A Other Methods

We detail a few other methods which we explored but do not include for analysis in Sec. 6. We found the methods we included in the paper had the best results overall, on the standard and hard examples.

- Logistic regression on pre-trained Imagenet features: We found this performed noticeably worse than ERM across the board, with a slight improvement in loss on hard positives (but worse than most other methods).
- ERM trained from scratch: We found this to perform slightly worse across the board than ERM from a pre-trained model.
- ERM with learning rate decay: We explored several learning rate decays and found, with tuning, minor improvements in overall performance from ERM. Due to the level of tuning required, we chose not to include these results and just used a constant learning rate.
- ERM with data augmentation: We explored using data augmentations such as random affine transformations, cropping, and horizontal flips, but found they did not improve performance in ERM.
- Adaptive parameters [46, 60]: We explored the possibility of learning affine transformations after each ResNet block. With tuning, we found this improved performance on hard examples over ERM but not close to the level of the other robust methods shown.
- Auxiliary prediction: We explored adding a second readout head to the final layer and predicting the highest α-context variable as an auxiliary loss. We found this to perform nearly identically to ERM in all metrics, and occasionally slightly worse.

We include this list for completeness: this is not to say that no such approach can be useful for this problem, but at the time of writing we did not find compelling enough results from these methods to yield interesting or informative comparative analysis in Sec. 6.

B Data

B.1 Licences

- The images in the COCO dataset [36] are provided under the Flickr terms of use.
- The annotations in the COCO dataset [36] are provided under a Creative Commons Attribution 4.0 License.
- The annotations in the COCO-Stuff dataset [7] are provided under a Creative Commons Attribution 4.0 License.

B.2 Dataset Details

We use the COCO-Stuff dataset [7] (`https://github.com/nightrome/cocostuff`), which contains images and annotaitons from the COCO dataset [36], as well as adding additional annotations. Since COCO is a competitive benchmark, the test set is not publicly available, so we merge the provided training/validation sets and create our own new training/validation/test split using 70/10/20 percentages. This yields training/validation/test set sizes of 86302/12328/24657.

B.3 Creation of the NOOCH Suite

When choosing the tasks, we filtered out a couple of tasks which seemed subjectively too similar to ones already in the benchmark. Specifically, we removed `sheep`, `handbag`, `bottle`, and `wine-glass` due to the fact that other similar tasks had higher average difference in average NLL on all examples and average NLL on hard examples (specifically, `cow`, `backpack`, `cup`, and `cup` respectively). We show in Table 5 how many hard positives, hard negatives, and positives are present in the test set for each task in NOOCH.

Fig 9 shows that, under the CE criterion, most of the 171 object classes in COCO-Stuff have context cues for $\alpha \geq 0$. Note that a 0-context cue is fairly strong: this means that on average, in images where Y is present, C takes up at least as much area as Y does. This indicates that there may be a large

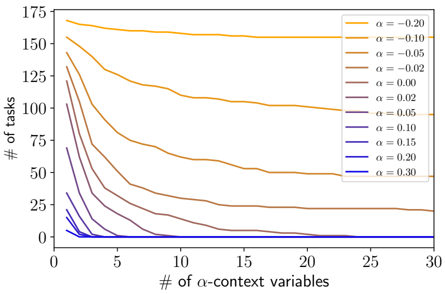

Figure 9: The number of tasks in the COCO-Stuff dataset which have context variables identifiable at each level of α. There are 171 total tasks. For instance, this figure shows that there are around 70 tasks with at least one context cue at $\alpha \geq 0.05$, and around 30 tasks with at least five context cues at $\alpha \geq 0$.

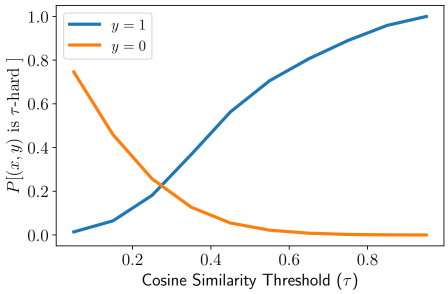

Figure 10: Using the gist criterion, the number of τ-hard positive/negative examples in the COCO-Stuff test set as a proportion of all positive/negative examples, where τ thresholds the cosine similarity to the average caption embedding of that category in the training set.

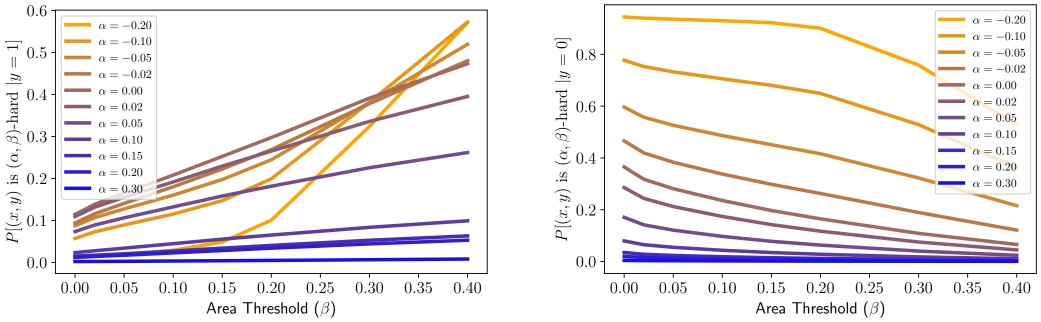

Figure 11: Across a range of α, β, the percentage of hard positive/negative examples (L/R) in the test set which are (α, β)-hard.

number of OOC examples hidden in many of these tasks. We find that the contexts returned using this method for large enough α are intuitive. Some examples of (label, 0.05-context) pairs are: (`car`, `road`), (`bowl`, `dining_table`), (`cow`, `grass`). Additionally, many tasks have more than one context variable: at $\alpha = 0$ about 50 tasks have 3, and at $\alpha = 0.05$ about 15 tasks have 3. Some examples of labels with multiple 0.05-context variables are: (`surfboard`, [`sea`, `sky-other`]), (`tennis-racket`, [`person`, `playing-field`]), (`traffic-light`, [`building-other`, `road`, `sky-other`, `tree`]). All of these groupings are intuitive, providing some evidence that this method of isolating context variables may be a useful one.

In Fig. 10 we see, across all values of τ, how many τ-hard examples there are in the test set by the gist criterion, across all 171 tasks in COCO-Stuff. As expected, the number of hard positives increases with τ, and the number of hard negatives decreased.

Fig 11 shows that hard positive/negative examples are numerous in the data. For instance, we see that at the (0.05, 0.1) level, about 10% of all possible (object, category) pairs are hard positives, and a similar number for hard negatives.

Fig. 12 shows the average loss incurred by (α, β)-hard positive and negative examples under the CE criterion as α, β vary.

Task	# Hard +	# Hard -	# +
car	1539	949	2585
bowl	944	1084	1463
boat	361	756	649
fire-hydrant	189	1577	332
airplane	258	3622	635
cow	194	2360	445
backpack	607	6413	1183
cup	743	6852	1926
surfboard	147	4011	750
tie	136	6347	778
sports-ball	151	6574	860
kite	55	6725	451

Table 5: Counts of hard positive, hard negative & positive examples in the test set per NOOCH task.

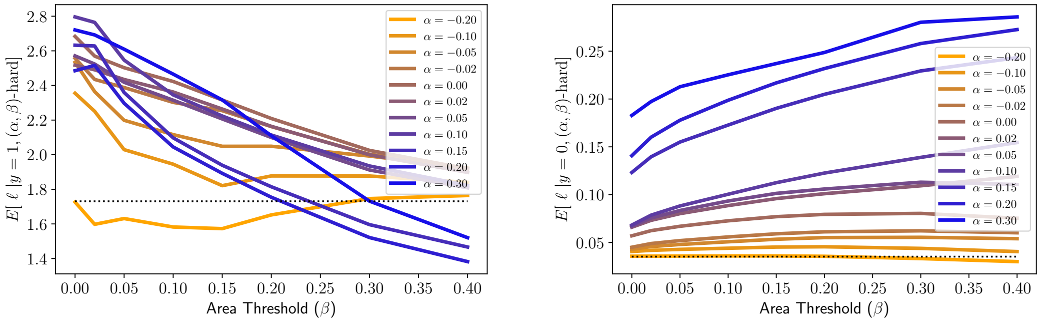

Figure 12: Across a range of α, β, the average test set loss of positive/negative examples (L/R) in the test set which are (α, β)-hard.

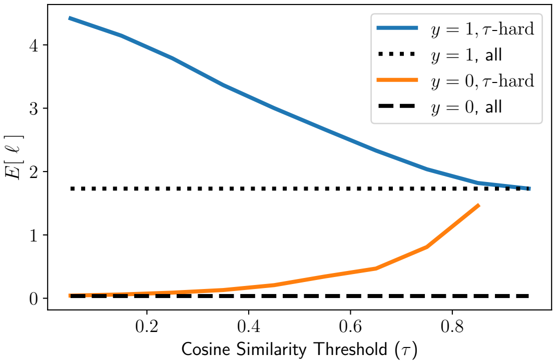

Figure 13: Caption

Fig. 13 shows the average loss incurred by (τ)-hard positive and negative examples under the gist criterion as τ varies.

In the paper, we choose $\alpha, \beta = 0.05, 0.1$ for the CE criterion, and then set τ in each task such that the number of hard positive and negative examples in each task is the same. Tables 6 and 7 show these values of τ for each for the 12 NOOCH tasks.

Task	Hard Positives	Hard Negatives
car	0.48	0.48
bowl	0.64	0.57
boat	0.73	0.43
fire-hydrant	0.74	0.35
airplane	0.81	0.19
cow	0.73	0.24
backpack	0.39	0.3
cup	0.46	0.23
surfboard	0.66	0.21
tie	0.26	0.23
sports-ball	0.47	0.12
kite	0.57	0.19

Table 6: Values for τ used for the gist criterion throughout the paper (test set).

Task	Hard Positives	Hard Negatives
car	0.47	0.5
bowl	0.65	0.57
boat	0.72	0.42
fire-hydrant	0.76	0.35
airplane	0.81	0.19
cow	0.71	0.23
backpack	0.39	0.31
cup	0.46	0.23
surfboard	0.68	0.21
tie	0.25	0.23
sports-ball	0.44	0.12
kite	0.58	0.19

Table 7: Values for τ used for the gist criterion throughout the paper (validation set).

Method	Hyperparameter Name	Values
GDRO	K	[5, 30, 60, 240]
IRM	λ	[0.1, 1, 3, 10]
CVaR	p	[0.05, 0.1, 0.15]
Focal	γ	[0.2, 0.5, 0.7, 1]
Label Reweighting	α	[0.2, 0.5, 0.7, 1]
Label Undersampling	α	[0.2, 0.5, 0.7, 1]
Environment Reweighting	α	[0.5, 1, 1.5, 2]
Environment Undersampling	α	[0.2, 0.5, 0.7, 1]

Table 8: Hyperparameter lists for each method (ERM has no method-specific hyperparameters).

C Experimental Details

C.1 Licences

The Resnet-50 is provided under an Apache 2.0 licence.

C.2 General Hyperparameters

We train (where not otherwise indicated) using SGD with learning rate of 1e-4, momentum of 0.9, L2 weight regularization of 1e-4, and batch size 32 (the largest we could fit on the available GPUs). For all methods, we use early stopping on the validation loss with a patience of 3 epochs. All images were resized to 321×321.

Training Parameters for ERM in Sec. 3. For the experiments which refer to all 171 tasks (rather than just the 12 we focus on in NOOCH , we use an Adam optimizer with 1e-3 learning rate and batch size 16. All other training parameters are the same as noted elsewhere.

Sentence Embeddings. To computer sentence embeddings, we use `stsb-distilroberta-base-v2` from the `sentence_transformers` package (`https://www.sbert.net/`).

