In-context Learning in Presence of Spurious Correlations

Anonymous authors
Paper under double-blind review

Abstract

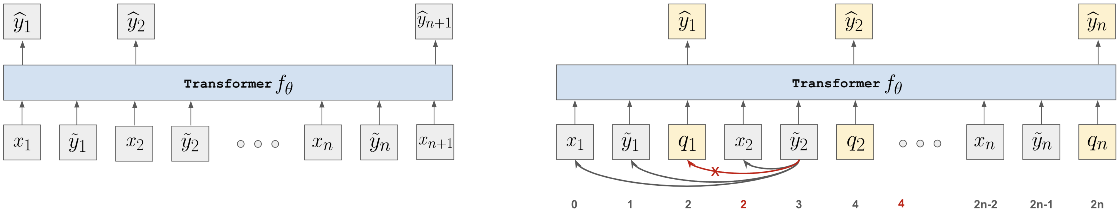
Large language models exhibit a remarkable capacity for in-context learning, where they learn to solve tasks given a few examples. Recent work has shown that transformers can be trained to perform simple regression tasks in-context. This work explores the possibility of training an in-context learner for classification tasks involving spurious features. We find that the conventional approach of training in-context learners is susceptible to spurious features. Moreover, when the meta-training dataset includes instances of only one task, the conventional approach leads to in-weights learning and fails to produce a model that leverages context for predictions. Based on these observations, we propose a novel technique to train such a learner for a given classification task. Remarkably, this in-context learner matches and sometimes outperforms strong methods like ERM and GroupDRO. However, unlike these algorithms, it does not generalize well to other tasks. We show that it is possible to obtain an in-context learner that generalizes to unseen tasks by training on a diverse dataset of synthetic in-context learning instances.

1 Introduction

Large language models, such as GPT-3, have the ability of in-context learning (ICL), wherein they learn to solve a task given a few examples in the context (Brown et al., 2020). The most significant aspect of in-context learning is that the learning occurs during the forward pass on the context and query, without updating network parameters. In order to study in-context learning in isolation, a number of studies considered training transformers (Vaswani et al., 2017) from scratch to solve simple learning tasks in-context. In particular, Garg et al. (2022) show empirically that transformers can be trained to perform in-context learning of simple regression functions, such as dense or sparse linear functions, two-layer ReLU neural networks, and small decision trees.

Training on ICL instances can be seen as an instance of meta-learning (Schmidhuber, 1987; Naik & Mammone, 1992; Thrun & Pratt, 1998), where the goal is to learn a learning algorithm. What exact algorithm is learned when training transformers on ICL instances is an open problem. Akyürek et al. (2023) and Von Oswald et al. (2023) show that transformers can implement a single gradient descent step of ordinary least squares and update the closed-form solution of ridge regression when a new example is added. Additionally, they provide evidence that transformers trained on ICL instances of linear regression learn algorithms that closely match predictions of the known algorithms, such as gradient descent on the ordinary least squares objective and ridge regression. However, there is evidence that the learned algorithm may vary with model scale, depth, and pretraining task diversity (Akyürek et al., 2023; Raventós et al., 2024; Goddard et al., 2025). In particular, Raventós et al. (2024) demonstrate that in the setting of in-context learning of linear regression tasks with insufficient pretraining task diversity, the learned algorithm behaves like a Bayesian estimator with the pretraining task distribution as the prior, and hence fails to generalize well to unseen tasks. Yadlowsky et al. (2023) show that when trained on ICL instances where the regression function belongs to a union of distinct function classes, the learned algorithm fails to generalize beyond the pretraining function classes. Ahuja & Lopez-Paz (2023) show that in-context learning ability diminishes under strong distribution shifts.

In this work, we explore the limits of in-context learning further by studying it in a more challenging setting. In particular, motivated by results indicating that in-context learning is susceptible to shortcuts and spurious correlations (Tang et al., 2023; Zhou et al., 2024; Song et al., 2024), we pose the following question:



(a) Naive approach

(b) Proposed approach

Figure 1: In-context learning transformer architectures of the naive and proposed approaches. The proposed approach allows arbitrary query tokens after each learning example. Token positions and the attention mask are modified so that these intermediate queries have no effect on other tokens.

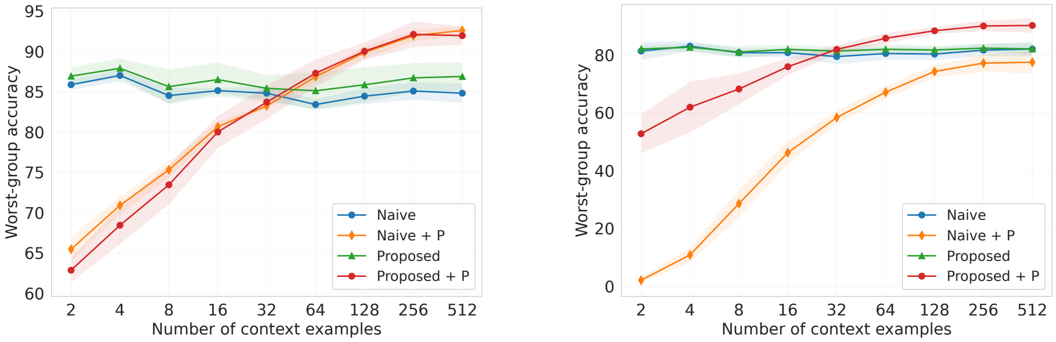
Can we obtain an effective classification algorithm that is robust to spurious features by training an in-context learner on a suitable meta-training dataset, rather than designing the learning algorithm manually?

To address this question in isolation, we consider visual classification tasks where some features are *spuriously correlated* with the label. Such features are predictive of the label but are not causally related to it, due to which their correlation might not hold at test time. A prominent example is the cow vs. camel classification task, where the background often correlates with the label, since cows are typically photographed in pastures, while camels are typically photographed in deserts (Beery et al., 2018). It is well known that neural networks trained with gradient-based methods to minimize empirical risk can exploit spurious features, causing performance degradation under distribution shifts affecting these correlations (Torralba & Efros, 2011; Ribeiro et al., 2016; Gururangan et al., 2018; Zech et al., 2018; McCoy et al., 2019; Geirhos et al., 2019; 2020; Xiao et al., 2021).

We start our analysis in the standard setting of having a single classification task with spurious features. We consider the conventional approach of obtaining an in-context learner, wherein a transformer is trained on sequences of form $(x_1, y_1, \ldots, x_k, y_k, x_{k+1})$ to predict the label y_{k+1} of the query example x_{k+1}. We find that this conventional approach leads to *in-weights learning* (Chan et al., 2022), wherein models perform classification *ignoring the context*, essentially memorizing the task. Furthermore, these models lack robustness to changes in the correlation between the label and spurious features. In particular, we observe a significant performance drop when the query follows a distribution in which the label and spurious feature correlation is zero. We propose an effective approach to reduce in-weights learning and improve in-context learning. Namely, we find that in-weights learning can be greatly mitigated by randomly permuting input embedding dimensions for each training sequence. To address the issue of spurious features, we propose a novel way of forming ICL instances and a suitable transformer architecture, which work together to simulate distribution shift with respect to spurious features in the context. Overall, our proposed techniques lead to strong in-context learners that outperform established methods such as 1-NN, empirical risk minimization (ERM), and GroupDRO (Sagawa* et al., 2020), suggesting that the in-context learner implements a more specialized algorithm.

Despite being trained on instances of a single task, the learned algorithm generalizes to other tasks *without* spurious features. However, it fails to generalize to unseen tasks with spurious features. For this reason, we next explore training an in-context learner that generalizes to unseen tasks with spurious features. We create a dataset of in-context learning instances for various binary classification tasks with varying spurious features. We demonstrate the efficacy of the proposed techniques on this dataset too and find that it can be improved further by passing spurious feature annotations as input and injecting occasional queries requesting the label of a preceding context example to promote learning induction heads. The resulting model generalizes perfectly to unseen tasks, as long as the data generating process is similar. However, generalization to unseen tasks with possibly different data generating process depends on the severity of the challenge posed by spurious features, indicating that the learned algorithm is more brittle to severe distribution shifts than conventional algorithms. The source code for reproducing our experiments is available at `[anonymized]`.

We summarize our main contributions as follows.



(a) `Waterbirds`

(b) `Waterbirds-severe`

Figure 2: Worst-group test accuracy on `Waterbirds` and `Waterbirds-severe` as a function of context size for the naive and proposed methods with or without permuting input dimensions. Shaded regions show standard deviation across 5 training runs.

(i) We show that the conventional approach of training an in-context learner is susceptible to presence of spurious features and also leads to in-weights learning in case of a single task.

(ii) We propose a suite of novel techniques of forming in-context training data to reduce in-weights learning and increase robustness to spurious features, leading to in-context learners that outperform established learning algorithms.

(iii) We demonstrate that it is possible to obtain more general-purpose robust in-context learners by training on a diverse set of synthetic classification tasks involving spurious features.

2 In-context learning based on a single task

We start by considering the common setting of having a single classification task with spurious features. For simplicity, we focus on label-balanced binary classification tasks in presence of a single binary spurious feature, although what follows next applies to label-imbalanced multiclass settings as well. Let $\mathcal{D}_{\text{train}}$ be a set of training examples for the task, where each example is a triplet (x, s, y) of input $x \in \mathbb{R}^d$, spurious feature value $s \in \{0, 1\}$, and label $y \in \{0, 1\}$. Similarly, let $\mathcal{D}_{\text{test}}$ be a set of test examples. Importantly, we do not make any assumptions on the data generating process, except that x has some information about s and s is predictive of y on the training set, but their correlation does not hold on the test set. For an example (x, s, y), we define its *group* $g = 2y + s$. In a binary classification task with a single binary spurious feature, there are four groups. Without loss of generality, we assume that for majority of training examples we have $y = s$. Hence, we refer to groups 0 and 3 as majority groups, while referring to groups 1 and 2 as minority groups.

Training a transformer to perform linear regression in-context requires millions of ICL training instances, even for small dimensional cases. For example, Garg et al. (2022) use 32 million training instances for 20-dimensional inputs. Next, we consider ways of generating so many ICL instances from a single task.

2.1 A naive approach of constructing ICL instances

The standard approach to constructing an ICL instance is to sample a subset of $n+1$ examples $\{(x_i, s_i, y_i)\}_{i=1}^{n+1}$ from $\mathcal{D}_{\text{train}}$ and form a sequence $S = (x_1, \tilde{y}_1, x_2, \tilde{y}_2, \ldots, x_n, \tilde{y}_n, x_{n+1})$, where $\tilde{y}_i \in \mathbb{R}^d$ is a fixed random representation of either y_i or g_i (this distinction will be elaborated on later). Then one trains a transformer $f_\theta : \cup_k \mathbb{R}^{k \times d} \to [0, 1]$ to predict y_i given $S_i \triangleq (x_1, \tilde{y}_1, \ldots, x_{i-1}, \tilde{y}_{i-1}, x_i)$ (see Figure 1a), optimizing the following loss function:

$$\frac{1}{n+1} \sum_{i=1}^{n+1} \text{CE}(y_i, f_\theta(S_i)), \tag{1}$$

where $\text{CE}(y, \widehat{y}) = -y \log \widehat{y} - (1-y)\log(1-\widehat{y})$ is the binary cross-entropy loss. We explore two options of setting $\tilde{y}_i$. In the first option, we set $\tilde{y}_i$ to represent y_i with a constant vector or its negation in $\mathbb{R}^d$. In this case we aim to obtain an in-context learner that is robust to spurious features without receiving spurious feature annotations as input. ERM is one such learner that minimizes average loss on training examples and does not require spurious feature annotations. In the second option, we set $\tilde{y}_i$ to represent g_i as a sum of two constant vectors in $\mathbb{R}^d$, one representing the class and the other representing the spurious feature. In this case we aim to obtain an in-context learner that does robust classification with respect to a specified spurious feature. GroupDRO is one such learner that minimizes worst-group loss, therefore requiring spurious feature annotations at training time.

Unfortunately, the simple approach of (1) has several issues. First, as the mapping from inputs to labels is the same for all ICL training instances, the model can do in-weights learning, wherein the model encodes this stable mapping in its weights and classifies queries without using context examples. In other words, the learned algorithm does no in-context learning and predict y_i based solely on x_i, essentially memorizing the task. Second, as all $n+1$ examples of a sequence S are sampled from the training set and the spurious correlation holds for all of them, there is nothing preventing usage of spurious features in making predictions. To confirm these two issues, we consider the `Waterbirds` dataset (Sagawa* et al., 2020), which is a landbird versus waterbird image classification task where image background (sea or land) is correlated with the label in the training set (4,795 examples), but not in the validation and test sets. A robust classifier should predict `waterbird` or `landbird` without relying on image background. To separate out the representation learning challenge, we represent images with a pretrained and frozen DINOv2 ViT-B/14 distilled (Oquab et al., 2023). This way each image is embedded in $\mathbb{R}^{768}$. While using powerful pretrained representations increases overall performance under distribution shifts (Radford et al., 2021; Mehta et al., 2022), we note that it does not eliminate the problem of spurious correlations. Representations obtained via large-scale self-supervised pretraining are likely rich enough to capture information about both the label and spurious feature. Furthermore, many works have indicated that the main contribution to the out-of-domain generalization error comes from the classification head (rather than the representation learning module) and called for designing better methods of training the classification head (Galstyan et al., 2022; Menon et al., 2021; Kirichenko et al., 2023; Izmailov et al., 2022; Shi et al., 2023).

