Supplementary Material for "Using Language to Extend to Unseen Domains"

Here we include experimentation details such as hyperparameters and their search spaces (Section A), dataset statistics (Section B), as well as additional analysis and visualizations (Section C, D, E).

A HYPERPARAMETERS

We provide the hyperparameters used in Section 4 of the main paper in Table 3, as well as include additional training details.

For the linear probing methods, we sweep over learning rates [0.1, 0.05, 0.01, 0.005, 0.001] and weight decay [0.5, 0.05, 0.005, 0.0005]. We also sweep over the same search space when deciding the learning rate and weight decay for the augmentation network (Aug). We train logistic regression on embeddings for 400 epochs.

Dataset	Method	LR	WD	Aug LR	Aug WD	α
CUB-Paintings	*LADS*	0.001	0.05	0.001	0.05	0.5
CUB-Paintings	CLIP LP	0.001	0.05	-	-	-
DomainNet	*LADS*	0.0001	0.05	0.0001	0.05	0.5
DomainNet	CLIP LP	0.0001	0.05	-	-	-
Colored MNIST	*LADS*	0.001	0.05	0.005	0.05	1
Colored MNIST	CLIP LP	0.005	0.05	-	-	-
Waterbirds	*LADS*	0.001	0.05	0.005	0.05	0.75
Waterbirds	CLIP LP	0.001	0.05	-	-	-

Table 3: **Experiment Hyperparameters.** "Aug" LR and WD refers to the learning rate and weight decay of the augmentation network used in *LADS*. Note that we use $\alpha = 1$ for Colored MNIST since CLIP Zero-Shot does poorly on classifying MNIST.

B DATASET STATISTICS

We provide the training, validation and test (ID and OOD) splits for each dataset in Table 4.

	Training		Validation		Testing	
Dataset	ID	OOD	ID	OOD	ID	OOD
CUB-Paintings	5,994	0	2,897	0	2,897	3,047
DomainNet	5,537	0	1,200	0	1,200	11,468
Colored MNIST	30,000	0	5,000	0	5,000	5,000
Waterbirds	4,795	0	600	0	2,897	2,897

Table 4: **Dataset Statistics.** Counts of ID and OOD samples in each split of the data.

C EXTENDED DOMAIN ACCURACY VS AMOUNT OF ID TEST DATA

While we report extended domain accuracy as the overage of ID and OOD accuracy, we plot the extended domain accuracy as a function of the proportion of data in the extended domain that belongs to D_{training} in Figure 6. For example, if we assume 70% of our extended domain is from D_{training}, our extended domain accuracy would be $0.7\times$ ID Acc $+ 0.3\times$ OOD Acc. We can see that as the

proportion of the extended domain belonging to D_{training} increases, the extended domain accuracy of fine-tuning methods and *LADS* is more favorable than zero-shot.

