Benchmarking Zero-Shot Robustness of Multimodal Foundation Models: A Pilot Study

Chenguang Wang[1]*, Ruoxi Jia[2], Xin Liu[3], Dawn Song[4]
[1]Washington University in St. Louis, [2]Virginia Tech, [3]UC Davis, [4]UC Berkeley
chenguangwang@wustl.edu, ruoxijia@vt.edu, xinliu@ucdavis.edu, dawnsong@berkeley.edu

Abstract

Pre-training image representations from the raw text about images enables zero-shot vision transfer to downstream tasks. Through pre-training on millions of samples collected from the internet, multimodal foundation models, such as CLIP, produce state-of-the-art zero-shot results that often reach competitiveness with fully supervised methods without the need for task-specific training. Besides the encouraging performance on classification accuracy, it is reported that these models close the robustness gap by matching the performance of supervised models trained on ImageNet under natural distribution shift. Because robustness is critical to real-world applications, especially safety-critical ones, in this paper, we present a comprehensive evaluation based on a large-scale robustness benchmark covering 7 natural, 3 synthetic distribution shifts, and 11 adversarial attacks. We use CLIP as a pilot study. We show that CLIP leads to a significant robustness drop compared to supervised ImageNet models on our benchmark, especially under synthetic distribution shift and adversarial attacks. Furthermore, data overlap analysis suggests that the observed robustness under natural distribution shifts could be attributed, at least in part, to data overlap. In summary, our evaluation shows a comprehensive evaluation of robustness is necessary; and there is a significant need to improve the robustness of zero-shot multimodal models.

1. Introduction

The common recipe of current state-of-the-art multimodal foundation models is the pre-training that learns representations from images and raw text. At test time, a standardized interface of natural language prompts enables task-agnostic architectures to zero-shot transfer to downstream datasets without the need for dataset-specific training or architecture modifications. For example, multimodal foundation models such as CLIP [37] learn image representations on a pre-training dataset of hundreds of millions of samples collected from the web, and is competitive across many computer vision tasks zero-shot, even comparable to task-specific fully supervised methods.

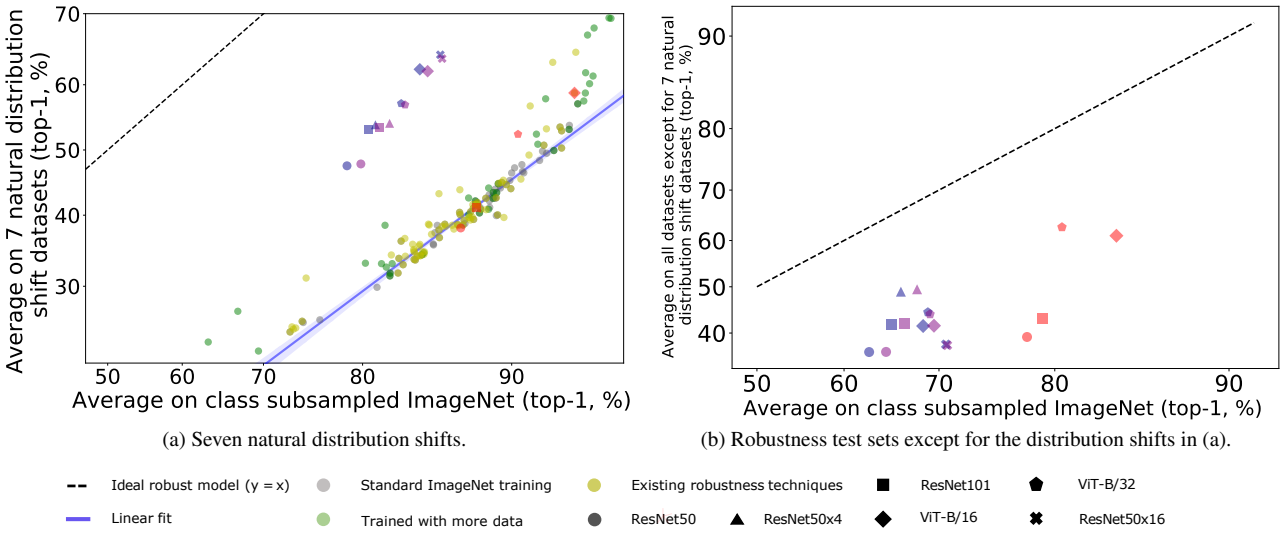
Besides the superior performance, multimodal models have reportedly made similar breakthroughs in robustness. For example, as one of the pioneer models, zero-shot CLIP has closed the robustness gap by up to 75% while matching the performance of a standard model trained on ImageNet. However, the robustness is often tested on natural distribution shifts, which contain natural (or unmodified) images collected from the web. While investigating their robustness to natural distribution shifts is important, it remains unclear whether these models are robust to synthetic distribution shifts, such as noise corruptions [18] and spatial transformations [3], and adversarial examples [16]. This is essential for these models to be deployed in safety-critical applications.

In this work, we establish a comprehensive robustness benchmark for zero-shot image classification, called RoZ (Robustness on Zero-shot), using CLIP as a pilot study. Specifically, we make the following contributions.

- To systematically evaluate robustness of image classification models, we construct a comprehensive robustness benchmark, RoZ, that spans 7 natural distribution shifts, 3 synthetic distribution shifts, and 11 adversarial attack models.
- Using RoZ, we evaluate the robustness of zero-shot multimodal foundation models, specifically, CLIP. We consider various vision encoders in CLIP, as well as a modified CLIP with automatic prompt generation, as shown in Figure 1.
- While our evaluation of CLIP under natural distribution shift shows robust performance as reported earlier (Figure 1 (a)), a careful examination of data overlap suggests the observed robustness could be attributed, at least in part, to data overlap.
- Under synthetic distribution shifts and adversarial attacks, we show that CLIP leads to significantly downgraded ro-

* Corresponding author.

The code and datasets are available at https://github.com/wang-research-lab/roz.



(a) Seven natural distribution shifts.

(b) Robustness test sets except for the distribution shifts in (a).

Figure 1. Summary of results on our RoZ benchmark. An ideal robust model (dashed line) performs equally well on the ImageNet distribution and other distributions. Multimodal models such as CLIP fail to improve robustness on test sets in (b) of our benchmark except for the test sets in (a). Red: standard ImageNet models. Blue: zero-shot CLIP models. Purple: CLIP-Auto models.

bustness (averaging -11.8%) compared to the corresponding standard ImageNet models (Figure 1 (b)).

- We introduce a new robustness test set based on the idea of typographic attacks [14], which targets the unique learning paradigm of multimodal learning. CLIP shows a robustness drop of 34.7%.

Our extensive results and analysis suggest that systematic benchmarking in robustness is important to multimodal applications. Our benchmark can be used to evaluate other multimodal models. Furthermore, there is a significant need to improve the robustness of zero-shot multimodal foundation models.

2. The RoZ Benchmark

We introduce a benchmark, called RoZ, to test the robustness of zero-shot multimodal foundation models. The benchmark provides a suite of existing and new robustness datasets. We focus on the zero-shot CLIP model as a pilot study throughout the rest of this work.

2.1. Datasets and Attacks

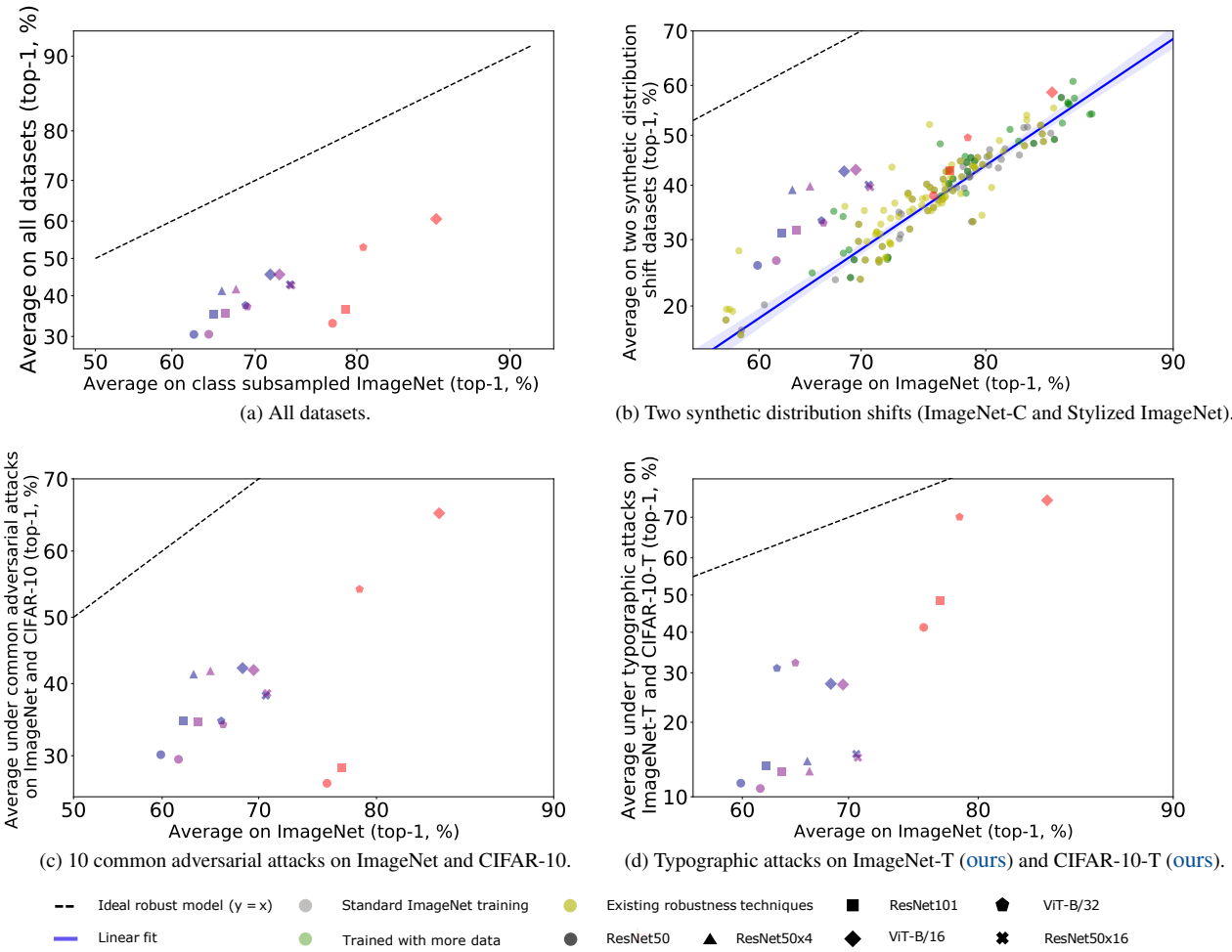
The RoZ benchmark includes common robustness test sets and adversarial attacks. We also create new test sets based on typographic attacks for the benchmark.

Distribution Shifts We follow [46] to distinguish two types of distribution shifts. Natural distribution shift refers to the dataset that relies only on natural or unmodified images, while synthetic distribution shift involves modifications of existing images. Image examples of distribution shifts are shown in the appendix.

- **Natural Distribution Shifts.** We measure on seven natural distribution shifts as in [46]: ImageNetV2 [40], ImageNet Sketch [54], Youtube-BB [39], ImageNet-Vid [43], ObjectNet [3], ImageNet Adversarial [19], and ImageNet Rendition [20].
- **Synthetic Distribution Shifts.** We also test on three most widely used synthetic distribution shift datasets: ImageNet-C and ImageNet-P [18], and Stylized ImageNet [13].

Adversarial Attacks Besides the distribution shifts, we test the robustness to potentially worst-case noises, and adversarial examples.

- **Common Attacks.** We use 10 widely used image attack methods following [9] for the robustness evaluation, including (i) white-box attacks: FGSM [16], DeepFool [31], BIM [26], and MIM [8]; (ii) transfer-based attacks: FGSM, BIM, MIM, and DIM [55]; and (iii) black-box attacks: NES [22] and SPSA [49]. Note that for transfer-based attacks, we use white-box methods on a substitute model to craft adversarial examples. We evaluate the performance on CIFAR-10 [25] and ImageNet [41]. For CIFAR-10, we utilize the test set of CIFAR-10 containing 10,000 images. For ImageNet, we randomly choose 1,000 images from the ImageNet validation set for evaluation. We focus on untargeted adversarial attacks. We also provide a technical description of the above attack methods in the appendix.
- **Typographic Attacks.** Multimodal models are often vulnerable to a kind of non-programmatic adversarial attack, i.e., the typographic attack [14], where adding adversar-



(a) All datasets.