We train a causal decoder-only GPT-J transformer (Wang & Komatsuzaki, 2021) with 80M parameters on 2M in-context learning sequences with $n = 512$ and $\tilde{y}_i$ representing labels, constructed from the training set of `Waterbirds`. We use balanced sampling of classes and set the minority group proportion to 10% within each class. We use the ADAM optimizer (Kingma & Ba, 2014) ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) with 32 batch size and no weight decay. The learning rate is selected from $\left\{3 \cdot 10^{-5}, 6 \cdot 10^{-5}, 10^{-4}\right\}$ based on average test performance over 5 runs. Concretely, we evaluate on 8192 sequences where the context part is n training examples, while the query is a sampled from the test set with equal group distribution. Exact metric definitions and missing details are provided in Section A. Note that with 512 context length and 10% minority group ratio within each class, the expected value of the number of context examples from each of the 2 minority groups is about 25. For reference, the smallest minority group has only 56 examples in the `Waterbirds` training set.

Figure 2a plots worst-group test accuracy as a function of context size n. We see that the naive approach results in models that ignore context – worst-group accuracy with 512 context examples is essentially the same as with 2 examples (see the *naive* curve). This confirms the in-weights learning issue. Figure 17a also shows that majority-group test accuracy of the naive approach is considerably higher compared to worst-group accuracy, confirming the non-robustness issue.

2.2 The proposed approach of constructing ICL instances

To prevent in-weights learning and induce in-context learning instead, potentially enabling generalization to other tasks, one approach is to increase the number of training tasks/mappings. We propose rotating input embeddings in each ICL instance independently as a simple approach of deriving many tasks from a single source task. When training with rotated input embeddings, every ICL sequences receives a different permutation and thus represents a different mapping from inputs to labels. This naturally encourages in-context learning, where the model uses context examples to infer the mapping on the fly, instead of encoding a single mapping. We found that generating random rotation matrices on-the-fly is computationally expensive and slows down training. We tried generating and storing 10K rotation matrices, but this resulted in fewer than 50M different training examples that were still possible to memorize to some extent. A more effective and efficient alternative is to apply random permutations to image embedding dimensions (for brevity, this technique is denoted by *+P* in figures and tables; please see Figure 12 for an illustration of this technique). We found this approach to be very effective in inducing in-context learning (see *naive + P* in Figure 2a). We also see that the difference between majority-group and worst-group accuracies decreases, although an approximately 5 p.p. gap remains.

When training an ICL transformer, ideally, we would like to simulate the situation of making a test prediction based on a context of training examples. Importantly, we would like to simulate the case where the test distribution has balanced groups (i.e., the spurious correlation does not hold). Given access to spurious feature annotations *for the training set*, we can simulate this scenario using only training examples. In particular, we can form ICL instances of form $(x_1, \tilde{y}_1, \ldots, x_n, \tilde{y}_n, x_{n+1})$, where the context examples $(x_1, \ldots, x_n)$ are sampled in such a way that the spurious feature is correlated with the label, while the query x_{n+1} is sampled to have a uniform group distribution. However, if we again optimize the loss of (1), for context lengths less than n, the network will be allowed to make predictions using the spurious feature, which is undesirable. Please refer to Figure 19 of Section C for evidence of this. To address this, one can compute loss only on the query token, ignoring all intermediate predictions. Unfortunately, this approach leads to reduced sample efficiency and slower learning.

Instead, we introduce a notion of *intermediate queries*, which allows us to simulate making predictions on test examples sampled from a balanced group distribution at all context lengths at once. Given training examples $\{(x_i, \tilde{y}_i)\}_{i=1}^n$, we sample test examples $\{q_i\}_{i=1}^n$ from $\mathcal{D}_{\text{train}} \setminus \{x_1, \ldots, x_n\}$ with uniform group distribution. We name these test examples $\{q_i\}_{i=1}^n$ *intermediate queries*. Then, we form a sequence $S = (x_1, \tilde{y}_1, q_1, x_2, \tilde{y}_2, q_2, \ldots, x_n, \tilde{y}_n, q_n)$. Our goal is to do a single forward pass on this sequence S and obtain the predictions of form $\hat{y}_i \triangleq f_\theta((x_1, \tilde{y}_1, x_2, \tilde{y}_2, \ldots, x_i, \tilde{y}_i, q_i))$ for all $i \in [n]$. To this end we make two modifications, essentially making representations of context tokens $x_1, \tilde{y}_1, \ldots, x_n, \tilde{y}_n$ agnostic to the presence of the intermediate queries $q_1, \ldots, q_n$. First, we modify the causal attention matrix to disable attention to query tokens, unless a query token is attending to itself. See Figure 11 for an illustration for $n = 3$. Formally, if we enumerate tokens from 1 to $3n$ and define $M_{i,j}$ as the attention mask for token i attending to token j, then we set

$$M_{i,j} = \begin{cases} 0, & i < j, \\ 0, & i > j \text{ and } j \equiv 0 \mod 3, \\ 1, & \text{otherwise.} \end{cases} \tag{2}$$

Second, we use modified token positions for computing positional encodings, in order to discount intermediate query tokens. Namely, for the sequence $(x_1, \tilde{y}_1, q_1, x_2, \tilde{y}_2, \ldots, x_n, \tilde{y}_n, q_n)$, position indices are set to $(0, 1, 2, 2, 3, 4, 4, \ldots, 2n-2, 2n-1, 2n)$. Formally, enumerating tokens from 1 to $3n$, the position index of the i-th token is set to $2\left\lfloor \frac{i-1}{3} \right\rfloor + (i-1) \mod 3$. Please refer to Figure 1 for an illustration.

Finally, once the predictions on the intermediate queries are obtained, we optimize the average loss on the *query* examples:

$$\frac{1}{n} \sum_{i=1}^{n} \text{CE}(y_{q_i}, \hat{y}_i), \tag{3}$$

where y_{q_i} is the label of query q_i. Hereafter, we refer to this approach as simply "proposed approach".

Figure 2a compares the proposed and naive approaches with and without input dimension permutations. Without random permutations, the proposed approach marginally outperforms the naive approach. However, the same is not true with random permutations. We found that image embeddings of DINOv2 have a bias toward representing objects more than backgrounds, alleviating the challenge posed by the spuriously correlated background in `Waterbirds`. In fact, with 512 training examples, the linear probing accuracy of the spurious feature is only ≈ 65%, while that of the label is ≈ 95%. For comparison, with ResNet-50 (He et al., 2016) embeddings, the linear probing accuracies of the spurious feature and label are ≈ 81% and ≈ 85% respectively.

For this reason, we create a modified version of Waterbirds by adding a constant vector $\tilde{s}$ or $-\tilde{s}$ to image embeddings based on the spurious feature s. We scale $\tilde{s}$ to have its norm equal to the average norm of image embeddings and verify that the linear probing accuracy of the spurious feature becomes 100%. On this modified `Waterbirds` dataset, which we call `Waterbirds-severe`, we see a large separation between the naive and proposed approaches (see Figure 2b). We also see that without permutations, both the naive and proposed approaches perform identically, indicating no robustness to the spurious correlation. This is expected, because in the absence of in-context learning, we can think of the naive and proposed approaches as standard and reweighted empirical risk minimization with a complex classification head, respectively. Sample reweighting has been observed to be ineffective in overparameterized settings, as all training examples will be perfectly fitted (Byrd & Lipton, 2019; Menon et al., 2021).

2.3 Comparison with conventional learning algorithms

Now that we have established the efficacy of the proposed technique, we compare it to a few established algorithms, such as 1-NN, ERM, and GroupDRO, of which the latter two have been historically hard to outperform (Gulrajani & Lopez-Paz, 2021; Koh et al., 2021). A comparison to more methods designed for robustness to spurious correlations is outside of the goal of this work, namely, studying limits of in-context learning. In our comparisons, we follow the evaluation recipe used for the in-context learners. We evaluate each baseline on 8192 sequences by training on the context part of the sequence and making a prediction on the single query. More information about hyperparameters and model selection is presented in Section A.

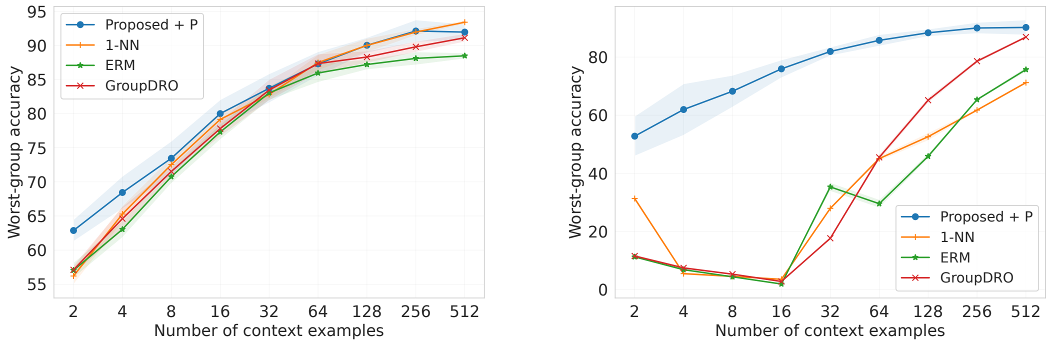
Figures 3a and 3b compare the proposed and baseline approaches on `Waterbirds` and `Waterbirds-severe` respectively. On `Waterbirds`, the proposed method outperforms ERM and GroupDRO on almost all context lengths, but is better than 1-NN only for short context lengths. The good performance of 1-NN is due to the bias in DINOv2 representations. On `Waterbirds-severe`, the proposed method outperforms the baselines at all context lengths. From these results, we conclude that in-context learners obtained with the proposed approach implement none of these algorithms.

It should be noted that worst-group accuracies of baselines at $n = 512$ are actually *higher* than what we get when training on the entire dataset. For example, on `Waterbirds`, 1-NN gets only 90.03 % worst-group accuracy, while ERM gets 84.23 ± 0.17 % and GroupDRO gets 92.43 ± 0.24 %. This is due to balanced class sampling and setting the minority ratio to 10% within each class, which is higher than the minority ratio of $\approx 5\%$ in the original `Waterbirds` dataset. One can think of our resampling as a weaker form of down-sampling which has been found to be helpful in presence of spurious correlations (Nagarajan et al., 2021; Menon et al., 2021; Idrissi et al., 2022).

Additionally, we verify our findings on one more image classification task `CelebA` (Liu et al., 2015) and on a natural language inference task `MultiNLI` (Williams et al., 2018a). `CelebA` is blond vs non-blond person classification, with sex being a spurious variable. Unlike `Waterbirds`, the spurious feature is asymmetric in `CelebA`, as blond and non-blond women are equally represented, while blond men are significantly infrequent compared to non-blond men. In `MultiNLI`, given a pair of sentences, a premise and a hypothesis, the task is to determine whether the hypothesis is entailed by, neutral with, or contradicts the premise. Prior work has observed that there is a spurious correlation between contradictions and the presence of the negation words *nobody*, *no*, *never*, and *nothing* (Gururangan et al., 2018). For our experiments, we frame a binary classification task *entails or neutral vs. contradicts*, with a binary spurious feature, which is 1 if and only if the hypothesis has one of the four negation words. For both datasets, we verify the two shortcomings of the conventional approach and demonstrate the efficacy of the proposed techniques compared to the baselines. The results for `CelebA` are presented in Table 4 and Figure 20a, while the results for `MultiNLI` are presented in Table 6 and Figure 20b.

2.4 Generality of the learned algorithm

Since we train in-context learners on ICL instances of a single task, a natural question arises whether the learned algorithm can generalize to unseen tasks. Without permuting input dimensions, the model does not learn to do in-context learning. Thus, we cannot expect any generality without permuting input dimensions. We take the model obtained with the "Proposed + P" approach and probe its in-context learning generality



(a) Waterbirds

(b) Waterbirds-severe

Figure 3: Worst-group test accuracies on Waterbirds and Waterbirds-severe for the proposed approach and conventional methods such as 1-NN, ERM, and GroupDRO. Majority-group accuracies are reported in Figure 18 of Section C.

