Dataset	LLaVA Caption prompt
ImageNet	“What is in this photo?”
Caltech	“What is in this photo?”
Pets	“What type of pet is in this photo?”
Cars	“What type of car is in this photo?”
Flowers	“What type of flower is in this photo?”
Food	“What type of food is in this photo”
Aircraft	“What type of aircraft is in this photo?”
SUN	“What is in this photo?”
DTD	“Describe the texture in this photo.”
EuroSAT	“What type of land use is in this satellite photo?”
UCF	“What type of action is in this photo?”
ImageNet	“Question: What is in this photo? Answer: A photo of ”
ImageNet-V2	“What is in this photo?”
ImageNet-Sketch	“What is in this photo?”
ImageNet-A	“What is in this photo?”
ImageNet-R	“What is in this photo?”

Table 4. Prompts used to generate the first LLaVA caption. The other two captions are generated by using the phrases “be specific” or “be concise” in the prompt to elicit a diverse set of responses from the VL model. For example, the two other ImageNet prompts are: “What is in this photo? Be specific.” and “What is in this photo? Be concise.”

A. Experiment Implementation Details

ICE. For each method ICE is paired with, we simply take the image probability distribution output by our baseline method, and apply ICE with the computed caption scores. The hyperparameters we use to implement ICE in all experiments are $K = 5$, $\xi = 0.08$, $\epsilon = 1 \times 10^{-12}$, $\upsilon = 3$, and $P = \{$“a”, “a photo of”, “a photo containing”$\}$. We find the best empirical performance when using the centroid of the embeddings of 3 differently prompted captions and dynamically computing the caption scores weight λ using Equation 3.

Baselines. We make a good-faith attempt to tune the hyperparameters of each few-shot baseline. We use batch size 64 and SGD with momentum. Training data is sampled in a round robin fashion to maximize class diversity within each mini-batch. CLIPood trains the vision encoder for 750 iterations at learning rate 1×10^{-5} with adaptive margin value of 0.1. CoOp trains 3 prompt tokens initialized with “a photo of” for 1250 iterations at learning rate 2×10^{-4} with cross-entropy loss. MaPLe trains 3 prompt tokens prepended to each of the first 3 layers on both encoders for 750 iterations at learning rate 1 with cross-entropy loss.

Caption prompts. For all ICE results, we use a set of 3 prompt templates to generate 3 diverse captions per image. These captions are stored in our GitHub repository under the “captions” folder for easy reference. Caption prompts are model specific due to the difference in pretraining.

CoCa caption prompts are dataset-agnostic: $\{$“a”, “a photo of”, “a photo containing”$\}$. BLIP-2 and LLaVA are optimized for VQA, so we found that the best classification results are achieved by prompting with a question and answer format, using dataset-specific questions. These prompts are listed in Tables 4 and 5

B. Quantitative Results for Section 4.3

We provide the quantity of images in each category of Fig. 4 as a percentage of total test images in Table 6, for each dataset. In all the datasets, most images are either correctly predicted by both methods or incorrectly predicted by both methods. This is expected, since every dataset contains a large amount of easy images and a large amount of impossible images given a constant model capacity. However, we do observe that in all datasets except Caltech and Pets, the percentage of images where ICE successfully reclassifies an initially incorrect prediction exceeds the percentage of images where ICE incorrectly reclassifies an initially correct prediction. Furthermore, the percentage of failed re-classifications is small when compared to the percentage of successful re-classifications on datasets such as DTD (0.5% compared to 3.1%) and EuroSAT

Dataset	BLIP-2 Caption prompt
ImageNet	“Question: What is in this photo? Answer: A photo of ”
Caltech	“Question: What is in this photo? Answer: A photo of ”
Pets	“Question: What type of pet is in this photo? Answer: A photo of ”
Cars	“Question: What type of car is in this photo? Answer: A photo of ”
Flowers	“Question: What type of flower is in this photo? Answer: A photo of ”
Food	“Question: What type of food is in this photo? Answer: A photo of ”
Aircraft	“Question: What type of aircraft is in this photo? Answer: A photo of ”
SUN	“Question: What is in this photo? Answer: A photo of ”
DTD	“Question: Describe the texture in this photo. Answer: A photo of ”
EuroSAT	“Question: What type of land use is in this satellite photo? Answer: A photo of ”
UCF	“Question: What type of action is in this photo? Answer: A photo of ”
ImageNet	“Question: What is in this photo? Answer: A photo of ”
ImageNet-V2	“Question: What is in this photo? Answer: A photo of ”
ImageNet-Sketch	“Question: What is in this photo? Answer: A photo of ”
ImageNet-A	“Question: What is in this photo? Answer: A photo of ”
ImageNet-R	“Question: What is in this photo? Answer: A photo of ”

Table 5. Prompts used to generate the first BLIP-2 caption. The other two captions are generated by using the phrases “be specific” or “be concise” in the prompt to elicit a diverse set of responses from the VL model. For example, the two other ImageNet prompts are: “Question: What is in this photo? Be specific. Answer: A photo of ” and “Question: What is in this photo? Be concise. Answer: A photo of ”

