RARITY-AWARE VISION-LANGUAGE BENCHMARKING AND PROMPTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Prompt learning for vision-language models is predominantly evaluated under balanced supervision, leaving the effects of class imbalance on recognition and transfer insufficiently understood. We introduce **TailPromptBench**, a benchmark for long-tailed prompt generalization that varies class imbalance under a fixed annotation budget and evaluates recognition across frequency groups, base-to-new transfer, and robustness to visual domain shift. This design separates unequal supervision from changes in training-set size. We further propose **RAP**, a rarity-aware training parameterization for prompt learning based on a shared residual context. Its contribution increases with class rarity, providing stronger auxiliary corrections for underrepresented classes without introducing class-specific parameters. Joint optimization allows rarity-dependent contexts to shape the prompt parameters retained for deployment. The residual is discarded at inference, preserving the original structure and cost without requiring frequency estimates for unseen categories. Experiments show stronger tail-class recognition and average base-to-new harmonic-mean gains of up to 2.8 percentage points, with benefits extending to shifted domains. Ablations support the importance of aligning residual adaptation with the actual class-frequency structure.

1 INTRODUCTION

Large-scale vision-language models (VLMs) such as CLIP recognize categories specified through natural-language descriptions, providing a flexible foundation for visual recognition beyond a fixed training vocabulary (Radford et al., 2021; Jia et al., 2021; Zhai et al., 2022). Prompt learning adapts these models to downstream tasks by optimizing a small set of context tokens while keeping the pretrained encoders frozen (Zhou et al., 2022a). Its appeal lies in learning from limited labeled examples while retaining the ability to recognize categories absent from adaptation. Under long-tailed supervision, however, the available examples represent some categories far more extensively than others. Adaptation must then contend with unequal supervision while preserving the transfer capabilities that make VLMs useful in the first place.

Existing prompt-learning research commonly studies transfer through the base-to-new protocol, in which prompts are learned on base classes and evaluated on both those classes and disjoint, unseen new classes. Approaches such as conditional prompt learning improve this transfer under predominantly class-balanced supervision (Zhou et al., 2022b). Recent work has also investigated long-tailed VLM adaptation through frequency-aware objectives and representation constraints (Shi et al., 2024b; Guo et al., 2026a). Nevertheless, rare-class recognition, transfer to new classes, and robustness to visual domain shift are distinct objectives: improvement in one does not establish improvement in the others. The question we address is how reallocating the same supervision budget affects recognition and transfer together. Comparisons that change the total amount of training data cannot isolate this effect.

To examine these effects, we introduce **TailPromptBench**, a benchmark for long-tailed prompt generalization that progressively redistributes a fixed annotation budget across base classes. It combines base-to-new evaluation and class-frequency analysis on six source datasets with domain-shift evaluation of ImageNet-trained models on four target datasets. As illustrated in Figure 1, CoOp on ImageNet exhibits substantial head–tail disparities under severe imbalance, while its harmonic mean declines in both the source and shifted domains as imbalance increases. These observations motivate

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

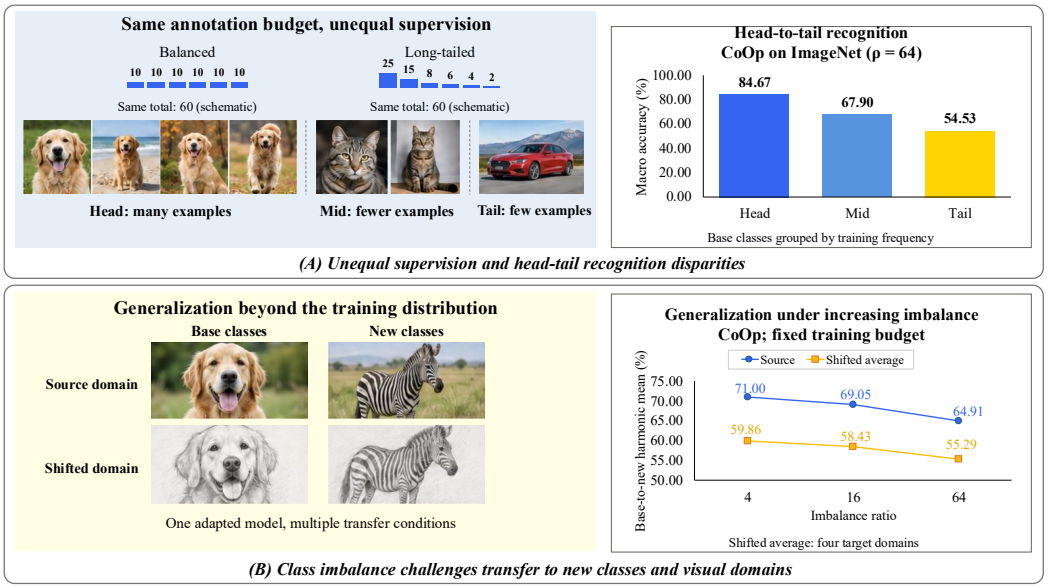


Figure 1: Long-tailed prompt generalization across classes and domains. (A) Balanced and long-tailed allocations use the same annotation budget (left); under severe imbalance, recognition accuracy differs substantially across head, medium, and tail base classes (right). (B) Evaluation covers source-domain base/new classes and shifted visual domains (left). As imbalance increases, the harmonic mean declines in the source domain and on average across four shifted domains (right).

a training mechanism that responds to supervision imbalance while retaining a prompt applicable beyond the source vocabulary.

We propose **RAP** a rarity-aware training parameterization for prompt learning based on a shared residual context. Class rarity controls the strength of this residual, allowing larger auxiliary adjustments for underrepresented classes. Sharing the residual pools its learning across categories and keeps the additional parameter count independent of vocabulary size. Joint optimization of the original prompt parameters and residual allows these rarity-conditioned contexts to shape the prompt learner retained for deployment. Source-class frequencies, however, describe the allocation of training supervision and provide no corresponding estimates for unseen categories. We therefore remove the residual after training and retain the optimized original prompt learner. This incorporates frequency information into adaptation while preserving the original inference structure and cost.

Across six datasets in TailPromptBench, RAP consistently improves the base-to-new harmonic mean of both evaluated prompt learners under severe imbalance, with average gains of up to 2.8 percentage points. These improvements are accompanied by stronger tail-class recognition and smaller head–tail disparities, and harmonic-mean gains extend to all four shifted domains. Ablations further show that conditioning the residual on the true class-frequency assignment yields higher harmonic means than constant or shuffled conditioning, supporting the importance of aligning adaptation with the distribution of supervision.