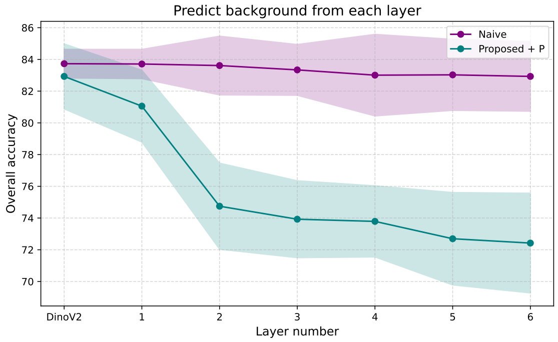


Figure 4: Linear probing accuracy of the background variable at various layers of in-context learner transformers trained on Waterbirds.

by evaluating on various tasks. We start by swapping the labels of two classes in Waterbirds at evaluation and observe ≈ 2 p.p. overall accuracy drop and ≈ 5 p.p. worst-group accuracy drop. Despite the worsened performance, this indicates that the model treats class labels symbolically, which is remarkable as labels had consistent semantics during training. However, when we evaluate on Waterbirds-severe, it gets 100% accuracy on the majority groups and *0% accuracy on minority groups.* Additionally, when we switch the task to predicting the background in the original Waterbirds dataset (now the class becomes a spurious feature), the overall test accuracy drops to 54.4%, while the worst-group accuracy drops to 9.3%.

It is worth noting that the learned algorithm is not completely useless for other tasks and works well in absence of spurious features, even on unseen tasks. For example, evaluating on binary classification tasks derived from the CUB-200 (Welinder et al., 2010) dataset, from where the bird images of Waterbirds were taken, we get 99.7% accuracy at context size 100 (the accuracy is so high because most pairs of classes are easy to distinguish). We also test on binary classification tasks derived from classes belonging to *Amphibia* and *Mammalia* supercategories of the iNaturalist (Van Horn et al., 2018) dataset. At context length 512, the overall accuracy is 98.5%.

These OOD evaluation results indicate that the learned algorithm does something specific to the spurious feature of Waterbirds. We hypothesize that it learns to ignore this particular spurious feature. To test this, we evaluate on *group-balanced* Waterbirds sequences, with the task set to predicting background, and get 58.5% overall accuracy and 41.3% worst-group accuracy. Additionally, we do a forward pass on 1024 ICL sequences and collect final query representations at various layers of the transformer. We then do a linear probing (512 examples for probe training and 512 for validation) to measure predictability of the background

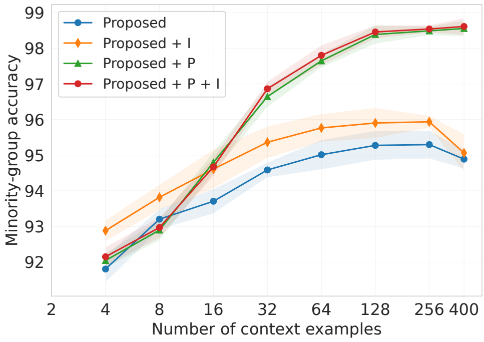


Figure 5: Minority-group accuracy on the OOD test set of `iNaturalist` for the proposed approach with or without permuting input dimensions and promoting induction heads.

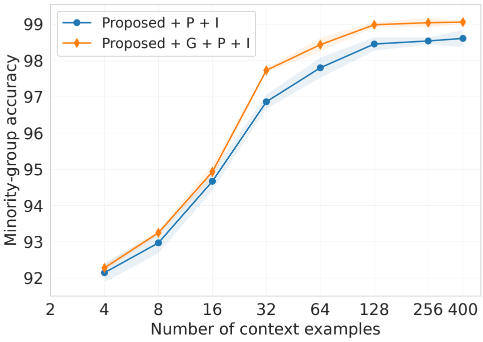


Figure 6: Minority-group accuracy on the OOD test set of `iNaturalist` for the best proposed approach with or without passing group information as input.

variable. We find that the “Proposed + P” approach reduces background information effectively as we sweep from input to the final layer, while the “Naive” fails to reduce background probing accuracy (see Figure 4).

One potential way of improving generality and possibly also performance is passing example groups as input, i.e., setting $\tilde{y}_i$ to represent g_i. We did not observe performance improvements or an increase in generality of the learned algorithm when passing groups as input (see Tables 1 and 2 of Section C). Thus, we conclude that when all ICL instances are derived from a single task, the learned algorithm is inherently tied to the spurious feature of that task.

3 In-context learning based on a diverse set of tasks

In Section 2, we showed that it is possible to obtain a good in-context learner for a given task, but it fails to generalize to tasks with different spurious features. A better in-context learner should detect spurious features from context and make predictions without employing them. In this section, we explore the possibility of obtaining such a learner by training on a diverse set of ICL tasks. Since there exist few suitable datasets, we synthesize binary classification tasks with a single binary spurious feature, aiming to capture “structure” present in existing datasets. In short, given a standard binary classification task, say cat vs. dog classification, for a sampled minority of cats we overwrite some of their features with those of random dogs. Similarly, we do an analogous operation for a sampled minority of dogs. In this way some cats share dog features and vice versa. To create a diverse pool of in-context learning instances, we vary the two classes and the subset of grafted features. Please refer to Figure 16 for an illustration of this grafting operation.

More concretely, we consider the `iNaturalist` dataset (Van Horn et al., 2018), which contains images from 5,089 natural fine-grained categories and filter out categories that have fewer than 500 images. For testing purposes, from remaining 239 categories we set apart the ones belonging to *Amphibia* and *Mammalia* supercategories, along with 10% of random categories. We denote the set of these 48 categories as $\mathcal{C}_{\text{ood}}$, and the set of remaining 191 categories as $\mathcal{C}_{\text{id}}$, which we use to create ICL instances for training. For each category in $\mathcal{C}_{\text{id}}$, we hold out half of the examples as in-distribution validation set. To generate an ICL instance, we randomly sample two distinct classes from $\mathcal{C}_{\text{id}}$ and sample $n/2$ images from the training split of each class uniformly at random without replacement. Please refer to Figure 15 for an illustration of our preprocessing of `iNaturalist`. We then do the grafting operation, setting the minority group ratio within each class to 10%. We select the grafted features randomly, by first picking a subset size k uniformly at random from 0 to 199, and then sampling a random subset of embedding dimensions of size k. With this we get n examples that form the context part of the instance. Abandoning the naive approach and focusing on the proposed one, for each class we sample $n/2$ queries from the remaining examples uniformly at random with replacement and do the grafting operation with 50% minority group ratio.

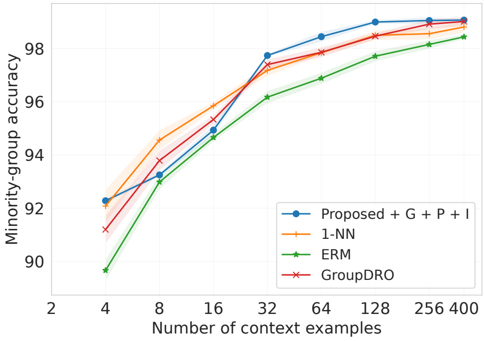


Figure 7: Minority-group accuracy on the OOD test set of `iNaturalist` for the best variant of proposed approach and conventional methods such as 1-NN, ERM, and GroupDRO.

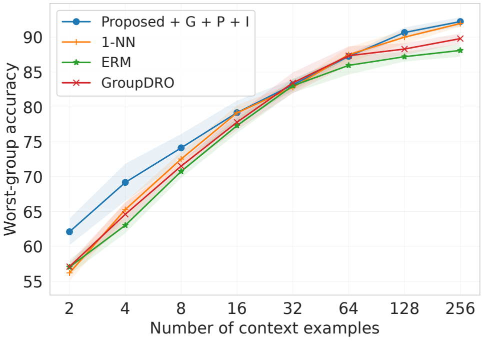


Figure 8: Worst-group test accuracy on `Waterbirds` for the best variant of proposed approach trained on `iNaturalist` and for methods such as 1-NN, ERM, and GroupDRO.

Following the experiments in Section 2, we train the same transformer with the proposed approach on 4M ICL instances with $n = 400$ context examples. We use the same optimizer and sweep the learning rate in the same range, selecting the best value based on the average *minority-group accuracy* (defined exactly in Section C) on instances where both categories belong to $\mathcal{C}_{\text{ood}}$ and thus were not observed during training. The results presented in Figure 5 indicate a major difference compared to the results in the single-task regime, namely, the proposed approach learns to do in-context learning to some extent without permuting embedding dimensions. As expected, we see much better performance with permuted embedding dimensions. Notably, comparing majority-group and minority-group accuracies of the proposed approach with permutations (Figure 23), we see almost no sign of reliance on spurious features.

Promoting emergence of induction heads. In-context learning ability has been linked to induction heads, which are a specific type of circuit found within large language models that implement the operation of looking back over the sequence for finding previous instances of the current token and copying what comes after that (Olsson et al., 2022). Inspired by this, we propose a data preparation technique that promotes learning of induction heads. With probability p, we replace each intermediate query independently with a random example from the preceeding part of the context (see Figure 14 for an illustration of this technique). Note that this type of “hinting” is not possible in the naive approach and is enabled by the introduction of intermediate queries. In all experiments with this technique enabled, we just set $p = 0.25$. We observed that training of typical runs escapes the initial loss plateau faster with this technique (in about 3k iterations compared to about 10k iterations). Moreover, we see modest performance gains in `iNaturalist` experiments (see Figure 5, where *+I* stands for this technique).

Passing example groups as input. In contrast to the findings in the single-task setting of Section 2, we observed that setting $\tilde{y}_i$ to represent group improves the proposed approach, even in addition of permuting input dimensions and promoting induction heads. One case of this is presented in Figure 6, while more cases can be found in the complete results presented in Section C. For brevity, we mark passing groups as inputs with *+G* in figures and tables. Please see Figure 13 for an illustration.

Comparison with conventional learning algorithms. Similar to the experiments in Section 2, we compare the best variant of the proposed approach (G + P + I) to 1-NN, ERM, and GroupDRO. The results presented in Figure 7 show that the learned algorithm is on par with or outperforms the baselines starting at context length 32. The results at context lengths below 20 are not as informative, since our implementation of the grafting operation implies that no examples are grafted when there are less than 10 examples in a class.

Generality of the learned algorithm. To test the generality of the learned algorithm, we report evaluation results on `Waterbirds` (Figure 8) and `Waterbirds-severe` (Figure 9). We see that the learned algorithm outperforms baselines on `Waterbirds` and performs comparably to the model trained on `Waterbirds` itself. However, the learned algorithm fails completely on `Waterbirds-severe`, while the baselines give

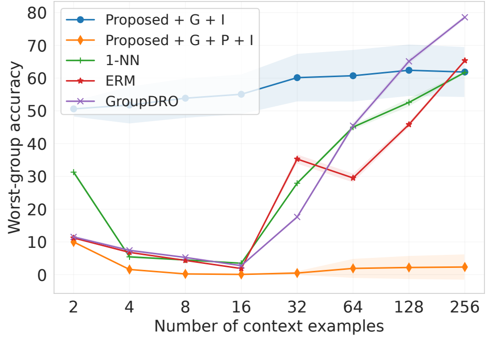


Figure 9: Worst-group test accuracy on `Waterbirds-severe` for the best variant of proposed approach trained on `iNaturalist` and for methods such as 1-NN, ERM, and GroupDRO.

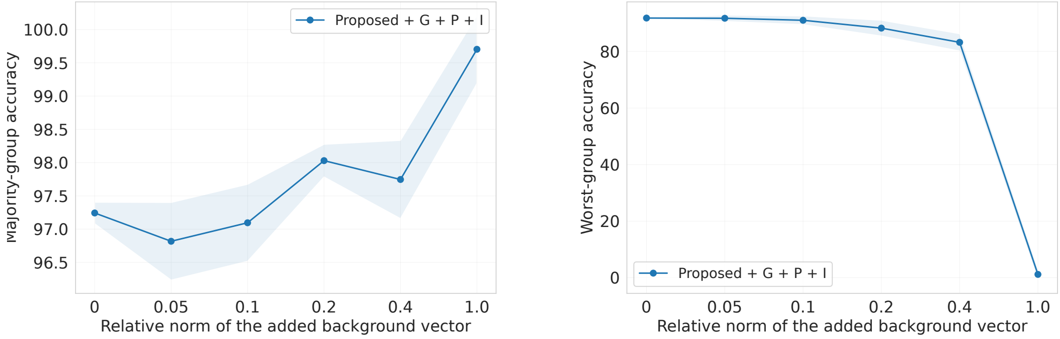


Figure 10: Majority-group and worst-group test accuracies of a proposed model (G + P + I) trained on `iNaturalist`, but evaluated on a modified variants of `Waterbirds` where we add a vector representing the spurious feature (background). The x-axis is the relative norm of the added vector compared to the average `Waterbirds` image embedding norm. Relative norm of 0 corresponds to `Waterbirds`, while relative norm of 1 corresponds to `Waterbirds-severe`. Shaded regions show standard deviation across 5 training runs.

meaningful results starting at context length 32. We hypothesize that the challenge posed by the spurious features in `Waterbirds-severe` is significantly more severe compared to that in `iNaturalist`. By varying the norm of the added background vector, we interpolate between `Waterbirds` and `Waterbirds-severe`, and we see good generalization until the norm of the added vector is ≈40% of the average embedding norm (see Figure 10). It is possible that the learned algorithm specializes to spurious features planted with the grafting operation, but fails to handle a strong, dense, additive spurious feature.

