Hold-One-Shot-Out (HOSO) for Validation-Free Few-Shot CLIP Adapters

Supplementary Material

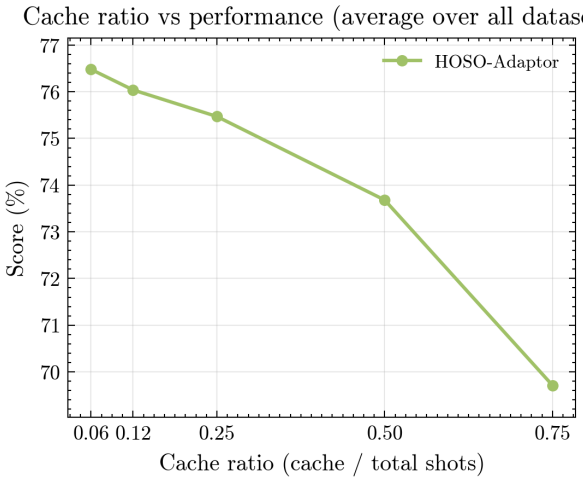

Figure S1. **A single hold-out shot is optimal.** An ablation study of the hold-out cache size in the 16-shot setting using ViT-B/16. The average accuracy across ten datasets is plotted against the cache size, expressed as a ratio of the total shots. Performance peaks at a cache size of one (ratio ≈ 0.06), validating the core design choice. Larger cache sizes lead to performance declines, indicating a trade-off between allocating samples to learn the blending ratio and retaining samples for adapter training.

A. Related Work Extended

CLIP few-shot adaptation methods traditionally fall into two categories: prompt learning and adapter-based approaches. Prompt learning optimises global text or visual prompts for downstream tasks, yielding substantial improvements over zero-shot baselines. Prompt learning requires backpropagation through the entire text encoder and access to its weights [32]. Adapter-based techniques operate in feature space and avoid requiring access to pre-trained model weights because they do not rely on backpropagation. Within adapter-based approaches, training-based variants either train a linear classifier with CLIP features as input or a shallow Multi-Layer Perceptron with a blending ratio. Later works introduce training-free methods that populate a key-value cache using the few-shot examples [30].

Recently, validation-free few-shot has emerged as a distinct setting in which methods must follow a strict few-shot protocol: hyperparameters may not be set using a validation set and must be held constant across datasets [24]. We adopt this setting because it mirrors real-world scenarios in which the goal is to adapt a large Vision-Language Model (VLM) to a specific domain of interest with limited labelled data. Previous works have used various terms to refer to the blending hyperparameter, including blending ratio, mixing coefficient, weighting parameter, or blending ratio. For clarity and consistency, this work uses the term "blending ratio" to refer to the scalar weight that linearly combines the CLIP- and adapted features.

CLIP-Adapter appends lightweight, learnable bottleneck linear layers to the image branch of CLIP, while keeping the main backbone frozen during few-shot fine-tuning. Adding extra layers can cause overfitting in few-shot settings. To enhance robustness and prevent overfitting, CLIP-Adapter uses a residual connection that integrates new, fine-tuned knowledge with pre-trained representations from the original CLIP backbone. Within this design, the blending ratio α serves as a critical hyperparameter; CLIP-Adapter performs a comprehensive search over α for each dataset to select the best-performing configuration. By ablation, Gao *et al.* [10] find the optimal α is 0.6 for the fine-grained Describable Textures Dataset (DTD) and 0.2 for the generic ImageNet. They observe that fine-grained benchmarks favour greater integration of new knowledge. In contrast, generic datasets rely more on existing priors. Setting α to 0 recovers zero-shot CLIP with no new knowledge, while $\alpha = 1$ makes classification depend entirely on the adapted features [10]. The proposed HOSO-Adapter keeps the architecture and hyperparameters identical to those of CLIP-Adapter. The only difference is that HOSO-Adapter learns the blending ratio, enabling CLIP-Adapter-style methods to compete under the validation-free setting. TipAdapter's prediction function comprises two terms: one adapts and summarises information from the few-shot training set, and the other retains prior knowledge from the original CLIP classifier. The balance between these terms is controlled by a blending ratio α. Zhang *et al.* find that α should be set higher when there is a substantial domain gap between the pre-trained and few-shot tasks, since more information from the few-shot set is required in such cases; otherwise, a lower α suffices [30]. Proto-Adapter [13] adapts CLIP via a single-layer class-prototype adapter and linearly blends adapter logits with CLIP zero-shot logits using a blending ratio α. The paper reports per-dataset values of α

without detailing the selection procedure. It is observed that the hardcoded α values provided in the codebase vary greatly between datasets (*e.g.* Caltech-101 0.8; FGVC-Aircraft 0.2; Oxford 102 Flowers 0.6; and EuroSAT 1.0). Retentive CLIP Adapter Tuning (RCAT) combines CLIP with the temporal processing of a Retentive Network for few-shot video recognition. RCAT employs a specialised adapter-tuning mechanism that modifies the original CLIP architecture to align with the temporal and spatial characteristics of video sequences, thereby improving predictive performance and interpretability. RCAT systematically evaluates the effect of the blending ratio α on bimodal feature alignment by performing a grid search over α from 0.1 to 0.9 in increments of 0.1 [29]. The works listed are a subset of all CLIP adapters that use a grid-searched blending ratio per dataset and thus fall outside the strict few-shot setting. CLIP-Adapter and Proto-Adapter report dataset-dependent optimal values of α: fine-grained benchmarks favour larger α while generic datasets prefer smaller α; TipAdapter prescribes larger α when the domain gap is substantial.

To the best of our knowledge, only two works have proposed methods to select the blending ratio of CLIP adapters in a validation-free manner, namely SVL-Adapter [21] and PathCLIP [17]. The SVL-Adapter combines the complementary strengths of vision-language pretraining and self-supervised representation learning and introduces a fully automatic method for selecting the blending hyperparameter α without requiring held-out, labelled validation data. The method computes α from CLIP's average prediction confidence on the N test images of a dataset: $\alpha = \frac{1}{N}\sum_{i=1}^{N} \max_k P(y_i = k \mid x_i)$, which directly reflects CLIP's confidence in its zero-shot predictions. This mechanism assumes that when CLIP is not confident, the influence of low-shot learning should increase, automatically adjusting its contribution relative to zero-shot CLIP. The adaptive blending, therefore, leverages the strengths of both zero-shot and low-shot learning to support improved task adaptation in the absence of validation labels [21].