(b) Two synthetic distribution shifts (ImageNet-C and Stylized ImageNet).

(c) 10 common adversarial attacks on ImageNet and CIFAR-10.

(d) Typographic attacks on ImageNet-T (ours) and CIFAR-10-T (ours).

Figure 2. Zero-shot multimodal CLIP fails to significantly improve the robustness over standard ImageNet models on our RoZ benchmark. Red: standard ImageNet models. Blue: zero-shot CLIP models. Purple: CLIP-Auto models. The notable outlier to this trend is CLIP on natural distribution shifts. In particular, we observe a significant performance drop in robustness on our ImageNet-T and CIFAR-10-T. The original CLIP and CLIP-Auto perform similarly on all the test sets.

ial text to images can cause them to be systematically misclassified. This is because these models consist of multimodal neurons which respond to both images and texts for a given concept. We therefore leverage typographic attacks to specifically examine the robustness of the zero-shot CLIP models. We generate the attacks using the same number of randomly chosen coordinates and using a consistent font style, and focus on targeted adversarial attacks. We choose a target class for each image uniformly over all other classes except its true class at random. The target class name is added to each image. For each image in the ImageNet validation set and CIFAR-10 test set, we perform the above image manipulation, resulting in two new robustness datasets: ImageNet Typographic (ImageNet-T) and CIFAR-10 Typographic (CIFAR-10-T). We use 8 and 4 coordinates for the construction of ImageNet-T and CIFAR-10-T, respectively. An example dataset is shown in Figure 3. To the best of our knowledge, this is the first publicly available benchmark based on the concept of typographic attack [14].

2.2. Multimodal Models

CLIP relies on manual natural language prompts to synthesize the zero-shot image classifier. Besides revisiting zero-shot CLIP, we present CLIP-Auto, which automatically learns prompts for enhanced classification performance.

Zero-Shot CLIP CLIP [37] consists of two components: image encoder and text encoder. At a high level, the image encoder is a computer vision backbone that computes a feature representation for the image. The text encoder is a hypernetwork that generates the weights of a linear classi-

Figure 3. Our ImageNet-T samples. We show the gold class (upper) and the target class (lower) of each sample.

fier based on the text specifying the visual concepts that the classes represent. Below is the main architecture of CLIP: (i) Text encoder architecture: Transformer [50] is adopted with the architecture modifications in [36]. (ii) Image encoder architecture: There are two architectures. The first architecture is ResNet [17]. The second architecture is Vision Transformer [10]. CLIP jointly trains the image encoder and the text encoder to predict the correct pairings of a batch of (image, text) training examples via contrastive learning. At test time, the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes. For zero-shot classification on a dataset, CLIP uses the names of all the classes in the dataset as the set of possible text pairs and predicts the most likely (image, text) pairings. CLIP relies on prompt engineering and ensembling to provide manual prompts. Basically, different classifiers are computed based on various manual prompts such as "A photo of a large {label}" and "A photo of a little {label}". CLIP ensembles 80 different prompts over the embedding space [37]. We use CLIP in short for the zero-shot CLIP in the remaining paper except noted otherwise.

CLIP-Auto While writing prompts is not only time-consuming, it is unclear whether it is optimal for robustness improvements. Motivated by the need for prompts that aim to enhance the robustness of zero-shot language models, we adapt the AutoPrompt [45] method to learn better language descriptions for CLIP. Different from language models that only deal with text, multimodal CLIP handles both images and text for the image classification task. Therefore, our automated prompt combines the image label names with a collection of text trigger tokens, which are learned using a variant of the gradient-based search [45] with respect to image classification loss. The basic idea is that, at each search step, we select a candidate text trigger token to replace a current trigger token that leads to the smallest image classification loss. The classifier computed based on the learned prompt is used to classify the images. We follow [37] to ensemble a set of automated prompts to further improve the performance of CLIP. More details are described in the appendix.

Radford et al. [37] have released model settings based on ResNet50, ResNet101, ViT-B/32, ViT-B/16, ResNet50x4, and ResNet50x16. For each setting, we include the corresponding CLIP and CLIP-Auto versions in our evaluations.

2.3. Comparison Models

We compare the robustness of the CLIP models to standard models; i.e., image classification models trained on the ILSVRC 2012 dataset [46]. We focus on the standard models that are parts of the CLIP release for a fair comparison, including: (i) ResNet [17]: ResNet50 and ResNet101; and (ii) Vision Transformer [10]: ViT-B/32 and ViT-B/16. Note that there is no standard model corresponding to CLIP ResNet50x4 and CLIP ResNet50x16.

We mainly report the results of the above models on our benchmark. We additionally include results of 78 standard models, 86 robust models, and 30 models trained with more data from [46] on distribution shift datasets.

2.4. Metrics

We follow [46] to consider two types of robustness: effective and relative robustness. For a model m, we denote two accuracy values: $\mathrm{acc}_1(m)$ and $\mathrm{acc}_2(m)$ on a standard test set and a robustness test set, respectively. Effective and relative robustness are defined as in [46].

Effective Robustness Instead of directly comparing accuracy, effective robustness aims to measure how much higher accuracy on the robustness test sets is compared to the accuracy on the standard test set. Formally, the effective robustness of a model is defined as: $\mathrm{acc}_2(m) - \beta(\mathrm{acc}_1(m))$, where $\beta(\cdot)$ is the baseline accuracy on a robustness test set for a given accuracy on the standard set. Graphically, effective robustness corresponds to a model being above the trend (blue line) given by a set of standard ImageNet models in Figure 1a. $\beta(\cdot)$ indicates the blue line.

Relative Robustness Effective robustness does not measure the improvements brought by a robustness technique. Therefore, relative robustness directly measures the improvements on the robustness test sets. Formally, given a model

Attack Setting / Model		White-Box Attacks				Transfer-Based Attacks				Black-Box Attacks	
		FGSM	DeepFool	BIM	MIM	FGSM	BIM	MIM	DIM	NES	SPSA
ResNet50	Standard	**0.003 / 8.30**	**0.0020 / 0.10**	**0.002 / 0.10**	**0.002 / 0**	**0.045 / 54.40**	**0.040 / 56.20**	0.030 / 47.90	0.032 / 49.60	**0.027 / 57.90**	**0.028 / 59.60**
	CLIP	0.001 / 5.90	0.0002 / 0	0.001 / 0	0.001 / 0	0.050 / 53.20	0.060 / 54.30	**0.047 / 52.30**	**0.049 / 51.80**	0.003 / 30.50	0.003 / 29.80
	CLIP-Auto	0.001 / 6.40	0.0002 / 0	0.001 / 0	0.001 / 0	0.047 / 52.30	0.057 / 53.40	0.042 / 51.50	0.047 / 51.10	0.003 / 31.90	0.003 / 32.00
ResNet101	Standard	0.003 / 8.40	**0.0022 / 0.20**	**0.002 / 0**	**0.025 / 0**	0.035 / 52.00	0.033 / 51.70	0.025 / 44.50	0.027 / 46.50	**0.029 / 57.50**	**0.030 / 59.80**
	CLIP	**0.001 / 8.60**	0.0003 / 0	0.001 / 0	0.001 / 0	0.079 / 55.60	0.078 / 56.70	0.059 / 54.80	0.061 / 54.50	0.004 / 36.20	0.004 / 35.80
	CLIP-Auto	0.001 / 8.00	0.0003 / 0	0.001 / 0	0.001 / 0	**0.064 / 56.00**	**0.091 / 57.40**	**0.062 / 55.10**	0.067 / 55.50	0.005 / 37.20	0.005 / 36.60
ViT-B/32	Standard	**0.006 / 22.00**	**0.0049 / 9.30**	**0.004 / 2.10**	**0.004 / 2.00**	**0.452 / 76.30**	**0.748 / 77.00**	**0.446 / 76.50**	**0.450 / 76.50**	**0.089 / 72.00**	**0.087 / 71.90**
	CLIP	0.001 / 10.20	0.0008 / 0	0.001 / 0	0.001 / 0	0.117 / 61.00	0.222 / 62.80	0.123 / 61.10	0.139 / 61.50	0.008 / 41.00	0.007 / 40.50
	CLIP-Auto	0.001 / 10.10	0.0009 / 0.10	0.001 / 0	0.001 / 0	0.111 / 61.80	0.243 / 64.00	0.143 / 61.50	0.145 / 62.60	0.010 / 43.70	0.010 / 43.10
ViT-B/16	Standard	**0.005 / 16.10**	**0.004 / 3.70**	**0.004 / 0.70**	**0.004 / 0.50**	**0.474 / 78.90**	**0.800 / 80.20**	**0.471 / 79.30**	**0.452 / 79.20**	**0.095 / 75.50**	**0.098 / 75.60**
	CLIP	0.001 / 6.70	0.0009 / 0	0.001 / 0	0.001 / 0	0.169 / 62.70	0.221 / 64.60	0.149 / 62.60	0.152 / 62.70	0.011 / 42.80	0.013 / 43.50
	CLIP-Auto	0.002 / 6.30	0.0010 / 0	0.001 / 0	0.001 / 0	0.171 / 64.10	0.2265 / 65.10	0.151 / 64.40	0.1535 / 64.40	0.011 / 40.30	0.012 / 39.80
ResNet50x4	CLIP	**0.001 / 12.20**	0.0003 / 0	0.001 / 0	0.001 / 0	**0.115 / 57.50**	**0.122 / 60.40**	**0.080 / 56.50**	**0.085 / 57.20**	0.006 / 37.60	0.006 / 37.30
	CLIP-Auto	0.001 / 11.70	**0.0004 / 0.10**	0.001 / 0	0.001 / 0	0.079 / 56.90	0.119 / 58.80	0.074 / 56.10	0.078 / 56.60	**0.006 / 40.70**	**0.007 / 40.60**
ResNet50x16	CLIP	0.001 / 14.60	0.0005 / 0	0.001 / 0	0.001 / 0	0.216 / 63.40	0.188 / 64.60	0.113 / 61.80	0.122 / 62.40	**0.014 / 49.00**	**0.014 / 48.40**
	CLIP-Auto	**0.001 / 15.10**	0.0005 / 0	0.001 / 0	0.001 / 0	**0.213 / 64.20**	**0.173 / 66.40**	**0.108 / 63.30**	**0.125 / 63.60**	0.012 / 48.70	0.013 / 48.20

Table 1. Model results against individual untargeted adversarial attacks under the l_∞ norm on ImageNet. Each entry consists of the median l_∞ distance of the minimum adversarial perturbations over all samples on the left, and the model accuracy for the perturbation budget $\epsilon = 8/255$ on the right. Note that there are no corresponding standard ResNet50x4 and ResNet50x16 available. We highlight the results based on accuracy.

m and its robustness enhanced version m', the relative robustness is $\text{acc}_2(m') - \text{acc}_2(m)$.

A robust model should be able to improve both effective and relative robustness. To help analyze the effective and relative robustness, we report the average accuracy across the corresponding datasets and average accuracy across the corresponding class subsets of ImageNet (Figure 1 and Figure 2). We use the average of pm-0 and pm-10 accuracy for Youtube-BB and ImageNet-Vid. We average over the five severities for each corruption in ImageNet-C. We use mFR and mT5D on ImageNet-P. We also report the median l_∞ distance of the minimum adversarial perturbations across all samples for the adversarial attacks, and the success rate on ImageNet-T and CIFAR-10-T.

3. Results

In this section, we show that our RoZ benchmark provides a comprehensive robustness evaluation of multimodal CLIP models. Except for natural distribution shifts, CLIP generally fails to improve robustness over the corresponding standard models on our benchmark. We first summarize the main results (Sec. 3.1), then describe the breakdown results with a focus on synthetic distribution shifts (Sec. 3.2) and adversarial attacks (Sec. 3.3). Details about the additional experimental setups and results are described in the appendix.