2 Related Work

Prompt Learning for Vision-Language Models. Prompt learning provides a lightweight way to adapt pretrained vision-language models to downstream recognition tasks. CoOp (Zhou et al., 2022a) learns continuous textual contexts, while CoCoOp (Zhou et al., 2022b) introduces image-conditioned prompts to improve transfer beyond the training categories. Subsequent work pursues generalization through several complementary strategies. TCP (Yao et al., 2024) and ATPrompt (Li et al., 2025) enrich prompts with class-level textual knowledge and transferable attributes, respectively. DePT (Zhang et al., 2024) and MMRL (Guo & Gu, 2025) separate adaptation-specific information from representations intended to support generalization. Other approaches improve the learning procedure: PromptKD (Li et al., 2024) distills a larger CLIP teacher into a smaller stu-

dent using unlabeled domain images, while PROMISE (Liang et al., 2026) uses meta-finetuning to improve robustness across prompt formulations. These studies address how prompt design and optimization preserve transfer during adaptation. Our work examines the complementary effect of unequal class supervision on this process.

Long-Tailed Adaptation of Vision-Language Models. Pretrained VLMs offer semantic representations that can support recognition when labeled examples are unevenly distributed across classes. Research in this area extends beyond prompt design to the choice of adaptation strategy. LIFT (Shi et al., 2024a) shows that heavy fine-tuning can harm tail-class performance and advocates lightweight adaptation, while Sage (Peng et al., 2026) incorporates textual semantics into visual adapters and compensates for class-conditional distribution mismatch. For long-tailed multi-label recognition, LMPT (Xia et al., 2024) learns class-specific contexts with an embedding objective that accounts for class imbalance. Closely related work explicitly considers transfer beyond the training categories. Candle (Shi et al., 2024b) combines logit adjustment, cross-modal feature adaptation, and virtual prototypes for unseen classes. LAMP (Wang et al., 2026) uses multimodal prompt pools and load balancing, while cluster-aware neural-collapse prompt tuning (Guo et al., 2026b) improves tail-class separation through local geometric constraints while preserving generalization. The latter evaluates imbalance by fixing the maximum per-class sample count, which allows the total training budget to change with the distribution. Our benchmark instead holds the total annotation budget fixed and evaluates base-to-new transfer alongside domain generalization. Methodologically, RAP conditions a single shared residual on class rarity during training and removes it at inference, retaining the original prompt learner's deployment structure and cost.

3 TAILPROMPTBENCH FOR LONG-TAILED PROMPT GENERALIZATION

We introduce **TailPromptBench**, a benchmark for long-tailed prompt generalization that allows for studying how unequal supervision affects the generalization of adapted VLMs. A central challenge in evaluating this effect is that changing class imbalance can also change the total amount of training data. TailPromptBench controls this confounding factor by redistributing a fixed annotation budget across base classes. Built on existing recognition datasets, it combines base-to-new evaluation on six source datasets with domain-shift evaluation of ImageNet-trained models on four target datasets. Figure 2 summarizes the supervision distributions and evaluation coverage.

3.1 CONTROLLED SUPERVISION ALLOCATION

Each source vocabulary is partitioned into disjoint base and new classes following the standard base-to-new protocol. Only source-domain images from base classes are available during adaptation; images from new classes and shifted domains are excluded from training. We take the corresponding balanced few-shot training set as the reference annotation budget and redistribute that budget across the base classes. On ImageNet (Deng et al., 2009), the reference contains 16 images for each of 500 base classes, giving a fixed total of 8,000 training images.

Let B be the number of base classes and N the annotation budget. For each random seed, we randomly permute the base-class identities to obtain a frequency ranking π, ordered from head to tail. We retain this ranking across imbalance levels within the seed, so a change in imbalance does not also change which classes occupy each rank. A nominal imbalance ratio ρ specifies the following continuous allocation:

$$q_{\pi(r)} = N \frac{\rho^{-r/(B-1)}}{\sum_{j=0}^{B-1} \rho^{-j/(B-1)}}, \qquad r = 0, \ldots, B-1. \tag{1}$$

We evaluate nominal ratios of 1, 4, 16, and 64, with 1 recovering the balanced allocation. The continuous counts are converted to integers, constrained to retain at least one image per base class and not exceed the available source images, and adjusted to preserve the total budget. The nominal ratio controls the allocation before integerization; the realized maximum-to-minimum count ratio may differ after these constraints are applied. At the most imbalanced ImageNet setting, 187 of the 500 base classes receive at most four training images, although the mean remains 16 images per class. This makes limited supervision for rare classes explicit without reducing the overall annotation budget. Appendix A provides the construction details and realized distribution statistics.

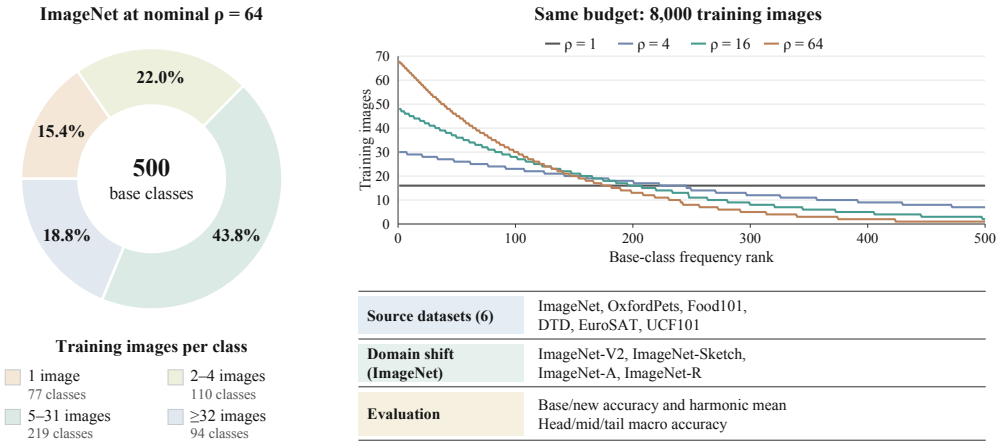


Figure 2: Overview of TailPromptBench. Left: the proportion of ImageNet base classes in each sample-count interval under severe imbalance. Upper right: class-frequency profiles generated by the allocation rule, each with the same budget of 8,000 training images across 500 base classes. Lower right: the source datasets, ImageNet domain variants, and evaluation metrics.

3.2 Evaluation Coverage and Metrics

The source datasets are ImageNet, OxfordPets, Food101, DTD, EuroSAT, and UCF101, covering object, fine-grained, texture, satellite-image, and action recognition. Each is treated as a separate adaptation task. For domain-shift evaluation, models trained on ImageNet are further evaluated on ImageNet-V2 (Recht et al., 2019), ImageNet-Sketch (Wang et al., 2019), ImageNet-A (Hendrycks et al., 2021b), and ImageNet-R (Hendrycks et al., 2021a), without using target-domain images for adaptation.

For the source-domain evaluation, base and new classes retain their training-defined identities. For each shifted dataset, the domain-shift protocol independently partitions its available class vocabulary into two halves. We refer to these as target-local base and new subsets; membership in a target-local subset does not, by itself, establish whether a category was observed during source training.