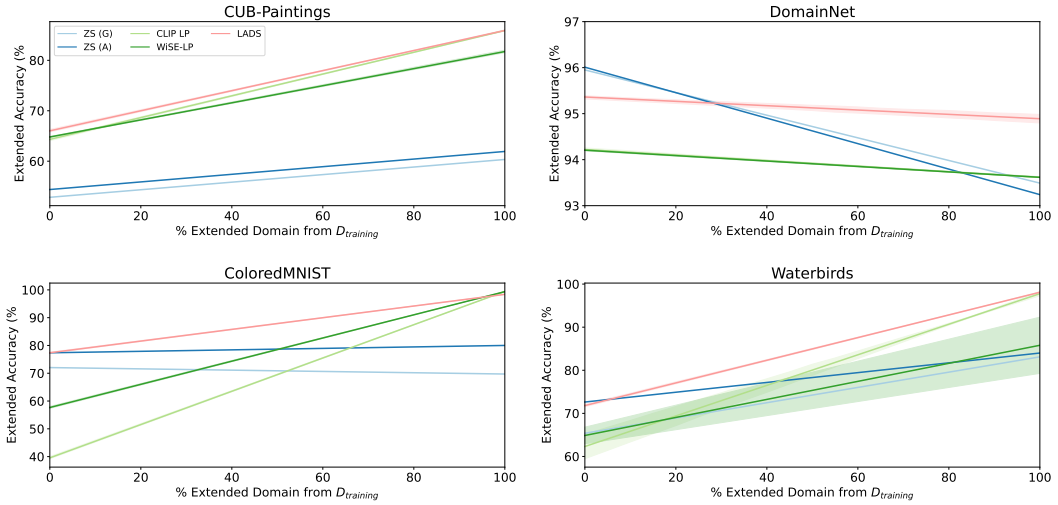

Figure 6: **Extended Domain Accuracy vs Amount of ID Test Data.** For each dataset, we compute the extended accuracy as a weighted average of the proportion of testing data from D_{training} and D_{unseen}. As the proportion of the extended domain belonging to D_{training} increases, the extended domain accuracy of fine-tuning methods and *LADS* is more favorable than zero-shot.

D Additional Analysis and Visualizations

D.1 Examples of the augmented images produced by VQGAN+CLIP

Figure 7 shows examples of augmented images produced by VQGAN+CLIP on DomainNet. We notice that in some cases the quality of the generated images can be rather poor.

Figure 7: **VQGAN+CLIP Augmentation** We give examples from the VQGAN+CLIP method. Beginning with a sketch of a bus (above) and an airplane (below), we attempt to augment them to the clipart, painting, and real domains. We repeat this procedure for approximately 15% of images in the sketch domain, picked randomly, before training a classifier on the original data and the augmented data.

D.2 AUGMENTATION QUALITY FOR CUB-PAINTINGS, COLORED MNIST, AND WATERBIRDS

Similar to Section 4.5 of the main paper, where we assess the quality of the augmented image embeddings, Figures 8 and 9 show nearest-neighbor visualizations for the CUB-Paintings and Colored MNIST datasets. Table 5 shows the domain alignment and class consistency scores for CUB-Paintings, Colored MNIST, and Waterbirds. We notice that the domain is well-aligned when applying *LADS* but the class consistency appears to be lower than on DomainNet. We hypothesize that this is due to the fact that CLIP zero-shot has rather weak performance on CUB-Paintings and MNIST, limiting the capabilities of the class consistency loss. In the future we hope that a class consistency loss which is not dependent on the CLIP zero-shot performance on the task will result in higher class consistency.

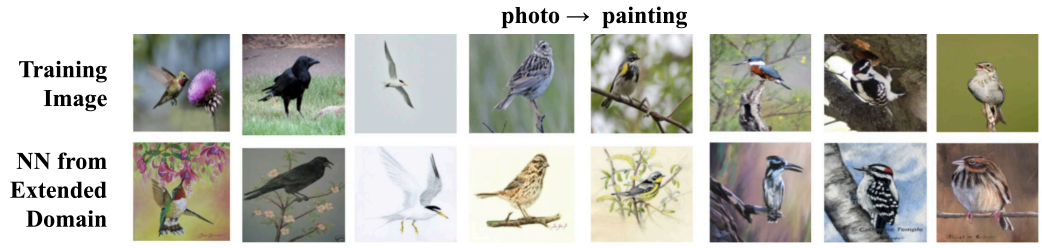

Figure 8: **Nearest Neighbors for *LADS* on CUB-Paintings.** The top row shows training images with the label being the intended domain augmentation for that embedding. The bottom row shows the nearest neighbor of the augmentation in the extended domain. *LADS* is able to augment photos to paintings while retaining class-specific information.

Figure 9: **Nearest Neighbors for *LADS* on Colored MNIST.** The top row shows training images with the the intended domain augmentation for that embedding. The bottom row shows the nearest neighbor of the augmentation in the extended domain. Note that because we do not have a class consistency loss for Colored MNIST due to CLIPS poor performance, the nearest neighbors are often of a different class.

	CUB-Paintings		ColoredMNIST		Waterbirds	
Method	DA	CC	DA	CC	DA	CC
CLIP LP	99.88±0.13%	73.26±0.85%	98.52±0.51%	96.20±0.76%	91.98±0.84%	95.66±0.26%
LADS	99.64±0.27%	54.84±2.69%	73.32±1.48%	55.08±1.79%	90.64±1.00%	95.28±0.66%

Table 5: **Augmentation Quality for CUB-Paintings, Colored MNIST, and Waterbirds.** The domain alignment(DA) and class consistency(CC) scores over 1000 randomly sampled training embeddings and their nearest neighbor in the test set. Note that for Colored MNIST, while the domain is somewhat well aligned, the class of digit often changes, likely due to the fact that we set $\alpha = 1$.

