Robust Cross-Modal Representation Learning with Progressive Self-Distillation Supplementary Materials

Alex Andonian[*]
MIT CSAIL
andonian@mit.edu

Shixing Chen, Raffay Hamid
Amazon Prime Video
{shixic, raffay}@amazon.com

1. Summary of Datasets

1.1. Pretraining Datasets

Table 1 lists the exact number of examples used for pretraining from each dataset. During COCO pretraining, we randomly select one of five unique captions assigned to each image, effectively multiplying the total number of image-text pairs by a factor of 5. Unlike COCO, the Conceptual Captions datasets are not stored in one central location and not all of the provided URLs are still valid. Therefore, the actual number of pretraining examples is less than the advertised amount by as much as 1M+ in the case of CC12M.

Dataset	COCO	CC3M	CC12M
#	118,287	2,884,940	10,707,814

Table 1. **Pretraining Dataset Sizes –** Exact sizes of pretraining datasets employed in this work (no. of image-text pairs).

1.2. Evaluation Datasets

In Table 2, we list additional details for the evaluation datasets in this study including the number of classes and the sizes of the training-testing splits. For the last 5 rows (ObjectNet and ImageNet-{A,O,R,V2}), we list only the number of classes and the testing set size as these have been designated as "testing-only" datasets.

Dataset	Classes	Train Size	Test Size
Cifar10	10	50,000	10,000
Cifar100	100	50,000	10,000
Caltech101	102	2,863	8,677
Food101	101	75,750	25,250
OxfordPets	37	3,680	3,669
Birdsnap	500	38,344	1,900
ImageNet	1000	1,281,167	50,000
Places365	365	1,803,460	36,489
ObjectNet	313	-	50,000
ImageNet-A	200	-	7,500
ImageNet-O	200	-	2,000
ImageNet-R	200	-	30,000
ImageNetV2	1000	-	30,000

Table 2. **Evaluation Dataset Details –** The number of classes, training and testing examples present in the evaluation datasets. The size of the training set size for the last 5 rows is omitted because these datasets are designated as "testing-only" benchmarks.

2. Detailed Experimental Settings

2.1. Implementation Details

Our pretraining implementation largely follows CLIP [8] with significant deviations motivated by computational constraints or empirical observations. Table 3 summarizes common hyperparameters settings shared across experiments. Notable differences from [8] include a reduced batch size, learning rate and weight decay, but increased number of training epochs and warm-up iterations. Unlike CLIP, which computes sharded, intra-GPU embedding similarities only, we perform a global all-gather operation to compute all pair-wise similarities within a batch.

Hyperparameter	Value
Batch size	4096
Vocabulary size	49408
Training epochs	100
Initial temperature τ	0.07
Teacher temperature $\tilde{\tau}$	0.1
Weight decay	0.001
Warm-up iterations (%)	0.2
Learning rate	1×10^{-5}
Adam β_1	0.9
Adam β_2	0.99
Adam ϵ	10^{-5}

Table 3. **Common hyperparameters –** used for both baseline CLIP pretraining and our method.

[*] This work was done when the author was an intern at Amazon.

2.2. Complexity Analysis

In order to underscore how our method improves data-efficiency without incurring additional computation cost, we note that the number of trainable parameters in the ViT-B/32 (151.3M) and RN50 (102.0M) CLIP models are precisely the same under our method as the architectures remain unmodified with no additional parameters required. Also, the average time per training iteration (*e.g.,* 0.253s for our COCO experiments using 8 Nvidia V100 GPUs) and memory requirements are effectively identical as well. Finally, we contrast our self-distillation approach to traditional knowledge-distillation methods and other choices of teacher network (*e.g.,* a momentum teacher design), which may provide similar data-efficiency improvements, but incur significant computational costs of at least 2x compute and memory requirements or more.

2.3. Prompt Engineering

Table 4 lists the prompt templates used for zero-shot classification on the evaluation datasets. For each template, string interpolation replaces the placeholder symbol {} with a text representation of the category name, and a grammatical correction is applied to the preceding article, i.e., a → an for categories that start with a vowel.

