Appendix

A Additional experiments and analysis

A.1 COCO-Counterfactuals Improve Model Robustness to Counterfactual Changes

By design, COCO-Counterfactuals may offer greater improvements to the robustness of models to minimal or counterfactual changes in images. Such examples are unlikely to be present in the datasets used previously to evaluate OOD generalization. Therefore, we also evaluate the performance of models on a withheld test set of COCO-Counterfactuals to determine their image-text retrieval capabilities on in-domain counterfactual examples. Specifically, we withhold 30% of the original-counterfactual paired examples in COCO-Counterfactuals for testing and train the pre-trained CLIP, BridgeTower, and Flava models on the remainder, with 56% of the total dataset used for training and 14% used as a development set.

Table 5 compares the performances of CLIP, BridgeTower, and Flava models trained on COCO-Counterfactuals to those trained on an equivalent amount of real examples from MS-COCO and to their pre-trained versions[10]. We observe that training on COCO-Counterfactuals results in a mean improvement of 11.83, 21.55, and 11.47 relative to the pre-trained CLIP, BridgeTower, and Flava models, respectively. This represents an average relative improvement of 24.3% for each model over the performance of its pre-trained version. In addition, the CLIP, BridgeTower, and Flava models that were trained on COCO-Counterfactuals achieve a mean absolute improvement of 6.06, 10.08, and 5.28, respectively, relative to those that were trained on MS-COCO. The greater magnitude of these performance gains relative to our OOD image-text retrieval evaluations (Table 3) suggests that training on COCO-Counterfactuals improves model robustness to counterfactual changes, which are not present in our (non-counterfactual) OOD evaluation datasets.

		Text Retrieval			Image Retrieval			
Pre-trained Models	**Training dataset**	**R@1**	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**	**Mean**
CLIP	None (pre-trained CLIP)	50.96	79.33	86.45	47.89	77.19	85.73	71.26
	MS-COCO	57.17	84.23	90.66	55.45	84.00	90.65	77.03
	COCO-CFs	65.03	90.26	94.99	64.09	89.52	94.62	83.09
BridgeTower	None (pre-trained BridgeTower)	35.26	65.31	76.73	28.77	56.63	68.46	55.19
	MS-COCO	41.78	71.78	81.88	44.68	75.38	84.48	66.66
	COCO-CFs	54.37	83.08	90.53	56.63	84.48	91.36	76.74
Flava	None (pre-trained Flava)	34.40	66.63	78.02	51.55	80.64	88.24	66.58
	MS-COCO	46.70	76.36	85.68	52.55	81.08	88.43	71.80
	COCO-CFs	54.39	83.35	90.27	57.97	85.11	91.38	77.08

Table 5: Image-text retrieval performance on a withheld COCO-CFs test set.

A.2 Analysis of Differences in OOD Generalization on Image Recognition Datasets

To better understand the differences in OOD generalization performance across datasets, we measured the frequency in which the altered subjects used to produce COCO-Counterfactuals overlapped with class labels. Specifically, we define the COCO-CFs Label Frequency for each image recognition dataset as the total number of COCO-Counterfactuals in which one or more of the dataset's labels matched one of the two altered subjects used to produce the counterfactual pair.

Table 6 provides the COCO-CFs Label Frequency for each image recognition dataset along with the change in OOD performance relative to pre-trained CLIP after training on various sizes of

[10] Note that the image-text retrieval performance of the three pre-trained models (CLIP, BridgeTower, and Flava) on the in-domain COCO-Counterfactuals test set in Table 5 are higher than the respective values on the entire COCO-Counterfactuals dataset provided in Tables 2 and 13. This is expected because the retrieval space of the in-domain COCO-Counterfactuals test set is only 30% of the entire COCO-Counterfactuals dataset.

IR Dataset	COCO-CFs Label Frequency	COCO-CFs $_{\text{base}}\Delta$	COCO-CFs $_{\text{medium}}\Delta$	COCO-CFs $_{\text{all}}\Delta$
CIFAR100	3446	2.50	2.63	1.80
Caltech101	354	2.31	2.55	2.45
Caltech256	744	1.78	1.52	1.16
CIFAR10	398	0.65	0.36	-0.29
ImageNet	887	0.41	-0.03	-0.37
Food101	28	-1.04	-2.05	-2.11

Table 6: Frequency of class label occurrence in COCO-CFs and absolute change (Δ) in performance relative to pre-trained CLIP after training on various sizes of COCO-CFs

Error category	% present in sampled COCO-CFs
Failure to generate subject/object	27%
Failure to generate fine-grained details	23%
Hyponymy relationship between altered subjects	15%
Human annotation error	15%
Failure to accurately depict spatial relationships	7%
Failure to generate correct number of objects	6%
Both altered subjects are present in the image	4%
Failure to bind attribute	3%

Table 7: Image-text retrieval performance on the in-domain COCO-CFs test set.

