A. Additional Ablations

In Section 5, we performed various ablations to understand the empirical effectiveness of FLYP . Here we present some additional ablations like (a) effect of changing the number of prompt templates used, (b) effect of batch-size, and (c) jointly optimizing both cross-entropy and contrastive loss. We observe that FLYP 's performance is quite robust to changes in number of prompt templates used or the batch-size as discussed in detail below.

Number of prompt templates. We test whether FLYP's gains come from sampling different text prompts while training. To do so, we experiment on ImageNet using a single text-description template instead of 80 templates as used in Wortsman et al. (2021), and observe that finetuning accuracy of FLYP is not affected by the number of text-templates used. We also note that for datasets like PatchCamelyon, SST2, and Flowers, experiments in Section 4 use only a single template as provided in Radford et al. (2021), and still outperform all the baselines.

Recall from Section 3 that for every given image-label pair (x, y), we construct a corresponding image-text pair (I, T), where T is sampled from a set of text-descriptions $\mathcal{T}_y$. For example, for every class, possible text descriptions in $\mathcal{T}_y$ can be "a photo of a {class}", "a painting of a {class}", "{class} in wild", etc. For all our experiments on all the datasets, we use the same text-templates as used in CLIP (Radford et al., 2021) and WiseFT (Wortsman et al., 2021). Figure 4 compares FLYP with baselines when only a single text-description template is used on ImageNet (the default template of "a photo of a {class}" as given in Radford et al. (2021)). We observe no change in the accuracy of FLYP both ID and OOD (without zeroshot ensembling) when using a single template or 80 templates. Note that since the corresponding zeroshot head is also constructed using a single text-description, there is a slight drop in ensembling accuracy (due to a decrease in the zeroshot model's performance) for all the baselines as expected.

Adding cross-entropy loss to FLYP . In Section 5 we observed that updating both the encoders using cross-entropy loss degrades the performance. Here we compare the performance of FLYP when the cross-entropy loss is added to FLYP's objective (i.e. the contrastive loss). On ImageNet, as shown in Figure 5, the weight ensembling curve for FLYP (orange) completely dominates (lies above and to the right) those of when the cross-entropy loss is added to FLYP 's objective under various regularization strengths. Similarly, on iWILDCam, adding cross-entropy loss (in equal weightage) leads to a *drop of* 2.64% ID and 0.5% OOD, as shown in Table 4. The performance degrades further, as the weight of cross-entropy loss is increased.

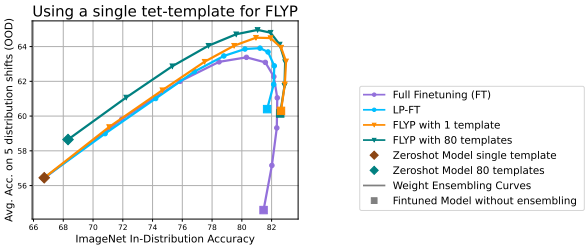
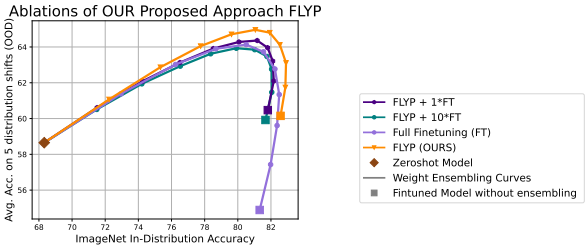


Figure 4. FLYP 's performance is unaffected by the number of prompt templates used. Here we compare using a single template versus 80 templates for text-descriptions on the ImageNet dataset. Observe that FLYP with a single template gives the same ID and OOD accuracy as FLYP with 80 templates, without ensembling. Note that the zeroshot model is also constructed using a single template, which causes a drop in it's accuracy, similar to the observations in Radford et al. (2021).



(a) Adding cross-entropy to FLYP (b) Legend

Figure 5. Adding cross-entropy loss to FLYP's objective degrades the performance on ImageNet.

FLYP with	iWILDCam		FMOW	
CE loss	ID	OOD	ID	OOD
FLYP	53.03	38.09	68.56	40.76
FLYP + FT	50.39	37.59	69.02	39.45
FLYP + 10 * FT	48	35.35	68.45	39.45
FT	47.99	34.77	68.47	39.53

Table 4. Adding cross-entropy loss to FLYP's objective (with various regularization) decreases the performance both ID and OOD.

Effect of batch-size on FLYP . Batch-size has been observed to cause variations in performance when using contrastive loss (Chen et al., 2020). As mentioned in Section 4, we use a fixed batch size of 512 for ImageNet and 256 for the rest of the datasets. We perform additional experiments with a lower batch-size (half of the previous). On ImageNet, a lower batch-size of 256 gives similar ID accuracy as batch-size of 512. However we observe a slight drop of 1% in the OOD accuracy. On iWILDCam, with a lower batch-size of 128, we get a similar ID and OOD accuracy as the batch-size of 256. However, on FMOW, we again observe a drop in OOD accuracy of 0.5%.

B. Additional Results

B.1. SOTA on WILDS-iWILDCam

Table 5 compares FLYP with the leaderboard on iWILDCam benchmark (Koh et al.). As discussed in Section 4, FLYP gives gains of 2.3% ID and 2.7% OOD over the top of the leaderboard, outperforming compute heavy ModelSoups (Wortsman et al., 2022), which ensembles 70+ different models finetuned using different augmentations and hyperparameters.