PathCLIP [17] adapts CLIP for pathology and targets learning the blending ratio for its domain of interest. PathCLIP uses Residual Feature Refinement (RFR) as a lightweight adaptation mechanism that tailors CLIP representations to pathology images. As a shallow residual module, RFR is designed to capture salient spatial cues and morphological patterns characteristic of pathological specimens. The self-adaptive blending ratio dynamically balances CLIP knowledge with task-specific features, introducing only a small number of additional trainable parameters. The self-adaptive fusion component, Dual-view Vision Contrastive (DVC), draws inspiration from self-supervised learning approaches such as SimCLR [4]. DVC employs two types of augmentations: *weak*, denoting standard flip-and-shift transformations, and *strong*, which consist of more substantial appearance perturbations using methods such as RandAugment [6] and CTAugment [2]. DVC measures the distance between representations obtained from weak and strong augmentations, using cosine similarity. Specifically, the DVC metric calculates the average cosine similarity between encoder outputs for each augmentation across the support set, providing an indicator of the model's consistency under different perturbations. In addition to DVC, support set accuracy serves as a complementary metric for assessing model learning. Since accuracy alone may be insufficient due to the risk of overfitting, particularly when adapting large models to tasks with limited data, PathCLIP combines DVC and accuracy via a bound-constrained mechanism to determine a dynamic blending ratio, α. In this work, we use CLIP's average zero-shot prediction confidence (from SVL-Adapter) and PathCLIP's DVC as baselines. Our method differs in that it sets α as a learnable logit and optimises it independently of the adapter weights via a hold-one-shot-out cache.

Following [24], for the validation-free version of CLIP-Adapter, we set the hyperparameter α to 0.2 for all datasets, as it is the best value reported on ImageNet in the original paper. Similarly, TIP-Adapter sets β and α to 1, as recommended in the official repository (https://github.com/gaopengcuhk/Tip-Adapter/issues/13).

B. PathCLIP Reimplementation

The PathCLIP [17] reimplementation employs an online adaptive strategy that adjusts the blending weight α at every training step. Two signals from the current mini-batch guide the update: (1) *Dual-view Vision Contrastive*, which measures the cosine similarity between the CLIP features of weakly and strongly augmented views of the same images, and (2) the mini-batch classification accuracy. A target α is computed as a weighted combination of the DVC score and the inverse of the accuracy. The model's internal α is then updated towards this target value using an exponential moving average, allowing it to dynamically balance the contributions of the CLIP encoder and the adapter throughout training.

PathCLIP is reimplemented as a CLIP-Adapter variant that learns an adapter and blends it with frozen CLIP image features using an adaptive weight α. The implementation loads an official CLIP backbone and freezes both the visual and text encoders. The only trainable component is a lightweight MLP adapter (two

Table S1. **Main results with RN50 backbone, average over three runs.** Performance of HOSO-Adapter across different few-shot settings.

Method	Caltech101	DTD	EuroSAT	FGVC	Food101	ImageNet	Flowers	Pets	Cars	SUN397	UCF101	Average
HOSO-Adapter (2-shot)	89.10	47.83	53.27	19.13	79.63	59.63	73.90	84.10	58.00	62.10	66.60	63.03
HOSO-Adapter (4-shot)	90.17	56.40	35.00	23.47	79.80	60.70	85.50	87.73	62.43	65.03	70.20	65.13
HOSO-Adapter (8-shot)	91.40	61.47	77.83	27.80	80.63	61.75	91.90	87.73	67.13	67.67	74.27	71.78
HOSO-Adapter (16-shot)	93.03	66.77	83.27	34.60	80.93	62.93	95.07	89.47	73.80	69.83	78.03	75.25

Table S2. **Ablation on design choices for HOSO-Adapter (RN50, 16-shot).** The baseline updates the ratio every epoch using a cache of size one per class, removing cached samples from the training set.

Method	Caltech101	DTD	EuroSAT	FGVC	Food101	Flowers	Pets	Cars	SUN397	UCF101	Average
Baseline (remove cached samples)	93.0	66.6	83.0	34.8	81.1	95.1	89.5	73.7	69.6	77.9	**76.43**
Keep cache in training set	91.2	63.9	84.2	34.4	73.5	94.2	79.5	70.8	65.5	76.3	73.35
Cache size of two per class	92.6	66.5	82.1	33.9	80.8	94.6	89.0	73.6	69.5	77.8	76.04
Cache size of eight per class	91.7	62.6	80.9	29.3	80.7	92.5	88.0	69.1	67.3	74.7	73.68

linear layers with a bottleneck and ReLU), sized to the CLIP embedding dimensionality (1024 for RN50, 512 for ViT variants). Given an input image, the pipeline computes base CLIP features, passes them through the adapter, and blends in feature space as $\text{image}_{\text{fused}} = \alpha \cdot \text{adapter}_{\text{out}} + (1-\alpha) \cdot \text{base}_{\text{features}}$. It then normalises image and text features and computes scaled logits using the CLIP logit scale. Similar to CLIP-Adapter, it uses cross-entropy loss and uses dataset-specific single templates as text prompts. The online update of α works as follow: at each training step the procedure computes on GPU, (i) a data variation consistency score as the mean cosine similarity between normalised CLIP features of weak and strong views (weak: horizontal flip plus small translation; strong: RandAugment plus a minimal CTAugment [2] module), and (ii) the current accuracy. A target $\alpha = w_{\text{dvc}} \cdot \text{dvc} + w_{\text{acc}} \cdot (1 - \text{acc})$ is clamped to $[\alpha_{\min}, \alpha_{\max}]$ and applied with exponential moving average smoothing ($\alpha \leftarrow (1 - \text{smooth}) \cdot \alpha + \text{smooth} \cdot \text{target}$). All augmentation and feature extraction for DVC run on GPU.