4 Related Work

In-weights vs in-context learning. We observe two modes of learner behavior in our experiments. In the first mode, the learner acts like a standard supervised classifier, ignoring context examples. This mode appears when training on ICL instance of a single task without permuting input embedding dimensions. In the second mode, the learner does proper in-context learning. Our experiments indicate that both permuting embedding dimensions and increasing the number of training tasks are reliable ways of steering the model towards the in-context learning mode. The former is akin to the method of randomly projecting inputs proposed by Kirsch et al. (2022) for obtaining general-purpose in-context classifiers. Prior work has made a distinction between these two modes of learning, naming them in-weights and in-context learning. In particular, Chan et al. (2022) demonstrate that certain distributional properties of data, such as long-tail of class frequencies and bursty distribution of context example classes, can promote in-context learning when meta-training on few-shot classification instances. Singh et al. (2024) show that in-context learning behavior is not persistent and decays away with overtraining, indicating a trade-off between in-weights and in-context learning mechanisms. Moreover, they find that this in-context learning skill decay can be prevented by applying weight decay of embeddings and MLP layers, slowing down in-weights learning. Anand et al. (2024) make similar observations about these two modes of learning and propose active forgetting of token embeddings as an effective way of steering towards the in-context learning mode.

Many shot ICL. One ancillary finding of this work is that transformers can be trained to do in-context learning of visual classification tasks when good image embeddings are provided. This is remarkable because the input dimensionality we considered is much higher than what was considered in the pioneering works of Garg et al. (2022) and Akyürek et al. (2023) (784 vs 20). Furthermore, we observe predictable performance gains from longer context sizes. The number of "shots" we consider (up to 512 examples) is well beyond what is typically considered in ICL works (up to a few dozen examples). Our findings are complementary to those of Agarwal et al. (2024), Jiang et al. (2024), and Li et al. (2024) who find that multimodal large language models, such as Gemini-1.5 Pro and GPT-4o, can benefit from large number of in-context demonstrations (up to 1000 demonstrations).

In-context learning and out-of-distribution generalization. Closest to our work are the works that propose to make use of in-context learning for out-of-distribution generalization. Han et al. (2023) test multimodal large language models (MLLMs) on a variety of visual classification tasks. They propose to leverage in-context learning abilities of MLLMs to improve performance on specialized domains and on tasks with significant corruptions. However, they only consider the case where both context examples and query are from the target domain. Zhang et al. (2024) make similar observations, but additionally study robustness of in-context learning to distribution shifts, such as domain shifts, label shifts, and spurious correlations. They find that in-context learning is highly susceptible to label shifts and presence of spurious correlations. This is in agreement with the findings of (Ahuja & Lopez-Paz, 2023) and our work that in-context learners are brittle and don't generalize under severe distribution shifts. Finally, Gupta et al. (2024) propose to address the problem of domain generalization (Muandet et al., 2013) by training an in-context learner that can take examples from a domain/environment and adapt to that domain in-context.

Task Diversity. The important of task diversity for out-of-distribution generalization is well-studied outside of the in-context learning literature. In a transfer learning setting, Tripuraneni et al. (2020) derive theoretical results highlighting the importance of task diversity in addition to samples per task in obtaining transferable representations. Hendrycks et al. (2020) find that pretrained then fine-tuned models generalize to out-of-distribution examples better than only fine-tuned models, connecting the success to pretraining task diversity. Ramanujan et al. (2023) provide more direct evidence of this connection for image classification. Task diversity has also been one of the primary drivers of success of instruction tuning (Wei et al., 2022). Similarly, the excellent out-of-distribution generalization performance of the CLIP (Radford et al., 2021) has been attributed to its pretraining dataset diversity (Fang et al., 2022). Similar observations have been made for training in-context learners (Raventós et al., 2024; Goddard et al., 2025; Raparthy et al., 2024). Our findings of Section 3 complement this literature and suggest that for training a general in-context learner robust to spurious features, one likely needs a dataset of diverse tasks with diverse spurious features.

5 Discussion and conclusion

We showed that it is possible to train an effective in-context learner tailored to a particular classification task with spurious features. We did this by introducing two key techniques: (a) permuting input embedding dimensions and (b) forming ICL sequences with intermediate queries simulating distribution shift. We provided evidence that the learned algorithm is highly competitive on the task it was trained on. However, we found that while it generalizes to other tasks without spurious features, it does not work for tasks with other spurious features. Understanding this failure mechanistically and exploring techniques for enabling better generalization are key future research directions.

We next explored training on synthetic ICL instances of diverse tasks and showed that it is possible to obtain an in-context learner that generalizes to unseen tasks, even with different data generating processes. We established the usefulness of two more techniques: (c) passing example groups as input and (d) promoting learning of induction heads by occasionally querying past context examples. We believe there is room for improving in-context learning via improved strategies of choosing intermediate queries and possibly optimizing worst-group loss. Understanding why the learned algorithm fails under extreme distribution shifts and why variants with permutations fail more (see Figure 10) is an interesting question to explore. Another interesting direction to explore is to find out what exact algorithm is learned in the process of training on diverse tasks. Based on the results presented in this work, we conclude that the learned algorithm is neither 1-NN, ERM, or GroupDRO.

Our work has several limitations. First, training a transformer-based in-context learner with high-dimensional image embeddings is computationally costly (see Section A for information on compute resources), although it is faster than the baselines during inference. For this reason, we did not explore more datasets and pretrained image embeddings. We believe the main conclusions of our work will remain unchanged and provide an experiment on `CelebA` with a larger network in Section C. Second, we experimented with only one model size, width, and depth. Larger models might behave differently (Wei et al., 2023). Third, in our `iNaturalist` experiments, we considered only one "type" of spurious features. This choice is likely to have a significant effect on the learned algorithm and its generality. Future research should explore more ways of synthesizing spurious features and consider varying the severity of the challenge posed by spurious features. The latter can be done by considering multiple spurious features, introducing label imbalance, varying the magnitude of spurious correlations, and varying the margin spurious features provide.

Finally, we acknowledge that the proposed approach of *training* robust in-context learners requires spurious feature annotations, which are typically costly to obtain. As we have shown, this limitation can be addressed by creating synthetic data, in which case spurious annotations are readily available. At inference time, the learned algorithm does not require spurious annotations if it is trained with $\tilde{y}_i$ set to represent y_i (i.e., ERM-like algorithm), but requires them when it is trained with passing example groups as input (i.e., $\tilde{y}$ set to represent g_i; GroupDRO-like algorithm). It is important to note that as we consider classification problems where the learner receives training data only from a single *environment*, spurious annotations are generally necessary to disambiguate core and spurious features.

Broader Impact Statement

Spurious correlations can appear in various domains, often in complex and unexpected ways. Given the increasing reliance on few-shot prompting in large language models, understanding and mitigating the negative effects of spurious correlations is important for ensuring the reliability and safety of these models. We hope that our work can lead to new methods and training strategies for making LLMs more robust to spurious correlations present in the context.

References

Rishabh Agarwal, Avi Singh, Lei M Zhang, Bernd Bohnet, Stephanie Chan, Ankesh Anand, Zaheer Abbas, Azade Nova, John D Co-Reyes, Eric Chu, et al. Many-shot in-context learning. *arXiv preprint arXiv:2404.11018*, 2024.

Kartik Ahuja and David Lopez-Paz. A closer look at in-context learning under distribution shifts. *arXiv preprint arXiv:2305.16704*, 2023.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2023.

Suraj Anand, Michael A Lepori, Jack Merullo, and Ellie Pavlick. Dual process learning: Controlling use of in-context vs. in-weights strategies with weight forgetting. *arXiv preprint arXiv:2406.00053*, 2024.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 456–473, 2018.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Jonathon Byrd and Zachary Lipton. What is the effect of importance weighting in deep learning? In *International conference on machine learning*, pp. 872–881. PMLR, 2019.

Stephanie Chan, Adam Santoro, Andrew Lampinen, Jane Wang, Aaditya Singh, Pierre Richemond, James McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. *Advances in Neural Information Processing Systems*, 35:18878–18891, 2022.

Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. Data determines distributional robustness in contrastive language image pre-training (clip). In *International conference on machine learning*, pp. 6216–6234. PMLR, 2022.

Tigran Galstyan, Hrayr Harutyunyan, Hrant Khachatrian, Greg Ver Steeg, and Aram Galstyan. Failure modes of domain generalization algorithms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19077–19086, 2022.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11): 665–673, 2020.

Chase Goddard, Lindsay M. Smith, Vudtiwat Ngampruetikorn, and David J. Schwab. When can in-context learning generalize out of task distribution? In *Forty-second International Conference on Machine Learning*, 2025.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

Sharut Gupta, Stefanie Jegelka, David Lopez-Paz, and Kartik Ahuja. Context is environment. In *The Twelfth International Conference on Learning Representations*, 2024.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 107–112, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2017.

Zhongyi Han, Guanglin Zhou, Rundong He, Jindong Wang, Xing Xie, Tailin Wu, Yilong Yin, Salman Khan, Lina Yao, Tongliang Liu, et al. How well does gpt-4v (ision) adapt to distribution shifts? a preliminary investigation. *arXiv preprint arXiv:2312.07424*, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. Pretrained transformers improve out-of-distribution robustness. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2744–2751, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.244.

Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning*, pp. 336–351. PMLR, 2022.

Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew G Wilson. On feature learning in the presence of spurious correlations. *Advances in Neural Information Processing Systems*, 35:38516–38532, 2022.

Yixing Jiang, Jeremy Irvin, Ji Hun Wang, Muhammad Ahmed Chaudhry, Jonathan H Chen, and Andrew Y Ng. Many-shot in-context learning in multimodal foundation models. *arXiv preprint arXiv:2405.09798*, 2024.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *The Eleventh International Conference on Learning Representations*, 2023.

Louis Kirsch, James Harrison, Jascha Sohl-Dickstein, and Luke Metz. General-purpose in-context learning by meta-learning transformers. *arXiv preprint arXiv:2212.04458*, 2022.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. Long-context llms struggle with long in-context learning. *arXiv preprint arXiv:2404.02060*, 2024.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334.

Raghav Mehta, Vítor Albiero, Li Chen, Ivan Evtimov, Tamar Glaser, Zhiheng Li, and Tal Hassner. You only need a good embeddings extractor to fix spurious correlations. *arXiv preprint arXiv:2212.06254*, 2022.

Aditya Krishna Menon, Ankit Singh Rawat, and Sanjiv Kumar. Overparameterisation and worst-case generalisation: friend or foe? In *International Conference on Learning Representations*, 2021.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International conference on machine learning*, pp. 10–18. PMLR, 2013.

Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization. In *International Conference on Learning Representations*, 2021.

Devang K Naik and Richard J Mammone. Meta-neural networks that learn by learning. In *[Proceedings 1992] IJCNN International Joint Conference on Neural Networks*, volume 1, pp. 437–442. IEEE, 1992.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

Ofir Press, Noah Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021.

Vivek Ramanujan, Thao Nguyen, Sewoong Oh, Ali Farhadi, and Ludwig Schmidt. On the connection between pre-training data diversity and fine-tuning robustness. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 66426–66437. Curran Associates, Inc., 2023.

Sharath Chandra Raparthy, Eric Hambro, Robert Kirk, Mikael Henaff, and Roberta Raileanu. Learning to solve new sequential decision-making tasks with in-context learning, 2024.

Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. *Advances in Neural Information Processing Systems*, 36, 2024.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

Jürgen Schmidhuber. *Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook*. PhD thesis, Technische Universität München, 1987.

Yuge Shi, Imant Daunhawer, Julia E Vogt, Philip Torr, and Amartya Sanyal. How robust is unsupervised representation learning to distribution shift? In *The Eleventh International Conference on Learning Representations*, 2023.

Aaditya Singh, Stephanie Chan, Ted Moskovitz, Erin Grant, Andrew Saxe, and Felix Hill. The transient nature of emergent in-context learning in transformers. *Advances in Neural Information Processing Systems*, 36, 2024.

Rui Song, Yingji Li, Lida Shi, Fausto Giunchiglia, and Hao Xu. Shortcut learning in in-context learning: A survey. *arXiv preprint arXiv:2411.02018*, 2024.

Ruixiang Tang, Dehan Kong, Longtao Huang, and Hui Xue. Large language models can be lazy learners: Analyze shortcuts in in-context learning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 4645–4657, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.284. URL `https://aclanthology.org/2023.findings-acl.284/`.

Sebastian Thrun and Lorien Pratt. Learning to learn: Introduction and overview. In *Learning to learn*, pp. 3–17. Springer, 1998.

Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pp. 1521–1528. IEEE, 2011.

Nilesh Tripuraneni, Michael Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. *Advances in neural information processing systems*, 33:7852–7862, 2020.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8769–8778, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pp. 35151–35174. PMLR, 2023.

Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. `https://github.com/kingoflolz/mesh-transformer-jax`, May 2021.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023.

P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

Florian Wenzel, Andrea Dittadi, Peter Gehler, Carl-Johann Simon-Gabriel, Max Horn, Dominik Zietlow, David Kernert, Chris Russell, Thomas Brox, Bernt Schiele, et al. Assaying out-of-distribution generalization in transfer learning. *Advances in Neural Information Processing Systems*, 35:7181–7198, 2022.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122. Association for Computational Linguistics, 2018a.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122. Association for Computational Linguistics, 2018b. URL `http://aclweb.org/anthology/N18-1101`.

Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *International Conference on Learning Representations*, 2021.

Steve Yadlowsky, Lyric Doshi, and Nilesh Tripuraneni. Pretraining data mixtures enable narrow model selection capabilities in transformer models. *arXiv preprint arXiv:2311.00871*, 2023.

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.

Xingxuan Zhang, Jiansheng Li, Wenjing Chu, Junjia Hai, Renzhe Xu, Yuqing Yang, Shikai Guan, Jiazheng Xu, and Peng Cui. On the out-of-distribution generalization of multimodal large language models. *arXiv preprint arXiv:2402.06599*, 2024.

Yuqing Zhou, Ruixiang Tang, Ziyu Yao, and Ziwei Zhu. Navigating the shortcut maze: A comprehensive analysis of shortcut learning in text classification by language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 2586–2614, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.146. URL `https://aclanthology.org/2024.findings-emnlp.146/`.

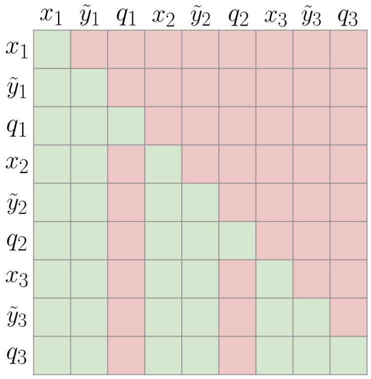
A Further experimental details

Baselines. For empirical risk minimization as a baseline, we tune 2 hyperparameters: learning rate (0.01 or 0.001) and number of epochs (100 or 200). For GroupDRO we additionally tune its parameter that controls adaptiveness of group weights (0.01, 0.1, or 1) and we also try an optional strong L2 regularization (1.0 weight decay), as it has been observed to be useful for small datasets (Sagawa* et al., 2020).

Transformer-based methods. In all transformer-based approaches, we train a causal decoder-only GPT-J transformer with 80M parameters that has 6 transformer layers with 8 multi-head attention, 768 model dimensionality, and 3072 hidden dimensionality. When training on `iNaturalist`, we add a layer normalization (Ba et al., 2016) on transformer input, as we expect input norms to change when we evaluate on `Waterbirds`-based datasets. The transformer input sequence in the proposed approach consists of 3 types of tokens: context image embeddings, query image embeddings, and label/group annotations. While the network can rely on positions and content to distinguish image embeddings from annotations, we found it to be helpful to encode token types explicitly. We do this by setting the first 3 dimensions of a token to be a one-hot vector representing token type (context image embedding, query image embedding, or annotation). When permuting dimensions, we do the permutation before encoding token types to keep the location of token-type information consistent. In our preliminary experiments and development, we used $n = 128$ context length. Apart from improved performance, we did not observe significant qualitative differences when we switched to larger context lengths for final experiments.

Evaluation and model selection. For all transformer-based approaches and baselines, we do a grid search to find the best combination of hyperparameters. In particular, we train each configuration with 5 different random seeds and select the one with the highest average test performance. Importantly, for baseline methods model selection is done for each context length independently, while for transformer-based methods model selection is done once with respect to the test performance at maximum context length observed during training. All evaluations are done on 8192 sequences, where the first n examples are sampled from the corresponding train set while the query is sampled from the test set with a balanced group distribution. Finally, even when training transformers on permuted image embeddings, we do not apply permutations during evaluation. In all figures throughout this work, shaded regions show standard deviation across the 5 training runs.

Note that the most principled model selection approach would be selecting models based on a metric calculated on a dataset similar to the training set (e.g., a held-out part of training set), rather than the test set. For example, in the case of experiments on `Waterbirds` or `Waterbirds-severe`, the principled approach would be to select based on performance on sequences where the context part is sampled from the training set, while the final query is sampled from a held-out validation set with balanced group distribution. We tried this way of model selection and did not observe significant changes. In the case of experiments on `iNaturalist`, the principled approach would be to select based on performance on sequences where the context part is sampled from the training set, while the final query is sampled from the hold-out part the training set. We observed that this in-distribution metric is always around 99.5%-100%, and can be non-informative for model selection. This is a typical scenario in OOD generalization (see for example (Gulrajani & Lopez-Paz, 2021) or (Wenzel et al., 2022)).



	x_1	$\tilde{y}_1$	q_1	x_2	$\tilde{y}_2$	q_2	x_3	$\tilde{y}_3$	q_3
x_1	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
$\tilde{y}_1$	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥
q_1	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥
x_2	🟩	🟩	🟥	🟩	🟥	🟥	🟥	🟥	🟥
$\tilde{y}_2$	🟩	🟩	🟥	🟩	🟩	🟥	🟥	🟥	🟥
q_2	🟩	🟩	🟥	🟩	🟩	🟩	🟥	🟥	🟥
x_3	🟩	🟩	🟥	🟩	🟩	🟥	🟩	🟥	🟥
$\tilde{y}_3$	🟩	🟩	🟥	🟩	🟩	🟥	🟩	🟩	🟥
q_3	🟩	🟩	🟥	🟩	🟩	🟥	🟩	🟩	🟩

Figure 11: The attention mask in the proposed approach with $n = 3$ context examples. Green cells indicate allowed attention pairs, i.e., $M_{i,j} = 1$.

Definitions of metrics. Given a set of predictions on `Waterbirds` or `Waterbirds-severe`, worst-group accuracy is defined as the lowest accuracy of predictions among the 4 groups. Note that worst-group accuracy is not applicable to `iNaturalist`, as different ICL sequences correspond to different classification tasks and hence form different groups. For this reason, we introduce minority-group and majority-group accuracies. Given a triplet $(C, q, \widehat{y})$, where C is a context, q is query, and $\widehat{y}$ is a prediction on q, we call $\widehat{y}$ a minority (majority) prediction, if q is among the least (most) represented group(s) of the context C. Given a list of triplets $(C, q, \widehat{y})$, we define minority (majority) group accuracy as the accuracy among minority (majority) predictions.

Compute resources. We used NVIDIA A100 GPUs with 40GB memory to train transformer-based methods. The network we considered is small enough to fit on one GPU with batch size 32 when $n = 400$ (`iNaturalist` experiments) and batch size 24 when $n = 512$ (`Waterbirds` and `Waterbirds-severe` experiments). We did mixed 16-bit training to save compute and did not notice any quality degradation. A single training takes around 12 hours for `iNaturalist` experiments and around 18 hours for `Waterbirds` experiments. We used a mix of CPUs and weaker GPUs to train baselines, as they are not computationally as demanding.

B Additional illustrations

In this section, we present illustrations of the main techniques and data preparation steps involved in this work.

Figure 11 plots the attention mask M of (2) in the proposed approach for $n = 3$.

Figure 12 presents an illustration of the proposed technique of permuting image embedding dimensions (denoted by *+P* throughout the paper). Figure 13 presents an illustration of the proposed approach with passing example groups as input (denoted by *+G* throughout the paper). Figure 14 presents an illustration of the proposed approach with promoting emergence of induction heads (denoted by *+I* throughout the paper). Figure 15 presents an illustration of our preprocessing of the `iNaturalist` dataset. Figure 16 presents an illustration of the grafting operation for creating spurious features.

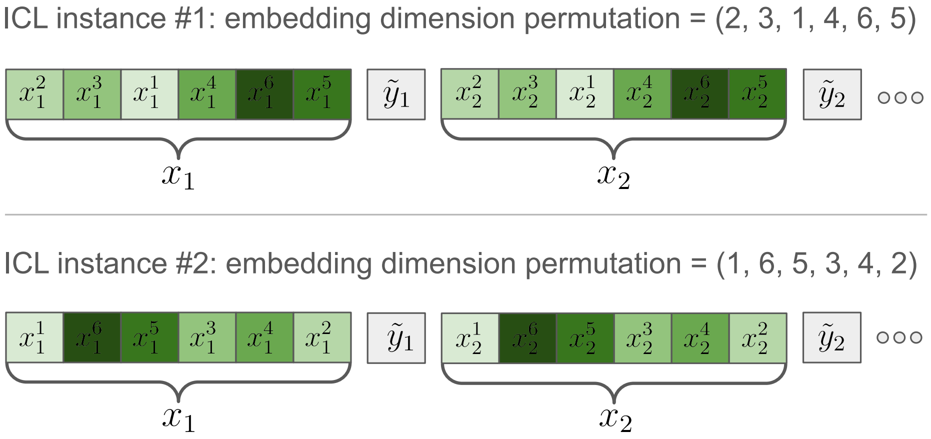

ICL instance #1: embedding dimension permutation = (2, 3, 1, 4, 6, 5)
x_1^2 x_1^3 x_1^1 x_1^4 x_1^6 x_1^5
$\tilde{y}_1$
x_2^2 x_2^3 x_2^1 x_2^4 x_2^6 x_2^5
$\tilde{y}_2$
x_1
x_2
ICL instance #2: embedding dimension permutation = (1, 6, 5, 3, 4, 2)
x_1^1 x_1^6 x_1^5 x_1^3 x_1^4 x_1^2
$\tilde{y}_1$
x_2^1 x_2^6 x_2^5 x_2^3 x_2^4 x_2^2
$\tilde{y}_2$
x_1
x_2

Figure 12: An illustration of the proposed technique of permuting image embedding dimensions (denoted by $+P$ throughout the paper). Note that in each ICL instance we sample a new permutation, but the same permutation is used to permute dimensions of all image embeddings within one ICL instance. Whenever we use the proposed approach of forming ICL sequences (see Figure 1b), the dimensions of intermediate queries are also permuted.

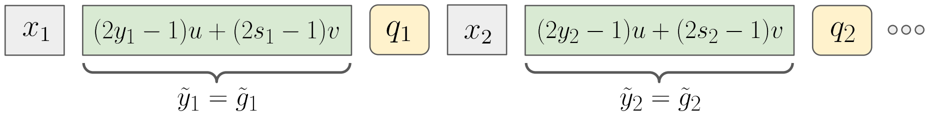

x_1
$(2y_1 - 1)u + (2s_1 - 1)v$
q_1
x_2
$(2y_2 - 1)u + (2s_2 - 1)v$
q_2
$\tilde{y}_1 = \tilde{g}_1$
$\tilde{y}_2 = \tilde{g}_2$

Figure 13: An illustration of the proposed approach with passing example groups as input (denoted by $+G$ throughout the paper).

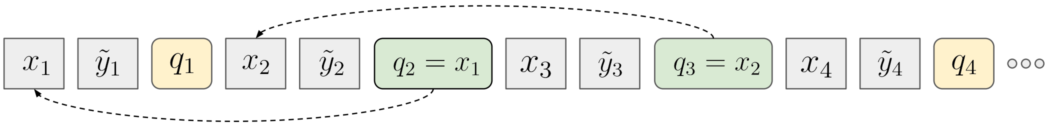


Figure 14: An illustration of the proposed approach with promoting emergence of induction heads (denoted by $+I$ throughout the paper). Intermediate queries that are randomly selected to be one of the previous context examples are shown in green.

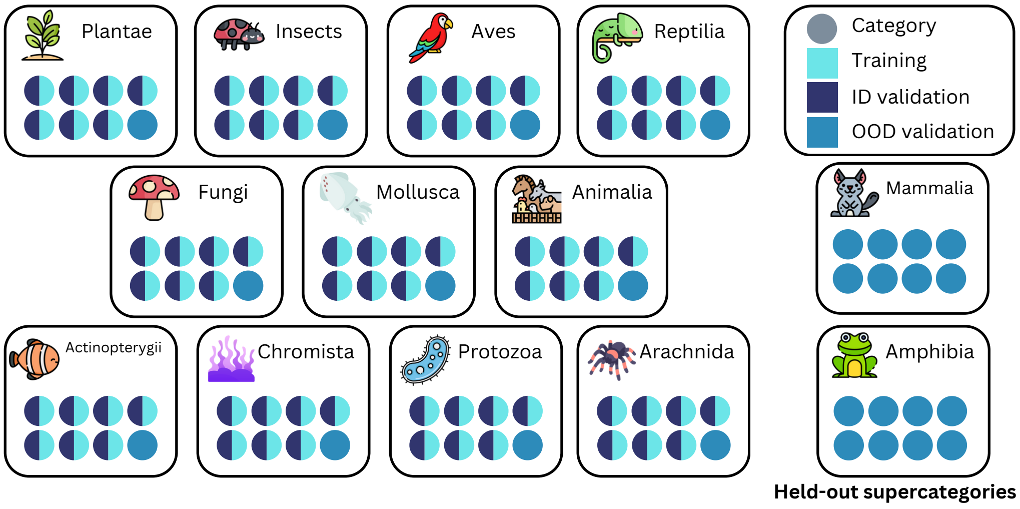
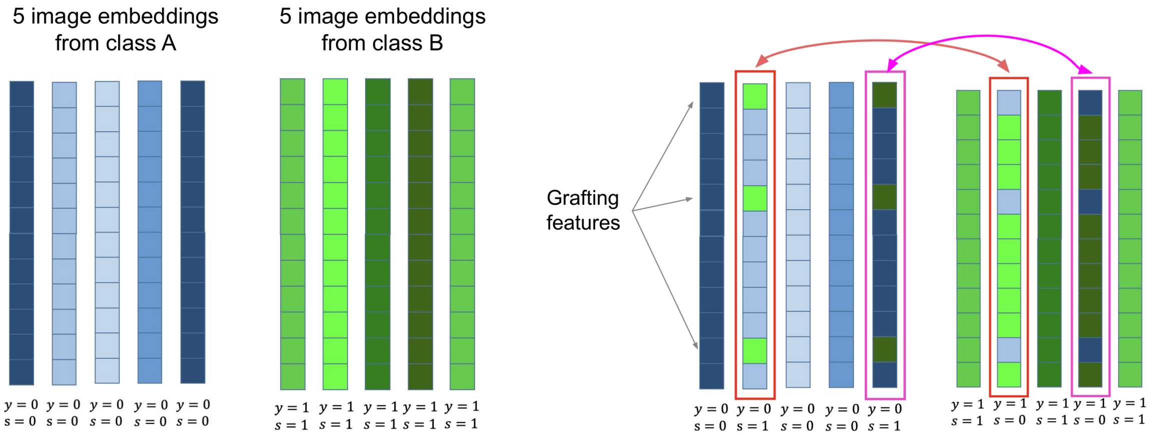


Figure 15: An illustration of our preprocessing of the iNaturalist dataset.



