Rethinking Few-Shot Adaptation of Vision-Language Models in Two Stages

Supplementary Material

This Supplementary Material aims to expand and complement the work's main body. We structure it as follows:

- **Generalization to other PEFT techniques.** Appendix A shows that the benefit of the two-stage design of 2SFS is not limited to Layer Normalization. In Appendix A.1 we experiment with LoRA, with particular emphasis on the relationship with CLIP-LoRA [47]; Appendix A.2 experiments with Prompt Learning techniques: CoOp [52] and "Independent Vision-Language Prompting" (IVLP);
- **Additional visual backbones.** Appendix B complements Sec. 5.1, by reporting results in the *base-to-novel* setting also for the ViT-B/32 and ViT-L/14 backbones;
- **Robustness to shots availability.** Appendix C reports the results of the main paper, with varying numbers of available shots. We further experiment with k=$\{4, 8\}$ shots;
- **Hyperparameter analysis.** Appendix D analyzes the impact of the total number of allowed iterations m;
- **Extended preliminary analysis.** Appendix E reports additional evidence for the preliminary observations of Sec. 3.2, with a focus on identifying the "best" and "worst" cases. We devote explanations for the slight failures of Fig. 5 with Oxford Pets [32] and Food-101 [1];
- **Limitations** are discussed in Appendix F.

A. 2SFS with different PEFT techniques

A.1. LoRA

We expand Sec. 5 reporting results of Tab. 1 and Tab. 2 when LoRA [12] is applied in the first stage. Note that 2SFS$_{\texttt{LoRA}}$ can be seen as a two-stage variant of the recently introduced CLIP-LoRA strategy, hence we are particularly interested in understanding the benefits, if any, relative to it.

Implementation details. Following Sec. 5, we perform a sweep for α in $[0.2, 0.8]$ with a step size of 0.1, only for the base-to-novel setting with the ViT-B/16 backbone, validating on ImageNet. We obtain an optimal value of $\alpha = 0.3$, which aligns with the behavior highlighted by Fig. 2 (*i.e.*, LoRA tends to saturate more quickly than LayerNorm). We transfer it to all other experiments in this Supplementary Material when 2SFS$_{\texttt{LoRA}}$ is specified. We strictly follow the CLIP-LoRA recipe for plugging low-rank modules in CLIP.

A.1.1. Base-to-novel generalization

Extended results for the base-to-novel setup are given in Tab. 3. We observe that, w.r.t. CLIP-LoRA, 2SFS$_{\texttt{LoRA}}$ provides a notable improvement, especially pronounced in the **Novel** metric ($+3.02\%$ on average) and, more in general, it boosts performance on 32 out of 33 $\{$dataset, metric$\}$ combinations. Overall, these results take a step further in confirming the hypothesis that the two-stage design is beneficial for other PEFT strategies, since 2SFS$_{\texttt{LoRA}}$ exhibits the 2nd greatest Harmonic Mean on average, only outperformed by MMA [42] among published works (excluding our own alternative 2SFS$_{\texttt{LayerNorm}}$).

A.1.2. All-to-all adaptation

Tab. 4 reports results for the all-to-all scenario. Following Sec. 5.2, we experiment with ViT-B/16, ViT-B/32, and ViT-L/14, inheriting hyperparameters from the base-to-novel setup. Here, 2SFS$_{\texttt{LoRA}}$ outperforms, on average, all strategies for all backbones, including 2SFS$_{\texttt{LayerNorm}}$. This further aligns with the evidence of Sec. 3.2, where LoRA is shown to incorporate more helpful knowledge to discriminate among available categories. Compared to CLIP-LoRA, 2SFS$_{\texttt{LoRA}}$ outperforms it in 27 out of 33 $\{$dataset, backbone$\}$ combinations, further supporting the benefits of the two-stage design.

A.2. Prompt Learning

Inspired by "prefix tuning" [23] for Language Models, Prompt Learning has become arguably the most widely adopted approach to adapt VLMs [2, 15, 16, 36, 43, 51, 52] in recent years. We do, hence, explore here if 2SFS also successfully integrates with Prompt Learning techniques.

Implementation details. To do so, we focus on CoOp [52] and "Independent Vision-Language Prompting" (IVLP), a baseline introduced in [16]. We do not conduct any hyperparameter tuning for these PEFT methods, but we set α=0.3 as used with LoRA, simply leveraging prior knowledge that excessive Prompt Learning severely harms generalization [51]. To exactly compare with both approaches, we use the same batch size, optimizer setup, and number of epochs suggested in the original papers.[1] When switching to the second stage of 2SFS, we train the linear classifier with the same optimizer setup described in the main paper.

Results with the ViT-B/16 backbone and k=16 shots are given in Tab. 5 (*base-to-novel* generalization). For all benchmarks, *wrapping prompt learning approaches in the two-stage design improves the Harmonic Mean* between seen/unseen semantic categories, providing further evidence to support our findings. The gap is particularly evident with CoOp ($+3.56$ overall HM), although also IVLP significantly benefits from this design ($+1.23$ overall HM). We also emphasize that the design of 2SFS is computationally friendlier than the original approaches: only the gradient w.r.t. the classifier is required for the second stage.

[1] with the only exception of ImageNet, for which we only train for 10 epochs to save computational resources.

Table 3. Experiments in *base-to-novel* generalization with the ViT-B/16 visual backbone. All methods use k=16 shots per base class. "CLIP" refers to zero-shot performance with dataset-specific templates, *e.g.*, "*a photo of a* {}, *a type of flower*" for Oxford Flowers. To highlight the benefits of the two-stage design, we report an additional line with the absolute improvement of 2SFS$_{\text{LoRA}}$ relative to its single-stage counterpart CLIP-LoRA [47]. In each table, the best performer is **bold**, and the second best is underlined.

Average across datasets.

Method	Base	Novel	HM
CLIP [33]	69.34	74.22	71.70
CoOP [52]	82.69	63.22	71.66
CoCoOp [51]	80.47	71.69	75.83
MaPLe [16]	82.28	75.14	78.55
ProGrad [53]	82.48	70.75	76.16
KgCoOp [43]	80.73	73.60	77.00
CLIP-LoRA [47]	85.32	70.63	77.28
MMA [42]	83.20	76.80	79.87
2SFS$_{\texttt{LayerNorm}}$	85.55	75.48	**80.20**
2SFS$_{\text{LoRA}}$	85.97	73.65	79.33
	+0.65	+3.02	+2.05

ImageNet

Method	Base	Novel	HM
CLIP [33]	72.43	68.14	70.22
CoOp [52]	76.47	67.88	71.92
CoCoOp [51]	75.98	70.43	73.10
MaPLe [16]	76.66	70.54	73.47
ProGrad [53]	77.02	66.66	71.46
KgCoOp [43]	75.83	69.96	72.78
CLIP-LoRA [47]	77.58	68.76	72.91
MMA [42]	77.31	71.00	74.02
2SFS$_{\texttt{LayerNorm}}$	77.71	70.99	74.20
2SFS$_{\text{LoRA}}$	77.70	71.60	**74.53**
	+0.12	+2.84	+1.62

Caltech101

Method	Base	Novel	HM
CLIP [33]	96.84	94.00	95.40
CoOP [52]	98.00	89.81	93.73
CoCoOp [51]	97.96	93.81	95.84
MaPLe [16]	97.74	94.36	96.02
ProGrad [53]	98.02	93.89	95.91
KgCoOp [43]	97.72	94.39	96.03
CLIP-LoRA [47]	98.19	93.05	95.55
MMA [42]	98.40	94.00	96.15
2SFS$_{\texttt{LayerNorm}}$	98.71	94.43	**96.52**
2SFS$_{\text{LoRA}}$	98.45	94.47	96.42
	+0.26	+1.42	+0.87

Oxford Flowers

Method	Base	Novel	HM
CLIP [33]	72.08	77.80	74.83
CoOP [52]	97.60	59.67	74.06
CoCoOp [51]	94.87	71.75	81.71
MaPLe [16]	95.92	72.46	82.56
ProGrad [53]	95.54	71.87	82.03
KgCoOp [43]	95.00	74.73	83.65
CLIP-LoRA [47]	97.91	68.61	80.68
MMA [42]	97.77	75.93	85.48
2SFS$_{\texttt{LayerNorm}}$	98.29	76.17	**85.83**
2SFS$_{\text{LoRA}}$	98.04	70.95	82.32
	+0.13	+2.34	+1.64

Oxford Pets

Method	Base	Novel	HM
CLIP [33]	91.17	97.26	94.12
CoOP [52]	93.67	95.29	94.47
CoCoOp [51]	95.20	97.69	96.43
MaPLe [16]	95.43	97.76	96.58
ProGrad [53]	95.07	97.63	96.33
KgCoOp [43]	94.65	97.76	96.18
CLIP-LoRA [47]	94.36	95.71	95.03
MMA [42]	95.40	98.07	**96.72**
2SFS$_{\texttt{LayerNorm}}$	95.32	97.82	96.55
2SFS$_{\text{LoRA}}$	95.50	97.13	96.31
	+1.14	+1.42	+1.28

Stanford Cars

Method	Base	Novel	HM
CLIP [33]	63.37	74.89	68.65
CoOP [52]	78.12	60.40	68.13
CoCoOp [51]	70.49	73.59	72.01
MaPLe [16]	72.94	74.00	73.47
ProGrad [53]	77.68	68.63	72.88
KgCoOp [43]	71.76	75.04	73.36
CLIP-LoRA [47]	83.93	65.54	73.60
MMA [42]	78.50	73.10	75.70
2SFS$_{\texttt{LayerNorm}}$	82.50	74.80	**78.46**
2SFS$_{\text{LoRA}}$	83.87	70.64	76.69
	−0.06	+5.10	+3.09