We report accuracy on each evaluation subset using its corresponding candidate labels, together with the harmonic mean of the two accuracies. To assess recognition across supervision levels, we additionally divide source-domain base classes into three approximately equal-sized groups according to their assigned frequency ranks. We report macro-averaged accuracy for the head, medium, and tail groups and the head–tail accuracy gap. This separates performance on classes with different supervision levels from aggregate base-to-new and domain-shift results.

4 Method

Long-tailed prompt generalization requires adapting to unequal supervision while retaining a prompt that remains applicable beyond the training vocabulary. A shared context couples the adaptation of all classes, although frequent classes supply many more training images than rare ones. Source-class frequencies expose this imbalance, but provide no corresponding statistic for categories absent from training. We propose RAP, a training-time reparameterization that uses this information to shape prompt optimization without making it a requirement for deployment. A shared residual introduces class-dependent context variations whose strength increases with rarity. Its parameters are learned across classes and optimized jointly with the original prompt parameters. After training, we discard the residual and retain the optimized prompt learner. Figure 3 illustrates the training and inference procedures.

4.1 Rarity-Aware Training Contexts

Let $V_\theta(x) \in \mathbb{R}^{M\times d}$ denote the general context produced by a prompt learner with trainable parameters θ, where M is the context length and d is the token dimension. The context may be shared

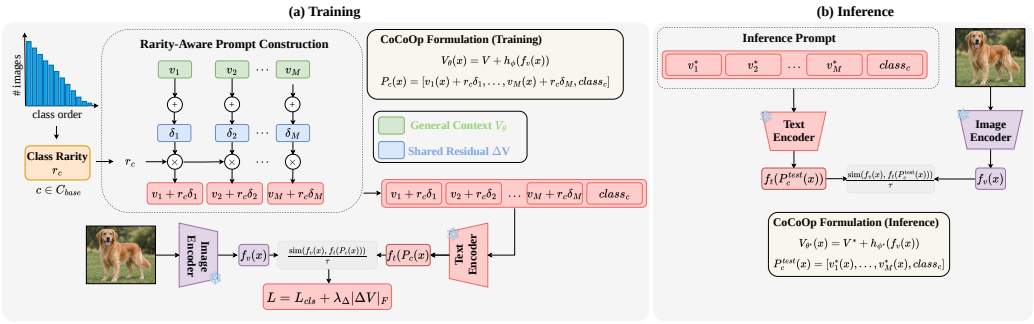


Figure 3: Overview of RAP. During training, each candidate class uses the general context augmented by a shared residual scaled by its rarity. The original prompt parameters and residual are optimized jointly with a penalty on the residual magnitude, while both VLM encoders remain frozen. At inference, the residual and rarity conditioning are removed, and the optimized original prompt learner is retained, including the image-conditioning network when used. The deployed model requires no class-frequency estimates and has the same architecture and inference cost as the underlying prompt learner.

across images or conditioned on the input image. We introduce a learnable residual $\Delta V \in \mathbb{R}^{M\times d}$, shared across all classes and images. For a base class c, the training context and corresponding text prompt are

$$U_c(x) = V_\theta(x) + r_c \Delta V, \qquad P_c(x) = [U_c(x), \text{class}_c], \tag{2}$$

where class_c denotes the class-name tokens and $r_c \in [0, 1]$ encodes the rarity of class c.

The residual defines a common direction of variation in context-token space, and rarity determines each class's displacement along that direction. For the same image, the context difference between two classes is $(r_c - r_j)\Delta V$: classes with similar supervision receive similar auxiliary corrections, while rarer classes receive a larger residual contribution. A constant coefficient would yield the same context correction for every class; conditioning on rarity instead introduces variation aligned with the supervision distribution. Sharing the residual pools its learning across classes and introduces only Md parameters, independent of vocabulary size.

We compute rarity from the realized training counts. Let n_c be the number of training images for class c, with $n_{\max}$ and $n_{\min}$ the largest and smallest counts over base classes. We use

$$r_c = \begin{cases} \dfrac{\log(n_{\max}+1) - \log(n_c+1)}{\log(n_{\max}+1) - \log(n_{\min}+1)}, & n_{\max} > n_{\min}, \\ 0, & n_{\max} = n_{\min}. \end{cases} \tag{3}$$

The logarithm compresses differences in raw sample counts before normalization. Under imbalanced supervision, the most frequent classes have zero residual contribution, and the rarest classes receive the full residual. The coefficients are computed once from the source training set and remain fixed throughout optimization. For balanced supervision, setting all coefficients to zero recovers the original prompt parameterization.

This formulation covers both shared and image-conditioned prompt learners. For CoOp, $V_\theta(x) = V$ is a learned context shared across inputs. For CoCoOp, $V_\theta(x) = V + h_\phi(f_v(x))$, where the image-dependent bias is added to each context token and θ includes both V and the meta-network parameters ϕ. Image conditioning captures variation across inputs, while the residual introduces variation across classes according to their supervision.

4.2 Learning Through the Shared Residual

The image and text encoders, f_v and f_t, remain frozen. Training compares each image representation $f_v(x)$ with the text representations $f_t(P_c(x))$ of all base classes using temperature-scaled cosine similarities. A softmax over these logits defines $p_{\theta,\Delta V}(c \mid x)$. We jointly optimize the original prompt parameters and residual using

$$\mathcal{L}(\theta, \Delta V) = -\frac{1}{N} \sum_{(x,y)\in \mathcal{D}^s_{\text{train}}} \log p_{\theta,\Delta V}(y \mid x) + \lambda_\Delta \|\Delta V\|_F, \tag{4}$$

where N is the number of training images and $\|\cdot\|_F$ denotes the unsquared Frobenius norm.

How rarity changes learning. For a single training example, let ℓ be its classification loss and $G_c = \partial\ell/\partial U_c(x)$ the gradient through the context of candidate class c. The additive construction gives

$$\frac{\partial \ell}{\partial V_\theta(x)} = \sum_{c \in \mathcal{C}_{\text{base}}} G_c, \qquad \frac{\partial \ell}{\partial \Delta V} = \sum_{c \in \mathcal{C}_{\text{base}}} r_c G_c. \tag{5}$$

The general context aggregates contributions from all candidate classes, whereas rarity weights their contributions to the residual update. Both sums include positive and negative classes in the softmax objective. During joint optimization, the scores and gradients used to update θ are evaluated at rarity-dependent contexts. Source-frequency information therefore influences the learning of the retained parameters even though its explicit conditioning is later removed.

Controlling the auxiliary correction. We initialize ΔV to zero, so the initial training prompts coincide with those of the original learner. The residual penalty controls the magnitude of the learned correction. Since every rarity coefficient lies in $[0, 1]$, the context displacement introduced by the residual satisfies $\|U_c(x) - V_\theta(x)\|_F = r_c\|\Delta V\|_F \le \|\Delta V\|_F$. Regularization therefore also limits the context-level discrepancy created when the residual is removed for deployment.