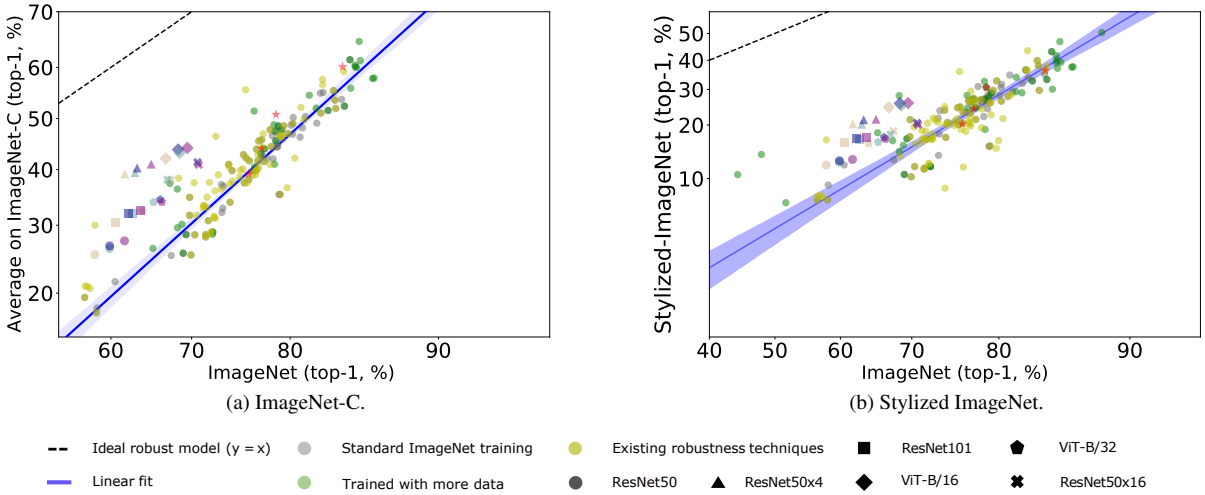
3.1. Main Results

CLIP fails to improve the robustness over the corresponding standard models in image classification. In Figure 2a, we compare the average accuracies of the zero-shot CLIP models with their CLIP-Auto versions and standard ImageNet models on all datasets. We find that most robustness improvements of CLIP are due to the significant improvements on natural distribution shift, in particular the effective robustness. The result on natural distribution shift is similar to that reported in [37]. Note that one vision model is evaluated for robustness in [37] while we evaluate multiple vision models and show that they all demonstrate similar behaviors. In Figure 1a, we summarize the performance of zero-shot CLIP models compared to existing ImageNet models and CLIP-Auto models on natural distribution shifts. The details are shown in Table 2. All CLIP models improve the effective robustness over standard ImageNet models on natural distribution shifts. These CLIP models also improve the relative robustness of the standard models on the natural distribution shifts except for one dataset, ImageNetV2. ImageNetV2 follows the original creation process of ImageNet, which suggests that the distribution is likely to be similar to the ImageNet distribution, and thus the standard models in general work well.

However, we draw contrary conclusions on the rest of our benchmark: synthetic distribution shifts and adversarial examples. CLIP has lower average robustness on these test sets. In particular, CLIP models are much more vulnerable to typographic attacks than standard models, resulting in a substantial 34.74% performance drop on average. We find that CLIP-Auto does not make much difference in the robustness performance compared to CLIP. This is in contrast to the conclusion in pre-trained language models (e.g., T5 [38] and GPT-3 [5]). The reason is that the image representation learned from large-scale pre-trained data is the key differentiator of the classification performance. The optimization of prompts that synthesize the linear classifiers on top of the image representation has limited impact. We find Vision Transformer is at least as robust as CLIP. We also find that while the effective robustness is comparable, CLIP actually reduces the relative robustness by a considerable amount compared to the corresponding standard model.

Model		ImageNet	ImageNetV2	ImageNet-R	ObjectNet	ImageNet-Sketch	ImageNet-A	Youtube-BB	ImageNet-Vid
ResNet50	Standard	**76.13**	**62.70**	35.05	35.77	22.20	0.81	50.10	60.14
	CLIP	59.85	52.64	**60.51**	**39.82**	35.46	22.75	**56.69**	**64.92**
	CLIP-Auto	61.78	54.53	60.48	39.21	**35.48**	**23.81**	56.68	64.61
ResNet101	Standard	**76.20**	**64.30**	37.70	32.60	25.20	2.70	53.10	62.80
	CLIP	62.32	56.08	**68.01**	44.17	**41.09**	29.44	61.75	71.28
	CLIP-Auto	63.83	56.95	67.37	**45.01**	40.89	**30.36**	**62.11**	**71.64**
ViT-B/32	Standard	**78.73**	**71.16**	44.69	**43.76**	32.55	**33.64**	**63.62**	**77.72**
	CLIP	63.40	55.73	**69.29**	43.60	**42.42**	31.35	60.98	74.03
	CLIP-Auto	65.16	57.51	68.10	43.16	42.15	32.21	61.01	74.80
ViT-B/16	Standard	**84.20**	**74.13**	50.89	51.12	38.10	**50.64**	**64.76**	**81.79**
	CLIP	66.94	62.52	**77.82**	**53.45**	**48.47**	49.39	64.07	80.26
	CLIP-Auto	69.51	63.11	76.83	53.08	48.43	49.56	63.28	79.76
ResNet50x4	CLIP	66.28	59.34	**72.63**	49.97	**44.75**	**41.53**	59.53	**72.50**
	CLIP-Auto	**66.48**	**59.75**	71.33	**50.36**	44.55	40.95	**59.75**	72.14
ResNet50x16	CLIP	70.67	**64.14**	**79.18**	58.86	**50.66**	**56.41**	**62.96**	**78.45**
	CLIP-Auto	**70.81**	63.90	78.33	**59.04**	49.62	55.93	61.97	78.04

Table 2. Model accuracy on ImageNet and seven natural distribution shifts.



(a) ImageNet-C. (b) Stylized ImageNet.

Figure 4. Model accuracies on two synthetic distribution shifts. Different from the results on natural distribution shifts, we show that CLIP fails to improve the robustness compared to standard models. Red: standard ImageNet models. Blue: zero-shot CLIP models. Purple: CLIP-Auto models.

3.2. Synthetic Distribution Shifts

In Figure 2b and Figure 4, we summarize the model accuracies on two synthetic distribution shifts, ImageNet-C and Stylized ImageNet. We show detailed results on ImageNet-C and ImageNet-P in the appendix. We observe downgraded robustness of CLIP compared to the corresponding standard models. We also find that the performance of CLIP and CLIP-Auto are comparable, again suggesting that prompt enhancement does not improve its robustness. The reason is that the key ingredient to the robustness of multimodal CLIP is pre-trained image representations, while the impact of language prompts is limited. This is different from the conclusion in language models, as language prompts improve robustness in a single modal setup. Also, we suspect that this is due to images of synthetic distribution shifts are not in the pre-training data and present a further analysis in Sec. 4. Improving the zero-shot robustness to synthetic distribution shifts via regularization techniques is one of our future investigations.

3.3. Adversarial Attacks

Common Attacks We test the robustness to adversarial examples, which is crucial for safety-critical applications. We compare the average accuracies under common adversarial attacks in Figure 2c. On CIFAR-10, we use zero-shot CLIP, CLIP-Auto, and linear probe standard models for the attacks. The CLIP models are more vulnerable when compared to the standard models. We illustrate results on ImageNet in Table 1. The trends on ImageNet and CIFAR-10 are similar. CLIP-Auto does not improve the adversarial robustness over CLIP. The reason is that the pre-training stage does not include adversarial examples and is not robust to adversarial attacks. A further discussion is presented in Sec. 4. Improving the zero-shot robustness to adversarial attacks via adversarial prompt learning is an important future direction.

Model		CIFAR-10-T Success rate / Accuracy	ImageNet-T Success rate / Accuracy
ResNet50	Standard	**8.16 / 13.90**	**0.03 / 68.66**
	CLIP	99.40 / 0.22	40.42 / 22.61
	CLIP-Auto	99.46 / 0.18	44.86 / 21.50
ResNet101	Standard	**6.49 / 27.03**	**0.03 / 69.70**
	CLIP	97.97 / 1.61	37.76 / 25.25
	CLIP-Auto	98.07 / 1.45	42.47 / 23.93
ViT-B/32	Standard	**3.73 / 65.12**	**0.02 / 75.03**
	CLIP	77.47 / 17.08	13.07 / 45.06
	CLIP-Auto	75.24 / 17.67	12.64 / 47.01
ViT-B/16	Standard	**2.10 / 67.42**	**0.02 / 80.05**
	CLIP	89.84 / 9.51	25.07 / 45.59
	CLIP-Auto	90.54 / 8.79	25.14 / 45.99
ResNet50x4	CLIP	98.75 / 1.06	**44.21 / 27.14**
	CLIP-Auto	**99.13 / 0.72**	48.28 / 24.90
ResNet50x16	CLIP	97.85 / 2.07	**50.22 / 27.99**
	CLIP-Auto	**97.59 / 2.30**	52.29 / 26.80

Table 3. Model accuracies and success rates under typographic attacks on ImageNet-T and CIFAR-10-T. We highlight the results based on accuracy. CLIP causes a significant robustness drop compared to the corresponding standard models.

Additional results are shown in the appendix.

Typographic Attacks CLIP is extremely vulnerable to our new robustness test sets (ImageNet-T and CIFAR-10-T) that are based on a new kind of non-programmatic attack named typographic attacks [14]. In Figure 2d and Figure 5, we find that CLIP reduces both effective and relative robustness by a large amount (-34.74% in average accuracy) compared to the standard models on ImageNet-T and CIFAR-10-T. The attack success rate is also much higher for the CLIP models (Table 3). The underlying reason is that, different from standard models, multimodal CLIP learns to respond to both images and text given a concept. Adding adversarial text to images can fool the CLIP models. This also applies to CLIP-Auto, as learnable prompts still correspond to a visual concept. We plan to improve the zero-shot robustness to typographic attacks via regularization techniques to force CLIP to only focus on image representation. The results indicate ImageNet-T and CIFAR-10-T are important test sets for understanding zero-shot robustness.

4. Data Overlap Analysis

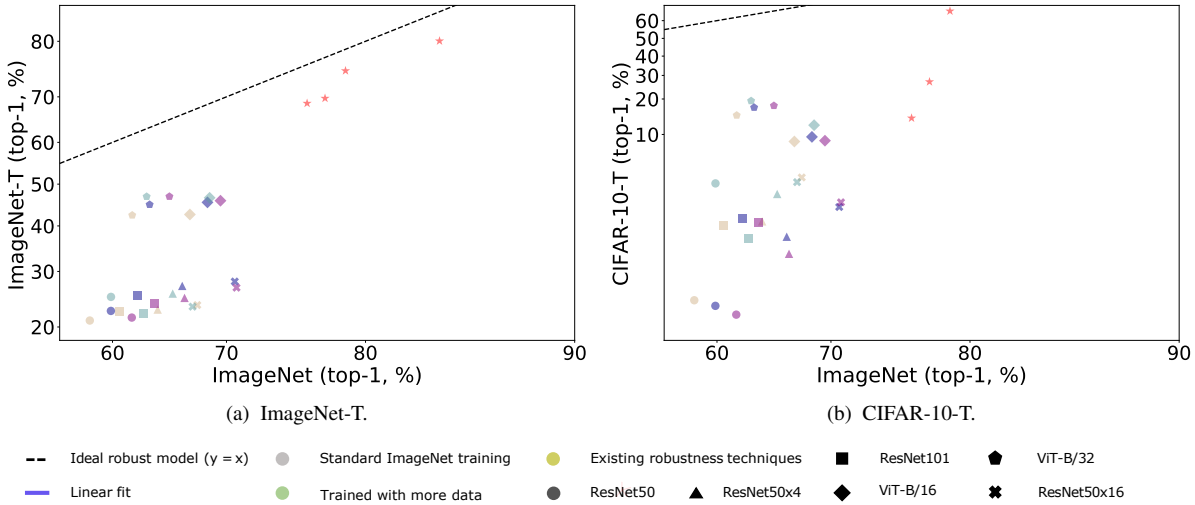
As shown in Sec. 3, while CLIP achieves improved robustness on natural distribution shifts, it fails to transfer to other robustness test sets in RoZ benchmark. A growing problem when training high-capacity models on large-scale datasets is data contamination, where the pre-training dataset can potentially include content from the test datasets because such content is on the web. We suspect that the data contamination issue in the pre-training data actually results in the performance on natural distribution shifts.