C.3 Compute

We used GPUs on internal clusters, utilizing Nvidia Titan XP, T4, and P100 GPUs. The total number of jobs run to produce our main results, i.e. Table **??** and all associated figures, is

$$\begin{aligned} &\text{num\_tasks} \times \text{num\_seeds} \times \text{num\_methods} \times \text{num\_hyperparameters\_per\_method} && (1)\\ =&12 \times 3 \times 5 \times 3.2 = 576 && (2)\end{aligned}$$

where we calculate num_hyperparameters_per_method from Table 8. The job runtime lengths varied widely due to the use of early stopping.

C.4 Method-Specific Hyperparameters

Each of the methods we run other than ERM has a hyperparameter which trades off better overall performance and OOC performance. For each method, we swept over 4 values of these parameters (3 for CVaR), running 3 seeds at each value. The values of these hyperparameters are shown in Table 8. They were chosen through some manual experimentation to get a sense of the useful ranges for each.

C.5 Classification Error Results for Experimental Section

In Sec. 6, we show results using AUC, NLL and ECE metrics - here we show analogous plots for classification error.

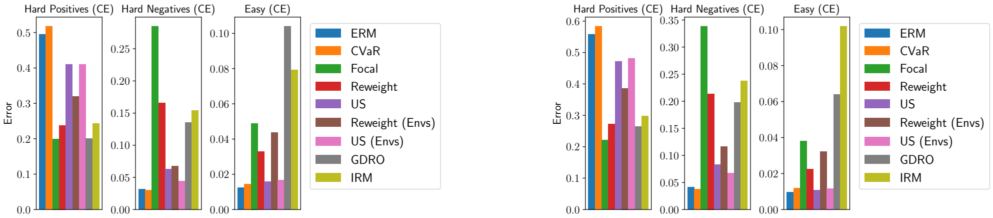

Figure 14: Classification error achieved on hard positive, hard negative, and easy examples.

Task	ERM	CVaR	Focal	Reweight	US	Reweight (Envs)	US (Envs)	GDRO	IRM
car	0.769 (0.002)	0.787 (0.007)	0.759 (0.005)	0.773 (0.001)	0.773 (0.003)	**0.886 (0.002)**	0.846 (0.013)	**0.891 (0.006)**	0.868 (0.004)
bowl	0.749 (0.003)	0.781 (0.013)	0.734 (0.004)	0.751 (0.004)	0.759 (0.007)	**0.864 (0.003)**	0.814 (0.006)	**0.865 (0.011)**	0.828 (0.023)
boat	0.869 (0.002)	0.888 (0.018)	0.823 (0.005)	0.866 (0.004)	0.877 (0.006)	**0.954 (0.002)**	0.923 (0.002)	**0.945 (0.012)**	0.925 (0.004)
fire-hydrant	0.913 (0.004)	0.933 (0.003)	0.908 (0.003)	0.927 (0.001)	0.913 (0.001)	0.933 (0.002)	0.920 (0.002)	**0.946 (0.003)**	**0.942 (0.004)**
airplane	0.986 (0.002)	0.984 (0.002)	0.983 (0.001)	0.986 (0.000)	0.985 (0.002)	**0.991 (0.000)**	0.986 (0.001)	**0.991 (0.001)**	0.986 (0.003)
cow	0.935 (0.003)	0.937 (0.003)	0.932 (0.001)	0.939 (0.002)	0.938 (0.003)	**0.963 (0.001)**	0.948 (0.002)	**0.963 (0.004)**	0.943 (0.001)
backpack	0.812 (0.004)	0.806 (0.010)	0.809 (0.002)	0.816 (0.003)	0.812 (0.008)	0.813 (0.002)	0.816 (0.014)	**0.871 (0.008)**	0.844 (0.018)
cup	**0.870 (0.004)**	0.863 (0.008)	0.867 (0.004)	**0.876 (0.000)**	**0.873 (0.005)**	**0.879 (0.006)**	**0.875 (0.001)**	0.825 (0.019)	**0.869 (0.023)**
surfboard	0.939 (0.002)	0.947 (0.002)	0.933 (0.001)	0.940 (0.001)	0.939 (0.003)	**0.960 (0.001)**	0.940 (0.002)	**0.960 (0.004)**	**0.957 (0.008)**
tie	0.742 (0.002)	0.752 (0.008)	0.728 (0.006)	0.760 (0.005)	0.763 (0.009)	0.756 (0.006)	0.761 (0.003)	**0.806 (0.015)**	**0.775 (0.028)**
sports-ball	0.867 (0.002)	0.869 (0.007)	0.871 (0.001)	0.869 (0.002)	0.868 (0.006)	**0.911 (0.001)**	0.890 (0.002)	**0.911 (0.005)**	0.894 (0.003)
kite	0.932 (0.001)	0.932 (0.003)	0.937 (0.002)	0.940 (0.001)	0.928 (0.002)	0.949 (0.001)	0.936 (0.001)	**0.960 (0.005)**	**0.950 (0.012)**
Average	0.865 (0.001)	0.873 (0.002)	0.857 (0.001)	0.870 (0.001)	0.869 (0.002)	0.905 (0.001)	0.888 (0.002)	**0.911 (0.003)**	0.899 (0.004)

Task	ERM	CVaR	Focal	Reweight	US	Reweight (Envs)	US (Envs)	GDRO	IRM
car	0.766 (0.002)	0.785 (0.003)	0.763 (0.005)	0.768 (0.005)	0.773 (0.006)	0.845 (0.001)	0.823 (0.004)	**0.863 (0.008)**	0.832 (0.017)
bowl	0.642 (0.004)	**0.686 (0.015)**	0.635 (0.002)	0.656 (0.006)	0.678 (0.009)	**0.692 (0.007)**	**0.698 (0.006)**	**0.704 (0.013)**	**0.670 (0.021)**
boat	0.826 (0.002)	0.855 (0.019)	0.771 (0.008)	0.823 (0.004)	0.823 (0.002)	**0.897 (0.002)**	0.868 (0.003)	0.884 (0.008)	0.872 (0.002)
fire-hydrant	0.840 (0.007)	**0.865 (0.008)**	0.822 (0.005)	0.849 (0.002)	0.841 (0.007)	0.834 (0.005)	0.836 (0.007)	0.823 (0.007)	0.845 (0.004)
airplane	**0.977 (0.001)**	**0.979 (0.000)**	0.975 (0.001)	**0.978 (0.000)**	0.976 (0.001)	**0.978 (0.001)**	0.977 (0.001)	**0.981 (0.003)**	**0.977 (0.002)**
cow	0.901 (0.004)	**0.912 (0.001)**	0.877 (0.002)	0.903 (0.002)	0.906 (0.004)	0.908 (0.002)	**0.908 (0.003)**	**0.911 (0.006)**	0.896 (0.008)
backpack	0.716 (0.004)	0.717 (0.013)	0.725 (0.002)	0.731 (0.003)	**0.749 (0.004)**	0.729 (0.004)	0.736 (0.003)	**0.732 (0.012)**	**0.750 (0.010)**
cup	0.733 (0.006)	0.738 (0.007)	0.727 (0.000)	0.735 (0.005)	0.742 (0.004)	0.759 (0.005)	**0.771 (0.005)**	0.742 (0.015)	**0.755 (0.011)**
surfboard	0.913 (0.002)	**0.920 (0.002)**	0.900 (0.002)	0.912 (0.002)	0.914 (0.000)	0.909 (0.001)	0.912 (0.001)	0.882 (0.006)	0.894 (0.011)
tie	**0.822 (0.003)**	0.816 (0.007)	0.813 (0.004)	**0.829 (0.004)**	**0.824 (0.002)**	**0.829 (0.002)**	**0.831 (0.005)**	**0.835 (0.009)**	**0.840 (0.017)**
sports-ball	0.830 (0.003)	0.832 (0.008)	0.828 (0.001)	0.830 (0.002)	0.822 (0.004)	0.866 (0.003)	0.859 (0.002)	**0.880 (0.001)**	0.855 (0.011)
kite	**0.942 (0.001)**	**0.939 (0.005)**	**0.945 (0.002)**	**0.947 (0.001)**	0.936 (0.003)	**0.947 (0.002)**	**0.939 (0.002)**	**0.947 (0.007)**	**0.943 (0.010)**
Average	0.826 (0.001)	0.837 (0.003)	0.815 (0.001)	0.830 (0.001)	0.832 (0.001)	**0.849 (0.001)**	0.847 (0.001)	**0.849 (0.003)**	0.844 (0.003)

Table 9: AUC on hard test examples for all 12 NOOCH stress tests, after hyperparameter selection. Bold numbers have overlapping standard deviations. L: NOOCH-CE. R: NOOCH-Gist.

C.6 Detailed Results

For each metric of AUC, NLL, Error, and ECE, each of the 12 tasks in NOOCH , and both NOOCH-CE and NOOCH-Gist, we show results below with standard deviation across 3 seeds. There are the same results as Figures 4, 14, 5, are 6, but shown in more detail.

Task	ERM	CVaR	Focal	Reweight	US	Reweight (Envs)	US (Envs)	GDRO	IRM
car	**0.988 (0.000)**	0.987 (0.000)	0.986 (0.000)	**0.988 (0.000)**	0.987 (0.000)	0.939 (0.001)	0.961 (0.004)	0.912 (0.014)	0.959 (0.003)
bowl	**0.983 (0.000)**	**0.981 (0.002)**	0.982 (0.000)	**0.983 (0.000)**	0.980 (0.002)	0.913 (0.004)	0.968 (0.004)	0.881 (0.037)	0.943 (0.007)
boat	0.995 (0.000)	**0.996 (0.001)**	**0.997 (0.000)**	0.996 (0.000)	0.995 (0.000)	0.961 (0.003)	0.982 (0.001)	0.961 (0.011)	0.984 (0.005)
fire-hydrant	0.979 (0.001)	0.982 (0.003)	**0.985 (0.001)**	**0.985 (0.001)**	0.978 (0.001)	0.965 (0.002)	0.977 (0.004)	0.968 (0.004)	0.981 (0.001)
airplane	0.997 (0.001)	**0.999 (0.001)**	**0.998 (0.000)**	0.997 (0.000)	0.996 (0.000)	0.995 (0.000)	0.996 (0.000)	**0.997 (0.002)**	**0.998 (0.002)**
cow	0.995 (0.001)	0.993 (0.002)	**0.997 (0.000)**	0.997 (0.000)	0.995 (0.001)	0.989 (0.001)	0.995 (0.001)	0.981 (0.007)	0.995 (0.001)
backpack	0.932 (0.001)	0.910 (0.005)	**0.937 (0.001)**	**0.937 (0.001)**	0.931 (0.003)	**0.935 (0.002)**	**0.932 (0.006)**	0.877 (0.011)	**0.928 (0.008)**
cup	**0.966 (0.001)**	0.962 (0.001)	**0.966 (0.001)**	**0.966 (0.000)**	**0.965 (0.001)**	0.964 (0.001)	0.963 (0.001)	0.922 (0.007)	0.946 (0.004)
surfboard	**0.998 (0.000)**	0.998 (0.000)	**0.998 (0.000)**	**0.998 (0.000)**	**0.998 (0.000)**	0.960 (0.002)	0.997 (0.000)	0.980 (0.004)	0.989 (0.003)
tie	0.987 (0.001)	0.985 (0.002)	**0.989 (0.000)**	0.987 (0.000)	0.980 (0.002)	0.987 (0.000)	0.981 (0.002)	0.983 (0.002)	**0.987 (0.003)**
sports-ball	0.993 (0.000)	**0.993 (0.000)**	**0.994 (0.000)**	**0.994 (0.000)**	0.991 (0.001)	0.885 (0.004)	0.991 (0.001)	0.963 (0.002)	0.984 (0.000)
kite	0.999 (0.000)	**0.999 (0.000)**	**0.999 (0.000)**	**0.999 (0.000)**	**0.999 (0.000)**	0.999 (0.000)	**0.999 (0.000)**	0.996 (0.000)	0.997 (0.001)
Average	0.984 (0.000)	0.982 (0.001)	**0.986 (0.000)**	**0.986 (0.000)**	0.983 (0.000)	0.958 (0.001)	0.978 (0.001)	0.952 (0.004)	0.974 (0.001)