4.3 Frequency-Agnostic Inference

Class rarity describes the allocation of source training supervision. Categories absent from that supervision have no corresponding observed training frequency, so retaining rarity conditioning would require an additional rule for assigning their coefficients. We avoid this dependence by discarding ΔV after optimization and constructing every inference prompt as

$$P_c^{\text{test}}(x) = [V_{\theta^\star}(x), \text{class}_c], \tag{6}$$

for any candidate class c, whether base or new. For CoOp, this retains the optimized shared context; for CoCoOp, it retains both that context and the optimized image-conditioning network.

The deployed model requires no class-frequency estimates and has the same architecture and inference cost as the underlying prompt learner. It can be evaluated on both source and shifted images without target-domain adaptation.

5 Experiments

5.1 Experimental Setup

Where the recognition gains occur. Figure 5 resolves the changes by source-class frequency. Tail accuracy improves in all 12 comparisons, by 10.77 points on average for CoOp and 9.14 for CoCoOp. These gains accompany mean head-accuracy decreases of 4.68 and 2.42 points, respectively. The mean head–tail gap falls from 32.23 to 16.79 for CoOp and from 29.98 to 18.42 for CoCoOp. Together with the improvements in H, this shows a shift toward stronger rare-class recognition, with an explicit cost on frequent classes. Full group accuracies and gaps are reported in Appendix B.

Evaluation protocol. We follow TailPromptBench (Section 3) and evaluate each prompt learner with and without RAP under the same fixed-budget training protocol. The main comparison uses the most imbalanced setting, $\rho = 64$; additional experiments vary ρ to examine when rarity-aware adaptation is useful. We report source-domain base/new accuracy and their harmonic mean (H), target-domain H under the specified target-local splits, and macro accuracy across head, medium, and tail base classes. Dataset definitions and sampling rules are given in the benchmark section, with realized training statistics in Appendix A.

Implementation and reporting. We instantiate RAP with CoOp (Zhou et al., 2022a) and CoCoOp (Zhou et al., 2022b), using CLIP ViT-B/16 with both encoders frozen. The four general-context tokens are initialized from ``a photo of a''; the four residual tokens are initialized to zero. We retain the optimization settings of the corresponding baseline and use an unsquared Frobenius penalty with $\lambda_\Delta = 0.1$ for CoOp and 0.001 for CoCoOp. Each coefficient is shared across datasets and selected using the reported six-dataset mean H in the regularization sweep. Results are means over three random seeds, obtained on a single NVIDIA RTX 6000 Ada GPU. All

Table 1: Base-to-new generalization on TailPromptBench at $\rho = 64$. RAP improves the base-to-new harmonic mean across all six datasets for both prompt learners.

Dataset	CoOp			CoOp + RAP			ΔH
	Base	New	H	Base	New	H	
ImageNet	69.07	61.23	64.91	74.57	66.43	70.26	+5.35
OxfordPets	90.03	94.83	92.36	94.13	96.60	95.35	+2.99
Food101	84.47	87.83	86.11	88.50	89.23	88.86	+2.75
DTD	68.60	47.57	56.18	68.27	**52.00**	59.03	+2.85
EuroSAT	70.67	59.53	64.62	**72.53**	**60.90**	**66.21**	+1.59
UCF101	77.20	65.00	70.57	**79.13**	65.00	71.37	+0.80
Average	76.67	69.33	72.46	**79.52**	**71.69**	**75.18**	+2.72
Dataset	CoCoOp			CoCoOp + RAP			ΔH
	Base	New	H	Base	New	H	
ImageNet	72.30	66.70	69.39	**75.40**	**69.70**	**72.43**	+3.04
OxfordPets	92.90	95.23	94.05	**95.37**	**97.63**	**96.49**	+2.44
Food101	87.53	90.30	88.89	**90.57**	**91.40**	**90.98**	+2.09
DTD	65.67	46.37	54.35	**70.60**	51.60	**59.62**	+5.27
EuroSAT	54.17	53.73	53.94	64.37	47.53	54.68	+0.74
UCF101	73.60	68.47	70.94	77.87	**70.63**	**74.07**	+3.13
Average	74.36	70.13	71.93	79.03	71.42	74.71	+2.79

Table 2: Performance under visual domain shift at $\rho = 64$. RAP consistently improves the harmonic mean across all four target domains for both prompt learners.

	Source	Target				
Method	ImageNet	V2	Sketch	A	R	Average
CoOp	64.91	58.65	43.82	46.82	71.89	55.30
CoOp + RAP	70.26	63.87	48.51	50.82	76.35	59.89
ΔH	+5.35	+5.22	+4.69	+4.00	+4.46	+4.59
CoCoOp	69.39	63.12	48.06	50.05	75.43	59.17
CoCoOp + RAP	**72.43**	**66.16**	**50.82**	**52.82**	**78.14**	**61.99**
ΔH	+3.04	+3.04	+2.76	+2.77	+2.71	+2.82

gains are absolute percentage-point differences. Dataset/domain averages and paired gains are computed before final rounding.

5.2 Generalization Under Unequal Supervision

Base-to-new transfer. Table 1 shows that RAP improves H for both prompt learners on every source dataset, yielding 12 positive paired comparisons. The mean gains are 2.72 points for CoOp and 2.79 points for CoCoOp. Both base and new accuracy improve on average, indicating that the aggregate gains extend beyond fitting the observed classes. Individual trade-offs remain: on EuroSAT, CoCoOp gains 10.20 points on base classes but loses 6.20 on new classes, resulting in a smaller H improvement of 0.74. Thus, the consistent improvement concerns the joint base-to-new metric, rather than every individual accuracy.

Transfer under visual domain shift. We next evaluate ImageNet-trained models on four shifted domains without additional adaptation (Table 2). At $\rho = 64$, RAP improves H on every target domain for both learners, with average gains of 4.59 and 2.82 points for CoOp and CoCoOp, respectively. CoCoOp with RAP achieves the highest target H among the four evaluated configurations. These results show that the benefits of rarity-aware training extend to shifted visual inputs. The target-local class partitions measure domain-shift performance; they do not independently establish

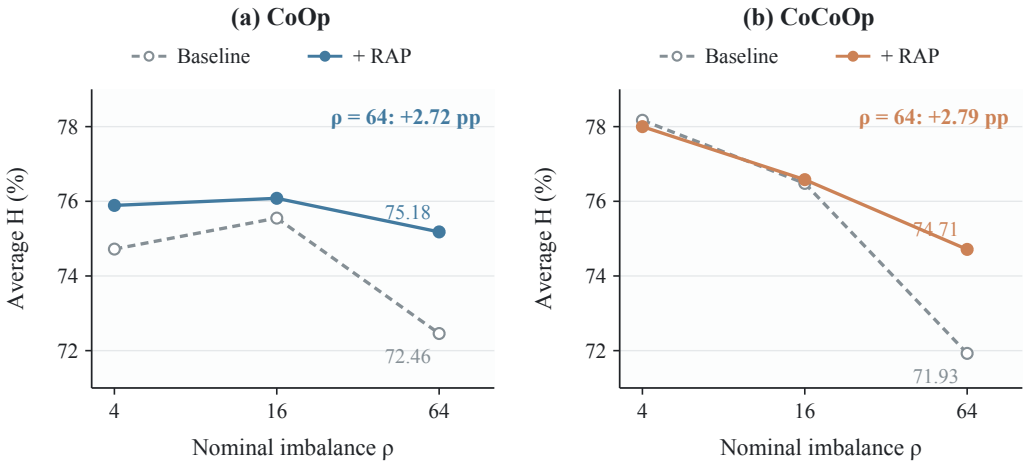


Figure 4: Effect of imbalance severity at a fixed training budget. RAP achieves its largest harmonic-mean gains under the most imbalanced allocation for both prompt learners.