Although CLIP [37] has conducted data overlap analysis, we find the analysis is not rigorous since it assumes the overlapped images share the same distribution with the test sets. We propose to rigorously measure the data overlap between the CLIP pre-training data and the robustness test sets. The main idea is to remove image examples that are the same or similar to training examples from test sets. The "cleaned" test sets can be used for robustness re-evaluation. In particular, we use the image encoder of ResNet50x16 as the duplication detector as it is trained on the same distribution as the pre-training set. The deduplication threshold is defined as the cosine similarity between image representations. We consider the images where the similarity between them are beyond the deduplication threshold as overlapped images, which are then removed from the test sets. Since the entire pre-training set of CLIP has not been released, we use a subset of it, YFCC100M dataset [48]. We focus on the comparison between natural and synthetic distribution shifts.

Figure 6a and Figure 6b show the overall comparison and a case study. Importantly, while performance drops on ImageNetV2, we see an accuracy improvement on Stylized ImageNet. This indicates that the natural distribution shift benefits from the data overlap as the pre-training set contains similar images. While the above analysis is based on a small subset of original CLIP training data, more severe data overlap issues may exist when testing on the full training data. Besides, we simulate the pre-training data using Google Images, and find that very similar (or even the same) images are found by Google Images through querying with an image in ImageNetV2. The results are presented in the appendix. We argue that it is crucial to clean the pre-training data to test robustness, which is a common practice when training large-scale models. For example, GPT-3 [5] has reported significantly inflated results due to the data overlap issue. Our probe is an initial attempt to understand the role of data overlap on robustness. We hope that the analysis promotes further research along this line.

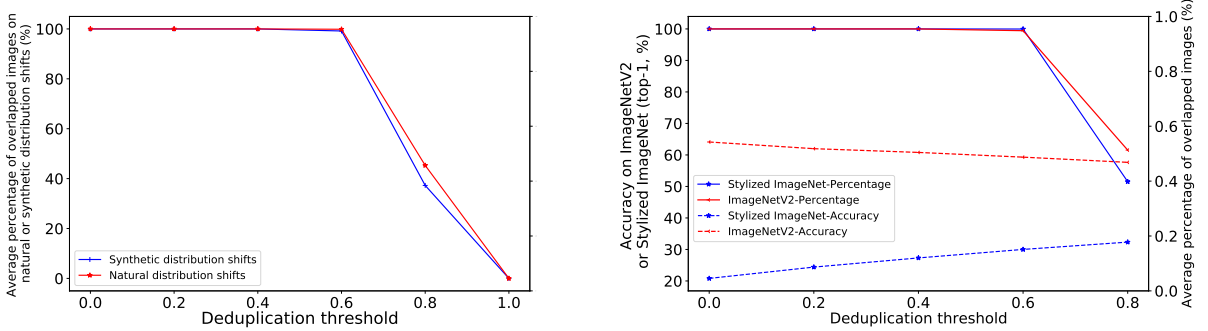
5. Related Work

Radford et al. [36] focus on the robustness of CLIP to natural distribution shifts. Taori et al. [46] conduct a comprehensive robustness study of the image classification models on both natural distribution shifts and synthetic distribution shifts. Dong et al. [9] evaluate the adversarial robustness of image classification models. Koh et al. [24] propose a new natural distribution shifts datasets. Shen et al. [44] propose a new vision-language STEM understanding dataset. Feuer et al. [11, 12] evaluate robustness to various distributions. Different from these studies, we focus on conducting a comprehensive study of the zero-shot robustness in the domain of image classification, considering all natural distribution shifts, synthetic distribution shifts, and worst-case adversarial examples. In addition, we benchmark the robustness under a new kind of attack, typographic attacks. To the best of our knowledge, this is the first work to quantitatively



(a) ImageNet-T. (b) CIFAR-10-T.

Figure 5. Model accuracies under typographic attacks on our ImageNet-T and CIFAR-10-T. Red: standard ImageNet models. Blue: zero-shot CLIP models. Purple: CLIP-Auto models.



(a) Percentage of overlapped images on natural or synthetic distribution shifts.

(b) Percentage of overlapped images and corresponding accuracies on ImageNetV2 and Stylized ImageNet. Left y-axis: Accuracy. Right y-axis: Percentage.

Figure 6. Data overlap between YFCC100M (pre-training data) and distribution shifts.

measure the robustness under typographic attacks.

Radford et al. [37] test the zero-shot robustness using the default natural language prompts [6]. We measure the robustness by disentangling the effect of the natural language prompts in the evaluation. Zero-shot robustness of language models has found pre-training improves the relative robustness [21], but little evidence of effective robustness improvements [30]. In comparison, we test the zero-shot robustness in image classification.

Li et al. [27] use the learned visual n-grams to perform zero-shot image classification. VirTex [7], ICMLM [42] and ConVIRT [56] have demonstrated the usage of natural language in learning image representations. Gomez et al. [15] and Joulin et al. [23] also introduce methods that learn visual representations from natural language supervision. In this work, we focus on the current state-of-the-art model learned from the text, CLIP. There are similar models including GLIP [29], GLIDE [32], and BLIP [28]. We leave the study of the robustness of other work as a future direction.

6. Conclusion

We construct a comprehensive benchmark to study the zero-shot robustness of multimodal foundation models using CLIP as a pilot study. Our results show that CLIP is not robust under synthetic distribution shifts and adversarial attacks, and its previously reported robustness under natural distribution shifts might be attributed, at least in part, to data overlap. The finding differs from the original finding in the CLIP paper, where they conclude that the model is more robust than standard models trained on ImageNet. In order to benefit safety-critical applications, our results suggest that it is crucial to conduct comprehensive robustness evaluations. We hope our results will foster further research into the zero-shot robustness of foundation models.

References

[1] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access*, pages 14410–14430, 2018. 11

[2] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *ICML*, pages 284–293, 2018. 15

[3] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *NeurIPS*, pages 9453–9463, 2019. 1, 2

[4] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017. 15

[5] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 5, 7, 16

[6] Nicholas Crispino, Kyle Montgomery, Fankun Zeng, Dawn Song, and Chenguang Wang. Agent instructs large language models to be general zero-shot reasoners. *arXiv preprint arXiv:2310.03710*, 2023. 8

[7] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. *arXiv preprint arXiv:2006.06666*, 2020. 8

[8] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *CVPR*, pages 9185–9193, 2018. 2, 11

[9] Yinpeng Dong, Qi-An Fu, Xiao Yang, Tianyu Pang, Hang Su, Zihao Xiao, and Jun Zhu. Benchmarking adversarial robustness on image classification. In *CVPR*, pages 321–331, 2020. 2, 7, 11

[10] 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. *arXiv preprint arXiv:2010.11929*, 2020. 4

[11] Benjamin Feuer, Ameya Joshi, and Chinmay Hegde. A meta-analysis of distributionally-robust models. *arXiv preprint arXiv:2206.07565*, 2022. 7

[12] Benjamin Feuer, Ameya Joshi, Minh Pham, and Chinmay Hegde. Distributionally robust classification on a data budget. *arXiv preprint arXiv:2308.03821*, 2023. 7

[13] 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. *arXiv preprint arXiv:1811.12231*, 2018. 2, 11

[14] Gabriel Goh, Nick Cammarata †, Chelsea Voss †, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 2021. 2, 3, 7, 14, 15

[15] Lluis Gomez, Yash Patel, Marçal Rusiñol, Dimosthenis Karatzas, and CV Jawahar. Self-supervised learning of visual features through embedding images into text topic spaces. In *CVPR*, pages 4230–4239, 2017. 8

[16] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1, 2, 11

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 4, 11

[18] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 1, 2

[19] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *arXiv preprint arXiv:1907.07174*, 2019. 2

[20] 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. *arXiv preprint arXiv:2006.16241*, 2020. 2

[21] Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. Pretrained transformers improve out-of-distribution robustness. *arXiv preprint arXiv:2004.06100*, 2020. 8

[22] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *ICML*, pages 2137–2146, 2018. 2, 11

[23] Armand Joulin, Laurens Van Der Maaten, Allan Jabri, and Nicolas Vasilache. Learning visual features from large weakly supervised data. In *ECCV*, pages 67–84, 2016. 8

[24] 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*, pages 5637–5664. PMLR, 2021. 7

[25] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 2

[26] Alexey Kurakin, Ian Goodfellow, Samy Bengio, et al. Adversarial examples in the physical world, 2016. 2, 11

[27] Ang Li, Allan Jabri, Armand Joulin, and Laurens van der Maaten. Learning visual n-grams from web data. In *ICCV*, pages 4183–4192, 2017. 8

[28] 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. 8

[29] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, pages 10955–10965, 2022. 8

[30] John Miller, Karl Krauth, Benjamin Recht, and Ludwig Schmidt. The effect of natural distribution shift on question answering models. In *ICML*, pages 6905–6916, 2020. 8

[31] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool

deep neural networks. In *CVPR*, pages 2574–2582, 2016. 2, 11

[32] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, pages 16784–16804, 2022. 8

[33] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. 16

[34] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against deep learning systems using adversarial examples. *arXiv preprint arXiv:1602.02697*, page 3, 2016. 11

[35] Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. How context affects language models' factual predictions. *arXiv preprint arXiv:2005.04611*, 2020. 16

[36] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, page 9, 2019. 4, 7

[37] 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. *arXiv preprint arXiv:2103.00020*, 2021. 1, 3, 4, 5, 7, 8, 15

[38] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019. 5

[39] Esteban Real, Jonathon Shlens, Stefano Mazzocchi, Xin Pan, and Vincent Vanhoucke. Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video. In *CVPR*, pages 5296–5305, 2017. 2

[40] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, pages 5389–5400, 2019. 2

[41] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, pages 211–252, 2015. 2

[42] Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. Learning visual representations with caption annotations. *arXiv preprint arXiv:2008.01392*, 2020. 8

[43] Vaishaal Shankar, Achal Dave, Rebecca Roelofs, Deva Ramanan, Benjamin Recht, and Ludwig Schmidt. Do image classifiers generalize across time? *arXiv preprint arXiv:1906.02168*, 2019. 2

[44] Jianhao Shen, Ye Yuan, Srbuhi Mirzoyan, Ming Zhang, and Chenguang Wang. Measuring vision-language stem skills of neural models. *arXiv preprint arXiv:2402.17205*, 2024. 7

[45] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*, 2020. 4

[46] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *NeurIPS*, 2020. 2, 4, 7, 11, 12

[47] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 16

[48] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 7

[49] Jonathan Uesato, Brendan O'donoghue, Pushmeet Kohli, and Aaron Oord. Adversarial risk and the dangers of evaluating against weak attacks. In *ICML*, pages 5025–5034, 2018. 2, 11

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 4

[51] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing NLP. In *EMNLP*, pages 2153–2162, 2019. 15

[52] Chenguang Wang, Xiao Liu, and Dawn Song. Language models are open knowledge graphs. *arXiv preprint arXiv:2010.11967*, 2020. 16

[53] Chenguang Wang, Xiao Liu, Zui Chen, Haoyun Hong, Jie Tang, and Dawn Song. Deepstruct: Pretraining of language models for structure prediction. In *ACL*, 2022. 16

[54] Haohan Wang, Songwei Ge, Eric P Xing, and Zachary C Lipton. Learning robust global representations by penalizing local predictive power. *arXiv preprint arXiv:1905.13549*, 2019. 2

[55] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *CVPR*, pages 2730–2739, 2019. 2, 11

[56] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020. 8, 16

[57] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022. 16

A. The RoZ Benchmark Details

Distribution Shifts Settings

- **Natural Distribution Shifts**. We follow the settings in [46]. For ImageNetV2, we report results on imagenetv2-matched-frequency-format-val. For ImageNet-A, ImageNet-R, Youtube-BB, ImageNet-Vid, and ObjectNet, we only take predictions that are also in the category of the ImageNet validation set. We use the same setup for ImageNet Sketch as in [46].
- **Synthetic Distribution Shifts**. For ImageNet-C, we use all 15 common corruption types including: gaussian noise (on disk), shot noise (on disk), impulse noise (on disk), defocus blur (on disk), glass blur (on disk), motion blur (on disk), zoom blur (on disk), snow (on disk), frost (on disk), fog (on disk), brightness (on disk), contrast (on disk), elastic transform (on disk), pixelate (on disk), jpeg compression (on disk). For each corruption, we average over the five severities. For ImageNet-P, we use 10 common perturbations: gaussian noise, shot noise, motion blur, zoom blur, snow, brightness, translate, rotate, tilt, and scale. We also use the Stylized ImageNet dataset [13].

