

One Last Attention for Your Vision-Language Model

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

Abstract

Pretrained vision-language models (VLMs), such as CLIP, achieve remarkable zero-shot performance, yet their downstream potential hinges on effective fine-tuning. Most adaptation methods typically focus on refining representation from separate modalities (text or vision) but neglect the critical role of their fused representations in the decision-making process, i.e. rational matrix that drives the final prediction [5]. To bridge the gap, we propose a simple yet effective **Rational Adaptaion (RAda)** to explicitly exploit the final fused representation during fine-tuning. RAda employs a learned mask, obtained from a lightweight attention layer attached at the end of a VLM, to dynamically calibrate the contribution of each element in the rational matrix, enabling targeted adjustments to the final cross-modal interactions without incurring costly modifications to intermediate features. Experiments in different settings (i.e. updating, or freezing pretrained encoders in adaptation, and test-time training that can only access the unlabeled test data) show that RAda serves as a versatile fine-tuning technique, improving the baseline with minimal code and performing comparably against current arts in most settings. Code is available at github.com/khufia/RAda.

1. Introduction

Recent foundation models trained on multiple modalities (e.g. vision, language, audio) have demonstrated exceptional generalization across diverse tasks. Among these, vision-language models (VLMs) like CLIP [30] and ALIGN [19], pretrained on large-scale image-text pairs, achieve remarkable zero-shot classification by aligning input images with text prompts. This alignment is measured by the similarity between image and text represen-

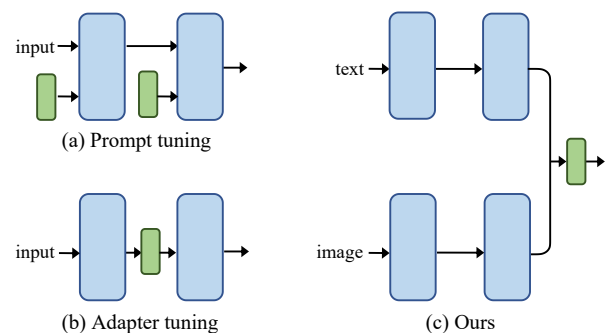


Figure 1. Comparisons between fine-tuning ideas for VLMs that incorporate new parameters. Blue and green blocks denote fixed encoders and learnable parameters. Unlike previous arts that focus on intermediate features from separate modalities, we aim at the fused representations in the final decision-making process, adjusting the corresponding rational matrix [5] to achieve adaptation.

tations, with the closest match determining the prediction. Thanks to the rich supervision provided by the diverse data pairs, such “zero-shot” classifiers can reason about open-vocabulary visual concepts and obtain impressive robustness to many distribution shifts. Nevertheless, in many occasions, it is still beneficial that pretrained VLMs can be adapted to the given data distribution through fine-tuning.

There are several different fine-tuning strategies applicable for VLMs when the source data is available. Besides fine-tuning within the standard practice of transfer learning, which updates all parameters [15, 23] or only the classification head, some recent methods suggest altering the original text or visual representations by tuning newly introduced parameters. Ideas are achieved by either incorporating learnable prompts as an addition to alter the original input embeddings [20, 33, 48, 54, 55] (i.e. prompt tuning, as seen in Figure 1 (a)), or inserting lightweight learnable parameters within the encoders to modify their outputs [8, 12, 14, 26, 36, 51] (i.e. adapter tuning, as seen in Figure 1 (b)). Despite their distinct settings, most methods treat the text or visual features in isolation, overlook-

^{*}Equal technical contribution. [†]Project lead. [‡]Correspondence.

ing a fundamental aspect of the VLM decision-making: the two features are not independently meaningful, as the final prediction emerges from the interaction of both modalities. This motivates the needs to explicitly emphasize the fused representation in fine-tuning.

A recent art [21] attempts to address this issue by fostering mutual synergy between the two modalities on the basis of combining vision [20] and text [55] prompt tunings. While the art shows leading performance in specific scenarios, it incurs relatively large computational overhead due to its requirement for both encoders to actively participate in the intermediate prompt updating steps. More critically, the reliance on both encoders will restrict its applicability when confronting non-ViT-based [11] architectures or in a standard transfer learning paradigm [23], diminishing its practical utility in diverse fine-tuning settings.

To address these challenges, we suggest leveraging fused information from the final decision-making process to achieve lightweight and encoder-agnostic adaptation. Unfortunately, in conventional VLMs like CLIP, fused representation at the final stage will be buried within the similarity calculation, rendering it difficult to be accessed directly. To explicitly surface the buried information, we extend the concept of rational matrix [5], which is originally developed for the classical classifying system, to VLMs. In its classical form, the rational matrix is defined as the entry-wise product between the feature vector and classifier weights, aiming to describe associations between each feature element and classifier weights during prediction. Extended to VLMs, this concept depicts the fine-grained interaction between visual and textual representations, where the text encoder’s output (acting as classifier weights) interacts with the image encoder’s output (the feature vector) to form a fused representation that governs the final prediction, making it inherently suited for capturing fused information at the VLM decision-making stage. Through a rational matrix adaptation (see Figure 1 (c)), we can thus learn complementary information from a cross-modal perspective, resulting in a more holistic understanding of the data compared to ideas that consider the different modalities in isolation (*empirical and theoretical supports are provided in Sec. 5.1 and the supplementary material, respectively*).

We adopt a streamlined implementation for RAda. Specifically, we propose using a mask, obtained via a lightweight attention layer at the end, to dynamically adjust the contribution of each rational elements for achieving adaptation¹. In our design, a multi-query setting is employed for the attention layer to fully exploit all available information, where multiple queries (*i.e.*, image features, text features, and the rational matrix) are applied to the same key and value (*i.e.* the rational matrix), with outputs averaged across all queries. The initial prediction can then be adapted

¹In the original CLIP, contributions are all 1 for different elements.

by taking the entry-wise product between the original rational matrix and the learned mask. Our method is easy to implement, requiring minimal code upon the baseline CLIP (see Algorithm 1). Unlike [21], RAda does not modify intermediate features or require encoder participation, which not only reduces computational overhead but also ensures broader applicability in various fine-tuning settings (detailed comparisons with [21] are provided in Sec. 5.2).