2.4. Linear Probe Details

The linear probe evaluation involves training a logistic regression classifier on the frozen visual features extracted using the model's image encoder. Following CLIP [8], we train the logistic regression for a maximum of 1000 iterations using the L-BFGS optimization algorithm provided by scikit-learn [7]. We use the train/test split sizes listed in Table 2.

3. Comparison to Concurrent Works

In this section, we provide a comparison of our work to unpublished concurrent works.

While the concurrent works discussed in this section are aimed at improving on CLIP, they altogether vary along several dimensions including pretraining dataset, backbone architecture, hyperparameters and training details. Due to the large number of unique experimental configurations, it is not feasible to precisely replicate the setting of each concurrent work; therefore, we provide our most closely matched experiments while acknowledging that the comparisons are not exactly one-to-one analogs.

Table 5 lists the commonly reported ImageNet zero-shot Top 1 accuracy achieved by concurrent methods aimed at reproducing CLIP and/or addressing its limitations. Each work additionally provides a re-implementation of CLIP (listed as CLIP impl. followed by a citation). Even for architecture and dataset matched experiments, the difference in accuracy can differ by as much as 4.8% (compare [6] and [8]), further highlighting the challenges of providing meaningful comparison. From the perspective of raw performance, our method achieves the highest absolute top 1 accuracy on zero-shot ImageNet classification out of all approaches and architectures that use at most 15M pretraining examples.

Dataset	Method	Architecture	ImageNet
YFCC (15M)	CLIP impl. [9]	ViT-B/32	30.4
	FILIP [9]	ViT-B/32	37.8
	CLOOB [3]	RN50	35.7
	CLOOB [3]	RN101	37.1
	CLOOB [3]	RN50x4	39.0
	CLIP impl. [6]	RN50	35.9
	DeCLIP [6]	RN50	41.9
	OpenAI CLIP [8]	RN50	31.3
	OpenCLIP [4]	RN50	32.7
	OpenCLIP [4]	RN101	34.8
CC3M	OpenCLIP [4]	RN50x4	22.2
	CLIP impl. [3]	RN50	23.9
	CLOOB [3]	RN50	25.6
	DeCLIP [6]	RN50	27.8
	CLIP impl. (ours)	ViT-B/32	23.5
	Our method	ViT-B/32	**28.0**
CC12M	CLIP impl. (ours)	ViT-B/32	37.8
	DeCLIP [6]	RN50	41.0
	Our method	ViT-B/32	**42.2**

Table 5. **Zero-Shot ImageNet Classification Comparison –** Zero-shot Top1 accuracy (%) of our method compared to concurrent works.

4. Additional Quantitative Results

4.1. ResNet50 Visual Backbone Results

In order to demonstrate the effectiveness of our method to a broader experimental setting, we present results from experiments that utilize the widely adopted ResNet50 (RN50) visual backbone in Table 6. These results are highly consistent with performance trends observed with the ViT backbone (even with minimal hyperparameter tuning), suggesting that our method is amenable to CNN-based backbones as well.

4.2. Ablation Studies for α-scheduling

Our method is fairly robust to the choice of α scheduling. For example, replacing cosine annealing with a simple linear schedule produces a network with near identical performance on downstream evaluations. Specifically, for COCO the two scheduling methods result in an absolute mean difference in top1 classification of 0.13% and relative difference of 0.67% for our method using a ViT backbone. Similarly, for CC3M we observe an absolute mean difference of -0.22% and relative difference of -0.63%.