Common Adversarial Attack Methods Based on the different levels of knowledge of the target model, we consider the following attack scenarios from white-box attacks that have access to the model architectures and parameters, to transfer-based attacks and black-box attacks that only have access to the training data or model outputs. There are two typical strategies for creating adversarial examples with small perturbations. The first results in adversarial examples with a constrained perturbation, while the second strategy produces an adversarial example with an optimized perturbation.

- **White-Box Attacks**. White-box attacks rely on detailed information of the target model. White-box attacks craft adversarial examples based on the gradient of the input. We include the following widely-used attack methods: fast gradient sign method (FGSM) [16], basic iterative method (BIM) [26], DeepFool [31], and momentum iterative method (MIM) [8]. We empirically set the number of iterations to 12 and 20 for both BIM and MIM in two strategies respectively. We set the maximum number of iterations as 50 for DeepFool in two strategies.
- **Transfer-Based Attacks**. The attacks have access to the training data and leverage the adversarial transferability [34], aiming to obtain a substitute model from which the adversarial examples are created. We craft adversarial examples from the above white-box methods on a substitute model including FGSM, BIM, and MIM. Besides, we incorporate the diverse inputs method (DIM) [55] to improve adversarial transferability. We use a ResNet-152 model [17] as the substitute model following [9]. We set the number of iterations to 12 and 10 for all methods in both strategies.
- **Black-Box Attacks**. Black-box attacks, in particular, score-based black-box attacks only have access to output probabilities via querying the target model. Therefore the gradient can be estimated by gradient-free methods. We include NES [22] and SPSA [49] that conduct gradient estimation based on random samples and the corresponding loss. We set the number of iterations to 12 and 20 for both methods in two evaluation settings.

To ease the reproductivity, we use the same hyperparameters as [9] for all methods. We refer readers to [1] for a survey of the attack methods.

ResNet50	Standard	**20.24**
	CLIP	12.71
	CLIP-Auto	12.92
ResNet101	Standard	**24.25**
	CLIP	17.01
	CLIP-Auto	17.20
ViT-B/32	Standard	**30.58**
	CLIP	17.19
	CLIP-Auto	16.80
ViT-B/16	Standard	**36.36**
	CLIP	25.76
	CLIP-Auto	25.86
ResNet50x4	CLIP	21.32
	CLIP-Auto	**21.43**
ResNet50x16	CLIP	**20.52**
	CLIP-Auto	20.07

Table 4. Model accuracies on Stylized ImageNet. CLIP is not able to improve the robustness performance over the corresponding standard models. CLIP and CLIP-Auto perform similarly.

Implementation Details For all distribution shift datasets, we leverage the image classification testbed [46][1] to evaluate the results. The testbed includes all the standard models. We integrate the released CLIP models[2] and the corresponding CLIP-Auto models into the testbed. For adversarial attack experiments, we implement all adversarial attack methods for the evaluation. For all experiments, we use: 4 GeForce GTX 1080 with a batch size of 32 per GPU. The average runtime is approximately 60 minutes on seven natural distribution shifts, 10 minutes on ImageNet-P and Stylized-ImageNet, 450 minutes on ImageNet-C, and 45 minutes under 11 adversarial attacks.

B. More Results

Sec. 3 provides a high-level summary of the robustness results, we show the breakdown results on each individual dataset in our RoZ benchmark in this section.

[1] https://modestyachts.github.io/imagenet-testbed/
[2] https://github.com/openai/CLIP

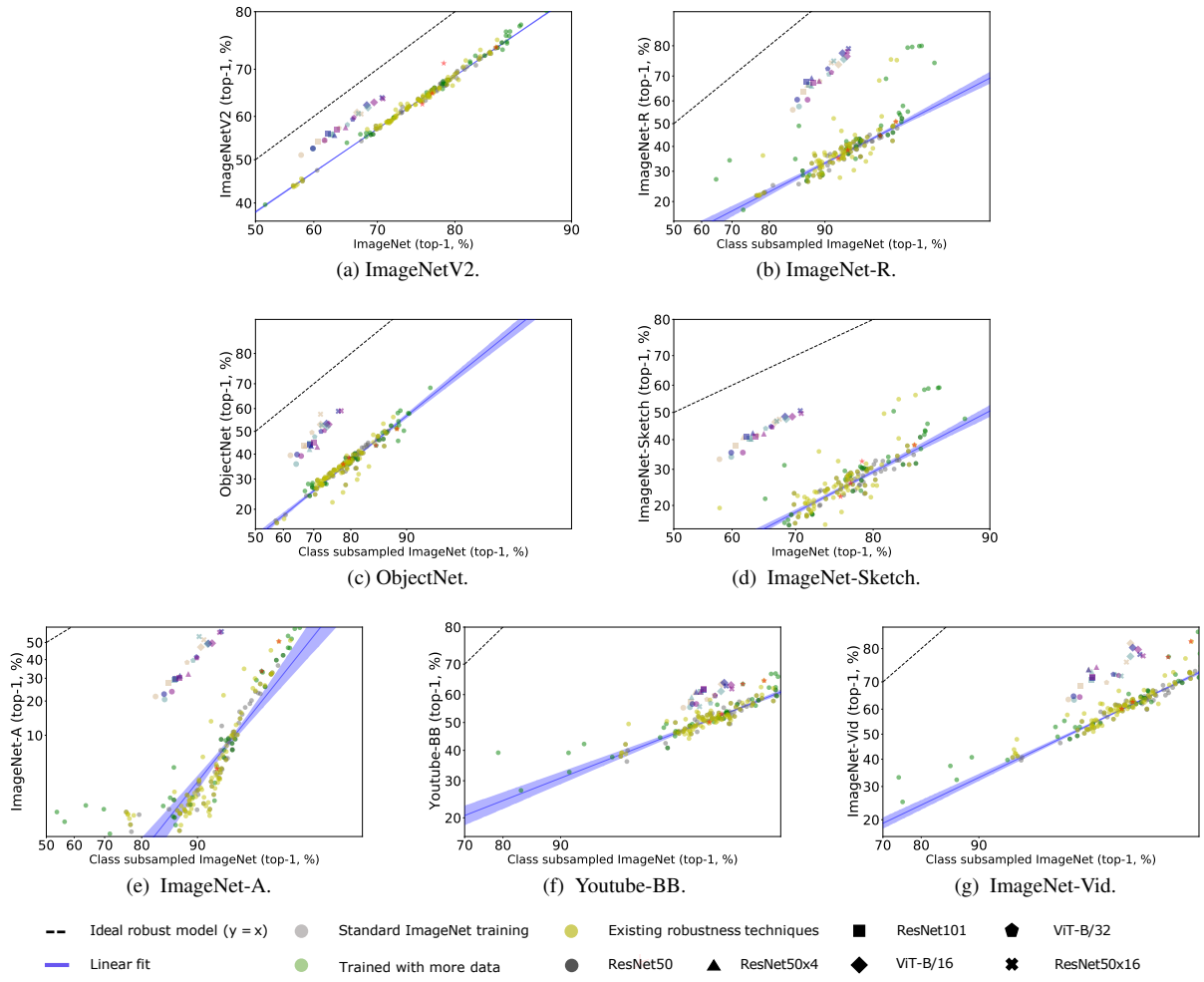


Figure 7. Model accuracies on the seven natural distribution shifts. Red: standard ImageNet models. Blue: zero-shot CLIP models. Purple: CLIP-Auto models.

B.1. Distribution Shifts

Natural Distribution Shifts Performance per dataset in natural distribution shifts is shown in Figure 7. The CLIP results are compared to the results of standard ImageNet models in the testbed [46]. We find that CLIP models obtain state-of-the-art effective robustness on all datasets. Compared to CLIP models, CLIP-Auto obtains comparable performance. The reason is that the pre-trained image features are the key to the performance, while the prompts that only synthesize the classifiers based on the image features have limited impact.

Synthetic Distribution Shifts We show the performance on each individual dataset of synthetic distribution shifts. Figure 4a illustrates that CLIP models fail to improve the robustness over corresponding standard ImageNet models on ImageNet-C, which is in contrast to the results on natural distribution shifts. This also confirms the findings in [46] that robustness under synthetic distribution shifts does not imply that the corresponding model has robustness on natural distribution shifts. Similar to the observation on natural distribution shifts, CLIP-Auto achieves comparable effective and relative robustness with CLIP, suggesting that the impact of CLIP models on the robustness is limited.

In Table 6, we compare two common metrics on ImageNet-P: mean flip rate (mFR) and mean top-5 distance (mT5D). We find that all CLIP and CLIP-Auto models generally reduce the performance compared to the standard models. In addition, we show the results on Stylized ImageNet in Figure 4b and Table 4. We draw the same conclusion that CLIP does not improve robustness. CLIP and CLIP-Auto produce comparable performance.

Model				Noise			Blur				Weather				Digital			
		Original Acc	Avg Acc	Gauss	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixel	JPEG
ResNet50	Standard	**76.13**	**39.17**	**29.29**	**27.03**	**23.81**	**38.75**	**26.79**	**38.67**	**36.24**	**32.53**	**38.14**	**45.83**	**68.02**	**39.06**	**45.25**	**44.79**	**53.41**
	CLIP	59.85	26.67	19.63	18.09	12.23	26.06	14.78	25.36	23.61	22.68	26.98	38.55	50.35	32.89	28.29	28.13	32.42
	CLIP-Auto	61.78	27.46	19.93	18.31	11.83	27.09	14.90	26.29	24.02	23.26	27.82	39.46	52.38	33.95	29.08	29.58	33.97
ResNet101	Standard	**77.37**	**44.10**	**34.77**	**32.63**	**29.02**	**44.27**	**32.60**	**44.31**	**40.78**	**36.49**	**41.73**	**49.14**	**69.85**	**42.98**	**50.11**	**53.51**	**59.30**
	CLIP	62.32	32.07	24.81	23.34	19.03	31.17	18.77	30.14	26.90	28.32	31.53	43.38	54.57	39.01	31.69	38.91	39.49
	CLIP-Auto	63.83	32.57	25.31	23.60	18.93	31.81	18.75	30.85	27.51	28.20	31.80	44.14	55.68	39.89	32.11	39.44	40.46
ViT-B/32	Standard	**78.73**	**50.79**	**44.05**	**41.40**	**41.30**	**48.59**	**45.81**	**54.48**	**43.20**	**31.50**	**38.99**	**56.81**	**67.77**	**59.72**	**57.45**	**66.59**	**64.13**
	CLIP	63.40	40.31	37.21	35.36	33.65	40.34	29.18	41.17	32.13	34.29	37.13	46.52	57.62	45.22	40.96	47.26	46.60
	CLIP-Auto	65.16	41.04	37.77	35.69	33.95	41.12	29.44	42.19	32.65	34.65	37.43	47.27	59.14	46.15	41.97	48.35	47.81
ViT-B/16	Standard	**84.20**	**60.12**	**54.11**	**52.74**	**52.14**	**54.28**	**51.4**	**62.84**	**55.21**	**60.48**	**59.03**	**61.48**	**78.18**	**55.70**	**63.15**	**71.46**	**69.63**
	CLIP	68.43	43.87	39.48	37.82	35.05	42.03	33.22	44.75	36.60	43.52	42.72	51.53	62.33	47.92	41.27	49.76	50.01
	CLIP-Auto	69.51	44.19	39.79	38.09	35.08	42.44	33.17	45.08	36.83	43.93	42.94	51.76	63.17	48.09	41.30	50.24	50.91
ResNet50x4	CLIP	66.28	**34.43**	**27.76**	**26.57**	**23.59**	**31.36**	**18.70**	**31.67**	**28.34**	**30.77**	**34.39**	**45.72**	57.81	**39.80**	**32.57**	42.40	44.94
	CLIP-Auto	**66.48**	33.98	27.05	25.82	22.84	30.60	18.04	30.85	27.75	30.44	33.98	45.15	**57.83**	39.78	32.21	42.17	45.23
ResNet50x16	CLIP	70.67	**41.39**	**38.22**	**37.16**	**35.15**	**36.60**	**23.49**	**38.34**	**34.08**	**38.08**	**39.33**	**50.88**	**62.37**	**46.23**	**37.68**	49.67	53.63
	CLIP-Auto	**70.81**	40.93	37.63	36.34	34.68	36.22	22.55	37.55	32.95	37.22	38.93	50.56	62.42	46.00	37.13	**50.05**	**53.66**

Table 5. Individual top-1 accuracy scores on all the corruption types of ImageNet-C. "Original Acc" refers to the accuracy of the clean ImageNet validation set. "Avg Acc" denotes the average accuracy of 15 common corruptions.