To comprehensively evaluate RAda, besides the settings with source data (*i.e.* (1) full fine-tuning (FFT) that updates all parameters; (2) efficient fine-tuning (EFT) that updates only the rational adapter with frozen encoders), we also evaluate it in a setting when source data is unavailable (*i.e.* (3) test-time training (TTT) that tunes the rational adapter with unlabeled test data). We observe that RAda exhibits consistent improvements over the baseline in all three settings and perform favorably against existing arts in most scenarios, demonstrating its versatility and effectiveness in the VLM fine-tuning task. Meanwhile, we also show that existing fine-tuning ideas and RAda are not mutually exclusive, rather, their integration shows further improvements.

Main contributions of this work are three-fold:

- We propose *rational adaptation*, a novel method that extends the concept of rational matrix from the classical classifying system to VLMs, to favorably focus on the fused text and visual representations at the final decision-making process during fine-tuning.
- We offer a simple and lightweight implementation for RAda, fulfilled by attaching a single attention layer at the end to learn contributions for different rational elements. This design can be seamlessly integrated into most training pipelines with minimal code.
- We conduct extensive experiments in three mainstream fine-tuning settings to evaluate RAda. We observe that RAda is a versatile fine-tuning idea that can consistently benefit the baseline, and it can obtain comparable performance against existing arts in most settings.

2. Related Work

Vision-Language models. Previous studies have demonstrated the effectiveness of using text supervision for various vision tasks [19, 24, 30, 45, 46, 49, 52]. Attributed to the large-scale training data from the web, current VLMs can achieve astonishing results on a wide spectrum of vision tasks without any fine-tuning [19, 30]. Similar to prevailing fine-tuning methods [14, 15, 33, 54], our implementation builds on the pretrained CLIP, aiming to enhance its performance across various fine-tuning strategies.

Fine-tuning strategies. We briefly review some of the fine-tuning arts by categorizing them into the following types.

The first common paradigm for fine-tuning is FFT. While it can effectively adapt models to a new distribution, the overfitting problem remains the primary concern in the liter-

ature, causing compromised robustness in diverse tasks [4, 6, 7, 23]. Several ideas have been proposed lately to mitigate the issue, including LP-FT [23] that conducts linear probing (LP) and fine tuning (FT) in a sequential manner, and weight ensemble [40] that combines weights of both the fine-tuned and pretrained models. We extend the sequential training strategy in LP-FT for RAda, where we first train the rational adapter as initialization and then finetune all parameters. Our experiments indicate that RAda can better help ease the overfitting problem than previous arts while maintaining comparable effectiveness in the training distribution.

The second idea that has been widely explored is EFT, which fixes encoders during updating. The key in EFT is to introduce new learnable parameters for adaptation. Inspired by CoOp [55], a wide range of nascent studies use learnable prompts in the vision or text encoders as additional inputs [20, 21, 33, 48, 54]. Adapter tuning also shows its effectiveness in fine-tuning [8, 12, 14, 26, 32, 36, 51], aiming to modify the original model by inserting layers to act on the input representations, unlike prompt tuning which modifies the inputs themselves. RAda can be extended to EFT by tuning only the rational adapter. Unlike previous ideas, it specifically focuses on the fused representations, aiming for a more holistic understanding of the new data.

The last is the emerging TTT that updates the model in the test phase. The idea is to leverage a self-supervised task to update the model with test samples on the fly [25, 35, 38]. The concept has been explored in some recent studies to fully uncover the zero-shot potential of CLIP [1, 33, 47, 53], with most approaches focusing on updating newly introduced parameters (*i.e.* prompts) for CLIP using the combinations of a basic entropy minimization task [38] and other hand-crafted objectiveness, such as distribution alignment [1] or advanced feedback from a larger model [53]. In line with these efforts, we update the rational adapter during test using the same entropy minimization objective.

To the best of our knowledge, this work represents a pioneering effort to explore and adapt a fine-tuning idea that can contribute effectively in all three fine-tuning settings.

3. Methodology

3.1. Preliminary

Our method is built upon a representative VLM, CLIP, which includes two parallel encoders for mapping the text and visual inputs into feature vectors. We denote the text and visual encoders as \mathcal{F}_t and \mathcal{F}_v and their pretrained parameters as θ_t and θ_v , respectively. Considering a K -class classification task in the fine-tuning process, the visual representation $\mathbf{f} \in \mathbb{R}^D$ for a given input image $\mathbf{I} \in \mathbb{R}^{C \times H \times W}$ can be simply obtained via $\mathbf{f} = \mathcal{F}_v(\mathbf{I}, \theta_v)$, and the text representations $\mathbf{h} \in \mathbb{R}^{K \times D}$ are decided upon shared prefixed text prompts \mathbf{p} for different predictions.

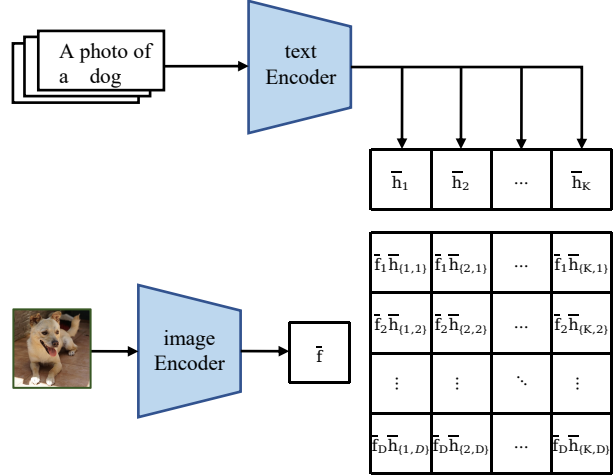


Figure 2. Rational matrix [5] in the CLIP decision-making process for a given image, where predictions (*i.e.* logits) are computed by summing each column. It fuses text and visual features and provides a fine-grained characterization of different predictions.

For instance, a commonly-used \mathbf{p} = “a photo of a” can be adopted as the prefix prompts, and each text representation $\mathbf{h}_i \in \mathbb{R}^D$ is then obtained via $\mathbf{h}_i = \mathcal{F}_t([\mathbf{p}, \mathcal{Y}_i], \theta_t)$, given $\mathcal{Y} = \{\mathcal{Y}_1, \mathcal{Y}_2, \dots, \mathcal{Y}_K\}$ the category-specific texts for all K classes, and $[\mathbf{p}, \mathcal{Y}_i] = \text{“a photo of a [class]”}$ being a direct concatenate of \mathbf{p} and \mathcal{Y}_i .