COCO-CFs (see Appendix B.4.1 for a definition of dataset sizes). We observe that datasets having a higher COCO-CFs Label Frequency generally achieve larger improvements in OOD generalization performance. The Pearson correlation coefficient between COCO-CFs Label Frequency and the 18 performance change measurements in Table 6 is 0.522 with a p-value of 0.026, indicating statistically significant positive correlation.

These results suggest that a major contributor to the variation in OOD generalization performance across datasets is the overlap between the evaluation dataset domain and the set of subjects which are altered in COCO-Counterfactuals. Food101, the only dataset which saw no improvement in performance on our best-performing COCO-CFs training dataset, had only 28 cases of overlap between its label set and the subject alterations in COCO-CFs. In contrast, the greatest performance improvements were achieved on CIFAR100, for which 3446 COCO-CFs had subject alterations matching at least one label from the dataset. These findings point to the potential usefulness of targeting counterfactual changes for task-specific datasets.

A.3 Analysis of Errors in COCO-Counterfactuals Identified by Human Annotators

In this section, we analyze errors in COCO-Counterfactuals using the labels assigned by human annotators (Section 4.1). Specifically, we consider an error to be any image-text pair from the COCO-Counterfactuals dataset for which the human annotator did not select the correct caption for the corresponding image.

A.3.1 Manual Categorization of Errors

To investigate potential failure cases in our counterfactual generation approach, we randomly sampled and categorized 100 image-text pairs which were identified as errors by the human annotators. Table 7 provides the percentage of sampled COCO-Counterfactuals which were assigned to various error categories. Additionally, Tables 8 and 9 provide examples of counterfactual pairs which were assigned to the top-six most frequent error categories.

We found that 66% of the sampled errors can be attributed to known limitations of existing text-to-image diffusion models (Chefer et al., 2023; Samuel et al., 2023; Cho et al., 2022), which include the categories for failure to generate a subject or object (e.g., Table 8, row 1), failure to generate fine-grained details (e.g., Table 8, row 2), failure to accurately depict spatial relationships (e.g.,

	Original	Counterfactual
Failure to generate subject/object	 *A cat walking through a **kitchen** by a eating tray*	 *A cat walking through a **field** by a eating tray.*
Failure to generate fine-grained details	 *A **man** playing Wii in a dirty room*	 *A **kid** playing Wii in a dirty room*
Hyponymy relationship between altered subjects	 *Two **kids** in pink and purple jackets standing by a fence*	 *Two **girls** in pink and purple jackets standing by a fence*

Table 8: Examples of failure cases identified by manual error analysis

	Original	**Counterfactual**
Human annotation error	 *Two people dressed in red **skiing** across a snowy landscape*	 *Two people dressed in red **race** across a snowy landscape*
Failure to accurately depict spatial relationships	 *A **woman** lies on the ground under a suitcase.*	 *A **man** lies on the ground under a suitcase.*
Failure to generate correct number of objects	 *A bathroom sink with two **toothbrush** holders on it*	 *A bathroom sink with two **cup** holders on it*

Table 9: Additional examples of failure cases identified by manual error analysis

Altered Subjects	Count	Altered Subjects	Count	Altered Subjects	Count
woman → girl	126	man → boy	125	people → men	116
person → man	93	person → woman	42	person → boy	37
couple → group	36	people → guy	35	people → kid	33
person → girl	33	girl → woman	32	man → woman	30
men → people	29	people → student	27	woman → man	24
man → person	24	building → house	23	men → boy	21
women → girl	21	boy → man	21		

Table 10: Frequency of altered subjects which appeared at least 20 times in errors identified by human annotators

Table 9, row 2), failure to generate the correct number of objects described in the prompt (e.g., Table 9, row 3), and failure to bind attributes such as color.

In many cases, these failures do not negatively impact the depiction of the counterfactual change in the two images because the inaccuracies pertain to details other than the altered subjects. For example, the first row of Table 8 shows the counterfactual pair associated with an image which was categorized as a failure to generate a subject/object; in this case, the altered subjects (kitchen → field) are depicted correctly, but both images lack the *eating tray* described in the prompt. Similarly, the counterfactual pair shown in the second row of Table 8 lacks fine-grained details in the prompt (e.g., *dirty* room), but still depicts the altered subjects correctly (man → kid).