Cifar{10,100}	**Caltech101**	**ImageNet+**
”a photo of a {}.”, ”a blurry photo of a {}.”, ”a black and white photo of a {}.”, ”a low contrast photo of a {}.”, ”a high contrast photo of a {}.”, ”a bad photo of a {}.”, ”a good photo of a {}.”, ”a photo of a small {}.”, ”a photo of a big {}.”, ”a photo of the {}.”, ”a blurry photo of the {}.”, ”a black and white photo of the {}.”, ”a low contrast photo of the {}.”, ”a high contrast photo of the {}.”, ”a bad photo of the {}.”, ”a good photo of the {}.”, ”a photo of the small {}.”, ”a photo of the big {}.”	“a photo of a {}.”, “a painting of a {}.”, “a plastic {}.”, “a sculpture of a {}.”, “a sketch of a {}.”, “a tattoo of a {}.”, “a toy {}.”, “a rendition of a {}.”, “a embroidered {}.”, “a cartoon {}.”, “a {} in a video game.”, “a plushie {}.”, “a origami {}.”, “art of a {}.”, “graffiti of a {}.”, “a drawing of a {}.”, “a doodle of a {}.”, “a photo of the {}.”, “a painting of the {}.”, “the plastic {}.”, “a sculpture of the {}.”, “a sketch of the {}.”, “a tattoo of the {}.”, “the toy {}.”, “a rendition of the {}.”, “the embroidered {}.”, “the cartoon {}.”, “the {} in a video game.”, “the plushie {}.”, “the origami {}.”, “art of the {}.”, “graffiti of the {}.”, “a drawing of the {}.”, “a doodle of the {}.”	“{}” “A photo of {}” “A photo the {}” “itap of a {}.” “a bad photo of the {}.” “a origami {}.” “a photo of the large {}.” “a {} in a video game.” “art of the {}.” “a photo of the small {}.”

Table 4. **Prompt templates for zero-shot evaluation –** The placeholder symbol {} is replaced with a string representation of the category name. The last column “ImageNet+” corresponds to the templates used for all other datasets that appear in this work, including all of the ImageNet variants.

Pretraining Dataset	Method	*Cifar10*	*Cifar100*	*Caltech101*	*Places365*	*ObjectNet*	*ImageNet-R*	*ImageNet-O*	*Imagenet-A*	*ImageNetV2*	*ImageNet*	*Average*
COCO	CLIP	31.12	7.66	27.01	10.56	2.64	5.77	8.9	2.31	5.92	6.59	10.84
	Ours	**37.85**	**9.97**	**33.87**	**11.17**	**3.38**	**6.48**	**11.7**	**2.56**	**6.87**	**7.48**	**13.12**$^{+2.27}$
CC3M	CLIP	36.46	14.20	44.81	19.29	3.24	19.28	24.50	3.53	16.95	18.12	20.04
	Ours	**56.69**	**24.44**	**61.93**	**24.67**	**5.27**	**25.96**	**31.65**	**4.57**	**21.65**	**22.918**	**27.98**$^{+7.94}$

Table 6. **Zero-Shot Image Classification Comparison with RN50 backbone –** Zero-shot Top1 accuracy (%) of our method compared to baseline CLIP on numerous downstream benchmark datasets. Note: Results are derived from published hyperparameters for the baseline and minimal hyperparameter tuning of our method to account for differences in backbone architecture.

5. Additional Qualitative Results

5.1. Additional Text-Image Retrievals

In Figures 1 and 2, we provide additional text-image retrieval results computed on the COCO test set. Consistent with the retrievals shown in Figure 6 from our main paper, our method more consistently captures the full semantic extent of the query caption, whereas the baseline CLIP model tends to narrowly focus on one particular aspect. For example, in the first row of Figure 1, CLIP does retrieve a “black and white cat,” but also retrieves a black and white dog, a black and white zebra and a black and white photograph of boxes, whereas our method retrieves images containing a cat in 9 out of the 10 top retrievals.

5.2. Additional Robust Classification Examples

In Figure 3, we show qualitative examples of the instances where our method shows improved robustness over its CLIP counterpart for out-of-distribution images drawn from the ImageNet-R dataset. While CLIP struggles to handle certain artistic styles (angular shark depicted in row 1, goldfinch in row 4, tattooed tree frog in row 6), strong color patterns (black and white colors in row 3), or subjects in unusual contexts (rows 5,7,8), our method is able to more consistently provide a reasonable set of top predictions, which is consistent with the quantitative improvement of nearly 8.5% on average when considering Conceptual Captions pretraining as reported in Table 1 of our main paper.

5.3. Similarity Matrix Visualization

In Figure 4, we show a matrix of pairwise cosine similarity scores assigned to a batch of images and corresponding text snippets by our method compared to its CLIP counterpart. These similarity matrices present concrete examples of the trends captured by the distributions shown in Figure 5 of the main paper. Namely, it shows that CLIP has been optimized to assign high similarity scores along the diagonal (positive pairs) and low similarity to off diagonal elements (negatives), even when there is non-negligible semantic similarity between unpaired instances (*e.g.,* the text “a black-and-white silhouette...” and the black-and-white image of a photographer dressed in black clothing) . In contrast, our method yields elevated scores for negative pairings that show this amount of secondary similarity. As a consequence, we empirically observe that our learned representations produce larger scores for ground truth positive pairs and lead to more robust zero-shot classification performance.