Task	ERM	CVaR	Focal	Reweight	US	Reweight (Envs)	US (Envs)	GDRO	IRM
car	**0.990 (0.000)**	0.989 (0.000)	0.988 (0.000)	0.988 (0.000)	0.989 (0.001)	0.956 (0.001)	0.976 (0.001)	0.890 (0.011)	0.968 (0.004)
bowl	**0.991 (0.000)**	**0.990 (0.001)**	0.989 (0.000)	**0.990 (0.000)**	0.986 (0.001)	0.953 (0.003)	0.970 (0.002)	0.931 (0.023)	0.981 (0.001)
boat	**0.999 (0.000)**	**0.999 (0.000)**	0.999 (0.000)	**0.999 (0.000)**	0.998 (0.000)	0.996 (0.000)	0.998 (0.000)	0.995 (0.001)	0.997 (0.001)
fire-hydrant	0.999 (0.000)	**0.998 (0.001)**	0.999 (0.000)	0.999 (0.000)	0.997 (0.000)	0.999 (0.000)	0.998 (0.000)	**0.999 (0.001)**	**0.999 (0.000)**
airplane	1.000 (0.000)	**1.000 (0.000)**	1.000 (0.000)	**1.000 (0.000)**	**1.000 (0.000)**	1.000 (0.000)	**1.000 (0.000)**	**1.000 (0.000)**	1.000 (0.000)
cow	1.000 (0.000)	0.998 (0.001)	**1.000 (0.000)**	1.000 (0.000)	0.999 (0.000)	0.999 (0.000)	0.998 (0.000)	0.998 (0.001)	0.999 (0.000)
backpack	**0.972 (0.001)**	0.952 (0.006)	**0.972 (0.001)**	**0.972 (0.000)**	0.961 (0.003)	**0.973 (0.001)**	0.966 (0.001)	0.958 (0.002)	**0.968 (0.005)**
cup	**0.995 (0.001)**	0.989 (0.001)	**0.995 (0.000)**	**0.995 (0.000)**	0.991 (0.002)	0.989 (0.001)	0.984 (0.001)	0.950 (0.010)	0.979 (0.008)
surfboard	0.999 (0.000)	0.999 (0.000)	**0.999 (0.000)**	0.999 (0.000)	0.999 (0.000)	0.999 (0.000)	0.999 (0.000)	0.994 (0.002)	0.996 (0.001)
tie	0.978 (0.001)	0.976 (0.002)	**0.981 (0.000)**	0.980 (0.000)	0.972 (0.001)	0.979 (0.000)	0.969 (0.002)	0.979 (0.001)	**0.981 (0.002)**
sports-ball	0.999 (0.000)	0.998 (0.000)	**0.999 (0.000)**	0.999 (0.000)	0.999 (0.000)	0.898 (0.005)	0.989 (0.001)	0.968 (0.004)	0.990 (0.002)
kite	0.998 (0.000)	**0.999 (0.000)**	**0.999 (0.000)**	**0.999 (0.000)**	0.999 (0.000)	0.999 (0.000)	0.999 (0.000)	0.997 (0.000)	0.997 (0.001)
Average	**0.993 (0.000)**	0.991 (0.001)	**0.993 (0.000)**	**0.993 (0.000)**	0.991 (0.000)	0.978 (0.000)	0.987 (0.000)	0.972 (0.002)	0.988 (0.001)

Table 10: AUC on easy test examples for all 12 NOOCH stress tests, after hyperparameter selection. Bold numbers have overlapping standard deviations. L: NOOCH-CE. R: NOOCH-Gist.

Task	ERM	CVaR	Focal	Reweight	US	Reweight (Envs)	US (Envs)	GDRO	IRM
car	0.435 (0.012)	0.476 (0.030)	0.197 (0.008)	0.301 (0.003)	0.357 (0.017)	0.253 (0.013)	0.365 (0.020)	**0.136 (0.011)**	0.253 (0.041)
bowl	0.581 (0.032)	0.627 (0.054)	**0.162 (0.007)**	0.336 (0.014)	0.406 (0.021)	0.314 (0.005)	0.369 (0.015)	0.239 (0.007)	0.286 (0.023)
boat	0.353 (0.009)	0.362 (0.028)	**0.120 (0.009)**	0.183 (0.002)	0.307 (0.002)	0.199 (0.002)	0.332 (0.022)	**0.145 (0.020)**	0.225 (0.055)
fire-hydrant	0.460 (0.011)	0.406 (0.010)	**0.101 (0.011)**	0.217 (0.004)	0.418 (0.020)	0.383 (0.005)	0.397 (0.007)	0.268 (0.026)	0.268 (0.070)
airplane	0.137 (0.008)	0.125 (0.013)	**0.054 (0.000)**	**0.054 (0.003)**	0.112 (0.005)	0.103 (0.002)	0.114 (0.004)	**0.056 (0.007)**	0.090 (0.013)
cow	0.357 (0.009)	0.445 (0.027)	**0.115 (0.010)**	0.192 (0.002)	0.340 (0.013)	0.259 (0.005)	0.314 (0.019)	0.187 (0.016)	0.249 (0.015)
backpack	0.822 (0.003)	0.883 (0.039)	**0.239 (0.005)**	**0.237 (0.007)**	0.588 (0.042)	0.599 (0.014)	0.638 (0.011)	**0.238 (0.014)**	**0.271 (0.082)**
cup	0.678 (0.011)	0.704 (0.013)	**0.189 (0.010)**	**0.186 (0.006)**	0.431 (0.007)	0.279 (0.011)	0.480 (0.014)	0.223 (0.018)	0.310 (0.022)
surfboard	0.372 (0.014)	0.388 (0.061)	**0.184 (0.017)**	**0.175 (0.012)**	0.345 (0.008)	**0.184 (0.006)**	0.313 (0.028)	**0.161 (0.018)**	0.220 (0.037)
tie	0.831 (0.006)	0.863 (0.043)	0.547 (0.025)	0.512 (0.009)	0.757 (0.012)	0.752 (0.009)	0.760 (0.014)	**0.380 (0.028)**	**0.348 (0.018)**
sports-ball	0.521 (0.022)	0.563 (0.081)	0.285 (0.014)	0.258 (0.009)	0.477 (0.037)	**0.199 (0.000)**	0.446 (0.030)	**0.205 (0.053)**	0.243 (0.040)
kite	0.388 (0.034)	0.370 (0.031)	0.206 (0.009)	0.206 (0.009)	0.382 (0.000)	0.315 (0.023)	0.394 (0.023)	**0.176 (0.043)**	**0.152 (0.017)**
Average	0.495 (0.005)	0.518 (0.012)	**0.200 (0.003)**	0.238 (0.002)	0.410 (0.006)	0.320 (0.003)	0.410 (0.005)	**0.201 (0.007)**	0.243 (0.012)

Task	ERM	CVaR	Focal	Reweight	US	Reweight (Envs)	US (Envs)	GDRO	IRM
car	0.459 (0.013)	0.495 (0.029)	**0.207 (0.009)**	0.323 (0.006)	0.380 (0.014)	0.279 (0.015)	0.386 (0.024)	**0.251 (0.044)**	0.278 (0.057)
bowl	0.616 (0.031)	0.670 (0.059)	**0.170 (0.008)**	0.370 (0.017)	0.443 (0.021)	0.405 (0.022)	0.416 (0.016)	0.300 (0.006)	0.276 (0.027)
boat	0.419 (0.015)	0.458 (0.040)	**0.129 (0.009)**	0.211 (0.000)	0.357 (0.006)	0.283 (0.009)	0.400 (0.021)	0.211 (0.040)	0.361 (0.135)
fire-hydrant	0.612 (0.015)	0.543 (0.002)	**0.150 (0.010)**	0.307 (0.011)	0.550 (0.028)	0.527 (0.002)	0.524 (0.009)	0.399 (0.024)	0.376 (0.083)
airplane	0.200 (0.013)	0.186 (0.021)	**0.087 (0.004)**	**0.087 (0.002)**	0.177 (0.005)	0.168 (0.004)	0.177 (0.002)	0.098 (0.002)	0.116 (0.011)
cow	0.459 (0.011)	0.526 (0.043)	**0.144 (0.018)**	0.249 (0.006)	0.426 (0.009)	0.361 (0.004)	0.431 (0.033)	0.268 (0.022)	0.320 (0.033)
backpack	0.881 (0.008)	0.919 (0.044)	**0.318 (0.002)**	**0.316 (0.006)**	0.666 (0.035)	0.686 (0.005)	0.715 (0.008)	0.340 (0.006)	**0.337 (0.073)**
cup	0.758 (0.008)	0.779 (0.016)	0.301 (0.004)	**0.281 (0.011)**	0.537 (0.005)	0.323 (0.024)	0.576 (0.011)	**0.247 (0.030)**	0.369 (0.048)
surfboard	0.517 (0.017)	0.542 (0.081)	**0.213 (0.020)**	**0.206 (0.012)**	0.449 (0.010)	0.367 (0.010)	0.497 (0.024)	0.329 (0.031)	0.420 (0.067)
tie	0.848 (0.009)	0.870 (0.054)	0.387 (0.015)	0.395 (0.009)	0.713 (0.010)	0.657 (0.012)	0.755 (0.023)	**0.299 (0.018)**	**0.321 (0.054)**
sports-ball	0.554 (0.017)	0.614 (0.070)	0.360 (0.008)	0.344 (0.005)	0.600 (0.019)	**0.245 (0.011)**	0.503 (0.029)	**0.243 (0.047)**	**0.243 (0.027)**
kite	0.382 (0.030)	0.400 (0.015)	**0.182 (0.015)**	**0.176 (0.017)**	0.364 (0.015)	0.333 (0.009)	0.406 (0.017)	**0.188 (0.052)**	**0.164 (0.030)**
Average	0.559 (0.005)	0.583 (0.013)	**0.221 (0.003)**	0.272 (0.003)	0.472 (0.005)	0.386 (0.004)	0.482 (0.006)	0.264 (0.009)	0.298 (0.018)

Table 11: Classification error on hard positive test examples for all 12 NOOCH stress tests, after hyperparameter selection. Bold numbers have overlapping standard deviations. L: NOOCH-CE. R: NOOCH-Gist.

