One Last Attention for Your Vision-Language Model Supplementary Material

Liang Chen[*†]
MBZUAI
liangchen527@gmail.com

Ghazi Shazan Ahmad[*]
MBZUAI
ghazi.ahmad@mbzuai.ac.ae

Tianjun Yao
MBZUAI
tianjun.yao@mbzuai.ac.ae

Lingqiao Liu
The University of Adelaide
lingqiao.liu@adelaide.edu.au

Zhiqiang Shen[‡]
MBZUAI
zhiqiang.shen@mbzuai.ac.ae

In this supplementary material, we provide,

1. Theoretical support for using fused information in Sec. 1.
2. Visualizations for RAda in Sec. 2.
3. Detailed settings of RAda in the FFT setting in Sec. 3.
4. Extending other fine-tuning ideas in FFT in Sec. 4.
5. Ablation studies on $\mathcal{L}_{reg}$ in Sec. 5.
6. Experiments with different backbones in Sec. 6.
7. Experiments with different VLMs in Sec. 7.
8. Experiments for using more attention layers in Sec. 8.

1. Theoretical Support for Utilizing Fused Information Over Isolated Representation

Our empirical observations indicate that adapting the rational matrix yields better performance than adapting the different modalities in isolation. In this section, we provide theoretical explanations for justification the selection of utilizing the final fused information rather than the isolated representations in fine-tuning. In particular, we demonstrate the advantage of RAda against three fine-tuning ideas (*i.e.* separately adapt the image features or text features in isolation or jointly adapt both features). The explanation framework is grounded in information theory and statistical sufficiency. Below is a step-by-step proofs.

Given the random sampled class label $\mathbf{Y}$, image embedding $\mathbf{f} \in \mathbb{R}^D$, text embedding $\mathbf{h} \in \mathbb{R}^{K\times D}$, rational matrix $\mathbf{R} \in \mathbb{R}^{K\times D}$ (each element defined as $\mathbf{R}_{i,j} = \mathbf{f}_j \cdot \mathbf{h}_{i,j}$ [1]), we present the adaptation of image or text embedding as learning task-specific transformations: $\mathbf{f} \mapsto \mathbf{f}\circ\mathbf{M}_\mathbf{f}$, $\mathbf{h} \mapsto \mathbf{h}\circ\mathbf{M}_\mathbf{h}$, and similar for the rational matrix: $\mathbf{R} \mapsto \mathbf{R} \circ \mathbf{M}$, with $\circ$ element-wise product and $\mathbf{M}_\mathbf{f} \in \mathbb{R}^D, \mathbf{M}_\mathbf{h} \in \mathbb{R}^{K\times D}$, and $\mathbf{M} \in \mathbb{R}^{K\times D}$ being learnable parameters. We first have,

Lemma 1.1. *The rational matrix $\mathbf{R}$ is a sufficient statistic of $\mathbf{Y}$. Formally, by the definition,*

$$p(\mathbf{Y}|\mathbf{f},\mathbf{h}) = p(\mathbf{Y}|\mathbf{R}), \tag{1}$$

where $p(\mathbf{Y}|\mathbf{f},\mathbf{h})$ denotes that the prediction in CLIP relies on both the image and the text embeddings.

Proof. In the CLIP model, the prediction rule depends only on the inner products $\sum_j \mathbf{R}_{i,j}$, which are functions of $\mathbf{R}$. Thus, the likelihood $p(\mathbf{Y}|\mathbf{f},\mathbf{h})$ depends on $\mathbf{f}$ and $\mathbf{h}$ only through $\mathbf{R}$. Therefore, $\mathbf{R}$ is a sufficient statistic for $\mathbf{Y}$. □

Given $\mathbf{R}$ is sufficient for $\mathbf{Y}$, we thus have equality between mutual informations: $I(\mathbf{Y};\mathbf{R}) = I(\mathbf{Y};\mathbf{f},\mathbf{h})$.