6. Ethical Considerations

6.1. Impact on ML and Related Scientific Fields

A primary motivation driving this work is to increase the robustness and efficiency of a vision-language pretraining (VLP) method that has recently given rise to a set of so-called *foundation models* [1]. Due to the anticipated role that foundation models are to play in the immediate de-

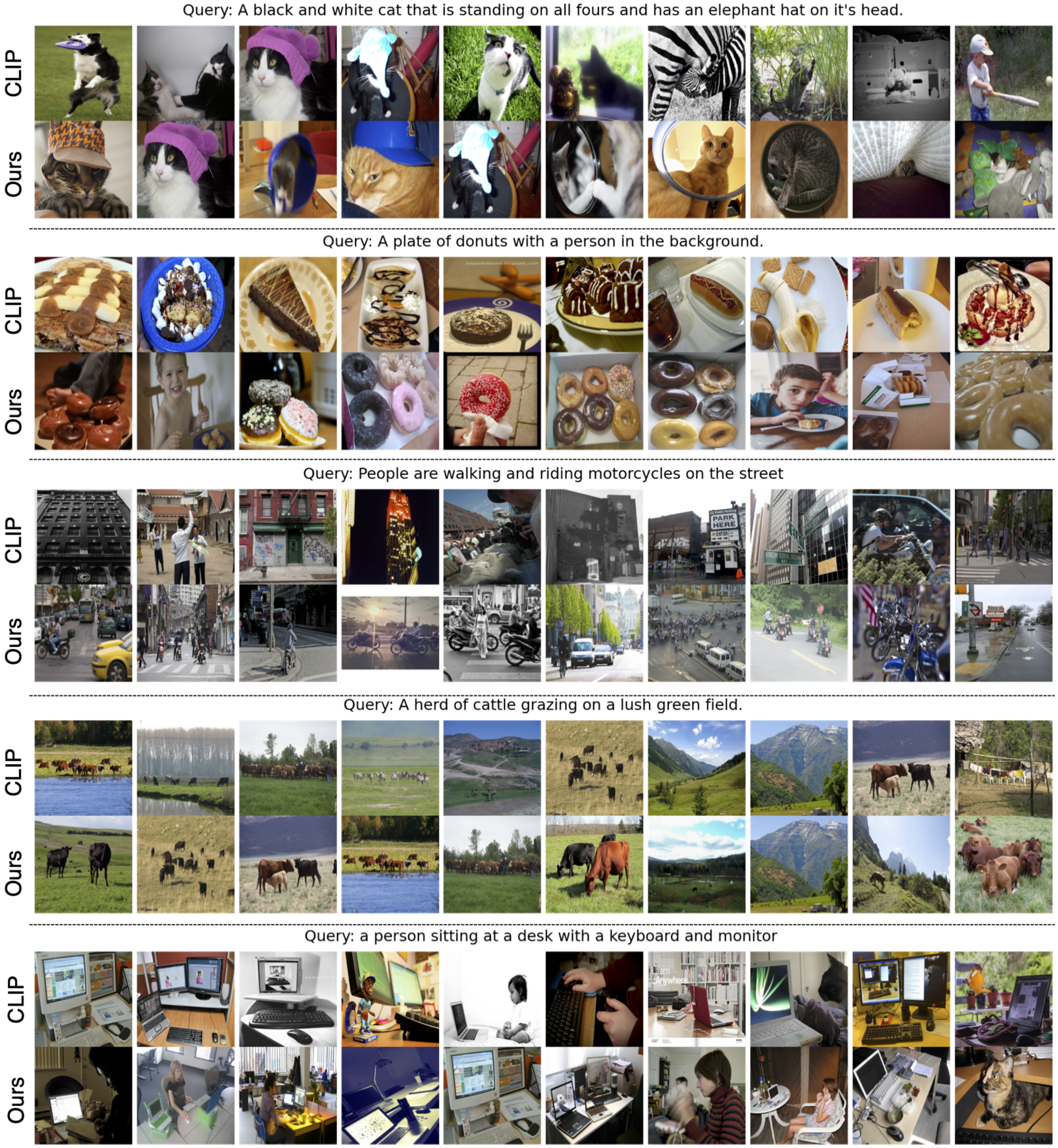

Query: A black and white cat that is standing on all fours and has an elephant hat on it's head.
CLIP
Ours
Query: A plate of donuts with a person in the background.
CLIP
Ours
Query: People are walking and riding motorcycles on the street
CLIP
Ours
Query: A herd of cattle grazing on a lush green field.
CLIP
Ours
Query: a person sitting at a desk with a keyboard and monitor
CLIP
Ours

Figure 1. **Example Text-Image Retrievals –** Given a text query, we display the top ten most semantically related images (ranked left to right) retrieved by CLIP and our method. Compared to CLIP, our method continues to retrieve images that more holistically match the text description, even after the ground truth image has appeared in the ranking.

velopment of AI systems, contributions to advancing the core training method will have far-reaching impacts on the field and downstream application areas by definition. Since improving the efficiency and lowering the computational/environmental cost associated with this VLP method is a primary objective of our work, we would like our work to assist in providing greater accessibility to the study, development and deployment of these VLP methods.