Task	ERM	CVaR	Focal	Reweight	US	Reweight (Envs)	US (Envs)	GDRO	IRM
car	0.118 (0.009)	**0.084 (0.019)**	0.554 (0.017)	0.269 (0.006)	0.195 (0.027)	0.132 (0.009)	0.122 (0.013)	0.176 (0.029)	0.158 (0.053)
bowl	**0.087 (0.016)**	**0.113 (0.043)**	0.628 (0.024)	0.274 (0.017)	0.200 (0.024)	0.142 (0.010)	0.163 (0.021)	0.220 (0.030)	0.191 (0.027)
boat	**0.052 (0.004)**	**0.037 (0.022)**	0.622 (0.023)	0.266 (0.028)	0.088 (0.008)	**0.064 (0.006)**	**0.037 (0.009)**	0.130 (0.044)	**0.113 (0.090)**
fire-hydrant	**0.004 (0.001)**	**0.003 (0.001)**	0.330 (0.006)	0.119 (0.007)	0.009 (0.003)	0.008 (0.001)	0.009 (0.003)	0.041 (0.003)	0.117 (0.078)
airplane	**0.002 (0.000)**	**0.002 (0.001)**	0.039 (0.003)	0.038 (0.004)	**0.003 (0.000)**	0.004 (0.000)	**0.003 (0.001)**	0.012 (0.002)	0.010 (0.002)
cow	**0.011 (0.001)**	**0.010 (0.001)**	0.175 (0.003)	0.068 (0.002)	**0.010 (0.001)**	0.014 (0.001)	**0.010 (0.002)**	0.051 (0.012)	0.045 (0.007)
backpack	**0.018 (0.002)**	**0.022 (0.019)**	0.307 (0.004)	0.274 (0.002)	0.069 (0.012)	0.075 (0.005)	0.051 (0.005)	0.225 (0.009)	0.266 (0.035)
cup	**0.037 (0.001)**	0.060 (0.011)	0.225 (0.003)	0.217 (0.008)	0.091 (0.001)	0.171 (0.016)	0.067 (0.001)	0.303 (0.051)	0.226 (0.061)
surfboard	**0.013 (0.001)**	**0.008 (0.003)**	0.092 (0.004)	0.080 (0.002)	0.014 (0.001)	0.037 (0.001)	**0.013 (0.002)**	0.051 (0.022)	0.023 (0.006)
tie	0.014 (0.000)	**0.010 (0.003)**	0.239 (0.024)	0.178 (0.012)	0.024 (0.004)	0.039 (0.004)	0.021 (0.003)	0.245 (0.038)	0.468 (0.061)
sports-ball	0.020 (0.001)	**0.013 (0.004)**	0.141 (0.003)	0.141 (0.003)	0.048 (0.008)	0.121 (0.007)	0.027 (0.002)	0.142 (0.041)	0.173 (0.041)
kite	**0.004 (0.001)**	**0.003 (0.001)**	0.066 (0.007)	0.061 (0.008)	0.004 (0.001)	0.006 (0.001)	0.004 (0.001)	0.031 (0.012)	0.063 (0.004)
Average	**0.032 (0.002)**	**0.030 (0.005)**	0.285 (0.004)	0.165 (0.003)	0.063 (0.003)	0.068 (0.002)	0.044 (0.002)	0.136 (0.008)	0.154 (0.014)

Task	ERM	CVaR	Focal	Reweight	US	Reweight (Envs)	US (Envs)	GDRO	IRM
car	0.115 (0.008)	**0.086 (0.019)**	0.521 (0.016)	0.262 (0.009)	0.187 (0.029)	0.156 (0.014)	0.125 (0.013)	0.188 (0.078)	0.218 (0.031)
bowl	**0.141 (0.024)**	**0.147 (0.046)**	0.785 (0.031)	0.403 (0.014)	0.305 (0.032)	0.359 (0.021)	0.292 (0.036)	0.471 (0.039)	0.561 (0.039)
boat	**0.067 (0.007)**	**0.045 (0.021)**	0.672 (0.021)	0.343 (0.026)	0.114 (0.008)	0.116 (0.006)	**0.069 (0.004)**	0.210 (0.045)	0.186 (0.087)
fire-hydrant	0.013 (0.002)	**0.004 (0.001)**	0.427 (0.009)	0.193 (0.008)	0.024 (0.005)	0.019 (0.002)	0.026 (0.007)	0.127 (0.026)	0.212 (0.116)
airplane	**0.002 (0.000)**	**0.003 (0.001)**	0.048 (0.004)	0.047 (0.004)	0.003 (0.000)	0.006 (0.000)	0.003 (0.000)	0.029 (0.002)	0.078 (0.025)
cow	0.019 (0.000)	**0.014 (0.001)**	0.291 (0.005)	0.118 (0.003)	0.025 (0.001)	0.034 (0.003)	0.021 (0.004)	0.105 (0.018)	0.146 (0.037)
backpack	**0.027 (0.003)**	**0.032 (0.019)**	0.387 (0.006)	0.338 (0.004)	0.101 (0.015)	0.104 (0.006)	0.076 (0.005)	0.323 (0.015)	0.332 (0.032)
cup	**0.055 (0.001)**	0.085 (0.009)	0.392 (0.009)	0.376 (0.012)	0.152 (0.004)	0.305 (0.014)	0.121 (0.003)	0.405 (0.032)	0.340 (0.051)
surfboard	**0.015 (0.001)**	**0.010 (0.004)**	0.122 (0.005)	0.110 (0.003)	0.022 (0.001)	0.084 (0.003)	**0.016 (0.003)**	0.094 (0.034)	0.102 (0.026)
tie	0.013 (0.000)	**0.009 (0.003)**	0.173 (0.016)	0.142 (0.008)	0.022 (0.004)	0.036 (0.004)	0.020 (0.002)	0.216 (0.025)	0.413 (0.044)
sports-ball	0.023 (0.002)	**0.014 (0.005)**	0.174 (0.003)	0.172 (0.005)	0.034 (0.004)	0.168 (0.016)	0.035 (0.004)	0.168 (0.049)	0.194 (0.014)
kite	0.004 (0.001)	**0.003 (0.001)**	0.071 (0.007)	0.065 (0.007)	0.005 (0.001)	0.008 (0.001)	0.005 (0.001)	0.041 (0.015)	0.069 (0.005)
Average	**0.041 (0.002)**	**0.038 (0.005)**	0.338 (0.004)	0.214 (0.003)	0.083 (0.004)	0.116 (0.003)	0.067 (0.003)	0.198 (0.011)	0.238 (0.015)

Table 12: Classification error on hard negative test examples for all 12 NOOCH stress tests, after hyperparameter selection. Bold numbers have overlapping standard deviations. L: NOOCH-CE. R: NOOCH-Gist.

Task	ERM	CVaR	Focal	Reweight	US	Reweight (Envs)	US (Envs)	GDRO	IRM
car	0.020 (0.001)	**0.019 (0.001)**	0.066 (0.002)	0.031 (0.001)	0.026 (0.002)	0.071 (0.004)	0.030 (0.002)	0.465 (0.104)	0.068 (0.013)
bowl	**0.018 (0.001)**	**0.019 (0.001)**	0.087 (0.006)	0.034 (0.001)	0.028 (0.001)	0.070 (0.001)	0.036 (0.002)	0.107 (0.020)	0.077 (0.005)
boat	0.004 (0.000)	**0.004 (0.000)**	0.039 (0.002)	0.016 (0.001)	0.006 (0.000)	0.058 (0.002)	0.006 (0.001)	0.089 (0.012)	0.092 (0.059)
fire-hydrant	0.003 (0.000)	**0.002 (0.000)**	0.058 (0.002)	0.020 (0.001)	0.003 (0.000)	0.005 (0.000)	0.004 (0.001)	0.027 (0.001)	0.040 (0.029)
airplane	**0.001 (0.000)**	**0.001 (0.000)**	0.008 (0.000)	0.007 (0.001)	**0.001 (0.000)**	0.002 (0.000)	**0.001 (0.000)**	0.031 (0.009)	0.003 (0.001)
cow	0.003 (0.000)	**0.003 (0.000)**	0.023 (0.000)	0.009 (0.000)	0.004 (0.000)	0.006 (0.000)	0.004 (0.000)	0.026 (0.003)	0.009 (0.001)
backpack	**0.028 (0.000)**	0.038 (0.005)	0.124 (0.002)	0.109 (0.002)	0.045 (0.003)	0.038 (0.002)	0.037 (0.000)	0.135 (0.013)	0.146 (0.024)
cup	**0.039 (0.000)**	0.052 (0.004)	0.098 (0.003)	0.094 (0.002)	0.047 (0.001)	0.105 (0.003)	0.048 (0.001)	0.145 (0.015)	0.125 (0.010)
surfboard	**0.004 (0.000)**	**0.005 (0.001)**	0.014 (0.001)	0.011 (0.000)	**0.004 (0.000)**	0.033 (0.002)	**0.005 (0.000)**	0.041 (0.017)	0.016 (0.005)
tie	0.011 (0.000)	0.013 (0.002)	0.037 (0.002)	0.032 (0.002)	**0.011 (0.000)**	**0.010 (0.001)**	0.012 (0.000)	0.080 (0.016)	0.260 (0.035)
sports-ball	0.013 (0.000)	0.014 (0.001)	0.022 (0.001)	0.021 (0.002)	**0.011 (0.000)**	0.124 (0.011)	0.015 (0.001)	0.079 (0.017)	0.093 (0.024)
kite	**0.002 (0.000)**	**0.002 (0.000)**	0.011 (0.001)	0.010 (0.001)	**0.002 (0.000)**	0.003 (0.000)	0.003 (0.000)	0.024 (0.008)	0.021 (0.002)
Average	**0.012 (0.000)**	0.014 (0.001)	0.049 (0.001)	0.033 (0.000)	0.016 (0.000)	0.044 (0.001)	0.017 (0.000)	0.104 (0.009)	0.079 (0.007)

Task	ERM	CVaR	Focal	Reweight	US	Reweight (Envs)	US (Envs)	GDRO	IRM
car	**0.019 (0.001)**	**0.017 (0.001)**	0.067 (0.003)	0.030 (0.000)	0.025 (0.001)	0.068 (0.004)	0.029 (0.002)	0.131 (0.025)	0.109 (0.034)
bowl	**0.014 (0.000)**	0.016 (0.001)	0.080 (0.006)	0.026 (0.002)	0.022 (0.001)	0.041 (0.003)	0.028 (0.002)	0.092 (0.019)	0.238 (0.040)
boat	0.003 (0.000)	**0.002 (0.000)**	0.037 (0.001)	0.013 (0.001)	0.005 (0.000)	0.055 (0.002)	0.003 (0.000)	0.085 (0.012)	0.194 (0.109)
fire-hydrant	**0.001 (0.000)**	**0.001 (0.000)**	0.051 (0.002)	0.014 (0.001)	0.001 (0.000)	0.003 (0.000)	0.002 (0.000)	0.020 (0.002)	0.033 (0.027)
airplane	0.001 (0.000)	**0.000 (0.000)**	0.006 (0.000)	0.005 (0.001)	**0.000 (0.000)**	0.001 (0.000)	**0.001 (0.000)**	0.008 (0.001)	0.013 (0.004)
cow	**0.002 (0.000)**	0.002 (0.000)	0.010 (0.000)	0.004 (0.000)	**0.002 (0.000)**	0.003 (0.000)	0.002 (0.000)	0.020 (0.003)	0.014 (0.003)
backpack	**0.023 (0.000)**	0.034 (0.005)	0.099 (0.003)	0.083 (0.001)	0.031 (0.001)	0.025 (0.000)	0.026 (0.001)	0.095 (0.004)	0.120 (0.021)
cup	0.028 (0.000)	0.039 (0.001)	0.027 (0.002)	0.026 (0.002)	**0.018 (0.000)**	0.049 (0.004)	0.022 (0.001)	0.103 (0.024)	0.076 (0.028)
surfboard	**0.003 (0.000)**	**0.003 (0.001)**	0.007 (0.001)	0.005 (0.000)	**0.003 (0.000)**	0.023 (0.001)	**0.003 (0.000)**	0.031 (0.015)	0.043 (0.021)
tie	**0.011 (0.000)**	**0.013 (0.001)**	0.055 (0.003)	0.046 (0.003)	0.012 (0.000)	**0.012 (0.001)**	0.012 (0.000)	0.091 (0.015)	0.279 (0.042)
sports-ball	0.011 (0.000)	0.013 (0.002)	**0.009 (0.001)**	**0.009 (0.001)**	**0.010 (0.001)**	0.107 (0.007)	0.012 (0.001)	0.069 (0.015)	0.085 (0.030)
kite	**0.002 (0.000)**	**0.002 (0.000)**	0.010 (0.001)	0.009 (0.001)	**0.002 (0.000)**	0.002 (0.000)	0.002 (0.000)	0.020 (0.007)	0.019 (0.001)
Average	**0.010 (0.000)**	0.012 (0.000)	0.038 (0.001)	0.022 (0.000)	0.011 (0.000)	0.032 (0.001)	0.012 (0.000)	0.064 (0.004)	0.102 (0.012)

Table 13: Classification error on easy test examples for all 12 NOOCH stress tests, after hyperparameter selection. Bold numbers have overlapping standard deviations. L: NOOCH-CE. R: NOOCH-Gist.