C. SVL-Adapter Reimplementation

SVL-Adapter determines the blending weight α data-dependently before training begins. The method first calculates the average zero-shot confidence of the pretrained CLIP model over all images in the few-shot training set. This average confidence, denoted as λ, serves as a proxy for zero-shot performance on the target task. The blending weight is set to $\alpha = 1 - \lambda$, thereby giving the adapter more weight when CLIP yields lower zero-shot confidence. The computed α remains fixed for the entire duration of adapter training and subsequent evaluation. We follow the same reimplementation approach as in PathCLIP: freeze CLIP, train only the MLP adapter, and blend the adapter and base features with α.

D. Per-Dataset Ablation Study

Per-dataset ablations for cache policy and size are reported in Table S2. Updating the ratio at each epoch with a per-class cache of 1 and removing cached samples from training yields the best average (76.43), with gains across most datasets. Retaining cached items in training reduces the average by 3.08 to 73.35, with marked drops on Food101 and Pets; EuroSAT slightly improves but cannot offset broader declines. A cache of two per class is near-neutral (76.04), whilst a cache of eight hurts performance (73.68). A minimal cache with cached samples removed from the training set is the optimal configuration.

E. Blending Ratio Over Training

HOSO-Adapter treats the blending ratio as a dynamic regulariser to curb overfitting. This analysis tracks α under two regimes: the decoupled HOSO approach and a naive joint-training baseline. In the naive case, co-optimising α and the adapter on the same few-shot data steadily increases α, overweighting the expressive adapter and leading to overfitting. With HOSO, α is tuned on a hold-out set. It thus reflects generalisation: when the adapter starts to overfit, the features yield lower scores on the hold-out cache, the optimiser lowers α and assigns more weight to the robust CLIP prior. This pattern holds across all datasets except EuroSAT. We posit that this discrepancy arises from the substantial domain gap between CLIP's pretraining data (natural images) and EuroSAT's satellite images.

F. Overfitting Analysis

Figure S5 compares the extent of overfitting for the proposed HOSO-Adaptor with a baseline that directly learns the blending ratio on the full training set. The vertical axis shows the training-test accuracy gap, a

Table S3. **Performance of HOSO-Adapter with data augmentation.** Results use an RN50 backbone.

Shots	Caltech101	DTD	EuroSAT	FGVCAircraft	Food101	Flowers102	OxfordPets	StanfordCars	SUN397	UCF101	Average
2 shots	88.5	47.7	50.7	19.1	78.8	73.6	84.7	57.5	62.0	65.7	62.83
4 shots	89.4	56.4	32.6	22.9	79.4	85.1	87.1	61.7	65.2	69.5	64.93
8 shots	91.2	62.1	77.6	28.2	80.2	91.4	86.9	66.7	67.5	74.0	72.58
16 shots	**92.5**	**66.3**	**81.6**	**34.9**	**80.7**	**94.6**	**89.2**	**73.3**	**69.4**	**78.0**	**76.05**

Table S4. **ViT-B/16 16-shot performance comparison.** CLIP-Adapter results are shown with a fixed blending ratio ($\alpha = 0.2$) and with the best-performing ratio selected per dataset. HOSO-Adapter results are averaged over three runs.

Method	Caltech101	DTD	EuroSAT	FGVCAircraft	Food101	ImageNet	Flowers102	OxfordPets	StanfordCars	SUN397	UCF101	Average
CLIP-Adapter ($\alpha = 0.2$)	94.90	59.70	70.50	34.10	89.10	**71.50**	93.10	92.60	73.90	74.20	80.40	75.82
CLIP-Adapter (best α)†	**95.90**	**71.70**	**85.80**	**45.80**	**89.30**	**71.50**	**97.40**	**92.70**	**82.10**	**75.60**	**84.00**	**81.07**
HOSO-Adapter (ours)	95.40	70.67	85.30	43.23	88.97	70.93	97.23	92.27	81.50	74.67	83.43	80.33

† Best blending ratio from the set {0.2, 0.4, 0.6, 0.7, 0.8} selected independently per dataset using the test set.

direct measure of overfitting, over the full training run. Across the nine benchmark datasets, HOSO-Adaptor (green) consistently maintains a smaller generalisation gap than the baseline (red). The baseline exhibits severe overfitting across several datasets, including Describable Textures, Food101, SUN397, StanfordCars and FGVC Aircraft, where the train-test accuracy gap exceeds 40 percentage points. In contrast, HOSO-Adaptor effectively regularises the model, keeping the gap significantly lower throughout training. This demonstrates the critical role of the decoupled optimisation strategy in reducing overfitting to the few-shot training data and improving generalisation.

G. Data Description

In Table S5 we list the 11 datasets and templates used to embed the class labels. These datasets span both fine-grained and coarse-grained domains. We use the same datasets and templates as in CLIP-Adapter.

H. Blending Ratio Grid Search

Increasing the blending weight on the adapter with a ViT-B/16 backbone in the 16-shot setting has variable effects. Average accuracy rises from 74.82 at $\alpha = 0.2$ to a peak of 80.62 at $\alpha = 0.6$, then slightly decreases at $\alpha = 0.8$. Fine-grained and distribution-shifted datasets (DTD, EuroSAT, FGVC Aircraft, Stanford Cars) benefit most from higher adapter contribution, while tasks with strong CLIP priors or broader scene cues (OxfordPets, SUN397, UCF101) favour lower to mid weights (around $\alpha = 0.4$). Caltech101 and Flowers102 remain stable across different values of α in the 0.4 - 0.8 range.

Following the ViT-B/16 sweep, we repeat the blending-weight grid search for the RN50 backbone under the same 16-shot protocol. We interpolate between the frozen CLIP features and the adapter outputs using $\alpha \in \{0.2, 0.4, 0.6, 0.8\}$, where a higher α increases the adapter contribution. Compared to ViT-B/16, RN50's weaker visual prior makes the optimal α more dataset-dependent: larger blending ratios often help on distribution-shifted or fine-grained tasks (e.g., EuroSAT, FGVC Aircraft, Stanford Cars), while smaller ratios suffice when CLIP priors are strong (e.g., Food101, Pets). Table S6 reports the resulting per-dataset accuracies.