D.3 NEAREST NEIGHBOR VISUALIZATIONS ABLATING DIFFERENT LOSSES

Figures 10 and 11 compare the nearest neighbors obtained with our full approach and those obtained with the ablations of *LADS* for Waterbirds and CUB-Paintings, as defined in Section 4.6 of the main paper. The images with red outlines fail to augment to the intended domains, and the ones with red labels are augmented to a different class. We see how the domain alignment loss leads to consistent domains (but shifting appearances), while the class consistency losses lead to more similar looking birds (but not necessarily in the right domain). Combining both losses leads to the best accurate nearest neighbors in terms of both, domain alignment and class consistency. Tables 6 and 7 report accuracies, domain alignment and class consitency scores from these experiments.

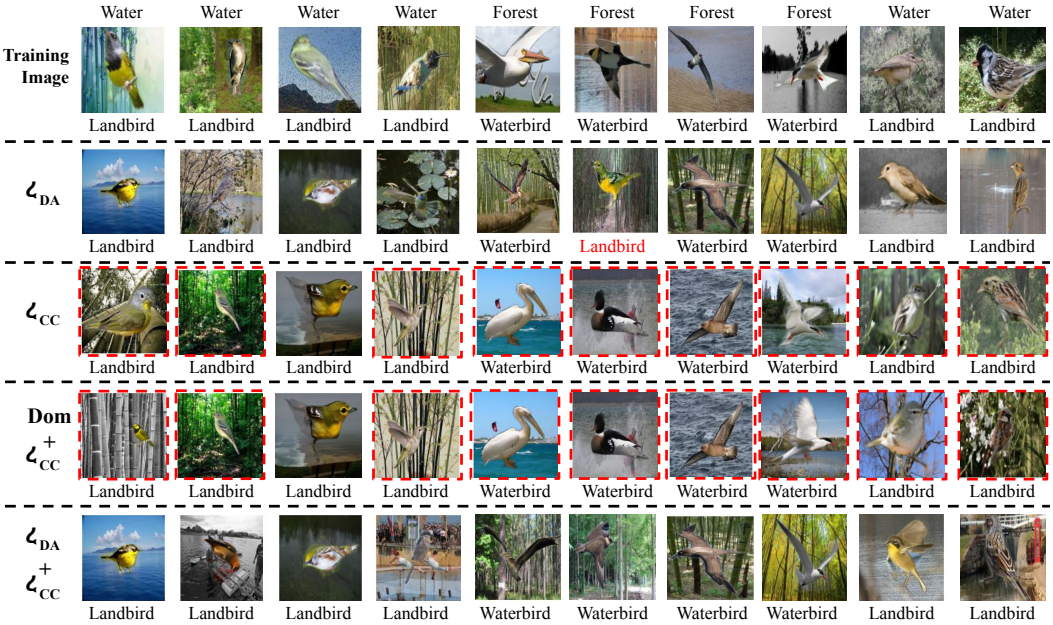

Figure 10: **Nearest Neighbors for *LADS* when ablating the loss on Waterbirds.** The top row shows training images with the label on top being the intended domain augmentation for the image's embedding, and the label on the bottom being the class of the image. The following rows correspond to choices of loss functions as in Section 4.6 and show the nearest neighbor of the embedding in the test set. The images with red outlines fail to augment to the intended domains, and the ones with red labels are augmented to a different class. Frequently the domain alignment loss is able to augment the embedding into the intended domain although sometimes it does not retain class-specific information. The class consistency loss (both generic and domain specific) retains class-specific information but often fails to augment the embedding to the intended domain.