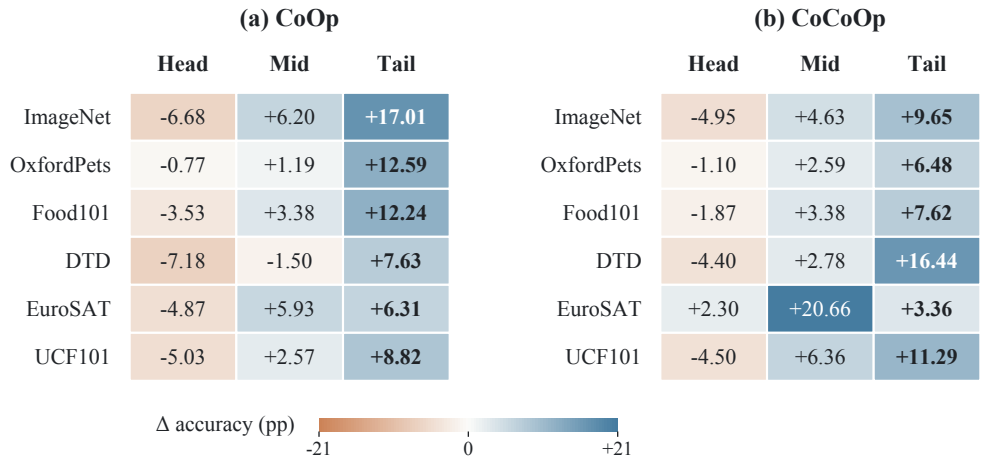


Figure 5: Recognition changes across class-frequency groups. RAP consistently improves tail-class recognition, while head-class accuracy decreases in most settings.

transfer to classes unseen during source training. Complete results at all imbalance levels appear in Appendix C.

Dependence on imbalance severity. Figure 4 compares the same methods as supervision becomes more unequal. The largest gains occur at $\rho = 64$ for both learners. At smaller ratios, the benefit is more limited and depends on the learner; CoCoOp changes little at $\rho = 16$ and slightly declines at $\rho = 4$. This pattern supports the usefulness of rarity conditioning when the allocation of supervision is strongly skewed, while showing that the added training flexibility is not uniformly beneficial under mild imbalance.

5.3 Analyzing Rarity-Aware Optimization

Role of the rarity assignment. Table 3 compares the original learner, a constant residual coefficient, shuffled rarity coefficients, and their true class-frequency assignment. True rarity achieves the highest H for both learners. Relative to constant conditioning, it improves H by 2.61 points for CoOp and 1.54 for CoCoOp; relative to shuffled conditioning, the gains are 3.47 and 1.93 points. These controls test whether the residual benefits from conditioning that matches the supervision distribution. The results support that alignment, although constant conditioning gives the highest new-class accuracy for CoCoOp.

Table 3: Effect of the rarity assignment. Conditioning the shared residual on true class rarity yields the highest base-to-new harmonic mean for both prompt learners.

CoOp	Base	New	H	ΔH
No residual ($r_c = 0$)	76.67	69.33	72.46	0.00
Constant ($r_c = 0.5$)	76.33	69.95	72.57	+0.11
Shuffled rarity	74.57	69.70	71.71	-0.75
True rarity (RAP)	**79.52**	**71.69**	**75.18**	+2.72
CoCoOp	Base	New	H	ΔH
No residual ($r_c = 0$)	74.36	70.13	71.93	0.00
Constant ($r_c = 0.5$)	75.12	**71.81**	73.17	+1.24
Shuffled rarity	74.42	71.72	72.78	+0.85
True rarity (RAP)	**79.03**	71.42	**74.71**	+2.79

Residual regularization. Figure 6 varies λ_Δ from 10^{-4} to 10. The highest observed six-dataset mean H occurs at 0.1 for CoOp and 0.001 for CoCoOp, which are the settings used above. Increasing the penalty suppresses the learned residual; at $\lambda_\Delta = 10$, H falls to 73.11 and 72.63, respectively. This is consistent with an auxiliary correction that needs room to affect optimization, while the different response curves indicate that its appropriate scale depends on the underlying learner. The full sweep and parameter-selection scope are reported in Appendix D.

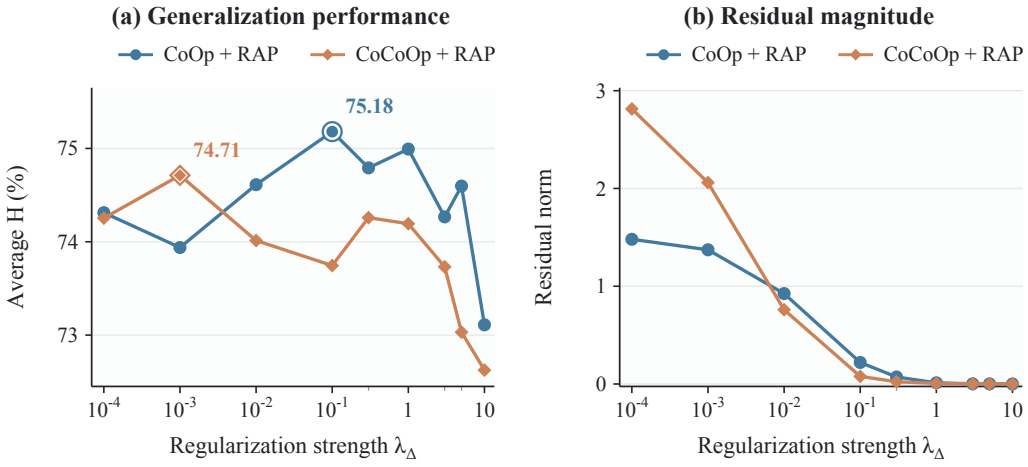


Figure 6: Effect of residual regularization. Excessive regularization drives the residual toward zero and reduces the harmonic mean for both prompt learners.

Parameter overhead. With four context tokens of dimension 512, RAP adds 2,048 trainable parameters. The residual is removed after training, so the deployed model retains the architecture and inference cost of the corresponding baseline.