Food 101

Method	Base	Novel	HM
CLIP [33]	90.10	91.22	90.66
CoOP [52]	88.33	82.26	85.19
CoCoOp [51]	90.70	91.29	90.99
MaPLe [16]	90.71	92.05	**91.38**
ProGrad [53]	90.37	89.59	89.98
KgCoOp [43]	90.50	91.70	91.09
CLIP-LoRA [47]	86.84	86.67	86.76
MMA [42]	90.13	91.30	90.71
2SFS$_{\texttt{LayerNorm}}$	89.11	91.34	90.21
2SFS$_{\text{LoRA}}$	88.47	89.96	89.21
	+1.61	+3.29	+2.45

FGVC Aircraft

Method	Base	Novel	HM
CLIP [33]	27.19	36.29	31.09
CoOP [52]	40.44	22.30	28.75
CoCoOp [51]	33.41	23.71	27.74
MaPLe [16]	37.44	35.61	36.50
ProGrad [53]	40.54	27.57	32.82
KgCoOp [43]	36.21	33.55	34.83
CLIP-LoRA [47]	50.10	26.03	34.26
MMA [42]	40.57	36.33	38.33
2SFS$_{\texttt{LayerNorm}}$	47.48	35.51	**40.63**
2SFS$_{\text{LoRA}}$	51.00	31.37	38.85
	+0.90	+5.34	+ 4.59

SUN 397

Method	Base	Novel	HM
CLIP [33]	69.36	75.35	72.23
CoOP [52]	80.60	65.89	72.51
CoCoOp [51]	79.74	76.86	78.27
MaPLe [16]	80.82	78.70	79.75
ProGrad [53]	81.26	74.17	77.55
KgCoOp [43]	80.29	76.53	78.36
CLIP-LoRA [47]	81.11	74.53	77.68
MMA [42]	82.27	78.57	80.38
2SFS$_{\texttt{LayerNorm}}$	82.59	78.91	**80.70**
2SFS$_{\text{LoRA}}$	82.43	79.05	**80.70**
	+1.32	+4.52	+3.02

DTD

Method	Base	Novel	HM
CLIP [33]	53.24	59.90	56.37
CoOP [52]	79.44	41.18	54.24
CoCoOp [51]	77.01	56.00	64.85
MaPLe [16]	80.36	59.18	68.16
ProGrad [53]	77.35	52.35	62.45
KgCoOp [43]	77.55	54.99	64.35
CLIP-LoRA [47]	83.95	62.84	71.39
MMA [42]	83.20	65.63	73.38
2SFS$_{\texttt{LayerNorm}}$	84.60	65.01	**73.52**
2SFS$_{\text{LoRA}}$	84.53	63.53	72.54
	+0.58	+0.69	+1.15

EuroSAT

Method	Base	Novel	HM
CLIP [33]	56.48	64.05	60.03
CoOP [52]	92.19	54.74	68.69
CoCoOp [51]	87.49	60.04	71.21
MaPLe [16]	94.07	73.23	82.35
ProGrad [53]	90.11	60.89	72.67
KgCoOp [43]	85.64	64.34	73.48
CLIP-LoRA [47]	97.04	62.50	76.03
MMA [42]	85.46	82.34	**83.87**
2SFS$_{\texttt{LayerNorm}}$	96.91	67.09	79.29
2SFS$_{\text{LoRA}}$	97.05	64.59	77.56
	+0.01	+2.09	+1.53

UCF101

Method	Base	Novel	HM
CLIP [33]	70.53	77.50	73.85
CoOP [52]	84.69	56.05	67.46
CoCoOp [51]	82.33	73.45	77.64
MaPLe [16]	83.00	78.66	80.77
ProGrad [53]	84.33	74.94	79.35
KgCoOp [43]	82.89	76.67	79.65
CLIP-LoRA [47]	87.52	72.74	79.45
MMA [42]	86.23	80.03	82.20
2SFS$_{\texttt{LayerNorm}}$	87.85	78.19	**82.74**
2SFS$_{\text{LoRA}}$	88.59	76.82	82.28
	+1.07	+4.08	+2.83

Table 4. *All-to-all* experiments, where train/test categories coincide, with the ViT-B/16 (top), ViT-B/32 (middle), and ViT-L/14 (bottom) backbones. All methods use $k = 16$ shots per class. To highlight the benefits of the two-stage design, we report an additional line with the absolute improvement of $2SFS_{LoRA}$ relative to its single-stage counterpart CLIP-LoRA [47]. In each group, the best performer is marked by **bold text**; the second best is underlined.

BACKBONE	METHOD	IMAGENET	SUN	AIR	ESAT	CARS	FOOD	PETS	FLWR	CAL	DTD	UCF	MEAN
ViT-B/16	*Zero-Shot*	66.7	62.6	24.7	47.5	65.3	86.1	89.1	71.4	92.9	43.6	66.7	65.1
	CoOp [52] (ctx=16)	71.9	74.9	43.2	85.0	82.9	84.2	92.0	96.8	95.8	69.7	83.1	80.0
	CoCoOp [51]	71.1	72.6	33.3	73.6	72.3	**87.4**	93.4	89.1	95.1	63.7	77.2	75.4
	TIP-Adapter-F [50]	73.4	76.0	44.6	85.9	82.3	86.8	92.6	96.2	95.7	70.8	83.9	80.7
	CLIP-Adapter [10]	69.8	74.2	34.2	71.4	74.0	87.1	92.3	92.9	94.9	59.4	80.2	75.5
	PLOT++ [2]	72.6	76.0	46.7	92.0	84.6	87.1	93.6	97.6	96.0	71.4	85.3	82.1
	KgCoOp [43]	70.4	73.3	36.5	76.2	74.8	87.2	93.2	93.4	95.2	68.7	81.7	77.3
	TaskRes [45]	73.0	76.1	44.9	82.7	83.5	86.9	92.4	97.5	95.8	71.5	84.0	80.8
	MaPLe [16]	71.9	74.5	36.8	87.5	74.3	**87.4**	93.2	94.2	95.4	68.4	81.4	78.6
	ProGrad [53]	72.1	75.1	43.0	83.6	82.9	85.8	92.8	96.6	95.9	68.8	82.7	79.9
	LP++ [13]	73.0	76.0	42.1	85.5	80.8	87.2	92.6	96.3	95.8	71.9	83.9	80.5
	CLIP-LoRA [47]	73.6	76.1	**54.7**	92.1	86.3	84.2	92.4	**98.0**	**96.4**	72.0	86.7	83.0
	MMA [42]	73.2	76.6	44.7	85.0	80.2	87.0	**93.9**	96.8	95.8	72.7	85.0	81.0
	2SFS LayerNorm	73.7	**77.0**	50.0	92.4	85.4	86.1	93.7	97.7	**96.4**	73.2	86.6	82.9
	2SFS LoRA	**73.8**	76.9	54.6	**92.7**	**86.9**	85.7	93.8	**98.0**	**96.4**	**73.5**	**87.3**	**83.6**
		+0.2	+0.8	-0.1	+0.6	+0.6	+1.5	+1.4	0.0	0.0	+1.5	+0.6	+0.6
ViT-B/32	*Zero-Shot*	61.9	62.0	19.3	45.1	60.4	80.5	87.5	67.0	91.1	42.6	62.2	61.8
	CoOp [52] (ctx=16)	66.8	72.2	32.9	83.3	76.0	78.6	88.7	95.4	94.9	65.3	78.6	75.7
	CoCoOp [51]	66.0	69.8	22.6	70.4	64.6	**81.9**	91.0	82.5	94.3	59.7	75.3	70.7
	TIP-Adapter-F [50]	68.4	74.1	34.8	83.4	77.0	81.7	90.4	94.3	95.1	68.0	80.5	77.1
	CLIP-Adapter [10]	64.9	71.8	26.7	64.7	68.9	**81.9**	90.1	88.7	94.8	58.1	76.5	71.6
	PLOT++ [2]	67.4	73.4	36.3	91.1	77.4	79.7	89.1	96.3	94.9	67.0	81.5	77.6
	KgCoOp [43]	65.4	71.0	23.7	70.1	67.3	81.7	90.8	86.1	94.4	65.1	77.5	72.1
	TaskRes [45]	68.2	73.6	37.0	77.7	78.0	81.4	89.4	95.5	95.7	68.3	80.6	76.9
	MaPLe [16]	66.7	72.0	28.0	83.3	66.9	82.1	**91.7**	89.0	95.1	63.4	77.3	74.1
	ProGrad [53]	66.9	73.2	33.3	81.0	76.1	80.1	89.3	95.1	95.0	65.8	79.6	75.9
	LP++ [13]	68.1	74.0	34.3	82.8	75.2	81.8	90.5	93.9	95.0	67.8	80.1	76.7
	CLIP-LoRA [47]	68.4	74.0	**44.9**	91.8	79.7	78.2	88.8	96.2	95.2	68.2	**82.8**	78.9
	MMA [42]	68.0	74.0	34.0	80.1	73.5	81.4	91.5	94.3	95.6	68.9	81.7	76.7
	2SFS LayerNorm	68.4	**74.8**	40.2	92.1	80.2	80.8	90.3	96.3	**95.8**	**70.4**	82.3	79.2
	2SFS LoRA	**68.6**	74.5	44.4	**92.2**	**81.3**	80.2	90.0	**96.4**	95.7	69.6	82.5	**79.6**
		+0.2	+0.5	-0.5	+0.4	+1.6	+2.0	+1.2	+0.2	+0.5	+1.4	-0.3	+0.7
ViT-L/14	*Zero-Shot*	72.9	67.6	32.6	58.0	76.8	91.0	93.6	79.4	94.9	53.6	74.2	72.2
	CoOp [52] (ctx=16)	78.2	77.5	55.2	88.3	89.0	89.8	94.6	**99.1**	97.2	74.4	87.3	84.6
	CoCoOp [51]	77.8	76.7	45.2	79.8	82.7	91.9	95.4	95.3	97.4	71.4	85.2	81.7
	TIP-Adapter-F [50]	79.3	79.6	55.8	86.1	88.1	91.6	94.6	98.3	97.5	74.0	87.4	84.8
	CLIP-Adapter [10]	76.4	78.0	46.4	75.8	83.8	91.6	94.3	97.3	97.3	71.3	86.1	81.7
	PLOT++ [2]	78.6	79.1	44.1	92.2	87.2	90.2	93.6	98.8	97.5	75.0	87.1	83.9
	KgCoOp [43]	76.8	76.7	47.5	83.6	83.2	91.7	95.3	96.4	97.4	73.6	86.4	82.6
	TaskRes [45]	78.1	76.9	55.0	84.3	87.6	91.5	94.7	97.8	97.3	74.4	86.6	84.0
	MaPLe [16]	78.4	78.8	46.3	85.4	83.6	**92.0**	95.4	97.4	97.2	72.7	86.5	83.1
	ProGrad [53]	78.4	78.3	55.6	89.3	88.8	90.8	94.9	98.7	97.5	73.7	87.7	84.9
	LP++ [13]	79.3	79.7	54.6	89.3	87.7	91.7	94.9	98.5	97.4	76.1	88.1	85.2
	CLIP-LoRA [47]	79.6	79.4	66.2	93.1	90.9	89.9	94.3	99.0	97.3	76.5	89.9	86.9
	MMA [42]	**79.9**	80.2	56.4	76.3	88.0	**92.0**	**95.5**	98.4	**97.6**	75.8	88.0	84.4
	2SFS LayerNorm	79.4	80.3	64.1	92.9	90.3	91.1	**95.5**	**99.1**	97.5	**78.0**	89.5	87.1
	2SFS LoRA	79.7	**80.7**	**66.5**	**93.2**	**91.2**	90.8	**95.5**	99.0	97.5	77.2	**90.3**	**87.4**
		+0.1	+1.3	+0.3	+0.1	+0.3	+0.9	+1.2	0.0	+0.2	+0.7	+0.4	+0.5