(a) Before grafting

(b) After grafting

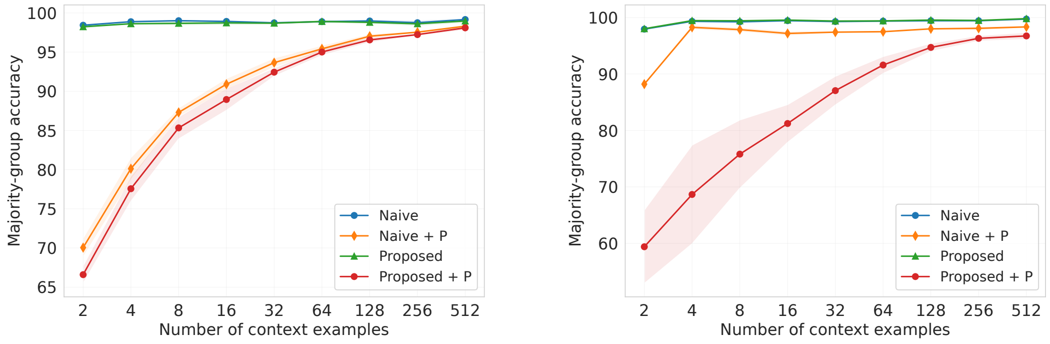
Figure 16: An illustration of the grafting operation for creating spurious features. The figure (a) depicts two classes of examples, each having 5 examples given by 12-dimensional embeddings. In this example, the grafting operation selects 3 embedding dimensions to become spurious features. For this end, these 3 features of examples 2 and 5 of class A are swapped with those of examples 2 and 4 of class B, respectively. Figure (b) depicts the embeddings after the grafting operation.

C Additional results

In addition to the figures presented in the main text, here we provide the exact experimental results for multiple transformer-based and baseline approaches, some of which were not included in the main text due to space constraints. Recall that +P means permuting input dimensions, +I means promoting learning of induction heads, and +G means passing example groups as input to in-context learning transformers.

Table 1 presents worst-group accuracies on the test set of `Waterbirds` for 3 sets of approaches: (a) in-context learners trained on `Waterbirds` itself, (b) in-context learners trained on `iNaturalist`, and (c) baselines. Similarly, Table 2 presents worst-group accuracies on the test set of `Waterbirds-severe` for 3 sets of approaches: (a) in-context learners trained on `Waterbirds-severe` itself, (b) in-context learners trained on `iNaturalist`, and (c) baselines. As RoPE-based transformers are not good at length extrapolation (Press et al., 2021), we do not attempt evaluating models trained on `iNaturalist` with context size 400 on 512-long sequences of `Waterbirds` or `Waterbirds-severe`. Finally, Table 3 presents minority-group accuracy on out-of-distribution classes of `iNaturalist` for two sets of approaches: (a) in-context learners trained on `iNaturalist` itself and (b) baselines.

Figure 17 presents *majority-group* accuracies on `Waterbirds` and `Waterbirds-severe` for the naive and proposed methods. Figure 18 presents *majority-group* accuracies on `Waterbirds` and `Waterbirds-severe` for the proposed approach and conventional methods.



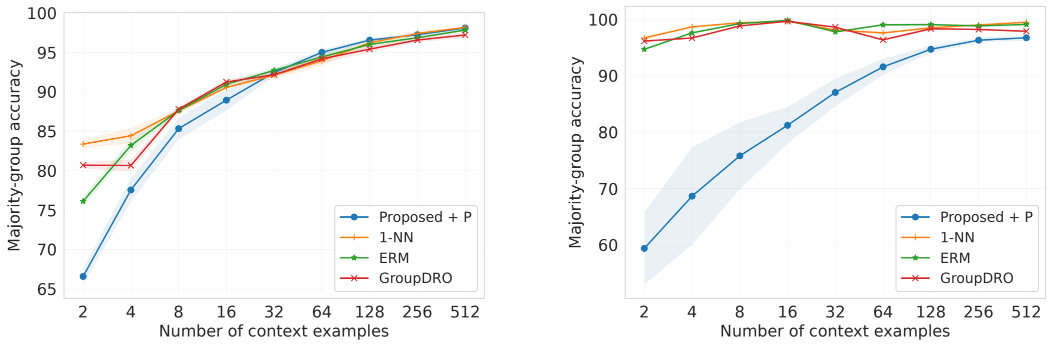
(a) `Waterbirds`

(b) `Waterbirds-severe`

Figure 17: Majority-group accuracies on `Waterbirds` and `Waterbirds-severe` as a function of context size for the naive and proposed approaches with or without permuting input dimensions. Shaded regions show standard deviation across 5 training runs.

Figure 19 demonstrates that it is important to simulate distribution shift at all intermediate query locations, not just for the main query. This makes sure that the learned network can make robust predictions for all intermediate context sizes.

Experiments on CelebA. To further verify our main findings presented in Section 2, we conduct experiments on `CelebA` (Liu et al., 2015). Here the task is to classify blond vs non-blond persons from face images, with sex being a spuriously correlated variable. Notably, the spurious correlation is asymmetric, in the sense that blond and non-blond women are almost equally represented, while blond men are much less represented compared to non-blond men. We follow the design of `Waterbirds` experiments in our `CelebA` experiments, with the only difference that we set the group distribution of context examples to $(0.25, 0.25, 0.05, 0.45)$, where group 0 are non-blond men, group 1 are non-blond women, group 2 are blond men, and group 4 are blond women. Table 4 presents worst-group accuracies on the test set of `CelebA` for 2 sets of approaches: (a) in-context learners trained on `CelebA` itself and (b) baseline algorithms. As in our `Waterbirds` experiments, we see that it is essential to permute input embeddings *and* to form ICL sequences in the proposed fashion. Unlike `Waterbirds`, comparing "Proposed + P" with "Proposed + P + G" we see that providing spurious annotations in-context provides significant gains. Figure 20a demonstrates that both of these approaches outperform 1-NN, ERM, and GroupDRO.



(a) `Waterbirds`

(b) `Waterbirds-severe`

Figure 18: Majority-group test accuracies on `Waterbirds` and `Waterbirds-severe` for the proposed approach and conventional methods such as 1-NN, ERM, and GroupDRO. Shaded regions show standard deviation across 5 training runs.

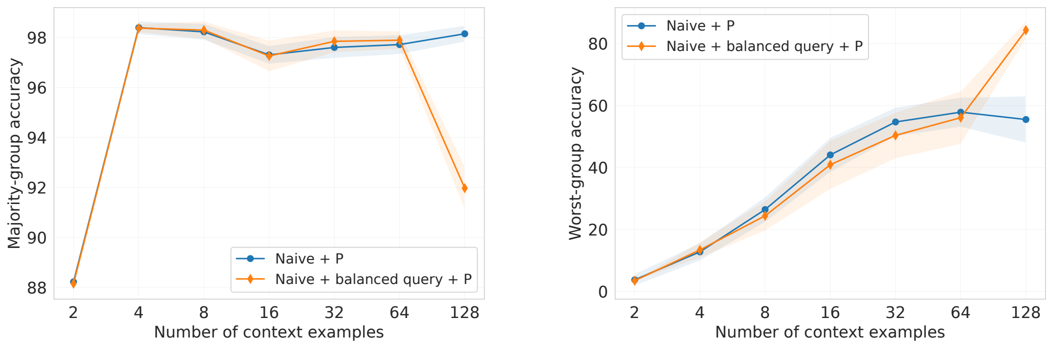
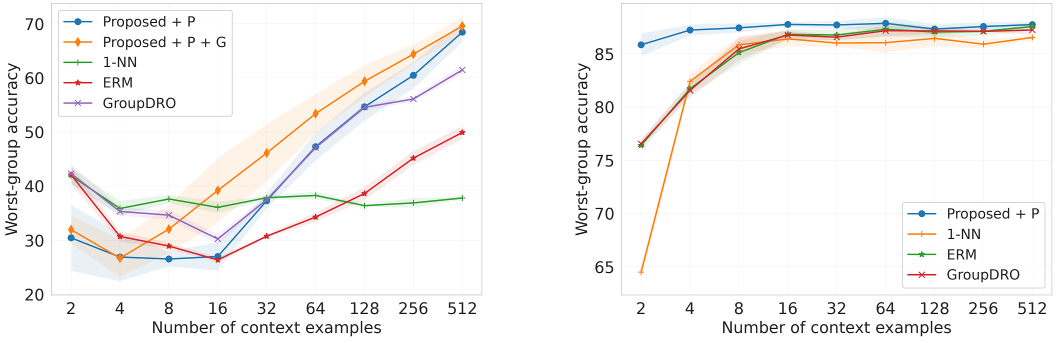


Figure 19: Majority-group and worst-group test accuracies on `Waterbirds-severe` as a function of context size for the naive approach with a single modification of making the last example (query) *group-balanced*. Shaded regions show standard deviation across 5 training runs. As expected, at intermediate context lengths this method performs similarly to the naive approach, but is much better at the training context length.

Experiments on MultiNLI. Beyond visual classification tasks, we also conduct experiments on a natural language inference (NLI) task. In particular, we consider the `MultiNLI` dataset (Williams et al., 2018b), where given a pair of sentences, a premise and a hypothesis, the task is to determine whether the hypothesis is entailed by, neutral with, or contradicts the premise. Prior work has observed that there is a spurious correlation between contradictions and the presence of the negation words *nobody*, *no*, *never*, and *nothing* (Gururangan et al., 2018). For our experiments, we frame a binary classification task *entails or neutral vs. contradicts*, with a binary spurious feature, which is 1 if and only if the hypothesis has one of the four negation words. We follow the design of `Waterbirds` experiments and set the group the group distribution of context examples to $(0.45, 0.05, 0.05, 0.45)$. Table 6 presents worst-group accuracies on the test set of `MultiNLI` for 2 sets of approaches: (a) in-context learners trained on `CelebA` itself and (b) baseline algorithms. Similar to `Waterbirds` and `CelebA` results, we see that the proposed approach of intermediate queries with balance group distribution helps to decrease the reliance on the spurious feature. However, in contrast to `Waterbirds` and `CelebA` experiments, permuting input embeddings has smaller effect on preventing in-weight learning. Figure 20b demonstrates the "Proposed + P" method outperforms 1-NN, ERM, and GroupDRO.

Experiments with a larger network. To verify that our findings generalize to larger models, we repeat CelebA experiments but with a transformer architecture of 12 layers with 12 multi-head attention (instead



(a) CelebA

(b) MultiNLI

Figure 20: Worst-group test accuracies on CelebA and MultiNLI for the proposed approach and conventional methods such as 1-NN, ERM, and GroupDRO. Shaded regions show standard deviation across 5 training runs.

of 6 layers with 8 multi-head attention). Due to memory increase, we decrease the batch size from 24 to 8. Besides these two changes, we keep all other experimental details the same. The complete results presented in Table 5 are qualitatively the same compared to the smaller network case (Table 4), with the difference that the results of transformer-based entries are lower. Furthermore, the standard deviation of the $+P$ approaches is significantly higher, indicating difficulties in optimization. We hypothesize that this is due to reusing learning rate and training length that were tuned for the smaller network with 3 times larger batch size.