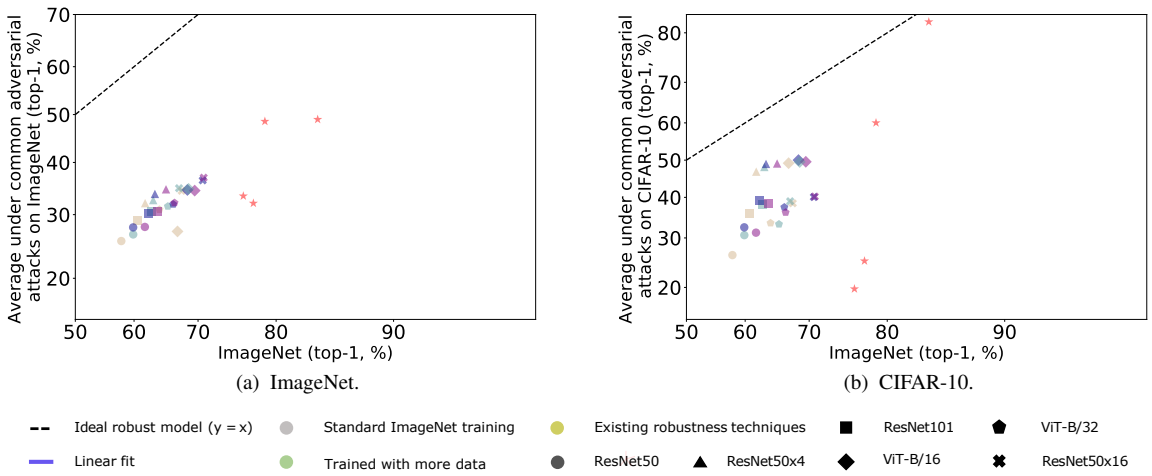
Model			Noise		Blur		Weather		Digital			
		mFR	Gauss	Shot	Motion	Zoom	Snow	Bright	Translate	Rotate	Tilt	Scale
ResNet50	Standard	**57.90**	**59.00**	**58.00**	**64.00**	**72.00**	**63.00**	**62.00**	**44.00**	**52.00**	**57.00**	**48.00**
	CLIP	113.69	99.59	93.34	129.19	147.51	118.32	148.67	99.13	102.11	129.09	70.00
	CLIP-Auto	110.62	95.18	90.29	125.63	143.40	117.65	143.62	96.41	99.66	125.83	68.55
ResNet101	Standard	**53.02**	**55.10**	**51.70**	**53.77**	**63.25**	**58.80**	**57.32**	**42.25**	**48.96**	**53.65**	**45.35**
	CLIP	98.91	82.82	78.80	108.99	129.43	104.07	129.08	89.22	90.17	113.72	62.79
	CLIP-Auto	97.23	81.51	77.38	108.10	127.41	103.15	127.22	86.67	88.12	110.71	62.08
ViT-B/32	Standard	**35.26**	**28.23**	**27.76**	**32.70**	**46.46**	**34.86**	**56.44**	**35.79**	**45.69**	**43.05**	**44.54**
	CLIP	74.53	64.56	58.40	66.85	93.94	71.13	94.87	67.26	76.61	88.34	63.35
	CLIP-Auto	72.14	61.17	56.20	65.02	92.26	70.00	90.61	64.49	74.81	85.04	61.81
ViT-B/16	Standard	**33.15**	**34.13**	**33.86**	**27.17**	**39.63**	**19.90**	**47.63**	**26.03**	**31.84**	**36.30**	**35.06**
	CLIP	64.85	57.37	52.92	60.26	85.93	54.98	79.75	57.62	67.08	79.96	52.59
	CLIP-Auto	63.58	56.11	52.03	59.67	83.72	54.28	77.24	56.95	66.13	78.00	51.72
ResNet50x4	CLIP	90.07	**71.05**	**68.37**	107.20	125.19	**101.92**	117.69	72.61	79.64	101.59	55.47
	CLIP-Auto	**89.42**	71.15	68.57	**106.49**	**122.51**	102.64	**116.73**	**72.52**	**78.80**	**99.77**	**55.03**
ResNet50x16	CLIP	76.78	57.59	55.10	92.47	111.45	**88.87**	102.99	56.70	67.19	89.35	46.08
	CLIP-Auto	**76.23**	**56.67**	**54.61**	**92.22**	**110.02**	89.08	**101.74**	**56.51**	**67.06**	**88.75**	**45.60**

(a) The mean flip rate (mFR) across all perturbations.

Model			Noise		Blur		Weather		Digital			
		mT5D	Gauss	Shot	Motion	Zoom	Snow	Bright	Translate	Rotate	Tilt	Scale
ResNet50	Standard	**78.20**	**82.00**	**79.00**	**84.00**	**89.00**	**80.00**	**84.00**	**64.00**	**73.00**	**80.00**	**67.00**
	CLIP	113.36	100.52	94.79	128.12	144.02	118.11	144.64	97.28	102.42	124.14	79.60
	CLIP-Auto	112.58	99.03	94.11	126.98	142.5	118.68	142.34	97.22	102.13	123.32	79.52
ResNet101	Standard	**74.37**	**80.78**	**75.62**	**74.82**	**81.54**	**77.05**	**79.30**	**62.45**	**70.72**	**76.21**	**65.22**
	CLIP	102.95	88.59	84.83	113.54	130.41	107.79	130.73	91.34	94.19	113.95	74.17
	CLIP-Auto	102.60	88.13	84.42	113.51	129.98	107.99	130.05	90.82	93.75	113.49	73.87
ViT-B/32	Standard	**66.58**	**42.42**	**42.20**	**45.85**	**56.53**	**45.55**	**69.09**	**189.21**	**59.54**	**55.92**	**59.53**
	CLIP	81.19	72.78	66.85	76.38	98.42	78.51	99.52	71.46	82.40	91.56	74.04
	CLIP-Auto	80.04	70.46	66.04	75.39	97.20	77.84	96.97	70.83	81.85	90.29	73.54
ViT-B/16	Standard	**53.26**	**58.81**	**57.83**	**45.97**	**56.56**	**33.92**	**68.96**	**43.05**	**52.33**	**60.05**	**55.17**
	CLIP	77.13	70.69	65.90	73.60	94.08	67.82	91.99	68.16	79.38	91.22	68.49
	CLIP-Auto	77.17	70.20	66.06	73.85	93.90	68.25	90.79	68.67	79.74	91.53	68.72
ResNet50x4	CLIP	98.25	80.76	78.25	114.14	128.89	**109.35**	124.89	80.48	88.09	108.25	69.42
	CLIP-Auto	**97.74**	**80.19**	**77.62**	**113.99**	**127.88**	110.22	**123.88**	**79.87**	**87.47**	**107.30**	**69.00**
ResNet50x16	CLIP	90.52	71.21	68.30	107.04	123.43	**102.17**	117.19	69.26	80.68	103.20	62.69
	CLIP-Auto	**89.73**	**70.28**	**67.59**	**106.36**	**122.45**	102.30	**115.45**	**68.37**	**80.15**	**102.08**	**62.30**

(b) The mean top-5 distance (mT5D) across all perturbations.

Table 6. Results on ImageNet-P. CLIP reduces the robustness compared to standard models.



(a) ImageNet. (b) CIFAR-10.

Figure 8. Model accuracies under common adversarial attacks on ImageNet and CIFAR-10. Similar to the results on synthetic distribution shifts, CLIP is more vulnerable to adversarial attacks than standard models. Red: standard ImageNet models. Blue: zero-shot CLIP models. Purple: CLIP-Auto models.

Attack Setting		White-Box Attacks				Transfer-Based Attacks				Black-Box Attacks	
Model		FGSM	DeepFool	BIM	MIM	FGSM	BIM	MIM	DIM	NES	SPSA
ResNet50	Standard	0.001 / 34.60	0.0001 / 0	0.001 / 0	0.001 / 0	0.019 / 44.74	0.026 / 42.16	0.025 / 36.00	0.027 / 37.74	0.001 / 1.10	0.001 / 1.10
	CLIP	0.001 / 35.30	0.0002 / 0	0.001 / 0	0.001 / 0	**0.086 / 61.40**	**0.138 / 61.80**	**0.103 / 60.00**	**0.115 / 70.20**	0.002 / 17.70	**0.002 / 18.40**
	CLIP-Auto	**0.001 / 36.40**	0.0002 / 0	0.001 / 0	0.001 / 0	0.081 / 61.30	0.136 / 61.30	0.101 / 59.30	0.116 / 59.20	0.002 / 17.10	0.002 / 17.60
ResNet101	Standard	0.001 / 43.13	0.0002 / 0	0.001 / 0	0.001 / 0	0.025 / 56.36	0.041 / 50.27	0.040 / 45.90	0.047 / 49.03	0.002 / 3.00	0.002 / 2.40
	CLIP	**0.001 / 45.70**	0.0004 / 0	0.001 / 0	0.001 / 0	**0.225 / 69.60**	**0.281 / 70.30**	**0.174 / 69.40**	**0.220 / 70.20**	**0.004 / 33.90**	**0.005 / 33.10**
	CLIP-Auto	0.001 / 44.30	0.0004 / 0	0.001 / 0	0.001 / 0	0.222 / 68.50	0.274 / 69.20	0.189 / 68.20	0.227 / 68.40	0.005 / 32.40	0.005 / 32.50
ViT-B/32	Standard	**0.016 / 37.60**	**0.0106 / 9.40**	**0.006 / 0.80**	**0.007 / 0.6**	**0.576 / 94.10**	**0.876 / 94.60**	**0.591 / 94.10**	**0.594 / 94.30**	**0.093 / 87.30**	**0.094 / 87.30**
	CLIP	0.005 / 26.79	0.0020 / 0.25	0.002 / 0	0.002 / 0	0.692 / 83.98	0.870 / 88.30	0.582 / 83.90	0.592 / 84.12	0.023 / 61.10	0.023 / 61.30
	CLIP-Auto	0.007 / 26.24	0.0023 / 0.10	0.002 / 0.06	0.002 / 0	0.687 / 84.23	0.875 / 87.23	0.583 / 84.18	0.592 / 86.00	0.024 / 61.60	0.023 / 61.20
ViT-B/16	Standard	**- / 91.30**	0.0026 / 0	**- / 91.10**	**- / 91.20**	**- / 93.80**	**- / 95.10**	**- / 94.30**	**- / 93.90**	**- / 83.90**	**- / 83.80**
	CLIP	0.004 / 30.20	0.0020 / 0	0.002 / 0	0.002 / 0	0.5005 / 84.80	0.573 / 88.20	0.454 / 85.20	0.503 / 85.40	0.024 / 62.80	0.025 / 63.50
	CLIP-Auto	0.005 / 30.10	0.0030 / 0	0.002 / 0	0.002 / 0	0.543 / 84.40	0.573 / 87.80	0.464 / 85.30	0.5065 / 85.30	0.026 / 61.30	0.026 / 61.30
ResNet50x4	CLIP	**0.001 / 54.00**	0.0003 / 0	0.001 / 0	0.001 / 0	**0.274 / 63.20**	**0.312 / 64.80**	**0.262 / 62.70**	**0.301 / 61.50**	**0.004 / 34.90**	0.004 / 33.50
	CLIP-Auto	0.001 / 50.90	0.0003 / 0	0.001 / 0	0.001 / 0	0.266 / 60.70	0.306 / 63.40	0.256 / 59.60	0.275 / 60.30	0.004 / 32.90	**0.005 / 33.60**
ResNet50x16	CLIP	0.001 / 62.80	0.0003 / 0	0.001 / 0	0.001 / 0	0.443 / 69.60	0.451 / 68.10	0.342 / 68.80	0.408 / 68.40	**0.004 / 31.40**	**0.004 / 31.70**
	CLIP-Auto	**0.001 / 62.90**	0.0003 / 0	0.001 / 0	0.001 / 0	**0.447 / 69.80**	**0.494 / 69.00**	**0.361 / 69.20**	**0.404 / 69.50**	0.006 / 30.40	0.006 / 31.40

Table 7. Model results against individual untargeted adversarial attacks under the l_∞ norm on CIFAR-10. Each entry consists of the median l_∞ distance of the minimum adversarial perturbations over all samples on the left, and the model accuracy for the perturbation budget $\epsilon = 8/255$ on the right.