Similar to that in pretraining CLIP, the objective in fine-tuning is still contrastive learning. Take the updating of θ_v as an example, the goal is to align the input \mathbf{f} with its corresponding text description \mathbf{h}_* while away from others, which can be formulated as minimizing the following,

$$\mathcal{L}_{main} = -\log \frac{\exp(\langle \bar{\mathbf{f}}, \bar{\mathbf{h}}_* \rangle)}{\sum_{i=1}^K \exp(\langle \bar{\mathbf{f}}, \bar{\mathbf{h}}_i \rangle)}, \quad (1)$$

where $\bar{\mathbf{f}}$ and $\bar{\mathbf{h}}$ are the l_2 normalized \mathbf{f} and \mathbf{h} , \langle, \rangle returns the inner product for two vectors. In evaluation, the class k that with the largest logit $\langle \bar{\mathbf{f}}, \bar{\mathbf{h}}_k \rangle$ is considered the true prediction. Note that in the pretraining process, \mathcal{L}_{main} should be considered with another contrastive case where each text representation would correspond to different visual features, and the alignment is to ensure they are paired with the correct visual targets, same as that in Eq. (1).

3.2. Rational Matrix in CLIP

The default design in Eq. (1) cannot characterize fine-grained associations between \mathbf{f} and \mathbf{h} as it leads directly to the coarse final result. To surface the fused information from the two modalities, we suggest extending the concept of rational matrix [5] to CLIP. The rational matrix is regarded as a fine-grained characterization of the decision-making process in [5]. The concept is originally introduced within the classical classification system, where a linear classifier is involved for computing logits from the obtained

image feature \mathbf{f} : given the classifier $\mathbf{W} \in \mathbb{R}^{D \times K}$ and rational matrix $\hat{\mathbf{R}} \in \mathbb{R}^{K \times D}$, the logit value for the i -th class is $\mathbf{o}_i = \langle \mathbf{f}, \mathbf{W}_{\{i\}} \rangle = \sum_{j=1}^D \mathbf{f}_j \mathbf{W}_{\{j,i\}} = \sum_{j=1}^D \hat{\mathbf{R}}_{\{i,j\}}$.

In CLIP, the text feature $\bar{\mathbf{h}}$ plays the same role as the “classifier” \mathbf{W} when calculating the similarity (*i.e.* the logits computation process). When studying the decision-making process in CLIP (with the same \mathbf{f}), we can correspondingly depict it with a rational matrix $\mathbf{R} \in \mathbb{R}^{K \times D}$ as,

$$\mathbf{R}^\top = \begin{bmatrix} \bar{\mathbf{f}}_1 \bar{\mathbf{h}}_{\{1,1\}} & \bar{\mathbf{f}}_1 \bar{\mathbf{h}}_{\{2,1\}} & \dots & \bar{\mathbf{f}}_1 \bar{\mathbf{h}}_{\{K,1\}} \\ \bar{\mathbf{f}}_2 \bar{\mathbf{h}}_{\{1,2\}} & \bar{\mathbf{f}}_2 \bar{\mathbf{h}}_{\{2,2\}} & \dots & \bar{\mathbf{f}}_2 \bar{\mathbf{h}}_{\{K,2\}} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{\mathbf{f}}_D \bar{\mathbf{h}}_{\{1,D\}} & \bar{\mathbf{f}}_D \bar{\mathbf{h}}_{\{2,D\}} & \dots & \bar{\mathbf{f}}_D \bar{\mathbf{h}}_{\{K,D\}} \end{bmatrix}. \quad (2)$$

Similarly, the i -th logit in the CLIP result can be represented as $\sum_{j=1}^D \mathbf{R}_{\{i,j\}}$. An example of obtaining \mathbf{R} for an image is provided in Figure 2. \mathbf{R} encodes the interactions between the visual and textual features in the final decision-making process of CLIP, manifesting the fine-grained fused information at the final stage. Compared to the isolated \mathbf{f} and \mathbf{h} , it offers a more holistic understanding of the data by integrating information from both modalities.

By depicting the final-decision making process with the rational matrix, objective in Eq. (1) can be rewritten as,

$$\mathcal{L}_{main} \triangleq -\log \frac{\exp(\langle \mathbf{1}_D, \mathbf{R}_* \rangle)}{\sum_{i=1}^K \exp(\langle \mathbf{1}_D, \mathbf{R}_i \rangle)}, \quad (3)$$

where $\mathbf{1}_D$ is a D -dimensional all-one vector. *Note that reformulating Eq. (1) to Eq. (3) does not introduce any additional parameters.* We show in the following that the new formulation in Eq. (3) can provide a new perspective for adaptation in the fine-tuning process with minimum cost.

3.3. Rational Adaptation in Fine-Tuning CLIP

Unlike previous methods [14, 20, 26, 54] that use various forms to extract complementary text or visual information from the new data while processing them in isolation, we suggest a new fine-tuning idea by adapting the corresponding rational matrix, which can specifically leverage fused representations from different modalities. Specifically, we adopt a learned mask \mathbf{M} , which is with continuous values and with the same shape as \mathbf{R} , to dynamically calibrate contributions of each rational element in making the final predictions. Formally, with the adaptive \mathbf{M} , the contrastive objective in Eq. (3) is evolved into,

$$\mathcal{L}_{adapt} = -\log \frac{\exp(\langle \mathbf{1}_D, (\mathbf{M} \circ \mathbf{R})_* \rangle)}{\sum_{i=1}^K \exp(\langle \mathbf{1}_D, (\mathbf{M} \circ \mathbf{R})_i \rangle)}, \quad (4)$$

where \circ denotes the Hadamard product.

