Anchor-based Robust Finetuning of Vision-Language Models
Supplementary Material

A. Additional Experiments

Additional experiments are conducted to assess the efficacy of our Anchor-based Robust Finetuning (ARF) in regularizing the finetune process using auxiliary image-text-pair anchors with rich semantics. Specifically, we provide quantitative results of baselines and our ARF with weight ensemble in Section A.1. Furthermore, we explore different ways of retrieval in Section A.2 and different utilization of anchors in Section A.3 for domain shift and zero-shot learning scenarios. Our results demonstrate that the introduced anchors consistently exert a positive influence. Additionally, our ARF, employing a simple approach, attains exceptional performance in downstream tasks without compromising OOD generalization.

Methods	ImageNet						
	ID	Im-V2	Im-R	Im-A	Im-Sketch	ObjectNet	Avg. OOD
CLIP	68.3	61.9	77.7	50.0	48.3	54.2	58.4
LP	80.0	70.3	72.4	47.8	48.1	53.4	58.4
FT	82.5	72.8	74.9	48.1	51.9	56.5	60.8
LP-FT	82.1	72.8	75.3	50.1	51.7	56.3	61.2
FLYP	**82.9**	**73.4**	74.2	51.9	51.2	56.1	61.4
ARF	82.8	73.2	**77.2**	**52.5**	**53.1**	**57.0**	**62.6**

Table 1. Domain shift results (%) of state-of-the-art conventional finetuning and robust finetuning approaches with weight ensemble. We employ ImageNet as the finetuning datasets, while the others serve as domain shift evaluation datasets. The numbers represent the top-1 accuracy and the best results are marked in **Black**.

A.1. Quantitative Results of Weight Ensemble

We compare the performance of our ARF to several baselines in a weight ensembling scenario on domain shift benchmarks. Ten coefficients, ranging from 0 to 1, are utilized to interpolate the model weights and we select the mixing coefficient with the highest ID validation accuracy. As illustrated in Table 1, our ARF, combined with weight ensembling, achieves the highest domain shift accuracy of 62.6%, surpassing the previous state-of-the-art method (*i.e.*, FLYP [1]) by 1.2%. These findings demonstrate the effectiveness of our ARF when integrated with weight ensemble following Wise-FT [2].

Method	ImageNet		Zero-Shot
	ID	Domain Shift	Avg. Acc
$ARF_{v\to v}$	82.4	61.0	57.1
$ARF_{t\to t}$	82.6	61.2	55.6
$ARF_{t\to v}$	82.5	61.2	55.7
$ARF_{v\to t}$	82.7	61.3	55.6

Table 2. Ablation study for different ways of retrieval in our ARF. We use the first letter in the subscript to denote the modality in the finetune set, and the second letter in the subscript to represent the modality in the candidate set. For instance, the image-to-text retrieval of our ARF is denoted as $ARF_{v\to t}$.

A.2. Different Ways of Retrieval

In addition to cross-modal retrieval capacity, CLIP can also search for image-text pairs relevant to the downstream task from the candidate set using uni-modal retrieval. Specifically, as depicted in Fig.1, four different combinations of retrieval between images and texts from the downstream dataset and the candidate set are investigated to identify appropriate content for preserving the OOD generalization capabilities of CLIP. We denote the image-to-text retrieval as $ARF_{v\to t}$, which utilizes images from the finetune set to search for the most similar texts in the candidate set.

As presented in Table 2, it can be observed that $ARF_{v\to v}$ achieves the highest accuracy in zero-shot learning but exhibits the lowest performance in ID and domain shift scenarios. This is because uni-modal retrieval, using merely visual information, may search for image-text pairs unrelated to the downstream task while possessing more diverse semantics, thereby improving zero-shot prediction at the expense of ID and domain shift accuracies. Regarding the other three retrieval combinations, they identify appropriate image-text pairs and yield similar results, demonstrating the exceptional retrieval capacity of CLIP. For the highest ID performance, we choose $ARF_{v\to t}$ as our approach.