Method	ID Acc	OOD Acc	Extended	DA score	CC score
CLIP LP	97.77±0.65%	62.31±3.79%	80.04±1.59%	91.98±0.84%	95.66±9.26%
$\mathcal{L}_{DA}$	98.70±0.41%	59.29±9.35%	79.00±4.47%	86.00±1.78%	94.90±0.86%
$\mathcal{L}_{CC}$	98.06±0.72%	59.37±7.28%	78.71±3.29%	50.46±1.93%	96.30±0.19%
Domain specific $\mathcal{L}_{CC}$	98.18±0.83%	65.46±1.32%	81.82±0.78%	52.04±1.15%	96.20±0.39%
$\mathcal{L}_{DA}$ + $\mathcal{L}_{CC}$	98.09± 0.10%	71.80±0.38%	84.95±0.17%	90.64±1.00%	95.28±0.66%

Table 6: **Effect of the Loss Functions on Waterbirds**. We report the results of training with just the domain alignment loss, the class consistency loss, a domain-specific class consistency loss, and the domain alignment + class consistency loss, on Waterbirds, corresponding to Figure 5. The DA loss leads to a high DA score but low accuracy. The CC loss leads to a low DA score and does not improve the OOD accuracy; the domain-specific CC variant brings negligible gains. Our final design (DA+CC losses) works best.

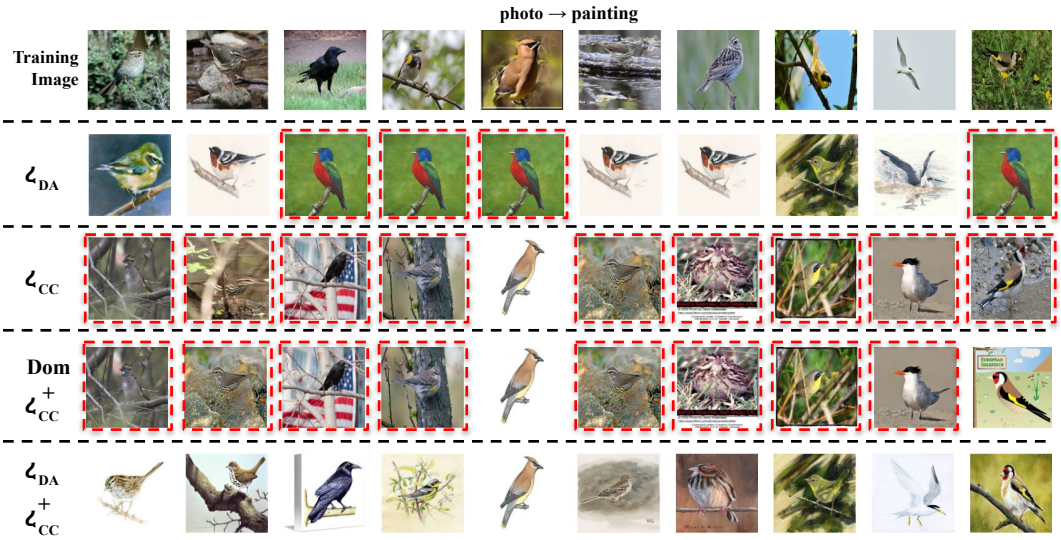

Figure 11: **Nearest Neighbors for *LADS* when ablating the loss on CUB-Paintings.** The images with red outlines fail to augment to the intended domains. While the domain alignment loss is usually able to augment the image into the intended domain, it does not retain class-specific information and often augments to similar features. The class consistency loss (both generic and domain specific) retains class-specific information but fails to convert the image to the intended domain.

Method	ID Acc	OOD Acc	Extended	DA score	CC score
CLIP LP	95.03±0.07%	93.75±0.02%	94.39±0.04%	91.42±0.47%	73.32±1.35%
$\mathcal{L}_{DA}$	83.87± 0.15%	53.74±0.76%	68.80±0.41%	37.48±2.03%	89.36±1.60%
$\mathcal{L}_{CC}$	85.88±0.14%	64.59±0.41%	75.24±0.27%	9.5±0.95%	57.46±3.21%
Domain specific $\mathcal{L}_{CC}$	85.87±0.05%	65.29±0.19%	75.58±0.12%	54.68±0.54%	56.56±1.13%
$\mathcal{L}_{DA}$ + $\mathcal{L}_{CC}$	86.14±0.29%	66.38±0.25%	76.16±0.23%	99.64±0.27%	54.84±2.69 %

Table 7: **Effect of the Loss Functions on CUB-Paintings**. We report the results of training with just the domain alignment loss, the class consistency loss, a domain-specific class consistency loss, and the domain alignment + class consistency loss, on CUB-Paintings. The DA loss leads to high DA score but low accuracy. The CC loss leads to a low DA score and does not improve the OOD accuracy; the domain-specific CC variant brings negligible gains. Our final design (DA+CC losses) works the best.