Eq. (4) can be seamlessly extended to the entropy minimization task in test-time fine-tuning [38] as well, where

Algorithm 1 PyTorch-style pseudocode for RAda in EFT.

```
# CLIP_encoder: Include vision and text encoders
# Attn: Attention layer
# I[BxHxWxC]: Batch of visual inputs
# T[KxL]: Text inputs in forms of "a photo of a []"

# extract and normalize features of each modality
f, h = CLIP_encoder(I, T) #BxD, KxD
f_n, h_n = l2_norm(f, 1), l2_norm(h, 1)

# compute the rational matrix and the mask
f_e = f_n.unsqueeze(1).repeat(1,K,1) #BxKxD
h_e = h_n.unsqueeze(0).repeat(B,1,1) #BxKxD
R = f_e * h_e
M = Attn(query=[f_e, h_e, R], key=R, value=R) #BxKxD

# obtain adapted results, baseline CLIP is with M=1
logits = torch.sum(M * R, -1) #BxK

# compute the loss
L_main = cross_entropy_loss(logits, label)
L_reg = mse_loss(M, 1)
```

the ground-truth \mathbf{h}_* is unavailable. By tuning \mathbf{M} , we can formulate the main objective in TTT as,

$$\mathcal{L}_{ttt} = -\sum_{j=1}^K p_j \log p_j, \quad \text{s.t. } p_j = \frac{\exp(\langle \mathbf{1}_D, (\mathbf{M} \circ \mathbf{R})_j \rangle)}{\sum_{i=1}^K \exp(\langle \mathbf{1}_D, (\mathbf{M} \circ \mathbf{R})_i \rangle)}. \quad (5)$$

Adaptively Learning \mathbf{M} . To adapt \mathbf{R} w.r.t different inputs, we suggest obtaining \mathbf{M} through a learnable function \mathcal{F}_m , with θ_m denoted as its parameter. Considering that the original “classifier”, *i.e.* the text embeddings $\bar{\mathbf{h}}$, does not account for the presence of other to-be-distinguished classes when making a prediction, we suggest incorporating an attention mechanism [37] for implementing \mathcal{F}_m , with the original \mathbf{R} as the input query, key, and value, enabling interactions between rationale elements from different classes. Meanwhile, to account for the original information embedded in different modalities, we suggest also using the image and text features $\{\bar{\mathbf{f}}, \bar{\mathbf{h}}\}$ as additional queries for \mathcal{F}_m , helping it to learn more nuanced relationships and dependencies across the different rational elements. In our implementation, we use different projectors for the different queries $\{\bar{\mathbf{f}}, \bar{\mathbf{h}}, \mathbf{R}\}$. The result is then computed from the average attention weight of these queries,

$$\begin{aligned} \mathbf{M}' &= \mathcal{F}_m(\{\bar{\mathbf{f}}, \bar{\mathbf{h}}, \mathbf{R}\}, \theta_m) \\ &= \left(\text{sfm}\left(\frac{Q_{\bar{\mathbf{f}}} K^\top}{\sqrt{d_K}}\right) + \text{sfm}\left(\frac{Q_{\bar{\mathbf{h}}} K^\top}{\sqrt{d_K}}\right) + \text{sfm}\left(\frac{Q_{\mathbf{R}} K^\top}{\sqrt{d_K}}\right) \right) \frac{V}{3} \quad (6) \\ \text{s.t. } Q_x &= W_x^\top x, \quad K = W_k^\top \mathbf{R}, \quad V = W_v^\top \mathbf{R}, \end{aligned}$$

where $\text{sfm}(\cdot)$ denotes the softmax function, W_x is the linear projector for the corresponding query x ; W_k and W_v are the key and value projectors. We omit the last projection layer in \mathcal{F}_m , which are zero-initialized in all three settings, and we use $\mathbf{M} = \mathbf{M}' + \mathbf{1}_{K \times D}$ for implementation. These designs ensure $\mathbf{M} = \mathbf{1}_{K \times D}$ in the first updating step, and it does not affect the initial predictions of CLIP.

Regularization for \mathbf{M} . Although fine-tuning can improve the in-domain (ID) accuracy, it inevitably diminishes the strong zero-shot capabilities of the original CLIP model. As

Table 1. **Evaluations in the FFT setting.** Results with † are reevaluated in our device, others are from FLYP [15]².

Methods	Training in Imagenet						
	ID	Im-V2	Im-R	Im-A	Im-S	Object	OOD Avg.
CLIP	68.3	61.9	77.7	50.0	48.3	55.4	58.7
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
RAda	78.1	68.3	72.8	47.3	46.9	53.8	57.8
RAda-FT	81.4	71.9	75.5	51.7	50.4	56.8	61.3

such, we suggest using a regularization for \mathbf{M} to explicitly maintain the zero-shot performance. Specifically, we introduce a smoothness constraint on the mask, designed to prevent significant deviations from its original setting (*i.e.* ensuring each element in \mathbf{R} contributes equally in the initial decision-making process), which formally gives,

$$\mathcal{L}_{reg} = \|\mathbf{M} - \mathbf{1}_{K \times D}\|^2. \quad (7)$$

Notably, \mathcal{L}_{reg} is applied only when the pretrained encoders are frozen during fine-tuning. Since in other situations, \mathbf{R} will be different from its initial state, and the pretrained information cannot be retained even with $\mathbf{M} = \mathbf{1}_{K \times D}$.

Overall Algorithm. Objectives in three different fine-tuning settings are summarized as, 1) in FFT, we minimize \mathcal{L}_{adapt} regarding all learnable parameters; 2) in EFT, we minimize $\mathcal{L}_{adapt} + \mathcal{L}_{reg}$ w.r.t θ_m ; 3) in TTT, the objective is to minimize $\mathcal{L}_{tnt} + \mathcal{L}_{reg}$ regarding θ_m . Pseudo code for a representative fine-tuning setting EFT is illustrated in Algorithm 1. As seen, the proposed method is extremely simple, as it only adds a few lines on the baseline CLIP.

4. Experiments

We conduct experiments in three fine-tuning strategies to evaluate RAda on a same NVIDIA A100 (40GB RAM). The CLIP ViT-B/16 from OpenAI is used as backbone.

4.1. RAda in FFT

We extend the training paradigm in LP-PT [23] for RAda in this setting, where we first train θ_m (*i.e.* referred to as RAda) and then jointly updates all parameters (*i.e.* referred to as RAda-FT). Note in FFT, a linear classifier is used to replace the text encoder, and its weight, initialized by the text features, can then be regarded as the evolving text information. We compare with several different FFT ideas, namely LP that only updates the new classifier, FT that updates both

²Results are diverse in the ObjectNet dataset for CLIP is mainly because FLYP uses a different version of the original dataset, as also noted in their official project <https://github.com/locuslab/FLYP>.

the image encoder and the new classifier, LP-FT [23], L2-SP [42] that ensures similarity between pretrained and fine-tuned models, and FLYP [15] that mimics the pretraining pipeline of CLIP in FFT. Cross-entropy loss is utilized, except for FLYP where the original contrastive loss is adopted.