We found that 15% of the sampled errors could be attributed to a hyponym relationship between the altered subjects which caused both captions to be equally valid for a given image. For example, the third row of Table 8 shows a counterfactual pair where the counterfactual image was incorrectly labeled by the human annotator because both captions were valid descriptions of the image (i.e., *girls* can also be referred to as *kids*). Nevertheless, this example is still a valid counterfactual pair considering that the counterfactual caption does not accurately describe the original image and is more descriptive of the counterfactual image than the original caption.

An additional 15% of the sampled errors appeared to be valid image-text pairs without any significant deficiencies. We therefore concluded that such cases were human annotation errors (see Table 9 row 1 for an example). Finally, 4% of the sampled images had equally valid caption choices because both of the altered subjects appeared in the image that was annotated.

The results of this error analysis suggest that the quality of counterfactuals produced by our approach may improve as the capabilities of text-to-image diffusion models advance. New models which overcome known limitations of existing models could be used as a substitute for Stable Diffusion in our approach to produce higher-quality counterfactuals. Additionally, errors associated with hyponymy relationships could be addressed in future work through a refinement of our subject alteration process. For example, ontologies could be used to avoid noun substitutions where it can be determined that a hyponymy relationship exists between the noun candidates. Finally, additional constraints on the image generation process could be explored to prevent both altered subjects from appearing in the same image.

A.3.2 Taxonomic Analysis of Errors

To better understand the relationship between the altered subjects in our counterfactuals and potential failure cases, we conducted a taxonomic analysis of the altered subjects which occurred most frequently among errors identified by human annotators. Table 10 provides the frequency of altered subject pairs which occurred at least 20 times in the error cases identified by human annotators. Interestingly, we observe that 19 of these 20 most frequent altered subject pairs belong to the *human* taxonomy.

We further analyzed this *human* taxonomy in COCO-Counterfactuals by constructing a list of human-related words, which consists of 'girl', 'boy', 'man', 'men', 'woman', 'guy', 'kid', 'person',

			Text Retrieval			Image Retrieval			
Training dataset	$\lvert D_{\text{train}} \rvert$	$\lvert D^{\text{CF}}_{\text{train}} \rvert$	**R@1**	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**	**Mean**
MS-COCO + COCO-CFs	34,313	20,385	75.91	93.95	96.90	77.66	94.51	97.20	89.36

Table 11: Mean image-text retrieval performance on the OOD Flickr30k test set using only COCO-Counterfactuals which were correctly labeled by humans, measured across 25 different random seeds.

			Text Retrieval			Image Retrieval			
Training dataset	$\lvert D_{\text{train}} \rvert$	$\lvert D^{\text{CF}}_{\text{train}} \rvert$	**R@1**	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**	**Mean**
None (pre-trained CLIP)	0	0	50.12	75.04	83.6	30.73	56.28	67.18	60.49
MS-COCO	13,928	0	$57.33_{0.3}$	$81.28_{0.2}$	$88.71_{0.2}$	$41.13_{0.1}$	$68.46_{0.1}$	$78.45_{0.1}$	$69.23_{0.1}$
MS-COCO + COCO-CFs	13,928	6,939	$56.91_{0.3}$	$80.70_{0.2}$	$87.82_{0.2}$	$39.92_{0.1}$	$67.01_{0.1}$	$77.15_{0.1}$	$68.25_{0.1}$
MS-COCO + COCO-CFs	34,820	20,894	$\underline{\mathbf{58.06}}_{0.3}$	$\underline{\mathbf{81.39}}_{0.2}$	$\underline{\mathbf{88.91}}_{0.2}$	$\underline{41.63}_{0.2}$	$\underline{68.64}_{0.1}$	$\underline{78.85}_{0.1}$	$\underline{69.58}_{0.1}$
MS-COCO + COCO-CFs	41,784	27,853	$\underline{58.02}_{0.3}$	$\underline{81.39}_{0.2}$	$88.78_{0.2}$	$\underline{\mathbf{41.82}}_{0.1}$	$\underline{\mathbf{68.79}}_{0.1}$	$\underline{\mathbf{78.89}}_{0.1}$	$\underline{\mathbf{69.62}}_{0.1}$

Table 12: Image-text retrieval performance on the in-domain MS-COCO test set. All other settings are identical to Table 3.

'people', 'child', 'children', 'couple', 'group', and 'lady'. An image-text pair is said to be related to this human taxonomy if the altered subject of its caption belong to this list. We find that there are 4117 image-text pairs in COCO-Counterfactuals that are related to the human taxonomy, among which 1864 were identified as errors by human annotators. The corresponding error rate for altered subjects related to the human taxonomy is 44.3%, which indicates that generating counterfactual pairs involving human altered subjects is more challenging for our approach. This suggests that a promising direction for future work is the exploration of improvements to the generation of images involving human subjects.