I. Design Choices Ablation

Table S8 shows that learning the blending weight yields the highest average accuracy, slightly improves on a fixed blending weight, and clearly outperforms random sampling. HOSO-Adapter achieves the best mean accuracy, while CLIP-Adapter with a fixed blending weight leads on some datasets. The fixed-value CLIP-Adapter (0.6) is shown as an upper bound for fixed-ratio performance, as it is selected based on the best test-set performance across all datasets via grid search. This ablation demonstrates that learning the blending ratio as the default design choice yields clear improvement over random sampling and even outperforms the fixed-ratio upper bound in the 16-shot setting.

There are two strategies for combining knowledge from the pre-trained CLIP model and a trained adapter. By experimentation, we determine whether it is more effective to blend information in the feature space or in the final logit space. The comparison in Table S9 uses a fixed blending ratio weight of $\alpha = 0.7$. The two methods are as follows: Feature Blending, which performs a linear interpolation between the feature vector from the original CLIP image encoder, $\mathbf{f}_{\text{CLIP}}$, and the output of the adapter, $\mathbf{f}_{\text{adapter}}$. The approach then normalises the resulting blended feature vector, $\mathbf{f}_{\text{blend}} = \alpha \cdot \mathbf{f}_{\text{adapter}} + (1 - \alpha) \cdot \mathbf{f}_{\text{CLIP}}$, and computes the final logits. This method integrates the adapter's adjustments directly into the image representation before the classification head. The second approach, Logit

Table S5. **Details of the 11 datasets used for few-shot evaluation.**

Dataset	Classes	Splits (Tr/V/Te)	Task Category	Prompt Template
ImageNet	1000	1.28M / – / 50k	objects	`"a photo of a {}."`
Caltech101	100	4128 / 1649 / 2465	objects	`"a photo of a {}."`
OxfordPets	37	2944 / 736 / 3669	pets	`"a photo of a {}, a type of pet."`
StanfordCars	196	6509 / 1635 / 8041	cars	`"a photo of a {}."`
Flowers102	102	4093 / 1633 / 2463	flowers	`"a photo of a {}, a type of flower."`
Food101	101	50.5k / 20.2k / 30.3k	food	`"a photo of {}, a type of food."`
FGVCAircraft	100	3334 / 3333 / 3333	aircrafts	`"a photo of a {}, a type of aircraft."`
SUN397	397	15.9k / 4.0k / 19.9k	scenes	`"a photo of a {}."`
DTD	47	2820 / 1128 / 1692	textures	`"{} texture."`
EuroSAT	10	13.5k / 5.4k / 8.1k	satellite	`"a centered satellite photo of {}."`
UCF101	101	7639 / 1898 / 3783	actions	`"a photo of a person doing {}."`

Table S6. **CLIP-Adapter ViT-B/16 blending search (16-shot).** Classification accuracy for different blending weights (α). A low α favours CLIP features, while a high α favours the adapter.

Dataset	$\alpha = 0.2$	$\alpha = 0.4$	$\alpha = 0.6$	$\alpha = 0.8$
Caltech101	94.9	**95.7**	**95.7**	95.3
DTD	59.7	69.7	71.5	**71.6**
EuroSAT	70.5	80.3	84.1	**85.8**
FGVCAircraft	34.1	40.9	44.1	**45.8**
Flowers102	93.1	97.0	**97.3**	96.8
OxfordPets	92.6	**92.7**	92.2	90.7
StanfordCars	73.9	79.7	**81.9**	81.7
SUN397	74.2	**75.6**	75.1	73.7
UCF101	80.4	**84.0**	83.7	83.5
Average	74.82	79.51	**80.62**	80.54

Table S7. **CLIP-Adapter RN50 blending ratio search (16-shot).**

Dataset	$\alpha = 0.2$	$\alpha = 0.4$	$\alpha = 0.6$	$\alpha = 0.8$
Caltech101	91.4	**92.7**	92.6	91.8
DTD	59.2	66.1	**67.0**	66.2
EuroSAT	65.4	81.4	83.3	**84.0**
FGVCAircraft	24.5	32.1	35.0	**35.6**
Food101	**81.3**	79.9	77.9	75.7
Flowers102	89.7	94.0	94.7	**94.8**
OxfordPets	**88.6**	87.9	85.1	83.9
StanfordCars	67.6	73.5	**74.7**	73.2
SUN397	69.8	**70.3**	69.4	68.0
UCF101	76.2	78.8	**78.9**	77.9

Blending, operates at the end of the pipeline. It first computes two sets of logits: one using the original CLIP features ($\mathbf{l}_{\text{CLIP}}$) and another using the adapter's features ($\mathbf{l}_{\text{adapter}}$). The final prediction is a weighted average of these two logit vectors: $\mathbf{l}_{\text{final}} = \alpha \cdot \mathbf{l}_{\text{adapter}} + (1-\alpha) \cdot \mathbf{l}_{\text{CLIP}}$. This constitutes a form of model ensembling at the prediction level. As shown in Table S9, neither method is universally superior. However, feature blending shows a slight advantage on several datasets, which motivates its use in the primary experiments.

Table S10 indicates that the blending ratio optimiser exhibits relatively low sensitivity to the learning rate in the 16-shot RN50 setting, with a stable plateau from 0.001 to 0.4. The average accuracy peaks at 0.1, matching the per-dataset optimum on DTD, EuroSAT and UCF101. These results motivate choosing 0.1 as the learning rate for the blending ratio.

Figure S1 shows that HOSO-Adapter achieves peak average accuracy when the hold-out cache uses one shot, approximately six per cent of the sixteen-shot budget. Increasing the cache beyond one shot consistently reduces accuracy across ten datasets, indicating a trade-off between reliable blending-ratio estimation and sufficient adapter training data. We therefore propose a hold-one-shot-out cache.