D.4 Effect of domain descriptions.

A consideration when using language to describe domains is that there are different ways to say the same thing. As such we evaluate the robustness of *LADS* when using different wordings to describe the same domain shift. Results (obtained on the Waterbirds dataset) are shown in Table 8. The first two rows, which have similar prompts, have also similar results. In contrast, the bottom two rows shows the results when a target prompt does not match the extended domain. The OOD and Extended domain results are worse with this target prompt.

D.5 Effect of vision and language model

We also experiment over several different CLIP models. Table 9 displays how *LADS* consistently obtains high accuracy across model sizes, architectures, and pretraining datasets.

t_{training}	t_{unseen}	ID	OOD	Extended
"a photo of a {} on forest."	"a photo of a {} on water."	98.09±0.10%	71.80±0.38%	84.95±0.17%
"there is a picture of a {} on forest."	"there is a picture of a {} on water."	98.03±0.04%	72.43±0.56%	85.23±0.26%
"a photo of a {} at disco."	"a photo of a {} at Supreme Court."	97.51±0.12%	67.80±0.98%	82.66±0.45%
"{} threaten old seminar."	"{} thank stable student."	98.66±0.08%	56.21±2.39%	77.43±1.16%

Table 8: **Prompt Ablation.** On the Waterbirds dataset, we experiment with similar descriptions of the training and unseen test domain (top two rows) as well as two nonsense prompts (bottom two rows). As show, *LADS* produces similar results when given prompts with the same meaning, and obtains comparable or worse results to CLIP LP when given nonsense prompts.

Method	RN50 (OpenAI)	ViT-L14 (OpenAI)	ViT-H14 (LAION-2B)
CLIP ZS (G)	47.96%	70.88%	64.96%
CLIP ZS (A)	54.44%	78.68%	81.64%
CLIP LP	56.95±0.06%	69.50±0.25%	86.50±0.32%
WiSE-FT	60.76±0.31%	78.52±0.23%	92.14±0.19%
LADS	**69.73±0.14%**	**87.88±0.13%**	**97.12±0.05%**

Table 9: Extended Accuracy on **Colored MNIST** using models of ResNet50 and ViT-L14 from OpenAI and ViT-H14 pretained on LAION-2B from OpenCLIP (Ilharco et al., 2021).

E FAILURE CASE: MNIST-TO-SVHN ADAPTATION

We show results of adapting MNIST to SVHN in Table 10, another domain adaptation benchmark. While *LADS* outperforms the fine-tuning baselines, it has poor performance when compared to zero-shot. We suspect this is because CLIP is bad at differentiating an MNIST digit from an SVHN digit, achieving only around 20% accuracy at classifying the digit domain.

Method	MNIST (ID)	SVHN (OOD)	Extended
CLIP ZS	80.60%	41.44%	61.02%
CLIP ZS (A)	97.02%	**57.3%**	**77.16%**
CLIP LP	**97.2±0.00%**	49.63 ±0.10%	73.36±0.10%
WiSE	**97.2±0.00%**	49.63 ±0.10%	73.36±0.10%
LADS	**97.2±0.00%**	51.99±0.12%	74.60±0.06%

Table 10: In-domain (ID), out-of-domain (OOD) and extended domain accuracy on MNIST to SVHN.

F EFFECT OF α

While our loss ablation (Sec 4.6) is exploring $\alpha = 0$ (only $\mathcal{L}_{\text{CC}}$) and $\alpha = 1$ (only $\mathcal{L}_{\text{DA}}$), we explore other values of α on CUB Paintings and Waterbirds in Table 12 and Table 11. Note that the domain alignment (DA) and class consistency (CC) scores are the proportion of nearest neighbors of the augmented embeddings that match the desired (target) domain and class.

As shown in Sec 4.6, when alpha = 0, the class consistency is high and the domain alignment is low, while the opposite is seen when alpha = 1. However, we see that we can often achieve a better domain alignment score when alpha is between 0 and 1. We believe that this is because the augmented embeddings drift too far from the unaugmented embeddings without the CLIP supervision that the class consistency score provides.