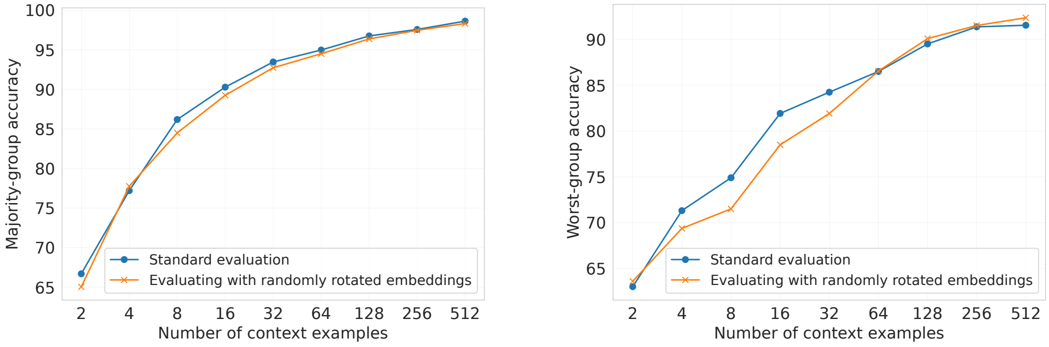


Figure 21: Majority-group and worst-group test accuracies on Waterbirds as a function of context size for a "Proposed + P" run evaluated on ICL sequences with randomly rotated input embeddings. Largely unchanged evaluation results fail to confirm that there is any data leakage when input embeddings are permuted during training.

On data leakage in single task regime. In the single task setting of Section 2, there is a potential for data leakage, not in the sense that individual examples might be leaked (we always evaluate on unseen examples), but in the sense that the learner effectively observes more data from the single task than its context length at evaluation. Indeed, when we do not permute input embeddings, we observe task memorization (i.e., data leakage) and the model does very well at evaluation with even close to empty context. To verify that there is no data leakage when we enable permuting input embeddings (+P), we take one of the "Proposed + P" runs trained on Waterbirds and evaluate it on ICL sequences where input embeddings of each sequence are rotated with a random *rotation matrix*. As the set of permutation matrices is a measure-zero subset of general rotation matrices, we expect that in case of data leakage we would observe degraded performance, as the model would be expecting randomly permuted embeddings of some memorized embedding space. In results presented in Figure 21, we see that under this new evaluation the results are the same (up to statistical noise), failing to confirm that there is any data leakage when input embeddings are permuted during training.

Finally, note that data leakage is not a concern in the multiple task setting of Section 3, because we evaluate on either unseen categories of `iNaturalist` or on unseen tasks such as Waterbirds and `Waterbirds-severe`.

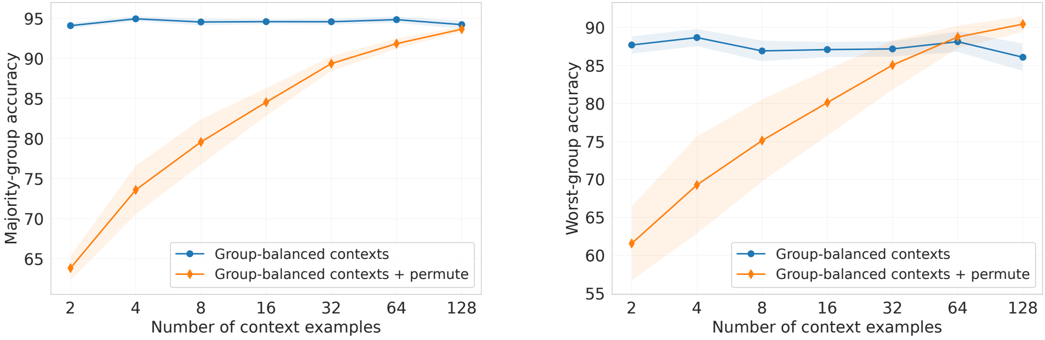


Figure 22: Majority-group and worst-group test accuracies on `Waterbirds` as a function of context size for the naive approach trained and evaluated on *group-balanced* contexts. The training is done on ICL sequences with 128 context examples. Shaded regions show standard deviation across 5 training runs.

Experiments with group-balanced contexts. As noted in Section 5, the proposed approach of *training* an in-context learner requires spurious annotations. Given access to spurious annotations, one can simply train an in-context learner on sequences with balanced groups. While in-context learners obtained this way will not be useful for new tasks for which we do not have spurious annotations (and thus cannot form group-balanced contexts), it is still useful to compare how well this approach does in the single task setting of Section 2. For this end, we train in-context learners on balanced-group sequences consisting of 128 `Waterbirds` examples. This way each group is represented with 32 context examples. Note that in our main `Waterbirds` experiments with 512 context examples but group-imbalanced contexts, the minority groups are represented with even fewer, 25 examples. As the group-balanced sampling context breaks the correlation between the label and spurious feature, we only consider the naive approach of forming ICL sequences (Figure 2a). The results presented in Figure 22 show that, as expected, group-balanced sampling improves worst-group accuracy. The naive approach, which again leads to in-weights learning, reaches 86.08 ± 1.87 worst-group accuracy with 128 group-balanced context examples, compared to 84.82 ± 1.26 worst-group accuracy on 512 *group-imbalanced* context examples (see Table 1). This positive effect of downsampling has been also observed in standard (not in-context) training settings (Nagarajan et al., 2021; Menon et al., 2021; Idrissi et al., 2022). Furthermore, we again see that the proposed technique of permuting embedding dimensions induces strong in-context learning and reaches 90.44 ± 1.10 worst-group accuracy with 128 group-balanced context examples. This is just a bit below the 91.95 ± 1.20 worst-group accuracy we get training “Proposed + P” on 512 *group-imbalanced* context examples (see Table 1).

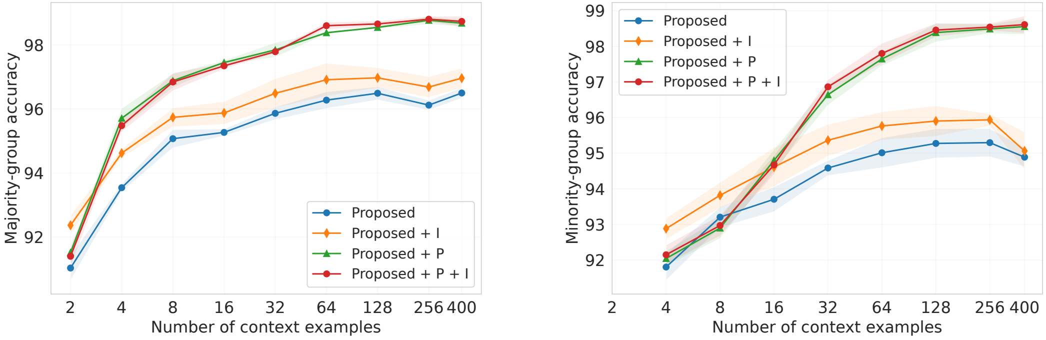


Figure 23: Majority-group and minority-group accuracies on the OOD test set of `iNaturalist` for the proposed approaches with or without permuting input dimensions and promoting induction heads. Shaded regions show standard deviation across 5 training runs.

Table 1: Complete results on `Waterbirds`. Reported numbers are average worst-group test accuracies, along with their standard deviation. The top half of in-context learners were trained on `Waterbirds` itself, while the ones in the bottom half were training on `iNaturalist`.

Method / Context size	4	8	16	32	64	128	256	512
Naive	87.02 (0.79)	84.52 (1.00)	85.14 (0.42)	84.82 (0.89)	83.41 (0.75)	84.45 (1.04)	85.08 (1.15)	84.82 (1.26)
Naive + P	70.92 (1.18)	75.32 (1.11)	80.66 (0.68)	83.24 (0.35)	86.87 (0.62)	89.87 (0.85)	91.94 (0.75)	92.60 (0.59)
Proposed	87.91 (1.29)	85.63 (2.20)	86.51 (2.17)	85.42 (1.73)	85.12 (2.34)	85.86 (2.22)	86.72 (1.89)	86.89 (1.82)
Proposed + I	88.18 (1.07)	85.89 (1.31)	86.68 (1.02)	86.01 (1.39)	84.82 (1.02)	85.92 (1.23)	86.07 (1.27)	86.46 (1.57)
Proposed + P	68.44 (2.40)	73.46 (2.53)	80.00 (2.06)	83.71 (2.15)	87.30 (1.79)	90.02 (1.16)	92.11 (1.65)	91.95 (1.20)
Proposed + P + I	68.05 (1.51)	72.47 (1.80)	78.97 (1.12)	82.58 (0.68)	86.39 (0.69)	90.00 (0.60)	91.78 (0.56)	92.17 (0.86)
Proposed + G	88.74 (1.01)	87.00 (1.60)	87.62 (1.58)	86.86 (1.31)	86.18 (1.33)	86.91 (0.98)	87.26 (1.11)	86.95 (1.21)
Proposed + G + I	88.89 (0.53)	87.49 (0.69)	87.70 (0.74)	86.90 (0.95)	86.03 (0.71)	86.64 (0.72)	87.29 (0.77)	87.35 (1.00)
Proposed + G + P	68.47 (2.32)	73.74 (2.00)	79.21 (1.68)	82.85 (1.33)	86.55 (1.17)	89.98 (0.72)	92.00 (0.82)	93.05 (0.40)
Proposed + G + P + I	68.24 (1.88)	73.78 (1.67)	80.23 (0.94)	83.02 (1.22)	86.94 (1.31)	89.89 (0.91)	92.46 (1.00)	92.69 (1.15)
1-NN	65.29 (1.23)	72.53 (1.11)	79.15 (1.16)	82.81 (0.63)	87.49 (1.18)	90.00 (1.05)	91.96 (0.51)	93.40 (0.27)
ERM	63.04 (1.22)	70.76 (1.01)	77.32 (1.16)	83.04 (1.09)	85.95 (1.38)	87.20 (0.77)	88.10 (0.98)	88.48 (0.45)
GroupDRO	64.61 (1.79)	71.52 (0.73)	77.81 (1.19)	83.45 (1.57)	87.34 (1.42)	88.30 (0.91)	89.79 (0.81)	91.12 (0.62)
Naive	69.77 (1.37)	77.98 (1.51)	79.23 (0.83)	81.20 (1.35)	82.57 (1.52)	83.85 (1.56)	84.21 (1.19)	-
Naive + P	66.47 (1.17)	73.12 (1.44)	77.85 (1.74)	81.76 (1.49)	86.36 (0.86)	88.02 (1.25)	89.68 (0.77)	-
Proposed	69.75 (5.51)	77.51 (3.01)	79.20 (2.11)	81.39 (1.49)	82.04 (1.29)	83.51 (0.97)	84.63 (0.80)	-
Proposed + I	70.73 (1.42)	77.10 (1.76)	78.90 (1.49)	80.86 (1.74)	82.22 (1.72)	84.22 (1.45)	84.69 (1.47)	-
Proposed + P	66.09 (1.49)	73.71 (1.17)	78.33 (0.69)	82.75 (0.83)	86.32 (0.52)	88.85 (0.72)	89.98 (1.35)	-
Proposed + P + I	65.51 (2.16)	70.91 (2.32)	75.94 (3.04)	81.51 (1.90)	86.41 (1.50)	89.39 (0.98)	91.08 (0.75)	-
Proposed + G	70.98 (2.52)	78.41 (1.25)	79.67 (1.26)	81.59 (1.42)	82.42 (1.28)	83.91 (1.64)	84.31 (1.31)	-
Proposed + G + I	71.94 (2.70)	78.56 (1.65)	80.62 (1.66)	82.31 (1.76)	83.52 (1.57)	84.52 (1.32)	85.35 (1.20)	-
Proposed + G + P	67.55 (0.78)	73.79 (0.33)	78.32 (0.93)	82.56 (1.31)	86.01 (1.09)	89.40 (1.22)	90.99 (1.15)	-
Proposed + G + P + I	69.18 (2.76)	74.13 (2.06)	79.18 (1.81)	83.17 (0.85)	87.25 (0.37)	90.67 (0.80)	92.23 (0.69)	-

Table 2: Complete results on `Waterbirds-severe`. Reported numbers are average worst-group test accuracies, along with their standard deviation. The top half of in-context learners were trained on `Waterbirds-severe` itself, while the ones in the bottom half were training on `iNaturalist`.