Lemma 1.2. *Adapting $\mathbf{R}$ achieves mutual information with $\mathbf{Y}$ no less than adapting $\mathbf{f}$ or $\mathbf{h}$. In particular,*

$$I(\mathbf{Y};\mathbf{R}\circ\mathbf{M}) \geq \max\{I(\mathbf{Y};\mathbf{f}\circ\mathbf{M}_\mathbf{f},\mathbf{h}), I(\mathbf{Y};\mathbf{h}\circ\mathbf{M}_\mathbf{h},\mathbf{f})\}. \tag{2}$$

Proof. Revisiting the first term in RHS of Eq. (2), we can represent it as: $I(\mathbf{Y};\mathbf{f}\circ\mathbf{M}_\mathbf{f},\mathbf{h}) = I(\mathbf{Y};\mathbf{R}\circ(\mathbf{M}_\mathbf{f}\otimes\mathbf{1}_K^\intercal))$, with $\otimes$ denotes the Kronecker product, and $\mathbf{M}_\mathbf{f}\otimes\mathbf{1}_K^\intercal$ refers replicate $\mathbf{M}_\mathbf{f}$ across rows. By the data processing inequality (DPI) [2], we have,

$$I(\mathbf{Y};\mathbf{R}\circ\mathbf{M}) \geq I\left(\mathbf{Y};\mathbf{R}\circ(\mathbf{M}_\mathbf{f}\otimes\mathbf{1}_K^\intercal)\right), \tag{3}$$

since $\mathbf{R}\circ(\mathbf{M}_\mathbf{f}\otimes\mathbf{1}_K^\intercal)$ can be regarded as a deterministic function of $\mathbf{R}\circ\mathbf{M}$ (by constraining $\mathbf{M}$ to be column-wise). If we constrain $\mathbf{M} = \mathbf{M}_\mathbf{f}\otimes\mathbf{1}_K^\intercal$, where the task-relevant information in $\mathbf{R}$ is uniformly distributed across rows within each column, then: $I(\mathbf{Y};\mathbf{R}\circ\mathbf{M}) = I\left(\mathbf{Y};\mathbf{R}\circ(\mathbf{M}_\mathbf{f}\otimes\mathbf{1}_K^\intercal)\right)$.

The same goes for adapting $\mathbf{h}$ with $\mathbf{M}_\mathbf{h}$, with $I(\mathbf{Y};\mathbf{h}\circ\mathbf{M}_\mathbf{h},\mathbf{f}) = I(\mathbf{Y};\mathbf{R}\circ\mathbf{M}_\mathbf{h})$, due to the uniform natural of text embeddings across all samples, $\mathbf{R}\circ\mathbf{M}_\mathbf{h}$ can be regarded as

[*]Equal technical contribution. [†]Project lead. [‡]Correspondence.
[1]The two embeddings are both normalized in this section.

a deterministic function of $\mathbf{R} \circ \mathbf{M}$ (by constraining $\mathbf{M}$ to be sample-wise). We thus have,

$$I(\mathbf{Y}; \mathbf{R} \circ \mathbf{M}) \geq I(\mathbf{Y}; \mathbf{h} \circ \mathbf{M}_\mathbf{h}, \mathbf{f}). \tag{4}$$

In all, equality in Eq. (2) holds for these constrained cases. Otherwise, we will have LHS larger than RHS in Eq. (2). □

Lemma 1.3. *Adapting* $\mathbf{R}$ *achieves higher mutual information with* $\mathbf{Y}$ *than adapting* $\mathbf{h}$ *and* $\mathbf{f}$ *jointly. In particular,*

$$I(\mathbf{Y}; \mathbf{R} \circ \mathbf{M}) \geq I(\mathbf{Y}; \mathbf{h} \circ \mathbf{M}_\mathbf{h}, \mathbf{f} \circ \mathbf{M}_\mathbf{f}). \tag{5}$$

Proof. By the DPI, we have,

$$I(\mathbf{Y}; \mathbf{h}, \mathbf{f} \circ \mathbf{M}_\mathbf{f}) \geq I(\mathbf{Y}; \mathbf{h} \circ \mathbf{M}_\mathbf{h}, \mathbf{f} \circ \mathbf{M}_\mathbf{f}), \tag{6}$$

where the equality holds when $\mathbf{M}_\mathbf{h}$ is invertible. Combining with Lemma 1.2, we thus can complete the proof. □

Lemma 1.2 and 1.3 demonstrate that adapting the rational matrix results in mutual information no less than adapting $\mathbf{f}$, $\mathbf{h}$, or both $\mathbf{f}$ and $\mathbf{h}$. According to the information bottleneck principle [12], a higher mutual information between the label and an intermediate representation generally correlates with better predictive performance. Given the compression performance (*i.e.* generalizibility) of the model can be largely preserved via an all-one regularization for $\mathbf{M}$ in our implementation, it is not surprise that leveraging the fused representation can be more beneficial than utilizing the individual modalities in isolation. These observations align with our empirical observations in the ablation studies, and they further justify our motivation for adapting the rational matrix to achieve effective fine-tuning.