J. Detailed Results

Table S11 shows that HOSO-Adapter matches CLIP-Adapter on ImageNet with an RN50 backbone across few-shot settings. Both methods improve as the number of shots increases, and their averages are effectively identical (61.25 versus 61.26). However, CLIP-Adapter in this case uses the test set to grid-search the optimal blending ratio for each dataset. In contrast, HOSO-Adapter learns the blending ratio under the strict few-shot protocol.

Table S1 shows that HOSO-Adapter with an RN50 backbone improves steadily with more shots, raising the average accuracy from 63.03 at 2-shot to 75.25 at 16-shot. The largest gains occur for EuroSAT, Flowers102, DTD and Stanford Cars, with improvements of +30.0, +21.17, +18.94 and +15.80 percentage points, respectively, while Food101 and ImageNet change modestly.

Table S8. **Comparison of different blending ratio strategies for CLIP-Adapter (ViT-B/16, 16-shot).** HOSO-Adapter uses a dynamically learned blending ratio.

Dataset	**Fixed** $\alpha = 0.6$	**Random** $\alpha \in [0, 1]$	**HOSO-Adapter**
Caltech101	95.7	95.4	**96.1**
DTD	71.5	67.0	**72.5**
EuroSAT	84.1	79.6	**86.6**
FGVCAircraft	44.1	37.3	**46.4**
Flowers102	**97.3**	94.0	97.1
OxfordPets	**92.2**	88.3	90.4
StanfordCars	**81.9**	69.9	80.4
SUN397	**75.1**	68.2	74.4
UCF101	**83.7**	79.4	83.3
Average	80.62	75.46	**80.80**

Table S9. **CLIP-Adapter ViT-B/16 feature vs. logit blending (16-shot; top-1 accuracy, %).** Both blending methods use a blending weight of $\alpha = 0.7$.

Dataset	Feature Blend	Logit Blend
Caltech101	**95.9**	94.3
DTD	**71.7**	70.7
EuroSAT	85.1	**85.5**
FGVC Aircraft	45.4	**46.4**
Flowers102	**97.4**	96.9
OxfordPets	**91.6**	90.5
Stanford Cars	82.1	**82.4**
SUN397	**74.4**	72.5
UCF101	**83.7**	82.5
Average	**80.8**	80.2

Table S10. **Ablation on the learning rate for the blending ratio optimiser.** Experiments use RN50 with 16-shots.

Dataset	**0.001**	**0.01**	**0.1**	**0.4**
DTD	65.5	66.5	**66.7**	66.6
EuroSAT	83.9	83.6	**84.0**	83.0
UCF101	76.2	77.4	**78.1**	77.9
Average	75.20	75.83	**76.27**	75.83

Performance on FGVC Aircrafts remains comparatively low but increases with shot count.

Table S12 shows that HOSO-Adapter scales reliably with additional supervision: the mean accuracy improves from 69.76 at 2-shot to 80.33 at 16-shot. The most significant gains occur on EuroSAT (+26.07) and DTD (+20.27). Consistent improvements are observed on fine-grained datasets such as FGVC Aircraft and Stanford Cars. Performance on datasets already aligned with CLIP, including Food101 and Caltech101, remains high with only marginal changes. EuroSAT displays a non-monotonic outcome, with 4-shot underperforming 2-shot. Inspection of the logs and code does not reveal an apparent reason for the decrease with the 4-shot setting. As this occurs only for this dataset and the same lower results are observed over three randomised seed runs, the outcome likely reflects variability in performance due to the significant domain gap between satellite images and CLIP's pretraining images.

Table S4 shows that HOSO-Adapter achieves an average of 80.33 at 16-shot, outperforming CLIP-Adapter (validation-free, hence $\alpha = 0.2$) by more than four percentage points while trailing the per-dataset-tuned CLIP-Adapter by less than one point. Together with the trend in Table S12, these results indicate that HOSO-Adapter delivers strong 16-shot performance without per-dataset hyperparameter search on the test set.

K. Class-level Perspective

Figure S2 illustrates the classification accuracy for each class within the EuroSAT dataset, comparing performance across different blending ratios (α). A ratio of zero corresponds to the zero-shot CLIP model, while a ratio of one relies solely on features from the trained adapter. The top panel (1-shot) shows significant performance variability, indicating that the optimal blending ratio is highly class-dependent. For instance, accuracy for classes such as "Residential Buildings" is high with ratios ranging from 0.7 to 0.9, whereas for "Herbaceous Vegetation Land" the best performance is around 0.8, despite a very low zero-shot baseline. This indicates that there is no single optimal ratio for all classes within a dataset, especially in low-data regimes. The bottom panel (16-shot) shows a marked improvement in overall accuracy and reduced sensitivity to the exact ratio.

Table S11. **ImageNet accuracy with RN50 backbone.** Average of three runs.

Method	2-shot	4-shot	8-shot	16-shot	Average
CLIP-Adapter	59.63	60.67	61.67	63.07	61.26
HOSO-Adapter	59.63	60.70	61.75	62.93	61.25

Table S12. **Main results with ViT-B/16 backbone, averaged over three runs.** Performance of HOSO-Adapter across different few-shot settings.

Method	Caltech101	DTD	EuroSAT	FGVC	Food101	ImageNet	Flowers	Pets	Cars	SUN397	UCF101	Average
HOSO-Adapter (2-shot)	94.20	50.40	59.23	27.73	88.20	67.53	82.07	90.00	67.73	67.03	73.27	69.76
HOSO-Adapter (4-shot)	94.67	59.57	45.10	32.57	87.97	68.67	89.43	91.63	70.90	70.13	77.10	71.61
HOSO-Adapter (8-shot)	94.83	65.40	79.87	37.47	88.90	69.83	95.17	91.57	75.37	72.60	80.73	77.43
HOSO-Adapter (16-shot)	**95.40**	**70.67**	**85.30**	**43.23**	**88.97**	**70.93**	**97.23**	**92.27**	**81.50**	**74.67**	**83.43**	**80.33**

With more shots, the adapter becomes more robust, and most ratios above 0.4 yield strong, comparable results that consistently outperform the zero-shot baseline.