Task	ERM	CVaR	Focal	Reweight	US	Reweight (Envs)	US (Envs)	GDRO	IRM
car	1.072 (0.041)	0.669 (0.064)	0.419 (0.009)	0.673 (0.009)	0.899 (0.113)	0.462 (0.016)	0.862 (0.038)	**0.309 (0.012)**	0.461 (0.068)
bowl	1.385 (0.088)	0.702 (0.005)	**0.415 (0.007)**	0.720 (0.032)	0.940 (0.138)	0.580 (0.015)	0.827 (0.120)	0.480 (0.011)	0.529 (0.030)
boat	1.171 (0.033)	0.817 (0.049)	**0.297 (0.006)**	0.454 (0.012)	0.964 (0.019)	0.395 (0.008)	1.038 (0.065)	**0.285 (0.021)**	0.413 (0.060)
fire-hydrant	1.639 (0.040)	0.996 (0.069)	**0.316 (0.006)**	0.429 (0.005)	1.488 (0.083)	1.090 (0.010)	1.453 (0.218)	0.472 (0.020)	0.437 (0.076)
airplane	0.485 (0.025)	0.420 (0.069)	**0.130 (0.002)**	0.144 (0.002)	0.469 (0.013)	0.318 (0.003)	0.463 (0.033)	**0.150 (0.032)**	0.272 (0.050)
cow	1.272 (0.037)	1.102 (0.064)	**0.290 (0.006)**	0.497 (0.008)	1.219 (0.108)	0.793 (0.014)	1.154 (0.096)	0.410 (0.036)	0.584 (0.027)
backpack	2.035 (0.049)	0.719 (0.003)	**0.506 (0.006)**	**0.492 (0.010)**	1.519 (0.284)	1.225 (0.043)	1.453 (0.124)	**0.511 (0.021)**	**0.523 (0.094)**
cup	1.538 (0.073)	0.698 (0.001)	**0.439 (0.012)**	**0.427 (0.015)**	0.986 (0.043)	0.528 (0.023)	1.001 (0.039)	0.479 (0.009)	0.573 (0.028)
surfboard	1.496 (0.076)	1.163 (0.115)	**0.405 (0.013)**	0.530 (0.006)	1.418 (0.054)	**0.410 (0.016)**	1.300 (0.019)	**0.371 (0.083)**	0.525 (0.066)
tie	3.357 (0.034)	1.447 (0.148)	0.737 (0.014)	1.054 (0.035)	3.158 (0.022)	2.360 (0.106)	3.287 (0.143)	**0.631 (0.049)**	**0.672 (0.001)**
sports-ball	1.991 (0.031)	1.127 (0.062)	0.526 (0.011)	0.663 (0.030)	1.936 (0.321)	**0.366 (0.017)**	1.402 (0.044)	**0.374 (0.062)**	0.469 (0.065)
kite	1.925 (0.014)	1.753 (0.113)	**0.426 (0.033)**	0.516 (0.048)	1.789 (0.082)	1.275 (0.022)	1.798 (0.087)	**0.561 (0.163)**	**0.396 (0.019)**
Average	1.614 (0.014)	0.968 (0.022)	**0.409 (0.004)**	0.550 (0.007)	1.399 (0.041)	0.817 (0.010)	1.336 (0.029)	**0.419 (0.017)**	0.488 (0.016)

Task	ERM	CVaR	Focal	Reweight	US	Reweight (Envs)	US (Envs)	GDRO	IRM
car	1.126 (0.045)	0.690 (0.068)	**0.431 (0.009)**	0.708 (0.011)	0.941 (0.116)	0.511 (0.017)	0.916 (0.040)	**0.456 (0.039)**	**0.487 (0.083)**
bowl	1.486 (0.089)	0.705 (0.005)	**0.435 (0.007)**	0.769 (0.035)	1.002 (0.147)	0.727 (0.036)	0.927 (0.136)	0.533 (0.010)	0.675 (0.010)
boat	1.386 (0.043)	1.004 (0.070)	**0.343 (0.004)**	0.517 (0.019)	1.134 (0.025)	0.608 (0.025)	1.274 (0.082)	0.416 (0.049)	0.616 (0.133)
fire-hydrant	2.300 (0.060)	1.382 (0.085)	**0.404 (0.005)**	0.646 (0.008)	2.066 (0.101)	1.595 (0.004)	2.035 (0.288)	0.707 (0.034)	0.617 (0.100)
airplane	0.771 (0.038)	0.628 (0.079)	**0.190 (0.004)**	0.210 (0.003)	0.763 (0.023)	0.532 (0.006)	0.740 (0.046)	0.247 (0.016)	0.269 (0.035)
cow	1.634 (0.037)	1.339 (0.093)	**0.348 (0.009)**	0.623 (0.011)	1.559 (0.062)	1.137 (0.021)	1.517 (0.135)	0.576 (0.071)	0.669 (0.034)
backpack	2.363 (0.060)	0.725 (0.005)	**0.594 (0.003)**	0.631 (0.009)	1.809 (0.313)	1.458 (0.044)	1.740 (0.159)	0.635 (0.006)	**0.603 (0.077)**
cup	1.898 (0.068)	0.702 (0.002)	0.579 (0.006)	0.590 (0.022)	1.259 (0.041)	**0.572 (0.033)**	1.222 (0.038)	**0.522 (0.023)**	0.633 (0.035)
surfboard	2.045 (0.144)	1.578 (0.184)	**0.490 (0.019)**	0.629 (0.010)	1.852 (0.141)	0.919 (0.052)	1.851 (0.020)	0.764 (0.112)	0.781 (0.028)
tie	2.891 (0.027)	1.374 (0.118)	0.652 (0.009)	0.846 (0.028)	2.678 (0.058)	1.961 (0.094)	2.807 (0.116)	**0.557 (0.032)**	0.669 (0.003)
sports-ball	2.458 (0.013)	1.282 (0.078)	0.601 (0.015)	0.877 (0.034)	2.425 (0.053)	**0.415 (0.026)**	1.780 (0.108)	**0.442 (0.081)**	**0.507 (0.076)**
kite	1.672 (0.039)	1.612 (0.165)	**0.362 (0.032)**	**0.435 (0.047)**	1.633 (0.036)	1.219 (0.032)	1.720 (0.044)	**0.614 (0.222)**	**0.439 (0.083)**
Average	1.836 (0.019)	1.085 (0.028)	**0.452 (0.004)**	0.623 (0.007)	1.594 (0.035)	0.971 (0.011)	1.544 (0.036)	0.539 (0.024)	0.581 (0.020)

Table 14: NLL on hard positive test examples for all 12 NOOCH stress tests, after hyperparameter selection. Bold numbers have overlapping standard deviations. L: NOOCH-CE. R: NOOCH-Gist.

Task	ERM	CVaR	Focal	Reweight	US	Reweight (Envs)	US (Envs)	GDRO	IRM
car	**0.330 (0.019)**	0.444 (0.061)	0.819 (0.008)	0.643 (0.014)	0.491 (0.050)	0.397 (0.016)	**0.341 (0.029)**	0.487 (0.038)	0.471 (0.106)
bowl	**0.241 (0.033)**	0.665 (0.005)	0.890 (0.024)	0.629 (0.031)	0.491 (0.030)	0.377 (0.015)	0.415 (0.031)	0.550 (0.039)	0.510 (0.033)
boat	0.165 (0.007)	**0.158 (0.056)**	0.903 (0.017)	0.682 (0.044)	0.246 (0.016)	0.169 (0.013)	**0.123 (0.020)**	0.340 (0.089)	0.409 (0.117)
fire-hydrant	**0.023 (0.003)**	0.050 (0.013)	0.637 (0.007)	0.290 (0.010)	0.034 (0.006)	0.040 (0.002)	**0.034 (0.010)**	0.178 (0.023)	0.343 (0.108)
airplane	**0.009 (0.001)**	**0.013 (0.003)**	0.165 (0.011)	0.138 (0.012)	0.011 (0.000)	0.017 (0.001)	**0.010 (0.001)**	0.047 (0.012)	0.038 (0.008)
cow	**0.034 (0.001)**	0.054 (0.011)	0.453 (0.005)	0.201 (0.005)	**0.034 (0.002)**	0.042 (0.002)	**0.036 (0.007)**	0.156 (0.033)	0.131 (0.013)
backpack	**0.077 (0.005)**	0.644 (0.007)	0.597 (0.004)	0.617 (0.008)	0.187 (0.029)	0.206 (0.010)	0.153 (0.018)	0.457 (0.005)	0.544 (0.045)
cup	**0.106 (0.004)**	0.670 (0.005)	0.501 (0.006)	0.538 (0.018)	0.247 (0.002)	0.455 (0.029)	0.191 (0.006)	0.597 (0.039)	0.518 (0.042)
surfboard	**0.031 (0.003)**	**0.032 (0.009)**	0.288 (0.009)	0.237 (0.004)	**0.036 (0.003)**	0.128 (0.008)	**0.034 (0.005)**	0.156 (0.050)	0.088 (0.018)
tie	**0.051 (0.002)**	0.201 (0.014)	0.554 (0.027)	0.466 (0.025)	0.074 (0.009)	0.122 (0.009)	0.064 (0.007)	0.551 (0.066)	0.673 (0.004)
sports-ball	**0.057 (0.002)**	0.128 (0.018)	0.419 (0.016)	0.452 (0.017)	0.104 (0.012)	0.425 (0.020)	0.078 (0.002)	0.402 (0.068)	0.481 (0.079)
kite	**0.014 (0.002)**	**0.013 (0.003)**	0.208 (0.017)	0.180 (0.019)	**0.016 (0.002)**	0.024 (0.002)	**0.014 (0.002)**	0.098 (0.032)	0.190 (0.017)
Average	**0.095 (0.003)**	0.256 (0.007)	0.536 (0.004)	0.423 (0.006)	0.164 (0.006)	0.200 (0.004)	0.124 (0.004)	0.335 (0.014)	0.366 (0.018)