2. Visualization

We present 2D t-SNE plots [13] of $\mathbf{M} \circ \mathbf{R}$ (corresponding to RAda) and $\mathbf{R}$ (representing the original CLIP) to illustrate the behavior of the learned mask $\mathbf{M}$. As shown in Figure 1, the adapted $\mathbf{R}$ in the EuroSAT dataset exhibits enhanced differentiability, with tighter clusters indicating improved class separability compared to that of CLIP. Additionally, the plot of $\mathbf{M} \circ \mathbf{R}$ for OxfordPets reveals 19 clusters, matching the total class count in the dataset, while the original $\mathbf{R}$ from CLIP shows only 18 classes. This distinction aligns with the enhanced classification performance achieved through $\mathbf{M}$. These findings validate the effectiveness of adapting the decision-making process in achieving improved predictions within a VLM.

We also present the distribution of the values in the mask $\mathbf{M}$ for the two datasets. The plots in Figure 2 show that, for both datasets, the mask values exhibit a mean value of approximately 1.0, with the majority of the weights centered around 1 to form a normal distribution, and the mask value

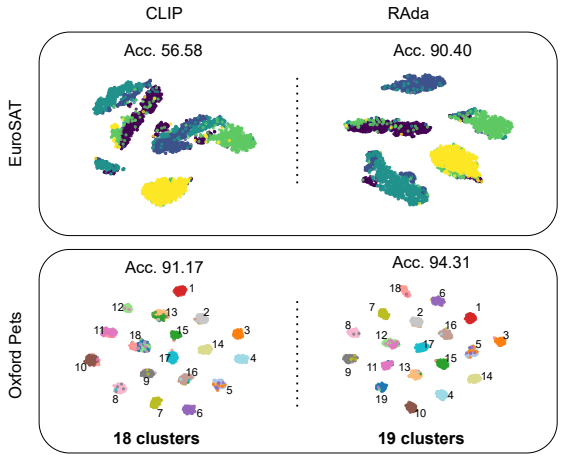


Figure 1. T-SNE [13] plots of the Rational Matrix from CLIP and RAda in the EuroSat (5 classes) and OXfordPets (19 classes) datasets. The adapted rational matrix in RAda shows clearer and more precise separation than that in the original CLIP.

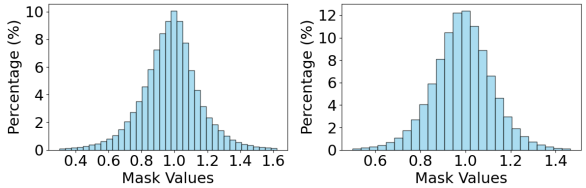
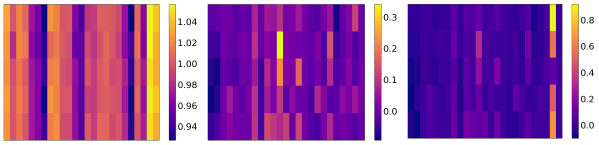


Figure 2. Distributions of values in the learned mask $\mathbf{M}$ for EuroSAT (*i.e.* left figure) and Oxford Pets (*i.e.* right figure) datasets.



(a) $\mathbf{M}$ (b) $\mathbf{R}$ (c) $\mathbf{M} \circ \mathbf{R}$

Figure 3. Heatmaps of the rational adaptation process.

can reach as high as 3, indicating the varying contributions of different rational elements after adaptation. We hope this finding can inspire future research to develop more effective learning objective for adapting the rational matrix. We present an example of the adaptation process to better illustrate how the original decision matrix got shifted by the process. We show heatmaps of the mask $\mathbf{M}$, the rational matries $\mathbf{R}$, and $\mathbf{M} \circ \mathbf{R}$ in Figure 3 (left to right), where $\mathbf{M} \circ \mathbf{R}$ shows more evident classification clue than $\mathbf{R}$ with larger values in the 2nd last column, suggesting the rational adaptation process helps the model to capture more confident outputs.

3. Detailed Settings in FFT

Our implementation in the FFT setting consists of two consecutive parts, first updating the rational adapter (*i.e.* RAda) and then updating all learnable parameters (*i.e.* RAda-FT).

Table 1. Extending CLIP-Adapter [4] in the FFT setting. Results with † are reevaluated in our device, others are from FLYP [5].