	Architecture	ID Macro F1	OOD Macro F1
FLYP	ViTL-336px	**59.9 (0.7)**	**46.0 (1.3)**
Model Soups	ViTL	57.6 (1.9)	43.3 (1.)
ERM	ViTL	55.8 (1.9)	41.4 (0.5)
ERM	PNASNet	52.8 (1.4)	38.5 (0.6)
ABSGD	ResNet50	47.5 (1.6)	33.0 (0.6)

Table 5. FLYP (with ensembling) achieves highest reported accuracy both ID and OOD on WILDS-iWILDCam. We compare FLYP with the top 4 entries on the leaderboard (Koh et al.).

B.2. Few-shot classification using CLIP ViT-L/14

In Section 4.2.1, we considered a challenging task of binary few-shot classification on 2 datasets of PatchCamelyon and SST2, using CLIP ViT-B/16. Here we perform a similar comparison, although using a much bigger model of CLIP ViT-L/14. Table 6 compares FLYP with baselines on 2 datasets of SST2 and PatchCamelyon. We observe that similar to the case of using smaller CLIP ViT-B/16 (Table 2), FLYP outperforms the baselines when a larger model of CLIP ViT-L/14 is used. For example, under 32-shot classification, FLYP outperforms LP-FT by 4.2% on SST2 and 3.1% on PatchCamelyon.

Methods	SST2			PatchCamelyon		
	4 Shot	16 Shot	32 Shot	4 Shot	16 Shot	32 Shot
Zeroshot	68.9 (-)	68.9 (-)	68.9 (-)	62 (-)	62 (-)	62 (-)
LP	69.2 (0.2)	70.2 (0.4)	71.0 (0.4)	66.9 (0.8)	71.2 (1.1)	74.1 (0.9)
FT	69.3 (0.1)	69.5 (0.2)	70.0 (0.4)	65.9 (0.7)	69.9 (0.7)	71.8 (0.8)
LP-FT	69.3 (0.3)	70.7 (0.4)	71.3 (0.4)	67.5 (0.8)	72.9 (1.0)	76.0 (0.5)
FLYP	**69.8 (0.7)**	**73.2 (0.8)**	**75.5 (0.6)**	**67.8 (1.2)**	**75.8 (0.9)**	**79.1 (0.7)**

Table 6. Binary few-shot classification using CLIP ViT-L/14. FLYP continues to outperforms the baselines. For example, under 32-shot classification, FLYP outperforms LP-FT by 4.2% on SST2 and 3.1% on PatchCamelyon.

B.3. ImageNet Distribution Shifts - Detailed Results

Table 7 gives the detailed results on each individual associated distribution shifts with the ImageNet dataset, in the same experiment setting as Section 4.1. Observe that with weight ensembling, FLYP consistently outperforms the baselines across all the distribution shifts.

C. Hyperparameter details

For all algorithms on all the datasets (apart from ImageNet), we perform a hyper-parameter sweep over 5 learning rates in $\{1e^{-2}, 1e^{-3}, ..., 1e^{-6}\}$ and 5 weight decay in $\{0.0, 0.1, ..., 0.4\}$, with a fixed batch-size of 256. For ImageNet, due to computational cost, we perform a sweep over 3 learning rates in {1e-4, 1e-5, 1e-6} and 2 weight decay in $\{0.0, 0.1\}$ and use a batch-size of 512. L2-SP requires tuning a additional regularization term weight $\lambda \in \{1e^{-1}, 1e^{-2}, ..., 1e^{-4}\}$.

We early stop and choose the best hyper-parameter based on the ID Validation accuracy. For the datasets which do not have a publicly available validation split, we split the training dataset in 80:20 ratio to make a training and a validation set. In all the cases, note that the OOD dataset is used only for evaluation purposes.

For the k few-shot classification setting (where $k \in \{4, 16, 32\}$), we randomly sample k training and k validation points from the respective full datasets and repeat the process 50 times, due to increased variance caused by a small training and validation set. We finally report the mean test accuracy over the 50 runs, corresponding to the hyperparameter with the lowest mean validation loss over the 50 runs.

Methods	ImageNet Distribution Shifts							ImageNet Distribution Shifts (with ensembling)						
	ID	Im-V2	Im-R	Im-A	Sketch	ObjectNet	Avg. OOD	ID	Im-V2	Im-R	Im-A	Sketch	ObjectNet	Avg. OOD
Zeroshot	68.3	61.9	77.7	50.0	48.3	55.4	58.7	68.3	61.9	77.7	50.0	48.3	55.4	58.7
LP	79.9	69.8	70.8	46.4	46.9	52.1	57.2	80.0	70.3	72.4	47.8	48.1	52.8	58.3
FT	81.3	71.2	66.1	37.8	46.1	53.3	54.9	82.5	72.8	74.9	48.1	51.9	59.0	61.3
L2-SP	81.7	71.8	70.0	42.5	48.5	56.2	57.8	82.2	72.9	75.1	48.6	51.4	58.9	61.4
LP-FT	81.7	72.1	**73.5**	47.6	**50.3**	58.2	**60.3**	82.1	72.8	75.3	50.1	51.7	59.2	61.8
FLYP	**82.6**	**73.0**	71.4	**48.1**	49.6	**58.7**	60.2	**82.9**	**73.5**	**76.0**	**53.0**	**52.3**	**60.8**	**63.1**

Table 7. FLYP compared with baselines on each of the associated distribution shifts with the ImageNet dataset. Note that with weight ensembling, FLYP consistently outperforms all the baselines on all the distribution shift benchmarks as well as ID.