A.4 Training Data Augmentation with Only Correctly-annotated COCO-Counterfactuals

We investigate the potential impact of COCO-Counterfactuals which were incorrectly labeled by humans on training data augmentation. Table 11 provides the OOD image-text retrieval performance in this setting, where COCO-Counterfactuals were filtered to only include those which were correctly labeled by the human annotators. Overall we find similar performance as our previous experiments using the full COCO-Counterfactuals dataset (Table 3), suggesting that filtering our synthetic data using human evaluations is not necessary for data augmentation applications.

A.5 COCO-Counterfactuals Improve In-domain Performance

We evaluate the same models trained with counterfactual data augmentation described in Section 5 on the MS-COCO test set. The results of this in-domain evaluation are provided in Table 12. Similar to the OOD image-text retrieval setting, we find that data augmentation with $20,892$ COCO-Counterfactuals provides statistically significant performance improvements relative to training without counterfactual data augmentations. Notably, previous work has observed that counterfactual data augmentation can degrade performance on withheld in-domain test sets (Wang and Culotta, 2021; Howard et al., 2022), whereas data augmentation with our COCO-Counterfactuals actually increases in-domain performance on MS-COCO.

A.6 COCO-Counterfactuals for Model Evaluation Experiments

We further investigate whether our COCO-Counterfactuals (COCO-CFs) can serve as a challenging test set for state-of-the-art multimodal vision-language models such as CLIP, Flava (Singh et al., 2022), BridgeTower (Xu et al., 2022) and ViLT (Kim et al., 2021) for the zero-shot image-text

		Text Retrieval			Image Retrieval		
HuggingFace Pre-trained Models	**Evaluated Dataset**	**R@1**	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**
Clip	COCO-CFs	37.65 (**-21%**)	64.89 (-9%)	74.57 (-7%)	34.98 (+5%)	62.29 (+7%)	72.43 (+4%)
	human-evaluated-COCO-CFs	43.25 (**-9%**)	70.4 (-2%)	79.37 (-1%)	40.14 (+21%)	67.86 (+16%)	77.66 (+11%)

Table 13: Image-text retrieval performance on COCO-CFs and human-evaluated COCO-CFs for CLIP model. Largest drops of performance against the baseline are in boldface.

retrieval and image-text matching tasks. We employed the following HuggingFace implementations of these models via the transformers library:

- **CLIP**: We used the pre-trained model clip-vit-base-patch32
- **Flava**: We used the pre-trained model flava-full
- **BridgeTower**: We used the pre-trained model bridgetower-large-itm-mlm-itc
- **ViLT**: We used the pre-trained model vilt-b32-finetuned-coco

Zero-shot Image-text Retrieval. In Section 4, we evaluated the zero-shot image-text retrieval (ITR) performance of pre-trained Flava and BridgeTower models on COCO-CFs and *human-evaluated COCO-CFs* that consists of only image-text pairs that were correctly matched in human evaluation in Section 4.1. Since a pre-trained CLIP model was employed in our counterfactual image generation process (see Section 3.2), CLIP models are not suitable for the zero-shot ITR evaluation. Hence, we only report evaluation of pre-trained CLIP model for ITR task here for completeness.

Table 13 reports ITR performance (i.e., Recall at 1, 5, and 10) on COCO-CFs and human-evaluated-COCO-CFs for the pre-trained CLIP model. Similar to Table 2, the percentages enclosed within parentheses indicate the change in performance of the CLIP model on an evaluated dataset versus the performance of that model on MS-COCO (baseline).

We observe that on both COCO-CFs and human-evaluated-COCO-CFs datasets, while the performance of the pre-trained CLIP model degrades marginally on Text Retrieval task, its performance increases for Image Retrieval task. We attribute this to potential data contamination due to how we employed a pre-trained CLIP model in our counterfactual image generation process (see Section 3.2). As a result, COCO-Counterfactuals includes image-text pairs for which CLIP achieves high image-text retrieval performance.

B Dataset and experiment details

B.1 URL to Access COCO-Counterfactuals Dataset and Code

During review, COCO-Counterfactuals and its accompanying code can be accessed via the following link:

`https://drive.google.com/drive/folders/1nHKuYCOyU1JH4cNiKa3lNUA4ENvsL51F`

This link leads to a Google Drive that includes two folders:

- Folder *COCO-Counterfactuals-Dataset* includes our zipped COCO-Counterfactuals dataset and a README file.
- Folder *COCO-Counterfactuals-SourceCode* includes a zip file and a README file. The zip file includes all of data and implementations that can be used to re-produce our generated COCO-Counterfactuals dataset and experimental results presented in the paper.