Task	ERM	CVaR	Focal	Reweight	US	Reweight (Envs)	US (Envs)	GDRO	IRM
car	**0.321 (0.016)**	0.439 (0.060)	0.784 (0.007)	0.619 (0.021)	0.472 (0.054)	0.475 (0.018)	**0.360 (0.031)**	0.539 (0.075)	0.575 (0.049)
bowl	**0.361 (0.050)**	0.671 (0.004)	1.026 (0.026)	0.876 (0.033)	0.701 (0.036)	0.819 (0.043)	0.722 (0.036)	0.786 (0.016)	0.701 (0.005)
boat	**0.200 (0.010)**	**0.182 (0.056)**	0.971 (0.015)	0.838 (0.051)	0.310 (0.013)	0.299 (0.017)	**0.197 (0.016)**	0.499 (0.107)	0.468 (0.097)
fire-hydrant	**0.044 (0.004)**	0.064 (0.011)	0.721 (0.009)	0.438 (0.012)	0.065 (0.011)	0.092 (0.006)	0.075 (0.019)	0.369 (0.038)	0.465 (0.121)
airplane	**0.011 (0.001)**	**0.016 (0.004)**	0.177 (0.011)	0.153 (0.013)	0.013 (0.001)	0.023 (0.001)	**0.012 (0.001)**	0.098 (0.011)	0.230 (0.039)
cow	**0.056 (0.001)**	**0.064 (0.009)**	0.618 (0.009)	0.331 (0.005)	0.074 (0.005)	0.102 (0.004)	**0.066 (0.014)**	0.309 (0.046)	0.349 (0.082)
backpack	**0.103 (0.005)**	0.655 (0.004)	0.663 (0.005)	0.754 (0.007)	0.259 (0.034)	0.261 (0.013)	0.207 (0.014)	0.602 (0.019)	0.628 (0.048)
cup	**0.166 (0.008)**	0.679 (0.002)	0.677 (0.007)	0.854 (0.022)	0.391 (0.002)	0.681 (0.031)	0.314 (0.012)	0.693 (0.010)	0.632 (0.043)
surfboard	**0.038 (0.004)**	**0.037 (0.010)**	0.323 (0.008)	0.307 (0.004)	0.058 (0.001)	0.242 (0.013)	**0.042 (0.006)**	0.268 (0.079)	0.323 (0.065)
tie	**0.047 (0.002)**	0.193 (0.021)	0.462 (0.031)	0.395 (0.020)	0.070 (0.008)	0.110 (0.009)	0.062 (0.007)	0.517 (0.050)	0.650 (0.008)
sports-ball	**0.066 (0.003)**	0.130 (0.020)	0.449 (0.015)	0.516 (0.020)	0.079 (0.007)	0.502 (0.030)	0.096 (0.005)	0.441 (0.073)	0.484 (0.038)
kite	**0.015 (0.002)**	**0.013 (0.002)**	0.221 (0.018)	0.194 (0.021)	0.017 (0.002)	0.028 (0.003)	**0.016 (0.002)**	0.128 (0.042)	0.211 (0.018)
Average	**0.119 (0.005)**	0.262 (0.007)	0.591 (0.005)	0.523 (0.007)	0.209 (0.006)	0.303 (0.006)	0.181 (0.005)	0.438 (0.016)	0.476 (0.018)

Table 15: NLL on hard negative test examples for all 12 NOOCH stress tests, after hyperparameter selection. Bold numbers have overlapping standard deviations. L: NOOCH-CE. R: NOOCH-Gist.

Task	ERM	CVaR	Focal	Reweight	US	Reweight (Envs)	US (Envs)	GDRO	IRM
car	**0.059 (0.002)**	0.286 (0.057)	0.305 (0.007)	0.095 (0.001)	0.075 (0.008)	0.249 (0.009)	0.092 (0.004)	0.782 (0.084)	0.222 (0.028)
bowl	**0.051 (0.002)**	0.600 (0.003)	0.320 (0.011)	0.096 (0.004)	0.079 (0.005)	0.224 (0.003)	0.108 (0.008)	0.352 (0.067)	0.223 (0.011)
boat	**0.015 (0.000)**	0.033 (0.003)	0.293 (0.006)	0.057 (0.004)	0.020 (0.001)	0.255 (0.006)	0.021 (0.002)	0.295 (0.009)	0.306 (0.098)
fire-hydrant	**0.012 (0.000)**	0.023 (0.006)	0.289 (0.005)	0.059 (0.001)	0.013 (0.001)	0.028 (0.001)	0.016 (0.002)	0.140 (0.021)	0.161 (0.062)
airplane	**0.006 (0.000)**	**0.007 (0.001)**	0.056 (0.004)	0.035 (0.003)	**0.006 (0.000)**	0.010 (0.001)	**0.006 (0.000)**	0.101 (0.025)	0.013 (0.003)
cow	**0.011 (0.000)**	0.022 (0.005)	0.192 (0.003)	0.035 (0.000)	**0.012 (0.001)**	0.025 (0.000)	0.013 (0.001)	0.118 (0.020)	0.038 (0.002)
backpack	**0.086 (0.001)**	0.636 (0.007)	0.380 (0.002)	0.281 (0.003)	0.123 (0.003)	0.113 (0.005)	0.109 (0.004)	0.363 (0.032)	0.359 (0.026)
cup	**0.103 (0.002)**	0.651 (0.006)	0.268 (0.004)	0.229 (0.007)	0.123 (0.001)	0.297 (0.009)	0.125 (0.003)	0.398 (0.037)	0.329 (0.021)
surfboard	**0.014 (0.001)**	0.016 (0.001)	0.133 (0.006)	0.041 (0.001)	**0.014 (0.001)**	0.144 (0.012)	**0.015 (0.001)**	0.133 (0.043)	0.060 (0.013)
tie	**0.035 (0.001)**	0.167 (0.029)	0.327 (0.019)	0.121 (0.006)	0.039 (0.001)	0.039 (0.002)	0.040 (0.001)	0.275 (0.049)	0.609 (0.015)
sports-ball	0.033 (0.000)	0.086 (0.019)	0.238 (0.013)	0.081 (0.006)	**0.032 (0.001)**	0.430 (0.025)	0.041 (0.001)	0.264 (0.040)	0.305 (0.043)
kite	0.008 (0.000)	**0.008 (0.001)**	0.064 (0.006)	0.041 (0.005)	**0.007 (0.000)**	0.011 (0.000)	0.009 (0.001)	0.082 (0.024)	0.075 (0.007)
Average	**0.036 (0.000)**	0.211 (0.006)	0.239 (0.002)	0.098 (0.001)	0.045 (0.001)	0.152 (0.003)	0.050 (0.001)	0.275 (0.012)	0.225 (0.011)

Task	ERM	CVaR	Focal	Reweight	US	Reweight (Envs)	US (Envs)	GDRO	IRM
car	**0.056 (0.002)**	0.285 (0.057)	0.306 (0.007)	0.093 (0.001)	0.073 (0.008)	0.242 (0.009)	0.087 (0.004)	0.408 (0.074)	0.313 (0.074)
bowl	**0.041 (0.001)**	0.599 (0.003)	0.312 (0.011)	0.083 (0.004)	0.066 (0.005)	0.144 (0.008)	0.089 (0.009)	0.338 (0.068)	0.495 (0.108)
boat	**0.010 (0.000)**	0.029 (0.003)	0.290 (0.006)	0.051 (0.004)	0.016 (0.001)	0.248 (0.006)	0.012 (0.001)	0.288 (0.009)	0.453 (0.143)
fire-hydrant	**0.005 (0.000)**	0.019 (0.006)	0.282 (0.005)	0.047 (0.001)	0.006 (0.001)	0.020 (0.001)	0.008 (0.002)	0.125 (0.021)	0.151 (0.062)
airplane	**0.002 (0.000)**	0.004 (0.001)	0.053 (0.004)	0.032 (0.003)	**0.002 (0.000)**	0.006 (0.001)	**0.002 (0.000)**	0.053 (0.008)	0.067 (0.017)
cow	**0.005 (0.000)**	0.019 (0.005)	0.174 (0.004)	0.020 (0.001)	0.006 (0.000)	0.016 (0.000)	0.007 (0.001)	0.101 (0.022)	0.073 (0.026)
backpack	**0.065 (0.001)**	0.632 (0.008)	0.398 (0.004)	0.226 (0.003)	0.087 (0.001)	0.085 (0.003)	0.079 (0.006)	0.306 (0.024)	0.326 (0.020)
cup	0.063 (0.001)	0.647 (0.007)	0.255 (0.005)	0.095 (0.006)	**0.053 (0.001)**	0.204 (0.008)	0.066 (0.001)	0.358 (0.045)	0.280 (0.022)
surfboard	0.009 (0.000)	0.012 (0.001)	0.078 (0.001)	0.026 (0.001)	**0.007 (0.000)**	0.118 (0.011)	0.009 (0.001)	0.108 (0.038)	0.159 (0.062)
tie	**0.040 (0.001)**	0.170 (0.027)	0.296 (0.025)	0.147 (0.008)	0.044 (0.000)	0.046 (0.002)	0.044 (0.001)	0.287 (0.050)	0.617 (0.012)
sports-ball	**0.026 (0.000)**	0.084 (0.019)	0.226 (0.013)	0.056 (0.004)	**0.025 (0.001)**	0.401 (0.022)	0.032 (0.002)	0.249 (0.037)	0.303 (0.053)
kite	0.009 (0.000)	**0.008 (0.001)**	0.059 (0.005)	0.036 (0.004)	**0.007 (0.000)**	0.010 (0.000)	**0.008 (0.001)**	0.071 (0.020)	0.067 (0.006)
Average	**0.028 (0.000)**	0.209 (0.006)	0.227 (0.003)	0.076 (0.001)	0.033 (0.001)	0.128 (0.002)	0.037 (0.001)	0.224 (0.012)	0.275 (0.019)

Table 16: NLL on easy test examples for all 12 NOOCH stress tests, after hyperparameter selection. Bold numbers have overlapping standard deviations. L: NOOCH-CE. R: NOOCH-Gist.

Task	ERM	CVaR	Focal	Reweight	US	Reweight (Envs)	US (Envs)	GDRO	IRM
car	0.017 (0.001)	**0.005 (0.002)**	**0.004 (0.000)**	0.012 (0.000)	0.015 (0.002)	**0.004 (0.000)**	0.014 (0.001)	0.008 (0.001)	**0.004 (0.001)**
bowl	0.013 (0.001)	0.011 (0.000)	0.007 (0.001)	0.009 (0.000)	0.010 (0.002)	**0.004 (0.000)**	0.009 (0.002)	0.008 (0.002)	**0.005 (0.001)**
boat	0.004 (0.000)	0.002 (0.000)	0.007 (0.001)	0.004 (0.000)	0.004 (0.000)	**0.001 (0.000)**	0.004 (0.000)	0.002 (0.000)	0.005 (0.001)
fire-hydrant	0.002 (0.000)	**0.001 (0.000)**	0.003 (0.000)	**0.001 (0.000)**	0.002 (0.000)	**0.001 (0.000)**	0.002 (0.000)	0.004 (0.001)	0.007 (0.002)
airplane	0.001 (0.000)	**0.000 (0.000)**	0.010 (0.001)	0.005 (0.001)	0.001 (0.000)	**0.000 (0.000)**	0.001 (0.000)	0.001 (0.001)	0.001 (0.000)
cow	0.002 (0.000)	**0.002 (0.001)**	0.009 (0.000)	0.001 (0.000)	0.002 (0.000)	**0.001 (0.000)**	0.002 (0.000)	**0.001 (0.001)**	**0.001 (0.000)**
backpack	0.006 (0.001)	0.108 (0.003)	0.009 (0.001)	0.022 (0.001)	0.008 (0.003)	**0.002 (0.001)**	0.004 (0.000)	0.007 (0.003)	0.011 (0.003)
cup	0.009 (0.001)	0.113 (0.002)	0.012 (0.001)	0.018 (0.001)	0.009 (0.002)	0.018 (0.002)	**0.003 (0.000)**	0.021 (0.011)	0.034 (0.016)
surfboard	0.002 (0.000)	**0.001 (0.000)**	0.013 (0.001)	0.004 (0.000)	0.002 (0.000)	0.006 (0.001)	0.001 (0.000)	0.004 (0.001)	0.005 (0.001)
tie	**0.002 (0.000)**	0.039 (0.004)	0.027 (0.002)	0.010 (0.001)	0.003 (0.000)	**0.003 (0.000)**	**0.003 (0.000)**	0.009 (0.002)	0.023 (0.007)
sports-ball	**0.002 (0.000)**	0.023 (0.006)	0.028 (0.002)	0.017 (0.001)	0.006 (0.000)	0.040 (0.004)	**0.002 (0.001)**	0.021 (0.004)	0.035 (0.005)
kite	**0.001 (0.000)**	**0.001 (0.000)**	0.012 (0.001)	0.002 (0.000)	**0.000 (0.000)**	**0.001 (0.000)**	**0.000 (0.000)**	0.002 (0.001)	0.005 (0.002)
Average	0.005 (0.000)	0.025 (0.001)	0.012 (0.000)	0.009 (0.000)	0.005 (0.000)	0.007 (0.000)	**0.004 (0.000)**	0.007 (0.001)	0.011 (0.002)