Figure S3 presents a complementary view, plotting the accuracy of each class as a function of the blending ratio, α. This visualisation clarifies the impact of increasing the adapter's influence. In the 1-shot scenario (top), most classes exhibit a unimodal performance curve, starting at the zero-shot baseline ($\alpha = 0$), peaking at an intermediate ratio, and often declining as α approaches one. This pattern suggests that while the adapter provides crucial task-specific information, completely discarding the general-purpose features from the pre-trained CLIP encoder is detrimental. The optimal peak is class-specific, with some classes like "Forrest" peaking at lower adapter influence (*e.g.*, $\alpha \approx 0.2$) and others like "Permanent Crop Land" peaking at higher influence (*e.g.*, $\alpha \approx 1$). In the 16-shot setting (bottom), the performance curves show a steeper initial improvement, followed by a high-performance plateau for ratios above ≈ 0.5. This indicates that, with more data, training yields a sufficiently powerful adapter representation, making performance robust to a wide range of higher α values.

L. Augmentation Analysis

This analysis tests the possible addition of augmentations to HOSO-Adapter to make it more comparable with PathCLIP, which makes use of augmentations. In this approach, we create an additional, augmented view for each training image. The procedure applies weak and strong augmentations probabilistically, using policies such as RandAugment and RandomErasing. The original and the newly generated augmented views are passed to the model in a single forward pass, doubling the number of training views.

Comparing Hoso-Adapter with augmentations (Table S3) to the main RN50 baseline yields minor differences: -0.54, -0.64, -0.20, and -0.43 percentage points for the 2-, 4-, 8-, and 16-shot settings, respectively. The introduced augmentations cause small performance drops; hence, we did not investigate augmentations further.

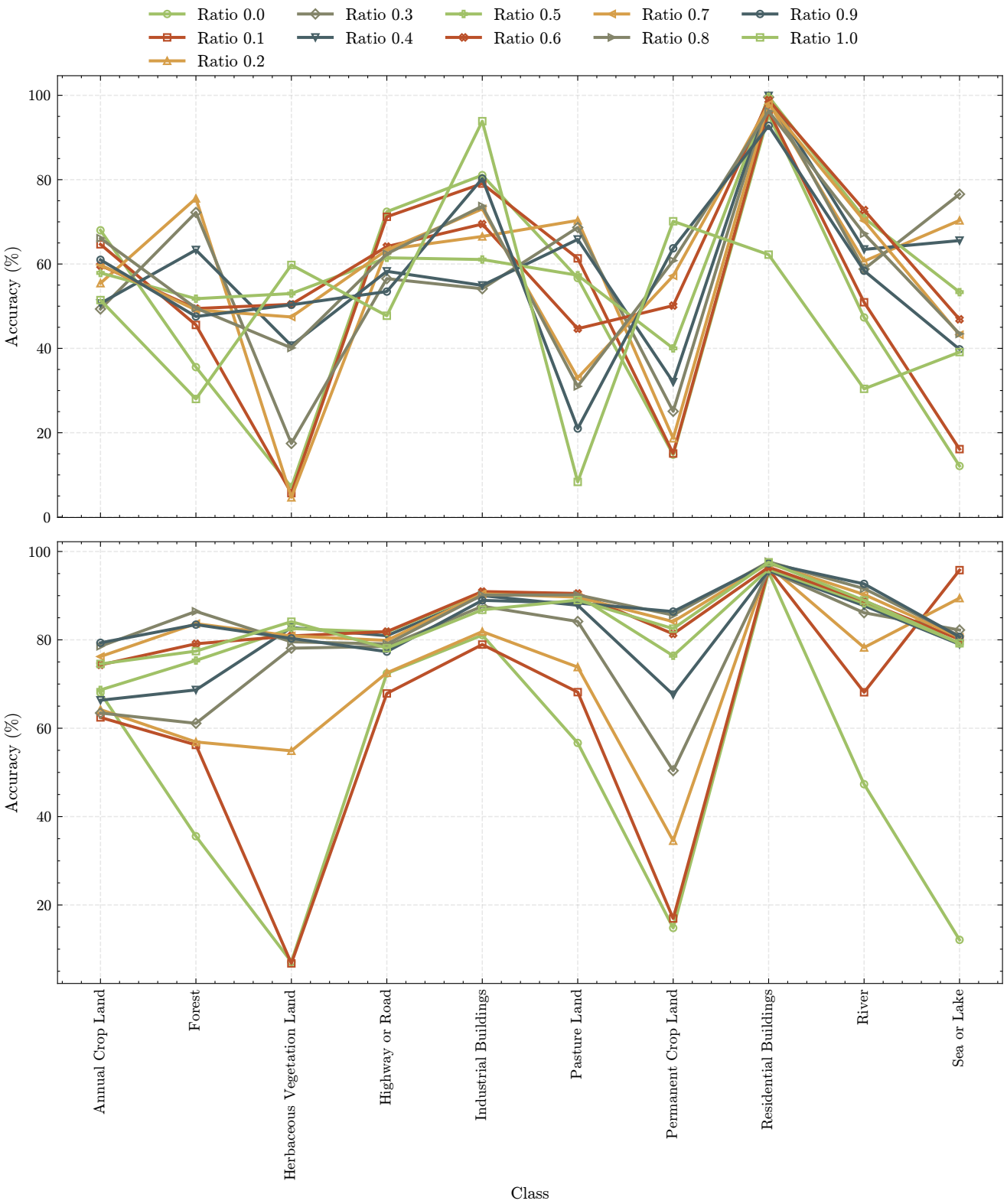

Figure S2. **Per-class accuracy as a function of class type for different blending ratios (α) on the EuroSAT dataset.** The top plot shows the results for 1-shot learning, while the bottom plot shows 16-shot learning. Each coloured line represents a fixed blending ratio. The analysis highlights that the optimal ratio varies significantly between classes, particularly in the 1-shot scenario.