B. Base-to-novel generalization with different backbones

This Appendix complements Sec. 5.1, where results are given for the ViT-B/16 backbone mimicking the experimental setup of [16, 42]. Specifically, here we focus on the comparison with the best competitor MMA [42] and expand the experimental evaluation to the ViT-B/32 and ViT-L/14 backbones further.

Implementation Details. To align with Sec. 5.1, we use layer normalization in the first stage as in the main body of

Table 5. Direct comparison between established prompt learning approaches (CoOp [52] and IVLP [16]) and their behavior when wrapped in the Two-Stage design of 2SFS. We examine the *base-to-novel* setting with ViT-B/16 and k=16 shots per class.

Average across datasets.

Method	Base	Novel	HM
CoOp [52]	82.69	63.22	71.66
2SFS $_{\text{CoOp}}$	83.49	68.27	**75.12**
IVLP	84.21	71.79	77.51
2SFS $_{\text{IVLP}}$	84.53	73.69	**78.74**

ImageNet

Method	Base	Novel	HM
CoOp [52]	76.47	67.88	71.92
2SFS $_{\text{CoOp}}$	77.44	71.11	**74.14**
IVLP	77.00	66.50	71.37
2SFS $_{\text{IVLP}}$	75.85	71.05	**73.37**

Caltech101

Method	Base	Novel	HM
CoOp [52]	98.00	89.81	93.73
2SFS $_{\text{CoOp}}$	98.00	91.99	**94.90**
IVLP	98.30	93.20	95.68
2SFS $_{\text{IVLP}}$	98.45	94.32	**96.34**

Oxford Flowers

Method	Base	Novel	HM
CoOp [52]	97.60	59.67	74.06
2SFS $_{\text{CoOp}}$	98.16	69.46	**81.35**
IVLP	97.97	72.10	83.07
2SFS $_{\text{IVLP}}$	98.1	72.93	**83.66**

Oxford Pets

Method	Base	Novel	HM
CoOp [52]	93.67	95.29	94.47
2SFS $_{\text{CoOp}}$	93.35	96.96	**95.12**
IVLP	94.90	97.20	96.04
2SFS $_{\text{IVLP}}$	95.46	97.61	**96.53**

Stanford Cars

Method	Base	Novel	HM
CoOp [52]	78.12	60.40	68.13
2SFS $_{\text{CoOp}}$	80.15	67.87	**73.50**
IVLP	79.53	71.47	75.28
2SFS $_{\text{IVLP}}$	81.69	73.5	**77.38**

Food 101

Method	Base	Novel	HM
CoOp [52]	88.33	82.26	85.19
2SFS $_{\text{CoOp}}$	88.06	88.68	**88.37**
IVLP	89.37	90.30	89.83
2SFS $_{\text{IVLP}}$	89.45	91.51	**90.47**

FGVC Aircraft

Method	Base	Novel	HM
CoOp [52]	40.44	22.30	28.75
2SFS $_{\text{CoOp}}$	44.60	29.91	**35.81**
IVLP	42.60	25.23	31.69
2SFS $_{\text{IVLP}}$	44.78	25.93	**32.85**

SUN 397

Method	Base	Novel	HM
CoOp [52]	80.60	65.89	72.51
2SFS $_{\text{CoOp}}$	79.16	70.32	**74.48**
IVLP	81.60	75.50	78.43
2SFS $_{\text{IVLP}}$	81.31	78.27	**79.76**

DTD

Method	Base	Novel	HM
CoOp [52]	79.44	41.18	54.24
2SFS $_{\text{CoOp}}$	81.40	49.11	**61.27**
IVLP	82.40	56.20	**66.82**
2SFS $_{\text{IVLP}}$	83.45	53.66	65.32

EuroSAT

Method	Base	Novel	HM
CoOp [52]	92.19	54.74	**68.69**
2SFS $_{\text{CoOp}}$	92.94	50.76	65.66
IVLP	96.73	67.83	79.74
2SFS $_{\text{IVLP}}$	95.01	74.51	**83.52**

UCF101

Method	Base	Novel	HM
CoOp [52]	84.69	56.05	67.46
2SFS $_{\text{CoOp}}$	85.04	64.67	**73.47**
IVLP	85.93	74.17	79.62
2SFS $_{\text{IVLP}}$	86.28	77.27	**81.53**

the paper, and make no hyperparameter changes. Results for these backbones are not available in the official article of [42], hence we used the open-source implementation of the authors with no modifications (the repository already integrates with different CLIP variants) as done for the all-to-all experiments of Sec. 5.2.

Results are given in Tab. 6 for ViT-B/32 and in Tab. 7 for ViT-L/14. For both backbones, 2SFS largely outperforms MMA on average, exhibiting larger improvements than those emerging with the ViT-B/16 visual encoder ($+1.70\%$ and $+1.83\%$, respectively), confirming that the effectiveness of 2SFS does not depend on a specific backbone.

C. Varying Shots

In Sec. 5 of the main body, results are given for the most popular FSA scenario in which k=16 shots are available per category. Here, we test the robustness of 2SFS in extreme data scarcity, working with both k=4 and k=8 shots for both all-to-all and base-to-novel cases. The results are discussed below.

Base-to-novel generalization. In line with Appendix B, we focus on the comparison with MMA [42]. We experiment with all backbones, and report results in Tab. 11 and Tab. 12 for ViT-B/16, Tab. 13 and Tab. 14 for ViT-B/32, and Tab. 15 and Tab. 16 for ViT-L/14, with 4 and 8 shots, respectively.[2]

On average, *2SFS outperforms MMA for all backbones and all shots setups*. Importantly, we observe that the performance gap increases as the shots decrease, up to large gaps such as $+3.98\%$ and $+4.45\%$ HM with ViT-B/16 and ViT-B/32 using 4 shots. We speculate this behavior stems from the reduced amount of learnable parameters of 2SFS, which better accommodates a smaller number of examples. To ground the discussion in some numbers: summing up LayerNorm instances totals around 61k parameters for ViT-B backbones, while MMA introduces 674k new parameters.

All-to-all adaptation. Results for the all-to-all setup are given in Tab. 9 and Tab. 10 for 4 and 8 shots. We include numbers from all 11 competitors of Sec. 5.2, following the reported results of [47], and reproducing when unavailable. Also in this case, *2SFS outperforms all competitors on average for all* $\{$*backbones, shots*$\}$ *combinations*.

In summary, looking at both scenarios, 2SFS appears to be a stronger approach w.r.t. to the comparison suite, regard-

[2]Please note that results for CoOp, CoCoOp, ProGrad, and KgCoOp with the ViT-B/16 backbone and $k\in\{4, 8\}$ are given in the supplementary material of [43], which we omit to avoid excessively dense tables. 2SFS largely outperforms all methods with available results.