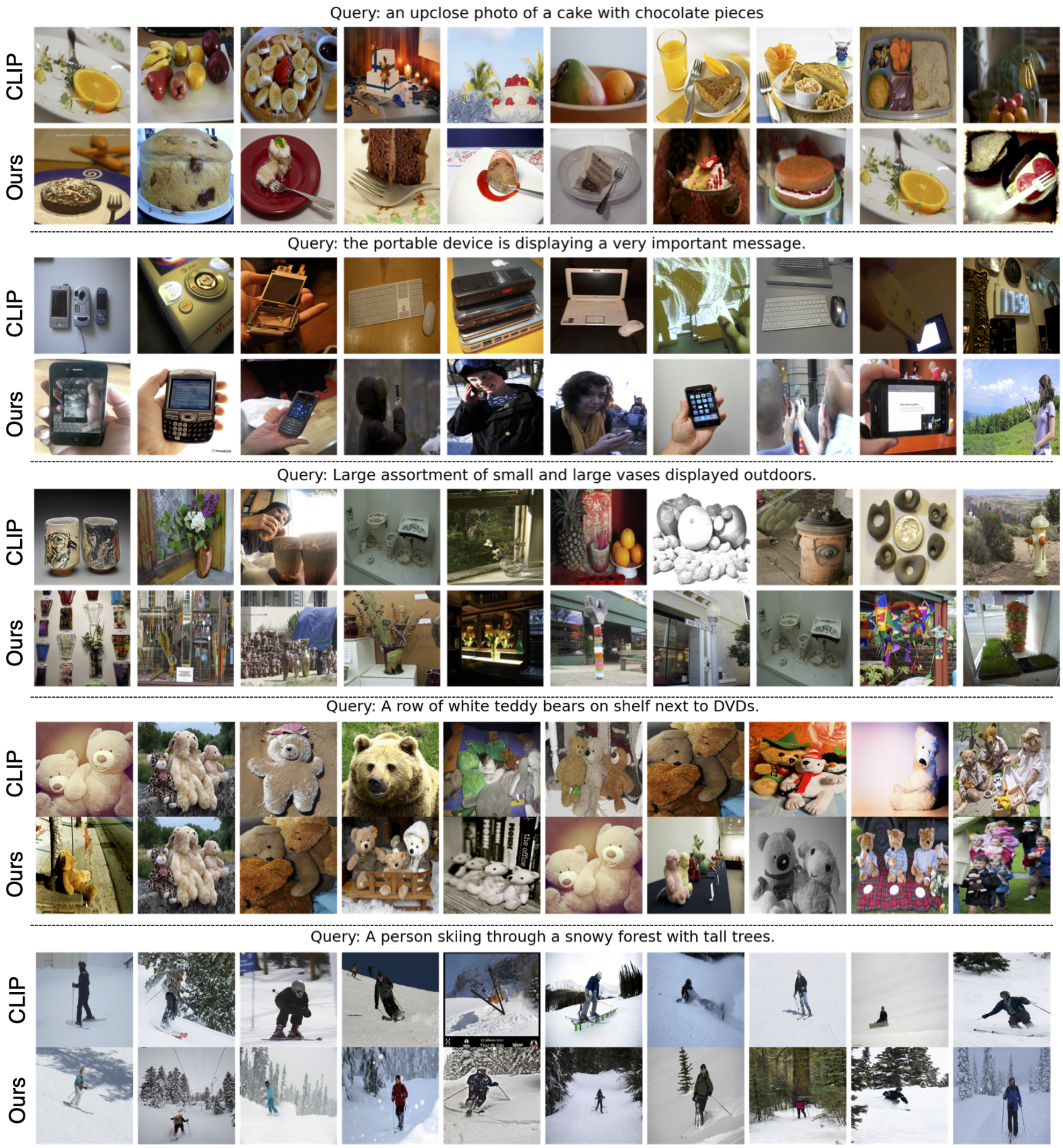

Query: an upclose photo of a cake with chocolate pieces
CLIP
Ours
Query: the portable device is displaying a very important message.
CLIP
Ours
Query: Large assortment of small and large vases displayed outdoors.
CLIP
Ours
Query: A row of white teddy bears on shelf next to DVDs.
CLIP
Ours
Query: A person skiing through a snowy forest with tall trees.
CLIP
Ours

Figure 2. **Additional Text-Image Retrievals –** Given a text query, we display the top ten most semantically related images (ranked left to right) retrieved by CLIP and our method. Compared to CLIP, our method continues to retrieve images that more holistically match the text description, even after the ground truth image has appeared in the ranking.

6.2. Impact on Society

Robustness to challenging, novel, and even adversarial examples is rapidly becoming an extremely important part of modern computer vision systems, which are now starting to be deployed in sensitive contexts such as autonomous vehicles [5] and medical applications [2] with life and death consequences. Additionally, the increasing diversity of data sources, ranging from massive and cumbersome datasets to extremely limited and highly sensitive information, poses several practical and environmental challenges to consistently training robust and reliable machine learning systems. Our proposed framework aims to address these aspects jointly.

References

[1] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 3

[2] Thomas Davenport and Ravi Kalakota. The potential for artificial intelligence in healthcare. *Future healthcare journal*, 6(2):94, 2019. 5

[3] Andreas Fürst, Elisabeth Rumetshofer, Viet Tran, Hubert Ramsauer, Fei Tang, Johannes Lehner, David Kreil, Michael Kopp, Günter Klambauer, Angela Bitto-Nemling, et al. Cloob: Modern hopfield networks with infoloob outperform clip. *arXiv preprint arXiv:2110.11316*, 2021. 2

[4] Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. If you use this software, please cite it as below. 2

[5] Shantanu Ingle and Madhuri Phute. Tesla autopilot: semi autonomous driving, an uptick for future autonomy. *International Research Journal of Engineering and Technology*, 3(9):369–372, 2016. 5

