Supplementary materials of Post-pre-training for Modality Alignment in Vision-Language Foundation Models

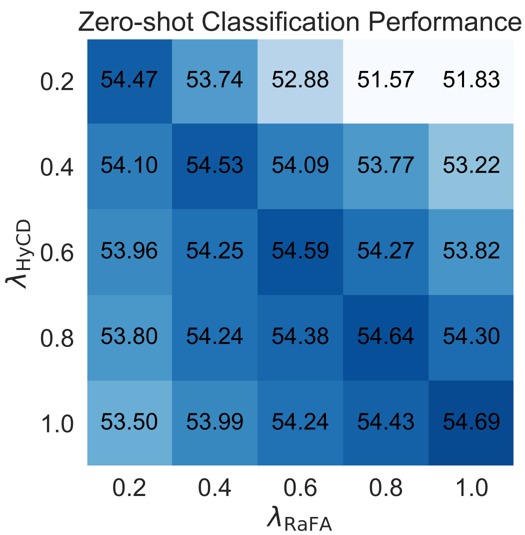

Figure I. Zero-shot classification accuracy averaged on 12 datasets when varying balancing parameters between $\mathcal{L}_{\text{RaFA}}$ and $\mathcal{L}_{\text{HyCD}}$ (ViT-B/32).

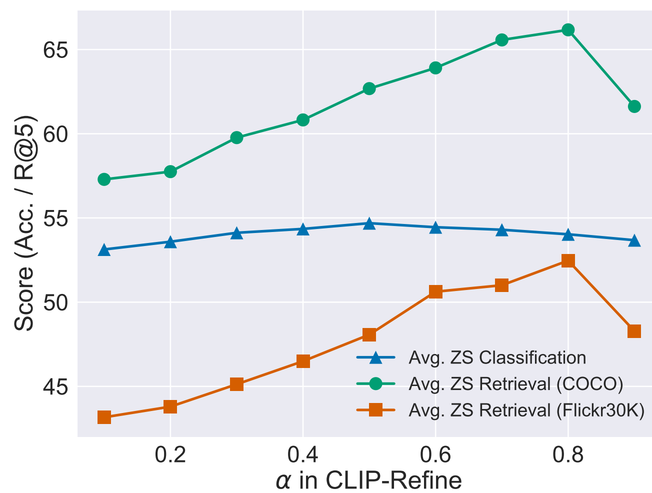

Figure II. Zero-shot performance on 12 classification datasets and retrieval datasets when varying α in $\mathcal{L}_{\text{HyCD}}$ (ViT-B/32).

A. Effects of Hyperparameters

In the main paper, we fixed the contributions of $\mathcal{L}_{\text{RaFA}}$, $\mathcal{L}_{\text{HyCD}}$ in CLIP-Refine and the hyperparameter of α in Eq. (1) for HyCD, and epochs for post-pre-training. Here, we confirm the effects of varying them on the performance.

Trade-off between $\mathcal{L}_{\text{RaFA}}$ and $\mathcal{L}_{\text{HyCD}}$ We evaluate balancing $\mathcal{L}_{\text{RaFA}}$ and $\mathcal{L}_{\text{HyCD}}$ in Eq (1) by introducing hyperparameters λ_{RaFA} and λ_{HyCD} as follows:

$$\min_{\theta_{\text{V}},\theta_{\text{T}}} \lambda_{\text{RaFA}}\mathcal{L}_{\text{RaFA}}(\theta_{\text{V}},\theta_{\text{T}}) + \lambda_{\text{HyCD}}\mathcal{L}_{\text{HyCD}}(\theta_{\text{V}},\theta_{\text{T}}).$$

We varied λ_{RaFA} and λ_{HyCD} in $\{0.2, 0.4, 0.6, 0.8, 1.0\}$ and post-pre-trained CLIP ViT-B/32 on COCO Caption. Figure I illustrates the heatmap where each cell represents the zero-shot classification accuracy averaged on 12 datasets. We can see that the diagonal elements of the heatmap achieve higher performance, indicating that keeping the equal contribution of λ_{RaFA} and λ_{HyCD} is important for better zero-shot performance.