		Cross-dataset Evaluation Targets											**Domain Generalization Targets**				
	INet	Caltech	Pets	Cars	Flowers	Food	Aircraft	SUN	DTD	EuroSAT	UCF	Mean	INet-V2	Sketch	INet-A	INet-R	DG Mean
Base accuracy (image zero-shot)	75.1	97.6	93.8	92.7	77.3	87.5	36.6	73.6	57.2	58.5	73.4	74.8	67.5	63.5	53.8	87.0	67.9
ICE accuracy	75.6	97.0	93.5	93.0	77.6	87.6	40.0	73.9	59.8	61.1	74.3	75.8	67.7	63.8	54.4	87.4	68.3
Both correct (%)	73.4	96.8	92.5	92.1	75.4	86.3	34.1	72.1	56.6	57.7	72.0	73.6	65.5	61.6	51.8	86.3	66.3
Base incorrect, ICE correct (%)	2.2	0.2	1.0	0.9	2.2	1.3	5.9	1.8	3.1	3.4	2.3	2.2	2.2	2.2	2.6	1.1	2.0
Image correct, ICE incorrect (%)	1.7	0.8	1.2	0.6	1.9	1.2	2.5	1.5	0.5	0.8	1.4	1.2	1.9	1.9	2.0	0.7	1.7
Both incorrect (%)	22.7	2.2	5.2	6.4	20.5	11.2	57.5	24.6	39.7	38.1	24.2	23.0	30.4	34.3	43.6	11.9	30.0

Table 6. Detailed quantitative comparison of baseline zero-shot accuracy and ICE accuracy across target datasets. These results go along with Section 4.3 and Fig. 4 in the main paper. The grey row highlights the percentage of total samples where the textual cues from the captioner correctly adjusted the baseline prediction. In general, the number of samples where the captioner steers the prediction to the correct one is greater than the number of samples where the captioner steers the prediction away form the correct one.

(0.8% compared to 3.4%). Finally, we note that images incorrectly classified by ICE may be images that are hard to classify without additional context or ambiguous images. Future work can help identify which samples to apply ICE to decrease the number of correct-to-incorrect ICE re-classifications.

C. Test Sample Nearest Neighbors in LAION-2B

Firstly, we emphasize that the captioner component of CoCa is trained jointly with the text and image encoders on the same dataset. Therefore, comparisons with baselines in the main paper are fair in terms of training data exposure. Figure 6 shows the nearest neighbor of select ImageNet test images in the LAION-2B [23] dataset in CLIP embedding space. This shows that in general, the pretraining dataset used for CoCa is not contaminated by test data.

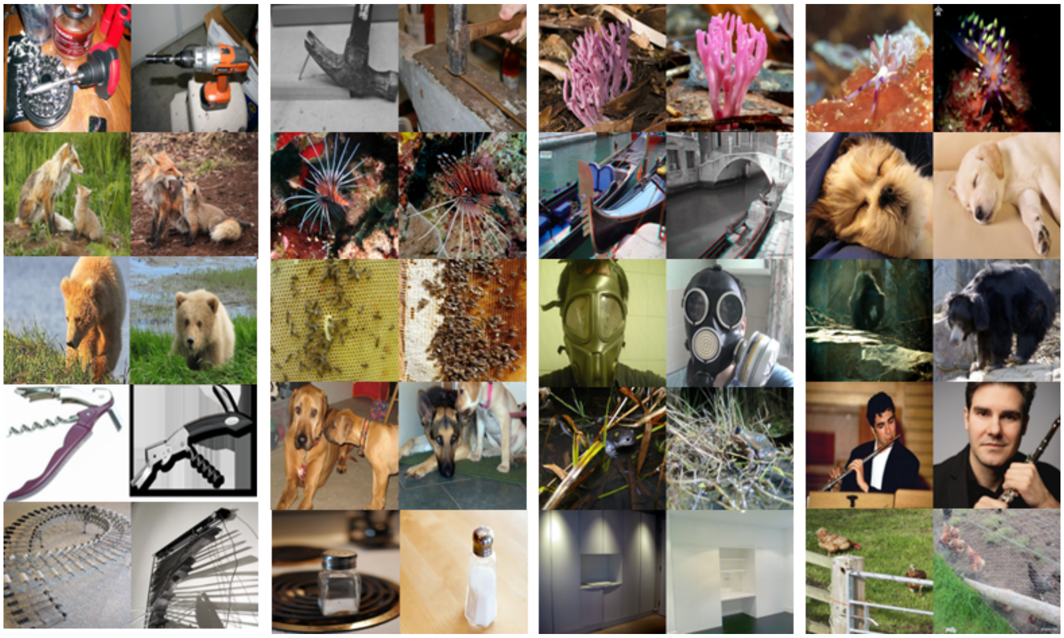

Figure 6. Test samples are generally not present in the pre-training dataset. Images on the left are a random subset of ImageNet test samples; images on the right are their nearest neighbors in LAION-2B.