While the README file in the former folder describes the structure of our zipped COCO-Counterfactuals dataset, that one in the latter folder details instructions to re-produce our generated COCO-Counterfactuals dataset and experimental results presented in the paper.

We will make COCO-Counterfactuals and the code for our counterfactual data generation pipeline publicly available upon publication.

B.2 Hyper-parameter Selection and Models Used to Generate COCO-Counterfactuals

In this section, we will detail hyper-parameters and pre-trained models used to our generate COCO-Counterfactuals dataset.

B.2.1 Creating Counterfactual Captions

Given an original caption from the MS-COCO dataset, we use Natural Language Toolkit (**NLTK**) (Bird et al., 2009) modules:

- *punkt* for sentence tokenizer, and
- *averaged_perceptron_tagger* for part-of-speech (POS) tagger

to identify all nouns as candidate words for substitution.

For each of the identified nouns, we create 10 candidate counterfactual captions by replacing only one noun with the [MASK] token and retrieving the top-10 most probable replacements via masked language modeling (MLM). For MLM, we used the pre-trained model *roberta-base* (Liu et al., 2019) implemented in the library *transformers* (Wolf et al., 2019)

In order to measure similarity between each candidate counterfactual caption and an original caption, we used the pre-trained model all-MiniLM-L6-v2, which is implemented within the library *sentence-transformers* (Reimers and Gurevych, 2019).

Among generated candidate counterfactual captions, we kept only those candidates which have a sentence similarity within the range $(0.8, 0.91)$. We selected this similarity range heuristically, observing that it produced best results after extensive experimentation.

Finally, we employed the pre-trained model gpt2-large, a *GPT-2* (Radford et al., 2018) model implemented in the transformers library, to score the perplexity and choose the candidate having the lowest perplexity as our counterfactual caption.

B.2.2 Counterfactual Image Generation

After creating a counterfactual caption, our next task is to generate synthetic images from the corresponding original caption and counterfactual caption, respectively. In order to do so, we have adopted an implementation from Instruct-Pix2Pix (Brooks et al., 2023) in which all hyperparameters are set to their default values.

Specifically, we over-generate 100 image pairs with Prompt-to-Prompt by randomly sampling values of the parameter $p \sim U(0.1, 0.9)$ (i.e., parameter p indicates the portion of denoising for which to fix self attention maps). The resulting 100 image pairs are filtered using CLIP (Radford et al., 2021) to ensure:

i. a minimum cosine similarity of 0.2 between the encoding of each caption and its corresponding generated image, and

ii. a minimum cosine similarity of 0.7 between the encoding of the two respective images in each generated image pair.

From remaining image pairs, the best image pair is chosen such that it has the highest directional similarity $CLIP_{dir}$ score. Selecting images with the highest $CLIP_{dir}$ improves the overall quality of our generated counterfactuals via greater consistency between the alterations made in both modalities.

B.3 Human Annotation Study

Professional annotation services for our human study were provided by Mindy Support. The total cost of this study was $1068.59 for 218 annotation hours. The instructions provided to annotators are depicted in Figure 4. We are unable to provide the hourly wages paid to workers as this is considered

Instructions:

Select the caption which best describes the image. In cases where both captions are valid for the image, please try to pick the one which is more descriptive or detailed. If both captions are valid and describe the image equally well, select "Both". If neither of the captions accurately describe the image, select "Neither".

- A woman standing in a kitchen by a window
- A man standing in a kitchen by a window
- Both
- Neither

Figure 4: Instructions provided to data annotators

proprietary information by Mindy Support. However, the following statement was provided by the vendor regarding compensation:

"We prioritize compliance with all standards of local and international legislation, ensuring fair treatment and equal opportunities for individuals of various backgrounds, ages, and other characteristics. We are committed to upholding the principles of fair wages, non-discrimination, and labor standards, including the prohibition of child labor. As an organization, we strictly adhere to legal requirements and strive to create an inclusive and ethical working environment for all. Rest assured that our compensation rates reflect market demands and provide fair remuneration for the work performed by our participants. We remain dedicated to abiding by all labor regulations and social and economic standards."

B.4 Training Data Augmentation Experiments

In this section, we detail how we constructed our training datasets and how we finetuned the pre-trained CLIP model for experiments described in Section 5.