Task	ERM	CVaR	Focal	Reweight	US	Reweight (Envs)	US (Envs)	GDRO	IRM
car	0.019 (0.001)	**0.006 (0.002)**	**0.003 (0.000)**	0.013 (0.001)	0.016 (0.002)	0.005 (0.000)	0.015 (0.001)	0.006 (0.003)	**0.003 (0.001)**
bowl	0.015 (0.001)	0.008 (0.001)	0.014 (0.001)	0.014 (0.000)	0.014 (0.001)	0.013 (0.000)	0.014 (0.002)	**0.005 (0.001)**	**0.004 (0.001)**
boat	0.005 (0.000)	**0.003 (0.000)**	0.009 (0.001)	0.006 (0.000)	0.005 (0.000)	**0.002 (0.000)**	0.004 (0.000)	**0.002 (0.001)**	**0.002 (0.000)**
fire-hydrant	0.003 (0.000)	**0.001 (0.000)**	0.004 (0.001)	0.004 (0.001)	0.003 (0.000)	0.002 (0.000)	0.003 (0.000)	0.005 (0.002)	0.006 (0.002)
airplane	0.001 (0.000)	0.001 (0.000)	0.009 (0.001)	0.004 (0.001)	0.001 (0.000)	**0.000 (0.000)**	0.001 (0.000)	0.003 (0.001)	0.008 (0.004)
cow	0.002 (0.000)	**0.002 (0.000)**	0.006 (0.000)	0.004 (0.000)	0.003 (0.000)	0.002 (0.000)	0.003 (0.000)	**0.003 (0.002)**	0.003 (0.001)
backpack	0.008 (0.000)	0.106 (0.003)	**0.005 (0.001)**	0.038 (0.001)	0.014 (0.002)	**0.004 (0.000)**	0.007 (0.002)	0.011 (0.003)	**0.012 (0.009)**
cup	**0.008 (0.002)**	0.105 (0.003)	**0.008 (0.000)**	0.046 (0.002)	0.015 (0.001)	0.026 (0.003)	**0.007 (0.002)**	0.012 (0.002)	0.016 (0.006)
surfboard	0.002 (0.000)	**0.001 (0.000)**	0.006 (0.001)	0.005 (0.000)	0.003 (0.000)	0.005 (0.001)	0.002 (0.000)	0.003 (0.000)	0.017 (0.003)
tie	0.002 (0.000)	0.037 (0.005)	0.028 (0.002)	0.006 (0.001)	0.003 (0.000)	**0.001 (0.000)**	0.003 (0.000)	0.013 (0.003)	0.022 (0.007)
sports-ball	0.004 (0.001)	0.023 (0.007)	0.020 (0.002)	0.023 (0.000)	**0.002 (0.000)**	0.035 (0.003)	**0.002 (0.000)**	0.014 (0.003)	0.025 (0.005)
kite	**0.001 (0.000)**	**0.001 (0.000)**	0.011 (0.001)	0.002 (0.000)	**0.000 (0.000)**	**0.001 (0.000)**	**0.000 (0.000)**	0.002 (0.001)	0.005 (0.002)
Average	0.006 (0.000)	0.024 (0.001)	0.010 (0.000)	0.014 (0.000)	0.007 (0.000)	0.008 (0.000)	**0.005 (0.000)**	0.007 (0.001)	0.010 (0.001)

Table 17: ECE on hard test examples for all 12 NOOCH stress tests, after hyperparameter selection. Bold numbers have overlapping standard deviations. L: NOOCH-CE. R: NOOCH-Gist.

Task	ERM	CVaR	Focal	Reweight	US	Reweight (Envs)	US (Envs)	GDRO	IRM
car	**0.008 (0.001)**	0.198 (0.039)	0.147 (0.005)	0.016 (0.002)	**0.009 (0.006)**	0.069 (0.004)	**0.011 (0.008)**	0.116 (0.082)	0.054 (0.013)
bowl	**0.007 (0.001)**	0.395 (0.001)	0.130 (0.002)	0.012 (0.001)	**0.009 (0.005)**	0.047 (0.003)	**0.012 (0.008)**	0.125 (0.029)	0.057 (0.016)
boat	**0.002 (0.000)**	0.022 (0.003)	0.185 (0.003)	0.014 (0.002)	0.003 (0.000)	0.119 (0.002)	0.004 (0.001)	0.093 (0.028)	0.119 (0.007)
fire-hydrant	**0.001 (0.000)**	0.016 (0.005)	0.151 (0.003)	0.010 (0.001)	0.002 (0.000)	0.011 (0.000)	**0.003 (0.002)**	0.069 (0.015)	0.069 (0.009)
airplane	**0.000 (0.000)**	0.002 (0.000)	0.032 (0.002)	0.015 (0.002)	**0.000 (0.000)**	0.004 (0.000)	**0.000 (0.000)**	0.020 (0.004)	0.004 (0.001)
cow	**0.000 (0.000)**	0.013 (0.005)	0.119 (0.003)	0.007 (0.000)	0.001 (0.000)	0.007 (0.000)	**0.001 (0.000)**	0.044 (0.016)	0.013 (0.001)
backpack	**0.003 (0.000)**	0.308 (0.001)	0.098 (0.002)	0.010 (0.002)	**0.005 (0.002)**	0.016 (0.002)	**0.008 (0.007)**	0.065 (0.009)	0.042 (0.026)
cup	**0.003 (0.001)**	0.295 (0.005)	0.048 (0.005)	**0.004 (0.001)**	**0.003 (0.001)**	0.027 (0.003)	**0.006 (0.004)**	0.080 (0.015)	0.068 (0.002)
surfboard	**0.001 (0.000)**	0.004 (0.001)	0.081 (0.005)	0.009 (0.001)	**0.001 (0.000)**	0.052 (0.009)	**0.001 (0.000)**	0.023 (0.003)	0.018 (0.002)
tie	**0.002 (0.000)**	0.097 (0.018)	0.167 (0.013)	0.033 (0.003)	0.002 (0.000)	0.008 (0.001)	**0.001 (0.000)**	0.077 (0.023)	0.131 (0.036)
sports-ball	**0.002 (0.000)**	0.042 (0.013)	0.127 (0.007)	0.018 (0.001)	**0.002 (0.001)**	0.112 (0.007)	0.004 (0.001)	0.053 (0.019)	0.071 (0.007)
kite	**0.000 (0.000)**	0.001 (0.000)	0.025 (0.002)	0.008 (0.001)	**0.000 (0.000)**	0.002 (0.000)	**0.001 (0.000)**	0.013 (0.004)	0.015 (0.005)
Average	**0.003 (0.000)**	0.116 (0.004)	0.109 (0.001)	0.013 (0.000)	**0.003 (0.001)**	0.039 (0.001)	0.004 (0.001)	0.065 (0.008)	0.055 (0.004)

Task	ERM	CVaR	Focal	Reweight	US	Reweight (Envs)	US (Envs)	GDRO	IRM
car	**0.010 (0.001)**	0.199 (0.039)	0.147 (0.005)	0.017 (0.002)	**0.010 (0.006)**	0.070 (0.004)	**0.012 (0.008)**	0.125 (0.053)	0.076 (0.026)
bowl	**0.009 (0.001)**	0.398 (0.001)	0.137 (0.002)	0.017 (0.001)	**0.012 (0.005)**	0.044 (0.004)	**0.015 (0.011)**	0.137 (0.030)	0.113 (0.086)
boat	**0.003 (0.000)**	0.023 (0.003)	0.187 (0.003)	0.017 (0.002)	0.004 (0.000)	0.120 (0.002)	0.004 (0.001)	0.095 (0.028)	0.079 (0.054)
fire-hydrant	**0.002 (0.000)**	0.016 (0.006)	0.157 (0.004)	0.014 (0.001)	**0.003 (0.000)**	0.011 (0.000)	**0.004 (0.001)**	0.069 (0.015)	0.072 (0.010)
airplane	**0.001 (0.000)**	0.002 (0.000)	0.034 (0.002)	0.016 (0.002)	**0.001 (0.000)**	0.004 (0.000)	**0.001 (0.000)**	0.028 (0.005)	0.032 (0.014)
cow	**0.001 (0.000)**	0.013 (0.005)	0.126 (0.003)	0.009 (0.000)	**0.001 (0.000)**	0.008 (0.000)	**0.001 (0.000)**	0.046 (0.015)	0.037 (0.018)
backpack	**0.007 (0.000)**	0.310 (0.002)	0.142 (0.002)	0.026 (0.001)	**0.006 (0.003)**	0.021 (0.002)	**0.012 (0.007)**	0.079 (0.013)	0.059 (0.022)
cup	**0.007 (0.001)**	0.303 (0.002)	0.128 (0.003)	0.027 (0.002)	**0.007 (0.002)**	0.065 (0.004)	**0.009 (0.003)**	0.110 (0.022)	0.091 (0.018)
surfboard	**0.001 (0.000)**	0.004 (0.001)	0.051 (0.001)	0.010 (0.001)	**0.001 (0.000)**	0.053 (0.008)	**0.001 (0.000)**	0.025 (0.002)	0.057 (0.017)
tie	**0.001 (0.000)**	0.098 (0.016)	0.128 (0.012)	0.028 (0.002)	0.002 (0.000)	0.009 (0.001)	**0.001 (0.000)**	0.068 (0.022)	0.120 (0.040)
sports-ball	**0.003 (0.000)**	0.044 (0.013)	0.135 (0.007)	0.025 (0.002)	**0.003 (0.000)**	0.117 (0.008)	0.005 (0.001)	0.062 (0.021)	0.080 (0.007)
kite	**0.000 (0.000)**	0.001 (0.000)	0.026 (0.002)	0.010 (0.001)	**0.000 (0.000)**	0.002 (0.000)	**0.000 (0.000)**	0.013 (0.004)	0.014 (0.005)
Average	**0.004 (0.000)**	0.118 (0.004)	0.116 (0.001)	0.018 (0.000)	**0.004 (0.001)**	0.044 (0.001)	0.006 (0.001)	0.071 (0.007)	0.069 (0.010)

Table 18: ECE on easy test examples for all 12 NOOCH stress tests, after hyperparameter selection. Bold numbers have overlapping standard deviations. L: NOOCH-CE. R: NOOCH-Gist.

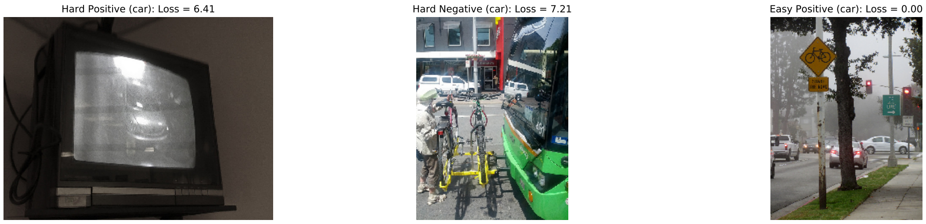
Hard Positive (car): Loss = 6.41 Hard Negative (car): Loss = 7.21 Easy Positive (car): Loss = 0.00

Figure 15: Examples from the `car` task.