Table 6. Experiments in *base-to-novel* generalization, with the ViT-B/32 visual backbone and $k = 16$ shots per base category, focusing on the comparison with MultiModal Adapter (MMA) [42]. "CLIP" refers to zero-shot performance with dataset-specific templates, *e.g.*, "*a photo of a* {}*, a type of flower*" for Oxford Flowers. Formatting follows Tab. 1.

Average across datasets.

Method	Base	Novel	HM
CLIP [33]	67.27	71.68	69.41
MMA [42]	78.69	71.04	74.67
2SFS	82.32	71.23	**76.37**

ImageNet

Method	Base	Novel	HM
CLIP [33]	67.49	64.06	65.73
MMA [42]	72.53	65.77	68.98
2SFS	72.52	66.62	**69.44**

Caltech101

Method	Base	Novel	HM
CLIP [33]	94.06	94.00	94.03
MMA [42]	97.20	92.63	94.86
2SFS	97.83	93.30	**95.51**

Oxford Flowers

Method	Base	Novel	HM
CLIP [33]	72.36	73.69	73.02
MMA [42]	95.50	71.57	**81.82**
2SFS	96.64	70.02	81.20

Oxford Pets

Method	Base	Novel	HM
CLIP [33]	90.64	96.87	93.65
MMA [42]	93.77	96.30	**95.02**
2SFS	93.18	95.56	94.35

Stanford Cars

Method	Base	Novel	HM
CLIP [33]	60.72	69.74	64.92
MMA [42]	73.73	69.27	71.43
2SFS	78.21	70.30	**74.04**

Food 101

Method	Base	Novel	HM
CLIP [33]	85.30	86.89	86.09
MMA [42]	85.77	87.13	**86.44**
2SFS	84.75	87.37	86.04

FGVC Aircraft

Method	Base	Novel	HM
CLIP [33]	21.25	29.27	24.62
MMA [42]	31.77	28.73	30.17
2SFS	39.12	30.85	**34.50**

SUN 397

Method	Base	Novel	HM
CLIP [33]	69.80	73.01	71.37
MMA [42]	80.27	76.57	78.38
2SFS	81.11	78.02	**79.53**

DTD

Method	Base	Novel	HM
CLIP [33]	54.17	58.21	56.12
MMA [42]	79.50	57.00	**66.40**
2SFS	80.09	54.63	64.95

EuroSAT

Method	Base	Novel	HM
CLIP [33]	55.14	69.77	61.60
MMA [42]	71.83	62.97	67.11
2SFS	96.80	62.68	**76.09**

UCF101

Method	Base	Novel	HM
CLIP [33]	69.08	72.96	70.97
MMA [42]	83.77	73.47	78.28
2SFS	85.25	74.15	**79.31**

Table 7. Experiments in *base-to-novel* generalization, with the ViT-L/14 visual backbone and $k = 16$ shots per base category, focusing on the comparison with MultiModal Adapter (MMA) [42]. "CLIP" refers to zero-shot performance with dataset-specific templates, *e.g.*, "*a photo of a* {}*, a type of flower*" for Oxford Flowers. Formatting follows Tab. 1.

Average across datasets.

Method	Base	Novel	HM
CLIP [33]	76.18	80.08	78.08
MMA [42]	85.70	79.06	82.25
2SFS	89.05	79.64	**84.08**

ImageNet

Method	Base	Novel	HM
CLIP [33]	79.18	74.04	76.53
MMA [42]	83.17	76.73	79.82
2SFS	83.11	76.98	**79.93**

Caltech101

Method	Base	Novel	HM
CLIP [33]	95.61	95.41	95.51
MMA [42]	98.60	95.97	97.27
2SFS	98.82	96.69	**97.74**

Oxford Flowers

Method	Base	Novel	HM
CLIP [33]	80.34	83.05	81.67
MMA [42]	99.00	80.20	88.61
2SFS	98.99	80.73	**88.93**

Oxford Pets

Method	Base	Novel	HM
CLIP [33]	93.78	96.53	95.14
MMA [42]	96.23	98.70	97.45
2SFS	96.74	98.64	**97.68**

Stanford Cars

Method	Base	Novel	HM
CLIP [33]	74.56	84.65	79.29
MMA [42]	85.27	83.80	84.53
2SFS	87.46	84.56	**85.99**

Food 101

Method	Base	Novel	HM
CLIP [33]	93.75	94.82	94.28
MMA [42]	94.23	95.10	**94.66**
2SFS	93.59	94.93	94.26

FGVC Aircraft

Method	Base	Novel	HM
CLIP [33]	37.52	44.21	40.59
MMA [42]	50.00	42.47	45.93
2SFS	59.76	43.59	**50.41**

SUN 397

Method	Base	Novel	HM
CLIP [33]	73.23	77.71	75.40
MMA [42]	85.03	81.77	83.37
2SFS	85.57	82.24	**83.87**

DTD

Method	Base	Novel	HM
CLIP [33]	59.14	67.87	63.21
MMA [42]	85.23	70.77	77.33
2SFS	87.35	70.73	**78.17**

EuroSAT

Method	Base	Novel	HM
CLIP [33]	70.93	82.90	76.45
MMA [42]	77.33	62.77	69.29
2SFS	98.41	64.69	**78.06**

UCF101

Method	Base	Novel	HM
CLIP [33]	79.94	79.66	79.80
MMA [42]	88.60	81.37	84.83
2SFS	89.80	82.24	**85.85**

M	Base	Novel	HM
$M = 100$	77.55	70.50	73.86
$M = 300$	77.71	70.99	74.20
$M = 500$	77.35	71.16	74.12

Table 8. Sweep on $M \in \{100, 300, 500\}$ when $\alpha = 0.6$ on the ImageNet validation set and CLIP ViT-B/16.

less of how many shots are available. Importantly, it does so by (i) not employing any external source of knowledge (such as LLMs to generate descriptions or Image Generators to craft new examples [4]), (ii) avoiding the usage of well-engineered templates for each dataset, which are likely to be unavailable in practice, and (iii) only leveraging a single template "*a photo of a* {}", in contrast to an ensemble of templates [17]. We speculate, however, that such orthogonal techniques may further improve 2SFS.

D. Total gradient steps allowed

This Appendix briefly analyzes the impact of increasing or reducing the total number of iterations m. For simplicity, we stick to the ViT-B/16 backbone and the ImageNet validation set. Recall that in Sec. 5, the total number of iterations is defined as $m = M \times k$, where, in our case, $M = 300$ and k is the number of shots. M was chosen so to match the number of gradient steps performed on ImageNet with ≈ 10 epochs (constant mini-batch size of 32, 16 shots for all categories). In essence, this means that the budget is expressed in terms of a constant number of gradient steps rather than epochs, following [47]. Here, we analyze the impact of varying M when α=0.6 as in Sec. 5. Results are given in Tab. 8. We observe that $M = 100$ likely allocates insufficient compute for learning a good feature extractor in the first stage (lowest "Novel" metric). In contrast, $M = 300$ and $M = 500$ exhibit more comparable behaviors, which leads to choosing $M = 300$ considering the reduced overall runtime.

E. Extended preliminary analysis

Here, we aim to enrich the preliminary analysis conducted in Sec. 3.2. Recall that Sec. 3.2 introduces the natural emergence of two distinct stages when training CLIP ViT-B/16 with three different PEFT techniques in the low-data regime of FSA, and does so by visualizing the learning dynamics on DTD [3] and FGVC Aircraft [29]. First, we show that such a dynamic is not limited to those datasets. Second, we identify a *saturating* behavior of Layer Normalization, which we link to the data-to-parameter ratio. Finally, we focus on Oxford Pets [32] and Food-101 [1], which were the only datasets (out of 11) leading to a slight performance degradation during the ablation study of Sec. 5.3.

Consistent behaviors. Fig. 6 shows that analogous and consistent patterns emerge also for UCF-101 [37] and EuroSAT [11] for all the PEFT techniques of our study. Particularly with EuroSAT, this behavior emerges to the extreme, with sharp breakpoints. In line with Sec. 3.2, BitFit tends to "break" earlier than both LoRA and LayerNorm.

Saturating behaviors. Fig. 7 shows consistent breakpoints for LoRA and BitFit further, displaying SUN397 [40] and ImageNet [34]. These two datasets have a trait in common w.r.t. the rest of the evaluation suite: a much larger label space. In FSA, where samples are constant per category, this inevitably entails a larger amount of examples. In parallel, LayerNorm instances total a reduced number of parameters w.r.t. to LoRA and BitFit (61k, 184k, 125k, respectively). We speculate that the more balanced data-to-parameter ratio of LayerNorm for these larger datasets has a regularizing effect, which avoids breaking and reaches a behavior similar to saturation, where the novel class accuracy remains constant.

Unexpected behaviors. Fig. 8 depicts the learning dynamics on Food-101 [1] and Oxford Pets [32]. These were the only two datasets where including a second stage did not appear beneficial in Fig. 5 of the main body. From the dynamics, the reason is evident: base and novel accuracy break together. For both datasets, *base* accuracy either decreases or saturates right after the breakpoint (pink line), implying overfitting since training data are available for base categories only. This suggests that α and M should be tailored to these datasets, to avoid training a classifier on overfitted features. However, we consider it fairer to transfer hyperparameters across datasets since, in practice, no annotated data except for the shots should be available in FSA, which raises concerns about the feasibility of tuning hyperparameters per dataset.