B.4.1 Training Dataset Preparation

Our training data augmentation experiments utilize various combinations of the MS-COCO validation set and our COCO-Counterfactuals dataset. For simplicity, a caption-image pair is referred to as a *sample*. We define a *counterfactual sample* as following. Given a sample (C, I) (i.e., caption C and image I) from our COCO-Counterfactuals dataset, a sample (C', I') from COCO-Counterfactuals dataset is called a counterfactual sample of (C, I) iff C' and C are counterfactual captions of each other. By this definition, COCO-Counterfactuals dataset includes 34,820 samples that correspond to 17,410 paired counterfactual samples.

For experiments in Section 5, we have prepared the following 4 datasets:

(a.) **MS-COCO** dataset. This is a subset of the 5K validation split of the 2017 MS-COCO dataset[11], achieved by filtering out all samples with captions which are not included in our COCO-Counterfactuals. This results in a dataset (referred to as the MS-COCO dataset used in experiments in Section 5) of 17,410 captions and their paired original images.

(b.) **[MS-COCO + COCO-CFs]$_{\textbf{base}}$** dataset. This dataset is a combination of:

- 50% random sampling (i.e., 8,705 caption-image pairs) of the MS-COCO dataset constructed in *(a.)*.
- 25% random sampling of paired counterfactual samples from our COCO-Counterfactuals dataset. This results in a total of 4,353 pairs of samples with their corresponding counterfactuals, for a total of 8,706 caption-image samples from our COCO-Counterfactuals dataset.

Overall, the [MS-COCO + COCO-CFs]$_{base}$ dataset consists of 17,411 captions and their paired original images, which is approximately equal in size to the MS-COCO dataset constructed in *(a.)*

(c.) **[MS-COCO + COCO-CFs]$_{\textbf{medium}}$** dataset. This dataset is a combination of:

- all samples (i.e., 17,410 caption-image pairs) from the MS-COCO dataset constructed in *(a.)*.
- 75% random sampling (i.e., 26,115 caption-image pairs) from our COCO-Counterfactuals dataset.

Overall, dataset [MS-COCO + COCO-CFs]$_{medium}$ consists of 43,525 captions and their paired original images.

(d.) **[MS-COCO + COCO-CFs]$_{\textbf{all}}$** dataset. This dataset is a combination of:

- all samples (i.e., 17,410 caption-image pairs) from the MS-COCO dataset constructed in *(a.)*.
- all samples (i.e., 34,820 caption-image pairs) from our COCO-Counterfactuals dataset.

Overall, dataset [MS-COCO + COCO-CFs]$_{all}$ consists of 52,230 captions and their paired original images.

Each of the datasets described above is split into a training set (80%) and a validation set (20%). In each experiment, the validation set is used to pick the best model checkpoint at the conclusion of training. Tables 3, 4, and 12 report experimental results for models trained using the train split of these four datasets. $|D_{\text{train}}|$ indicates the total number of samples (i.e., image-text pairs) included in the respective training set, while $|D^{\text{CF}}_{\text{train}}|$ indicates how many of those image-text pairs were sampled from the COCO-Counterfactuals dataset.

[11] `https://cocodataset.org/#download`

B.4.2 Finetuning CLIP with Data Augmentation

We use each of the four training sets constructed in Section B.4.1 to finetune the CLIP model *clip-vit-base-patch32*. We adopted a publicly-available finetuning script provided by HuggingFace[12].

We repeat each of our training experiments with 25 different *seeds* and *data_seed* from the ranges $[107, 131]$ and $[108, 132]$, respectively. In each experiment, we use a learning rate to 5e-7, weight decay of 0.001, training batch size of 128, and evaluation batch size of 128.

B.5 Compute Infrastructure Used In this Study

COCO-Counterfactuals was generated using an Intel AI supercomputing cluster comprised of Intel Xeon processors and Intel Habana Gaudi AI accelerators. Our dataset generation pipeline was parallelized across 512 accelerators and took approximately 3 days to complete.

Our training data augmentation experiments were run on an internal Slurm linux cluster with Nvidia RTX 3090 GPUs and varied in running time depending upon the size of the dataset, ranging between 2 to 10 hours.