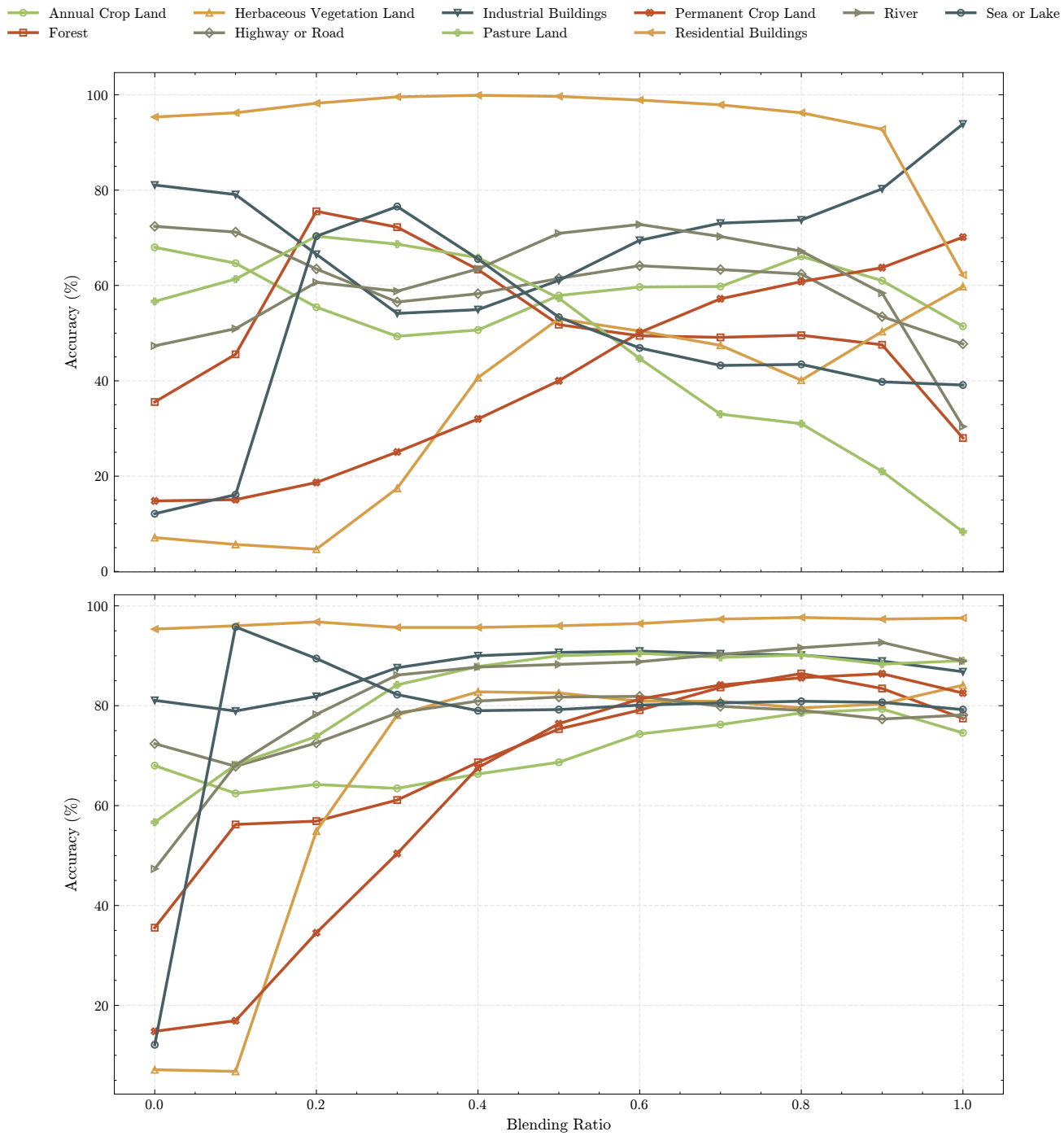

Figure S3. **Per-class accuracy as a function of the blending ratio** (α) on the EuroSAT dataset. The top plot corresponds to 1-shot and the bottom to 16-shot settings. Each coloured line represents a specific class. The plots show that performance generally improves from the 0-shot baseline ($\alpha = 0.0$) to an optimal intermediate ratio, with the 16-shot setting yielding a broader high-performance plateau.