6 Conclusion

We introduced TailPromptBench, a benchmark for studying long-tailed prompt generalization by varying the allocation of supervision under a fixed annotation budget. To address unequal supervision while retaining applicability to unseen categories, we proposed RAP for prompt learning, which conditions a shared residual context on class rarity during training and removes it at inference. This design allows frequency information to influence prompt optimization without changing the deployed model's architecture or computational cost. Experiments demonstrate improved base-to-new harmonic mean across six datasets and gains under visual domain shift, together with stronger tail-class recognition and reduced head–tail disparities. Ablations further show that aligning the residual contribution with the actual class-frequency structure matters beyond introducing additional parameters.

AI USE STATEMENT

We used ChatGPT to polish the writing of this manuscript, including grammar, clarity, and concision. The methodology, model training, evaluation, experimental design, interpretation of results, figures, tables, and code were developed and produced entirely by the authors. We reviewed all AI-assisted text before incorporating it and remain responsible for the final manuscript.

ETHICS STATEMENT

TailPromptBench reorganizes existing public image-classification benchmarks and involves no new data collection or human-subject recruitment. The frequency groups studied in this work are defined by class sample counts rather than demographic attributes, and improved tail-class recognition should not be interpreted as evidence of demographic fairness. Pretrained vision-language models and their datasets may contain social biases and privacy concerns that are not addressed by RAP. Evaluation on the selected benchmarks does not establish reliability in safety-critical settings, which would require application-specific validation. Our planned release will provide code and split definitions; access to the underlying images remains subject to the original datasets' terms.

REPRODUCIBILITY STATEMENT

Section 3 describes the TailPromptBench construction and evaluation protocol, with realized class-frequency statistics in Appendix A. Section 4 specifies the prompt parameterization, rarity computation, training objective, and inference procedure. Section 5 reports the experimental configuration and result-aggregation conventions, while Appendix D documents the regularization sweep and coefficient-selection procedure. We will release the implementation, training configurations, and exact training and evaluation split definitions to support independent replication.

REFERENCES

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Boyang Guo, Liang Li, Lin Peng, Yuhan Gao, Xichun Sheng, and Chenggang Yan. Cluster-aware neural collapse prompt tuning for long-tailed generalization of vision-language models, 2026a. URL https://arxiv.org/abs/2605.11939.

Boyang Guo, Liang Li, Lin Peng, Yuhan Gao, Xichun Sheng, and Chenggang Yan. Cluster-aware neural collapse prompt tuning for long-tailed generalization of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026b. URL https://arxiv.org/abs/2605.11939.

Yuncheng Guo and Xiaodong Gu. Mmrl: Multi-modal representation learning for vision-language models, 2025. URL https://arxiv.org/abs/2503.08497.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12:2217–2226, 08 2017. doi: 10.1109/JSTARS.2019.2918242.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization, 2021a. URL https://arxiv.org/abs/2006.16241.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15262–15271, June 2021b.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision, 2021. URL https://arxiv.org/abs/2102.05918.

Zheng Li, Xiang Li, Xinyi Fu, Xin Zhang, Weiqiang Wang, Shuo Chen, and Jian Yang. PromptKD: Unsupervised prompt distillation for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26617–26626, 2024.

Zheng Li, Yibing Song, Ming-Ming Cheng, Xiang Li, and Jian Yang. Advancing textual prompt learning with anchored attributes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3618–3627, 2025.

Haohui Liang, Runlin Huang, Yingjun Du, Yujia Hu, Weifeng Su, and Cees G. M. Snoek. PROMISE: Prompt-robust vision-language models via meta-finetuning. In *The Fourteenth International Conference on Learning Representations*, 2026.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3498–3505, 2012. doi: 10.1109/CVPR.2012.6248092.

Yufei Peng, Yonggang Zhang, and Yiu-ming Cheung. Semantic-guided fine-tuning of foundation model for long-tailed visual recognition. *International Journal of Computer Vision*, 134:322, 2026. doi: 10.1007/s11263-026-02874-4.

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, 2021. URL https://arxiv.org/abs/2103.00020.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet?, 2019. URL https://arxiv.org/abs/1902.10811.

Jiang-Xin Shi, Tong Wei, Zhi Zhou, Jie-Jing Shao, Xin-Yan Han, and Yu-Feng Li. Long-tail learning with foundation model: Heavy fine-tuning hurts. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 45014–45039. PMLR, 2024a.

Jiang-Xin Shi, Chi Zhang, Tong Wei, and Yu-Feng Li. Efficient and long-tailed generalization for pre-trained vision-language model. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2663–2673. Association for Computing Machinery, 2024b. doi: 10.1145/3637528.3671945.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild, 2012. URL https://arxiv.org/abs/1212.0402.

Haohan Wang, Songwei Ge, Eric P. Xing, and Zachary C. Lipton. Learning robust global representations by penalizing local predictive power, 2019. URL https://arxiv.org/abs/1905.13549.

Xucong Wang, Pengkun Wang, Zhe Zhao, Liheng Yu, and Yang Wang. LAMP: Long-tailed multimodal prompt learning for vision-language models, 2026. URL https://openreview.net/forum?id=amBzV6V3tQ.

Peng Xia, Di Xu, Ming Hu, Lie Ju, and Zongyuan Ge. Lmpt: Prompt tuning with class-specific embedding loss for long-tailed multi-label visual recognition, 2024. URL https://arxiv.org/abs/2305.04536.

Hantao Yao, Rui Zhang, and Changsheng Xu. TCP: Textual-based class-aware prompt tuning for visual-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23438–23448, 2024.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning, 2022. URL https://arxiv.org/abs/2111.07991.

Ji Zhang, Shihan Wu, Lianli Gao, Heng Tao Shen, and Jingkuan Song. Dept: Decoupled prompt tuning, 2024. URL https://arxiv.org/abs/2309.07439.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022a. ISSN 1573-1405. doi: 10.1007/s11263-022-01653-1. URL http://dx.doi.org/10.1007/s11263-022-01653-1.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models, 2022b. URL https://arxiv.org/abs/2203.05557.

A Additional Details of TailPromptBench

Allocation and sampling. We form each training set by assigning the base classes to frequency ranks using a seeded random permutation, allocating the fixed budget according to the exponential profile in Eq. 1, and converting the resulting counts to integers. Each base class retains at least one image. Counts exceeding the available source images are capped, and the remaining budget is reassigned to classes with capacity. The same class-to-rank assignment is retained across imbalance levels within a seed. Training images are sampled without replacement from each base class; new-class and target-domain images are excluded.

Dataset coverage and budgets. The six source tasks comprise ImageNet (Deng et al., 2009), OxfordPets (Parkhi et al., 2012), Food101 (Bossard et al., 2014), DTD (Cimpoi et al., 2014), EuroSAT (Helber et al., 2017), and UCF101 (Soomro et al., 2012). Each task has its own budget of 16 images per base class before redistribution. The resulting budgets are 8,000, 304, 816, 384, 80, and 816 images, respectively. These are separate adaptation tasks, not a combined training mixture.