A.3. Different Utilization of Anchors

Different Utilization of Text-Compensated Anchors. In our ARF, the captions generated by the Text-Compensated Anchor Generation module and class prompts are aligned

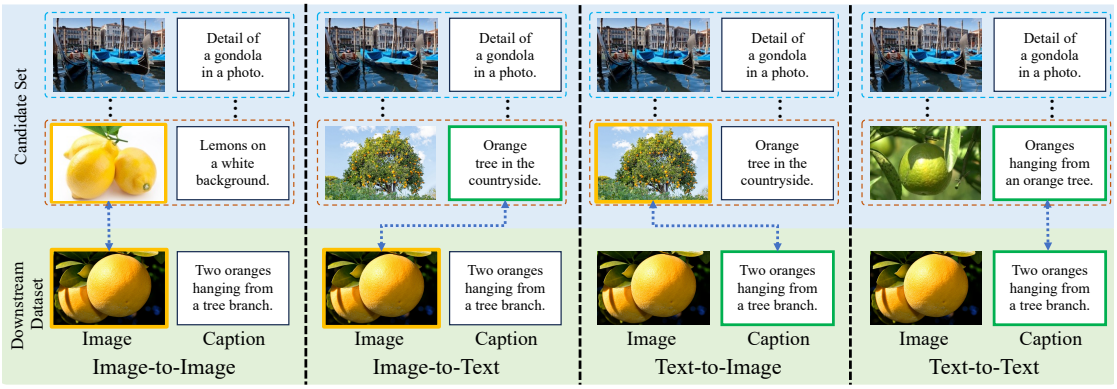

Figure 1. Comparison of various retrieval modality combinations. We search for the most similar image-text pairs in the candidate set to obtain the image-text-pair anchors related to the downstream task for preserving the original feature space of CLIP.

Method	ImageNet ID	ImageNet Domain Shift	Zero-Shot Avg. Acc
baseline	82.6	59.4	48.6
$merge(t_c, t^{\text{cap}})$	82.6	60.5	53.4
$sep(t_c, t^{\text{cap}})$	82.6	60.7	54.3
$merge(t^{\text{cap}}, t^{\text{ret}})$	82.7	61.3	54.3
$sep(t^{\text{cap}}, t^{\text{ret}})$	82.7	61.3	55.6

Table 3. Ablation study for different utilization of anchors in our ARF. The baseline only conducts visual-language contrastive learning for finetuning like FLYP [1].

separately with the images from the finetune set. This strategy effectively avoids overfitting by preventing the image from being drawn excessively close to its corresponding class prompt. We designate this approach as $sep(t_c, t^{\text{cap}})$, where t_c denotes the class prompts and t^{cap} represents the generated captions. Alternatively, we can also merge the class prompt with the caption to enhance the text representation, such as "a photo of a `[CLASS], [CAPTION]`", where "`[CLASS]`" indicates the class name and "`[CAPTION]`" refers to the text description generated by a pretrained captioner for each image. This approach of employing the text-compensated anchor is denoted as $merge(t_c, t^{\text{cap}})$.

As demonstrated in Table 3, we can see that both approaches significantly improve the performance of zero-shot learning and domain shift over the baseline, indicating the effectiveness of our strategy in mitigating overfitting through the use of text-compensated anchors. The separate alignment of captions and class prompts with images denoted as $sep(t_c, t^{\text{cap}})$, achieves better performance compared to $merge(t_c, t^{\text{cap}})$, with improvements of 0.9% and 0.2% in zero-shot learning and domain shift, respectively. Therefore, we employ $sep(t_c, t^{\text{cap}})$, this straightforward utilization of text-compensated anchors, in our ARF.

Different Utilization of Retrieved Image-Text Anchors. We apply two separate contrastive losses (*i.e.*, $\mathcal{L}_{\text{Cap}}$ and $\mathcal{L}_{\text{Ret}}$) to align the text-compensated anchors and the retrieved image-text anchors, denoting this approach as $sep(t^{\text{cap}}, t^{\text{ret}})$. Alternatively, we can merge the two types of anchors in a mini-batch and calculate a single contrastive loss for finetuning the CLIP model, referring to this strategy as $merge(t^{\text{cap}}, t^{\text{ret}})$.

As shown in Table 3, it is evident that $sep(t^{\text{cap}}, t^{\text{ret}})$ significantly improves the performance of zero-shot learning and domain shift compared to the addition of text-compensated anchors. In contrast, $merge(t^{\text{cap}}, t^{\text{ret}})$ merely boosts performance on domain shift. This finding suggests that aligning the retrieved image-text pairs separately using contrastive loss more effectively maintains the original feature space of CLIP while merging the retrieved image-text pairs with the image-caption pairs in a mini-batch mitigates its effectiveness. Consequently, we employ $sep(t^{\text{cap}}, t^{\text{ret}})$ in our ARF.

References

[1] Sachin Goyal, Ananya Kumar, Sankalp Garg, Zico Kolter, and Aditi Raghunathan. Finetune like you pretrain: Improved finetuning of zero-shot vision models. In *CVPR*, 2023. 1, 2

[2] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *CVPR*, 2022. 1