Methods	Training in Imagenet						
	ID	Im-V2	Im-R	Im-A	Im-S	Object	OOD Avg.
LP	79.9	69.8	70.8	46.4	46.9	52.1	57.2
FT	81.3	70.9	65.6	36.7	46.3	49.6	53.8
L2-SP	81.7	71.8	70.0	42.5	48.5	56.2	57.8
LP-FT	81.7	72.1	73.5	47.6	50.3	58.2	60.3
FLYP	82.6	73.0	71.4	48.1	49.6	58.7	60.2
CLIP†	68.3	61.9	**77.7**	49.9	48.2	54.2	58.4
FLYP†	**82.6**	**72.6**	71.8	48.5	49.8	54.6	59.5
Adapter†	81.5	71.7	74.3	50.3	50.1	55.3	60.3
RAda-FT	81.4	71.9	75.5	**51.7**	**50.4**	**56.8**	**61.3**

This section provides more details regarding the objectives and hyper-parameter settings for the two parts.

First, for RAda, the objective is,

$$\arg\min_{\theta_m} \|\mathbf{M}-\mathbf{1}_{K\times D}\|^2 - \log\frac{\exp(<\mathbf{1}_D,(\mathbf{M}\circ\mathbf{R}')_*>)}{\sum_{i=1}^{K}\exp(<\mathbf{1}_D,(\mathbf{M}\circ\mathbf{R}')_i>)},$$
$$\text{s.t. } \mathbf{R}'^{\top} = \begin{bmatrix} \bar{\mathbf{f}}_1\mathbf{W}_{\{1,1\}} & \bar{\mathbf{f}}_1\mathbf{W}_{\{1,2\}} & \dots & \bar{\mathbf{f}}_1\mathbf{W}_{\{1,K\}} \\ \bar{\mathbf{f}}_2\mathbf{W}_{\{2,1\}} & \bar{\mathbf{f}}_2\mathbf{W}_{\{2,2\}} & \dots & \bar{\mathbf{f}}_2\mathbf{W}_{\{2,K\}} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{\mathbf{f}}_D\mathbf{W}_{\{D,1\}} & \bar{\mathbf{f}}_D\mathbf{W}_{\{D,2\}} & \dots & \bar{\mathbf{f}}_D\mathbf{W}_{\{D,K\}} \end{bmatrix}, \tag{7}$$

where the first term is the smooth regularization for the mask, and the second term is the main classification loss. Since the text encoder is replaced with a linear classifier, we use the weight of the classifier (*i.e.* $\mathbf{W}\in\mathbb{R}^{K\times D}$, which is initialized by the text feature $\bar{\mathbf{h}}$) to compute the corresponding rational matrix $\mathbf{R}'$. We train it for 10 epochs with the learning rate of 0.004 and batch size of 512. Default settings from [14] are adopted, where the AdamW optimizer [10] is utilized; weight decay is set to be 0.1; the same warmup learning strategy is also utilized.

Second, for RAda-FT, the objective is,

$$\arg\min_{\{\theta'_m,\theta_v,\mathbf{W}\}} -\log\frac{\exp(<\mathbf{1}_D,(\mathbf{M}\circ\mathbf{R}')_*>)}{\sum_{i=1}^{K}\exp(<\mathbf{1}_D,(\mathbf{M}\circ\mathbf{R}')_i>)}, \tag{8}$$

where θ'_m is the rational adapter trained after 5 epochs with the objective in Eq. (7). We train RAda-FT with Eq. (8) for 5 epochs using the same settings as RAda, except for the learning rate, which is set as 0.000004.

4. Extending Other Fine-Tuning Ideas in FFT

Our experiments demonstrate that RAda can be seamlessly integrated into the FFT setting by building on the existing practice [9], effectively enhancing the baseline. This exploration has received limited attention in other alike fine-tuning approaches. To provide a comprehensive evaluation of our method, this section investigates extending the same concept to other fine-tuning approaches within the FFT framework. Note that not all fine-tuning approaches are suitable for the FFT setting. For instance, CoOp [17] relies on the presence of a text encoder, which will be replaced by a linear classifier in FFT. Meanwhile, given additional inserted prompts will require large computational resources when updating all parameters, this section will focus exclusively on experiments extending an adapter-based fine-tuning approach that operates within the FFT framework without requiring a text encoder. Specifically, we extend CLIP-Adapter [4] by applying the same training strategy as LP-FT [9]. In this extension, we first train the feature adapter and then use the weights obtained after 5 epochs as initialization to fine-tune all learnable parameters. To ensure fair comparisons, we adopt the same settings as RAda for CLIP-Adapter, except for the learning rate, which is adjusted by factors of $\times 0.1, \times 1, \times 10$ relative to the original values in our implementation. The learning rate yielding the best performance on the evaluation sets in the ID data is selected for reporting results.

