXiv:2302.03668*, 2023. 1

[80] Olivia Wiles, Isabela Albuquerque, and Sven Gowal. Discovering bugs in vision models using off-the-shelf image generation and captioning. *arXiv preprint arXiv:2208.08831*, 2022. 1, 3

[81] W.Liang and J.Zou. Metashift: A dataset of datasets for evaluating contextual distribution shifts and training conflicts. In *ICLR*, 2022. 8

[82] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. Errudite: Scalable, reproducible, and testable error analysis. In *ACL*, 2019. 2

[83] Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *ICLR*, 2021. 2

[84] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *TMLR*, 2022. 3, 8

[85] Michael Zhang and Christopher Ré. Contrastive adapters for foundation model group robustness. In *NeurIPS*, 2022. 3, 6, 7

[86] Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Ré. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. In *ICML*, 2022. 6

[87] Yuhui Zhang, Jeff Z HaoChen, Shih-Cheng Huang, Kuan-Chieh Wang, James Zou, and Serena Yeung. Drml: Diagnosing and rectifying vision models using language. In *ICLR*, 2023. 3

[88] Bowen Zhao, Chen Chen, Qian-Wei Wang, Anfeng He, and Shu-Tao Xia. Combating unknown bias with effective bias-conflicting scoring and gradient alignment. In *AAAI*, 2022. 3

[89] Dora Zhao, Jerone Andrews, and Alice Xiang. Men also do laundry: Multi-attribute bias amplification. In *ICML*, 2023. 2

[90] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*, 2017. 2