Datasets and implementation details. We use 6 datasets for evaluations: ImageNet [10] is considered the ID dataset for fine-tuning the model, and 5 standard out-of-distribution (OOD) datasets (*i.e.*, ImageNetV2 [31], ImageNet-R [17], ImageNet-A [18], ImageNet-Sketch [39], and ObjectNet [2]) are used for evaluations. Following [15, 40], we use a batch size of 512 and train for 10 epochs. For RAda, we report results after 10 epochs; for RAda-FT, we initialize the model with the pretrained weights of RAda after 5 epochs and then perform RAda-FT for another 5 epochs. Similar to [23], we use diverse learning rates for RAda and RAda-FT (*i.e.* 0.004 and 0.000004). Other settings, such as optimizers, weight decay, *etc.* are inherited from [40], same as [15]. Please see our supplementary material for details.

Experimental results. We list the results in Table 1. Similar to LP, RAda underperforms FT on ID data but shows better results across most OOD datasets, except for ImageNetV2 which is close to ID. This aligns with prior theory [23] that training later layers enhances generalization by preserving pretrained features, especially when ID and OOD distributions are substantially distinct. Since RAda operates in even later layers than the classifier, the feature-preserving theory can be further validated when comparing RAda and LP. Moreover, we notice that RAda-FT improves performance in all OOD datasets for FT and outperforms prior arts [15, 23]. This is because RAda-FT can better balance the tradeoff between overfitting and retaining both pretrained text and visual features, a benefit cannot be achieved by focusing different modalities in isolation. These findings underscore the effectiveness of prioritizing final decision-making process in VLMs as a targeted FFT strategy.

4.2. RAda in EFT

This experiment aims to evaluate whether the adapted rational matrix can contribute when the encoders are fixed during fine-tuning. Besides the baseline CLIP model, we compare our idea with some recent arts that specifically designed for EFT: CLIP-Adapter [14], CoOp [55], Co-CoOp [54], ProGrad [56], KgCoOp [44], MaPLe [21], DePT [50] on the basis of MaPLe, and MMA [43]. Results are directly cited from the paper except for [14], which is reimplemented in our device using the provided code.

Datasets and implementation details. We test the methods with the base-to-novel generalization setting. A total of 11 datasets are utilized, including, 2 used for classification on generic objects, *i.e.* ImageNet and Caltech101 [13]; 5 used for fine-grained classification, *i.e.* OxfordPets [29], StanfordCars [22], Flowers102 [28], Food101 [3], and FGV-

Table 2. Comparison with EFT methods in the base-to-new setting. All methods are learned from the base classes with 16 shots. RAda + Prompt indicates combining RAda with vision and text prompt tunings [20, 55], which are also utilized in [21]. RAda + Adapter indicates combining RAda with adapters within both encoders (the foundation skill also adopted in [43]). ‘HM’ denotes harmonic mean.

Methods	Average			ImageNet			Caltech101			OxfordPets		
	Base	New	HM	Base	New	HM	Base	New	HM	Base	New	HM
CLIP [30]	69.34	74.22	71.70	72.43	68.14	70.22	96.84	94.00	95.40	91.17	97.26	94.12
CoOp [55]	82.69	63.22	71.66	76.47	67.88	71.92	98.00	89.81	93.73	93.67	95.29	94.47
CoCoOp [54]	80.47	71.69	75.83	75.98	70.43	73.10	97.96	93.81	95.84	95.20	97.69	96.43
ProGrad [56]	82.48	70.75	76.16	77.02	66.66	71.46	98.02	93.89	95.91	95.07	97.63	96.33
CLIP-Adapter [14]	80.83	72.93	76.67	75.78	67.60	71.45	98.32	93.56	95.88	93.73	95.97	94.84
KgCoOp [44]	80.73	73.60	77.00	75.83	69.96	72.78	97.72	94.39	96.03	94.65	97.76	96.18
MaPLe [21]	82.28	75.14	78.55	76.66	70.54	73.47	97.74	94.36	96.02	95.43	97.76	96.58
DePT [50] + MaPLe	84.85	74.82	79.52	77.87	70.23	73.85	98.53	95.03	96.75	95.03	97.83	96.41
MMA [43]	83.20	76.80	79.87	77.31	71.00	74.02	98.40	94.00	96.15	95.40	98.07	96.72
RAda	82.16	74.14	77.94	75.50	68.41	71.78	98.39	94.32	96.31	94.31	96.03	95.16
RAda + Prompt	84.18	75.61	79.67	77.11	68.29	72.44	98.01	94.65	96.31	96.16	97.87	97.01
RAda + Adapter	84.32	76.25	80.08	77.96	70.23	73.89	98.06	93.56	95.76	95.43	97.99	96.69

Methods	StanfordCars			Flowers102			Food101			FGVCAircraft		
	Base	New	HM	Base	New	HM	Base	New	HM	Base	New	HM
CLIP [30]	63.37	74.89	68.65	72.08	77.80	74.83	90.10	91.22	90.66	27.19	36.29	31.09
CoOp [55]	78.12	60.40	68.13	97.60	59.67	74.06	88.33	82.26	85.19	40.44	22.30	28.75
CoCoOp [54]	70.49	73.59	72.01	94.87	71.75	81.71	90.70	91.29	90.99	33.41	23.71	27.74
ProGrad [56]	77.68	68.63	72.88	95.54	71.87	82.03	90.37	89.59	89.98	40.54	27.57	32.82
CLIP-Adapter [14]	73.64	71.50	72.55	96.77	71.56	82.28	90.16	90.96	90.56	35.65	32.27	33.87
KgCoOp [44]	71.76	75.04	73.36	95.00	74.73	83.65	90.50	91.70	91.09	36.21	33.55	34.83
MaPLe [21]	72.94	74.00	73.47	95.92	72.46	82.56	90.71	92.05	91.38	37.44	35.61	36.50
DePT [50] + MaPLe	80.93	71.73	76.06	98.03	73.17	83.79	90.33	91.53	90.93	44.53	32.80	37.78
MMA [43]	78.50	73.10	75.70	97.77	75.93	85.48	90.13	91.30	90.71	40.57	36.33	38.33
RAda	76.29	73.73	74.99	95.63	72.77	82.65	90.01	90.55	90.28	38.90	33.65	36.09
RAda + Prompt	79.15	73.93	76.45	97.25	69.86	81.31	90.33	91.17	90.75	41.83	35.03	38.13
RAda + Adapter	79.36	73.16	76.13	97.74	75.32	85.08	90.35	91.49	90.92	41.72	38.09	39.82