F. Limitations

In this work, we build on the finding that PEFT techniques learn good task-level features to design a simple and effective strategy for few-shot adaptation. For completeness, we identify and report three limitations of our work, which we hope can help construct future works.

Evaluating outside of our suite. While we successfully experiment with a variety of backbones (*i.e.*, ViT-B/16, ViT-B/32, ViT-L/14), datasets (*i.e.*, 11 different benchmarks), settings (*i.e.*, base-to-novel, all-to-all), PEFT techniques (*i.e.*, LayerNorm tuning and LoRA), and data availability conditions (*i.e.*, 4, 8, and 16 shots), as per most empirical observations, our results might not extend when tested with other (or future) PEFT strategies and on different benchmarks or additional models.

Expanding the variety of tasks. Our work focuses on downstream classification, following the established and recent field literature [2, 10, 16, 42, 43, 45, 47, 50–53]. How-

Table 9. *All-to-all* experiments with $\mathbf{k} = \mathbf{4}$ shots, using ViT-B/16, ViT-B/32, and ViT-L/14. Formatting follows Tab. 2.

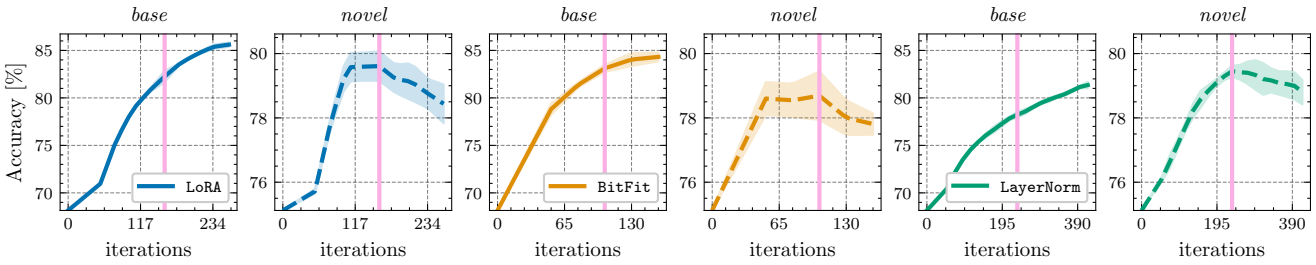
BACKBONE	METHOD	IMAGENET	SUN	AIR	ESAT	CARS	FOOD	PETS	FLWR	CAL	DTD	UCF	MEAN
ViT-B/16	*Zero-Shot* [33]	66.7	62.6	24.7	47.5	65.3	86.1	89.1	71.4	92.9	43.6	66.7	65.1
	CoOp [52] (ctx=16)	68.8	69.7	30.9	69.7	74.4	84.5	92.5	92.2	94.5	59.5	77.6	74.0
	CoCoOp [51]	70.6	70.4	30.6	61.7	69.5	86.3	92.7	81.5	94.8	55.7	75.3	71.7
	TIP-Adapter-F [50]	70.7	70.8	35.7	76.8	74.1	86.5	91.9	92.1	94.8	59.8	78.1	75.6
	CLIP-Adapter [10]	68.6	68.0	27.9	51.2	67.5	86.5	90.8	73.1	94.0	46.1	70.6	67.7
	PLOT++ [2]	70.4	71.7	35.3	83.2	76.3	86.5	92.6	92.9	95.1	62.4	79.8	76.9
	KgCoOp [43]	69.9	71.5	32.2	71.8	69.5	**86.9**	92.6	87.0	95.0	58.7	77.6	73.9
	TaskRes [45]	71.0	72.7	33.4	74.2	76.0	86.0	91.9	85.0	95.0	60.1	76.2	74.7
	MaPLe [16]	70.6	71.4	30.1	69.9	70.1	86.7	**93.3**	84.9	95.0	59.0	77.1	73.5
	ProGrad [53]	70.2	71.7	34.1	69.6	75.0	85.4	92.1	91.1	94.4	59.7	77.9	74.7
	LP++ [13]	70.8	73.2	34.0	73.6	74.0	85.9	90.9	93.0	95.1	62.4	79.2	75.6
	CLIP-LoRA [47]	**71.4**	72.8	37.9	84.9	77.4	82.7	91.0	93.7	95.2	63.8	81.1	77.4
	MMA [42]	70.5	72.9	35.0	42.4	73.3	86.0	92.9	91.3	94.5	60.1	79.0	72.5
	2SFS	71.1	**73.7**	**39.8**	**85.5**	**77.5**	85.9	92.6	**94.0**	**95.4**	**66.0**	**82.0**	**78.5**
ViT-B/32	*Zero-Shot* [33]	61.9	62.0	19.3	45.1	60.4	80.5	87.5	67.0	91.1	42.6	62.2	61.8
	CoOp [52] (ctx=16)	63.2	67.1	24.0	68.7	66.2	75.6	88.8	87.9	93.0	55.3	75.0	69.5
	CoCoOp [51]	65.2	67.8	17.3	58.5	62.0	81.1	89.8	74.6	93.2	52.3	71.6	66.7
	TIP-Adapter-F [50]	65.8	68.3	28.8	71.5	67.6	80.9	88.6	88.9	94.6	58.0	75.1	71.6
	CLIP-Adapter [10]	63.7	65.6	21.3	49.9	62.2	81.3	88.4	68.3	92.0	47.2	67.3	64.3
	PLOT++ [2]	64.6	69.2	26.2	81.6	68.5	77.8	89.1	90.2	93.9	57.2	75.6	72.2
	KgCoOp [43]	64.7	69.2	22.6	64.9	63.2	81.2	89.5	76.8	93.8	55.1	71.6	68.4
	TaskRes [45]	66.1	66.7	23.1	70.7	66.7	76.7	86.7	79.0	90.6	57.0	68.2	68.3
	MaPLe [16]	65.6	69.4	23.4	64.7	62.2	**81.4**	90.5	78.1	94.0	55.0	70.9	68.7
	ProGrad [53]	65.2	69.6	24.8	63.7	66.4	79.2	89.4	87.5	93.2	55.9	73.4	69.8
	LP++ [13]	66.1	70.5	26.0	73.5	67.3	80.0	88.9	90.2	94.0	59.3	74.8	71.9
	CLIP-LoRA [47]	**66.5**	70.3	27.7	**85.6**	68.3	75.6	86.3	90.1	94.3	60.3	76.5	72.9
	MMA [42]	64.7	70.4	25.6	36.0	66.3	80.5	**90.7**	86.1	94.0	55.6	74.6	67.7
	2SFS	66.0	**71.4**	**30.6**	82.6	**70.4**	80.2	89.4	**91.0**	**95.1**	**63.1**	**77.4**	**74.3**
ViT-L/14	*Zero-Shot* [33]	72.9	67.6	32.6	58.0	76.8	91.0	93.6	79.4	94.9	53.6	74.2	72.2
	CoOp [52] (ctx=16)	74.9	73.1	43.6	75.9	83.3	88.7	94.6	95.9	96.5	63.9	82.8	79.4
	CoCoOp [51]	77.0	74.7	41.0	74.7	79.7	91.3	94.9	89.8	97.1	64.9	82.6	78.9
	TIP-Adapter-F [50]	77.1	74.1	47.4	81.4	82.3	91.2	94.0	95.5	96.5	64.4	83.9	80.7
	CLIP-Adapter [10]	75.2	72.1	35.8	61.3	78.8	91.2	93.7	81.7	95.6	57.9	77.9	74.7
	PLOT++ [50]	76.4	75.2	43.2	81.3	82.6	87.7	94.2	95.9	96.9	66.8	83.8	80.4
	KgCoOp [43]	76.4	75.2	40.6	79.5	80.0	**91.5**	94.4	90.2	96.9	66.3	83.4	79.5
	TaskRes [45]	77.1	74.9	42.5	76.6	83.6	90.7	94.4	90.3	96.5	65.4	80.1	79.3
	MaPLe [16]	77.2	76.0	40.4	74.6	80.3	**91.5**	**95.0**	93.2	97.0	64.5	82.8	79.3
	ProGrad [53]	76.5	75.0	44.6	79.3	83.8	90.6	94.8	95.6	96.8	66.3	83.6	80.6
	LP++ [13]	77.4	76.9	45.9	83.1	82.7	91.0	93.8	97.2	**97.4**	68.3	85.3	81.7
	CLIP-LoRA [47]	**77.9**	76.7	48.9	86.4	**85.2**	89.6	93.9	97.4	97.2	70.4	86.4	82.7
	MMA [42]	77.7	77.1	45.2	55.3	83.3	91.4	94.3	95.1	97.0	63.8	83.2	78.5
	2SFS	77.3	**77.5**	**52.0**	**86.7**	84.9	90.9	**95.0**	**97.5**	**97.4**	**71.1**	**86.9**	**83.4**

ever, an additional intriguing direction to pursue is represented by tasks focusing on different challenges (*e.g.*, the spatial ones of semantic segmentation, and the temporal one of action recognition), which may require different adaptation strategies.