B.2. Adversarial Attacks

We briefly provide some additional numerical details of the zero-shot CLIP models under adversarial attacks.

Common Attacks In Figure 8, we see that all CLIP models obtain comparable average effective robustness with the standard models on both ImageNet and CIFAR-10. However, we see little evidence of relative robustness improvements. In particular, we find that the robustness drop on ImageNet is more significant than on CIFAR-10. The reason is that the gains of the standard models are mainly from the similar distribution in ImageNet. Note that on CIFAR-10, we use CLIP in a zero-shot manner, and use CLIP-Auto and linear-probe standard models for comparison. In Table 7, we show detailed results under each adversarial attack on CIFAR-10. We also show the median distance of the minimum adversarial perturbations. We find that the CLIP is more vulnerable under transfer-based attacks and black-box attacks than standard models. CLIP-Auto does not make much difference in the performance compared to CLIP.

Typographic Attacks CLIP consists of multimodal neurons which respond to both images and text for a given concept. Therefore this particular type of attack [14] is designed for zero-shot CLIP models. As all the CLIP models are built based on classifiers from the text, CLIP can be very vulnerable to such attacks. In Figure 5, we compare the model accuracies on our new robustness datasets: ImageNet-T and CIFAR-10-T. Unsurprisingly, all CLIP models are vulnera-

ble to typographic attacks and cause significant robustness drop compared to standard models. We find that typographic attacks reduce both the ImageNet in-distribution and out-of-distribution (ImageNet-T and CIFAR-10-T) performance by a large amount compared to the standard models. We show that the success rate is also much higher for the CLIP models. CLIP-Auto models are not able to significantly improve the robustness of the CLIP models. More details of ImageNet-T and CIFAR-10-T are described in Appendix C.

C. The ImageNet-T and CIFAR-10-T Robustness Test Sets

We use typographic attacks to create ImageNet-T and CIFAR-10-T.

ImageNet-T Design We aim to quantitatively evaluate the image classification robustness under the typographic attacks [14]. Attacks are automatically generated using the same (arbitrarily chosen) eight coordinates and using a consistent font style. As the setup is the targeted attack, we consider an attack to have succeeded if the predicted class is changed to the attack class. The attack text for each image is a target label text uniformly chosen over other classes except its true class at random. We use OpenCV[3] to attach the attack text to the images. The dataset contains 50,000 images, which equals the size of the ImageNet validation set. As documented in [14], the idea of typographic attacks is in general similar to work such as adversarial patches [4] and physical adversarial examples [2]. We plan to investigate more attacks along this line as one of the future directions.

CIFAR-10-T Design We aim to provide a small typographic attacks-based test set for quicker experimentation. The only difference from the ImageNet-T is that we use four coordinates instead of eight due to the lower resolution of the images in CIFAR-10. This results in 10,000 images in CIFAR-10-T, which is the size of the CIFAR-10 test set.

D. Data Overlap Examples

We provide similar image examples found by Google Images in Figure 10.

E. CLIP-Auto Details

Automated Prompt Generation We describe our method to generate automated prompts in detail. Formally, the goal is to learn a prompt $\mathbf{x}_{\text{prompt}} = z(\mathbf{x}_{\text{trig}}, \mathbf{x}_{\text{label}})$, where z is the template such as "`[T][T][T][T][C]`". `[T]` indicates a trigger token, and `[C]` indicates the label text. The idea is to add a set of trigger tokens (i.e., `[T]`) to the label text according to the template. The process is framed as a prompt search task. As shown in Figure 9, the trigger tokens are initialized as "A photo of a", then iteratively updated to minimize the classification loss over batches of training examples. At each search step, the change of loss corresponding to the replacement of a trigger token with another token in the vocabulary is computed by a first-order Taylor approximation [51]. For each trigger token, we keep the top-k candidate trigger tokens that lead to the smallest loss, formally: $\mathcal{T}_{cand} = \text{top-}k_{t \in \mathcal{V}}[\mathcal{L} - \nabla_{\mathbf{t}}\mathcal{L}]$. $\mathcal{L}$ is the loss, t is a candidate token from the vocabulary $\mathcal{V}$, $\mathbf{t}$ is the corresponding input embedding. We evaluate the updated prompt at the current step and retain the prompt with the highest probability in the next step. In practice, we perform a left-to-right beam search over the top-k candidate trigger tokens using the candidate sequences with the smallest loss at the current step. We use small beam sizes for efficiency consideration. For example, the trigger tokens converge to "lovely picture show the", which is used in the next step. The final prompt from the last step is returned. We use the ImageNet training set to find the prompts.

Rather than keeping the best candidate sequence of trigger tokens at each step as in the above setting, the best candidate from each step is concatenated and deduplicated. We select n candidate sequences that lead to the best performance on a validation set sampled from ImageNet containing 10,000 images. Instead of ensembling over the probability space of multiple classifiers, we follow [37] to ensemble over the embedding space of the text. We use this as the default setting for CLIP-Auto.

To keep the training as simple as possible, we use the same hyperparameters to search the automated prompts for all CLIP-Auto models. We use: top-k equals 20 of the candidate trigger tokens; beam size equals 5 according to [51]; 1 GeForce GTX 1080 with a batch size of 512; the number of training steps equals 2000; each step takes approximately 3 minutes. We generate all CLIP-Auto models on the ImageNet training set and randomly choose 1,000 samples from the validation set to validate the best number of prompts. The number of automated prompts is: 49, 8, 186, 84, 7, 22 for ResNet50, ResNet101, ViT-B/32, ViT-B/16, ResNet50x4, and ResNet50x16 respectively.

Case Study We show uncurated full sets of prompts for CLIP[4] and CLIP-Auto in Table 8 to Table 14. Automated prompts for each model are produced based on the same prompt template "`[T][T][T][T][C].`". We find that although the automated prompts are less interpretable, they are able to construct classifiers that predict comparably with the manual prompts, thanks to its flexibility in deriving customized prompts for each model. However, the impact of

[3] https://opencv.org/

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

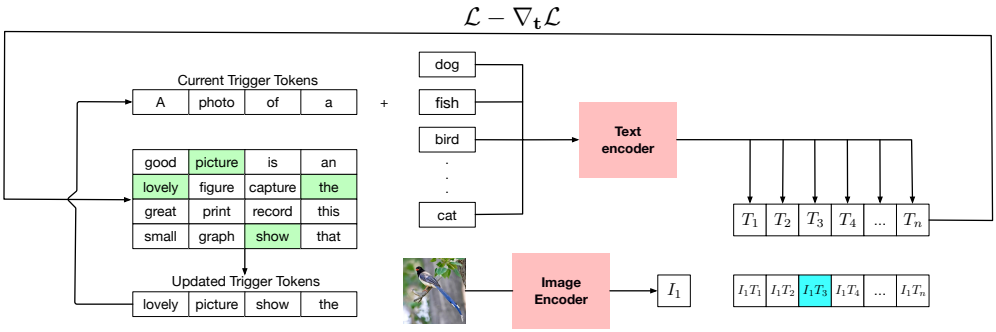
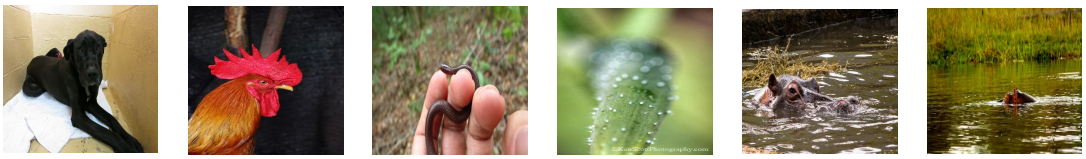


Figure 9. Summary of producing an automated prompt for CLIP-Auto. A prompt is created using a template that combines a set of trigger tokens with the label text. The trigger tokens are shared across all classes and decided via a gradient-based search. At each search step, we use the current trigger tokens and each label text to synthesize a linear classifier, then compute the gradients for candidate trigger tokens and update the trigger tokens with that lead to the smallest loss. After iteratively repeating this process, the trigger tokens converge and are returned to create a prompt.



(a) ImageNetV2.

(b) Google Images.

Figure 10. Data overlap examples. We use Google Images to find similar or same images in ImageNetV2.

prompts synthesizing the classifiers is limited by the pre-trained image representations. Therefore we do not see a significant difference in the robustness performance.

F. Additional Related Work

Natural language prompt engineering is useful for zero (or few)-shot performance in NLP, such as large language models [52, 53] including GPT-3 [5], GPT-4 [33], and Gemini [47]. Especially, the manually created prompts [35] are used for fact retrieval. While the primary usage of the natural language prompts aims to improve the performance of NLP tasks, we aim to use the natural language prompts to improve the robustness of performance in other domains, e.g., image classification. The prompt learning methods [56, 57] employ the continuous prompt method, which is less explainable. As concluded in Sec. 4, the dominant factor is the pre-trained image representation. Therefore, similar to our method, these prompt learning approaches mainly improve the efficiency of training (avoiding updating all model parameters) instead of robustness enhancement.

a bad photo of a {label}	the {label} in a video game	the plushie {label}
a photo of many {label}	a sketch of a {label}	a dark photo of a {label}
a sculpture of a {label}	a doodle of the {label}	itap of a {label}
a photo of the hard to see {label}	a origami {label}	graffiti of the {label}
a low resolution photo of the {label}	a low resolution photo of a {label}	a toy {label}
a rendering of a {label}	the toy {label}	itap of my {label}
graffiti of a {label}	a rendition of the {label}	a photo of a cool {label}
a bad photo of the {label}	a photo of the clean {label}	a photo of a small {label}
a cropped photo of the {label}	a photo of a large {label}	a tattoo of the {label}
a tattoo of a {label}	a rendition of a {label}	a photo of the dirty {label}
the embroidered {label}	a photo of a nice {label}	a jpeg corrupted photo of a {label}
a photo of a hard to see {label}	a photo of a weird {label}	a blurry photo of the {label}
a bright photo of a {label}	a blurry photo of a {label}	a photo of the {label}
a photo of a clean {label}	a cartoon {label}	a good photo of the {label}
a photo of a dirty {label}	art of a {label}	a rendering of the {label}
a dark photo of the {label}	a sketch of the {label}	a {label} in a video game
a drawing of a {label}	a embroidered {label}	a photo of one {label}
a photo of my {label}	a pixelated photo of a {label}	a doodle of a {label}
the plastic {label}	itap of the {label}	a close-up photo of the {label}
a photo of the cool {label}	a jpeg corrupted photo of the {label}	a photo of a {label}
a close-up photo of a {label}	a good photo of a {label}	the origami {label}
a black and white photo of the {label}	a plushie {label}	a painting of the {label}
a plastic {label}	a photo of the nice {label}	a painting of a {label}
a photo of the small {label}	the cartoon {label}	a pixelated photo of the {label}
a photo of the weird {label}	art of the {label}	a sculpture of the {label}
a bright photo of the {label}	a drawing of the {label}	a photo of the large {label}
a cropped photo of a {label}	a black and white photo of a {label}	
a photo of the large {label}		