Methods	SUN397			DTD			EuroSAT			UCF101		
	Base	New	HM	Base	New	HM	Base	New	HM	Base	New	HM
CLIP [30]	69.36	75.35	72.23	53.24	59.90	56.37	56.48	64.05	60.03	70.53	77.50	73.85
CoOp [55]	81.16	75.08	78.00	80.32	56.52	66.35	79.43	74.26	76.76	84.13	72.96	78.15
CoCoOp [54]	79.74	76.86	78.27	77.01	56.00	64.85	87.49	60.04	71.21	82.33	73.45	77.64
ProGrad [56]	81.26	74.17	77.55	77.35	52.35	62.45	90.11	60.89	72.67	84.33	74.94	79.35
CLIP-Adapter [14]	81.16	75.08	78.00	80.32	56.52	66.35	79.43	74.26	76.76	84.13	72.96	78.15
KgCoOp [44]	80.29	76.53	78.36	77.55	54.99	64.35	85.64	64.34	73.48	82.89	76.67	79.65
MaPLe [21]	80.82	78.70	79.75	80.36	59.18	68.16	94.07	73.23	82.35	83.00	78.66	80.77
DePT [50] + MaPLe	82.90	76.40	79.52	83.87	59.93	69.91	94.43	76.23	84.36	86.87	78.10	82.25
MMA [43]	82.27	78.57	80.38	83.20	65.63	73.38	85.46	82.34	83.87	86.23	80.03	82.20
RAda	80.38	75.97	78.11	79.17	58.70	67.42	90.40	74.72	81.82	84.80	76.74	80.57
RAda + Prompt	82.38	77.30	79.76	83.28	60.87	70.33	94.27	84.69	89.22	86.21	78.15	81.98
RAda + Adapter	82.58	78.77	80.63	82.06	67.15	73.86	96.48	74.69	84.20	85.78	78.26	81.85

CAircraft [27]; an scene recognition dataset SUN397 [41]; a action recognition dataset UCF101 [34]; a texture classification dataset DTD [9]; and a satellite imagery recognition EuroSAT [16]. Batch size, learning rate, and epoch are fixed as 1, 0.0009, and 13 for all datasets, and we use 16 shots per class for the source data, same as [54].

Experimental results. Results in Table 2 show that RAda increases the average base accuracy by nearly 13pp for the baseline CLIP without compromising its performance in unseen classes—a tradeoff observed in many other methods [14, 44, 54–56] where base accuracies are improved at

the cost of novel class performances. These results validate the effectiveness of RAda in fast adaptation to new data while preserving the generalizability of the original CLIP.

While RAda alone does not achieve state-of-the-art performance, it remains highly competitive: among the compared arts, RAda is outperformed only by arts that specifically modify intermediate representations [21, 43, 50]. When also interfering the intermediate representations within RAda, we show its performance can be further boosted. For instance, adopting prompt tuning for RAda (RAda + Prompt) achieves better performance than that uti-

Table 3. **Comparisons with representative TTT methods regarding DG performance in four distribution shifts.** Here “pretrained” denotes whether the model is pretrained in ImageNet. RAda shows comparable effectiveness against arts specialized in the TTT task.

	pretrained	ImageNet V2	ImageNet Sketch	ImageNet A	ImageNet R	OOD Avg.
CLIP	✗	60.86	46.09	47.87	73.98	57.20
TPT [33]	✗	64.35	47.94	54.77	77.06	60.81
CoOp [55]+TPT	✓	66.83	49.29	57.95	77.27	62.84
CoCoOp [54]+TPT	✓	64.85	48.27	58.47	78.65	62.61
MaPLe [21]+TPT	✓	64.87	48.16	58.08	78.12	62.31
PromptAlign [1]	✓	65.29	50.23	59.37	79.33	63.55
RAda	✗	64.10	49.36	61.17	79.35	63.50
RAda [†]	✓	65.10	49.45	62.72	79.75	64.26

lized in MaPLe, while combining adapter tuning for RAda (RAda + Adapter) yields best average accuracy. This synergy with complementary strategies highlights RAda is orthogonal to existing EFT paradigms (*e.g.* prompt or adapter tuning). Collectively, these results affirm RAda’s potential as a competitive approach in the EFT setting.

4.3. RAda in TTT

Different from other strategies, TTT can only access the unlabeled test data in updating. We compare RAda with two recent methods, *i.e.* TPT [33] and PromptAlign [1], within this setting. Both these two methods are developed based on the prompt tuning paradigm. Specifically, TPT extends CoOp by updating the text prompts with the entropy minimization objective, and PromptAlign extends the idea in [21] by including an additional distribution alignment regularization to refine both text and visual prompts.

Datasets and implementation details. Same as previous works [1, 33], we use the 4 OOD datasets (*i.e.* ImageNetV2, ImageNet-R, ImageNet-A, and ImageNet-Sketch) for evaluation. For every test sample, we obtain 63 of its augmented view using the same augmentation strategies in [1] to form a batch of 64 samples, among which, we select the top 10% confident predictions with the lowest entropy and compute the entropy loss in Eq. (5) for the sub-batch. The offline TTT updating strategy [35] is adopted where the weights are initialized to the original state for each sample, so that the order of the arrived data does not affect the result. The learning rate is fixed as 0.0008 for all datasets, and we perform three updating steps for each of the test sample.