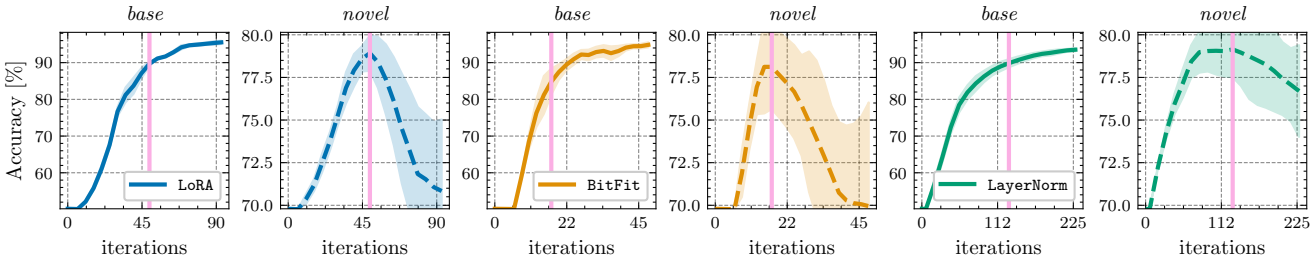
Validation-free stopping criterion. Finally, a core hyperparameter of our approach is α, regulating when to stop with the feature extractor training (*i.e.*, the first stage) and start with the second one (*i.e.*, classifier learning). As we have shown empirically with Oxford Pets [32] and Food-101 [1] a single α, tuned on a given dataset, may not be ideal for others. To this aim, future works may integrate (or investigate) stopping criteria not requiring a validation set [28], to dynamically understand or approximate, in an unsupervised manner, when to switch between the two stages.

Table 10. *All-to-all* experiments with $\mathbf{k} = \mathbf{8}$ shots, using ViT-B/16, ViT-B/32, and ViT-L/14. Formatting follows Tab. 2.

BACKBONE	METHOD	IMAGENET	SUN	AIR	ESAT	CARS	FOOD	PETS	FLWR	CAL	DTD	UCF	MEAN
ViT-B/16	*Zero-Shot* [33]	66.7	62.6	24.7	47.5	65.3	86.1	89.1	71.4	92.9	43.6	66.7	65.1
	CoOp [52] (ctx=16)	70.6	71.9	38.5	77.1	79.0	82.7	91.3	94.9	94.5	64.8	80.0	76.8
	CoCoOp [51]	70.8	71.5	32.4	69.1	70.4	87.0	**93.3**	86.3	94.9	60.1	75.9	73.8
	TIP-Adapter-F [50]	71.7	73.5	39.5	81.3	78.3	86.9	91.8	94.3	95.2	66.7	82.0	78.3
	CLIP-Adapter [10]	69.1	71.7	30.5	61.6	70.7	86.9	91.9	83.3	94.5	50.5	76.2	71.5
	PLOT++ [2]	71.3	73.9	41.4	88.4	81.3	86.6	93.0	95.4	95.5	66.5	82.8	79.6
	KgCoOp [43]	70.2	72.6	34.8	73.9	72.8	87.0	93.0	91.5	95.1	65.6	80.0	76.0
	TaskRes [45]	72.3	74.6	40.3	77.5	79.6	86.4	92.0	96.0	95.3	66.7	81.6	78.4
	MaPLe [16]	71.3	73.2	33.8	82.8	71.3	**87.2**	93.1	90.5	95.1	63.0	79.5	76.4
	ProGrad [53]	71.3	73.0	37.7	77.8	78.7	86.1	92.2	95.0	94.8	63.9	80.5	77.4
	LP++ [13]	72.1	75.1	39.0	78.2	76.4	86.8	91.8	95.2	95.5	67.7	81.9	78.2
	CLIP-LoRA [47]	72.3	74.7	**45.7**	**89.7**	**82.1**	83.1	91.7	**96.3**	95.6	67.5	84.1	80.3
	MMA [42]	71.9	74.7	38.9	69.7	76.8	86.4	92.9	94.6	95.6	66.9	82.9	77.4
	2SFS	**72.5**	**75.5**	44.3	89.1	81.9	86.1	92.9	95.9	**96.1**	**68.7**	**84.4**	**80.7**
ViT-B/32	*Zero-Shot* [33]	61.9	62.0	19.3	45.1	60.4	80.5	87.5	67.0	91.1	42.6	62.2	61.8
	CoOp [52] (ctx=16)	65.5	69.2	29.1	76.4	71.3	76.3	87.4	92.7	93.8	61.7	76.5	72.7
	CoCoOp [51]	65.8	68.9	20.3	58.1	63.4	81.6	90.1	77.3	93.8	57.4	72.4	68.1
	TIP-Adapter-F [50]	66.8	71.2	32.1	75.0	72.6	81.3	89.8	90.4	94.5	63.6	78.0	74.1
	CLIP-Adapter [10]	64.2	69.3	23.5	55.2	65.4	81.5	89.3	78.0	93.9	50.8	73.0	67.6
	PLOT++ [2]	66.2	71.0	31.7	87.1	73.5	78.2	88.4	**93.8**	94.4	62.9	79.1	75.1
	KgCoOp [43]	65.1	69.5	24.7	66.2	65.0	81.7	90.3	83.1	94.5	61.1	74.7	70.5
	TaskRes [45]	**67.4**	71.9	31.9	74.9	73.8	80.6	89.1	93.5	94.8	64.5	78.4	74.6
	MaPLe [16]	66.3	70.3	25.4	79.0	63.7	**81.9**	90.9	81.1	94.4	59.8	75.0	71.6
	ProGrad [53]	66.1	71.1	29.0	73.5	71.8	80.0	89.1	92.1	94.2	62.3	75.7	73.2
	LP++ [13]	67.1	72.2	30.3	78.8	71.2	81.5	89.3	92.4	94.6	64.2	78.4	74.5
	CLIP-LoRA [47]	67.2	72.1	**36.1**	**88.8**	74.4	76.7	87.7	92.4	94.8	63.7	80.1	75.8
	MMA [42]	66.7	72.2	29.6	56.2	70.4	81.0	**91.0**	90.7	94.6	64.4	78.7	72.3
	2SFS	67.2	**73.1**	35.2	88.7	**75.4**	80.4	90.4	93.4	**95.4**	**65.9**	**80.2**	**76.8**
ViT-L/14	*Zero-Shot* [33]	72.9	67.6	32.6	58.0	76.8	91.0	93.6	79.4	94.9	53.6	74.2	72.2
	CoOp [52] (ctx=16)	76.8	75.0	51.2	82.8	86.4	88.6	94.0	98.0	96.7	69.4	85.1	82.2
	CoCoOp [51]	77.4	75.6	43.3	77.0	81.4	**91.6**	**95.3**	93.0	97.0	67.9	84.5	80.4
	TIP-Adapter-F [50]	77.8	76.7	50.4	84.9	85.9	91.4	94.1	97.3	96.9	71.2	86.2	83.0
	CLIP-Adapter [10]	75.7	75.9	40.7	67.9	81.6	91.4	94.3	92.3	96.8	63.8	82.8	78.5
	PLOT++ [2]	77.8	77.0	43.2	87.0	84.6	89.6	93.3	96.3	96.8	69.5	84.8	81.8
	KgCoOp [43]	76.7	76.2	45.9	82.1	82.3	**91.6**	95.1	95.2	97.3	70.8	85.7	81.7
	TaskRes [45]	77.9	76.0	51.1	81.1	85.7	91.1	94.5	96.7	96.9	69.4	85.6	82.4
	MaPLe [16]	78.0	77.2	42.9	80.7	81.8	90.1	95.0	95.8	96.8	69.5	85.1	81.2
	ProGrad [53]	77.7	76.1	49.9	83.6	86.2	90.8	95.1	97.8	96.7	69.9	85.4	82.7
	LP++ [13]	78.4	78.4	50.8	85.0	85.2	91.4	94.4	97.9	**97.6**	72.1	86.0	83.4
	CLIP-LoRA [47]	78.5	78.0	57.5	**90.0**	**88.7**	89.7	94.2	98.0	97.0	72.2	88.3	84.7
	MMA [42]	**78.6**	78.8	50.9	61.4	85.8	91.5	95.1	97.7	97.1	71.9	86.2	81.4
	2SFS	**78.6**	**79.2**	**57.6**	89.8	88.2	91.4	95.2	**98.3**	97.2	**74.2**	**88.4**	**85.3**

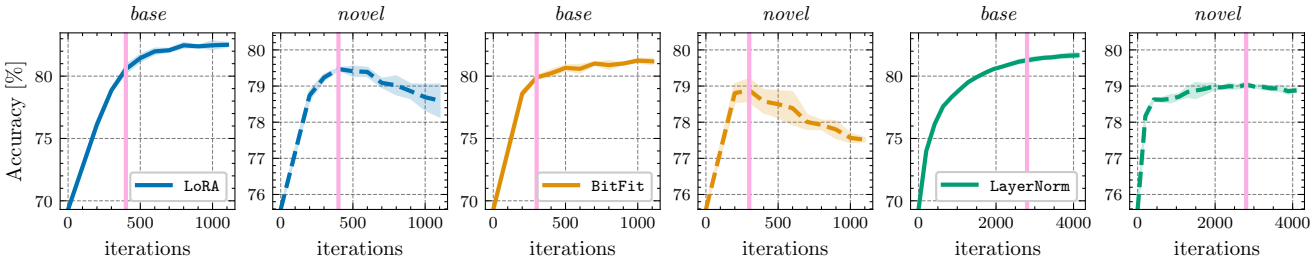


(a) UCF-101

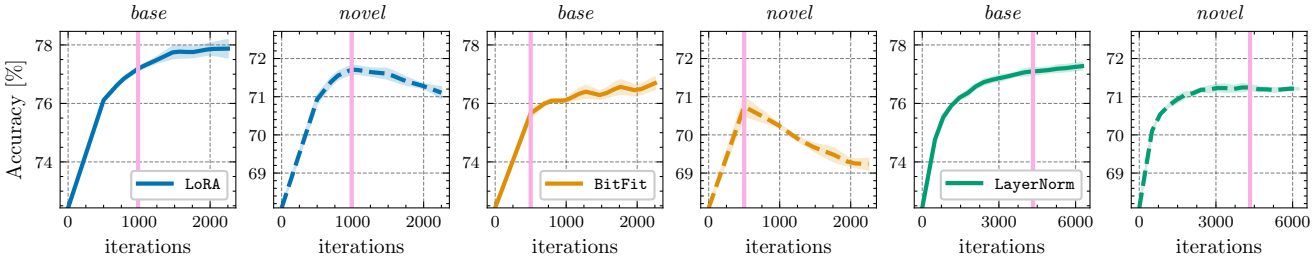


(b) EuroSAT

Figure 6. Breakpoints consistently emerging for UCF-101 [37] and EuroSAT [11], regardless of the PEFT technique used in our study. The pattern appears particularly evident with EuroSAT (bottom).