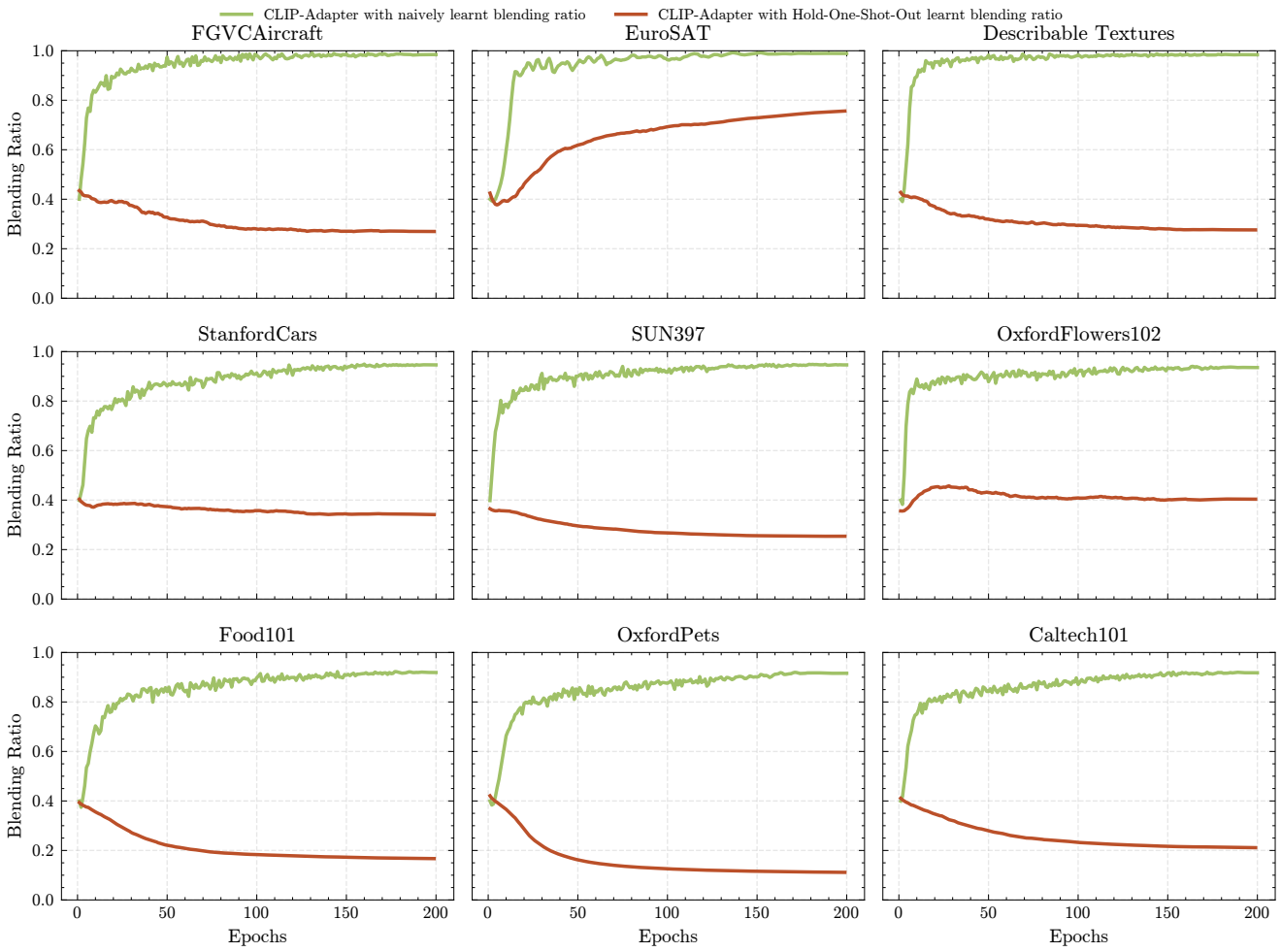

Figure S4. Most datasets exhibit a trend: *hold-one-shot-out* reduces the blending ratio (green) compared with a learnt blending ratio that leads to overfitting on the limited few-shot cases (red).

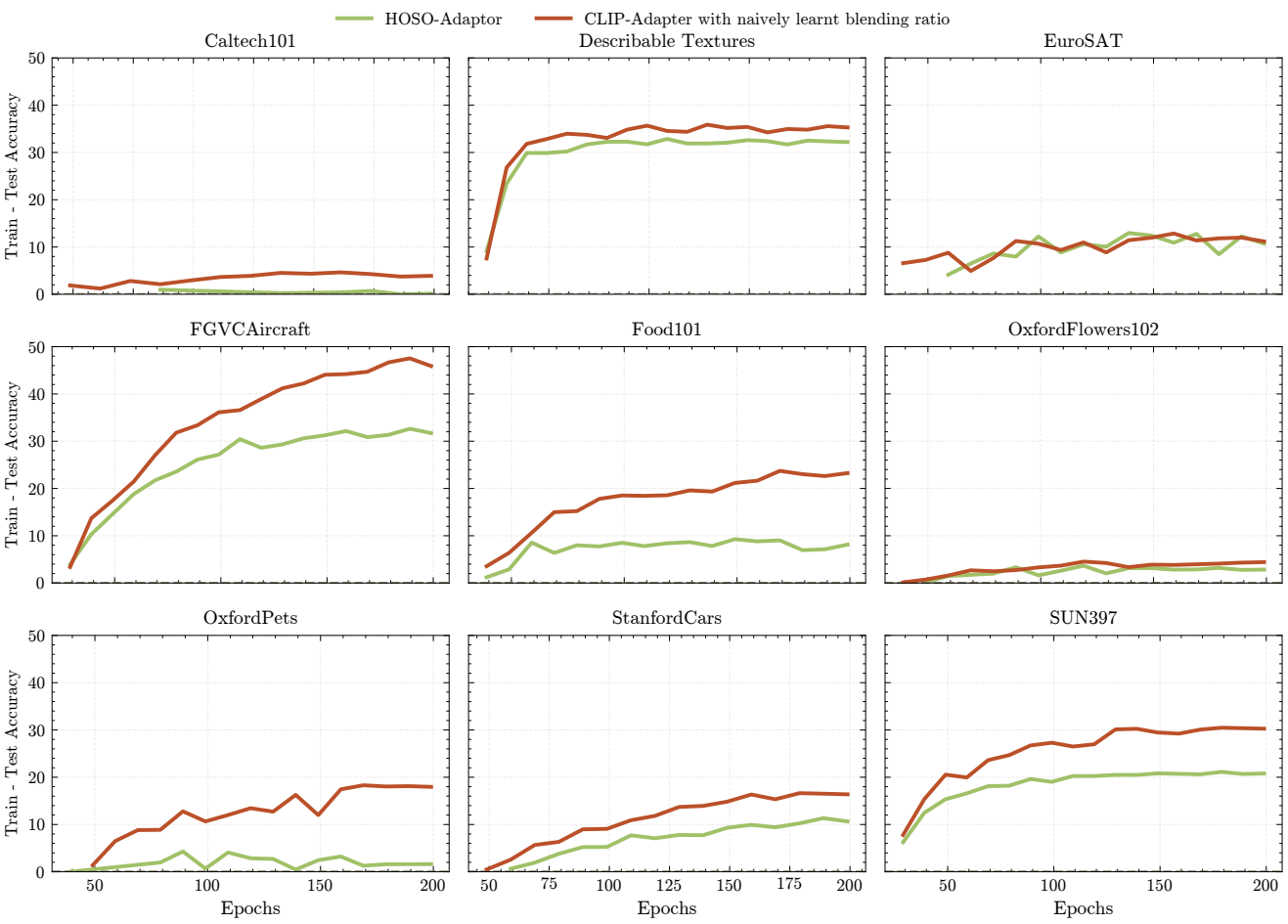

Figure S5. All datasets exhibit a consistent trend: *hold-one-shot-out* (green) exhibits less overfitting than the naively learnt blending ratio (red).