α	ID Acc	OOD Acc	Extended	DA score	CC score
0	97.72±0.23%	66.97±1.48%	82.34±0.66%	32.12±1.33%	95.46±0.70%
0.25	99.96±0.10%	72.95±0.40%	85.45±0.20%	67.22±2.95%	94.90±0.75%
0.5	98.03±0.06%	72.23±0.47%	85.13±0.21%	89.60±3.26%	92.86±1.52%
0.75	98.13± 0.03%	71.65±0.40%	84.89±0.18%	95.22±0.64%	89.50±1.21%
1.0	98.69± 0.35%	63.77±10.60%	81.23±5.13%	88.10±1.83%	94.30±1.23%

Table 11: **Effect of the α for *LADS* on Waterbirds**.

α	ID Acc	OOD Acc	Extended	DA score	CC score
0	85.88±0.14%	64.59±0.41%	75.24±0.27%	9.50±0.95%	57.46±3.21%
0.25	85.89±0.17%	65.04±0.73%	75.47±0.45%	89.16±3.21%	51.28±5.93%
0.5	86.14±0.29%	66.18±0.25%	76.16±0.23%	99.64±0.27%	54.84±2.69%
0.75	85.63± 0.09%	62.93±0.31%	74.28±0.18%	94.20±1.00%	44.56±2.27%
1.0	83.87± 0.15%	53.74±0.76%	68.80±0.41%	89.36±1.60%	37.48±2.03%

Table 12: **Effect of the α for *LADS* on CUB Paintings**.

G WHEN TO USE *LADS* OVER CLIP ZS

We analyze when to use LADS over CLIP ZS. In general, CLIP ZS should be used over LADS when linear probing the CLIP embeddings outperforms ZS on the majority of classes from the source domain.

For example, in Table13 we have the results for 2 splits of OfficeHome (Venkateswara et al., 2017): source domain of clipart and source domain of product, with the test domain being all 4 domains (art, clipart, product, real world). For both splits we see the same phenomenon: since CLIP ZS outperforms CLIP LP on 37/65 classes on the source domain, the overall OOD accuracy of CLIP ZS is higher than LADS. However, if we take the 28/65 classes where CLIP LP outperforms CLIP ZS, then we see that LADS OOD accuracy beats ZS and CLIP LP. We have added a section in the Appendix explaining this phenomenon.

		$D_{\text{training}} = Clipart$		$D_{\text{training}} = Product$	
Classes Subset	Method	ID	OOD	ID	OOD
All (65/65)	CLIP ZS (G)	72.07%	90.19%	94.61%	**86.13%**
All (65/65)	CLIP ZS (A)	74.75%	**91.34%**	94.49%	84.73%
All (65/65)	CLIP LP	**82.66±0.16%**	85.95±0.65%	95.46±0.03	78.53±0.40
All (65/65)	*LADS*	**82.55±0.28%**	87.97±0.50%	**96.21±0.20%**	81.20±0.22%
ZS Win (37/65)	CLIP ZS (G)	79.64%	89.70%	97.19%	**88.55%**
ZS Win (37/65)	CLIP ZS (A)	**80.95%**	**91.32%**	**97.42%**	87.86%
ZS Win (37/65)	CLIP LP	79.44±0.17%	81.92±0.69%	95.79±0.40%	75.17±0.64%
ZS Win (37/65)	*LADS*	79.15±0.61%	84.68± 0.77%	96.16±0.25%	78.79±0.37%
LP Win (28/65)	CLIP ZS (G)	62.08%	90.84%	90.72%	82.49%
LP Win (28/65)	CLIP ZS (A)	66.56%	91.36%	90.09%	80.04%
LP Win (28/65)	CLIP LP	**86.91±0.41%**	91.28±0.06%	94.97±0.67%	83.56±0.09%
LP Win (28/65)	*LADS*	**87.20±0.28%**	**92.31± 0.16%**	**96.29±0.32%**	**84.82±0.13%**

Table 13: In-domain (ID), out-of-domain (OOD) and extended domain accuracy on **OfficeHome** with a source domain of clipart and product and a test set of all 4 domains (art, clipart, product, real world) over different subsets of classes. On classes where ZS outperforms CLIP LP (37/65 classes) on the validation set, zero-shot outperforms *LADS* on OOD and extended accuracy. The opposite phenomenon is seen on the subset of classes where CLIP LP matches or beats ZS on the validation set.