Trade-off parameter α in $\mathcal{L}_{\text{HyCD}}$ We evaluate the trade-off parameter α in Eq.(8) for balancing learning of the new knowledge from post-pre-training and retaining of the past knowledge in the pre-trained CLIP models. We varied α in $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$. Figure II shows the trend of the averaged zero-shot classification and retrieval accuracy. We see that the trends in classification and retrieval are different; the classification performance is less sensitive

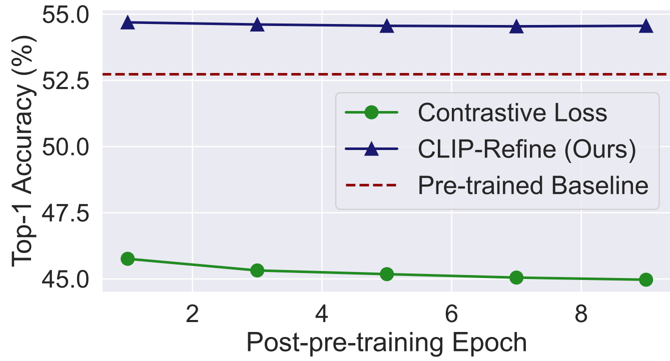

Figure III. Zero-shot classification accuracy averaged on 12 datasets when varying epochs in post-pre-training.

Table I. Robustness Evaluation on Zero-shot Classification.

Method	IN1K	V2	A	R	Sketch
Pre-trained	59.04	51.80	28.84	64.81	38.38
Contrastive	37.04	45.52	22.92	62.80	35.57
m^2-mix	59.06	46.32	22.51	63.42	35.59
Self-KD	51.88	52.01	28.65	65.08	38.52
HyCD+$\mathcal{L}_{\text{Align}}$	57.06	45.41	21.45	62.00	34.73
CLIP-Refine (Ours)	**60.92**	**53.51**	**30.68**	**67.05**	**41.46**

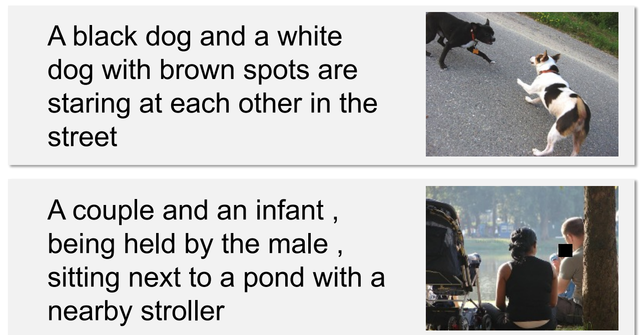

Figure IV. Correctly retrieved samples

than the retrieval performance, and an overly high value of α degrades both performances. This suggests that prioritizing new knowledge is important but balancing the new and past knowledge is crucial to achieve the best performance.

Post-pre-training Epochs We show the effect of increasing post-pre-training epochs from one, which is used in the main paper. Figure III shows the averaged zero-shot classification accuracy when varying the post-pre-training epoch in $\{1, 3, 5, 7, 9\}$. CLIP-Refine stably kept performance even when increasing epochs, while the contrastive loss slightly degraded the performance according to the epochs. This implies that our CLIP-Refine can provide stable performance improvements by avoiding catastrophic forgetting even in longer epochs. This also means that our CLIP-Refine has the practical advantage of not having to search for the appropriate epoch length in each case.

B. Additional Experiments

B.1. Robustness Evaluation

Here, we evaluate the robustness of our method through the evaluation on ImageNet variants including ImageNet-V2 [3], ImageNet-A [2], ImageNet-R [1], and ImageNet-Sketch [4]. Table I that our method robustly performs on these variants, supporting the general performance improvements of our method.

B.2. Visualization Study

We randomly selected samples of Flickr30K from which CLIP failed, but CLIP-Refine succeeded (Fig. IV). We see that CLIP-Refine can match complex text and image pairs with multiple attributes and object combinations. This highlights that the multi-modal alignment is enhanced by reducing the modality gap.

References

[1] 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. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 2

[2] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 2

[3] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In International conference on machine learning, 2019. 2

[4] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In Advances in Neural Information Processing Systems, pages 10506–10518, 2019. 2