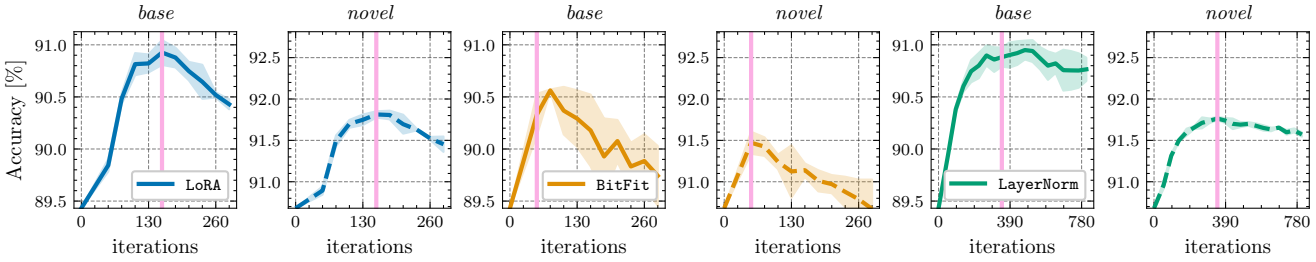


(a) SUN397

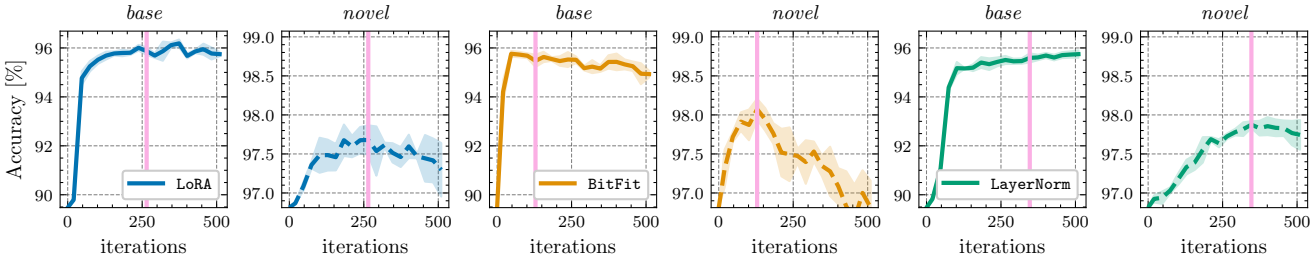


(b) ImageNet

Figure 7. Breakpoint further confirmed for both LoRA [12] and BitFit [46] on ImageNet [34] and SUN397 [40]. For Layer Normalization, we speculate that the more balanced data-to-parameter ratio, given the larger number of examples in these datasets and the smaller number of parameters of LayerNorm, has a regularizing effect, which avoids breaking and leads to saturation.



(a) Food-101



(b) Oxford Pets

Figure 8. Understanding the failure cases of Sec. 5.3 through the lens of breakpoints. On Oxford Pets [32] and Food-101 [1], base accuracy overfits or saturates right after degradation on novel accuracy, which leads the second stage of 2SFS to train a classifier on disrupted base features since α is fixed. These visualizations suggest that α and M should be tuned explicitly for these benchmarks, which we avoid to strive for an evaluation as realistic as possible.

Table 11. Experiments in *base-to-novel* generalization with the ViT-B/16 visual backbone $\mathbf{k}$=**4** shots per base class.

Average across datasets.

Method	Base	Novel	HM
CLIP [33]	69.34	74.22	71.70
MMA [42]	80.13	78.57	74.90
2SFS	84.64	78.53	**78.88**

ImageNet

Method	Base	Novel	HM
CLIP [33]	72.43	68.14	70.22
MMA [42]	75.37	70.10	72.64
2SFS	75.68	70.27	**72.87**

Caltech101

Method	Base	Novel	HM
CLIP [33]	96.84	94.00	93.73
MMA [42]	97.33	94.57	95.93
2SFS	98.13	94.21	**96.13**

Oxford Flowers

Method	Base	Novel	HM
CLIP [33]	72.08	77.80	74.83
MMA [42]	91.07	75.07	82.30
2SFS	94.94	76.26	**84.58**

Oxford Pets

Method	Base	Novel	HM
CLIP [33]	91.17	97.26	94.12
MMA [42]	91.30	97.07	94.10
2SFS	94.58	97.48	**96.01**

Stanford Cars

Method	Base	Novel	HM
CLIP [33]	63.37	74.89	68.65
MMA [42]	71.30	74.07	72.66
2SFS	74.45	75.98	**75.21**

Food101

Method	Base	Novel	HM
CLIP [33]	90.10	91.22	**90.66**
MMA [42]	89.77	91.10	90.43
2SFS	88.91	91.40	90.14

FGVC Aircraft

Method	Base	Novel	HM
CLIP [33]	27.19	36.29	31.09
MMA [42]	31.97	34.03	32.97
2SFS	39.36	35.67	**37.42**

SUN397

Method	Base	Novel	HM
CLIP [33]	69.36	75.35	72.23
MMA [42]	79.37	78.53	78.95
2SFS	80.23	78.46	**79.33**

DTD

Method	Base	Novel	HM
CLIP [33]	53.24	59.90	56.37
MMA [42]	63.50	63.57	63.53
2SFS	76.81	65.02	**70.42**

EuroSAT

Method	Base	Novel	HM
CLIP [33]	56.48	64.05	60.03
MMA [42]	50.13	69.93	58.40
2SFS	91.29	75.10	**82.41**

UCF101

Method	Base	Novel	HM
CLIP [33]	70.53	77.50	73.85
MMA [42]	80.13	78.57	79.34
2SFS	84.64	78.53	**81.47**

Table 12. Experiments in *base-to-novel* generalization with the ViT-B/16 visual backbone **k=8** shots per base class.

Average across datasets.

Method	Base	Novel	HM
CLIP [33]	69.34	74.22	71.70
MMA [42]	84.80	78.10	76.68
2SFS	86.37	78.58	**79.74**

ImageNet

Method	Base	Novel	HM
CLIP [33]	72.43	68.14	70.22
MMA [42]	76.43	70.07	73.11
2SFS	76.97	70.67	**73.69**

Caltech101

Method	Base	Novel	HM
CLIP [33]	96.84	94.00	93.73
MMA [42]	97.77	93.87	95.78
2SFS	98.13	94.18	**96.11**

Oxford Flowers

Method	Base	Novel	HM
CLIP [33]	72.08	77.80	74.83
MMA [42]	95.37	75.43	84.24
2SFS	96.68	76.36	**85.33**

Oxford Pets

Method	Base	Novel	HM
CLIP [33]	91.17	97.26	94.12
MMA [42]	94.90	97.50	96.18
2SFS	94.95	97.80	**96.35**

Stanford Cars

Method	Base	Novel	HM
CLIP [33]	63.37	74.89	68.65
MMA [42]	75.00	74.80	74.90
2SFS	78.99	75.45	**77.18**

Food101

Method	Base	Novel	HM
CLIP [33]	90.10	91.22	**90.66**
MMA [42]	89.53	91.07	90.29
2SFS	89.02	91.36	90.17

FGVC Aircraft

Method	Base	Novel	HM
CLIP [33]	27.19	36.29	31.09
MMA [42]	37.53	34.57	35.99
2SFS	42.96	36.39	**39.40**

SUN397

Method	Base	Novel	HM
CLIP [33]	69.36	75.35	72.23
MMA [42]	80.73	78.33	79.51
2SFS	81.25	78.76	**79.99**

DTD

Method	Base	Novel	HM
CLIP [33]	53.24	59.90	56.37
MMA [42]	77.07	64.47	70.21
2SFS	80.17	64.45	**71.46**

EuroSAT

Method	Base	Novel	HM
CLIP [33]	56.48	64.05	60.03
MMA [42]	50.30	69.97	58.53
2SFS	93.33	75.09	**83.23**

UCF101

Method	Base	Novel	HM
CLIP [33]	70.53	77.50	73.85
MMA [42]	84.80	78.10	81.31
2SFS	86.37	78.58	**82.29**

Table 13. Experiments in *base-to-novel* generalization with the ViT-B/32 visual backbone **k=4** shots per base class.

Average across datasets.