B.6 License Information of Assets Employed in This Study

- **NLTK** is open source software distributed under the terms of the Apache License Version 2.0.
- *Transformers* is released under the Apache License Version 2.0 and is available on GitHub at `https://github.com/huggingface/transformers`.
- Pre-trained model Roberta-base is released under the MIT License.
- Library *sentence-transformers* is licensed under the Apache License Version 2.0 and is available on GitHub at https://github.com/UKPLab/sentence-transformers.
- Pre-trained model *all-MiniLM-L6-v2* is licensed under the Apache License Version 2.0.
- Pre-trained *gpt2-large* model is license under the MIT License.
- *Instruct-Pix2Pix* is licensed under the MIT License and is available on GitHub at `https://github.com/timothybrooks/instruct-pix2pix`.
- Instruct-Pix2Pix further employs stable-diffusion-v1-5 that is released under CreativeML-Open-RAIL-M License.
- For the *MS-COCO* dataset:
  - The annotations in the dataset are released under the Creative Commons Attribution 4.0 License.
  - The use of the images in the dataset must abide by the Flickr Terms of Use.
- Pre-trained model *clip-vit-base-patch32* is licensed under the MIT License.
- Pre-trained model *flava-full* is licensed under the 3-Clause BSD License.
- Pre-trained model *BridgeTower large-itm-mlm-itc* is released under the MIT License.
- Pre-trained *vilt-b32-finetuned-coco* model is license under the Apache License Version 2.0.

C Datasheet for Dataset

C.1 Motivation

For what purpose was this dataset created? This dataset was created for the purpose of exploring the relevancy of counterfactual examples for multimodal vision-language models. Specifically, our

[12]The finetuning script can be accessed at `https://github.com/huggingface/transformers/blob/main/examples/pytorch/contrastive-image-text/run_clip.py`.

aim was to create a dataset which can serve both as a challenging evaluation dataset for existing models and as a resource for training data augmentation to improve multimodal models on downstream tasks. For additional discussion of our motivation and the intuition behind counterfactual examples, see Section 1.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? The dataset was created by the authors of this paper who are affiliated with Intel Labs, a research and development organization within Intel Corporation.

Who funded the creation of the dataset? The creation of this dataset was founded by Intel Corporation.

C.2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? The instances represent synthetically-generated images and accompanying text captions. The images depict a variety of different everyday scenarios.

How many instances are there in total (of each type, if appropriate)? COCO-Counterfactuals contains a total of 34,820 image-caption pairs.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? Yes, it contains all possible instances per our filtering criteria.

What data does each instance consist of? Each instance consists of a synthetically-generated image and an accompanying text caption.

Is there a label or target associated with each instance? No

Is any information missing from individual instances? No

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? Yes, instances which correspond to a single counterfactual pair are annotated as such in our dataset. Otherwise, there are no other relationships between individual instances.

Are there recommended data splits (e.g., training, development/validation, testing)? No

Are there any errors, sources of noise, or redundancies in the dataset? The automated methodology used to generate COCO-Counterfactuals introduces the possibility of noise and errors in the dataset. See Section 7 for additional discussion.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? Yes

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? No

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? Yes, the dataset may contain offensive material due to the manner in which it was automatically constructed. See Section 7 for additional discussion.

Does the dataset identify any subpopulations (e.g., by age, gender)? No

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? No

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? No

C.3 Collection Process

How was the data associated with each instance acquired? The data associated with each instance was acquired via our data generation methodology (see Section 3 for a detailed description).

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? Please see Section 3 for a complete description of our data generation methodology.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? Not applicable

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)? The COCO-Counterfactuals dataset was collected automatically, as detailed in Section 3. Human evaluation of COCO-Counterfactuals involved paid professional annotators employed by Mindy Support (see Appendix B.3 for details).

Over what timeframe was the data collected? The data was generated and evaluated over the course of approximately three months.

Were any ethical review processes conducted (e.g., by an institutional review board)? No, institutional review was not required.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)? No, the dataset was generated automatically and was not collected directly from individuals.

Were the individuals in question notified about the data collection? Not applicable

Did the individuals in question consent to the collection and use of their data? Not applicable

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? Not applicable

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? No, not applicable

C.4 Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? Yes, we apply extensive filtering to various stages of our data generation pipeline in order to improve the quality of the dataset. See Section 3 for a complete description of these methods.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? No. However, due to how our dataset is automatically constructed, raw data can be reproduced by running our code.

Is the software that was used to preprocess/clean/label the data available? Yes, we will make our code publicly available upon publication.

C.5 Uses

Has the dataset been used for any tasks already? Yes, we applied COCO-Counterfactuals to the task of model evaluation in Section 4 and to the task of training data augmentation in Section 5.

Is there a repository that links to any or all papers or systems that use the dataset? Our GitHub repository will contain links to papers and systems used by our data generation methodology. Additionally, this paper contains references to all such papers and systems that we utilized.

What (other) tasks could the dataset be used for? COCO-Counterfactuals is broadly applicable to tasks which require multimodal inputs consisting of images with paired text. One potential use case not explored during this study is large-scale pre-trianing of multimodal models, which could be improved through counterfactual data augmentation.

Is there anything about the composition of the dataset or the way it was collected and pre-processed/cleaned/labeled that might impact future uses? Due to the way in which COCO-Counterfactuals was generated automatically, it may contain errors, offensive material, or biases which are present in the models employed by our pipeline (e.g., Stable Diffusion). Users of the dataset should carefully consider how these limitations may impact their potential use case.