Experimental results. As shown in Table 3³, RAda enhances the baseline CLIP across all evaluated datasets, outperforming the naive TPT in 3 datasets and leading the average accuracy by 2.7pp. These observations validate the effectiveness of focusing the final decision-making process for adaptation in test-time. When compared to the recent PromptAlign, RAda demonstrates strengths on half of the evaluated datasets with comparable average results, which is achieved without leveraging pretrained information in

ImageNet. When using the same pretrained information, we observe that the performance of RAda can be further enhanced, leading all compared arts in average performance. Combined with the demonstrated efficiency of RAda in Table 5, these results validate RAda as a strong competitor for the TTT application, even against task-specific methods.

5. Analysis

5.1. Ablation Studies

We evaluate the effectiveness of our designs using the base-to-new generalization setting in EFT, where the settings are the same as that detailed in Sec. 4.2. Please refer to our supplementary material for more ablation studies.

Effectiveness of the regularization term. To assess the impact of the regularization term on the model performance, we evaluate RAda with and without adding \mathcal{L}_{reg} in the overall objective. As shown in the 2nd row in Table 4, the base accuracy for the baseline CLIP can be improved across both settings, indicating that the regularization term does not degrade performance on familiar classes. Meanwhile, we observe the inclusion of \mathcal{L}_{reg} significantly improves the generalization performance, *i.e.* the accuracy on novel classes, achieving an improvement of 2.5pp. These results highlight the importance of the regularization term in maintaining the training effectiveness of RAda without compromising the superior zero-shot capability of CLIP.

Different settings for implementing \mathcal{F}_m . We use a multi query attention layer for implementing \mathcal{F}_m . In this section, we assess the impact of adopting different settings for \mathcal{F}_m . Namely, we first try using an MLP layer to replace the attention-based rational adapter (*i.e.* MLP for \mathcal{F}_m), and then use the following query settings in the final attention layer: rational matrix \mathbf{R} , and its combination with either the image features $\bar{\mathbf{f}}$ or text features $\bar{\mathbf{h}}$.

As listed in 3rd-6th rows from Table 4, although using an MLP for \mathcal{F}_m can improve the baseline, it performs inferior to using the attention-based design. This is mainly because the attention layer encourages the interactions of different to-be-distinguished classes, while the MLP layer can only process different classes independently. These results jus-

³Results for some datasets are different for CLIP in Table 1 and 3 is due to the different prefixed text prompts \mathbf{p} used in these two settings.

Table 4. Comparisons of RAda with its different variants.

Variants	Base acc.	New acc.	HM
Baseline	69.34	74.22	71.70
W/O \mathcal{L}_{reg}	81.38	71.58	76.16
MLP for \mathcal{F}_m	77.51	73.18	75.28
query= $\{\mathbf{R}\}$	81.99	73.89	77.72
query= $\{\bar{\mathbf{h}}, \mathbf{R}\}$	81.93	74.07	77.80
query= $\{\bar{\mathbf{f}}, \mathbf{R}\}$	82.03	73.93	77.76
(a) \mathcal{F}_t + attn	82.56	69.62	75.53
(b) \mathcal{F}_v + attn	81.96	71.18	76.19
(a) + (b)	82.09	72.46	76.97
RAda (query= $\{\bar{\mathbf{h}}, \bar{\mathbf{f}}, \mathbf{R}\}$)	82.16	74.14	77.94

tify our motivation of using an attention layer to implement \mathcal{F}_m . Meanwhile, we note that incorporating additional information from either text or visual modality improves the performance than rely solely on the rational matrix to act as query, key, and value. This is because fusing $\bar{\mathbf{f}}$ and $\bar{\mathbf{h}}$ may obscure their specific patterns, resulting \mathbf{R} to not include all information. In comparison, our multi query setting offers best results in both base and novel class accuracies, underscoring the importance of leveraging all available information in the rational matrix calibration process.

Effectiveness of the fused information. To verify if leveraging fused information is superior to ideas that consider the different modalities in isolation, we compare RAda with variants that attach attention layers to encoders, which modify the original image or text features through a learned mask from the the attached attention layer. Specifically, three variants are compared: “ \mathcal{F}_t + attn” and “ \mathcal{F}_v + attn”, where an attention layer is attached at the end of the text and visual encoders, respectively; and a combined variant that applies separate attention layers to each encoder.

As seen in 7th-9th rows in Table 4, adding an attention layer enhances adaptation to the training distribution, whether attached to the vision or text encoder, as the baseline accuracies increase across all three variants. Notably, the variant with attention layers attached to both the text and visual encoders performs the best, likely due to the incorporation of adapted information from both modalities. However, despite using the same mask regularization, attaching attention layers solely to the encoders does not preserve the superior zero-shot ability of the original CLIP, as all these variants show marked declines in novel class accuracy compared to RAda. These results verifies the superiority of using the final fused information in fine-tuning compared with these variants that process different modalities in isolation. In the supplementary material, we further provide theoretical explanations for this observation.

5.2. Comparisons with MaPLE [21]

The closest conceptual counterpart to RAda is MaPLE [21], which similarly seeks to leverage fused information during fine-tuning. Key differences between these two methods are

Table 5. Efficiency comparisons with MaPLE [21] for a single updating step in fine-tuning CLIP with a batch size of 1.

Setting	Method	FPS(\uparrow)	Memory(\downarrow)	GFLOPs(\downarrow)
EFT	MaPLE	27.20	1.09GB	206.19
	RAda	50.11	0.49GB	17.67
TTT	MaPLE	34.51	1.31GB	191.75
	RAda	68.39	0.61GB	17.26

as follows. (1) **Efficiency.** MaPLE updates intermediate prompts across layers, incurring high memory and computational costs. In contrast, RAda’s adaptation is confined to the final output stage, requiring minimal memory and computational resources. Efficiency comparisons in Table 5 solidifying its lightweight advantage over MaPLE. (2) **Effectiveness.** The two methods exhibit divergent performance across the three mainstream fine-tuning settings. Specifically, while MaPLE achieves a 0.6pp average advantage over RAda in EFT (78.55 vs. 77.94), it is not applicable in the FFT setting, where RAda achieves leading performance, and it underperforms RAda by nearly 2pp in the TTT setting (62.31 vs. 64.26). These results highlight that the strengths of each method vary depending on the fine-tuning scenario and that RAda exhibits greater versatility, delivering consistently competitive results in most settings. (3) **Applicability.** Inheriting limitations from vision prompt tuning [20], MaPLE is restricted to the transformer-based image encoders. In contrast, RAda thrives with arbitrary encoder structures (see results in the supplementary material), further demonstrating its encoder-agnostic advantage.