Method	Base	Novel	HM
CLIP [33]	67.27	71.68	69.41
MMA [42]	77.20	74.43	70.55
2SFS	82.04	74.44	**75.00**

ImageNet

Method	Base	Novel	HM
CLIP [33]	67.49	64.06	65.73
MMA [42]	69.77	65.63	67.64
2SFS	70.51	65.91	**68.13**

Caltech101

Method	Base	Novel	HM
CLIP [33]	94.06	94.00	94.03
MMA [42]	96.90	93.03	94.93
2SFS	97.16	93.52	**95.31**

Oxford Flowers

Method	Base	Novel	HM
CLIP [33]	72.36	73.69	73.02
MMA [42]	87.03	70.83	78.10
2SFS	93.99	71.70	**81.35**

Oxford Pets

Method	Base	Novel	HM
CLIP [33]	90.64	96.87	93.65
MMA [42]	88.43	96.33	92.21
2SFS	92.50	95.25	**93.86**

Stanford Cars

Method	Base	Novel	HM
CLIP [33]	60.72	69.74	64.92
MMA [42]	66.23	69.17	67.67
2SFS	69.22	70.84	**70.02**

Food101

Method	Base	Novel	HM
CLIP [33]	85.30	86.89	**86.09**
MMA [42]	85.03	86.40	85.71
2SFS	84.04	86.95	85.47

FGVC Aircraft

Method	Base	Novel	HM
CLIP [33]	21.25	29.27	24.62
MMA [42]	25.27	27.90	26.52
2SFS	33.11	30.23	**31.61**

SUN397

Method	Base	Novel	HM
CLIP [33]	69.80	73.01	71.37
MMA [42]	77.37	76.40	76.88
2SFS	78.23	76.34	**77.27**

DTD

Method	Base	Novel	HM
CLIP [33]	54.17	58.21	56.12
MMA [42]	62.13	57.70	59.83
2SFS	74.23	56.76	**64.33**

EuroSAT

Method	Base	Novel	HM
CLIP [33]	55.14	69.77	61.60
MMA [42]	41.73	57.17	48.24
2SFS	87.91	68.35	**76.91**

UCF101

Method	Base	Novel	HM
CLIP [33]	69.08	72.96	70.97
MMA [42]	77.20	74.43	75.79
2SFS	82.04	74.44	**78.05**

Table 14. Experiments in *base-to-novel* generalization with the ViT-B/32 visual backbone **k=8** shots per base class.

Average across datasets.

Method	Base	Novel	HM
CLIP [33]	67.27	71.68	69.41
MMA [42]	81.67	73.83	72.06
2SFS	84.21	74.62	**75.88**

ImageNet

Method	Base	Novel	HM
CLIP [33]	67.49	64.06	65.73
MMA [42]	71.03	65.17	67.97
2SFS	71.39	66.24	**68.72**

Caltech101

Method	Base	Novel	HM
CLIP [33]	94.06	94.00	94.03
MMA [42]	97.13	92.57	94.80
2SFS	97.61	93.56	**95.54**

Oxford Flowers

Method	Base	Novel	HM
CLIP [33]	72.36	73.69	73.02
MMA [42]	92.97	71.63	80.92
2SFS	95.79	71.35	**81.78**

Oxford Pets

Method	Base	Novel	HM
CLIP [33]	90.64	96.87	93.65
MMA [42]	93.57	95.57	**94.56**
2SFS	92.79	95.58	94.16

Stanford Cars

Method	Base	Novel	HM
CLIP [33]	60.72	69.74	64.92
MMA [42]	69.83	69.80	69.81
2SFS	73.48	70.71	**72.07**

Food101

Method	Base	Novel	HM
CLIP [33]	85.30	86.89	**86.09**
MMA [42]	85.03	86.53	85.77
2SFS	84.15	87.33	85.71

FGVC Aircraft

Method	Base	Novel	HM
CLIP [33]	21.25	29.27	24.62
MMA [42]	28.63	27.67	28.14
2SFS	35.47	30.35	**32.71**

SUN397

Method	Base	Novel	HM
CLIP [33]	69.80	73.01	71.37
MMA [42]	78.83	76.10	77.44
2SFS	79.49	77.13	**78.29**

DTD

Method	Base	Novel	HM
CLIP [33]	54.17	58.21	56.12
MMA [42]	73.70	56.07	63.69
2SFS	75.85	55.23	**63.92**

EuroSAT

Method	Base	Novel	HM
CLIP [33]	55.14	69.77	61.60
MMA [42]	41.80	57.20	48.30
2SFS	94.18	68.24	**79.14**

UCF101

Method	Base	Novel	HM
CLIP [33]	69.08	72.96	70.97
MMA [42]	81.67	73.83	77.55
2SFS	84.21	74.62	**79.12**

Table 15. Experiments in *base-to-novel* generalization with the ViT-L/14 visual backbone **k=4** shots per base class.

Average across datasets.

Method	Base	Novel	HM
CLIP [33]	76.18	80.08	78.08
MMA [42]	82.70	81.60	80.25
2SFS	88.11	82.15	**82.82**

ImageNet

Method	Base	Novel	HM
CLIP [33]	79.18	74.04	76.53
MMA [42]	82.00	76.67	**79.25**
2SFS	81.35	75.90	78.53

Caltech101

Method	Base	Novel	HM
CLIP [33]	95.61	95.41	95.51
MMA [42]	97.30	97.30	97.30
2SFS	98.36	97.09	**97.72**

Oxford Flowers

Method	Base	Novel	HM
CLIP [33]	80.34	83.05	81.67
MMA [42]	92.93	81.87	87.05
2SFS	97.94	81.77	**89.13**

Oxford Pets

Method	Base	Novel	HM
CLIP [33]	93.78	96.53	95.14
MMA [42]	94.93	98.47	96.67
2SFS	96.46	98.56	**97.50**

Stanford Cars

Method	Base	Novel	HM
CLIP [33]	74.56	84.65	79.29
MMA [42]	79.83	85.03	82.35
2SFS	82.50	85.11	**83.79**

Food101

Method	Base	Novel	HM
CLIP [33]	93.75	94.82	**94.28**
MMA [42]	93.70	94.57	94.13
2SFS	93.11	94.76	93.93

FGVC Aircraft

Method	Base	Novel	HM
CLIP [33]	37.52	44.21	40.59
MMA [42]	42.57	42.40	42.48
2SFS	51.58	44.57	**47.82**

SUN397

Method	Base	Novel	HM
CLIP [33]	73.23	77.71	75.40
MMA [42]	82.17	81.80	81.98
2SFS	82.91	81.20	**82.05**

DTD

Method	Base	Novel	HM
CLIP [33]	59.14	67.87	63.21
MMA [42]	65.90	67.00	66.45
2SFS	80.86	70.29	**75.21**

EuroSAT

Method	Base	Novel	HM
CLIP [33]	70.93	82.90	76.45
MMA [42]	72.50	72.20	72.35
2SFS	92.92	67.15	**77.96**

UCF101

Method	Base	Novel	HM
CLIP [33]	79.94	79.66	79.80
MMA [42]	82.70	81.60	82.15
2SFS	88.11	82.15	**85.03**

Table 16. Experiments in *base-to-novel* generalization with the ViT-L/14 visual backbone **k=8** shots per base class.

Average across datasets.

Method	Base	Novel	HM
CLIP [33]	76.18	80.08	78.08
MMA [42]	86.30	80.73	81.54
2SFS	88.28	82.24	**83.66**

ImageNet

Method	Base	Novel	HM
CLIP [33]	79.18	74.04	76.53
MMA [42]	82.63	76.80	**79.61**
2SFS	82.43	76.46	79.34

Caltech101

Method	Base	Novel	HM
CLIP [33]	95.61	95.41	95.51
MMA [42]	98.30	96.60	97.44
2SFS	98.52	96.62	**97.56**

Oxford Flowers

Method	Base	Novel	HM
CLIP [33]	80.34	83.05	81.67
MMA [42]	97.97	80.30	88.26
2SFS	98.67	81.21	**89.09**

Oxford Pets

Method	Base	Novel	HM
CLIP [33]	93.78	96.53	95.14
MMA [42]	95.77	98.33	97.03
2SFS	96.15	98.47	**97.30**

Stanford Cars

Method	Base	Novel	HM
CLIP [33]	74.56	84.65	79.29
MMA [42]	82.63	84.20	83.41
2SFS	85.51	84.97	**85.24**

Food101

Method	Base	Novel	HM
CLIP [33]	93.75	94.82	94.28
MMA [42]	93.87	94.87	**94.37**
2SFS	93.81	94.76	94.28

FGVC Aircraft

Method	Base	Novel	HM
CLIP [33]	37.52	44.21	40.59
MMA [42]	46.50	41.23	43.71
2SFS	55.00	44.49	**49.19**

SUN397

Method	Base	Novel	HM
CLIP [33]	73.23	77.71	75.40
MMA [42]	83.67	81.40	82.52
2SFS	84.25	81.84	**83.03**

DTD

Method	Base	Novel	HM
CLIP [33]	59.14	67.87	63.21
MMA [42]	78.13	69.90	73.79
2SFS	83.91	70.01	**76.33**

EuroSAT

Method	Base	Novel	HM
CLIP [33]	70.93	82.90	76.45
MMA [42]	72.73	72.00	72.36
2SFS	94.50	71.68	**81.53**

UCF101

Method	Base	Novel	HM
CLIP [33]	79.94	79.66	79.80
MMA [42]	86.30	80.73	83.42
2SFS	88.28	82.24	**85.15**