Are there tasks for which the dataset should not be used? The dataset should not be used for a task if the limitations discussed above are unacceptable or potentially problematic for the inteded use case.

C.6 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? Yes, the dataset will be made open source and publicly available.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? The dataset will be distributed via the Hugging Face Hub.

When will the dataset be distributed? The dataset will be made available publicly upon publication of this paper.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? The dataset will be distributed under the CC BY 4.0 license.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? No

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? No

C.7 Maintenance

Who will be supporting/hosting/maintaining the dataset? The datasset will be hosted on the Hugging Face Hub. The authors of this paper will support and maintain the dataset via our public GitHub repository.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)? The corresponding author can be contacted via the e-mail address listed on the first page of this paper. Alternatively, an issue can be raised on our GitHub repository.

Is there an erratum? No

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? Although we do not anticipate the need to update this dataset in the future, we will respond to issues which are raised on our public GitHub repository for this project.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? Not applicable

Will older versions of the dataset continue to be supported/hosted/maintained? Yes. If the dataset is updated in the future, older versions will remain available.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? Yes, we make our dataset open source and welcome others to build on it. This can be done by making contributions to our GitHub repository and/or citing our dataset as appropriate when used in future work.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2023. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1493–1504.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402.

Nitay Calderon, Eyal Ben-David, Amir Feder, and Roi Reichart. 2022. Docogen: Domain counterfactual generation for low resource domain adaptation. *arXiv preprint arXiv:2202.12350*.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10.

Jaemin Cho, Abhay Zala, and Mohit Bansal. 2022. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative models. *arXiv preprint arXiv:2202.04053*.

Jacob Eisenstein. 2022. Uninformative input features and counterfactual invariance: Two perspectives on spurious correlations in natural language. *arXiv preprint arXiv:2204.04487*.

Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. 2020. Evaluating models' local decision boundaries via contrast sets. *arXiv preprint arXiv:2004.02709*.

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

Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. 2022. Is synthetic data from generative models ready for image recognition? *arXiv preprint arXiv:2210.07574*.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Phillip Howard, Gadi Singer, Vasudev Lal, Yejin Choi, and Swabha Swayamdipta. 2022. Neurocounterfactuals: Beyond minimal-edit counterfactuals for richer data augmentation. *arXiv preprint arXiv:2210.12365*.

Nitish Joshi and He He. 2022. An investigation of the (in) effectiveness of counterfactually augmented data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3668–3681.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.

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

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. 2022. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pages 423–439. Springer.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738.

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

Aengus Lynch, Gbètondji JS Dovonon, Jean Kaddour, and Ricardo Silva. 2023. Spawrious: A benchmark for fine control of spurious correlation biases. *arXiv preprint arXiv:2303.05470*.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language models are unsupervised multitask learners. *Technical Report, OpenAI*.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. 2019. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494.

Dvir Samuel, Rami Ben-Ari, Simon Raviv, Nir Darshan, and Gal Chechik. 2023. It is all about where you start: Text-to-image generation with seed selection. *arXiv preprint arXiv:2304.14530*.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. 2023. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22522–22531.

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

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. FLAVA: A foundational language and vision alignment model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15617–15629. IEEE.

Sahil Singla and Soheil Feizi. 2021. Salient imagenet: How to discover spurious features in deep learning? *arXiv preprint arXiv:2110.04301*.

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

Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. 2023. Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944*.

Joshua Vendrow, Saachi Jain, Logan Engstrom, and Aleksander Madry. 2023. Dataset interfaces: Diagnosing model failures using controllable counterfactual generation. *arXiv preprint arXiv:2302.07865*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. 2020. Causal mediation analysis for interpreting neural nlp: The case of gender bias. *arXiv preprint arXiv:2004.12265*.

Zhao Wang and Aron Culotta. 2021. Robustness to spurious correlations in text classification via automatically generated counterfactuals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14024–14031.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. *arXiv preprint arXiv:2101.00288*.

Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. 2020. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*.

Xiao Xu, Chenfei Wu, Shachar Rosenman, Vasudev Lal, and Nan Duan. 2022. Bridge-tower: Building bridges between encoders in vision-language representation learning. *arXiv preprint arXiv:2206.08657*.

Linyi Yang, Jiazheng Li, Padraig Cunningham, Yue Zhang, Barry Smyth, and Ruihai Dong. 2021. Exploring the efficacy of automatically generated counterfactuals for sentiment analysis. *arXiv preprint arXiv:2106.15231*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.