We list the experimental results in Table 1. We observe that CLIP-Adapter can improve the OOD performance for FT when with the same sequential updating strategy, and it performs better than LP-FT in 3 out of the 5 OOD datasets evaluated. This is mainly because CLIP-Adapter can preserve the text features and part of the visual information, which is helpful for generalization [9], as opposed to LP-FT, where the text information will be compromised for adaptation in the training data. However, as CLIP-Adapter cannot leverage fused representations from multiple modalities and inevitably distorts certain pretrained visual cues, it is inferior to RAda across all OOD datasets, with an average performance gap of 1pp. These results validate the effectiveness of our design of focusing on the final decision-making process during the fine-tuning process.

5. Weight and Loss Format Ablation for Mask Regularization

To assess the sensitivity of RAda to variations in the regularization term $\mathcal{L}_{reg}$, we analyze its performance across different values of the hyperparameter α in the regularization loss, formulated as $\mathcal{L}_{adapt} + \alpha * \mathcal{L}_{reg}$. The results, shown in Figure 4, indicate that RAda remains robust to different values of α as long as it is within a reasonable range (*i.e.* from 0.5 to 2.5), indicating that we do not need to specifically tune this hyper-parameter (the default $\alpha = 1$ suffices). In practice, for all experiments in the EFT setting, it is fixed as 1.5 as it yields relatively better results.

Moreover, since there are different alternatives for the loss terms, we also conduct experiments to analyze if the adopted L_2 regularization is the optimal choice by compring it with L_1 and L_∞. Results in Table 2 indicate that using the adopted L2 norm leads to better results than other alternatives.

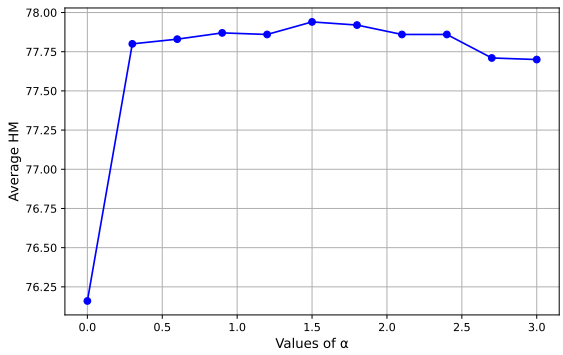


Figure 4. Evaluations of RAda with different weights (*i.e.* α) for the mask regularization term. Results are averaged over 11 datasets. RAda is insensitive regarding the different values of α as long as it is within a reasonable range (*i.e.* from 0.5 to 2.5).

Table 2. Ablation on loss format for the mask regularization term. Results are averaged over 11 datasets.

loss format	Base	Novel	**HM**
L_∞	80.06	73.87	76.84
L_1	81.50	73.3	77.21
L_2	82.16	74.14	77.94

6. Experiments with Different Backbones

Our default setting in the manuscript employs the ViT-B/16 backbone for the CLIP image encoder. In this section, we investigate whether RAda can maintain its effectiveness with alternative backbones. Specifically, we evaluate RAda with another transformer-based image encoder (*i.e.* ViT-B/32 [3]) and two ResNet-based variants of CLIP (*i.e.* ResNet-50 and ResNet-101 [6]). As presented in Table 3, RAda demonstrates consistent improvements over the baseline across different backbones, achieving significant enhancements in base accuracy while exhibiting slight reductions in novel class performance. In comparison with the established art CoOp [17], RAda consistently outperforms it across all backbones in terms of harmonic mean between the base and novel accuracies, with particularly notable advantages in the unseen novel classes. These findings affirm the robustness and adaptability of RAda across varying backbone architectures.

7. Experiments with Different VLMs

Following existing arts [4, 16], we conduct experiments only with CLIP in our manuscript. But note that the rational matrix [1] is applicable not only in CLIP's similarity-based structure, but any cases when there is contrastive (CT) or softmax losses (SM) , as it represents the inner product's intermediate state when computing these losses. This extends RAda also in other VLMs, such as ALIGN or SigLIP where CL and SM are involved. To validate the effectiveness of RAda also in these different VLMs, we conduct experiments in the EFT setting and present the results in Table 4. As seen, RAda is consistently effective even with different VLMs, indicating the broader applicability of RAda.