6. Discussion and Conclusion

Future works. While RAda has demonstrated its effectiveness in different fine-tuning settings, a promising extension is to apply it in the pretraining phase, where a richer fused representation from massive data can be utilized. Meanwhile, besides the classification task, the literature could also consider exploring RAda on other downstream applications within the VLM contexts, such as image captioning and visual question answering, where a fine-grained understanding of relationships between different modalities is essential for improving the performance.

Conclusion. This paper proposes a new rational adaptation method to effectively focus on the final decision-making process of CLIP, aiming to explicitly leverage fused representations from different modalities for improved performance. The idea is achieved by a simple implementation that attaches an additional attention layer at the end to learn a mask that can adaptively decide contributions for different rational elements. Through comprehensive experiments across various settings, we find the proposed idea can serve as a versatile fine-tuning strategy, consistently benefiting the baseline and competing favorably against existing arts.

Acknowledgment.

This research is supported by the MBZUAI-WIS Joint Program for AI Research.

References

- [1] Jameel Abdul Samadh, Mohammad Hanan Gani, Noor Hussein, Muhammad Uzair Khattak, Muhammad Muzammal Naseer, Fahad Shahbaz Khan, and Salman H Khan. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. In *NeurIPS*, 2023. 3, 7
- [2] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *NeurIPS*, 2019. 5
- [3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101—mining discriminative components with random forests. In *ECCV*, 2014. 5
- [4] Liang Chen, Yong Zhang, Yibing Song, Ying Shan, and Lingqiao Liu. Improved test-time adaptation for domain generalization. In *CVPR*, 2023. 3
- [5] Liang Chen, Yong Zhang, Yibing Song, Anton Van Den Hengel, and Lingqiao Liu. Domain generalization via rationale invariance. In *ICCV*, 2023. 1, 2, 3
- [6] Liang Chen, Yong Zhang, Yibing Song, Zhiqiang Shen, and Lingqiao Liu. Lfme: A simple framework for learning from multiple experts in domain generalization. *NeurIPS*, 2024. 3
- [7] Liang Chen, Yong Zhang, Yibing Song, Zhen Zhang, and Lingqiao Liu. A causal inspired early-branching structure for domain generalization. *IJCV*, 132(9):4052–4072, 2024. 3
- [8] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. In *ICLR*, 2023. 1, 3
- [9] Mircea Cimpoi, Subhansu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 6
- [10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5
- [11] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2
- [12] Constantin Eichenberg, Sidney Black, Samuel Weinbach, Letitia Parcalabescu, and Anette Frank. Magma—multimodal augmentation of generative models through adapter-based finetuning. In *EMNLP Findings*, 2022. 1, 3
- [13] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPRW*, 2004. 5
- [14] 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. 1, 2, 3, 4, 5, 6
- [15] 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, 5
- [16] 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(7):2217–2226, 2019. 6
- [17] 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. *ICCV*, 2021. 5
- [18] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, 2021. 5
- [19] 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. 1, 2
- [20] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, 2022. 1, 2, 3, 4, 6, 8
- [21] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *CVPR*, 2023. 2, 3, 5, 6, 7, 8
- [22] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCVW*, 2013. 5
- [23] 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. 1, 2, 3, 5
- [24] Yanguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *ICLR*, 2022. 2
- [25] Yuejiang Liu, Parth Kothari, Bastien Van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. Ttt++: When does self-supervised test-time training fail or thrive? In *NeurIPS*, 2021. 3
- [26] Teli Ma, Shijie Geng, Mengmeng Wang, Jing Shao, Jiasen Lu, Hongsheng Li, Peng Gao, and Yu Qiao. A simple long-tailed recognition baseline via vision-language model. *arXiv:2111.14745*, 2021. 1, 3, 4
- [27] Subhansu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv:1306.5151*, 2013. 6
- [28] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008. 5
- [29] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012. 5
- [30] 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. 1, 2, 6
- [31] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019. 5
- [32] Erica K Shimomoto, Edison Marrese-Taylor, Hiroya Takamura, Ichiro Kobayashi, Hideki Nakayama, and Yusuke Miyao. Towards parameter-efficient integration of pre-trained language models in temporal video grounding. In *ACL Findings*, 2023. 3
- [33] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *NeurIPS*, 2022. 1, 2, 3, 7
- [34] K Soomro. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv:1212.0402*, 2012. 6
- [35] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *ICML*, 2020. 3, 7
- [36] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *CVPR*, 2022. 1, 3
- [37] A Vaswani. Attention is all you need. In *NeurIPS*, 2017. 4
- [38] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *ICLR*, 2021. 3, 4
- [39] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, 2019. 5
- [40] 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, 5
- [41] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 6
- [42] LI Xuhong, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In *ICML*, 2018. 5
- [43] Lingxiao Yang, Ru-Yuan Zhang, Yan Chen Wang, and Xiaohua Xie. Mma: Multi-modal adapter for vision-language models. In *CVPR*, 2024. 5, 6
- [44] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *CVPR*, 2023. 5, 6
- [45] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. In *ICLR*, 2022. 2
- [46] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv:2111.11432*, 2021. 2
- [47] Maxime Zanella and Ismail Ben Ayed. On the test-time zero-shot generalization of vision-language models: Do we really need prompt learning? In *CVPR*, 2024. 3
- [48] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Unified vision and language prompt learning. *arXiv:2210.07225*, 2022. 1, 3
- [49] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *CVPR*, 2022. 2
- [50] Ji Zhang, Shihan Wu, Lianli Gao, Heng Tao Shen, and Jingkuan Song. Dept: Decoupled prompt tuning. In *CVPR*, 2024. 5, 6
- [51] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. In *ECCV*, 2022. 1, 3
- [52] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, 2022. 2
- [53] Shuai Zhao, Xiaohan Wang, Linchao Zhu, and Yi Yang. Test-time adaptation with CLIP reward for zero-shot generalization in vision-language models. In *ICLR*, 2024. 3
- [54] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022. 1, 2, 3, 4, 5, 6, 7
- [55] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022. 1, 2, 3, 5, 6, 7
- [56] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *ICCV*, 2023. 5, 6