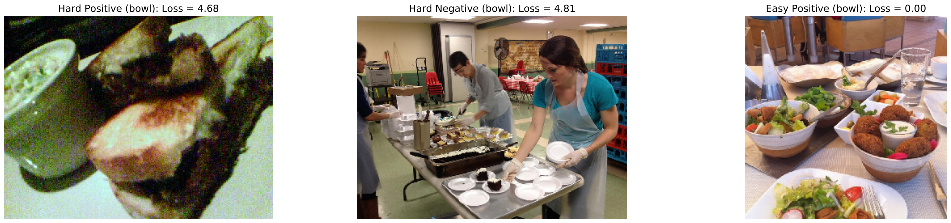
Hard Positive (bowl): Loss = 4.68 Hard Negative (bowl): Loss = 4.81 Easy Positive (bowl): Loss = 0.00

Figure 16: Examples from the `bowl` task.

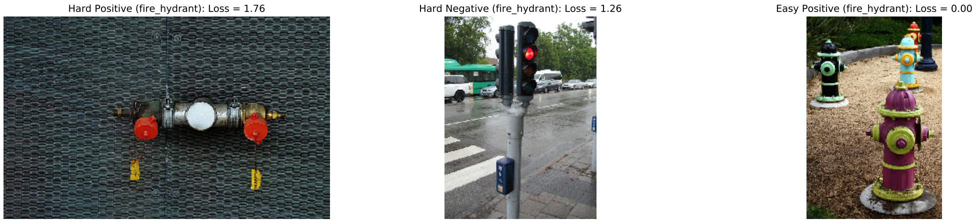
Hard Positive (fire_hydrant): Loss = 1.76 Hard Negative (fire_hydrant): Loss = 1.26 Easy Positive (fire_hydrant): Loss = 0.00

Figure 17: Examples from the `fire_hydrant` task.

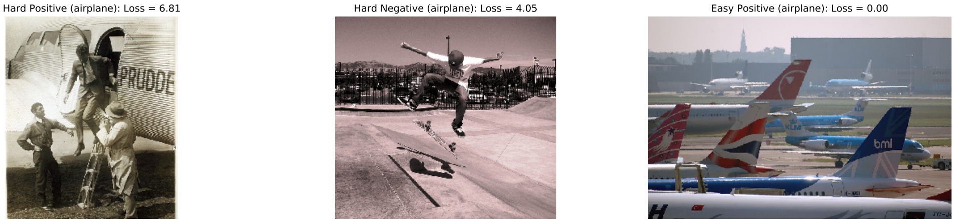
Hard Positive (airplane): Loss = 6.81 Hard Negative (airplane): Loss = 4.05 Easy Positive (airplane): Loss = 0.00

Figure 18: Examples from the `airplane` task.

D Qualitative Analysis

D.1 Examples

Here we show examples from NOOCH-CE: from left to right we show a hard positive, hard negative, and easy positive for each of the 12 tasks. We show the loss for a sample ERM model in the title of the caption. We choose these examples by finding the highest (lowest) loss ERM examples for each hard (easy) category — additionally, for ease of viewing, for hard positives, we choose examples for which the target object's area is larger than average for positive examples of that class. We note that label noise can be a problem in COCO, as in any dataset, and finding the highest loss examples can occasionally surface a mislabelled datapoint (see Fig. 23, the Hard Negative in the `tie` class, where a tie is present but is not labelled). Along with subjectivity in labels, mislabelling is an occasional hazard of any dataset requiring mass labelling — fortunately, we did not find too many occurrences of mislabelled points when examining hard positives and negatives in NOOCH.

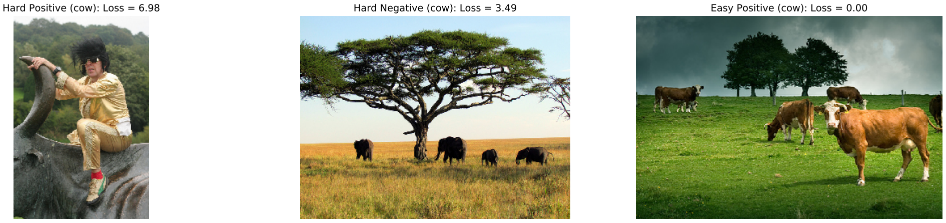
Hard Positive (cow): Loss = 6.98
Hard Negative (cow): Loss = 3.49
Easy Positive (cow): Loss = 0.00

Figure 19: Examples from the `cow` task.

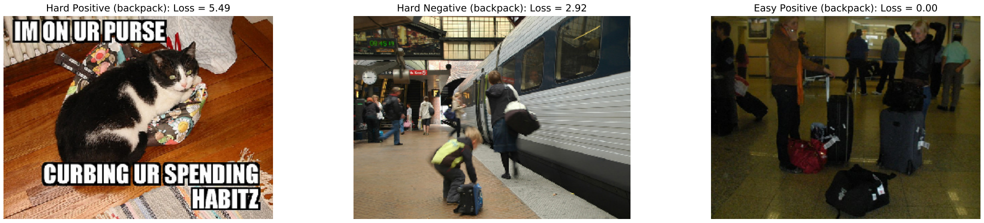
Hard Positive (backpack): Loss = 5.49
IM ON UR PURSE
CURBING UR SPENDING HABITZ
Hard Negative (backpack): Loss = 2.92
Easy Positive (backpack): Loss = 0.00

Figure 20: Examples from the `backpack` task.

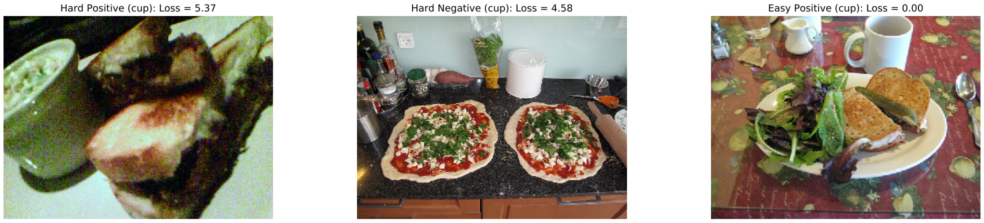
Hard Positive (cup): Loss = 5.37
Hard Negative (cup): Loss = 4.58
Easy Positive (cup): Loss = 0.00

Figure 21: Examples from the `cup` task.

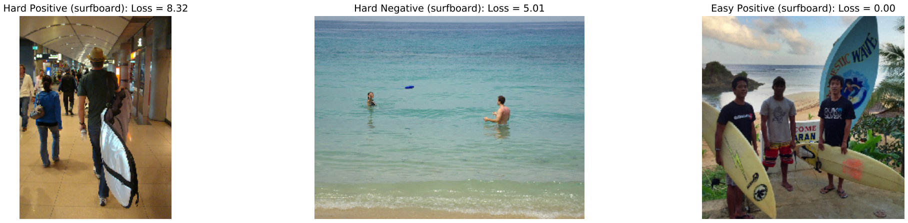
Hard Positive (surfboard): Loss = 8.32
Hard Negative (surfboard): Loss = 5.01
Easy Positive (surfboard): Loss = 0.00

Figure 22: Examples from the `surfboard` task.

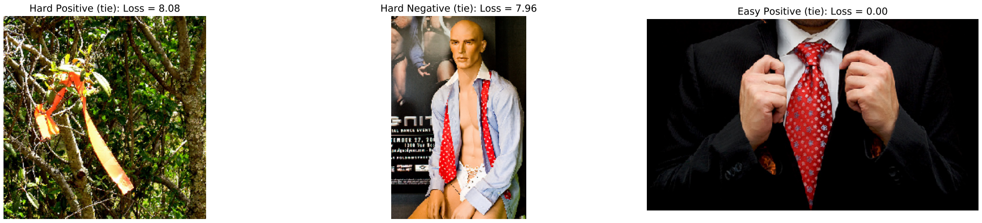
Hard Positive (tie): Loss = 8.08
Hard Negative (tie): Loss = 7.96
Easy Positive (tie): Loss = 0.00

Figure 23: Examples from the `tie` task.

D.2 Analyzing GDRO's Errors

In Fig. 7, we show the images with the largest gaps in NLL on the `car` task among two groups: hard positives and hard negatives (in NOOCH-CE), where GDRO is better than ERM. Here, in Fig. 26 we look at the images where GDRO most underperforms ERM. These help us consider some pitfalls

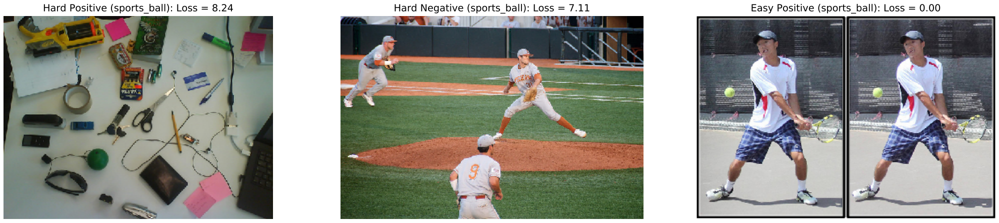
Hard Positive (sports_ball): Loss = 8.24
Hard Negative (sports_ball): Loss = 7.11
Easy Positive (sports_ball): Loss = 0.00

Figure 24: Examples from the `sports_ball` task.

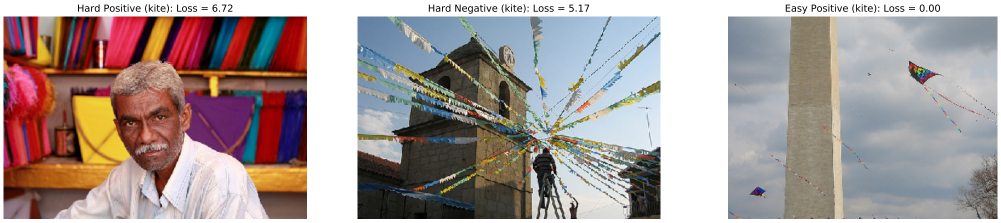
Hard Positive (kite): Loss = 6.72
Hard Negative (kite): Loss = 5.17
Easy Positive (kite): Loss = 0.00

Figure 25: Examples from the `kite` task.

of the GDRO objective, which incentivizes performance on two small subgroups: "cars without roads" and "roads without cars". GDRO's worst hard positive is reminiscent of a "road without a car": the background (labelled "pavement") looks road-like but the focal object is not a prototypical car (the car is in the top left, background). We hypothesize that GDRO's incentive to perform well on "roads without cars" pushed its prediction negative, whereas ERM was able to take advantage of context and correctly output a positive. GDRO's worst hard negative tells an opposite story: the snowy background looks *less* road-like, but the stop sign and black objects in the snow suggest a car may be present. We hypothesize that GDRO's incentive to perform well on "cars without roads", may have pushed its prediction wrongly *positive* in this instance.

Qualitatively, it appears that GDRO uses context in potentially more unintuitive ways than ERM does. While we hope that robust methods will be more interpretable since they may ignore irrelevant signals in the training data [22, 43, 51], these two goals may sometimes be at odds. Indeed, we observe that robust objectives may *unnecessarily* encourage predictions which diverge from the context.

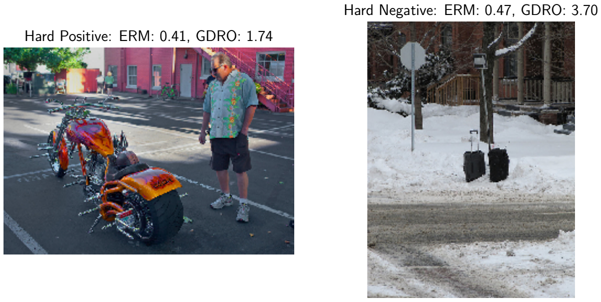
Hard Positive: ERM: 0.41, GDRO: 1.74
Hard Negative: ERM: 0.47, GDRO: 3.70

Figure 26: Test set examples with largest difference in NLL between ERM and GDRO methods, where ERM is better, among (L) hard positives and (R) hard negatives from the NOOCH-CE `car` category. Titles show NLL for each method on that image.