Table 3. Evaluations of RAda with different backbones for the CLIP image encoder. Results are averaged over 11 datasets.

Backbone	Base Acc.	Novel Acc.	Hamonic Mean
ResNet-50			
CLIP	65.29	69.01	67.09
CoOp	76.56	63.31	69.31
RAda	76.80	67.30	71.74
ResNet-101			
CLIP	64.53	69.82	67.07
CoOp	78.31	63.80	70.31
RAda	78.42	68.27	72.99
ViT-B/32			
CLIP	67.21	71.65	69.36
CoOp	78.55	66.08	71.78
RAda	76.31	70.45	73.26
ViT-B/16			
CLIP	69.34	74.22	71.70
CoOp	82.69	63.22	71.66
RAda	82.16	74.14	77.94

Table 4. Applying RAda in different VLMs. Results are averaged over 11 datasets.

VLM	Base Acc.	Novel Acc.	Harmonic Mean
CLIP [11]			
Zero Shot	69.34	74.22	71.70
RAda	78.42	68.27	72.99
OpenCLIP [7]			
Zero Shot	67.61	71.08	69.30
RAda	78.00	73.31	75.58
SigLIP [15]			
Zero Shot	78.28	73.75	75.95
RAda	84.13	74.98	79.29
ALIGN [8]			
Zero Shot	70.00	66.66	68.29
RAda	75.60	69.79	72.58

8. More Attention Layers at the End

RAda attaches one additional attention layer at the end of CLIP for adaptation. To evaluate if more layers can better help the performance, we compare the original implementation with variants that use different attention layers. For the consecutive layers, we use the combined mask output from all previous layers as the query, and the rational matrix is still served as key and values for the new layers. The residual connection is utilized for the multi-layer implementation, where the final mask is the combination result of the

Table 5. Evaluations of RAda with different attention layers attached at the end. Results are averaged over 11 datasets.

	Base Acc.	Novel Acc.	Hamonic Mean
Baseline	69.34	74.22	71.70
1 layer	82.16	74.14	77.94
2 layer	82.51	73.84	77.93
3 layer	82.33	71.81	76.71

masks obtained from all previous and current steps:

$$\begin{aligned} \mathbf{M}^n &= \mathbf{M}_0 + \mathbf{M}_1 + ... + \mathbf{M}_n, \\ \text{s.t. } \mathbf{M}_n &= \mathcal{F}_m^n(\{\mathbf{M}^{n-1}, \mathbf{R}\}, \theta_m^n), \end{aligned} \tag{9}$$

where $\mathbf{M}^n$ denotes using n layers to obtain the final mask, $\mathbf{M}_n$ is the mask from the n-th layer, θ_m^n is the parameter for the n-th attention layer, and $\{\mathbf{M}^{n-1}, \mathbf{R}\}$ denotes the query is from $\mathbf{M}^{n-1}$, key and value are from $\mathbf{R}$ in the attention layer. Note $\mathbf{M}_0$ is implemented with $\mathbf{M}_0 = \mathcal{F}_m^0(\{\overline{\mathbf{h}}, \bar{\mathbf{f}}, \mathbf{R}\}, \theta_m^0)$ (*i.e.* Eq. (6) in the manuscript) given there is no previous mask information. Similarly, we report average results across 11 datasets from the base-to-new experiments of the EFT setting. As shown in Table 5, using more attention layers can improve the base performance but decrease the zero-shot ability of CLIP, indicating a trade-off between adapting to seen categories and maintaining the model's capacity to generalize to unseen categories. Since using only one layer for RAda can obtain similar results as that with two layers, we thus attach only one attention layer for simplicity in our implementation.

References

[1] Liang Chen, Yong Zhang, Yibing Song, Anton Van Den Hengel, and Lingqiao Liu. Domain generalization via rationale invariance. In *ICCV*, 2023. 4

[2] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999. 1

[3] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 4

[4] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *IJCV*, 132(2):581–595, 2024. 3, 4

[5] 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. 3

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[7] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 4

[8] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 4

[9] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *ICLR*, 2022. 3

[10] I Loshchilov. Decoupled weight decay regularization. In *ICLR*, 2019. 3

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 4

[12] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000. 2

[13] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(11), 2008. 2

[14] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *CVPR*, 2022. 3

[15] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. 4

[16] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022. 4

[17] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022. 3, 4