Method / Context size	4	8	16	32	64	128	256	512
Naive	83.04	80.78	80.78	79.43	80.50	80.29	81.67	82.02
	(1.92)	(1.58)	(1.85)	(2.77)	(2.43)	(2.30)	(2.25)	(2.72)
Naive + P	10.89	28.61	46.23	58.40	67.13	74.28	77.18	77.49
	(2.71)	(4.98)	(4.17)	(2.46)	(2.34)	(2.25)	(3.11)	(4.08)
Proposed	82.64	81.01	81.90	81.36	81.94	81.70	82.35	82.09
	(1.56)	(2.23)	(1.80)	(1.69)	(1.91)	(1.62)	(1.72)	(2.15)
Proposed + I	83.23	80.76	81.65	81.46	81.63	81.34	81.46	82.24
	(1.30)	(1.93)	(2.38)	(2.11)	(2.40)	(2.01)	(2.32)	(3.49)
Proposed + P	61.94	68.23	75.94	81.93	85.76	88.36	90.01	90.20
	(8.91)	(5.53)	(3.13)	(1.53)	(2.03)	(1.30)	(1.98)	(2.65)
Proposed + P + I	64.01	72.22	78.45	82.00	85.86	88.13	90.09	90.59
	(4.05)	(4.43)	(2.79)	(2.20)	(1.64)	(1.39)	(1.73)	(1.54)
Proposed + G	82.02	81.15	83.11	81.22	81.30	81.90	82.48	82.44
	(3.37)	(3.56)	(1.84)	(2.08)	(1.72)	(1.62)	(1.62)	(1.34)
Proposed + G + I	82.61	80.48	81.20	80.13	81.09	80.84	81.61	81.84
	(3.42)	(2.69)	(3.55)	(3.18)	(2.86)	(2.47)	(2.36)	(2.51)
Proposed + G + P	59.11	64.44	71.30	79.46	85.21	88.60	90.65	91.38
	(2.89)	(5.67)	(3.74)	(0.83)	(1.54)	(1.36)	(1.01)	(1.14)
Proposed + G + P + I	64.26	70.05	77.76	82.38	86.56	89.09	90.75	90.82
	(5.81)	(4.01)	(1.77)	(1.66)	(0.88)	(1.02)	(0.96)	(0.73)
1-NN	5.44	4.50	3.49	27.92	45.04	52.58	61.74	71.20
	(0.60)	(0.43)	(0.21)	(0.54)	(0.88)	(1.39)	(0.48)	(0.58)
ERM	6.81	4.35	1.87	35.30	29.52	45.84	65.35	75.69
	(0.44)	(0.26)	(0.24)	(1.55)	(1.49)	(1.00)	(0.53)	(0.88)
GroupDRO	7.42	5.26	2.75	17.62	45.47	65.13	78.57	86.89
	(0.57)	(0.35)	(0.29)	(0.65)	(1.18)	(1.06)	(0.77)	(0.57)
Naive	48.18	49.39	48.71	52.58	54.10	56.41	56.86	-
	(3.52)	(3.28)	(6.49)	(4.56)	(6.04)	(5.03)	(4.75)	
Naive + P	0.88	0.06	0.00	0.13	0.19	0.13	0.02	-
	(0.45)	(0.05)	(0.00)	(0.29)	(0.43)	(0.29)	(0.04)	
Proposed	49.04	53.39	54.82	59.44	61.04	62.26	63.77	-
	(2.76)	(4.74)	(8.82)	(10.75)	(12.23)	(12.31)	(12.38)	
Proposed + I	48.45	52.44	54.74	58.67	60.37	62.42	63.27	-
	(6.15)	(11.15)	(10.69)	(11.38)	(9.19)	(8.69)	(9.15)	
Proposed + P	1.88	0.27	0.06	0.08	0.30	0.15	0.03	-
	(0.56)	(0.21)	(0.10)	(0.13)	(0.60)	(0.28)	(0.06)	
Proposed + P + I	2.27	0.66	0.15	1.18	2.50	1.14	0.50	-
	(0.74)	(0.49)	(0.14)	(1.09)	(2.49)	(0.88)	(0.20)	
Proposed + G	50.00	52.31	53.69	57.87	59.11	60.33	62.30	-
	(5.03)	(5.05)	(4.54)	(3.33)	(3.16)	(3.36)	(3.01)	
Proposed + G + I	51.78	53.87	55.07	60.15	60.73	62.40	61.86	-
	(5.76)	(6.15)	(6.20)	(7.40)	(8.02)	(8.01)	(7.77)	
Proposed + G + P	1.52	0.16	0.00	0.10	0.04	0.03	0.36	-
	(0.69)	(0.13)	(0.00)	(0.20)	(0.05)	(0.05)	(0.73)	
Proposed + G + P + I	1.59	0.23	0.08	0.50	1.91	2.19	2.34	-
	(0.17)	(0.16)	(0.10)	(0.69)	(3.05)	(3.67)	(4.00)	

Table 3: Complete results on `iNaturalist`. Reported numbers are average minority-group accuracies on the OOD test set of `iNaturalist`, along with their standard deviation.

Method / Context size	4	8	16	32	64	128	256	400
Naive	91.91	92.81	93.76	94.77	94.85	95.04	94.84	94.53
	(0.21)	(0.47)	(0.35)	(0.28)	(0.44)	(0.48)	(0.75)	(0.38)
Naive + P	92.19	92.79	94.42	96.22	97.29	98.05	98.16	98.18
	(0.34)	(0.25)	(0.20)	(0.19)	(0.16)	(0.19)	(0.07)	(0.16)
Proposed	91.80	93.20	93.71	94.58	95.01	95.27	95.30	94.89
	(0.39)	(0.29)	(0.35)	(0.22)	(0.42)	(0.42)	(0.40)	(0.27)
Proposed + I	92.88	93.82	94.61	95.36	95.76	95.90	95.94	95.06
	(0.31)	(0.37)	(0.56)	(0.45)	(0.40)	(0.44)	(0.18)	(0.54)
Proposed + P	92.04	92.90	94.80	96.64	97.65	98.39	98.49	98.55
	(0.22)	(0.30)	(0.32)	(0.30)	(0.20)	(0.27)	(0.14)	(0.23)
Proposed + P + I	92.15	92.97	94.67	96.86	97.80	98.46	98.54	98.61
	(0.28)	(0.30)	(0.28)	(0.21)	(0.29)	(0.20)	(0.11)	(0.25)
Proposed + G	92.48	93.27	93.88	94.91	94.99	95.29	95.13	94.64
	(0.45)	(0.72)	(0.43)	(0.63)	(0.38)	(0.45)	(0.33)	(0.43)
Proposed + G + I	92.59	93.80	94.18	95.50	95.82	95.83	95.82	95.28
	(0.33)	(0.23)	(0.38)	(0.33)	(0.41)	(0.34)	(0.55)	(0.60)
Proposed + G + P	91.90	92.84	94.69	97.28	98.29	98.70	98.85	99.00
	(0.17)	(0.19)	(0.15)	(0.31)	(0.13)	(0.19)	(0.19)	(0.11)
Proposed + G + P + I	92.28	93.25	94.93	97.73	98.44	98.99	99.04	99.06
	(0.10)	(0.09)	(0.22)	(0.07)	(0.20)	(0.09)	(0.14)	(0.07)
1-NN	92.08	94.56	95.84	97.17	97.84	98.49	98.55	98.80
	(0.64)	(0.39)	(0.16)	(0.23)	(0.12)	(0.20)	(0.23)	(0.21)
ERM	89.67	92.98	94.65	96.17	96.88	97.70	98.15	98.43
	(0.43)	(0.30)	(0.17)	(0.24)	(0.23)	(0.21)	(0.17)	(0.11)
GroupDRO	91.20	93.79	95.33	97.39	97.85	98.46	98.91	99.01
	(0.55)	(0.39)	(0.18)	(0.20)	(0.20)	(0.13)	(0.20)	(0.18)

Table 4: Complete results on CelebA. Reported numbers are average worst-group test accuracies, along with their standard deviation. All in-context learners were trained on CelebA itself.

Method / Context size	4	8	16	32	64	128	256	512
Naive	24.88	24.56	25.80	25.14	23.85	25.62	25.84	26.20
	(2.03)	(2.26)	(1.98)	(2.11)	(1.91)	(1.63)	(2.10)	(1.42)
Naive + P	20.72	17.27	12.43	14.85	13.56	16.17	20.16	26.03
	(2.21)	(2.38)	(2.04)	(2.33)	(1.60)	(2.83)	(4.14)	(5.13)
Proposed	25.83	25.42	26.85	25.80	25.18	26.65	26.89	27.53
	(1.77)	(1.66)	(1.52)	(1.60)	(1.06)	(2.11)	(1.41)	(1.31)
Proposed + P	26.90	26.54	27.00	37.30	47.29	54.66	60.48	68.45
	(4.56)	(1.46)	(2.72)	(1.55)	(2.67)	(2.67)	(2.55)	(1.99)
Proposed + G	23.87	24.67	25.60	24.95	24.42	25.77	25.70	26.55
	(1.50)	(1.51)	(1.30)	(1.12)	(1.17)	(1.24)	(0.81)	(1.39)
Proposed + G + P	26.71	32.06	39.21	46.13	53.41	59.37	64.39	69.58
	(3.57)	(4.62)	(6.06)	(5.38)	(3.73)	(3.07)	(1.84)	(1.72)
1-NN	35.87	37.63	36.08	37.86	38.28	36.40	36.90	37.81
	(1.48)	(0.86)	(1.13)	(0.45)	(0.80)	(0.32)	(0.99)	(0.36)
ERM	30.70	28.93	26.37	30.75	34.29	38.64	45.18	49.92
	(1.05)	(0.64)	(0.66)	(0.40)	(0.93)	(1.29)	(1.39)	(1.28)
GroupDRO	35.32	34.64	30.24	37.49	47.11	54.56	56.11	61.47
	(0.88)	(0.99)	(1.01)	(0.60)	(0.59)	(0.66)	(0.60)	(0.93)

Table 5: Complete results on CelebA, but with larger network of 120m parameters, consisting of 12 layers (instead of 6 layers) with 12 multi-head attention (instead of 8 heads). Reported numbers are average worst-group test accuracies, along with their standard deviation. All in-context learners were trained on CelebA itself.

Method / Context size	4	8	16	32	64	128	256	512
Naive	24.43	21.79	23.68	23.02	23.99	23.18	22.62	20.22
	(0.57)	(0.86)	(0.58)	(0.49)	(0.84)	(1.04)	(0.83)	(1.00)
Naive + P	21.34	14.64	12.56	13.35	13.76	15.73	17.69	21.67
	(1.40)	(0.66)	(0.74)	(1.54)	(2.58)	(2.41)	(3.27)	(4.55)
Proposed	22.90	20.86	23.11	21.93	23.35	21.82	21.75	19.54
	(2.30)	(2.40)	(2.66)	(2.50)	(2.80)	(3.11)	(2.32)	(2.06)
Proposed + P	35.13	31.54	30.89	35.19	41.63	47.81	51.60	55.53
	(5.19)	(1.73)	(4.40)	(5.28)	(6.74)	(9.40)	(11.00)	(11.70)
Proposed + G	23.08	20.89	22.74	21.73	22.70	21.24	21.50	18.90
	(2.47)	(2.66)	(2.54)	(3.23)	(3.25)	(3.45)	(2.61)	(2.01)
Proposed + G + P	31.44	32.09	36.54	43.67	49.62	54.77	58.75	61.74
	(4.27)	(5.24)	(1.90)	(2.72)	(2.93)	(4.24)	(5.94)	(6.10)
1-NN	35.87	37.63	36.08	37.86	38.28	36.40	36.90	37.81
	(1.48)	(0.86)	(1.13)	(0.45)	(0.80)	(0.32)	(0.99)	(0.36)
ERM	30.70	28.93	26.37	30.75	34.29	38.64	45.18	49.92
	(1.05)	(0.64)	(0.66)	(0.40)	(0.93)	(1.29)	(1.39)	(1.28)
GroupDRO	35.32	34.64	30.24	37.49	47.11	54.56	56.11	61.47
	(0.88)	(0.99)	(1.01)	(0.60)	(0.59)	(0.66)	(0.60)	(0.93)

Table 6: Complete results on `MultiNLI`. Reported numbers are average worst-group test accuracies, along with their standard deviation. All in-context learners were trained on `MultiNLI` itself.

Method / Context size	2	4	8	16	32	64	128	256	512
Naive	84.15 (0.30)	84.58 (0.37)	84.74 (0.37)	84.13 (0.45)	83.92 (0.25)	85.20 (0.24)	84.38 (0.32)	84.30 (0.33)	84.05 (0.39)
Naive + P	85.30 (1.58)	87.29 (0.95)	87.43 (1.05)	87.56 (0.77)	87.50 (0.57)	88.05 (0.38)	87.35 (0.27)	87.99 (0.20)	87.54 (0.34)
Proposed	86.62 (0.65)	86.02 (0.33)	87.45 (0.58)	87.54 (0.42)	86.99 (0.26)	86.51 (0.27)	86.11 (0.29)	87.35 (0.19)	86.89 (0.30)
Proposed + P	85.84 (1.12)	87.22 (0.55)	87.43 (0.48)	87.75 (0.24)	87.70 (0.43)	87.86 (0.62)	87.31 (0.53)	87.56 (0.46)	87.73 (0.30)
Proposed + G	87.15 (0.82)	86.23 (0.65)	87.23 (0.42)	87.37 (0.35)	87.02 (0.27)	86.51 (0.58)	86.47 (0.49)	87.69 (0.51)	87.50 (0.55)
Proposed + G + P	85.41 (1.38)	87.32 (0.76)	87.57 (0.34)	87.81 (0.56)	87.74 (0.50)	87.97 (0.30)	87.56 (0.27)	87.78 (0.31)	87.68 (0.12)
1-NN	64.49 (0.90)	82.41 (0.81)	85.84 (0.84)	86.39 (0.76)	86.01 (0.40)	86.04 (0.93)	86.44 (0.81)	85.90 (0.43)	86.53 (0.29)
ERM	76.40 (0.65)	81.71 (0.58)	85.10 (1.13)	86.82 (0.54)	86.76 (0.54)	87.30 (0.78)	87.04 (0.55)	87.10 (0.48)	87.55 (0.41)
GroupDRO	76.54 (0.42)	81.56 (0.75)	85.45 (1.10)	86.75 (0.47)	86.55 (0.30)	87.16 (0.55)	87.14 (0.37)	87.10 (0.14)	87.23 (0.49)