Table 8. Uncurated prompts of CLIP.

bri dexter suppose an {label}	annually ates] interesting {label}	potd enjo about an {label}
amazingly wri means an {label}	amazingly fascin about an {label}	among talking about an {label}
annually allegedly] a {label}	instead cro mesmerizing an {label}	amazingly uni tically an {label}
instead :] frifotos an {label}	amazingly sth discover an {label}	ably prett mous an {label}
atively factfriday !.. an {label}	(talking about an {label}	incredibly fascin about an {label}
instead aper frifotos an {label}	affection randomly about an {label}	singul thing factfriday a {label}
also .# pics numerous {label}	complete fascin about an {label}	ingh random)... an {label}
atively factfriday an {label}	also goog acquainted an {label}	commonly atio]: a {label}
ever cheerful about an {label}	ever fascin about an {label}	atively] awesome {label}
also intre acquainted an {label}	although photo of a {label}	hetero tional " beautiful {label}
ij favourites factfriday a {label}	rarely hein suppose an {label}	tious query !). googled {label}
unpopular typically] a {label}	ever genus admire an {label}	habit rained !). googled {label}
annually ates] typical {label}	rarely exc suppose an {label}	besides only !). an {label}
annually ates] this {label}	esper rained !). googled {label}	ever thing interesting an {label}
bi ant thousands interesting {label}	crazy factfriday .) an {label}	wonderfully kan)... an {label}
instead continuation frifotos an {label}	rarely easter suppose an {label}	
instead glorious frifotos an {label}		
esper sees !). googled {label}		

Table 9. Uncurated prompts of CLIP-Auto of ResNet50.

awesome coo ing led {label}	seated photo of a {label}	lovel ...: share interesting {label}
and awesome ==> a {label}	sson photo of a {label}	hamp ...: share famous {label}
classi ele primarily smaller {label}	relatively pictures ! a {label}	classi inged purposes smaller {label}
awesome hob ing led {label}	sively awesome ! a {label}	dana ...: ordinary large {label}
awesome wonderful thero led {label}	exceptionally awesome ! a {label}	...# ...: attractive large {label}
brightly ative - like {label}	potentially awesome ! a {label}	awesome lovely ca eled {label}
jou photo of a {label}	things awesome !). a {label}	awesome lovely ca oldest {label}
blu related typical smaller {label}	funfactfriday hooray !). a {label}	awesome lovely ini oldest {label}
brightly ative spo old {label}	or awesome !). a {label}	awesome gently be oldest {label}
awesome very ver oldest {label}	aside funfactfriday !). a {label}	awesome seemingly ver oldest {label}
super awesome neat a {label}	omfg funfactfriday !). a {label}	awesome neat ver oldest {label}
awesome lovely ca fiable {label}	just awesome neat a {label}	awesome european an nicest {label}
especially also !). smaller {label}	supposedly awesome) a {label}	awesome bic an nicest {label}
contributed potty ==> ordinary {label}	kidding awesome) a {label}	awesome impressive an yed {label}
.....# trivia !: smaller {label}	controversial awesome) a {label}	awesome cro an other {label}
potentially awesome - a {label}	funfactfriday awesome : a {label}	awesome cro ste other {label}
funfactfriday awesome ==> a {label}	obligatory awesome ==> a {label}	awesome hob uni led {label}
contributed ==> typical {label}	contributed ==> large {label}	amazingly hob ing led {label}
==> large {label}	contributed ==> which {label}	awesome delightful ing led {label}
unto picoftheday !). smaller {label}	contributed ==> unwanted {label}	awesome cool ing led {label}
awesome (!) () ous {label}	contributed ==> empty {label}	awesome wonderful thero important {label}
relatively awesome ! a {label}	contributed potty ==> empty {label}	awesome funfactfriday actual ous {label}
potentially awesome !). a {label}	contributed potty ==> wonderful {label}	awesome funfactfriday absolutely ous {label}
,# awesome neat a {label}	contributed ==> those {label}	small (!) some huge {label}
contributed ==> ordinary {label}	odd inspired ==> typical {label}	small ulously ǀ huge {label}
.....# trivia .. smaller {label}	presents inspired ==> typical {label}	small ative ǀ huge {label}
theworld ness !: smaller {label}	————————- inspired ==> typical {label}	brightly ative - smelly {label}
: ...: \' interesting {label}	————————- ij ==> smaller {label}	thest ative - gorgeous {label}
supposedly hob ing led {label}	————————- ij favourite smaller {label}	several ative - esque {label}
awesome gin ing led {label}	neh informative favourite smaller {label}	large ative - esque {label}
large two - covered {label}	pas related peoples smaller {label}	large single - esque {label}
exac latin - covered {label}	: related typical smaller {label}	large pra - covered {label}
funfactfriday omg !). a {label}	: related) smaller {label}	distinctive backward - covered {label}
quiz namesake) smaller {label}	highly namesake) smaller {label}	registered pee - covered {label}
commonly someday !). smaller {label}	often namesake) smaller {label}	commonly reasonable - covered {label}
of picoftheday .. smaller {label}	often similar !). smaller {label}	registered latin - covered {label}
awesome ...: approximately smaller {label}	often separately !). smaller {label}	and latin - covered {label}
awesome :- ca formed {label}	ima separately !). smaller {label}	exce dual - covered {label}
awesome neat an ering {label}	formally same !). smaller {label}	common continuous - covered {label}
awesome dy an ering {label}	commonly holidays !). smaller {label}	common µ - sized {label}
awesome wonderful ing led {label}	related – !). smaller {label}	common ities - sized {label}
awesome (!) wonderful ous {label}	grand – !). smaller {label}	recre ities - sized {label}
larger ative - like {label}	better – !). smaller {label}	recre wanna - sized {label}
seriously photos of a {label}	previous – !). smaller {label}	awesome macro ca fiable {label}
relatively photos of a {label}	taller lished !). smaller {label}	awesome hob tr led {label}
ously awesome ! a {label}	taller spoiled !). smaller {label}	awesome favourite ing led {label}
funfactfriday awesome !). a {label}	taller challenged !). smaller {label}	thest ative - like {label}
amazingly awesome neat a {label}	preten broader !). smaller {label}	ari ly - covered {label}
smaller awesome neat a {label}	preten later !). smaller {label}	common tively - covered {label}
contributed ==> favorite {label}	photo picoftheday !). smaller {label}	:@ ...: ¥ interesting {label}
contributed ==> ordinary {label}	of picoftheday smaller {label}	share ...: ¥ interesting {label}
contributed ==> nice {label}	_____ random .. smaller {label}	theo ...: / interesting {label}
contributed potty ==> those {label}	.....# trivia toftheday smaller {label}	aerop ...: share interesting {label}
naturally related) smaller {label}	.....# fascin !: smaller {label}	hamp ...: share large {label}
of picoftheday !). smaller {label}	!) fascin !: smaller {label}	...: share large {label} ...: share japanese {label}
theworld fascin !: smaller {label}	classi symbolic !: smaller {label}	awesome ...: out smaller {label}
theworld indication !: smaller {label}	classi introduced !: smaller {label}	awesome ...: largely smaller {label}
classi closely !: smaller {label}	awww impressive ca fiable {label}	awesome ...: ca other {label}
...: interesting {label}		awesome ...: ca formed {label}
awesome ...: share smaller {label}		awesome thero ca funded {label}
		beautiful Ê ca funded {label}
		gorgeous Ê ca funded {label}
		gorgeous ca funded {label}
		...# ...: awesome interesting {label}
		awesome Ê ca funded {label}
		...# ...: noisy interesting {label}
		...# ...: ordinary large {label}

Table 10. Uncurated prompts of CLIP-Auto of ViT-B/32.

various unlike ": a {label}
til photo – a {label}
unpopular introduce .âĢ¦ a {label}
completely familiar recognizable wonderful {label}
til correctly – a {label}
many other recognizable an {label}
adop circulating recognizable entire {label}
awesome orient ce sized {label}

Table 11. Uncurated prompts of CLIP-Auto of ResNet101.

wednesday goo explanations an {label}
sforsale fineartamerica nbd about {label}
neva awesome didyouknow a {label}
gee friend didyouknow a {label}
aun holidays didyouknow a {label}
kic holidays didyouknow a {label}
eco coalition – delightful {label}

Table 12. Uncurated prompts of CLIP-Auto of ResNet50x4.

then coolest)... a {label}	thatsmoly :) a {label}	deeply bilateral inviting a {label}
because unusual)... a {label}	thatsjust :) a {label}	recognizable particularly ugliest a {label}
neat awsome ..... a {label}	because neat)... a {label}	arbitractual oooooo a {label}
heavily etzinteresting a {label}	then craziest)... a {label}	incredibly actual oooooo a {label}
what awsome – a {label}	then fascinating :" a {label}	incredibly unusual a {label}
considered awesome :) a {label}	)... awesomrecognizable a {label}	incredibly behold a {label}
awesomeness !),)... a {label}	awesome wonderfully)(a {label}	incredibly behold :-) a {label}
.....awsome – a {label}	longest etzinteresting a {label}	)... behold :-) a {label}
awesomeness perhaps)... a {label}	neat :)) amazing a {label}	)... behold awesomeness a {label}
totally !),)... a {label}	awesomeness fer :) a {label}	)... behold exciting a {label}
a typical inviting a {label}	what awsome)... a {label}	ably awesomfascinating a {label}
ably awesom... a {label}	fascinating snazzy).. a {label}	ably awesominstance a {label}
wonderful vely interesting a {label}	supposedly awesome :) a {label}	cool wonderfully)(a {label}
fascinating renowned :) a {label}	fascinating ilove :) a {label}	instantly awsome admire a {label}
how awsome) a {label}	thatswanted :) a {label}	thus awsome admire a {label}
randomly !),)... a {label}	considered fascinating :) this {label}	incredibly awsome) a {label}
because fascinating)... a {label}	awesomeness fascinating :) a {label}	incredibly unusual oooooo a {label}
then awesomeness)... a {label}	fascinating jst ,) a {label}	wonderful etzinteresting a {label}
wonderful fortunately)... a {label}	because awesome)... a {label}	coolest justsaying :). a {label}
precisely awesomeness :" a {label}	because fyi)... a {label}	awesomeaping :) a {label}
a displainviting a {label}	then fancy)... a {label}	looodesperately admire a {label}
ably awesomsure a {label}	individually awesomeness :" a {label}	thus awsome beautiful a {label}
what awsome » a {label}	a photo of a {label}	actually awsome – a {label}
fairly interesting !!!!!! a {label}	a frontal bestoa {label}	fairly interesting).. a {label}
because hooray)... a {label}	a rentbestoa {label}	wow awesomeness why a {label}
)... awesomeness :) a {label}	a industrialbestoa {label}	a demoinviting a {label}
arbitrparticularly gosh a {label}	a typical bestoa {label}	the excinviting a {label}
		compare recognizable inviting a {label}
		because awesomeness)... a {label}
		wonderful ! a {label}

Table 13. Uncurated prompts of CLIP-Auto of ViT-B/16.

enigffect found '# {label}	whoa kineusing a {label}	umtom word : {label}
#:'topics another {label}	expectthageneric impressive {label}	undant j) terrifying {label}
tessenulogically cool {label}	pak,... random impressive {label}	those kalically amazing {label}
enigffect found '# {label}	dingh?) lovely {label}	spare sovereignically actual {label}
#:'topics another {label}	quite famili...!!! sized {label}	everydayphoto unmatched a {label}
tessenulogically cool {label}	inely similarities – sized {label}	oooo– coolest a {label}
small freaking spegir {label}	archelic :-) colorful {label}	pretty "# coolest a {label}
	yay perfectly !), amazing {label}	simultaneously almost coolest a {label}
	____ fact gratextinct {label}	weird ´ radinvolving {label}

Table 14. Uncurated prompts of CLIP-Auto of ResNet50x16.