Realized class frequencies. Table 4 reports the complete realized training-distribution statistics for ImageNet and the five downstream datasets. The nominal imbalance ratio ρ parameterizes the continuous allocation weights before integerization, whereas the realized Max/Min ratio is computed from the final per-class sample counts. Integerization and dataset-availability constraints can therefore cause the realized ratio to differ from its nominal value. For example, ImageNet at nominal $\rho = 16$ has per-class counts ranging from 2 to 48, corresponding to a realized Max/Min ratio of 24. We therefore distinguish between nominal and realized imbalance throughout the reported statistics.

Table 4: Realized training distributions under different nominal imbalance ratios ρ. All datasets use a fixed budget of 16 images per base class on average. Head and Tail denote the maximum and minimum realized per-class sample counts, respectively. Max/Min denotes their ratio.

	Setup			Samples per class			Distribution statistics			Number of classes		
Dataset	**Base/New**	**Budget**	ρ	**Head**	**Tail**	**Mean**	**Median**	**Std.**	**Max/Min**	$n_c = 1$	$n_c \leq 4$	$n_c \geq 32$
ImageNet	500/500	8000	1	16	16	16.0	16.0	0.0	1.00	0	0	0
			4	30	7	16.0	14.0	6.8	4.29	0	0	0
			16	48	2	16.0	11.0	12.5	24.00	0	95	76
			64	68	1	16.0	8.0	17.5	68.00	77	187	94
OxfordPets	19/18	304	1	16	16	16.0	16.0	0.0	1.00	0	0	0
			4	30	7	16.0	15.0	7.2	4.29	0	0	0
			16	46	2	16.0	12.0	13.0	23.00	0	4	3
			64	64	1	16.0	7.0	18.2	64.00	4	7	4
Food101	51/50	816	1	16	16	16.0	16.0	0.0	1.00	0	0	0
			4	30	7	16.0	15.0	6.8	4.29	0	0	0
			16	47	2	16.0	11.0	12.7	23.50	0	10	8
			64	67	1	16.0	9.0	17.7	67.00	8	19	10
DTD	24/23	384	1	16	16	16.0	16.0	0.0	1.00	0	0	0
			4	30	7	16.0	14.5	7.1	4.29	0	0	0
			16	47	2	16.0	12.0	13.0	23.50	0	5	4
			64	60	1	16.0	7.5	18.0	60.00	4	9	5
EuroSAT	5/5	80	1	16	16	16.0	16.0	0.0	1.00	0	0	0
			4	29	7	16.0	14.0	7.8	4.14	0	0	0
			16	42	2	16.0	10.0	14.5	21.00	0	1	1
			64	53	1	16.0	6.0	19.5	53.00	1	2	1
UCF101	51/50	816	1	16	16	16.0	16.0	0.0	1.00	0	0	0
			4	30	7	16.0	15.0	6.8	4.29	0	0	0
			16	47	2	16.0	11.0	12.7	23.50	0	10	8
			64	67	1	16.0	9.0	17.7	67.00	8	19	10

B Complete Head-to-Tail Results

Table 5 provides the absolute accuracies underlying Figure 5. Head, medium, and tail groups follow the frequency ranks assigned during construction and contain approximately equal numbers of base classes. The mean row weights datasets equally. All 12 tail-class comparisons improve, but head accuracy decreases in 11 cases. CoCoOp on EuroSAT is the exception, improving both head and tail accuracy; its remaining head–tail gap of 77.91 points also shows that rarity conditioning does not eliminate severe frequency disparities in every setting.

Table 5: Complete head-to-tail results at $\rho = 64$. Head, Mid and Tail are macro accuracies over frequency-rank groups. Gap is Head minus Tail (lower is better). Bold marks the better member of each baseline/RAP pair. Original per-dataset gap values are retained; rounding can cause a 0.01 difference from subtracting displayed accuracies.

Dataset	CoOp				CoOp + RAP			
	Head	Mid	Tail	Gap↓	Head	Mid	Tail	Gap↓
ImageNet	**84.67**	67.90	54.53	30.15	77.99	**74.10**	**71.54**	**6.45**
OxfordPets	**95.66**	93.92	79.69	15.98	94.89	**95.11**	**92.28**	**2.62**
Food101	**92.92**	86.23	74.30	18.61	89.39	**89.61**	**86.54**	**2.85**
DTD	**87.85**	**75.69**	42.25	45.60	80.67	74.19	**49.88**	**30.79**
EuroSAT	**93.59**	69.96	31.38	62.21	88.72	**75.89**	**37.69**	**51.03**
UCF101	**87.88**	74.32	67.04	20.84	82.85	**76.89**	**75.86**	**6.99**
Average	**90.43**	78.00	58.20	32.23	85.75	**80.97**	**68.97**	**16.79**
Dataset	CoCoOp				CoCoOp + RAP			
	Head	Mid	Tail	Gap↓	Head	Mid	Tail	Gap↓
ImageNet	**81.98**	70.10	64.77	17.21	77.03	**74.73**	**74.42**	**2.61**
OxfordPets	**97.47**	93.64	86.88	10.59	96.37	**96.23**	**93.36**	**3.01**
Food101	**92.68**	87.29	82.61	10.07	90.81	**90.67**	**90.23**	**0.58**
DTD	**82.41**	71.41	43.17	39.24	78.01	**74.19**	**59.61**	**18.40**
EuroSAT	89.28	45.25	10.31	78.97	**91.58**	**65.91**	**13.67**	**77.91**
UCF101	**86.20**	69.52	62.42	23.78	81.70	**75.88**	**73.71**	**7.99**
Average	**88.34**	72.87	58.36	29.98	85.92	**79.60**	**67.50**	**18.42**

C Complete Source and Shifted-Domain Results

Tables 6 and 7 report Base/New/H at all three long-tailed settings. Source labels preserve the training-defined base/new identity. Target labels follow the target-local partition in Section 3; target-new membership does not itself establish absence from source training. Target averages give equal weight to the four shifted domains and exclude the ImageNet source results.

For CoOp, mean target H gains increase from 0.91 points at $\rho = 4$ to 1.93 at $\rho = 16$ and 4.59 at $\rho = 64$. CoCoOp changes from a 0.10-point decline at $\rho = 4$ to gains of 0.88 and 2.82 points at the larger ratios. These results qualify the main conclusion: the most consistent cross-domain benefits occur under stronger supervision imbalance, while mild imbalance can produce small degradations.

Table 6: Complete source and shifted-domain results for CoOp. Each block varies the nominal imbalance ratio at a fixed training budget. Target Base/New labels denote target-local splits. Bold marks the better member of each pair for each metric. Target average excludes ImageNet; gains are calculated before rounding aggregate values.