[6] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021. 2

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 2

[8] 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, 2

[9] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021. 2

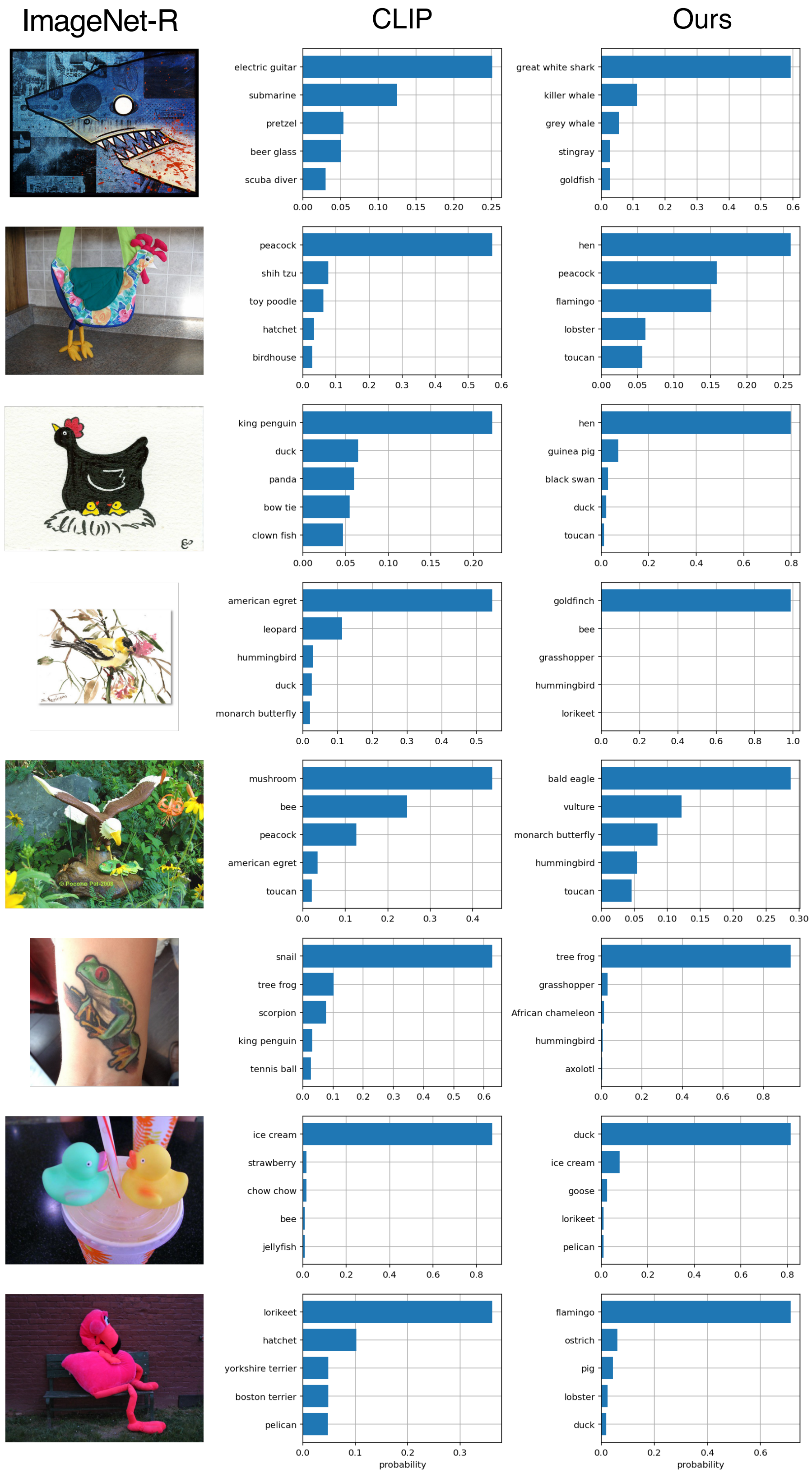
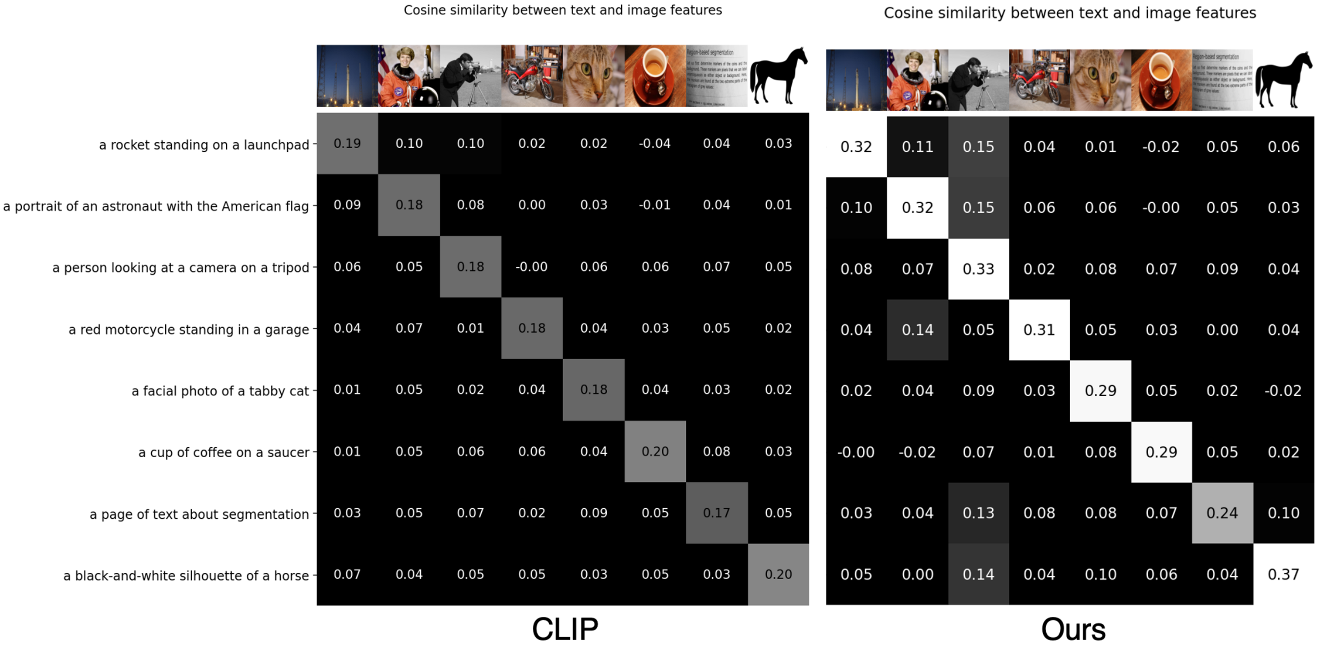