Domain	ρ	CoOp			CoOp + RAP			ΔH
		Base	New	H	Base	New	H	
ImageNet	4	75.67	66.87	71.00	**76.00**	**67.67**	**71.59**	+0.59
	16	73.10	65.43	69.05	**75.70**	**67.77**	**71.52**	+2.47
	64	69.07	61.23	64.91	**74.57**	**66.43**	**70.26**	+5.35
ImageNet-V2	4	69.73	59.60	64.27	**70.30**	**60.50**	**65.03**	+0.76
	16	67.67	58.63	62.83	**70.13**	**60.10**	**64.73**	+1.90
	64	64.03	54.10	58.65	**69.17**	**59.33**	**63.87**	+5.22
ImageNet-Sketch	4	47.70	49.67	48.67	**48.13**	**50.80**	**49.43**	+0.76
	16	46.07	47.70	46.87	**47.63**	**49.70**	**48.64**	+1.77
	64	43.10	44.57	43.82	**47.47**	**49.60**	**48.51**	+4.69
ImageNet-A	4	54.10	47.10	50.36	**55.00**	**48.23**	**51.39**	+1.03
	16	53.30	45.27	48.96	**54.90**	**48.00**	**51.22**	+2.26
	64	51.23	43.10	46.82	**55.47**	**46.90**	**50.82**	+4.00
ImageNet-R	4	74.27	78.10	76.14	**75.23**	**79.30**	**77.21**	+1.07
	16	72.83	77.40	75.05	**74.93**	**78.87**	**76.85**	+1.80
	64	70.13	73.73	71.89	**75.40**	**77.33**	**76.35**	+4.46
Target avg.	4	61.45	58.62	59.86	**62.17**	**59.71**	**60.77**	+0.91
	16	59.97	57.25	58.43	**61.90**	**59.17**	**60.36**	+1.93
	64	57.12	53.88	55.30	**61.88**	**58.29**	**59.89**	+4.59

Table 7: Complete source and shifted-domain results for CoCoOp. Each block varies the nominal imbalance ratio at a fixed training budget. Target Base/New labels denote target-local splits. Bold marks the better member of each pair for each metric. Target average excludes ImageNet; gains are calculated before rounding aggregate values.

Domain	ρ	CoCoOp			CoCoOp + RAP			ΔH
		Base	New	H	Base	New	H	
ImageNet	4	75.77	70.07	72.81	**75.87**	**70.17**	**72.91**	+0.10
	16	74.63	69.33	71.88	**75.70**	**70.07**	**72.78**	+0.90
	64	72.30	66.70	69.39	**75.40**	**69.70**	**72.43**	+3.04
ImageNet-V2	4	**70.70**	62.80	**66.52**	70.50	**62.93**	66.50	-0.02
	16	69.33	62.27	65.61	**70.80**	**62.80**	**66.56**	+0.95
	64	67.10	59.60	63.12	**70.03**	**62.70**	**66.16**	+3.04
ImageNet-Sketch	4	49.50	**53.90**	**51.61**	**49.57**	53.47	51.44	-0.17
	16	48.13	52.27	50.11	**49.13**	**53.00**	**51.00**	+0.89
	64	46.07	50.23	48.06	**48.87**	**52.93**	**50.82**	+2.76
ImageNet-A	4	55.60	**49.53**	52.39	**56.17**	49.10	**52.40**	+0.01
	16	55.13	48.33	51.51	**55.97**	**49.00**	**52.25**	+0.74
	64	53.23	47.23	50.05	**55.90**	**50.07**	**52.82**	+2.77
ImageNet-R	4	**76.93**	**80.77**	**78.80**	76.87	80.37	78.58	-0.22
	16	75.73	79.50	77.57	**76.73**	**80.40**	**78.52**	+0.95
	64	73.50	77.47	75.43	**76.20**	**80.20**	**78.14**	+2.71
Target avg.	4	63.18	**61.75**	**62.33**	**63.28**	61.47	62.23	-0.10
	16	62.08	60.59	61.20	**63.16**	**61.30**	**62.08**	+0.88
	64	59.98	58.63	59.17	**62.75**	**61.48**	**61.99**	+2.82

D Complete Regularization Sweep

Table 8 reports the full sweep underlying Figure 6. We use the unsquared Frobenius norm and one coefficient per learner across all datasets. The coefficients used in the main comparison were selected by the six-dataset mean H reported in this sweep; no independent validation set was used for this choice. The selected settings therefore reflect benchmark-level hyperparameter selection.

The coefficient maximizing the six-dataset mean does not maximize every individual dataset. For example, CoOp attains its highest ImageNet H at $\lambda_\Delta = 5$, whereas its highest six-dataset mean occurs at 0.1. CoCoOp similarly reaches its best ImageNet H at 1 but its best mean at 0.001. We retain a single coefficient for each learner rather than choosing a separate value for each dataset.

Table 8: Complete residual-regularization sweep at $\rho = 64$. Dataset columns report H. Average H is recomputed across the six displayed dataset values; average Base/New values are retained from the supplied results. Bold marks the best value within each learner and column. Shading indicates the setting used in the main comparison, selected by the reported six-dataset mean H.

	Dataset H						Average		
λ_Δ	ImageNet	Pets	Food101	DTD	EuroSAT	UCF101	Base	New	H
CoOp + RAP									
10^{-4}	68.79	94.66	**89.55**	58.15	**66.72**	68.00	78.40	70.96	74.31
10^{-3}	69.09	95.23	88.71	57.15	66.32	67.13	78.60	70.19	73.94
10^{-2}	69.59	**95.46**	89.06	58.49	66.10	68.97	79.05	71.01	74.61
10^{-1}	70.26	95.35	88.86	59.03	66.21	71.37	79.52	**71.69**	**75.18**
0.3	70.66	95.40	88.96	57.23	65.84	70.66	80.32	70.65	74.79
1	70.85	95.25	88.85	56.05	66.30	**72.66**	**80.53**	70.91	74.99
3	71.01	94.41	88.93	56.97	63.01	71.27	79.73	70.21	74.27
5	**71.16**	94.58	88.28	**59.05**	64.32	70.19	79.64	70.70	74.60
10	70.93	94.05	87.96	56.22	61.68	67.82	78.82	68.99	73.11
CoCoOp + RAP									
10^{-4}	72.32	96.19	**91.10**	58.34	53.89	73.68	78.84	70.83	74.25
10^{-3}	72.43	**96.49**	90.98	**59.62**	54.68	74.07	**79.03**	71.42	**74.71**
10^{-2}	72.30	96.33	91.06	58.24	51.29	**74.86**	78.97	70.49	74.01
10^{-1}	72.35	95.05	90.96	58.66	52.13	73.32	77.67	70.68	73.75
0.3	72.45	94.89	90.60	59.23	53.65	74.73	76.82	72.16	74.26
1	**72.58**	94.24	89.67	57.56	**57.95**	73.16	76.17	**72.65**	74.19
3	72.42	93.30	88.85	57.73	57.39	72.70	75.77	72.12	73.73
5	72.41	93.94	88.79	57.63	53.40	72.02	75.52	71.00	73.03
10	71.71	93.96	88.77	53.77	56.92	70.62	75.31	70.65	72.63