Figure 3. **Robustness Examples –** Given an image from the ImageNet-R dataset (left column), we compare the predictions of CLIP (middle column) to the predictions of our method (right column) by showing the probabilities assigned to the top 5 classes.



CLIP								
a rocket standing on a launchpad	0.19	0.10	0.10	0.02	0.02	-0.04	0.04	0.03
a portrait of an astronaut with the American flag	0.09	0.18	0.08	0.00	0.03	-0.01	0.04	0.01
a person looking at a camera on a tripod	0.06	0.05	0.18	-0.00	0.06	0.06	0.07	0.05
a red motorcycle standing in a garage	0.04	0.07	0.01	0.18	0.04	0.03	0.05	0.02
a facial photo of a tabby cat	0.01	0.05	0.02	0.04	0.18	0.04	0.03	0.02
a cup of coffee on a saucer	0.01	0.05	0.06	0.06	0.04	0.20	0.08	0.03
a page of text about segmentation	0.03	0.05	0.07	0.02	0.09	0.05	0.17	0.05
a black-and-white silhouette of a horse	0.07	0.04	0.05	0.05	0.03	0.05	0.03	0.20

Ours								
a rocket standing on a launchpad	0.32	0.11	0.15	0.04	0.01	-0.02	0.05	0.06
a portrait of an astronaut with the American flag	0.10	0.32	0.15	0.06	0.06	-0.00	0.05	0.03
a person looking at a camera on a tripod	0.08	0.07	0.33	0.02	0.08	0.07	0.09	0.04
a red motorcycle standing in a garage	0.04	0.14	0.05	0.31	0.05	0.03	0.00	0.04
a facial photo of a tabby cat	0.02	0.04	0.09	0.03	0.29	0.05	0.02	-0.02
a cup of coffee on a saucer	-0.00	-0.02	0.07	0.01	0.08	0.29	0.05	0.02
a page of text about segmentation	0.03	0.04	0.13	0.08	0.08	0.07	0.24	0.10
a black-and-white silhouette of a horse	0.05	0.00	0.14	0.04	0.10	0.06	0.04	0.37

Figure 4. Visualization of our method's similarity scores between a batch of eight image-text pairs. The baseline CLIP is optimized to maximize the diagonal scores and minimize off-diagonal scores, even when there exists non-negligible semantic similarity between unpaired instances. In contrast, our method yields elevated similarity scores on off-diagonal elements when there is increased semantic similarity between unpaired instances (e.g., photographer-to-astronaut pairing and the black-and-white-to-photographer-to-page of text pairing). As a result, we empirically observe larger similarity scores for ground truth positive pairs with our method, which coincides with improved downstream zero-shot classification performance.